MA4SP Stochastic Processes (2020)
Lecture Notes

J. Bröcker

May 10, 2021

Contents

Chapter 1

Reminder on probability and integration

1.1 Sigma algebras and probability measures

Video `Lecture2_V1` starts about here. Videos `Lecture2_V*` cover Sections 1.1 to 1.3.

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

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

The notation $A^{\complement}$ 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 1.1. Let Ω be a set. A system $\mathcal{A}$ of subsets of Ω is called an *algebra* if

1. $\varnothing \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^{\complement} \in \mathcal{A}$.
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 1.2. Let $\mathcal{A}$ be an algebra. A function $\mathbb{P} : \mathcal{A} \longrightarrow [0,1]$ is a probability if it satisfies

1. *Normalisation:* $\mathbb{P}(\Omega) = 1$
2. *Additivity:* If $A_1, \ldots, A_n \in \mathcal{A}$, with $A_i \cap A_j = \varnothing$ for $i \neq j$, then $\sum_{k=1}^{n} \mathbb{P}(A_k) = \mathbb{P}(\bigcup_{k=1}^{n} A_k)$.
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 $\mathbb{P}(A_k) \to 0$ 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 1.2). It is possible to construct examples of a probability on an algebra that is not continuous at $\varnothing$. Note that a probability satisfies $\mathbb{P}(\varnothing) = 0$ (Exercise 1.2).

Definition 1.3. A pair $(\Omega, \mathcal{A})$ with Ω a set and $\mathcal{A}$ a sigma algebra is called a *measurable space*. A triple $(\Omega, \mathcal{A}, \mathbb{P})$ with Ω a set, $\mathcal{A}$ a sigma algebra, and $\mathbb{P}$ 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 $\mathbb{P}$ just on an algebra (examples later).

Definition 1.4. 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 1.1 you will show that this concept is well defined.

Theorem 1.5 (The Measure Extension Theorem, also known as MET or Hahn-Carathéodory theorem)**.** *Let $\mathcal{A}$ be an algebra and $\mathbb{P}$ a probability on $\mathcal{A}$. Then there exists a unique probability $\tilde{\mathbb{P}}$ on $\sigma(\mathcal{A})$ with $\tilde{\mathbb{P}}|_{\mathcal{A}} = \mathbb{P}|_{\mathcal{A}}$. 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 $\tilde{\mathbb{P}}(A \triangle \bigcup_{k=1}^{n} A_k) \leq \epsilon$.*

Sketch of a proof, see e.g. [6]. For any $Y \subset \Omega$, put $\mathbb{P}^*(Y) = \inf \sum_{k=1}^{\infty} \mathbb{P}(A_k)$, inf taken over $A_1, A_2, \cdots \in \mathcal{A}$, with $Y \subset \bigcup_k A_k$. Now

1. $\mathbb{P}^*|_{\mathcal{A}} = \mathbb{P}|_{\mathcal{A}}$ ("$\leq$" 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 $\mathbb{P}^*(E) \geq \mathbb{P}^*(E \cap A) + \mathbb{P}^*(E \setminus A)$. One then proves that $\mathcal{M}$ is a $\sigma-$algebra with $\mathcal{M} \supset \mathcal{A}$.

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

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

□

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

Theorem 1.6 (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 $\mathbb{P}, \mathbb{Q}$ be two probabilities on $\sigma(\mathcal{A})$, the sigma algebra generated by $\mathcal{A}$. Then if $\mathbb{P}(A) = \mathbb{Q}(A)$ for any set $A \in \mathcal{A}$, they agree on $\sigma(\mathcal{A})$.*

For a proof see [1], 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 [7], Chapter 7.

Theorem 1.7 (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 [7] for a somewhat more general statement). It now follows from Theorem 1.5 that λ can be extended to a probability on $\sigma(\mathcal{A})$, which is the Borel algebra (see Definition 1.8).*

Exercises for Section 1.1

Exercise 1.1. 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 1.4 makes sense, i.e. there exist sigma algebras containing $\mathcal{A}$, and among these there exists a smallest possible one.

Exercise 1.2. Let Ω be a set, $\mathcal{A}$ an algebra, $\mathbb{P} : \mathcal{A} \to [0,1]$ a set function satisfying properties 1 and 2 in Definition 1.2.

1. Show that $\mathbb{P}(\emptyset) = 0$.
2. Show that property 3 in Definition 1.2 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 $\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \mathbb{P}(\bigcup_k A_k)$.
3. Show that property 3 in Definition 1.2 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 $\mathbb{P}(A_k) \to \mathbb{P}(A)$ for $k \to \infty$.
4. Show that property 3 in Definition 1.2 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 $\mathbb{P}(A_k) \to \mathbb{P}(A)$ for $k \to \infty$.
5. Show that for any series $A_1, A_2, \ldots$ of disjoint sets in $\mathcal{A}$, we have $\mathbb{P}(A_n) \to 0$ (in fact, $\mathbb{P}(A_n)$ must be summable).

1.2 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 $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.

Definition 1.8.
1. On $\mathbb{R}$ we define the Borel-algebra $\mathcal{B}$ as the smallest $\sigma-$algebra containing all open sets (see 1.4).
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 1.4).

Theorem 1.9.
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 1.8. $\inf_k f_{n+k}$ 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 $\int f \, \mathrm{d}\mathbb{P}$ 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 $x_n \uparrow x$ means that $x = \sup_n x_n$. Note that x might be infinite, but if it is finite, we have $x = \lim_{n\to\infty} x_n$. (We stress that the limit of a sequence is *always* finite per definition.) If $\{x_n\}$ is not increasing though, then $x_n \uparrow x$ is meaningless. In general, we will write $x_n \to x$ as a shorthand for $x = \lim_n x_n$.

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 [9, 6, 2, 3] for details.

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

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

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

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

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. If $f_1, \ldots, f_k$ are simple functions and $\phi : \mathbb{R}_k \to \mathbb{R}$ an arbitrary function, then the function $\omega \to \phi(f_1(\omega), \ldots, f_k(\omega))$ is simple. In particular, 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

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

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

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 step 4), then $\{\int f_n \, \mathrm{d}\mathbb{P}\}$ is an increasing sequence of real numbers and

$$\sup_n \int f_n \, \mathrm{d}\mathbb{P} = \sup_g \int g \, \mathrm{d}\mathbb{P}, \tag{1.1}$$

   where "$\sup_g$" is over all simple g with $f \geq g$. This will be proved in exercise 1.5. We define $\int f \ \mathrm{d}\mathbb{P}$ as either side of Equation (1.1). This might be a nonnegative real number or ∞. But if $\int f \, \mathrm{d}\mathbb{P} < \infty$, then $\int f_n \, \mathrm{d}\mathbb{P} \to \int f \ \mathrm{d}\mathbb{P}$ 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

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

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 $\int |f| \, \mathrm{d}\mathbb{P}$ is always well defined for any random variable f, and f is integrable if and only if $\int |f| \, \mathrm{d}\mathbb{P} < \infty$.

Lemma 1.10 (Properties of the integral)**.** *The integral enjoys the properties in step (6) if both $\int |f| d\mathbb{P} < \infty$ and $\int |g| \, \mathrm{d}\mathbb{P} < \infty$.*

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 $\int f \, \mathrm{d}\mathbb{P} \leq \int g \, \mathrm{d}\mathbb{P}$. To prove the additivity in the general case, observe first that $|f + g| \leq |f| + |g|$ and hence $\int |f + g| \, \mathrm{d}\mathbb{P} < \infty$ 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

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

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

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

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

Interchange of integral with a.s. limits

An important property of the integral is the relatively nice behaviour of the integral under pointwise limits (as opposed to using the Riemann integral).

Theorem 1.11 (Monotone Convergence)**.** *Suppose $\{f_n, n \in \mathbb{N}\}$ is an increasing sequence of nonnegative random variable, and $f_n \uparrow f$. Then*

$$\int f_n \, \mathrm{d}\mathbb{P} \uparrow \int f \, \mathrm{d}\mathbb{P}. \tag{1.2}$$

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,

$$g_n \leq f_n \leq f \qquad \text{for all } n. \tag{1.3}$$

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 (1.3) 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 (1.2) 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 $\int f_n d\mathbb{P} \geq 0$ rather than $f_n \geq 0$ for the theorem to hold, see [3].

Theorem 1.12 (Fatou Lemma)**.** *For $\{f_n\}$ a sequence of nonnegative random variables we have*

$$\int \liminf f_n \, \mathrm{d}\mathbb{P} \leq \liminf \int f_n \, \mathrm{d}\mathbb{P}. \tag{1.4}$$

Before proving this, a little example for illustration.

Example 1.13. 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 $\mathbb{P}(A) = \int_A \mathrm{d}x$. The integral with respect to $\mathbb{P}$ is of course the standard Lebesgue integral on the unit interval (or the Riemann integral if the integrand is continuous). Define

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

Now $\liminf f_n = \lim f_n = 0$, and hence the left hand side of Equation (1.4) is zero. But $\int f_n(x) \, \mathrm{d}x = 1$ and therefore $\liminf \int f_n(x) \quad \mathrm{d}x = 1$, 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 1.11) 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

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

so we take the inf over l and obtain

$$\int \inf_k f_{n+k} \, \mathrm{d}\mathbb{P} \leq \inf_k \int f_{n+k} \, \mathrm{d}\mathbb{P}. \tag{1.5}$$

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

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

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

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 1.13 is neither bounded nor monotone).

Theorem 1.14 (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 $\int f_n \, \mathrm{d}\mathbb{P} \longrightarrow \int f \, \mathrm{d}\mathbb{P}$.*

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 $\int |f| \, \mathrm{d}\mathbb{P} \leq C$, 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)

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

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

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

or after multiplying with -1:

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

In summary, we have shown that

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

completing the proof. □

Definition 1.15 (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 $\mathbb{P}(\Omega_1) = 1$. (One can check that this is indeed an equivalence relation.)

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

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

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

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

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 1.6, $f = 0$ 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 1.16 (The space L_1)**.**

1. The space of integrable functions (or strictly speaking, their equivalence classes) with the norm $\|.\|_1$ is denoted as $L_1(\Omega, \mathcal{A}, \mathbb{P})$ 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 1.17 (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 1.17 and also Definition 1.16,2, 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 1.2

Exercise 1.3. In this exercise, we fill in some details to Section 1.2. 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 1.4 to prove the remark after Definition 1.8: f is a random variable if $\{\omega \in \Omega; f(\omega) > a\} \in \mathcal{A}$ for any $a \in \mathbb{R}$.

Exercise 1.4. 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 1.5. 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 $\int f_n \, \mathrm{d}\mathbb{P}$. Therefore $c = \lim_n \int f_n \, \mathrm{d}\mathbb{P}$ exists. Show that the following statement implies item (7): If g is simple and $g \leq f$, then
$$\int g \, \mathrm{d}\mathbb{P} \leq c. \tag{1.6}$$
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:
$$\int f_n \, \mathrm{d}\mathbb{P} \geq \int f_n \cdot \mathbb{1}_{M_n} \, \mathrm{d}\mathbb{P} \geq \int g \cdot \mathbb{1}_{M_n} \, \mathrm{d}\mathbb{P} - \epsilon \mathbb{P}(M_n) \tag{1.7}$$

4. Use sigma additivity to establish that $\mathbb{P}(M_n) \to 1$, and that $\int g \cdot \mathbb{1}_{M_n} \, \mathrm{d}\mathbb{P} \to \int g \, \mathrm{d}\mathbb{P}$ (remember that g is simple). Using this in Equation (1.7) gives
$$c = \lim_n \int f_n \, \mathrm{d}\mathbb{P} \geq \int g \, \mathrm{d}\mathbb{P} - \epsilon$$
for any ϵ, establishing (1.6).

Exercise 1.6. Show that if f is a nonnegative random variable with $\int f \, \mathrm{d}\mathbb{P} = 0$, then $f = 0$ *almost surely*, that is $f(\omega) = 0$ for all ω in a set Ω_1 with $\mathbb{P}(\Omega_1) = 1$. Hint: Consider the sets $A_n = \{\omega : f(\omega) > 1/n\}$ and show that $n \cdot f \geq \mathbb{1}_{A_n}$ to get an upper bound on $\mathbb{P}(A_n)$. What can you now say about $\cup_{n=1}^{\infty} A_n$?

Exercise 1.7. In this exercise, we will introduce the concept of densities. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Let f be a nonnegative random variable, and suppose that $\int f \, \mathrm{d}\mathbb{P} = 1$. On $\mathcal{A}$, define the set function F by
$$F(A) = \int \mathbb{1}_A \cdot f \, \mathrm{d}\mathbb{P}.$$

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 $\mathbb{P}(A) = 0$ 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 1.6, 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_-.

1.3 Transformations

This short chapter is devoted to transformations, the pushforward 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 1.8 are just a special case of this, namely with $(\Omega_2, \mathcal{A}_2) = (\mathbb{R}, \mathcal{B})$. Let $\mathbb{P}$ be a measure on $(\Omega_1, \mathcal{A}_1)$. Then the formula $T_*\mathbb{P}(A) := \mathbb{P}(T^{-1}(A))$ for all $A \in \mathcal{A}_2$ defines a probability $T_*\mathbb{P}$ on $(\Omega_2, \mathcal{A}_2)$ called the *pushforward* of $\mathbb{P}$ under T. That the pushforward is indeed a probability will be proved in Exercise 1.8.

Theorem 1.18 (Transformation formula)**.** *If $f : (\Omega_2, \mathcal{A}_2) \to (\mathbb{R}, \mathcal{B})$ is a random variable, either positive or integrable w.r.t. $T_*\mathbb{P}$, then*

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

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

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

On the right hand side we obtain

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

establishing the transformation formula for simple functions. The rest of the proof is covered in Exercise 1.9 □

Exercises for Section 1.3

Exercise 1.8. This exercise fills in several details to the beginning of Section 1.3 in preparation of the transformation formula 1.18. Let $(\Omega_k, \mathcal{A}_k), k = 1, 2$ be measurable spaces, $\mathbb{P}$ 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 $T_*\mathbb{P}$ defined by $T_*\mathbb{P}(A) := \mathbb{P}(T^{-1}(A))$ 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 1.9. In this exercise, we actually prove the transformation formla 1.18. The same setup is as in theorem 1.18, 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 1.18 in the case that $f \geq 0$.
2. For integrable f prove theorem 1.18 by considering f_+ and f_- and using the previous item.

Video `Lecture3_V1` starts about here.

1.4 Products spaces and product measures, Fubini-Theorem

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

$$\Omega := \prod_{k \in \mathbb{N}} \Omega_k := \text{ sequences } (\omega_1, \omega_2, \dots) \text{ with } \omega_k \in \Omega_k \text{ for all } k \in \mathbb{N}. \quad (1.8)$$

A sigma algebra can be introduced on Ω as follows. A *finite dimensional rectangle* 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 (hence the term "finite dimensional"). Now let $\mathcal{A}$:= smallest *sigma* algebra on

Ω containing all finite dimensional rectangles. 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 finite dimensional rectangles is of course not needed then).

Video Lecture3_V2 starts about here.

Example 1.19. Let $\mathcal{B}(\mathbb{R})$ be the Borel algebra on $\mathbb{R}$ (see Def. 1.8). We define

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

and similarly

$$\mathbb{R}^d := \prod_{k=1}^{d} \mathbb{R}, \qquad \mathcal{B}_d := \bigotimes_{k=1}^{d} \mathcal{B}(\mathbb{R}),$$

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

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

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

Proof. Exercise 1.10. □

Definition 1.20. 1. For any finite $I \subset \mathbb{N}$, we define the *projections*

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

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

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

Video Lecture3_V3 starts about here.

3. $\mathbb{P}$ is called a *product probability* if for every finite dimensional rectangle

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

we have

$$\mathbb{P}(A) = \prod_{k\in\mathbb{N}} \mathbb{P}_{\{k\}}(A_k), \tag{1.9}$$

where $\mathbb{P}_{\{k\}}$ is the marginal for $I = \{k\}$. Note that in Equation 1.9, 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 $A_1 \in \mathcal{A}_1, \ldots, A_N \in \mathcal{A}_N$

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

Video `Lecture3_V3half` starts about here.

Theorem 1.21. *Let $\mathbb{P}, \mathbb{Q}$ 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*

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

Proof. The condition just means that $\mathbb{P} = \mathbb{Q}$ on finite dimensional rectangles. The rest of the proof is Exercise 1.12. □

Video `Lecture3_V4` starts about here.

