Conditioning and Markov properties

Anders Rønn-Nielsen

Ernst Hansen

Department of Mathematical Sciences
University of Copenhagen

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen

ISBN 978-87-7078-980-6

Contents

Preface

The present lecture notes are intended for the course "Beting". The purpose is to give a detailed probabilistic introduction to conditional distributions and how this concept is used to define and describe Markov chains and Bayesian networks.

The chapters 1–4 are mainly based on different sets of lecture notes written by Ernst Hansen but are adapted to suit students with the knowledge obtained in the courses MI, VidSand1, and VidSand2. Parts of these chapters also contain material inspired by lecture notes written by Martin Jacobsen and Søren Tolver Jensen. Chapter 5 is an adapted version of a set of lecture notes written by Søren Tolver Jensen and Martin Jacobsen. These notes themselves were inspired by earlier notes by Søren Feodor Nielsen. Chapter 6 contains some standard results on Bayesian networks - though both formulations and proofs are formulated in the rather general framework from chapter 3 and 4.

There are exercises in the end of each chapter and hints to selected exercises in the end of the notes. Some exercises are adapted versions of exercises and exam exercises from previous lecture notes. Others are inspired by examples and results collected from a large number of monographs on Markov chains, stochastic simulation, and probability theory in general.

I am grateful to both students and the teaching assistants from the last two years, Ketil Biering Tvermosegaard and Daniele Cappelletti, who have contributed to the notes by identifying mistakes and suggesting improvements.

Anders Rønn-Nielsen
København, 2014

Chapter 1

Conditional distributions

Let $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ be two measurable spaces. In this chapter we shall discuss the relation between measures on the product space $(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$ and measures on the two marginal spaces $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. More precisely we will see that measures on the product space can be constructed from measures on the two marginal spaces. A particularly simple example is a product measure $\mu \otimes \nu$ where μ and ν are measures on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively.

The two factors $\mathcal{X}$ and $\mathcal{Y}$ will not enter the setup symmetrically in the more general construction.

1.1 Markov kernels

Definition 1.1.1. *A* $(\mathcal{X}, \mathbb{E})$–Markov Kernel *on* $(\mathcal{Y}, \mathbb{K})$ *is a family of probability measures* $(P_x)_{x \in \mathcal{X}}$ *on* $(\mathcal{Y}, \mathbb{K})$ *indexed by points in* $\mathcal{X}$ *such that the map*

$$x \mapsto P_x(B)$$

is $\mathbb{E} - \mathbb{B}$*–measurable for every fixed* $B \in \mathbb{K}$.

Theorem 1.1.2. *Let* $(\mathcal{X}, \mathbb{E})$ *and* $(\mathcal{Y}, \mathbb{K})$ *be measurable spaces, let* ν *be a* σ*–finite measure on* $(\mathcal{Y}, \mathbb{K})$*, and let* $f \in \mathcal{M}^+(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$ *have the property that*

$$\int f(x, y)\, \mathrm{d}\nu(y) = 1 \qquad \text{for all } x \in \mathcal{X}\,.$$

Then $(P_x)_{x\in\mathcal{X}}$ *given by*

$$P_x(B) = \int_B f(x,y)\,\mathrm{d}\nu(y) \qquad \text{for all } B \in \mathbb{K},\ x \in \mathcal{X}$$

is a $(\mathcal{X},\mathbb{E})$*–Markov Kernel on* $(\mathcal{Y},\mathbb{K})$*.*

Proof. For each fixed set $B \in \mathbb{K}$ we need to argue that

$$x \mapsto \int 1_{\mathcal{X}\times B}(x,y) f(x,y)\,\mathrm{d}\nu(y)$$

is an $\mathbb{E}$–measurable function. That is a direct result of EH Theorem 8.7 applied to the function $1_{\mathcal{X}\times B} f$. □

Lemma 1.1.3. *If* $(\mathcal{Y},\mathbb{K})$ *has an intersection–stable generating system* $\mathbb{D}$*, then* $(P_x)_{x\in\mathcal{X}}$ *is a* $(\mathcal{X},\mathbb{E})$*–Markov Kernel on* $(\mathcal{Y},\mathbb{K})$ *if only*

$$x \mapsto P_x(D)$$

is $\mathbb{E}-\mathbb{B}$*-measurable for all fixed* $D \in \mathbb{D}$*.*

Proof. Define

$$\mathbb{H} = \{F \in \mathbb{K} \,:\, x \mapsto P_x(F) \text{ is } \mathbb{E}-\mathbb{B}\text{–measurable}\}$$

It is easily checked that $\mathbb{H}$ is a Dynkin Class. Since $\mathbb{D} \subseteq \mathbb{H}$, we have $\mathbb{H} = \mathbb{K}$. □

Lemma 1.1.4. *Let* $(P_x)_{x\in\mathcal{X}}$ *be a* $(\mathcal{X},\mathbb{E})$*–Markov kernel on* $(\mathcal{Y},\mathbb{K})$*. For each* $G \in \mathbb{E}\otimes\mathbb{K}$ *the map*

$$x \mapsto P_x(G^x)$$

is $\mathbb{E}-\mathbb{B}$*–measurable.*

Proof. Let

$$\mathbb{H} = \{G \in \mathbb{E}\otimes\mathbb{K} \,:\, x \mapsto P_x(G^x) \text{ is } \mathbb{E}-\mathbb{B}\text{–measurable}\}$$

and consider a product set $A \times B \in \mathbb{E}\otimes\mathbb{K}$. Then

$$(A\times B)^x = \begin{cases} \emptyset & \text{if } x \notin A \\ B & \text{if } x \in A \end{cases}$$

such that

$$P_x((A\times B)^x) = \begin{cases} 0 & \text{if } x \notin A \\ P_x(B) & \text{if } x \in A \end{cases} = 1_A(x)\cdot P_x(B)$$

This is a product of two $\mathbb{E}-\mathbb{B}$–measurable functions. Hence it is $\mathbb{E}-\mathbb{B}$–measurable. Thereby we conclude that $\mathbb{H}$ contains all product sets, and since this is a intersection stable generating system for $\mathbb{E}\otimes\mathbb{K}$, we have $\mathbb{H}=\mathbb{E}\otimes\mathbb{K}$, if we can show that $\mathbb{H}$ is a Dynkin class:

We already have that $\mathcal{X}\times\mathcal{Y}\in\mathbb{H}$ – it is a product set! Assume that $G_1\subseteq G_2$ are two sets in $\mathbb{H}$. Then obviously also $G_1^x\subseteq G_2^x$ for all $x\in\mathcal{X}$, and

$$(G_2\setminus G_1)^x = G_2^x\setminus G_1^x\,.$$

Then

$$P_x((G_2\setminus G_1)^x) = P_x(G_2^x) - P_x(G_1^x)$$

which is a difference between two measurable functions. Hence $G_2\setminus G_1\in\mathbb{H}$.

Finally, assume that $G_1\subseteq G_2\subseteq\cdots$ is an increasing sequence of $\mathbb{H}$–sets. Similarly to above we have $G_1^x\subseteq G_2^x\subseteq\cdots$ and

$$\left(\bigcup_{n=1}^{\infty} G_n\right)^x = \bigcup_{n=1}^{\infty} G_n^x\,.$$

Then

$$P_x\left(\left(\bigcup_{n=1}^{\infty} G_n\right)^x\right) = P_x\left(\bigcup_{n=1}^{\infty} G_n^x\right) = \lim_{n\to\infty} P_x(G_n^x)$$

This limit is $\mathbb{E}-\mathbb{B}$–measurable, since each of the functions $x\mapsto P_x(G_n^x)$ are measurable. Then $\bigcup_{n=1}^{\infty} G_n\in\mathbb{H}$. □

1.2 Integration of Markov kernels

Theorem 1.2.1. *Let μ be a probability measure on $(\mathcal{X},\mathbb{E})$ and let $(P_x)_{x\in\mathcal{X}}$ be a $(\mathcal{X},\mathbb{E})$–Markov kernel on $(\mathcal{Y},\mathbb{K})$. There exists a uniquely determined probability measure λ on $(\mathcal{X}\times\mathcal{Y},\mathbb{E}\otimes\mathbb{K})$ satisfying*

$$\lambda(A\times B) = \int_A P_x(B)\,\mathrm{d}\mu(x)$$

for all $A\in\mathbb{E}$ and $B\in\mathbb{K}$.

The probability measure λ constructed in Theorem 1.2.1 is called **the integration** of $(P_x)_{x\in\mathcal{X}}$ with respect to μ. The interpretation is that λ describes an experiment on $\mathcal{X}\times\mathcal{Y}$ that is performed in two steps: The first step is drawing $x\in\mathcal{X}$. The second step is drawing $y\in\mathcal{Y}$ according to a probability measure that is determined by x.

Proof. The uniqueness follows, since λ is determined on all product sets and these form an intersection stable generating system for $\mathbb{E} \otimes \mathbb{K}$.

In order to prove the existence, we define

$$\lambda(G) = \int P_x(G^x) \, \mathrm{d}\mu(x)$$

For each $G \in \mathbb{E} \otimes \mathbb{K}$ the integrand is measurable according to Lemma 1.1.4. It is furthermore non–negative, such that $\lambda(G)$ is well–defined with values in $[0, \infty]$.

Now let $G_1, G_2, \ldots$ be a sequence of disjoint sets in $\mathbb{E} \otimes \mathbb{K}$. Then for each $x \in \mathcal{X}$ the sets $G_1^x, G_2^x, \ldots$ are disjoint as well. Hence

$$\lambda\Big(\bigcup_{n=1}^{\infty} G_n\Big) = \int P_x\Big(\Big(\bigcup_{n=1}^{\infty} G_n\Big)^x\Big) \, \mathrm{d}\mu(x) = \int \sum_{n=1}^{\infty} P_x(G_n^x) \, \mathrm{d}\mu(x) = \sum_{n=1}^{\infty} \lambda(G_n)$$

In the second equality we have used that each P_x is a measure, and in the third equality we have used monotone convergence to interchange integration and summation. From this we have that λ is a measure. And since

$$\lambda(\mathcal{X} \times \mathcal{Y}) = \int P_x((\mathcal{X} \times \mathcal{Y})^x) \, \mathrm{d}\mu(x) = \int P_x(\mathcal{Y}) \, \mathrm{d}\mu(x) = \int 1 \, \mathrm{d}\mu(x) = 1$$

we obtain, than λ is actually a probability measure. Finally, it follows that

$$\lambda(A \times B) = \int P_x((A \times B)^x) \, \mathrm{d}\mu(x) = \int 1_A(x) P_x(B) \, \mathrm{d}\mu(x) = \int_A P_x(B) \, \mathrm{d}\mu(x)$$

for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$. □

Corollary 1.2.2. *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and let $(P_x)_{x \in \mathcal{X}}$ be a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. Let λ be the integration of $(P_x)_{x \in \mathcal{X}}$ with respect to μ. Then λ satisfies*

$$\begin{aligned} \lambda(A \times \mathcal{Y}) &= \mu(A) && \text{for all } A \in \mathbb{E} \\ \lambda(\mathcal{X} \times B) &= \int P_x(B) \, \mathrm{d}\mu(x) && \text{for all } B \in \mathbb{K} \end{aligned}$$

Proof. The second statement is obvious. For the first result just note that $P_x(\mathcal{Y}) = 1$ for all $x \in \mathcal{X}$. □

The probability measure on $(\mathcal{Y}, \mathbb{K})$ defined by $\lambda(\mathcal{X} \times B)$ is called **the mixture** of the Markov kernel with respect to μ.

Example 1.2.3. Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and let ν be a probability measure

on $(\mathcal{Y}, \mathbb{K})$. Define $P_x = \nu$ for all $x \in \mathcal{X}$. Then, trivially, $(P_x)_{x\in\mathcal{X}}$ is a $\mathcal{X}$–Markov kernel on $\mathcal{Y}$. Let λ be the integration of this kernel with respect to λ. Then for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$

$$\lambda(A \times B) = \int_A \nu(B)\, \mathrm{d}\mu(x) = \mu(A) \cdot \nu(B)\,.$$

The only measure satisfying this property is the product measure $\mu \otimes \nu$, so $\lambda = \mu \otimes \nu$. Hence a product measure is a particularly simple example of a measure constructed by integrating a Markov kernel. ∘

Example 1.2.4. Let μ be the Poisson distribution with parameter λ. For each $x \in \mathbb{N}_0$ we define P_x to be the binomial distribution with parameters (x, p). Then it is seen that $(P_x)_{x\in\mathbb{N}_0}$ is a $\mathbb{N}_0$–Markov kernel on $\mathbb{N}_0$.

Let ξ be the mixture of $(P_x)_{x\in\mathbb{N}_0}$ with respect to μ. This must be a probability measure on $\mathbb{N}_0$ and is thereby given by the probability function q. For $n \in \mathbb{N}_0$ we obtain

$$\begin{aligned} q(n) &= \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \frac{(\lambda p)^n}{n!} e^{-\lambda} \sum_{k=n}^{\infty} \frac{\big((1-p)\lambda\big)^{k-n}}{(k-n)!} \\ &= \frac{(\lambda p)^n}{n!} e^{-\lambda} e^{(1-p)\lambda} \\ &= \frac{(\lambda p)^n}{n!} e^{-\lambda p} \end{aligned}$$

Hence the mixture ξ is seen to be the Poisson distribution with parameter λp. ∘

Theorem 1.2.5 (Uniqueness of integration)**.** *Suppose that $(\mathcal{Y}, \mathbb{K})$ has a countable generating system that is intersection stable. Let μ and $\tilde{\mu}$ be two probability measures on $(\mathcal{X}, \mathbb{E})$ and assume that $(P_x)_{x\in\mathcal{X}}$ and $(\tilde{P}_x)_{x\in\mathcal{X}}$ are two $(\mathcal{X}, \mathbb{E})$–Markov kernels on $(\mathcal{Y}, \mathbb{K})$. Let λ be the integration of $(P_x)_{x\in\mathcal{X}}$ with respect to μ, and let $\tilde{\lambda}$ be the integration of $(\tilde{P}_x)_{x\in\mathcal{X}}$ with respect to $\tilde{\mu}$. Define*

$$E_0 = \{x \in \mathcal{X} \,:\, P_x = \tilde{P}_x\}$$

Then $\lambda = \tilde{\lambda}$ if and only if $\mu = \tilde{\mu}$ and $\mu(E_0) = 1$.

Proof. Let $(B_n)_{n\in\mathbb{N}}$ be a countable generating system for $(\mathcal{Y}, \mathbb{K})$. Then

$$E_0 = \bigcap_{n=1}^{\infty} \{x \in \mathcal{X} \,:\, P_x(B_n) = \tilde{P}_x(B_n)\}$$

from which we can conclude that $E_0 \in \mathbb{E}$.

Assume that $\mu = \tilde{\mu}$ and $\mu(E_0) = 1$. Then for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$ we have

$$\lambda(A \times B) = \int_{A \cap E_0} P_x(B)\, \mathrm{d}\mu(x) = \int_{A \cap E_0} \tilde{P}_x(B)\, \mathrm{d}\tilde{\mu}(x) = \tilde{\lambda}(A \times B)$$

and thereby $\lambda = \tilde{\lambda}$.

Assume conversely that $\lambda = \tilde{\lambda}$. According to Corollary 1.2.2 we have for all $A \in \mathbb{E}$

$$\mu(A) = \lambda(A \times \mathcal{Y}) = \tilde{\lambda}(A \times \mathcal{Y}) = \tilde{\mu}(A)$$

such that $\mu = \tilde{\mu}$.

The proof will be complete, if we can show that

$$\mu(\{x \in \mathcal{X} \,:\, P_x(B_n) \neq \tilde{P}_x(B_n)\}) = 0$$

for all $n \in \mathbb{N}$. Consider for this the set

$$E_n^+ = \{x \in \mathcal{X} \,:\, P_x(B_n) > \tilde{P}_x(B_n)\}\,.$$

Using this definition gives

$$\int_{E_n^+} (P_x(B_n) - \tilde{P}_x(B_n))\, \mathrm{d}\mu(x) = \lambda(E_n^+ \times B_n) - \tilde{\lambda}(E_n^+ \times B_n) = 0$$

and since the integrand is strictly positive on E_n^+, we can conclude that $\mu(E_n^+) = 0$. It is shown similarly that $\mu(E_n^-) = 0$, where

$$E_n^- = \{x \in \mathcal{X} \,:\, P_x(B_n) < \tilde{P}_x(B_n)\}\,.$$

□

1.3 Properties for the integration measure

In this section we will consider integration with respect to λ, where λ is the integrated measure of a Markov kernel $(P_x)_{x \in \mathcal{X}}$ with respect to some probability measure μ. We shall see, that such λ–integrals can be calculated by successive integration similar to what is known for product measures.

Lemma 1.3.1. *Let $(P_x)_{x \in \mathcal{X}}$ be a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$ and assume that $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ is $\mathbb{E} \otimes \mathbb{K}$–measurable. Then the map*

$$x \mapsto \int f(x, y)\, \mathrm{d}P_x(y) \tag{1.1}$$

is $\mathbb{E}$–measurable.

Proof. Firstly note that for fixed x then $f(x,y) = f \circ i_x(y)$ which is a $\mathbb{K}$–measurable function. Hence the integral in (1.1) is well–defined. Now assume that f is a simple function

$$f = \sum_{k=1}^{n} c_k 1_{G_k} \tag{1.2}$$

where $c_1, \ldots, c_n \in (0, \infty)$ and $G_1, \ldots, G_n$ are disjoint sets in $\mathbb{E} \otimes \mathbb{K}$. Since

$$1_{G_k}(x,y) = 1_{G_k^x}(y)$$

for all x and y, we obtain

$$\begin{aligned}\int f(x,y)\, \mathrm{d}P_x(y) &= \sum_{k=1}^{n} \int c_k 1_{G_k}(x,y)\, \mathrm{d}P_x(y) \\ &= \sum_{k=1}^{n} c_k \int 1_{G_k^x}(y)\, \mathrm{d}P_x(y) \\ &= \sum_{k=1}^{n} c_k P_x(G_k^x)\end{aligned}$$

According to Lemma 1.1.4 this is a linear combination of $\mathbb{E}$–measurable functions. Hence it is $\mathbb{E}$–measurable.