Probably the most important fact about product probabilities (with finitely many factors) is that integrals over such probabilites can be calculated iteratively, and the order of integration does not matter. This is the following

Theorem 1.22 (Fubini-Tonelli theorem)**.** *Consider $(\Omega, \mathcal{A}) = (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ with product measure $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$. 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 $\mathbb{P}_2$-integrable, then the function $\omega_1 \to \int f(\omega_1, \omega_2) \cdot d\mathbb{P}_2(\omega_2)$ is measurable.*
3. *If $f \geq 0$ then,*

$$\int f \, \mathrm{d}\mathbb{P} = \int [\int f(\omega_1, \omega_2) \cdot d\mathbb{P}_2(\omega_2)] d\mathbb{P}_1(\omega_1). \tag{1.10}$$

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

$$\int f \, \mathrm{d}\mathbb{P} = \int [\int f(\omega_1, \omega_2) \cdot d\mathbb{P}_2(\omega_2)] d\mathbb{P}_1(\omega_1). \tag{1.11}$$

To prove this theorem, we will use two lemmata.

Lemma 1.23. *Items (1,2) hold for indicators $\mathbb{1}_A$, $A \in \mathcal{A}$.*

Lemma 1.24.

$$\mathbb{P}(A) = \int \mathbb{1}_A \, \mathrm{d}\mathbb{P} = \int [\int \mathbb{1}_A(\omega_1, \omega_2) \cdot d\mathbb{P}_1] d\mathbb{P}_2. \tag{1.12}$$

Proof of Lemma 1.23. Put $\mathcal{D} :=$ set of all $A \subset \Omega$ so that (a),(b) hold for indicators $\mathbb{1}_A$. If $A = A_1 \times A_2$, $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$, then $\mathbb{1}_A = \mathbb{1}_{A_1}(\omega_1) \cdot \mathbb{1}_{A_2}(\omega_2)$ and (a, b) hold trivially. Thus $\mathcal{D} \supset$ all cyclinders.

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

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

so in particular for ω_1 fixed. Hence

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

Further

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

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 finite dimensional rectangles 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 1.24. It is trivial to verify 1.24 for finite dimensional rectangles. 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 Lemma 1.24 by the MET. □

Completion of proof of Theorem 1.22. We have shown that 1.22 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

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

respectively, are measurable and converge monotonically to

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

respectively (the second by monotone convergence). Finally, because $\int f_n \, \mathrm{d}\mathbb{P} = \int [f_n(\omega_1, \omega_2) \, \mathrm{d}\mathbb{P}_2] \, \mathrm{d}\mathbb{P}_1$ and both sides converge by monotone convergence to $\int f \, \mathrm{d}\mathbb{P}$ and $\int [\int f(\omega_1, \omega_2) \, \mathrm{d}\mathbb{P}_2] \, \mathrm{d}\mathbb{P}_1$, respectively, Theorem 1.22 is proved for random variable $f \geq 0$.

If f is integrable, Theorem 1.22 holds for f_+ and f_-. The only thing that needs proving is that

$$\omega_1 \longrightarrow \int f(\omega_1, \omega_2) \, \mathrm{d}\mathbb{P}_2 = \int f_+(\omega_1, \omega_2) \, \mathrm{d}\mathbb{P}_2 - \int f_-(\omega_1, \omega_2) \, \mathrm{d}\mathbb{P}_2 \qquad (1.13)$$

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

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

Similarly for $N_- = \{\omega_1 \in \Omega_1; \int f_-(\omega_1, \cdot) \, \mathrm{d}\mathbb{P}_2 = \infty\}$. So 1.13 is well defined apart from $\omega_1 \in N_+ \cap N_-$. which has $\mathbb{P}(N_+ \cap N_-) = 0$. □

Remark 1.

1. The integrability condition cannot be omitted. It's not hard to find cases where $\int |f| \, \mathrm{d}\mathbb{P} = \infty$ and then both sides of Equation (1.11) 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

$$\int |f| \, \mathrm{d}\mathbb{P} = \int \big(\int |f|(\omega_1, \omega_2) \, \mathrm{d}\mathbb{P}_2(\omega_2) \big) \, \mathrm{d}\mathbb{P}_1(\omega_1),$$

   even if one side (and then also the other) is infinite.

3. One can use the right hand side of Equation (1.12) to define $\mathbb{P}$ from the marginals. Indeed, we have shown that the right hand side is well defined for $A \in \mathcal{A}$ and is $\sigma-$additive. This allows to invoke the MET to show that

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

4. The Fubini-Tonelli theorem extends to finite products by induction.

Exercises for Section 1.4

Exercise 1.10. Prove the statement in Example 1.19.

Exercise 1.11. Show that the family $\mathcal{A}$ of sets that are finite unions of finite dimensional rectangles 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 finite dimensional rectangles, then $A \setminus B$ is a finite union of finite dimensional rectangles.)

Exercise 1.12. Demonstrate Theorem 1.21. (Hint: Show that the intersection of two finite dimensional rectangles is a finite dimensional rectangle. You can then use without proof the Theorem 1.6.)

Exercise 1.13 (**Assessed Exercise 1**)**.** Setup is as in Section 1.4

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 finite dimensional rectangle!) Show that B is nonetheless measurable.

2. Demonstrate that for a product probability $\mathbb{P}$ (see Definition 1.20, item 3) and with B as in the previous item, $\mathbb{P}(B) = \lim_{n\to\infty} \prod_{k=1}^{n} \mathbb{P}_k(A_k)$.

Exercise 1.14. Setup is as in Section 1.4. We might define the Borel algebra $\mathcal{B}(\mathbb{R}^d)$ for $d \in \mathbb{N}$ as the smallest sigma algebra containing all open sets of $\mathbb{R}^d$, similar to our definition of $\mathcal{B}(\mathbb{R})$ in Definition 1.8. Show that $\mathcal{B}(\mathbb{R}^d) = \mathcal{B}_d$ with $\mathcal{B}_d$ defined in Example 1.19. (Hint: Use Exercise 1.4.)

1.5 Distributions and independence

Video `Lecture4_V1` starts about here.

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 $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Random variables are measurable functions with values in $(\mathbb{R}^d, \mathcal{B}_d)$, where from now on, $d \in \mathbb{N} \cup \{\infty\}$. That is $d = \infty$ is permitted unless explicitly stated otherwise. Further, random variables are denoted by capital letters:

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

If $d = 1$, we put $\mathbb{E}(X) := \int X \, \mathrm{d}\mathbb{P}$ ("expectation value"). If $d < \infty$, $\mathbb{E}(X)$ is taken component wise.

Definition 1.25. 1. The *distribution* of a random variable $X : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}_d)$ is defined as $P_X := X_* \mathbb{P}$.

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

Video Lecture4_V2 starts about here.

Note that an I-marginal of X according to Definition 1.25 is the same as an I-marginal of $\mathbb{P}_X$ according to Definition 1.20 (Exercise 1.16).

Lemma 1.26. *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 $I = \{1, \dots, d\}$ 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 finite dimensional rectangle for some $I = \{n_1, \dots, n_k\} \subset \mathbb{N}$ and some $A_1, \dots, A_k \in \mathcal{B}_1$. Then

$$\begin{aligned} P_{X_I}(A_1 \times \cdots \times A_k) &= \mathbb{P}(\{\omega; X_{n_k} \in A_k, n_k \in I\}) \\ &= \mathbb{P}(X \in A) = P_X(A). \end{aligned} \tag{1.14}$$

and the same for Y. If the I-marginals agree, then Equation (1.14) shows that P_X and P_Y agree on finite dimensional rectangles, so Theorem 1.21 gives $P_X = P_Y$. If on the other hand $P_X = P_Y$, then Equation (1.14) (read from right to left) shows that the I-marginals agree. □

Video Lecture4_V3 starts about here.

Lemma 1.27. *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*

$$\mathbb{E}(f \circ X) = \int_{\mathbb{R}^d} f(x) \, \mathrm{d}P_X(x).$$

Proof. This is essentially the transformation formula, see Exercise 1.15. □

In Exercise 1.17, this result will be extended using densities.

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 $X_1, \dots, X_d$, each of them real valued. But "modelling" almost always means to merely specify the *distribution* of those random variables; the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which these random variables live, and in fact the variables themselves are usually not specified. The following lemma simply says that this is usually not a problem, because there is a canonical choice for these missing ingredients, the so–called coordinate representation, which can be found in the proof.

Lemma 1.28. *If μ is a probability on $(\mathbb{R}^d, \mathcal{B}_d)$, then there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a measurable random variable $X : \Omega \longrightarrow: \mathbb{R}^d$ so that $\mu = P_X$.*

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

Independence

Video `Lecture5_V1` starts about here.

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

Definition 1.29. 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

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

holds.

Definition 1.30. 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.

Exercises for Section 1.5

Exercise 1.15. In the setup of Lemma 1.27, show that $f \circ X$ is a random variable and prove the formula.

Exercise 1.16. For P_X the distribution of some $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B}_d)$, we have defined the concept of I–marginals in Definition 1.25. Show that Definition 1.20 however is also applicable and gives the same concept of I–marginals. (Hint: this is used in the proof of Lemma 1.26).

Exercise 1.17. 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 1.27)

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 1.18. 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 1.17. 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 $\mathbb{P}(\{\omega; X_1(\omega) \in B_1\}) = 0$ or 1.

Exercise 1.19. 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.

1.6 Conditional probabilities and Conditional expectations

Video `Lecture6_V1` starts about here.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ probability space. Consider L_2, the space of all random variables $f : \Omega \to \mathbb{R}$ so that $\int f^2 \, \mathrm{d}\mathbb{P} < \infty$. By a statement analogous to Theorem 1.17 but for L_2, this space is a Hilbert space with scalar product $\langle f, g\rangle := \int fg \, \mathrm{d}\mathbb{P}$ and norm $\|f\| := \sqrt{\int |f|^2 \, \mathrm{d}\mathbb{P}}$.

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

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

and equality occurs here if and only if $g = \hat{f}$ (see for instance Theorem 3.32 in [10]) 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

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

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

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

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

$$\int f \cdot g \, \mathrm{d}\mathbb{P} = \int \hat{f} \cdot g \, \mathrm{d}\mathbb{P} \quad \text{for all } g \in S. \tag{1.15}$$

We want to use this with for a special class of subspaces S, namely the space of all random variables g in L_2 that are measurable with respect to some sigma algebra $\mathcal{F}$ on Ω with $\mathcal{F} \subset \mathcal{A}$. It is clear that S is a subspace of L_2 but also a Hilbert space in its own right with the same scalar product as L_2. This implies that S is closed in L_2, hence we can find $\hat{f}$ so that (1.15) is correct. Note that for the special S we have chosen here, (1.15) is equivalent to

$$\int f \cdot \mathbb{1}_A \, \mathrm{d}\mathbb{P} = \int \hat{f} \cdot \mathbb{1}_A \, \mathrm{d}\mathbb{P} \tag{1.16}$$

for all $A \in \mathcal{F}$ by approximation. But Equation (1.16) makes sense even if only $\int |f| \, \mathrm{d}\mathbb{P} < \infty$, which is weaker that $\int f^2 \, \mathrm{d}\mathbb{P} < \infty$. This leads us to the following definition.

Video `Lecture6_V2` starts about here.

Definition 1.31. Let $\int |f| \, \mathrm{d}\mathbb{P} < \infty$, $\mathcal{F} \subset \mathcal{A}$, $\mathcal{F}$ 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 (1.16).

Lemma 1.32 (Properties of $\mathbb{E}(f|\mathcal{F})$)**.** *Suppose that f, g are random variables. Then (provided the conditional expectations appearing below exist), we have*

1. *$\mathbb{E}(.|\mathcal{F})$ is linear, that is $\mathbb{E}(af + bg|\mathcal{F}) = a\mathbb{E}(f|\mathcal{F}) + b\mathbb{E}(g|\mathcal{F})$ a.s.,*
2. *$\mathbb{E}(.|\mathcal{F})$ is monotone, that is $f \geq 0 \Rightarrow \mathbb{E}(f|\mathcal{F}) \geq 0$ a.s.,*

3. *If $\mathcal{G} \subset \mathcal{F} \subset \mathcal{A}$ are sigma-algebras, then*

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

(Law of the Iterated Expectations).

Proof. Exercise 1.21. □

Theorem 1.33. *Let $\int |f| \, \mathrm{d}\mathbb{P} < \infty$, $\mathcal{F} \subset \mathcal{A}$, $\mathcal{F}$ a sigma algebra.*

1. *There exists conditional expectation $\mathbb{E}(f|\mathcal{F})$.*
2. *Suppose $f^{(1)}, f^{(2)}$ are both conditional expectations of f given $\mathcal{F}$, then*

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

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

Proof. We start with item 2. Indeed, if $f^{(1)}, f^{(2)}$ are conditional expectations of f with respect to $\mathcal{F}$, then it follows from Equation (1.16) that

$$\int (f^{(1)} - f^{(2)}) \cdot \mathbb{1}_A \, \mathrm{d}\mathbb{P} = 0$$

for any $A \in \mathcal{F}$. Since $f^{(1)} - f^{(2)}$ is $\mathcal{F}$–measurable, this implies $f^{(1)} = f^{(2)}$ almost surely.

Video `Lecture6_V3` starts about here.

With regards to the existence of a conditional expectation, it follows from our discussion at the beginning of this section that $\mathbb{E}(f|\mathcal{F})$ exists if $\int f^2 \, \mathrm{d}\mathbb{P} < \infty$. We will now prove that this mapping is furthermore continuous in the L_1–norm. It follows from Lemma 1.32 that the mapping $f \to \mathbb{E}(f|\mathcal{F})$ is monotone (see Exercise 1.21). Therefore, since $f \leq |f|$ we obtain $\mathbb{E}(f|\mathcal{F}) \leq \mathbb{E}(|f||\mathcal{F})$. The same is true for $-f$ and we therefore get what might be termed a triangle inequality

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

We integrate over this inequality and use that $\mathbb{E}(\mathbb{E}(|f||\mathcal{F})) = \mathbb{E}(|f|)$ (this can be shown by using Equation (1.16) with $A = \Omega$). We obtain

$$\mathbb{E}\big(|\mathbb{E}(f|\mathcal{F})|\big) \leq \mathbb{E}(|f|).$$

Together with the linearity from Lemma 1.32, we get that the mapping $f \to \mathbb{E}(f|\mathcal{F})$ is a linear mapping from L_2 into L_1 which is uniformly continuous in the L_1–norm. Since L_2 is dense in L_1 (both contain simple functions, for instance), $\mathbb{E}(.|\mathcal{F})$ has a unique continuous extension onto the whole of L_1. □

Video Lecture7_V1 starts about here.

Remark 2 (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, Theorem 1.33 tells you that you have to check that

1. $\hat{f}$ is $\mathcal{F}$-measurable
2. $$\int f \cdot g \, \mathrm{d}\mathbb{P} = \int \hat{f} \cdot g \, \mathrm{d}\mathbb{P}$$
   for any function g which is $\mathcal{F}-$measurable and bounded (in fact, it suffices to check this for all g of the form $\mathbb{1}_A$ with $A \in \mathcal{F}$).

Definition 1.34.

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
   $$\int \mathbb{1}_B(x) \cdot \hat{f}(x) \, \mathrm{d}P_X(x) = \int \mathbb{1}_B \circ X(\omega) \cdot f(\omega) \, \mathrm{d}\mathbb{P}(\omega)$$
   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 1.35. $\hat{f}(x) = \mathbb{E}(f|X = x) \iff \hat{f}(X(\omega)) = \mathbb{E}(f|\sigma(X))(\omega)$.

Proof. Exercise 1.22. □

Video Lecture7_V2 starts about here.

Definition 1.36 (Conditional Probability)**.** The conditional expectation of an indicator function has a special interpretation. Let $A \in \mathcal{A}$.

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

3. Using the alternative concept of conditional expectation in Definition 1.34, item 3 we define $\mathbb{P}(A|X = x) := \mathbb{E}(\mathbb{1}_A|X = x)$.

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

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

In most practical applications, conditional expectations are calculated using the following result:

Lemma 1.37 (Bayes–Rule)**.** *Consider two random variables*

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

where $d_1 + d_2 < \infty$*. Suppose that* $Z = (X, Y)$ *has a density* $p : \mathbb{R}^2 \to \mathbb{R}; (x, y) \to p(x, y)$ *(see Exercise 1.7 for a discussion of densities). Then*

$$\mathbb{P}(X \in B|Y = y) = \frac{\int_B p(x, y) \, \mathrm{d}x}{\int_{\mathbb{R}^{d_1}} p(x, y) \, \mathrm{d}x} \tag{1.17}$$

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

Proof. Exercise 1.23. □

Video Lecture7_V3 starts about here.

Regular conditional probabilities

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

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

so that

1. for every $A \in \mathcal{A}$ the mapping $\omega \to \mu(A, \omega)$ is $\mathcal{F}-$measurable random variable.

We would also like to have that

2. for every $\omega \in \Omega$ the mapping $A \longrightarrow \mu(A, \omega)$ is a probability on A.

But there is a problem: Note that the relation

$$\lim_{n\to\infty} \sum_{k=1}^{n} \mu(A_k, \omega) = \mu(\bigcup_{k=1}^{\infty} A_k, \omega) \tag{1.18}$$

for pairwise disjoint $A_1, A_2, \cdots \in \mathcal{A}$ merely holds for $\omega \in \Omega_0$ with $\mathbb{P}(\Omega_0) = 1$. Although these are "almost all ω", the set Ω_0 where Equation (1.18) holds depends on $A_1, A_2 \ldots$. Now μ would have to be modified on $\Omega_0^{\complement}$ in order to render Equation (1.18) 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 or, worse still, might not even be measurable.

Video `Lecture7_V4` starts about here.

Theorem 1.38. *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*

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

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 $\mathbb{P}(\Omega_B) = 1$, and μ satisfies conditions (1,2) at the beginning of this paragraph.*

Proof. See [1], theorem 4.34. The structure of $\mathcal{B}_d$ enters in an essential way. □

Exercises for Section 1.6

Exercise 1.20. (Obsolete)

Exercise 1.21. Prove Lemma 1.32.

Exercise 1.22. Prove Lemma 1.35.

Exercise 1.23 (**Assessed Exercise 2**)**.** In this exercise you prove the Bayes Rule Lemma 1.37. Proceed along the following steps (you might want to look at the results of Exercise 1.7 as well):

1. Show using Fubini–Tonelli that both the numerator as well as the denominator on the right hand sides of Equation (1.17) are measurable functions of y.

2. Consider the set B_0 of all $y \in \mathbb{R}^{d_2}$ where the denominator on the right hand side of Equation (1.17) vanishes, and prove that $\mathbb{P}(\{Y \in B_0\}) = 0$.

3. Now using the defining property of the conditional expectation, show that Equation (1.17) holds.

Remember that conditional probabilities, like conditional expectations, are defined only "up to sets of measure zero", so you do not need to have Equation (1.17) satisfied for *all* y, but only for y in some set B_1 so that $\mathbb{P}(\{Y \in B_1\}) = 1$.

1.7 Literature on measure theory and integration

The following books cover measure theory and integration, mostly somewhat more general than in this chapter. [7] is nice and brief, strongly recommended. Some proofs are ommitted. [3] 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. [6] 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. [2] a very consise text which nonetheless covers everything that is important.

Concerning probability theory, I recommend the following. [9] 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. [1] 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. [4, 5] Feller's two books on probability theory are even more classic in probability theory than [1]. Again, does not cover measure theory and integration in detail.

For data assimilation, I believe that [8] 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`

Chapter 2

Stochastic processes in discrete time

2.1 Basic definitions

Video Lecture8_V1 starts about here.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.

Definition 2.1. A *stochastic process* with *state space* E is a sequence $\{X_n, n \in \mathbb{N}\}$ of random variables $X_k : \Omega \to E$.

For this definition to make sense, E has to be a space so that the notion of random variables with values in E makes sense.

Remark 3 (Convention regarding state space E). Many of the things we will say in these lecture notes about stationary stochastic processes as well as Markov processes (see Sec.2.2 and 2.5, respectively) remains true if we take for E to be a separable and complete metric space (so called *Polish space*). However, for simplicity's sake our state spaces will mostly be $E = \mathbb{R}^d$ with Borel algebra $\mathcal{B}(\mathbb{R}^d)$ where d is finite. We might demonstrate special results where E will be just the real line $\mathbb{R}$ or some interval with the standard Borel algebra, or even only a finite set, for instance $\{1, \ldots, d\}$ with the sigma algebra of all possible subsets.

For some set $I \subset \mathbb{N}$ we will use the shorthand $X_I := (X_i; i \in I)$. We also remind the reader of the cartesian product $(E^\infty, \mathcal{B}(E)_\infty)$.

2.2 Stationary processes and the ergodic theorem

For this lecture we fix a stochastic process $\{X_n, n \in \mathbb{N}\}$ with state space $(E, \mathcal{B}(E))$.

Definition 2.2. A stochastic process $\{X_n, n \in \mathbb{N}\}$ is *stationary* if its distribution is invariant under time shifts, more precisely, if it has the same distribution as the shifted process $\{Y_n, n \in \mathbb{N}\}$ defined through $Y_n = X_{n+1}$ for all $n \in \mathbb{N}$.

We remember that two stochastic processes $\{X_n, n \in \mathbb{N}\}$ and $\{Y_n, n \in \mathbb{N}\}$ have the same distribution if they have the same finite–dimensional distributions, that is if X_I and Y_I have the same distribution for all finite subsets $I \subset \mathbb{N}$ (see Lemma 1.26). We can therefore conclude that a process $\{X_n, n \in \mathbb{N}\}$ is stationary if and only if its finite–dimensional distributions are invariant with respect to shifting the time index. In particular the distribution of X_n does not depend on n, and the same is true for expectations like $\mathbb{E}(\phi(X_n))$ for measurable functions ϕ as long as the expectation exist.

Video Lecture8_V2 starts about here.

Given a stationary process, one may construct new ones:

Lemma 2.3. *Let $\phi : E^\infty \to \mathbb{R}^m$ be a measurable mapping, and $\{X_n, n \in \mathbb{N}\}$ a stationary process with state space E. Then the process $\{Y_n, n \in \mathbb{N}\}$ defined through $Y_n := \phi(X_n, X_{n+1}, \ldots)$ has state space $\mathbb{R}^m$ and is stationary.*

Note that ϕ in Lemma 2.3 might depend on finitely many components only, so that in particular processes like $\{\phi(X_k), k \in \mathbb{N}\}$ are stationary if $\{X_n, n \in \mathbb{N}\}$ is.

The most important result about stationary processes is the

Theorem 2.4 (Ergodic theorem)**.** *Suppose that $\phi : E \to \mathbb{R}$ is measurable and $\mathbb{E}(|\phi(X_1)|) < \infty$. Then there exists a random variable Y so that*

$$\frac{1}{n}\sum_{k=1}^{n} \phi(X_k) \to Y$$

almost surely as well as in L_1.

For a proof see [1], Theorem 6.21. The random variable Y that appears as the limit in the ergodic theorem has interesting properties which we want to describe.

Definition 2.5. Let $\{X_n, n \in \mathbb{N}\}$ be a stationary process with state space E.

1. A measurable mapping $\psi : E^\infty \to \mathbb{R}^m$ is called *invariant* with respect to $\{X_n\}$ if $\psi(X_1, X_2, \ldots) = \psi(X_2, X_3, \ldots)$ almost surely.
2. A random variable Y is called *invariant* with respect to $\{X_n\}$ if $Y = \psi(X_1, X_2, \ldots)$ for some invariant ψ as in the previous item.

Video Lecture8_V3 starts about here.

Corollary 2.6. *The random variable Y that appears in the Ergodic Theroem is invariant with respect to $\{X_n, n \in \mathbb{N}\}$. Furthermore $\mathbb{E}Y = \mathbb{E}\phi(X_1)$.*

We have formulated the Ergodic Theorem without any concept of "Ergodicity". This concept will be introduced now.

Definition 2.7. A stationary process $\{X_n, n \in \mathbb{N}\}$ is called *ergodic* if every invariant random variable is almost surely equal to a constant.

Exercise 2.1 (**Assessed Exercise 3**). 1. Prove Corollary 2.6.

2. Suppose that $\{X_n, n \in \mathbb{N}\}$ is stationary and ergodic. Show that the random variable Y in the Ergodic Theorem is equal to $\mathbb{E}\phi(X_1)$.

Exercise 2.2 (**Assessed Exercise 4**). 1. Prove Lemma 2.3.

2. Suppose that $\{X_n, n \in \mathbb{N}\}$ is stationary and ergodic. Show that the process $Y_n := \phi(X_n, X_{n+1}, \ldots)$ as in Lemma 2.3 is also ergodic.

2.3 Martingales

Video Lecture9_V1 starts about here.

In this section we consider martingales and related processes where the state space E is the real line. Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We first need the concept of a *filtration*:

Definition 2.8. 1. A *filtration* is a sequence $\{\mathcal{F}_n, n \in \mathbb{N}\}$ of sigma algebras so that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}$.

2. Given a filtration $\{\mathcal{F}_n, n \in \mathbb{N}\}$, we say that a stochastic process $\{X_n, n \in \mathbb{N}\}$ is *adapted* to the filtration if X_k is $\mathcal{F}_k$–measurable for all $k \in \mathbb{N}$.

Any stochastic process $\{X_n, n \in \mathbb{N}\}$ generates its "own" or *natural* filtration $\{G_n := \sigma(X_1, \ldots, X_n), n \in \mathbb{N}\}$ for all $n \in \mathbb{N}$. Clearly, any stochastic process is adapted to its natural filtration. We are now ready to define martingales and related processes:

Definition 2.9. Suppose we are given a filtration $\{\mathcal{F}_n, n \in \mathbb{N}\}$.

1. A stochastic process $\{X_n, n \in \mathbb{N}\}$ is a *martingale* with respect to the filtration $\{\mathcal{F}_n\}$ if it is adapted to the filtration and

$$\mathbb{E}(X_{k+1}|\mathcal{F}_k) = X_k \tag{2.1}$$

for all $k \in \mathbb{N}$.

2. A stochastic process $\{X_n, n \in \mathbb{N}\}$ is a *submartingale* with respect to the filtration $\{\mathcal{F}_n\}$ if it is adapted to the filtration and

$$\mathbb{E}(X_{k+1}|\mathcal{F}_k) \geq X_k \tag{2.2}$$

for all $k \in \mathbb{N}$. It is called a *supermartingale* if the opposite inequality holds in Equation (2.2) instead.

3. A stochastic process $\{X_n, n \in \mathbb{N}\}$ is a *martingale difference sequence* (MDS)[1] with respect to the filtration $\{\mathcal{F}_n\}$ if it is adapted to the filtration and

$$\mathbb{E}(X_{k+1}|\mathcal{F}_k) = 0$$

for all $k \in \mathbb{N}$.

We collect a few facts about these processes.

Video `Lecture9_V2` starts about here.

Lemma 2.10. *Fix a filtration $\{\mathcal{F}_n, n \in \mathbb{N}\}$. Unless stated otherwise, all martingales, submartingales, etc that appear in this Lemma are with respect to this filtration.*

1. *If $\{X_n, n \in \mathbb{N}\}$ is a martingale (resp. submartingale, supermartingale, MDS), then it is also martingale (resp. submartingale, supermartingale, MDS) with respect to its own filtration $\{\sigma(X_1, \ldots, X_n), n \in \mathbb{N}\}$.*

2. *If $\{X_n, n \in \mathbb{N}_0\}$ is a martingale, then $Y_n := X_n - X_{n-1}$ for $n \in \mathbb{N}$ defines an MDS.*

3. *If $\{X_n, n \in \mathbb{N}\}$ is an MDS, then $Y_n := \sum_{k=1}^{n} X_k$ for $n \in \mathbb{N}$ defines a martingale.*

4. *If $\{X_n, n \in \mathbb{N}\}$ is a martingale and $\phi : \mathbb{R} \to \mathbb{R}$ a convex function such that $\mathbb{E}|\phi(X_n)| < \infty$ for all $n \in \mathbb{N}$, then setting $Y_n := \phi(X_n)$ for all $n \in \mathbb{N}$ defines a submartingale.*

5. *If $\{X_n, n \in \mathbb{N}\}$ is a martingale, then for $m \leq n$ we have $\mathbb{E}(X_n|\mathcal{F}_m) = X_m$. For submartingales (resp. supermartingales) this holds with "$\geq$" (resp. "$\leq$") replacing "$=$".*

6. *If $\{X_n, n \in \mathbb{N}\}$ is a martingale, then $\mathbb{E}(X_n) = \mathbb{E}(X_1)$. For submartingales (resp. supermartingales), the sequence $\{\mathbb{E}(X_n), n \in \mathbb{N}\}$ is increasing (resp. decreasing).*

Proof. Exercise 2.3. □

[1] Also called *fair process*

Next we will discuss an important construction which may be seen as "integrating" a predictable process against a martingale, resulting in a new martingale.

Video `Lecture9_V3` starts about here.

Lemma 2.11. *Let $\{X_n, n \in \mathbb{N}_0\}$ be a martingale and $\{Y_n, n \in \mathbb{N}_0\}$ be a stochastic process, both with state space $\mathbb{R}$ and adapted to the same filtration $\{\mathcal{F}_n, n \in \mathbb{N}_0\}$. Furthermore, suppose that $\mathbb{E}|Y_{n-1} \cdot (X_n - X_{n-1})| < \infty$ for all $n \in \mathbb{N}$. Then*

$$Z_n := \sum_{k=1}^{n} Y_{k-1} \cdot (X_k - X_{k-1}) \qquad \text{for } n \in \mathbb{N} \tag{2.3}$$

forms a martingale with respect to the same filtration.

Proof. We just need to show that $R_n := Y_{k-1} \cdot (X_k - X_{k-1})$ for $n \in \mathbb{N}$ forms an MDS and invoke Lemma 2.10, item 4. We have $\mathbb{E}|R_n| < \infty$ by assumption for all $n \in \mathbb{N}$, so

$$\begin{aligned}
&\mathbb{E}(R_n|\mathcal{F}_{n-1}) \\
&= \mathbb{E}(Y_{n-1} \cdot (X_n - X_{n-1})|\mathcal{F}_{n-1}) \qquad \text{(Def. of } R_n\text{)} \\
&= Y_{n-1} \cdot \mathbb{E}(X_n - X_{n-1})|\mathcal{F}_{n-1}) \qquad \text{(Since } Y_{n-1} \text{ is } \mathcal{F}_{n-1} \text{ measbl.)} \\
&= Y_{n-1} \cdot (\mathbb{E}(X_n|\mathcal{F}_{n-1}) - X_{n-1}) \qquad \text{(Linearity of } \mathbb{E}(.|\mathcal{F}_{n-1})\text{)} \\
&= Y_{n-1} \cdot 0 = 0 \qquad \text{(Martingale property of } \{X_n\}\text{)}.
\end{aligned}$$

□

Equation (2.3) might be seen as a Riemann-Steltjes sum approximating the integral

$$Z_t = \int_0^t Y_s \, \mathrm{d}X_s. \tag{2.4}$$

Such *stochastic integrals* can indeed be defined for a martingale $\{X_t\}$ in continuous time and so–called *predictable processes* $\{Y_t\}$. The process $\{Z_t\}$ is then again a martingale. Continuous time martingales and stochastic integrals as in Equation (2.4) form a cornerstone of stochastic analysis and mathematical finance.

The reasons for this though can be explained in our discrete time framework. Suppose you want to make financial investments. The sigma–algebra $\mathcal{F}_n$ may be seen as the market information you hold at time n; if a random variable is measurable with respect to $\mathcal{F}_n$, you will known its value at time n. You will never forget any information and hence these sigma–algebras are growing. Let $\{X_n, n \in \mathbb{N}\}$ be a stochastic process representing the value of some financial asset. The condition Equation (2.1) then says that the

expected future value of a financial asset, given the information we have currently, is equal to its current value. This assumption basically means that the value of the asset has no predictable trend. This model is too simple for assets such as for instance shares. But martingales form important building blocks for more realistic models; we will use martingales here to explain the basic ideas.

At each time n you decide to buy an amount of Y_n of the asset X_n. The value at time $n+1$ will then be $Y_n \cdot X_{n+1}$ and you have made a gain (or loss) of $Y_n \cdot (X_{n+1} - X_n)$. (There is no way you can foresee the value of X_{n+1} at time n so this is why Y_n must be $\mathcal{F}_n$ measurable.) Summing up all your gains up to time n, you obtain Z_n as in Equation (2.3).

The fact that this is still a martingale according to Lemma 2.11 is remarkable as it means that whatever your "buying strategy" $\{Y_n\}$ your *expected* gains (or losses) will always be zero. This in fact also applies to strategies such as "quit while you are ahead", as we will see in the next section.

Exercise 2.3. Prove Lemma 2.10. For item 4, you need to invoke Jensen's inequality, see [1].