Now assume that f is a general function in $\mathcal{M}^+(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$. Then there exists a sequence $(f_n)_{n \in \mathbb{N}}$ of non–negative simple functions with $f_n(x,y) \uparrow f(x,y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. For fixed x we have from monotone convergence, that

$$\int f_n(x,y)\, \mathrm{d}P_x(y) \uparrow \int f(x,y)\, \mathrm{d}P_x(y)\,.$$

Hence the right hand side is the point–wise limit of $\mathbb{E}$–measurable functions. Thereby it is $\mathbb{E}$–measurable. □

Theorem 1.3.2 (Extended Tonelli)**.** *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$, and assume that $(P_x)_{x \in \mathcal{X}}$ is a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. Let λ be the integration of $(P_x)_{x \in \mathcal{X}}$ with respect to μ. For every $\mathbb{E} \otimes \mathbb{K}$–measurable function $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ it holds that*

$$\int f(x,y)\, \mathrm{d}\lambda(x,y) = \iint f(x,y)\, \mathrm{d}P_x(y)\, \mathrm{d}\mu(x)$$

Proof. The inner integral on the right hand side is according to Lemma 1.3.1 $\mathbb{E}$–measurable with values in $[0, \infty]$. Hence both the left hand side and the right hand side are well–defined.

Now assume that f is a simple function on the form (1.2). Then

$$\begin{aligned}\int f \, \mathrm{d}\lambda &= \sum_{k=1}^{n} c_k \lambda(G_k) \\ &= \sum_{k=1}^{n} c_k \int P_x(G_k^x) \, \mathrm{d}\mu(x) \\ &= \sum_{k=1}^{n} c_k \iint 1_{G_k^x}(y) \, \mathrm{d}P_x(y) \, \mathrm{d}\mu(x) \\ &= \sum_{k=1}^{n} c_k \iint 1_{G_k}(x,y) \, \mathrm{d}P_x(y) \, \mathrm{d}\mu(x) \\ &= \iint \sum_{k=1}^{n} c_k 1_{G_k}(x,y) \, \mathrm{d}P_x(y) \, \mathrm{d}\mu(x) \\ &= \iint f(x,y) \, \mathrm{d}P_x(y) \, \mathrm{d}\mu(x)\end{aligned}$$

which shows the result, when f is a simple function.

Now let f be a general function in $\mathcal{M}^+(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$. Then there exists a sequence $(f_n)_{n \in \mathbb{N}}$ of non–negative simple functions with $f_n \uparrow f$. Then from monotone convergence

$$\int f \, d\lambda = \lim_{n \to \infty} \int f_n \, d\lambda = \lim_{n \to \infty} \iint f_n(x,y) \, \mathrm{d}P_x(y) \, \mathrm{d}\mu(x)$$

But monotone convergence also yields

$$\int f_n(x,y) \, \mathrm{d}P_x(y) \uparrow \int f(x,y) \, \mathrm{d}P_x(y)$$

and applying monotone convergence once more then gives

$$\iint f_n(x,y) \, \mathrm{d}P_x(y) \, \mathrm{d}\mu(x) \uparrow \iint f(x,y) \, \mathrm{d}P_x(y) \, \mathrm{d}\mu(x) \,,$$

and this shows the theorem. □

Theorem 1.3.3 (Extended Fubini)**.** *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and assume that $(P_x)_{x \in \mathcal{X}}$ is a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. Let λ be the integration of $(P_x)_{x \in \mathcal{X}}$ with respect to μ. For every $\mathbb{E} \otimes \mathbb{K}$–measurable and λ–integrable function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ it holds that*

$$A_0 = \{x \in \mathcal{X} \, : \, \int |f(x,y)| \, \mathrm{d}P_x(y) < \infty\}$$

is $\mathbb{E}$–measurable with $\mu(A_0) = 1$. Furthermore it is fulfilled that the function

$$x \mapsto g(x) := \begin{cases} \int f(x,y) \, \mathrm{d}P_x(y) & x \in A_0 \\ 0 & x \notin A_0 \end{cases}$$

is $\mathbb{E}$–measurable and μ–integrable, and that

$$\int f(x,y)\,\mathrm{d}\lambda(x,y) = \int_{A_0}\int f(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\,.$$

Note: The extended Tonelli's Theorem can be applied to determine whether f is λ–integrable – that is whether $\int |f|\,d\lambda < \infty$.

Proof. It follows from Lemma 1.3.1 that $A_0 \in \mathbb{E}$. The extended Tonelli's Theorem gives

$$\iint |f(x,y)|\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x) = \int |f|\,d\lambda < \infty\,.$$

Hence the integral $\int |f(x,y)|\,\mathrm{d}P_x(y)$ must be finite for μ almost all $x \in \mathcal{X}$ such that $\mu(A_0) = 1$. For each $x \in A_0$ we have

$$\int f(x,y)\,\mathrm{d}P_x(y) = \int f^+(x,y)\,\mathrm{d}P_x(y) - \int f^-(x,y)\,\mathrm{d}P_x(y)$$

From this we see that the function g defined in the theorem is measurable according to Lemma 1.3.1. Furthermore we obtain from the extended Tonelli that

$$\begin{aligned}\int |g(x)|\mathrm{d}\mu(x) &= \int_{A_0}\left|\int f(x,y)\,\mathrm{d}P_x(y)\right|\,\mathrm{d}\mu(x)\\ &\le \iint 1_{A_0\times\mathcal{Y}}(x,y)|f(x,y)|\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\ &< \infty\,,\end{aligned}$$

showing that g is μ–integrable. Finally, we have from the extended Tonelli that

$$\begin{aligned}&\int_{A_0}\int f(x,y)\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\ &\quad= \int_{A_0}\int f(x,y)^+\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x) - \int_{A_0}\int f(x,y)^-\,\mathrm{d}P_x(y)\,\mathrm{d}\mu(x)\\ &\quad= \int 1_{A_0}(x)f^+(x,y)\,\mathrm{d}\lambda(x,y) - \int 1_{A_0}(x)f^-(x,y)\,\mathrm{d}\lambda(x,y)\\ &\quad= \int 1_{A_0}(x)f(x,y)\,\mathrm{d}\lambda(x,y)\\ &\quad= \int_{A_0\times\mathcal{Y}} f(x,y)\,\mathrm{d}\lambda(x,y)\,.\end{aligned}$$

But from Corollary 1.2.2 we have

$$\lambda(A_0\times\mathcal{Y}) = \mu(A_0)$$

so

$$\int_{A_0 \times \mathcal{Y}} f(x,y)\, \mathrm{d}\lambda(x,y) = \int f(x,y)\, \mathrm{d}\lambda(x,y)$$

□

1.4 Conditional distributions

In an experiment where two random variables X and Y are observed, it is often convenient to consider the probabilistic model in two steps: X is observed first, afterwards Y is observed. Here it is natural to believe that the mechanism that decides the value of Y depends on the drawn value of X. This two–step model can be constructed by considering the simultaneous distribution of X and Y as the integration of the conditional distribution of Y given X with respect to the distribution of X.

Definition 1.4.1. *Let X and Y be random variables defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Let $(P_x)_{x\in\mathcal{X}}$ be a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$. We say that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X, if the simultaneous distribution $(X,Y)(P)$ (a probability measure on $(\mathcal{X}\times\mathcal{Y}, \mathbb{E}\otimes\mathbb{K})$) is the integration of $(P_x)_{x\in\mathcal{X}}$ with respect to $X(P)$. That is, if*

$$P(X \in A, Y \in B) = X(P)(A \times B) = \int_A P_x(B)\, \mathrm{d}X(P)(x)$$

for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$.

Note that we say **the** conditional distribution although according to Theorem 1.2.5 the Markov kernel can be changed on nullsets (with respect to $X(P)$). The strictly correct term would be **a** conditional distribution. When stating the conditional distribution, it is not necessary to give the entire Markov kernel $(P_x)_{x\in\mathcal{X}}$. Since the Markov kernel is integrated with respect to $X(P)$ it will be enough to give $(P_x)_{x\in A_0}$, where $A_0 \in \mathbb{E}$ is a set with $P(X \in A_0) = 1$.

Conversely, a conditional distribution given by $(P_x)_{x\in A_0}$, where $P(X \in A_0) = 1$, can be extended to a "true" Markov kernel $(\tilde{P}_x)_{x\in\mathcal{X}}$ by the definition

$$\tilde{P}_x = \begin{cases} P_x & x \in A_0 \\ P_0 & x \notin A_0 \end{cases}$$

with P_0 is some probability measure on $(\mathcal{Y}, \mathbb{K})$. Note that $x \mapsto \tilde{P}_x(B)$ is measurable, since A_0 is a measurable set.

The interpretation of the conditional distribution of Y given X is that P_x describes the distribution of Y if we know that $X = x$. This interpretation is very useful although it should not be taken too seriously, since it may be difficult to give a strict mathematical description when the event $X = x$ is a nullset. Nevertheless, we will denote P_x **the conditional distribution of** Y **given** $X = x$.

This interpretation leads to the following alternative notation for a Markov kernel $(P_x)_{x\in\mathcal{X}}$ that is a conditional distribution of Y given X:

$$P_Y(B \,|\, X = x) = P_x(B) \qquad \text{for } B \in \mathbb{K}\,.$$

A more relaxed but useful notation will be simply talking about 'the distribution of $Y \mid X = x$ instead' of the longer 'the distribution P_x , when $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X'. We will also from time to time write expressions like $Y \mid X = x \sim \nu$.

Later in this chapter we will show the following very important result:

Theorem 1.4.2. *Let X and Y be random variables defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively, such that $\mathcal{Y}$ is a Borel space. Then there exists a conditional distribution of Y given X.*

This result is particularly important, since (as we shall show) $\mathbb{R}, \mathbb{R}^n$ and $\mathbb{R}^\infty$ are Borel sets.

The proof is not constructive in the sense that it is not in general clear how the Markov kernels should look like. The construction of the Markov kernels is however possible in a number of more concrete situations.

Theorem 1.4.3. *Assume that X and Y are random variables on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ such that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Then X and Y are independent if and only if P_x does not depend on x.*

That the Markov kernel is independent of x means, that it can be chosen such that

$$P_x = P_0$$

for all $x \in \mathcal{X}$. In the case of independence, then $P_x = P_0 = Y(P)$ for all $x \in \mathcal{X}$.

Proof. Suppose that X and Y are independent. Then for $A \in \mathbb{E}$ and $B \in \mathbb{K}$

$$P(X \in A, Y \in B) = X(P)(A) \cdot Y(P)(B) = \int_A Y(P)(B)\, \mathrm{d}X(P)(x)$$

which shows, that the constant Markov kernel $(Y(P))_{x\in\mathcal{X}}$ is the conditional distribution of Y given X.

Conversely, assume that $P_x = P_0$ for all $x \in \mathcal{X}$, where P_0 is some probability measure on $(\mathcal{Y}, \mathbb{K})$. Then for $Y \in \mathbb{K}$ we have

$$P(Y \in B) = P(X \in \mathcal{X}, Y \in B) = \int P_x(B)\, \mathrm{d}X(P)(x) = \int P_0(B)\, \mathrm{d}X(P)(x) = P_0(B)$$

which shows, that $Y(P) = P_0$. Furthermore for $A \in \mathbb{E}$ and $B \in \mathbb{K}$ we obtain

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A P_x(B)\, \mathrm{d}X(P)(x) = \int_A P_0(B)\, \mathrm{d}X(P)(x) \\ &= X(P)(A)P_0(B) = P(X \in A)P(Y \in B) \end{aligned}$$

leading to the conclusion that X and Y are independent. □

Hence independence between two variables X and Y is the same as the conditional distribution of Y given X being constant. If conversely the conditional distribution consists of very different probability measures, then it seems reasonable to believe that there is a strong dependence between X and Y.

In the following theorem it is seen that if X is a discrete random variable, then the conditional distribution is just given by elementary conditional probabilities.

Theorem 1.4.4. *Let X and Y be random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. Assume that $\mathcal{X}$ is finite or countable and that $\mathbb{E}$ is the paving that consists of all subsets of $\mathcal{X}$. Then the conditional distribution of Y given X is determined by*

$$P_x(B) = \frac{P(X = x, Y \in B)}{P(X = x)} \qquad \text{for } B \in \mathbb{K}, \tag{1.3}$$

for all $x \in \mathcal{X}$ with $P(X = x) > 0$.

Note that $P_x(B)$ is simply defined as the conditional distribution of $(Y \in B)$ given the set $(X = x)$:

$$P_x(B) = \frac{P(X = x, Y \in B)}{P(X = x)} = P(Y \in B \,|\, X = x)$$

Proof. Let $A_0 = \{x \in \mathcal{X} \ : \ P(X = x) > 0\}$ and note that $X(P)(A_0) = 1$ such that (1.3) defines a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$ – the measurability is not a problem, since all

functions on $\mathcal{X}$ are $\mathbb{E}$–measurable. For $A \subseteq \mathcal{X}$ and $B \in \mathbb{K}$ we have

$$\begin{aligned}\int_A P_x(B)\,\mathrm{d}X(P)(x) &= \int_{A\cap A_0} P_x(B)\,\mathrm{d}X(P)(x)\\ &= \sum_{x\in A\cap A_0} \frac{P(X=x, Y\in B)}{P(X=x)} P(X=x)\\ &= \sum_{x\in A\cap A_0} P(X=x, Y\in B)\\ &= P(X\in A\cap A_0, Y\in B)\\ &= P(X\in A, Y\in B)\end{aligned}$$

such that $(P_x)_{x\in\mathcal{X}}$ actually is the conditional distribution of Y given X. □

Example 1.4.5. Let X_1 and X_2 be independent random variables that are Poisson distributed with parameters λ_1 and λ_2. Then the distribution of $X = X_1 + X_2$ is a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$. We will find the conditional distribution of X_1 given X by indicating $P_x(\{n\})$ for all $x, n \in \mathbb{N}_0$. This must be sufficient, since all P_x are concentrated on $\mathbb{N}_0$. Using Theorem 1.3 yields for $x \in \mathbb{N}_0$ and $n = 0, 1, \ldots, x$ that

$$\begin{aligned}P_x(\{n\}) &= \frac{P(X_1=n, X_2=x-n)}{P(X=x)}\\ &= \frac{P(X_1=n)P(X_2=x-n)}{P(X=x)}\\ &= \frac{\frac{\lambda_1^n}{n!}e^{-\lambda_1}\frac{\lambda_2^{x-n}}{(x-n)!}e^{-\lambda_2}}{\frac{\lambda^x}{x!}e^{-\lambda}}\\ &= \binom{x}{n}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^n\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{x-n}\end{aligned}$$

Hence the conditional distribution of X_1 given $X = x$ is a binomial distribution with parameters $(x, \frac{\lambda_1}{\lambda_2+\lambda_2})$. ∘

Theorem 1.4.6. *Assume that X and Y are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Assume that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Furthermore let μ and ν be σ–finite measures on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively and assume that $X(P) = f \cdot \mu$. Finally assume that $(P_x)_{x\in\mathcal{X}}$ is a $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Y}, \mathbb{K})$ of the type constructed in Theorem 1.1.2: Assume that $P_x = g_x \cdot \nu$, where the function $(x, y) \mapsto g_x(y)$ is $\mathbb{E}\otimes\mathbb{K}$–measurable.*

Then the simultaneous distribution of X and Y is given by $(X, Y)(P) = h \cdot \mu\otimes\nu$, where

$$h(x, y) = f(x)\, g_x(y) \qquad \text{for all } x \in \mathcal{X},\ y \in \mathcal{Y}$$

Proof. Let $A \in \mathbb{E}$ and $B \in \mathbb{K}$. Then

$$\begin{aligned}(X,Y)(P)(A \times B) &= P(X \in A, Y \in B) \\ &= \int 1_A(x)\, P_x(B)\, \mathrm{d}X(P)(x) \\ &= \int 1_A(x) \Big(\int 1_B(y) g_x(y)\, \mathrm{d}\nu(y) \Big)\, f(x)\, \mathrm{d}\mu(x) \\ &= \int \int 1_{A\times B}(x,y)\, f(x) g_x(y) \mathrm{d}\nu(y) \mathrm{d}\mu(x) \\ &= \int 1_{A\times B}(x,y)\, h(x,y) \mathrm{d}(\mu \otimes \nu)(x,y)\end{aligned}$$

where the last equaility is due to Tonelli. We see that $(X,Y)(P)$ and $h \cdot \mu \otimes \nu$ coincide on all product sets, and thereby they must be equal. □

The theorem states that the simultaneous density is the product of the marginal density and the conditional densities. The next theorem gives the converse result: The densities for the conditional distribution is the fraction between the simultaneous density and the marginal density.

Theorem 1.4.7. *Assume that X and Y are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Furthermore let μ and ν be σ–finite measures on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively and assume that $(X,Y)(P) = h \cdot \mu \otimes \nu$. Then the conditional distribution of Y given X exists. The marginal distribution of X has density with respect to μ given by*

$$f(x) = \int h(x,y)\, \mathrm{d}\nu(y)$$

Let $A_0 = \{x \in \mathcal{X} : 0 < f(x) < \infty\}$. Then $X(P)(A_0) = 1$ and the conditional distribution $(P_x)_{x\in\mathcal{X}}$ of Y given X has density with respect to ν given by

$$g_x(y) = \frac{h(x,y)}{f(x)}$$

for all $x \in A_0$.