Exercise 2.4 (Assessed Exercise 5). Suppose that $\{X_n, n \in \mathbb{N}\}$ is a martingale or a nonnegative submartingale with the property that $\mathbb{E}(X_n^2) < \infty$ for all $n \in \mathbb{N}$. Show that for $m \leq n$ we have

$$\mathbb{E}(X_n - X_m)^2 \leq \mathbb{E}(X_n^2) - \mathbb{E}(X_m^2).$$

Proof.

$$\begin{aligned}
\mathbb{E}(X_n - X_m)^2 &= \mathbb{E}(X_n)^2 + \mathbb{E}(X_m)^2 - 2\mathbb{E}(X_n X_m)^2 \\
&= \mathbb{E}(X_n)^2 + \mathbb{E}(X_m)^2 - 2\mathbb{E}(\mathbb{E}(X_n|\mathcal{F}_m)X_m)^2 \\
&\leq \mathbb{E}(X_n)^2 + \mathbb{E}(X_m)^2 - 2\mathbb{E}(X_m X_m)^2 \\
&= \mathbb{E}(X_n - X_m)^2,
\end{aligned}$$

because $\mathbb{E}(X_n|\mathcal{F}_m) \geq X_m$. □

Exercise 2.5 (Assessed Exercise 6). Consider the situation of Lemma 2.11 and suppose that $\mathbb{E}(Y_{n-1}^2 \cdot (X_n - X_{n-1})^2)$ is finite for all $n \in \mathbb{N}$. Show that

$$\mathbb{E}(Z_n^2) = \sum_{k=1}^{n} \mathbb{E}(Y_{k-1}^2 \cdot (X_k - X_{k-1})^2).$$

Proof. Write $R_k = Y_{k-1}(X_k - X_{k-1})$ for all $k \in \mathbb{N}$. Then $Z_n = \sum_{k=1}^n R_k$ and $\mathbb{E}(R_k|\mathcal{F}_l) = 0$ for all $l < k$. Hence if $k > l$ we have $\mathbb{E}(R_k R_l) =$

$\mathbb{E}(\mathbb{E}(R_k|\mathcal{F}_l)R_l) = 0$. Therefore

$$\begin{aligned}\mathbb{E}(Z_n)^2 &= \sum_{k,l=1}^{n} \mathbb{E}(R_k R_l) \\ &= \sum_{k<l} \mathbb{E}(R_k R_l) + \sum_{k>l} \mathbb{E}(R_k R_l) + \sum_{k=1}^{n} \mathbb{E}(R_k^2) \\ &= \sum_{k=1}^{n} \mathbb{E}(R_k^2).\end{aligned}$$

□

2.4 Martingale convergence

Video Lecture10_V1 starts about here.

This lecture will only scatch the surface of this extremely important subject. The most general martingale convergence results are due to J.L.Doob (see [1] for a comprehensive discussion) but we will show simpler statements here following the presentation of [5]. Unless otherwise stated, we let $\{X_n, n \in \mathbb{N}\}$ be a martingale with respect to the filtration $\{\mathcal{F}_n, n \in \mathbb{N}\}$. Our aim is to prove that under suitable conditions, $\{X_n\}$ converges to a random variable Z as $n \to \infty$. We remind ourselves that with sequences of random variables, "convergence" may be understood in several different ways (for instance in L_1 or almost surely). Here we focus on mean square convergence and almost sure convergence.

Theorem 2.12. *Suppose that there exist $C > 0$ so that $\mathbb{E}(X_n^2) \leq C$ for all $n \in \mathbb{N}$. Then there is a random variable Z so that $X_n \to Z$ in mean square sense as well as almost surely.*

Note that $\mathbb{E}(X_n^2)$ being finite is not sufficient; it needs to be bounded.

Proof. To prove the convergence in the mean square sense, we need to show that there is a sequence $\{\alpha(n), n \in \mathbb{N}\}$ which goes to zero for $n \to \infty$ and so that $\alpha(n) \geq \mathbb{E}|X_{n+k} - X_n|^2$ for all $n, k \in \mathbb{N}$. This would mean that $\{X_n\}$ is a Cauchy sequence in the mean square norm. The result then follows because this norm is complete. Invoking Exercise 2.4 gives

$$\mathbb{E}|X_{n+k} - X_n|^2 \leq \mathbb{E}(X_{n+k}^2) - \mathbb{E}(X_n^2) \tag{2.5}$$

for all $n, k \in \mathbb{N}$. This shows that the sequence $\{\mathbb{E}(X_n^2), n \in \mathbb{N}\}$ is nondecreasing. Since it is bounded by assumption, it must converge to some $a \in \mathbb{R}$.

Setting $\alpha(n) := a - \mathbb{E}(X_n^2)$ we see that $\alpha(n) \to 0$ as $n \to \infty$. On the other hand, we obtain from Equation (2.5) that

$$\mathbb{E}|X_{n+k} - X_n|^2 \leq \alpha(n)$$

for all $n, k \in \mathbb{N}$, proving that $\{\alpha(n)\}$ has the desired property.

Video Lecture10_V2 starts about here.

To prove the almost sure convergence, let us recall that we need to find the set $\Omega_1 \subset \Omega$ so that $\{X_n(\omega), n \in \mathbb{N}\}$ is a Cauchy sequence whenever $\omega \in \Omega_1$, and then show that $\mathbb{P}(\Omega_1) = 1$. For $\epsilon > 0$ and $n, m \in \mathbb{N}$, we define the "good" sets

$$G_{n,m}(\epsilon) := \{\omega \in \Omega; \sup_{1 \leq k \leq m} |X_n(\omega) - X_{n+k}(\omega)| \leq \epsilon\}.$$

Now you need to check for yourself that $\{X_n(\omega)\}$ is a Cauchy sequence for a given ω if for every $\epsilon > 0$ there is an n such that $\omega \in G_{n,m}(\epsilon)$ for all $m \in \mathbb{N}$. In other words for every $\epsilon > 0$ we have $\omega \in G(\epsilon) := \cup_{n \in \mathbb{N}} \cap_{m \in \mathbb{N}} G_{n,m}(\epsilon)$. If you are happy with this, simply turn this around to find that $\{X_n(\omega)\}$ is *not* a Cauchy sequence if there is an $\epsilon > 0$ such $\omega \in B(\epsilon) := \cap_{n \in \mathbb{N}} \cup_{m \in \mathbb{N}} B_{n,m}(\epsilon)$, where each "bad" set

$$B_{n,m}(\epsilon) := \{\omega \in \Omega; \sup_{k \leq m} |X_n(\omega) - X_{n+k}(\omega)| > \epsilon\}.$$

is the complement of $G_{n,m}(\epsilon)$. Note that for n fixed, the bad sets are nested, in the sense that $B_{n,1}(\epsilon) \subset B_{n,2}(\epsilon) \subset \ldots$. We use this fact and the continuity of probability in the second equality below to get and estimate of $\mathbb{P}(B(\epsilon))$ (valid for all $n \in \mathbb{N}$):

$$\mathbb{P}(B(\epsilon)) \leq \mathbb{P}(\cup_{m \in \mathbb{N}} B_{n,m}(\epsilon)) = \lim_{m \to \infty} \mathbb{P}(B_{n,m}(\epsilon)). \tag{2.6}$$

We haven't used the martingale property yet; our estimate of $\mathbb{P}(B(\epsilon))$ in terms of of "bad" sets $B_{n,m}$ is valid for any stochastic process. For martingales, the probability of the $B_{n,m}$ can be estimated by Kolmogorov's inequality, which we will state and prove in Theorem 2.13 below. Applying Kolmogorov's inequality to the martingale $\{Y_k := X_{n+k} - X_n; k \in \mathbb{N}\}$ (where n is fixed) gives

$$\begin{aligned}
\mathbb{P}(B_{n,m}(\epsilon)) &= \mathbb{P}(\sup_{k \leq m} |X_n(\omega) - X_{n+k}(\omega)| > \epsilon) \\
&= \mathbb{P}(\sup_{k \leq m} |Y_k(\omega)| > \epsilon) \\
&\leq \frac{\mathbb{E}|Y_m|^2}{\epsilon^2} \qquad \text{(Kolmogorov's inequ.)} \\
&= \frac{\mathbb{E}|X_{n+m} - X_n|^2}{\epsilon^2} \\
&\leq \frac{\alpha(n)}{\epsilon^2}.
\end{aligned}$$

Replacing with this in Equation (2.6) and taking $n \to \infty$ gives $\mathbb{P}(B(\epsilon)) = 0$ since $\alpha(n) \to 0$. □

Video Lecture10_V3 starts about here.

Theorem 2.13 (Kolmogorov's inequality)**.** *Suppose that* $\{X_n, n \in \mathbb{N}\}$ *is a martingale with* $\mathbb{E}(X_k^2) < \infty$ *for all* k*. Then*

$$\mathbb{P}(\sup_{k \le m} |X_k| > \epsilon) \le \frac{\mathbb{E}X_m^2}{\epsilon^2} \tag{2.7}$$

for all $m \in \mathbb{N}$ *and* $\epsilon > 0$*.*

Proof. Consider the stochastic process defined for each $n \in \mathbb{N}$ through

$$Z_n := \sum_{k=1}^{n} Y_{k-1} \cdot (X_k - X_{k-1}),$$

where

$$Y_k := \begin{cases} 1 & \text{if} \quad |X_l| \le \epsilon \qquad \text{for all } l \le k, \\ 0 & \text{else} \end{cases}$$

for each $k \in \mathbb{N}$. According to Lemma 2.11, $\{Z_n, n \in \mathbb{N}\}$ is a martingale with $\mathbb{E}|Z_n|^p < \infty$ for all $n \in \mathbb{N}$. To understand what this process is doing, let τ be the first k where $|X_k| > \epsilon$ or equivalently the first k where $Y_k = 0$. This τ is a random variable which might even be ∞ if the event $|X_k| > \epsilon$ never happens. For all $n \le \tau$ we have $Z_n = X_n$ but for $n > \tau$ we have $Z_n = Z_\tau = X_\tau$, that is, Z_n "freezes" as soon as $|Z_n|$ exceeds ϵ for the first time. In particular, $\sup_{k \le m} |X_k| > \epsilon$ if and only if $|Z_m| > \epsilon$, therefore

$$\mathbb{P}(\sup_{k \le m} |X_k| > \epsilon) = \mathbb{P}(|Z_m| > \epsilon) \le \frac{\mathbb{E}(Z_m^2)}{\epsilon^2}, \tag{2.8}$$

the last inequality being Chebychev's inequality. Now calculate

Video Lecture10_V4 starts about here.

$$\begin{aligned} \mathbb{E}(Z_m^2) &= \sum_{k=1}^{m} \mathbb{E}(Y_{k-1}^2 \cdot (X_k - X_{k-1})^2) \qquad \text{(Exercise 2.5)} \\ &\le \sum_{k=1}^{m} \mathbb{E}((X_k - X_{k-1})^2) \qquad \text{(since } Y_k \le 1\text{)} \\ &\le \sum_{k=1}^{m} \mathbb{E}(X_k^2) - \mathbb{E}(X_{k-1}^2) \qquad \text{(Exercise 2.4)} \\ &= \mathbb{E}(X_m^2). \end{aligned}$$

Using this in Equation (2.8) gives Equation (2.7) and concludes the proof. □

Exercise 2.6 (**Assessed Exercise 7**)**.** Let $\{\mathcal{F}_n, n \in \mathbb{N}\}$ be a filtration and write $\mathcal{F}_\infty$ for the smallest sigma algebra containing all $\mathcal{F}_n, n \in \mathbb{N}$. Further, let Z be a random variable with $\mathbb{E}(Z^2) < \infty$.

1. Show that $\mathbb{E}(Z|\mathcal{F}_n)$ converges in mean square sense and almost surely to a random variable $\bar{Z}$. (Hint: prove that $Z_n := \mathbb{E}(Z|\mathcal{F}_n)$ for $n \in \mathbb{N}$ defines a martingale.)

2. Prove that $\bar{Z} = \mathbb{E}(Z|\mathcal{F}_\infty)$. (Hint: you may use without proof that if U, V are random variables measurable with respect to $\mathcal{F}_\infty$, and if $\mathbb{E}(U\mathbb{1}_A) = \mathbb{E}(V\mathbb{1}_A)$ for any $A \in \mathcal{F}_k$ and $k \in \mathbb{N}$, then $U = V$.)

Proof. We have $\mathbb{E}(Z_n|\mathcal{F}_{n-1}) = \mathbb{E}(\mathbb{E}(Z|\mathcal{F}_n)|\mathcal{F}_{n-1}) = \mathbb{E}(Z|\mathcal{F}_{n-1}) = Z_{n-1}$, hence $\{Z_n\}$ is a Martingale. Further $\mathbb{E}(Z_n^2) = \mathbb{E}(\mathbb{E}(Z|\mathcal{F}_n)^2) \leq \mathbb{E}(\mathbb{E}(Z^2|\mathcal{F}_n)) = \mathbb{E}(Z^2)$ so we can invoke Theorem 2.12 to show that $Z_n \to \bar{Z}$ both almost surely and in mean square sense. To prove the second item, pick $A \in \mathcal{F}_k$, then

$$\mathbb{E}(Z\mathbb{1}_A) = \mathbb{E}(\mathbb{E}(Z|\mathcal{F}_k)\mathbb{1}_A) = \mathbb{E}(Z_n\mathbb{1}_A) \tag{2.9}$$

if $n \geq k$. We know also that $Z_n \to \bar{Z}$ in mean square sense when taking $n \to \infty$; this implies that $\mathbb{E}(Z_n\mathbb{1}_A) \to \mathbb{E}(\bar{Z}\mathbb{1}_A)$ as is easy to see. Using this on the right hand side of Equation (2.9) and invoking the hint we find $Z = \bar{Z}$. □

2.5 Markov processes

In this section, we will consider another type of stochastic process called *Markov processes*. Again, we fix a stochastic process $\{X_n, n \in \mathbb{N}\}$ with state space E on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

Definition 2.14. The process $\{X_n\}$ is called a *Markov process* if for any $n \in \mathbb{N}$, the distribution of the *future* $\{X_{n+k}, k \in \mathbb{N}\}$ given the *past and present* $\{X_k, k \leq n\}$ only depends on the present X_n. More formally, for any $n \in \mathbb{N}$ and any set $A \in \sigma(X_{n+1}, X_{n+2}, \ldots)$ we have

$$\mathbb{P}(A|X_1, \ldots, X_n) = \mathbb{P}(A|X_n).$$

We will soon identify a lot of properties of Markov processes, but given that we have just discussed martingales, we first stress a couple of important differences between the two types of processes. Most importantly, the definition of a martingale just refers to the first conditional moment, while the definition of a Markov process refers to the whole distribution. Furthermore,

the definition of a martingale is linear; this means that if $\{X_n\}$ is a martingale with state space $\mathbb{R}^d$ and if $M : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a linear mapping, then $Y_n := MX_n$ for $n \in \mathbb{N}$ defines again a martingale. In particular, each coordinate of a martingale is again a martingale. The statement however becomes wrong if M is nonlinear; then $\{Y_n\}$ is no longer a martingale. Markov processes behave differently. Whether M is linear or nonlinear, $\{Y_n\}$ is in general not a Markov processes! The definition of a Markov process basically says that if you know the *full* current state of a Markov process, then additional information about the past will not improve your knowledge about what will happen in the future. This clearly is no longer true if you have only *partial* information about the current state (for instance if M is not a one–to–one mapping); in that case, additional information about the past can quite well improve your knowledge about what will happen in the future. A notable exception however is if M is one–to–one (linear or nonlinear). In that case, $\{Y_k, k \leq n\}$ and $\{X_k, k \leq n\}$ provide exactly the same information about the future of $\{X_n\}$ which in turn determines the future of $\{Y_n\}$.

In these lectures, we restrict attention to homogenous Markov processes (a definition will follow below). An important feature of Markov processes (homogenous ones in particular) is that their distribution is defined through relatively few objects. This renders Markov processes convenient for modelling dynamical phenomena. For the remainder of this section it is worth recapping Section 1.6, especially the last paragraph on regular conditional probabilities.

Definition 2.15. 1. A *Markov kernel* or *transition kernel* on a state space $(E, \mathcal{B}_E)$ is a mapping $K : \mathcal{B}_E \times E \to [0, 1]$ so that

   (a) for each $B \in \mathcal{B}_E$, the mapping $x \to K(B, x)$ is measurable;

   (b) for each $x \in E$, the mapping $B \to K(B, x)$ is a probability measure on $\mathcal{B}_E$.