Proof. Finding the marginal density for $X(P)$ is a well–known calculation. For $A \in \mathbb{E}$ we

have

$$\begin{aligned} X(P)(A) &= (X,Y)(P)(A\times\mathcal{Y}) \\ &= \int_{A\times\mathcal{Y}} h(x,y)\ \mathrm{d}(\mu\otimes\nu)(x,y) \\ &= \int_A \int h(x,y)\ \mathrm{d}\nu(y)\ \mathrm{d}\mu(x) \\ &= \int_A f(x)\ \mathrm{d}\mu(x) \end{aligned}$$

according to Tonelli. Thus $X(P)$ has density f with respect to μ.

Now define the sets

$$A_1 = \{x\in\mathcal{X} : f(x)=0\} \quad \text{and} \quad A_2 = \{x\in\mathcal{X} : f(x)=\infty\}\,.$$

Since $X(P)(\mathcal{X})=1$ we have

$$1 \geq X(P)(A_2) = \int_{A_2} f(x)\ \mathrm{d}\mu(x) = \infty\cdot\mu(A_2)$$

so $\mu(A_2)=0$. Clearly we have that $X(P)(A_1)=0$, such that $X(P)(A_0)=1$.

From Tonelli we have that $x\mapsto\int h(x,y)\mathrm{d}\nu(y)=f(x)$ is $\mathbb{E}-\mathbb{B}$–measurable. Then also

$$(x,y)\mapsto 1_{A_0}(x)\frac{h(x,y)}{f(x)} = 1_{A_0}(x)g_x(y)$$

is $\mathbb{E}\otimes\mathbb{K}-\mathbb{B}$–measurable, and we have from Theorem 1.1.2 that $(P_x)_{x\in A_0}$ is a Markov kernel, when $P_x = g_x\cdot\mu$. Finally we have for $A\in\mathbb{E}$ and $B\in\mathbb{K}$ that

$$\begin{aligned} \int_A P_x(B)\ \mathrm{d}X(P)(x) &= \int_{A\cap A_0}\left(\int_B g_x(y)\ \mathrm{d}\nu(y)\right) f(x)\ \mathrm{d}\mu(x) \\ &= \int_{A\cap A_0}\left(\int_B \frac{h(x,y)}{f(x)}\ \mathrm{d}\nu(y)\right) f(x)\ \mathrm{d}\mu(x) \\ &= \int_A\left(\int_B h(x,y)\ \mathrm{d}\nu(y)\right)\ \mathrm{d}\mu(x) \\ &= \int_{A\times B} h(x,y)\mathrm{d}(\mu\otimes\nu)(x,y) \\ &= (X,Y)(P)(A\times B) \\ &= P(X\in A, Y\in B)\,, \end{aligned}$$

which shows, that $(P_x)_{x\in A_0}$ is the conditional distribution for Y given X. □

1.5 Existence of conditional distributions

Assume that X and Z are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathbb{R}, \mathbb{B})$ respectively. Assume that $E|Z| < \infty$. Recall that the conditional expectation of Z given X exists and it satisfies 1)–3) in Theorem A.2.6. The fact that $E(Z|X)$ is $\sigma(X)$–measurable gives the existence of a measurable map $\phi : (\mathcal{X}, \mathbb{E}) \to (\mathbb{R}, \mathbb{B})$ such that

$$E(Z|X) = \phi(X) \quad P \text{ almost surely}\,.$$

If Z is an indicator function 1_F for some $F \in \mathbb{F}$, then Z is obviously integrable such that $E(1_F|X)$ is well–defined. We call this a conditional probability (given X) and use the notation

$$P(F|X) = E(1_F|X)$$

Note that $0 \leq P(F|X) \leq 1$ P–a.s., and we can choose a version of $P(F|X)$ such that it has values in $[0,1]$ for all $\omega \in \Omega$. Once again we can find a $\mathbb{E} - \mathbb{B}$–measurable map ϕ_F such that

$$P(F|X) = \phi_F(X)$$

We shall furthermore use the notation

$$P(F|X = x) = \phi_F(x)\,.$$

The conditional probability $P(F|X)$ is obviously determined by the map ϕ_F, so we shall also let this map be called the conditional probability of F given X. Hence ϕ_F with values in $[0,1]$ is a conditional probability of F given X if it is $\mathbb{E} - \mathbb{B}$–measurable and satisfies $\phi_F(X) = P(F|X)$.

Going back to the properties 1)–3) of Theorem A.2.6 that characterises conditional expectations we obtain

Theorem 1.5.1. *A $\mathbb{E}-\mathbb{B}$–measurable map ϕ_F with values in $[0,1]$ is a conditional probability of F given X, if and only if it holds that*

$$\int_A \phi_F(x)\,\mathrm{d}X(P)(x) = P((X \in A) \cap F) \tag{1.4}$$

for all $A \in \mathbb{E}$.

Proof. Assume that ϕ_F is $\mathbb{E} - \mathbb{B}$ measurable with values in $[0,1]$. Then according to the change-of-variable theorem (Theorem A.1.8) it holds

$$\int_{X \in A} \phi_B(X)\,dP = \int_A \phi_B(x)\,\mathrm{d}X(P)(x)$$

Hence equation (1.4) is fulfilled, if and only if

$$\int_{X\in A} \phi_B(X)\, dP = P((X \in A) \cap F) = \int_{(X\in A)} 1_F\, dP$$

is fulfilled for all $A \in \mathbb{E}$. And this is by definition fulfilled, if and only if $\phi_B(X)$ is the conditional expectation of 1_F given X. □

Corollary 1.5.2. *Let X and Y be random variables defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathbb{R}, \mathbb{B})$ respectively. Assume that $(P_x)_{x\in\mathcal{X}})$ is a conditional distribution of Y given X. Then for all $B \in \mathbb{B}$ the function ϕ defined by $\phi(x) = P_x(B)$ is a conditional probability of $(Y \in B)$ given X.*

Proof. Recall that $(P_x)_{x\in\mathcal{X}}$ satisfies

$$\int_A P_x(B)\, \mathrm{d}X(P)(x) = P(X \in A, Y \in B)$$

for all $A \in \mathbb{E}$ and $B \in \mathbb{B}$. Use Theorem 1.5.1. □

The following immediate consequence of properties for conditional expectations will be useful

Lemma 1.5.3. *The following results holds*

(a) Assume that $A \subseteq B$. Then

$$P(A \,|\, X = x) \le P(B \,|\, X = x)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

(b) If $(A_n)_{n\in\mathbb{N}}$ is an increasing sequence of sets, then

$$\lim_{n\to\infty} P(A_n \,|\, X = x) = P\Big(\bigcup_{n=1}^{\infty} A_n |\, X = x\Big)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

(c) If $(A_n)_{n\in\mathbb{N}}$ is a decreasing sequence of sets, then

$$\lim_{n\to\infty} P(A_n \,|\, X = x) = P\Big(\bigcap_{n=1}^{\infty} A_n |\, X = x\Big)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

Proof. (a) We have $1_A \leq 1_B$ so

$$P(A \,|\, X) = E(1_A|\, X) \leq E(1_B|\, X) = P(B \,|\, X) \quad P\text{–a.s.}.$$

If we write $\phi_A(X) = P(A \,|\, X)$ and $\phi_B(X) = P(B \,|\, X)$, then we have

$$\begin{aligned} &= X(P)(\{x \in \mathcal{X} \,:\, P(A \,|\, X = x) \leq P(B \,|\, X = x)\}) \\ &= X(P)(\{x \in \mathcal{X} \,:\, \phi_A(x) \leq \phi_B(x)\}) \\ &= P(\{\omega \in \Omega \,:\, \phi_A(X(\omega)) \leq \phi_B(X(\omega))\}) \\ &= P(P(A \,|\, X) \leq P(B \,|\, X)) = 1 \end{aligned}$$

(b) Let $A_0 = \bigcup_{n=1}^{\infty} A_n$ and note that $1_{A_n} \uparrow 1_{A_0}$. Then it follows from (7) in Theorem A.2.5 that

$$P(A_n|\, X) \to P(A_0|\, X) \qquad P\text{–a.s.}$$

and then the result can be concluded as in (a).

(c) Use that $1_{A_1} - 1_{A_n}$ is increasing and repeat the argument from (b). □

The idea is to use such conditional probabilities of the events $(Y \in A)$ given X to construct the conditional distribution of Y given X. Since we already know that the conditional probabilities $P(Y \in B|X)$ exists, then it seems really simple and obvious just to define $P_x(B) = P(Y \in B \,|\, X = x)$ for all $B \in \mathbb{K}$. The problem is that $P(Y \in B \,|\, X)$ is only determined almost surely and that the nullsets varies with B. We have, however, for B_1 and B_2 disjoint sets that

$$P(Y \in (B_1 \cup B_2) \,|\, X) = P(Y \in B_1 \,|\, X) + P(Y \in B_2 \,|\, X) \qquad P\text{–a.s.}$$

such that also

$$P(Y \in (B_1 \cup B_2) \,|\, X = x) = P(Y \in B_1 \,|\, X = x) + P(Y \in B_2 \,|\, X = x)$$

for $X(P)$ almost all $x \in \mathcal{X}$. This is of course necessary, if P_x should be a probability measure. The problem is, that for other sets $\tilde{B}_1$ and $\tilde{B}_2$, then the nullsets where the above equalities **are not true** may be different. If there are uncountably many sets in $\mathbb{K}$, then it may be impossible to choose P_x, such that e.g. the above equalities are true for all B_1 and B_2.

We are now ready to prove

Theorem 1.5.4. *Let X and Y be random variables defined on the probability space $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathbb{R}, \mathbb{B})$ respectively. Then there exists a conditional distribution of Y given X.*

Proof. We shall exploit the fact that probabilities on $(\mathbb{R}, \mathbb{B})$ are characterised by their distribution function, and that these in turn are completely determined by their values on the rational numbers $\mathbb{Q}$.

Let for $q \in \mathbb{Q}$,

$$P(Y \leq q \,|\, X)$$

be a conditional probability of $(Y \leq q)$ given X. If $q < r \in \mathbb{Q}$ then we have $1_{(Y \leq q)} \leq 1_{(Y \leq r)}$ so

$$P(Y \leq q \,|\, X) = E(1_{(Y \leq q)} | \, X) \leq E(1_{(Y \leq r)} | \, X) = P(Y \leq r \,|\, X) \quad P\text{–a.s.}\,.$$

Define

$$A_{qr} = \{x \in \mathcal{X} \,:\, P(Y \leq q \,|\, X = x) \leq P(Y \leq r \,|\, X = x)\} \in \mathbb{E}\,,$$

then we have $P(A_{qr}) = 1$ by Lemma 1.5.3 (a), such that also $P(A_0) = 1$, where $A_0 = \bigcap_{q<r\in\mathbb{Q}} A_{qr}$. For some fixed $x \in A_0$ we must have that the function

$$q \mapsto P(Y \leq q \,|\, X = x)$$

is increasing on $\mathbb{Q}$. In particular we must have, that for $x \in A_0$ the limits

$$\begin{aligned} L_-(x) &= \lim_{q\to-\infty, q\in\mathbb{Q}} P(Y \leq q \,|\, X = x)\,, \\ L_+(x) &= \lim_{q\to\infty, q\in\mathbb{Q}} P(Y \leq q \,|\, X = x) \end{aligned}$$

exists. Since $(Y \leq n) \uparrow \mathbb{R}$, we must have from Lemma 1.5.3 (b) that $X(P)(G_+) = 1$, where

$$G_+ = \{x \in \mathcal{X} \,:\, \lim_{n\to\infty} P(Y \leq n \,|\, X = x) = 1\} \in \mathbb{E}\,.$$

Hence it must hold that $L_+(x) = 1$ for $X \in A_0 \cap G_+$. It can be seen similarly that $X(P)(G_-) = 1$, where

$$G_- = \{x \in \mathcal{X} \,:\, \lim_{n\to-\infty} P(Y \leq n \,|\, X = x) = 0\} \in \mathbb{E}\,.$$

Altogether we have that $X(P)(M) = 1$, where $M = A_0 \cap G_+ \cap G_-$ and that for $x \in M$ the function $q \mapsto P(Y \leq q \,|\, X = x)$ is increasing on $\mathbb{Q}$ with the limit 1 at $+\infty$ and the limit 0 at $-\infty$.

Hence for each $x \in M$ the function $q \mapsto P(Y \leq q \,|\, X = x)$ looks like a distribution function – and we are looking for distributions indexed by x – but it needs to be defined on $\mathbb{R}$ instead of $\mathbb{Q}$. So define for $x \in M$

$$F(y \,|\, X = x) = \inf\{P(Y \leq q \,|\, X = x) \,:\, q \in \mathbb{Q}, q > y\}$$

and note, that (since $q \mapsto P(Y \leq q \,|\, X = x)$ is increasing) we have

$$F(y \,|\, X = x) = \lim_{n\to\infty} P(Y \leq q_n \,|\, X = x)$$

whenever $q_1 > q_2 > \cdots > y$ is a decreasing sequence with y as the limit. From this we see that $y \mapsto F(y \,|\, X = x)$ satisfies:

- it is increasing:
  Let $y_1 < y_2$ and choose $q_n \downarrow y_1$ and $r_n \downarrow y_2$ rational, such that $q_n < r_n$ for all $n \in \mathbb{N}$.
- it has limit 1 in $+\infty$:
  For $\epsilon > 0$ choose q_0 such that $P(Y \leq q \,|\, X = x) \geq 1 - \epsilon$ for $q \geq q_0$. Realise that $F(y \,|\, X = x) \geq 1 - \epsilon$.
- it has limit 0 in $-\infty$:
  For $\epsilon > 0$ choose q_0 such that $P(Y \leq q \,|\, X = x) \leq \epsilon$ for $q \leq q_0$. Then it is seen that $F(y \,|\, X = x) \leq \epsilon$ for $y \leq q_0 - 1$ by choosing $q_n \downarrow y$ such that $q_n \leq q_0$ for all $n \in \mathbb{N}$.
- it is right continuous:
  Let $y \in \mathbb{R}$ and $\epsilon > 0$. Choose q_0 such that $P(Y \leq q \,|\, X = x) \leq F(y \,|\, X = x) + \epsilon$ for $q \leq y + q_0$. Realise that $F(y' \,|\, X = x) \leq F(y \,|\, X = x) + \epsilon$ for $y' \leq y + q_0/2$ (just choose $q_n \downarrow y'$ with all $q_n \leq y + q_0$).

Hence it is seen that $y \mapsto F(y \,|\, X = x)$ is a distribution function for some distribution P_x on $(\mathbb{R}, \mathbb{B})$ for each $x \in M$. For $x \notin M$ define

$$P_x = P_0$$

with P_0 some probability on $(\mathbb{R}, \mathbb{B})$.

Now our claim is, that the constructed family of measures $(P_x)_{x \in \mathcal{X}}$ is the conditional distribution of Y given X. That P_x is a probability measure for each $x \in \mathcal{X}$ is fulfilled by the construction, so it is only needed to check that for each $B \in \mathbb{B}$

(i') $x \mapsto P_x(B)$ is $\mathbb{E} - \mathbb{B}$–measurable

(ii')
$$\int_A P_x(B) \, \mathrm{d}X(P)(x) = P(X \in A, Y \in B)$$
for all $A \in \mathbb{E}$.

Let

$$\mathbb{H} = \{B \in \mathbb{B} : \text{ (i) and (ii) are fulfilled}\}$$

Then it is shown rather easily that $\mathbb{H}$ is a Dynkin class, and we will have shown that $\mathbb{H} = \mathbb{B}$, if we can show that $\mathbb{D} \subseteq \mathbb{H}$, where $\mathbb{D}$ is the intersection stable generating system for $\mathbb{B}$ given by

$$\mathbb{D} = \{(-\infty, x] \,:\, x \in \mathbb{R}\}$$

So let $x \in \mathbb{R}$ and choose a decreasing rational sequence $q_n \downarrow x$. Then

$$P_x((-\infty, x]) = \begin{cases} F(y \,|\, X = x) & x \in M \\ P_0((-\infty, x]) & x \notin M \end{cases} = \begin{cases} \lim_{n\to\infty} P(Y \leq q_n \,|\, X = x) & x \in M \\ P_0((-\infty, x]) & x \notin M \end{cases} .$$

This must be measurable as a function of x, since $M \in \mathbb{E}$, and each function $x \mapsto P(Y \leq q_n \,|\, X = x)$ is $\mathbb{E} - \mathbb{B}$–measurable. Furthermore we obtain (using $X(P)(M) = 1$)

$$\begin{aligned} \int_A P_x((-\infty, x])\, \mathrm{d}X(P)(x) &= \int_{A\cap M} F(y \,|\, X = x)\, \mathrm{d}X(P)(x) \\ &= \lim_{n\to\infty} \int_{A\cap M} P(Y \leq q_n \,|\, X = x)\, \mathrm{d}X(P)(x) \\ &= \lim_{n\to\infty} P(X \in A \cap M, Y \leq q_n) \\ &= P(X \in A \cap M, Y \leq x) \\ &= P(X \in A, Y \leq x) \end{aligned}$$

In the third equality we have used Theorem 1.5.1, since the maps $x \mapsto P(Y \leq q_n \,|\, X = x)$ are conditional probabilities of $(Y \leq q_n)$ given X. Hence (i') and (ii') are show for $A = (-\infty, x]$. □

The result can be generalised to other random variables than the $\mathbb{R}$–valued

Definition 1.5.5. *A measurable space $\mathcal{Y}, \mathbb{K})$ is a Borel space, if there exist $B_0 \in \mathbb{B}$ and a bijective, bi–measurable map $\varphi : (\mathcal{Y}, \mathbb{K}) \to (B_0, B_0 \cap \mathbb{B})$.*

It is in particular required that both φ and φ^{-1} are measurable. Note that we consider B_0 as a subspace of $(\mathbb{R}, \mathbb{B})$ equipped with the σ–algebra $B_0 \cap \mathbb{B} = \{B_0 \cap B \,:\, B \in \mathbb{B}\}$. The bi–measurability then amounts to

$$\begin{aligned} \varphi^{-1}(B) \in \mathbb{K} \text{ for } B \in B_0 \cap \mathbb{B} \\ \varphi(A) \in B_0 \cap \mathbb{B} \text{ for } A \in \mathbb{K} \end{aligned}$$

Theorem 1.5.6. *If Y is a random variable taking values in a Borel space $(\mathcal{Y}, \mathbb{K})$, then there exists a conditional distribution of Y given X.*

Proof. Consider $Z = \varphi(Y)$ (φ is of course given as in Definition 1.5.5), so according to Theorem 1.5.4 there exists a conditional distribution $(\tilde{P}_x)_{x\in\mathcal{X}}$ of Z given X.