2. Let $\{X_n, n \in \mathbb{N}\}$ be a Markov process. A sequence $\{K_n, n \in \mathbb{N}\}$ of transition kernels on E will be referred to as *the transition kernels of the Markov process* $\{X_n\}$ if for each $n \in \mathbb{N}$ and any $B \in \mathcal{B}(E)$ we have

$$\mathbb{P}(X_{n+1} \in B | X_n) = K_n(B, X_n) \qquad \text{almost surely.} \tag{2.10}$$

3. A Markov process is called *homogenous* if the kernel can be chosen independent of n.

Lemma 2.16. *For any Markov process $\{X_n\}$ with state space E there exists a corresponding sequence $\{K_n, n \in \mathbb{N}\}$ of transition kernels.*

Proof. Due to our convention that $E = \mathbb{R}^d$ with d finite, this is a direct consequence of Theorem 1.38 (which in fact also holds for E a Polish space). □

From now on, we will consider only homogenous Markov processes. Call a pair (K, π), where K is a kernel and π a distribution on a state space E, a *Markov pair*, or *the Markov pair of the Markov process* $\{X_n, n \in \mathbb{N}\}$ if K is the transition kernel of that Markov process and $\pi = P_{X_1}$ is the distribution of X_1. Although this nomenclature is not standard, it will allow us to formulate our next theorem concisely. Before we do that though, two examples:

Example 2.17 (Random walk on the circle)**.** Consider $E = \{0, \ldots, d-1\}$ and define inductively

$$X_{n+1} = (X_n + R_{n+1}) \mod d, \qquad n \in \mathbb{N},$$

where $\{X_0, R_1, R_2, \ldots\}$ are independent, X_0 has some given distribution π on E, and $\mathbb{P}(R_m = 1) = \mathbb{P}(R_m = -1) = \frac{1}{2}$ for all $m \in \mathbb{N}$. The state space E can be imagined as d points arranged on a circle, while X_n denotes the position of a particle which jumps one step in either clockwise or anticlockwise direction with equal probability (see Figure 2.1). The point labelled d is identified with the point labelled 0. To see that this is a Markov process, fix $n \in \mathbb{N}$, $k \in E$ and $(m_0, \ldots, m_n) \in E^{n+1}$. Assuming that $\mathbb{P}(\{X_l = m_l, l = 0{:}n\}) > 0$ we calculate (using "$\equiv$" for "equal mod d")

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = k | X_l = m_l, l = 0{:}n) \\
&= \mathbb{P}(X_n + R_{n+1} \equiv k | X_l = m_l, l = 0{:}n) \\
&= \mathbb{P}(m_n + R_{n+1} \equiv k | X_l = m_l, l = 0{:}n) \\
&= \mathbb{P}(m_n + R_{n+1} \equiv k),
\end{aligned}$$

the last equation because R_{n+1} is independent of $\{X_l, l = 0{:}n\}$. Exactly the same argument gives

$$\mathbb{P}(X_{n+1} = k | X_n = m_n) = \mathbb{P}(m_n + R_{n+1} \equiv k),$$

proving the Markov property. Furthermore, we can calculate the kernel; according to the above, we have

$$\begin{aligned}
&K(\{k\}|j) \\
&= \mathbb{P}(X_{n+1} = k | X_n = j) \\
&= \mathbb{P}(j + R_{n+1} \equiv k) \\
&= \begin{cases} \frac{1}{2} \text{ if either } j+1 \equiv k \text{ or } j-1 \equiv k, \\ 0 \text{ else.} \end{cases}
\end{aligned}$$

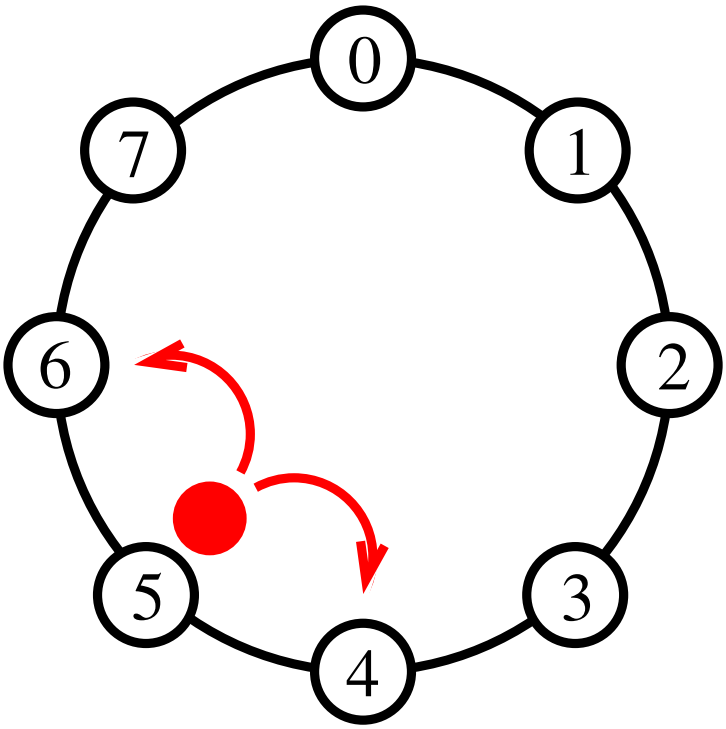


Figure 2.1: A particle jumps one step in either clockwise or anticlockwise direction with equal probability. The state space E is formed by d points arranged on a circle in "periodic" fashion, that is, the point d is identified with the point 0. Here $d = 8$.

This shows that the Markov process is homogenous.

Example 2.18 (Randomly selected measurable mappings)**.** Consider $E = \mathbb{R}^d$ and define inductively

$$X_{n+1} = f(X_n, R_{n+1}), \qquad n \in \mathbb{N},$$

where $f : E \times F \to E$ is a measurable mapping (and F is another state space). Further $\{X_0, R_1, R_2, \ldots\}$ are independent, X_0 has some given distribution π on E, and R_n has some given distribution ρ on F for all $n \in \mathbb{N}$. This system can be interpreted as a random (in fact independent and identically distributed) sequence of maps $f(., R_1), f(., R_2), \ldots$ applied in iterative fashion.

To see that $\{X_n\}$ is a Markov process, we first make a good guess how the Markov kernel K might look like. Fix $n \in \mathbb{N}$, $B \in \mathcal{B}(E)$ and $x \in E$. Now

imagine that $X_n = x$. Then we would have $X_{n+1} = f(x, R_{n+1})$, so

$$\begin{aligned}
&\mathbb{P}(X_{n+1} \in B | X_n = x) \\
&= \mathbb{P}(f(x, R_{n+1}) \in B | X_n = x) \\
&= \mathbb{P}(f(x, R_{n+1}) \in B) \qquad (\text{as } R_{n+1} \text{ is independent of } X_n) \\
&= \int \mathbb{1}_B(f(x,r)) \, \mathrm{d}\rho(r) \\
&=: K(B, x)
\end{aligned} \tag{2.11}$$

You will show in Exercise 2.8 that K is indeed a transition kernel in the sense of Definition 2.15. In order that $\{X_n, n \in \mathbb{N}\}$ is a Markov process, we need to show that

$$\mathbb{P}(X_{n+1} \in B | X_0, \ldots, X_n) \stackrel{?}{=} K(B, X_n)$$

almost surely for any $B \in \mathcal{B}(E)$ and any $n \in \mathbb{N}_0$. The left hand side is just a conditional expectation, so we have to check the properties of the conditional expectation from definition 1.31. It is clear that $K(B, X_n)$ is measurable with respect to $X_0, \ldots, X_n$ for any B fixed. Now let $\psi : E^{n+1} \to \mathbb{R}$ be any bounded measurable function. We need to show that

$$\mathbb{E}(\mathbb{1}_B(X_{n+1})\psi(X_0, \ldots, X_n)) \stackrel{?}{=} \mathbb{E}(K(B, X_n)\psi(X_0, \ldots, X_n)). \tag{2.12}$$

Our arguments will be very similar to those in the last example. As a shorthand, we will write P_n for the distribution of $X_0, \ldots, X_n$ on E^{n+1}.

$$\begin{aligned}
&\mathbb{E}(\mathbb{1}_B(X_{n+1}) \cdot \psi(X_0, \ldots, X_n)) \\
&= \mathbb{E}(\mathbb{1}_B(f(X_n, R_{n+1})) \cdot \psi(X_0, \ldots, X_n)) \\
&= \int\int \mathbb{1}_B(f(\xi_n, r)) \cdot \psi(\xi_0, \ldots, \xi_n) \, P_n(\mathrm{d}\xi_0, \ldots, \mathrm{d}\xi_n)\rho(\mathrm{d}r) \\
&\quad (\text{as } R_{n+1} \text{ is independent from } X_0, \ldots, X_n) \\
&= \int\int \mathbb{1}_B(f(\xi_n, r)) \, \rho(\mathrm{d}r) \, \psi(\xi_0, \ldots, \xi_n) \, P_n(\mathrm{d}\xi_0, \ldots, \mathrm{d}\xi_n) \\
&\quad (\text{Fubini}) \\
&= \int K(B, \xi_n) \, \psi(\xi_0, \ldots, \xi_n) \, P_n(\mathrm{d}\xi_0, \ldots, \mathrm{d}\xi_n) \\
&= \mathbb{E}(K(B, X_n) \cdot \psi(X_0, \ldots, X_n)),
\end{aligned}$$

demonstrating Equation (2.12). We also note that this Markov process is homogenous as K does not depend on n, effectively because the distribution of R_n is independent of n. This finishes the example.

Now a general theorem about Markov pairs and the distribution of Markov processes.

Theorem 2.19. *1. Given a Markov pair (K, π) there exist a Markov process $\{X_n, n \in \mathbb{N}\}$ with state space E so that (K, π) is the Markov pair of that process.*

2. Two Markov processes $\{X\}$ and $\{Y\}$ with the same Markov pair have the same distribution.

The proof of the first item is a sophisticated application of the Measture Extension Theorem (Thm. 1.5) which we will defer to Appendix A.1. First we need the following Lemma (the second item really belongs to the section on the conditional expectation).

Lemma 2.20. *1. Let K be a Markov kernel on E and $\phi : E \to \mathbb{R}$ be a bounded and measurable function. Then the function $K\phi : E \to \mathbb{R}, x \to \int_E \phi(z)K(\mathrm{d}z, x)$ is measurable.*

2. Let $\{Y_n, n \in \mathbb{N}\}$ be a sequence of nonnegative integrable random variables s.th. $Y_n \uparrow Y$, with Y integrable. Further, let $\mathcal{F} \subset \mathcal{A}$ be a sigma algebra. Then $\mathbb{E}(Y_n|\mathcal{F}) \uparrow \mathbb{E}(Y|\mathcal{F})$ almost surely. (Compare this with the Monotone Convergence Theorem).

Proof. Since K is a transition kernel, the function $K\phi$ is well defined. Let us first assume that ϕ is nonnegative and of the form $\phi(x) = \sum a_k \mathbb{1}_{B_k}(x)$ where the sum is finite (i.e. ϕ is simple). Then $K\phi(x) = \sum a_k K(B_k, x)$ which is clearly a measurable function. For ϕ nonnegative but not necessarily simple, by step 4 in the construction of the integral we can find simple functions $\{\phi_n, n \in \mathbb{N}\}$ so that $\phi_n \uparrow \phi$ as $n \to \infty$, while $K\phi_n$ is measurable as just discussed. By the monotone convergence theorem we have $K\phi_n(x) \uparrow K\phi(x)$ for every x . As the pointwise limit of measurable functions is measurable, $K\phi$ is measurable. For general ϕ, we may write $\phi = \phi_+ - \phi_-$ and note that by linearity $K\phi = K\phi_+ - K\phi_-$. Since both ϕ_+ and ϕ_- are nonnegative and measurable, the right hand side is measurable.

To prove the second item, we observe that since $\{Y_n, n \in \mathbb{N}\}$ is increasing, the same is true for $\{\mathbb{E}(Y_n|\mathcal{F}), n \in \mathbb{N}\}$. Therefore $\mathbb{E}(Y_n|\mathcal{F}) \uparrow \sup_n \mathbb{E}(Y_n|\mathcal{F}) =: \bar{Y}$. Since $Y \geq Y_n$ we have $\mathbb{E}(Y|\mathcal{F}) \geq \mathbb{E}(Y_n|\mathcal{F})$ so $\mathbb{E}(Y|\mathcal{F}) \geq \bar{Y}$. On the other hand

$$\begin{aligned}
&\mathbb{E}(\mathbb{E}(Y|\mathcal{F}) - \bar{Y}) \\
&= \mathbb{E}(\mathbb{E}(Y|\mathcal{F}) - \mathbb{E}(Y_n|\mathcal{F})) + \mathbb{E}(\mathbb{E}(Y_n|\mathcal{F}) - \bar{Y}) \\
&= \mathbb{E}(Y - Y_n) + \mathbb{E}(\mathbb{E}(Y_n|\mathcal{F}) - \bar{Y}).
\end{aligned}$$

Taking the limit $n \to \infty$ and applying the Monotone Convergence Theorem once again, the right hand side goes to zero and we can conclude $0 \leq \mathbb{E}(\mathbb{E}(Y|\mathcal{F}) - \bar{Y}) = 0$, which implies $\mathbb{E}(Y|\mathcal{F}) = \bar{Y}$ almost surely. $\square$

Proof of Theorem 2.19, item 2. Let now $\{X_n, n \in \mathbb{N}\}$ be a homogenous Markov process with Markov pair (K, π) and state space E. Our aim is to find expressions for the marginals of the distribution of $\{X_n\}$ that only involve the Markov pair (K, π). This implies that Markov processes with the same Markov pair will have the same marginals and therefore (by Thm. 1.21) the same distributions. Equipped with Lemma 2.20, we will now prove that for $\phi : E \to \mathbb{R}$ measurable and bounded, we have the identities

$$\mathbb{E}(\phi(X_{n+1})|X_1, \ldots, X_n) = \mathbb{E}(\phi(X_{n+1})|X_n) = K\phi(X_n). \tag{2.13}$$

Indeed, if $\phi = \mathbb{1}_B$ for some Borel set $B \in \mathcal{B}(E)$, then the identities follow directly from the Markov property and the definition of K, respectively. By linearity, the identities are true for ϕ simple. If ϕ is not simple but nonnegative, we may approximate ϕ by simple functions and invoke Lemma 2.20, item 2. For general ϕ, we again split into positive and negative part. Further, we conclude from Lemma 2.20, item 1 that $K\phi(X_n)$ is bounded and measurable.

Now let $B_1, \ldots, B_l$ be sets in $\mathcal{B}(E)$. We also let $\phi : E \to \mathbb{R}$ once again a bounded and measurable function. We then have

$$\begin{aligned}
&\mathbb{E}(\phi(X_{l+1}) \cdot \mathbb{1}_{B_l}(X_l) \cdot \ldots \cdot \mathbb{1}_{B_1}(X_1)) \\
&= \mathbb{E}\big(\mathbb{E}(\phi(X_{l+1})|X_l, \ldots, X_1) \cdot \mathbb{1}_{B_l}(X_l) \cdot \ldots \cdot \mathbb{1}_{B_1}(X_1)\big) \\
&\quad \text{(by def. of conditional expectation)} \\
&= \mathbb{E}\big(\mathbb{E}(\phi(X_{l+1})|X_l) \cdot \mathbb{1}_{B_l}(X_l) \cdot \ldots \cdot \mathbb{1}_{B_1}(X_1)\big) \\
&\quad \text{(by Eq. 2.13 first part)} \\
&= \mathbb{E}\big(K\phi(X_l) \cdot \mathbb{1}_{B_l}(X_l) \cdot \ldots \cdot \mathbb{1}_{B_1}(X_1)\big) \\
&\quad \text{(by Eq. 2.13 second part).}
\end{aligned} \tag{2.14}$$

We use this key identity as follows: first fix $k \in \mathbb{N}$ and use Equation (2.14) with $l = k$ and $\phi := \psi_k := \mathbb{1}_{B_k}$. This gives

$$\begin{aligned}
&\mathbb{P}(X_k \in B_k, \ldots, X_1 \in B_1) \\
&= \mathbb{E}(\mathbb{1}_{B_k}(X_k) \cdot \ldots \cdot \mathbb{1}_{B_1}(X_1)) \\
&= \mathbb{E}\big(K\psi_k(X_{k-1}) \cdot \mathbb{1}_{B_{k-1}}(X_{k-1}) \cdot \ldots \cdot \mathbb{1}_{B_1}(X_1)\big).
\end{aligned}$$

We could now again use Equation (2.14) with $l = k - 1$ and redefining $\phi := \psi_{k-1} := K\psi_k \cdot \mathbb{1}_{B_{k-1}}$. More generally, we set inductively $\psi_{l-1}(x) :=$

$K\psi_l(x) \cdot \mathbb{1}_{B_{l-1}}(x)$ for $l = k, \ldots, 2$ and find (using Eq. 2.14 in each induction step)

$$\mathbb{P}(X_k \in B_k, \ldots, X_1 \in B_1) = \mathbb{E}(\psi_1(X_1)) = \int_E \psi_1(x) \, \mathrm{d}\pi(x).$$

For given sets $B_1, \ldots, B_k$, the right hand side is entirely determined by the Markov pair (K, π). This finishes the proof; for completeness, we unroll the inductive definition of ψ_k to find a more explicit expression of the marginals:

$$\begin{aligned} &\mathbb{P}(X_k \in B_k, \ldots, X_1 \in B_1) \\ &= \int_E \cdots \int_E \mathbb{1}_{B_k}(x_k) K(\mathrm{d}x_k, x_{k-1}) \cdot \ldots \cdot \mathbb{1}_{B_2}(x_2) K(\mathrm{d}x_2, x_1) \cdot \mathbb{1}_{B_1}(x_1) \pi(\mathrm{d}x_1). \end{aligned} \tag{2.15}$$

□

Exercise 2.7. Let $X_n, n \in \mathbb{N}$ be a homogenous Markov process with pair (K, π_1) and let $\pi_n := P_{X_n}$. Prove the *Chapman–Kolmogorov–Identity*

$$\pi_{n+1} = K\pi_n, \tag{2.16}$$

(see Eq. 2.27 below for notation).