For all $x \in \mathcal{X}$ the probability measure $\tilde{P}_x$ is concentrated on B_0 and can therefore be viewed as a probability measure on $(B_0, B_0 \cap \mathbb{B})$. The probability measure obtained from this by the transformation φ^{-1} is a probability measure P_x on $(\mathcal{Y}, \mathbb{K})$. It is easily seen that $(P_x)_{x\in\mathcal{X}}$ is a conditional distribution of Y given X. □

Theorem 1.5.6 is useful because of the following fact:

Theorem 1.5.7. *For all $n \in \mathbb{N}$, $(\mathbb{R}^n, \mathbb{B}^n)$ is a Borel space. Furthermore, $(\mathbb{R}^\infty, \mathbb{B}^\infty)$ is a Borel space.*

Sketch of proof. We shall almost show that $([0,1]^2, [0,1]^2 \cap \mathbb{B}^2)$ is a Borel space. Consider the binary expansion of an arbitrary $x \in [0,1]$,

$$x = \sum_{n=1}^{\infty} x_n \frac{1}{2^n}, \qquad x < 1$$
$$x = 1 \cdot 2^0 + \sum_{n=1}^{\infty} 0 \cdot \frac{1}{2^n}, \qquad x = 1\,,$$

where all $x_n \in \{0,1\}$ and the expansion is uniquely determined by the requirement that the sequence (x_n) must contain infinitely many 0's.

Now define a map $\phi : [0,1] \to [0,1]^2$ by $\phi(x) = (y_1, y_2)$, where

$$y_1 = \sum_{n=1}^{\infty} x_{2n-1} \frac{1}{2^n}, \quad y_2 = \sum_{n=1}^{\infty} x_{2n} \frac{1}{2^n}$$

ϕ is surjective, but not injective (for instance, $\phi(1) = \phi(0.101010101\cdots)$), but by removing countably many points from $[0,1]$ we may turn ϕ into a bi–measurable bijection onto $([0,1]^2, [0,1]^2 \cap \mathbb{B}^2)$, the inverse of which, φ, satisfies the requirements given in Definition 1.5.5.

A more refined application of the same idea can be used to show that $([0,1]^\infty, [0,1]^\infty \cap \mathbb{B}^\infty)$ is a Borel space: Consider $\phi(x) = (y_1, y_2, \ldots)$, where y_n has the binary expansion given by the n'th row of

$$\begin{array}{lcccccc}
y_1: & x_0 & x_1 & x_3 & x_6 & \cdots \\
y_1: & x_2 & x_4 & x_7 & x_{11} & \cdots \\
y_1: & x_5 & x_8 & x_{12} & x_{17} & \cdots \\
y_1: & x_9 & x_{13} & x_{18} & x_{24} & \cdots \\
 & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

□

1.6 Exercises

Exercise 1.1. Assume that X_1 and X_2 are independent random variables that are both binomially distributed with parameters (n, p). Define the random variable $X = X_1 + X_2$. Find the conditional distribution of X_1 given X (Hint). ∘

Exercise 1.2. Let X and Y be random variables defined on $(\Omega, \mathbb{F}, P)$. Assume that

- X has the binomial distribution with parameters (n, p_1)
- The conditional distribution of Y given $X = x$ is binomial with parameters (n, p_2)

Find the marginal distribution of Y and try to give an intuitive explanation of the result. ∘

Exercise 1.3.

(1) Assume that X and Y are two random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively, and let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of Y given X. Let $f : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be a $\mathbb{E} \otimes \mathbb{K}$–measurable function. Show that

$$E[f(X, Y)] = \iint f(x, y)\, \mathrm{d}P_x(y)\, \mathrm{d}X(P)(x)$$

Hint: Use the extended Tonelli to calculate an integral with respect to $(X, Y)(P)$ as a double integral.

(2) Assume that X is uniformly distributed on $(0, 1)$. Assume that the conditional distribution $(P_x)_{x\in(0,1)}$ of Y given X fulfils that P_x is the exponential distribution with mean value x. Find EY (Hint).

∘

Exercise 1.4. Let $\mathcal{Y}$ be a finite or countable set, and let $\mathbb{K}$ consist of all subsets of $\mathcal{Y}$. Assume that Y is a random variable defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{Y}, \mathbb{K})$. Let $p(y)$

denote the probability function for Y. Let $\mathcal{X}$ be another finite or countable set, and assume that $t : \mathcal{Y} \to \mathcal{X}$ is some map. Define $X = t(Y)$.

(1) Show that X has probability function

$$r(x) = \sum_{y \in t^{-1}(x)} p(y)$$

(2) Show that $(P_x)_{x \in \mathcal{X}}$ is the conditional distribution of Y given X, where each P_x has probability function

$$q_x(y) = \frac{p(y) 1_{\{x\}}(t(y))}{r(x)}$$

∘

Exercise 1.5. Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $\{0, 1\}$. Assume that

$$P(Y_1 = 0) = 1 - p\,, \qquad P(Y_1 = 1) = p$$

for some $0 < p < 1$. Define $t : \{0, 1\}^n \to \{0, 1, \ldots, n\}$ by

$$t(y_1, \ldots, y_n) = y_1 + \cdots + y_n$$

Define $X = t(Y_1, \ldots, Y_n)$.

(1) Realise that X has the binomial distribution with parameters (n, p) and argue that $P(X = x) > 0$ for all $x = 0, 1, \ldots, n$.

(2) Show that $(P_x)_{x=0,\ldots,n}$ is the conditional distribution of $Y = (Y_1, \ldots, Y_n)$ given X, where P_x is the uniform distribution on $\{(y_1, \ldots, y_n) \in \mathbb{R}^n : y_1 + \cdots + y_n = x\}$.

∘

Exercise 1.6. Let X and Y be random variables defined on $(\Omega, \mathbb{F}, P)$. Assume that

- X has the binomial distribution with parameters (n, p_1).
- The conditional distribution of Y given $X = x$ is binomial with parameters (x, p_2).

Find the marginal distribution of Y and try to give an intuitive explanation of the result. ∘

Exercise 1.7. Let X and Y be random variables with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively, such that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Assume that μ and ν are σ–finite measures on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. Assume furthermore that $X(P)$ has density f with respect to μ, and that for each $x \in \mathcal{X}$ the probability P_x has density g_x with respect to ν, such that $(x, y) \mapsto g_x(y)$ is $\mathbb{E} \otimes \mathbb{K} - \mathbb{B}$–measurable.

(1) Show that
$$\ell(y) = \int g_x(y) f(x)\, \mu(dx)$$
is the density for the marginal distribution of Y with respect to ν.

(2) Show that $Y(P)(B_0) = 1$, where $B_0 = \{y \in \mathcal{Y} : 0 < \ell(y) < \infty\}$.

(3) Show that the conditional distribution of X given Y exists and is given by $(Q_y)_{y\in\mathcal{Y}}$, where Q_y has density with respect to ν given by
$$k_y(x) = \frac{g_x(y) f(x)}{\ell(y)}$$
for $y \in B_0$.

∘

Exercise 1.8. Assume that X is Gamma–distributed with parameters (λ, β) and that the conditional distribution of Y given X is given by $(P_x)_{x\in\mathbb{R}}$, where P_x is the Poisson distribution with parameter x.

(1) Show that the marginal distribution of Y is a negative binomial distribution and find the parameters.

(2) Show that the conditional distributions of X given Y are Γ–distributions.

∘

Exercise 1.9. Let X and Y be real valued random variables defined on $(\Omega, \mathbb{F}, P)$. Let $C \in \mathbb{B}$ be a fixed subset of $\mathbb{R}$. Consider the following game: We are told the value of X, and are based on this information supposed to guess whether $Y \in C$ or not.

It seems natural to expect that we in two different games, where the same value of X is observed, give the same guess of whether $Y \in C$ or not – we know the same in the two situations. Hence giving a rule for guessing must be the same as indicating a set A: If we observe $X \in A$ then we guess that $Y \in C$, and if we observe $X \notin A$, then we guess that $Y \notin C$.