Proof. For $n = 1$, this follows from Equation (2.15) by taking $k = 2$ and $B_1 = E$. Suppose it has been proved for all $l \leq n$. Then use Equation (2.15) with $k = n+1$ and $B_l = E$ for all $l \leq n$ while $B_{n+1} = B$ for some $B \in \mathcal{B}(E)$. Then the left hand side is just $\pi_{n+1}(B)$, while the right hand side, after applying the Chapman–Kolmogorov equation for all $l \leq n$, can be written as $K\pi_n$. □

Exercise 2.8. Prove that the object $K(B, x)$ defined in Equation (2.11) is indeed a transition kernel in the sense of Definition 2.15.

Proof. Since f is a measurable mapping in both arguments, the mapping $r \to f(x, r)$ is measurable as a mapping from F to E for each $x \in E$ fixed. Therefore $f(x, R_{n+1})$ is a random variable for each $x \in E$ fixed and hence $B \to \mathbb{P}(f(x, R_{n+1}) \in B)$ is the distribution of that random variable and thus a probability distribution for each $x \in E$ fixed. On the other hand, for fixed $B \in \mathcal{B}(E)$ the mapping $x \to \mathbb{P}(f(x, R_{n+1}) \in B)$ is measurable as a consequence of the Fubini–Tonelli theorem. □

2.6 Ergodic theory of Markov processes with finite state space

In this section we will consider Markov processes with finite state space $E = \{1, \dots, d\}$, as for instance the Random walk on the circle (Example 2.17). In this situation, Markov processes are also called *Markov chains*. In particular, we are interested in the "long term" behaviour of the Markov chain. To explain what we mean by this, note first that distributions over E are given by vectors $\pi = (\pi^{(1)}, \dots, \pi^{(d)}) \in \mathbb{R}^d$ with $\pi^{(k)} \geq 0$ for all $k \in E$ and $\sum_k \pi^{(k)} = 1$. The set of all such distributions will be written as $\mathcal{P}_E$. Furthermore, a Markov kernel on E has an equivalent description through a matrix M with elements $M_{k,l} := K(\{k\}, l)$ for $k, l \in E$. A square matrix $M \in \mathbb{R}^{d \times d}$ evidently defines a kernel on E if and only if M is a *Stochastic matrix*, that is any column of M is a distribution over E, or equivalently $M_{k,l} \geq 0$ for all $k, l \in E$ and $\sum_k M_{k,l} = 1$ for all $l \in E$.

Fix a Markov chain with pair (M, π_1) (where now M is a stochastic matrix and $\pi_1 \in \mathcal{P}_E$). We will again write $\pi_n := \mathbb{P}(X_n \in .)$ for the distribution of X_n, and we have the following version of the Chapman–Kolmogorov–Identity (2.16):

$$\pi_{n+1} := M\pi_n = \left(\sum_{l=1}^{d} M_{k,l} \pi_n^{(l)} \right)_{1 \leq k \leq d} . \tag{2.17}$$

The first question we have is whether M has eigenvectors $\bar{\pi} \in \mathcal{P}_E$ corresponding to the eigenvalue $\lambda = 1$. This would mean that $\bar{\pi} = M\bar{\pi}$, so if we put $\pi_1 = \bar{\pi}$ then $\pi_n = \bar{\pi}$ for all n according to Equation (2.17), meaning that the X_n would have distribution $\bar{\pi}$ for all $n \in \mathbb{N}$. Such a $\bar{\pi}$ will be referred to as an *invariant distribution* of M.

The second question is whether π_n converges to some π_∞ for $n \to \infty$ (both the limit π_∞ and whether this happens at all might depend on π_0). Note that π_∞ will be an invariant distribution. Indeed, since $\mathcal{P}_E$ is closed we will have $\pi_\infty \in \mathcal{P}_E$, and furthermore $\pi_\infty = \lim_{n\to\infty} \pi_n = \lim_{n\to\infty} M\pi_{n-1} = M \lim_{n\to\infty} \pi_{n-1} = M\pi_\infty$. As to why this section is called "Ergodic theory of Markov processes ...", this will only become clear in the next section.

The first question is sorted in the following theorem

Theorem 2.21. *A stochastic matrix M always has $\lambda = 1$ as an eigenvalue and a corresponding eigenvector $\pi \in \mathcal{P}_E$. Furthermore, any eigenvector λ of M satisfies $|\lambda| \leq 1$.*

Proof. On $\mathbb{R}^d$, introduce the norm $\|v\| := \sum_{k=1}^{d} |v^{(k)}|$. Then for any $v \in \mathbb{R}^d$

we have

$$\|Mv\| = \sum_{k=1}^{d} |\sum_{l=1}^{d} M_{k,l} v^{(l)}| \leq \sum_{k,l=1}^{d} M_{k,l} |v^{(l)}| = \sum_{l=1}^{d} |v^{(l)}| = \|v\|, \tag{2.18}$$

hence if the relation $\lambda v = Mv$ holds, we can conclude $|\lambda|\|v\| = \|Mv\| \leq \|v\|$ so $|\lambda| \leq 1$.

To find an eigenvector in $\mathcal{P}_E$, take any $\pi_0 \in \mathcal{P}_E$ and let $\pi_n := \frac{1}{n} \sum_{k=0}^{n-1} M^k \pi_0$. Then $\pi_n \in \mathcal{P}_E$ for all $n \in \mathbb{N}$ as can be easily checked. Since $\mathcal{P}_E$ is compact, $\{\pi_n, n \in \mathbb{N}\}$ has a convergent subsequence, or in other words there is a sequence $n_1 \leq n_2, \ldots$ in $\mathbb{N}$ so that $\frac{1}{n_l} \sum_{k=0}^{n_l-1} M^k \pi_0 \to \pi_\infty$ for $l \to \infty$, and we have

$$\begin{aligned}
M\pi_\infty &= M \lim_{l\to\infty} \frac{1}{n_l} \sum_{k=0}^{n_l-1} M^k \pi_0 \\
&= \lim_{l\to\infty} M \frac{1}{n_l} \sum_{k=0}^{n_l-1} M^k \pi_0 \\
&= \lim_{l\to\infty} \frac{1}{n_l} \sum_{k=0}^{n_l-1} M^{k+1} \pi_0 \\
&= \lim_{l\to\infty} \frac{1}{n_l} \sum_{k=1}^{n_l} M^k \pi_0 \\
&= \lim_{l\to\infty} \frac{1}{n_l} \sum_{k=0}^{n_l-1} M^k \pi_0 + \lim_{l\to\infty} \frac{1}{n_l} M^{n_l} \pi_0 - \lim_{l\to\infty} \frac{1}{n_l} M^0 \pi_0.
\end{aligned}$$

The first term converges to π_∞. The second term converges to zero because $|M^{n_l}\pi_0| \leq 1$ by Equation (2.18) and the fact that $|\pi_0| = 1$. The third term converges obviously to zero. This shows that π_∞ is the desired eigenvector. □

We now turn to the second question, but will not address this in full generality. Rather, we restrict ourselves to *irreducible* stochastic matrices.

Definition 2.22. A stochastic matrix M is *irreducible* if for any $k, l \in E$ there exists an $n \in \mathbb{N}$ such that $M^n_{k,l} > 0$.

We emphasise that $M^n_{k,l}$ refers to the element in row k and column l of the matrix M^n, which is the n'th power of M in terms of matrix multiplication.

We will illustrate this concept, first through the following Lemma

Lemma 2.23. *Let $\{X_n, n \in \mathbb{N}\}$ be a Markov chain with Markov pair (M, π) and consider $k_1, \ldots, k_n$ with $k_l \in E$ for all $l = 1, \ldots, n$. Then*

$$\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n) = M_{k_n,k_{n-1}} \cdot, \ldots, \cdot M_{k_2,k_1} \cdot \pi^{(k_1)}. \tag{2.19}$$

Furthermore, if $\pi^{(k_1)} > 0$, we have

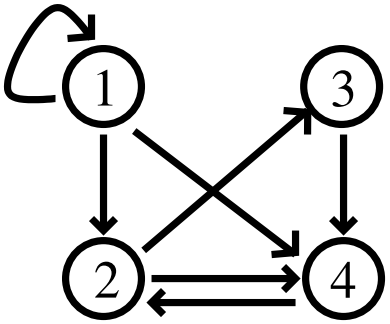
$$\mathbb{P}(X_n = k_n | X_1 = k_1) = M^n_{k_n,k_1}. \tag{2.20}$$

Proof. By basic probability calculus we have

$$\begin{aligned}\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n) = \mathbb{P}(X_n = k_n | X_{n-1} = k_{n-1}, \ldots, X_1 = k_1) \\ \cdot \mathbb{P}(X_{n-1} = k_{n-1} | X_{n-2} = k_{n-2}, \ldots, X_1 = k_1) \\ \vdots \\ \cdot \mathbb{P}(X_2 = k_2 | X_1 = k_1) \\ \cdot \mathbb{P}(X_1 = k_1).\end{aligned}$$

Using the Markov property and the definition of both M and π gives Equation (2.19). By summing Equation (2.19) over all possible $k_{n-1}, \ldots, k_1$ in E we obtain Equation (2.20). □

We now see that irreducibility, in conjuction with Equation (2.20), means that for any $k, l \in E$ there exists an $n \in \mathbb{N}$ such that $\mathbb{P}(X_n = k | X_1 = l) > 0$. If we interprete X_n as the position of a particle at time n (with the elements of E being the possible positions), irreducibility means a particle starting at position $l \in E$ has positive probability of arriving at $k \in E$ at *some* point in time, for any two states $k, l \in E$. A way to visualise the qualitative behaviour of a Markov chain (and in particular irreducibility) is via directed graphs (see Fig. 2.2 for an example). Given a stochastic matrix M on E, we first draw a circle for each element of E; these circles are called *nodes* Next, for each pair k, l so that $M_{k,l} > 0$, we draw an arrow from node l to node k (pointing from l to k, this is important!) Note that each arrow only points one way; if both $M_{k,l} > 0$ and $M_{k,l} > 0$, we draw two separate arrows pointing each way. Further, a node k gets an arrow pointing back at itself *only* if $M_{k,k} > 0$! In order to draw the directed graph, you don't actually need the matrix M, you just need to know which entries are zero and which not. Given the corresponding directed graph, it is easy to see if a stochastic matrix M is irreducible: if you are able to go from any given node to any other (using potentially many arrows but always in the right direction), then M is irreducible. The matrix in Figure 2.2 is *not* irreducible since there is no way to get from either 2, 3, or 4 to node 1. Between the nodes 2, 3, and 4 however, it is possible to get from any node to any other.



$$M = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 1 \\ 0 & 2/3 & 0 & 0 \\ 1/4 & 1/3 & 1 & 0 \end{pmatrix}$$

Figure 2.2: Directed graph corresponding to a stochastic matrix.

Theorem 2.24. *Suppose that M is an irreducible stochastic matrix. Then the eigenvector $\pi \in \mathcal{P}_E$ from Theorem 2.21 is unique an furthermore has only positive entries.*

Proof. We first assume that M is *mixing*, which means that there is a distribution $\rho \in \mathcal{P}_E$ and an $\alpha > 0$ such that $M_{k,l} \geq \alpha\rho_k$ for all k, l in E. (In particular, a stochastic matrix M is mixing if all entries are strictly positive.) Now define the subspace $V \subset \mathbb{R}^d$ of all $v \in \mathbb{R}^d$ such that $\sum_{k=1}^d v^{(k)} = 0$. It is easy to check that this is a subspace, and further that if $v \in V$ then $Mv \in V$, too. Now we use the same norm as in Theorem 2.21 and find for $v \in V$ that

$$(Mv)_k = \sum_{l \in E} M_{k,l} v_l = \sum_{l \in E} (M_{k,l} - \alpha\rho_k) v_l + \alpha \sum_{l \in E} \rho_k v_l = \sum_{l \in E} (M_{k,l} - \alpha\rho_k) v_l \tag{2.21}$$

by the property of v. Furthermore, $(M_{k,l} - \alpha\rho_k)_{k,l \in E}$ is still a matrix with nonnegative entries (albeit not a stochastic one). Therefore

$$\begin{aligned} \|Mv\| &= \sum_{k \in E} |\sum_{l \in E} (M_{k,l} - \alpha\rho_k) v_l| \\ &\leq \sum_{k,l \in E} (M_{k,l} - \alpha\rho_k) |v_l| \\ &= \sum_l \sum_k (M_{k,l} - \alpha\rho_k) |v_l| \\ &= (1 - \alpha) \|v\|. \end{aligned}$$

Note that $0 \leq (1 - \alpha) < 1$. If now M has two eigenvectors $\pi_1, \pi_2 \in \mathcal{P}_E$ corresponding to eigenvalue 1, then $\pi_1 - \pi_2 \in V$ and therefore

$$\|\pi_1 - \pi_2\| = \|M\pi_1 - M\pi_2\| = \|M(\pi_1 - \pi_2)\| \leq (1 - \alpha) \|\pi_1 - \pi_2\|.$$

This implies $0 = \|\pi_1 - \pi_2\|$ and hence $\pi_1 = \pi_2$.

If M is not mixing, we consider $T := \frac{1}{m}\sum_{l=0}^{m-1} M^l$, with m large enough so that all entries of T are strictly positive (this is possible because M is irreducible). Hence T is mixing. On the other hand, any eigenvector $\pi \in \mathcal{P}_E$ of M corresponding to eigenvalue 1 is also an eigenvector of T corresponding to eigenvalue 1. But since T is mixing, by our previous results π must be unique. Furthermore, it is easy to see that because T is a stochastic matrix with strictly positive entries and $\pi \in \mathcal{P}_E$, also $T\pi$ is a vector with strictly positive entries. Hence the relation $\pi = T\pi$ implies that π itself has strictly positive entries. □

Finally, we will investigate the question whether $\pi_n = M^n\pi_0 \to \pi_\infty$ for $n \to \infty$. It may already be evident that something like that can be shown for mixing stochastic matrices. Actually, a bit less will suffice:

Theorem 2.25. *Suppose that M is a stochastic matrix so that M^p is mixing for some $p \in \mathbb{N}$. Then $\pi_n := M^n\pi_0 \to \pi$ as $n \to \infty$ for any $\pi_0 \in \mathcal{P}_E$. Furthermore, π is the unique eigenvector in $\mathcal{P}_E$ of M for eigenvalue 1 (i.e. the unique stationary distribution). Finally, if λ is another eigenvalue of M, then $|\lambda| < 1$.*

Proof. If M^p is mixing, then a calculation similar to Equation (2.21) will show that

$$\|M^p v\| \leq (1-\alpha)\|v\| \tag{2.22}$$

for any $v \in V$ and some $\alpha > 0$. Any $n \in \mathbb{N}$ may be represented as $n = n_1 p + n_2$ with some $n_1 \in \mathbb{N}$ and $n_2 \in \{0, \ldots, p-1\}$ (divide n by p and take n_1 and n_2 as the integer part and the remainder, respectively). Therefore

$$\|M^n v\| = \|M^{n_1 p + n_2} v\| \leq (1-\alpha)^{n_1}\|M^{n_2} v\| \leq (1-\alpha)^{n_1}\|v\|, \tag{2.23}$$

where we have also used Equation (2.18) which is valid for any stochastic matrix. Since $n_1 \geq \frac{n}{p} - 1$, we may write Equation (2.23) as

$$\|M^n v\| \leq C\beta^n \|v\|, \tag{2.24}$$

with $\beta = (1-\alpha)^{1/p}$ and $C = \frac{1}{1-\alpha}$. Equation (2.24) implies that $\{\pi_n, n \in \mathbb{N}\}$ is a Cauchy sequence. Indeed, since $M^k\pi_0 - \pi_0 \in V$ we have

$$\|M^{n+k}\pi_0 - M^n\pi_0\| \leq C\beta^n\|M^k\pi_0 - \pi_0\| \leq C\beta^n(\|M^k\pi_0\| + \|\pi_0\|) \leq 2C\beta^n. \tag{2.25}$$

This shows that $\pi_n \to \pi$ and, as we have seen at the beginning of this section, π is a stationary distribution. If $\bar{\pi}$ is another stationary distribution, then

$\pi - \bar{\pi} \in V$, so Equation (2.22) implies $\|\pi - \bar{\pi}\| = \|M^p(\pi - \bar{\pi})\| \leq \beta\|\pi - \bar{\pi}\|$, and hence $\pi = \bar{\pi}$ since $\beta < 1$.

To show the last claim we use the following standard fact from linear algebra: if w is an eigenvector of M with eigenvalue $\lambda \neq 1$, then w must be perpendicular to any eigenvector ρ of M^{tr} with eigenvalue 1. But since we may take $\rho = (1, \ldots, 1)$, this implies $w \in V$. Therefore

$$|\lambda|^p\|w\| = \|\lambda^p w\| = \|M^p w\| = (1 - \alpha)\|w\|$$

so that $|\lambda| \leq \beta < 1$. □

Note that mixing stochastic matrices are not necessarily irreducible, and an irreducible stochastic matrix (or any power of it) might not be mixing (an example is the random walk on the circle with even number of states). Further, for mixing stochastic matrices the stationary distribution π might have zero entries, while this cannot happen for irreducible matrices according to Theorem 2.24.

We will now identify irreducible stochastic matrices M for which some power M^k is mixing. Fix $l \in E$. We define $\mathcal{R}(l) \subset \mathbb{N}$ as the set of all numbers $n > 0$ so that $M^n_{l,l} > 0$. In other words, $n \in \mathcal{R}(l)$ if there is nonzero probability that starting at l, the particle returns to l in n steps. Note that $\mathcal{R}(l)$ is not empty by irreducibility. Further, $\mathcal{R}(l)$ is *closed with respect to addition*, that means whenever n_1 and n_2 are in $\mathcal{R}(l)$, then so is $n_1 + n_2$. The *period* $\gamma(l)$ of l is the greatest common divisor of $\mathcal{R}(l)$. The following Lemma is remarkable, but we will not actually use it.

Lemma 2.26. *If M is irreducible, all states $l \in E$ will have the same period.*

Proof. (This proof can be skipped as we will not use the Lemma.) Pick two states l_1, l_2. Then there exist N_1, N_2 so that $M^{N_1}_{l_1,l_2} > 0$ and $M^{N_2}_{l_2,l_1} > 0$, that is the particle may travel from l_2 to l_1 in N_1 steps, and from l_1 to l_2 in N_2 steps. As a result, $N := N_1 + N_2$ is contained in $\mathcal{R}(l_2)$, therefore $\gamma(l_2)$ divides N. This means that $\gamma(l_2)$ divides all elements of $N + \mathcal{R}(l_2) := \{n + N, n \in \mathcal{R}(l_2)\}$ and is in fact the greatest common divisor of these numbers. On the other hand, $N + \mathcal{R}(l_2) \subset \mathcal{R}(l_1)$; this is because the particle might first travel from l_1 to l_2 in N_2 steps, then take a loop of n steps back to l_2 (for any $n \in \mathcal{R}(l_2)$), and then return to l_1 in N_1 steps. Hence $\gamma(l_1)$, too, divides all elements of $N + \mathcal{R}(l_2)$ so we obtain $\gamma(l_2) \geq \gamma(l_1)$. The argument is entirely symmetrical, so reversing the roles of l_1 and l_2 gives $\gamma(l_1) \geq \gamma(l_2)$, proving the claim. □

This motivates the following

Definition 2.27. The *period* of an irreducible matrix M is the period of the states $l \in E$ (which is the same for all l according to Lemma 2.26). Furthermore, an irreducible matrix M with period 1 is called *aperiodic*.

We are now ready to formulate the main theorem in this section:

Theorem 2.28. *Suppose that M is irreducible and aperiodic. Then there is a $q \in \mathbb{N}$ so that M^q is mixing. Therefore, the conclusions of both Theorems 2.24, 2.25 hold.*

Our proof will merely use that there is a state with period 1.

Proof. Fix a state $l \in E$ with period 1. We can find a finite set of elements $\{r_1, \ldots, r_p\} \subset \mathcal{R}(l)$ that have gcd equal to one. Using Euclid's algorithm, one may find numbers $\{a_1, \ldots, a_p\} \subset \mathbb{Z}$ such that $1 = a_1 r_1 + \ldots + a_p r_p$. Setting $s = r_1 + \ldots + r_p$, we may write any number n as $n = n_1 s + n_2$ with $0 \leq n_2 < s$. Therefore

$$n = n_1 s + n_2 \cdot 1 = n_1 s + n_2(a_1 r_1 + \ldots + a_p r_p) = \sum_{j=1}^{p} (n_1 + n_2 a_j) r_j.$$

Although the $\{a_j\}$ may contain negative numbers, there exists a $K \in \mathbb{N}$ so that if $n \geq K$, the coefficiens $n_1 + n_2 a_j$ are positive for all $j = 1, \ldots, p$. But then the right hand side is in $\mathcal{R}(l)$, so we conclude that any $n \geq K$ is in $\mathcal{R}(l)$. In other words, for any $n \geq K$ we have $M^n_{l,l} > 0$.

Now consider another state $j \neq l$. Since M is irreducible, we find m (depending on j and l) so that $M^m_{l,j} > 0$. Hence for $n \geq K$ we get $M^{n+m}_{l,j} \geq M^n_{l,l} M^m_{l,j} > 0$. Stated differently, for each $n \geq K + m$ we get $M^n_{l,j} > 0$. Since j was arbitrary, we can conclude (increasing K if necessary) that for any $n \geq K$ we have $M^n_{l,j} > 0$ for each $j \in E$. This means that the l'th row of M^K has only positive entries, which implies that M^K is mixing by Exercise 2.9. We set $q := K$ and the proof is complete. □

Exercise 2.9. Let M be a stochastic matrix with one row having only positive entries. Show that M is mixing. (Hint: If the l'th row of M has all entries larger than or equal to $\alpha > 0$, choose $\rho \in \mathcal{P}_E$ with $\rho^{(l)} = 1$ and $\rho^{(i)} = 0$ for $i \neq l$.)

Proof. We take ρ as in the hint. Then $M_{i,j} \geq \alpha \rho^{(i)}$ for all j and all $i \neq l$ because $\rho^{(i)} = 0$ for those i. Further, we have $M_{l,j} \geq \alpha = \alpha \rho^{(l)}$ by assumption and because we have set $\rho^{(l)} = 1$. Hence $M_{i,j} \geq \alpha \rho^{(i)}$ for all i, j, establishing the mixing condition from the proof of Theorem 2.24. □

Exercise 2.10.

Exercise 2.11 (**Assessed Exercise 8**)**.** Consider the Random walk on the circle (Example 2.17).

1. By analysing the Markov matrix M or otherwise, find an invariant distribution of this Markov process.

2. Demonstrate that this Markov process can have at most one invariant distribution.

Exercise 2.12 (**Assessed Exercise 9**)**.** 1. Consider a Markov matrix of the form

$$M = \begin{pmatrix} 0 & * & 0 & 0 \\ * & * & 0 & 0 \\ * & * & * & * \\ 0 & 0 & * & 0 \end{pmatrix}$$

where "$*$" indicates a nonzero entry. Show that M is not irreducible but nonetheless has a unique invariant distribution, which must be of the form $\pi = (0, 0, *, *)$, where again "$*$" indicates a nonzero entry. (Hint: show that M is mixing and revisit the proof of Thm. 2.24.)

2. Consider a Markov matrix of the form

$$M = \left(\begin{array}{c|c} M_1 & 0 \\ \hline 0 & M_2 \end{array}\right)$$

where the blocks M_1, M_2 are irreducible Markov matrices. Show that M is not irreducible. Furthermore, find at least two invariant distributions for M. (Hint: Use the invariant distributions π_1, π_2 of M_1, M_2, respectively).

3. Consider a Markov process $\{X_n, n \in \mathbb{N}\}$ on $E = \{1, 2\}$ with $X_0 = 1$ and a Markov matrix of the form

$$M = \left(\begin{array}{c|c} \alpha & 0 \\ \hline 1-\alpha & 1 \end{array}\right)$$

for some $0 < \alpha < 1$. Let τ be the index n at which $X_n = 2$ for the first time. Compute $P(\tau = n)$ as a function of n.

2.7 Introduction to ergodic theory of Markov processes with general state space

Ergodic theory of Markov processes with general state space is again a vast topic which we can only scrape the surface of. Fix a Markov pair (K, π) on

some state space $(E, \mathcal{B}_E)$, and let $\phi : E \to \mathbb{R}$ be a measurable function so that $\int_E |\phi(z)|K(\mathrm{d}z, x) < \infty$ for π–almost all x. We introduce the shorthand $K\phi$ for the function defined through

$$K\phi(x) = \int_E \phi(z)K(\mathrm{d}z, x) \tag{2.26}$$

for all $x \in E$, while $K\pi$ denotes the distribution defined through

$$K\pi(B) := \int_E K(B, x)\pi(\mathrm{d}x) \tag{2.27}$$

for all $B \in \mathcal{B}_E$.

Given that the entire distribution of a homogenous Markov process is uniquely defined through its Markov pair (K, π), it is probably not surprising that stationarity and ergodicity of such processes (which are concepts which usually refer to the entire distribution) can be characterised in terms of the Markov pair (K, π) only. This will be demonstrated in the next theorem after the following

Definition 2.29. Let (K, π) and ϕ be as above.

1. The distribution π is *invariant* with respect to the kernel K if
$$\pi = K\pi$$
2. The function ϕ is *invariant* with respect to the pair (K, π) if
$$\phi(x) = K\phi(x)$$
for π–allmost all $x \in E$.

Theorem 2.30. *Consider a Markov process $\{X\}$ with Markov pair (K, π).*

1. *The process is stationary if and only if π is invariant with respect to the kernel K*
2. *The process is ergodic if and only if π is invariant for K and every bounded invariant function ϕ is π–almost surely equal to a constant.*

Proof. To prove the first item, we use the same notation as in Exercise 2.7 (Chapman–Kolmogorov–Identity), that is, we let $\pi_n := P_{X_n}$. Clearly $\pi_0 = \pi$. We now invoke Equation (2.15) in the form

$$\begin{aligned}
&\mathbb{P}(X_{n+k} \in B_k, \dots, X_{n+1} \in B_1) \\
&= \mathbb{P}(X_{n+k} \in B_k, \dots, X_{n+1} \in B_1, X_n \in E, \dots X_1 \in E) \\
&= \int_E \cdots \int_E \mathbb{1}_{B_k}(x_{n+k})K(\mathrm{d}x_{n+k}, x_{n+k-1}) \cdot \ldots \cdot \mathbb{1}_{B_1}(x_{n+1})K(\mathrm{d}x_{n+1}, x_n) \\
&\quad \cdot K(\mathrm{d}x_n, x_{n-1}) \cdot \ldots \cdot K(\mathrm{d}x_2, x_1) \cdot \pi_1(\mathrm{d}x_1),
\end{aligned}$$

using that $\mathbb{1}_E = 1$. The integrals over $x_1, \ldots, x_{n-1}$ can be carried out using the Chapman-Kolmogorov-Identity. Using also the fact that $\pi_l = \pi$ for all $l \in \mathbb{N}$ we obtain

$$\begin{aligned}&\mathbb{P}(X_{n+k} \in B_k, \ldots, X_{n+1} \in B_1)\\ &= \int_E \cdots \int_E \mathbb{1}_{B_k}(x_{n+k}) K(\mathrm{d}x_{n+k}, x_{n+k-1}) \cdot \ldots \cdot \mathbb{1}_{B_1}(x_{n+1}) K(\mathrm{d}x_{n+1}, x_n) \cdot \pi(\mathrm{d}x_n).\end{aligned}$$

If we rename the integration variables $x_n, \ldots, x_{n+k}$ with, say $z_0, \ldots, z_k$, any mention of n on the right hand side disappears, and we obtain that the left hand side does not depend on n, showing that $\{X_n\}$ is stationary. If on the other hand $\{X_n\}$ is stationary, then $\pi_n = P_{X_n}$ does not depend on n.

To prove the second item, we need the following consequence of the Markov property, which is proved in Exercise 2.13. If $\psi : E^\infty \to \mathbb{R}$ is bounded and measurable, then

$$\mathbb{E}(\psi(X_{n+1}, X_{n+2}, \ldots)|X_1, \ldots, X_n) = \mathbb{E}(\psi(X_{n+1}, X_{n+2}, \ldots)|X_n). \tag{2.28}$$

Furthermore, for any bounded and measurable ψ we define $\bar{\psi}$ through

$$\bar{\psi}(x) := \mathbb{E}(\psi(X_{n+1}, X_{n+2}, \ldots)|X_n = x).$$

Note that $\bar{\psi}$ for given ψ only depends on the distribution of $\{X_n, X_{n+1}, \ldots\}$ and is therefore independent of n.

After these preliminaries, we now assume that any (K, π)–invariant function ϕ is π-a.s. constant and want to establish ergodicity of $\{X_n\}$. To this end, let $\psi : E^\infty \to \mathbb{R}$ be bounded, measurable and invariant for $\{X_n\}$. Then

$$\begin{aligned}\bar{\psi}(X_n) &= \mathbb{E}(\psi(X_{n+1}, X_{n+2}, \ldots)|X_n)\\ &= \mathbb{E}(\psi(X_{n+2}, X_{n+3}, \ldots)|X_n) \quad \text{(by invariance of } \psi)\\ &= \mathbb{E}\left(\mathbb{E}(\psi(X_{n+2}, X_{n+3}, \ldots)|X_{n+1}, X_n)|X_n\right) \quad \text{(by law of iterated exp.)}\\ &= \mathbb{E}\left(\mathbb{E}(\psi(X_{n+2}, X_{n+3}, \ldots)|X_{n+1})|X_n\right) \quad \text{(by Eq. 2.28)}\\ &= \mathbb{E}\left(\bar{\psi}(X_{n+1})|X_n\right) \quad \text{(by def. of } \bar{\psi})\\ &= K\bar{\psi}(X_n) \quad \text{(by Eq. 2.13).}\end{aligned}$$

We obtain that $\bar{\psi}$ is (K, π)-invariant and therefore $\bar{\psi}$ is π-almost surely equal to a constant c. Using the definition of $\bar{\psi}$ in conjuction with Equation (2.28) and the invariance of ψ once again we find

$$\begin{aligned}c &= \bar{\psi}(X_n)\\ &= \mathbb{E}(\psi(X_{n+1}, X_{n+2}, \ldots)|X_n)\\ &= \mathbb{E}(\psi(X_{n+1}, X_{n+2}, \ldots)|X_1, \ldots, X_n)\\ &= \mathbb{E}(\psi(X_1, X_2, \ldots)|X_1, \ldots, X_n) \quad \text{(by invariance of } \psi).\end{aligned}$$

Now we take the limit $n \to \infty$ and use the Martingale result in Exercise 2.6 on the right hand side to conclude that $\psi(X_1, X_2, \ldots)$ is almost surely equal to a constant.

To prove the other direction, start with assuming that $\{X_n, n \in \mathbb{N}\}$ is ergodic. If $\phi : E \to \mathbb{R}$ is measurable and bounded, the Ergodic Theorem implies that

$$\frac{1}{n} \sum_{k=0}^{n-1} \phi(X_k) \to c$$

in L_1 for some constant c. Take the conditional expectation with respect to X_1 on both sides. Exercise 2.10 shows that this can be interchanged with the L_1-limit in the Ergodic Theorem. We obtain

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}(\phi(X_k)|X_1) \overset{L_1}{\to} c. \tag{2.29}$$

On the other hand, it is easy to see from Equation (2.28) that $\mathbb{E}(\phi(X_k)|X_1) = K^{k-1}\phi(X_1)$ for any $k \in \mathbb{N}$. For a (K, π)–invariant ϕ, this means $\mathbb{E}(\phi(X_k)|X_1) = \phi(X_1)$. Using this in Equation (2.29), we find that ϕ is equal to c almost surely with respect to π. □

Exercise 2.13. If $\{X_n, n \in \mathbb{N}\}$ is a Markov process and $\psi : E^\infty \to \mathbb{R}$ is bounded and measurable, demonstrate Equation (2.28).

Proof. We recall that a *finite dimensional rectangle* in $\mathcal{B}(E^\infty)$ is a set of the form $\{x = (x_1, x_2, \ldots) \in E^\infty; x_1 \in B_1, \ldots, x_k \in B_k\}$ for some sets $B_1, \ldots, B_k \in \mathcal{B}(E)$ and some $k \in \mathbb{N}$. If $\psi = \mathbb{1}_B$ where B is a finite dimensional rectangle, Equation (2.28) follows (after some calculations) from the identity (2.15). If $\psi = \mathbb{1}_B$ for some general set $B \in \mathcal{B}(E^\infty)$, it follows from the MET (Thm. 1.5) that ψ can be approximated in L_1 by functions of the form $\sum_{l=1}^{n} \mathbb{1}_{B_n}$ where $B_1, \ldots, B_n$ are finite dimensional rectangles; since the identity (2.28) is linear, it also holds for functions of this form. We have established that Equation (2.28) holds for general characteristic functions, and again by linearity it holds for general simple functions. Since any bounded and measurable ψ can be approximated in L_1 by simple functions, the proof is complete. □

Exercise 2.14.

2.8 A simple sufficient condition for ergodicity

In the previous section, we saw that stationarity and ergodicity of a homogenous Markov process can be studied in terms of the corresponding Markov pair (K, π), checking if π is invariant (for stationarity) or, in addition, if every (K, π)–invariant bounded ϕ is π–almost surely equal to a constant. In the present section, we will prove a theorem which provides a simple criterion sufficient for ergodicity. It is worth stressing though that it is by no means necessary, and a plethora of different approaches have been investigated.

Theorem 2.31. *Consider a Markov kernel K on a state space E. Suppose that there exists a constant $c > 0$ and a probability measure ν on $\mathcal{B}_E$ such that $K(B, x) \geq c\nu(B)$ for all $B \in \mathcal{B}_E$ and all $x \in E$. Then there exists a unique K-invariant probability distribution π. Furthermore, a Markov process with pair (K, π) is ergodic.*

Definition 2.29 should make it plain that an invariant distribution π is a fixed point of K if we regard the latter as a mapping on the space of probability distributions over E into itself. Our proof will proceed by finding such a fixed point through Banach's fixed point theorem. With regards to ergodicity, a fundamental statement in ergodic says that if there is only one invariant measure, it must be ergodic, so given that Banach's theorem provides uniqueness of the invariant measure, ergodicity would follow. We will however give and independent and simple proof that an invariant function is almost surely constant.

Proof. Our proof will rely on results established in Exercises 2.15–2.18. We let $\mathcal{P}_E$ be the space of probability measures over $(E, \mathcal{B}_E)$ and equip $\mathcal{P}_E$ with the *total variation metric*

$$\mathrm{TV}(\pi, \rho) := \sup_{B \in \mathcal{B}_E} |\pi(A) - \rho(A)|.$$

It can be shown (see Ex. 2.15) that TV is indeed a metric and $(\mathcal{P}_E, \mathrm{TV})$ is a complete metric space. We now need to show that K is *contracting*, that is there exists an $\alpha < 1$ such that $\mathrm{TV}(K\pi, K\rho) \leq \alpha\, \mathrm{TV}(\pi, \rho)$ for any $\pi, \rho \in \mathcal{P}_E$. We first note that the lower bound on K in our conditions automatically define an *upper* bound as well, namely

$$K(B, x) = 1 - K(B^{\complement}, x) \leq 1 - c\nu(B^{\complement}) = 1 - c(1 - \nu(B)).$$

Next we fix $\pi, \rho \in \mathcal{P}_E$ and invoke Exercise 2.16 to get that for any $B \in \mathcal{B}_E$, we have

$$\begin{aligned} K\pi(B) - K\rho(B) &= \int_E K(B,x)\,(\pi(\mathrm{d}x) - \rho(\mathrm{d}x)) \\ &= \int_E (K(B,x) - c\nu(B) + c\nu(B))\,(\pi(\mathrm{d}x) - \rho(\mathrm{d}x)) \\ &= \int_E (K(B,x) - c\nu(B))\,(\pi(\mathrm{d}x) - \rho(\mathrm{d}x)) \\ &= \sup_{x\in E}(K(B,x) - c\nu(B))\,\mathrm{TV}(\pi,\rho), \end{aligned} \tag{2.30}$$

where in the third line we have used that $c\nu(B)$ does not depend on x. Our use of Exercise 2.16 is based on the fact that $x \to K(B,x) - c\nu(B)$ is a nonnegative function due to our assumptions. But using the upper bound we find

$$\sup_{x\in E}(K(B,x) - c\nu(B)) \le 1 - c(1-\nu(B)) - c\nu(B) \le 1 - c. \tag{2.31}$$

By combining Equations (2.30,2.31) and since B was arbitrary, we find the contraction property with $\alpha = 1 - c$.

To prove ergodicity, we introduce the space $\mathscr{C}$ of measurable and bounded functions ϕ so that $\pi(\phi) = 0$, equipped with the norm $\|\phi\| := \sup_{x,y\in\mathbb{E}} |\phi(x) - \phi(y)|$. Note that if $\|\phi\| = 0$ then ϕ must be constant, but since $\pi(\phi) = 0$ this constant must be zero. In Exercise 2.17 we prove that $(\mathscr{C}, \|.\|)$ is a Banach space. We let ϕ be a bounded measurable and invariant function. Since the function $\phi - b$, where b is a constant, is still an invariant function, by taking $b = \pi(\phi)$ we may assume $\phi \in \mathscr{C}$. We now want to show that $\phi = 0$. In Exercise 2.18 we show that for a *nonnegative*, measurable and bounded function ϕ we have $K\phi(x) - K\phi(y) \le (1-c)\sup_x \phi(x)$ for all $x \in E$. A function $\phi \in \mathscr{C}$ can be decomposed as $\phi = \phi_+ - \phi_-$, where $\phi_+(x) := \max\{\phi(x), 0\}$ and $\phi_-(x) := \max\{-\phi(x), 0\}$ are nonnegative, measurable and bounded functions. With these, we have

$$\begin{aligned} \phi(x) - \phi(y) &= K\phi(x) - K\phi(y) \qquad \text{(By invariance of } \phi) \\ &= K\phi_+(x) - K\phi_+(y) - (K\phi_-(x) - K\phi_-(y)) \qquad \text{(decomposing } \phi) \\ &\le (1-c)(\sup_z \phi_+(z) + \sup_z \phi_-(z)) \qquad \text{(Exercise 2.18)} \\ &\le (1-c)\|\phi\| \qquad \text{(Exercise 2.17.2).} \end{aligned}$$

Since x, y are arbitrary, we have $\|\phi\| \le (1-c)\|\phi\|$ which implies $\phi = 0$ since $1 - c < 1$. □

Exercise 2.15. Show that TV is a metric on $\mathcal{P}_E$, turning $(\mathcal{P}_E, \mathrm{TV})$ into a complete metric space.

Exercise 2.16. For $\pi, \rho \in \mathcal{P}_E$ and $\phi : E \to \mathbb{R}$ bounded, nonnegative and measurable, show that

$$|\int_E \phi(x)\pi(\mathrm{d}x) - \int_E \phi(x)\rho(\mathrm{d}x)| \leq \sup_{x\in E} \phi(x)\, \mathrm{TV}(\pi, \rho).$$

Proof. Write $c := \sup_{x\in E} \phi(x)$; we have

$$\int_E \phi(x)\pi(\mathrm{d}x) = \int_0^c \pi\{x; \phi(x) \geq z\}\, \mathrm{d}z,$$

and the same for ρ. Hence

$$\begin{aligned}
&|\int_E \phi(x)\pi(\mathrm{d}x) - \int_E \phi(x)\rho(\mathrm{d}x)| \\
&\leq |\int_0^c \pi(\phi(x) \geq z) - \rho(\phi(x) \geq z)\, \mathrm{d}z| \\
&\leq c\, \mathrm{TV}(\pi, \rho).
\end{aligned}$$

□

Exercise 2.17. Show that in the context of Theorem 2.31

1. $(\mathscr{C}, \|.\|)$ is a Banach space.
2. For an element $\phi \in \mathscr{C}$ it holds that $\|\phi\| = \sup_z \phi_+(z) + \sup_z \phi_-(z)$.

Partial solution to item 1. To show that $(\mathcal{C}, \|.\|)$ is complete, let $\{\phi_n, n \in \mathbb{N}\}$ be a Cauchy sequence in the norm $\|.\|$. Define the sequence $\{\psi_n\}$ of functions through $\psi_n(x, y) := \phi_n(x) - \phi_n(y)$. Due to how the norm $\|.\|$ is defined, we find that $\{\psi_n\}$ is a Cauchy sequence in the space $\mathcal{D}$ of bounded measurable functions on $E \times E$ with the usual sup norm $\|\!|f|\!\| := \sup_{x,y\in E} |f(x, y)|$. This space is complete, as can be proved exactly as was shown in the tutorial by group GOLF (here we don't even need to prove continuity). Hence there is a $\psi \in \mathcal{D}$ such that $\sup_{x,y\in E} |\psi(x, y) - (\phi_n(x) - \phi_n(y))| \to 0$.

We now define the function ϕ through $\phi(x) := \psi(x, z) - \int \psi(r, z)\pi(\mathrm{d}r)$, where z is an arbitrary element of E. Clearly, ϕ is bounded, measurable, and $\int \phi(r)\pi(\mathrm{d}r) = 0$, so $\phi \in \mathcal{C}$. We now claim that $\|\phi_n - \phi\| \to 0$. Indeed,

$$\begin{aligned}
&|\phi_n(x) - \phi_n(y) - (\phi(x) - \phi(y))| \\
&= |\phi_n(x) - \phi_n(y) - (\psi(x, z) - \psi(y, z))| \\
&= |\phi_n(x) - \phi_n(z) - \psi(x, z) - (\phi_n(y) - \phi_n(z) - \psi(y, z))| \\
&\leq |\phi_n(x) - \phi_n(z) - \psi(x, z)| + |\phi_n(y) - \phi_n(z) - \psi(y, z)| \\
&= |\psi_n(x, z) - \psi(x, z)| + |\psi_n(y, z) - \psi(y, z)|.
\end{aligned}$$

Taking the sup over x, y, z we find $\|\phi_n - \phi\| \le 2|||\psi_n - \psi|||$. The right hand side goes to zero, finishing the proof. □

Alternative solution to item 1. Since $\phi(x) = \int \phi(x) - \phi(y)\pi(\mathrm{d}y)$ for any $\phi \in \mathscr{C}$, we have $\sup_x |\phi(x)| \le \int \sup_{x,y} |\phi(x) - \phi(y)|\pi(\mathrm{d}y) = \|\phi\|$. So if we let $|||\phi||| := \sup_x |\phi(x)|$ be the usual sup norm, we have shown $|||\phi||| \le \|\phi\|$. On the other hand, by the triangle inequality $\sup_{x,y} |\phi(x) - \phi(y)| \le 2 \sup_x |\phi(x)|$, so $\|\phi\| \le 2|||\phi|||$. Hence the two norms are equivalent. Since the space of bounded measurable functions is complete with respect to the $|||.|||$-norm, we can conclude that there is a bounded measurable function ϕ so that $\|\phi_n - \phi\| \to 0$. It remains to show that $\int \phi(x)\pi(\mathrm{d}x) = 0$ but this follows from bounded convergence. □

Exercise 2.18. Show that in the context of Theorem 2.31, a nonnegative, measurable and bounded function ϕ satisfies the estimate

$$K\phi(x) - K\phi(y) \le (1 - c) \sup_x \phi(x) \qquad \text{for all } x, y \in E.$$

Proof. Define the probability kernel L through $L(B, x) := \frac{1}{1-c}(K(B, x) - c\nu(B))$ for all $B \in \mathcal{B}_E$ and $x \in E$; note that due to our assumptions, this is indeed a probability kernel. Further, $K\phi(x) - K\phi(y) = (1-c)(L\phi(x) - L\phi(y))$. According to Exercise 2.16 we have $L\phi(x) - L\phi(y) \le \sup_{x \in E} \phi(x)\, \mathrm{TV}(L(., x), L(., y))$. But $\mathrm{TV} \le 1$ on probabilities. Combining these facts, we get the result. □

Appendix A

Miscellaneous proofs

A.1 Proof of Theorem 2.19, item 1

.

A.2 Completeness of the total variation metric

Appendix B

Solutions to selected exercises

Exercise 1.1

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 $\emptyset$ is a subset of Ω, $\emptyset \in 2^\Omega$. Second, if $A \in 2^\Omega$ then $A \subset \Omega$ which implies that $A^\complement \subset \Omega$. This means that $A^\complement \in 2^\Omega$. 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 $\emptyset \in \mathcal{S}_1$ and $\emptyset \in \mathcal{S}_2$ we have that $\emptyset \in \mathcal{S}_1 \cap \mathcal{S}_2$. 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 1.2

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

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

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

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 $\cap B_n = \emptyset$. 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 $\cap A_k = \emptyset$, 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 $i \neq j$, $B_i \cap B_j = \emptyset$. 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 $A_1 \supset A_2 \supset \dots$. We define

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

   for all $n \in \mathbb{N}$. We have then that for $i \neq j$, $B_i \cap B_j = \emptyset$ 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,

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

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

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

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

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

   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 $A_1 \subset A_2 \subset \dots$ 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 $i \neq j$, $B_i \cap B_j = \emptyset$ 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 $B_1 \subset B_2 \subset \dots$ 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 $A_1, A_2, \dots$ 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 1.3

1. We check Definition 3.
   (a) $\emptyset \in \mathcal{B}$, and $f^{-1}(\emptyset) = \emptyset$, so $\emptyset \in \mathcal{A}_0$.
   (b) Let $A \in \mathcal{A}_0$. Then $\exists B \in \mathcal{B} : f^{-1}(B) = A$. Now $B^{\complement} \in \mathcal{B}$ and $f^{-1}(B^{\complement}) = f^{-1}(B)^{\complement} = A^{\complement}$, hence $A^{\complement} \in \mathcal{A}_0$.
   (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 $f^{-1}(\emptyset) = \emptyset \in \mathcal{A}$, $\emptyset \in \mathcal{B}_0$.
   (b) If $f^{-1}(B) \in \mathcal{A}$, then $f^{-1}(B^{\complement}) = f^{-1}(B)^{\complement} \in \mathcal{A}$. This shows $B \in \mathcal{B}_0 \implies B^{\complement} \in \mathcal{B}_0$.
   (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 1.5

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

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 $\bigcap_n M_n = \emptyset \implies \bigcup_n M_n = \Omega$.

3. We know that f_n, g and $\mathbb{1}_{M_n}$ are simple, so $f_n \cdot \mathbb{1}_{M_n}, g \cdot \mathbb{1}_{M_n}, \sum \mathbb{1}_{M_n}$ are too. Now $f_n \geq f_n \cdot \mathbb{1}_{M_n} \geq (g - \epsilon) \cdot \mathbb{1}_{M_n}$ due to the definition of M_n. Now the relation (1.7) 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 $g\mathbb{1}_{M_n} = \sum_{k=1}^{m} g_k \cdot \mathbb{1}_{B_k \cap M_n}$. 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 $\int \mathbb{1}_{M_n} g \mathrm{d}\mathbb{P} = \sum_{k=1}^{m} g_k \cdot \mathbb{P}(B_k \cap M_n) \xrightarrow{n\to\infty} \sum_{k=1}^{m} g_k \cdot \mathbb{P}(B_k) = \int g \mathrm{d}\mathbb{P}$.

Exercise 1.6

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 $n \cdot f(\omega) \geq \mathbb{1}_{A_n}(\omega)\ \forall \omega \in \Omega$. This gives $0 = n \int f \mathrm{d}\mathbb{P} \geq n \cdot \int \mathbb{1}_{A_n} \mathrm{d}\mathbb{P} = \mathbb{P}(A_n)$, 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 1.8

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 1.9

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 1.10

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 $\mathcal{A}$.

Bibliography

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

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

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

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

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

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

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

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

[9] Achim Klenke. *Probability Theory*. Springer, 2014.

[10] Bryan P. Rynne and Martin A. Youngson. *Linear functional analysis*. Springer Undergraduate Mathematics Series. Springer-Verlag London, Ltd., London, 2000.