Obviously, different choices of A may lead to more or less successful guessing rules (we define a guessing rule to be successful, if it often leads to the right guess...). Let $(P_x)_{x\in\mathbb{R}}$ be the conditional distribution of Y given X.

(1) Show that for a given guessing rule, then

$$P(\text{right guess}) = \int_A P_x(C)\,\mathrm{d}X(P)(x) + \int_{A^c} P_x(C^c)\,\mathrm{d}X(P)(x)$$

(2) Show that the optimal guessing rule corresponds to the set

$$A_0 = \{x \in \mathbb{R} : P_x(C) \geq \frac{1}{2}\}\,.$$

(3) How is the optimal guessing rule, if X and Y are independent?

(4) How is the optimal guessing rule, if $X = Y$?

∘

Exercise 1.10. Let X be a random variable with values in $(\mathcal{X}, \mathbb{E})$ that is defined on a probability space $(\Omega, \mathbb{F}, P)$. Let furthermore $F \in \mathbb{F}$ and consider the random variable 1_F.

(1) Find the Markov kernel $(P_z)_{z\in\{0,1\}}$ that is the conditional distribution of X given 1_F.

(2) Find the Markov kernel $(Q_x)_{x\in\mathcal{X}}$ that is the conditional distribution of 1_F given X (Hint).

∘

Chapter 2

Conditional distributions: Transformations and moments

2.1 Transformations of conditional distributions

In this section we shall present a series of transformation results for conditional distributions. They have a somewhat similar content: In a framework with three or more random variables, where we know some of the conditional distributions, then some other conditional distributions can be expressed.

It is complicated to understand how conditional distributions are specified in situations with three ore more random variables. The reader is encouraged to spend much time understanding the content of the results, rather than the proofs. The stated results are not very surprising if the content is understood. And the proofs are rather mechanical: Firstly, it is argued that some expression is a Markov kernel, and then it is shown that this Markov kernel is the right conditional distribution.

Assume in this section, that X, Y, X_1, X_2, Y_1 and Y_2 are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$, $(\mathcal{Y}, \mathbb{K})$, $(\mathcal{X}_1, \mathbb{E}_1)$, $(\mathcal{X}_2, \mathbb{E}_2)$, $(\mathcal{Y}_1, \mathbb{K}_1)$ and $(\mathcal{Y}_2, \mathbb{K}_2)$ respectively.

Theorem 2.1.1 (Substitution Theorem)**.** *Assume that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Let $(\mathcal{Z}, \mathbb{H})$ be a measurable space, and let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a measurable map. Define $Z = \phi(X, Y)$. Then the conditional distribution of Z given X exists and is*

determined by $(\tilde{P}_x)_{x\in\mathcal{X}}$*, where*

$$\tilde{P}_x = (\phi \circ i_x)(P_x)$$

Note that this is not at all surprising: If we know that $X = x$, then we have $Z = \phi(x, Y) = (\phi \circ i_x)(Y)$, and apparently we are allowed to plug the conditional distribution into this formula.

Proof. For a fixed $C \in \mathbb{H}$ we have

$$\tilde{P}_x(C) = P_x((\phi \circ i_x)^{-1}(C)) = P_x((\phi^{-1}(C))^x)\,,$$

which is a measurable function of x, since $(P_x)_{x\in\mathcal{X}}$ is a Markov kernel. Hence $(\tilde{P}_x)_{x\in\mathcal{X}}$ is a Markov kernel (each $\tilde{P}_x$ is a probability measure since it is the image measure by the function $\phi \circ i_x$).

Now let $A \in \mathbb{E}$ and $C \in \mathbb{H}$. Then

$$P(X \in A, Z \in C) = (X, Y)(P)((A \times \mathcal{Y}) \cap \phi^{-1}(C))\,.$$

It is seen that if $x \notin A$ then

$$((A \times \mathcal{Y}) \cap \phi^{-1}(C))^x = \emptyset$$

and if $x \in A$ we have

$$((A \times \mathcal{Y}) \cap \phi^{-1}(C))^x = (\phi^{-1}(C))^x = (\phi \circ i_x)^{-1}(C)\,.$$

Hence

$$P(X \in A, Z \in C) = \int_A P_x((\phi \circ i_x)^{-1}(C))\,\mathrm{d}X(P)(x) = \int_A \tilde{P}_x(C)\,\mathrm{d}X(P)(x)\,,$$

which is what we wanted to prove. □

Example 2.1.2. In this example we consider the p–dimensional normal distribution. Recall that this distribution is uniquely determined by the mean vector $\xi \in \mathbb{R}^p$ and the covariance matrix Σ which is a positive semidefinite $p \times p$ matrix. If X is a random vector in $\mathbb{R}^p$ with this distribution, we write $X \sim \mathcal{N}_p(\xi, \Sigma)$. If A is a $p \times p$ matrix and $b \in \mathbb{R}^p$ we have

$$AX + b \sim \mathcal{N}_p(b + A\xi, A\Sigma A^T) \tag{2.1}$$

where A^T denotes the transposition of A. Let $p = r + s$ with $1 \le r$, $1 \le s$, and let X_1 be the the first r coordinates of X. In the following we write

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{12}^T = \Sigma_{21}$, since Σ is positive semidefinite and thereby symmetric. It is furthermore a well–known property of the normal distribution that X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$. In the following we shall assume that Σ_{22} is positive define such that Σ_{22}^{-1} exists.

The aim will be to find the conditional distribution of X_1 given X_2. For this define $Z = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$. Then we have (with e.g. I_r the r–dimensional identity matrix)

$$\begin{pmatrix} Z \\ X_2 \end{pmatrix} = \begin{pmatrix} I_r & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_s \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Since

$$\begin{aligned} &\begin{pmatrix} I_r & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_s \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_r & 0 \\ -\Sigma_{12}\Sigma_{22}^{-1} & I_s \end{pmatrix} \\ =&\begin{pmatrix} I_r & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_s \end{pmatrix} \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{12} \\ 0 & \Sigma_{22} \end{pmatrix} \\ =&\begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}, \end{aligned}$$

it follows from (2.1) that

$$\begin{pmatrix} Z \\ X_2 \end{pmatrix} \sim \mathcal{N}_{r+s}\left(\begin{pmatrix} \xi_1 - \Sigma_{12}\Sigma_{22}^{-1}\xi_2 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right).$$

From this we see that Z and X_2 are independent and that

$$Z \sim \mathcal{N}_r(\xi_1 - \Sigma_{12}\Sigma_{22}^{-1}\xi_2\,,\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\,.$$

Hence this normal distribution is also the conditional distribution of Z given X_2 (Theorem 3). Then using the substitution $X_1 = Z + \Sigma_{12}\Sigma_{22}^{-1}X_2$ gives according to Theorem 2.1.1 and (2.1) that

$$X_1 \,|\, X_2 = x \sim \mathcal{N}_r(\xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x - \xi_2)\,,\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

∘

Example 2.1.3. Assume that X and Y a real valued variables such that the simultaneous distribution of (X, Y) is a Dirichlet distribution with parameters $(\lambda_1, \lambda_2, \lambda)$. Then the distribution of (X, Y) has density

$$f(x, y) = \frac{\Gamma(\lambda + \lambda_1 + \lambda_2)}{\Gamma(\lambda)\Gamma(\lambda_1)\Gamma(\lambda_2)}\, x^{\lambda_1 - 1} y^{\lambda_2 - 1}(1 - x - y)^{\lambda - 1}$$

on the set $\{x, y) \in \mathbb{R}^2 \,:\, 0 < x, 0 < y, x + y < 1\}$. It can be shown that the marginal distribution of X is a B–distribution with parameters $(\lambda_1, \lambda_2 + \lambda)$. Hence it has density

$$g(x) = \frac{\Gamma(\lambda + \lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2 + \lambda_2)}\, x^{\lambda_1 - 1}(1 - x)^{\lambda_2 + \lambda - 1}$$

for $x \in (0,1)$. The conditional distribution P_x of Y given $X = x$ for $x \in (0,1)$ must be concentrated on the interval $(0, 1-x)$ and have density

$$f_x(y) = \frac{(f(x,y)}{g(x)} = \frac{\Gamma(\lambda_2 + \lambda)}{\Gamma(\lambda)\Gamma(\lambda_2)} \left(\frac{y}{1-x}\right)^{\lambda_2 - 1} \left(1 - \frac{y}{1-x}\right)^{\lambda - 1} \frac{1}{1-x}\,.$$

If P_x is transformed by the map $y \to \frac{y}{1-x}$ then a B–distribution with parameters (λ_2, λ) is obtained. According to Theorem 2.1.1 the constant family consisting of B–distributions with parameters (λ_2, λ) indexed by $x \in (0,1)$ must be the conditional distribution of $\frac{Y}{1-X}$ given X. It follows from Theorem 1.4.3 that $\frac{Y}{1-X}$ and X are independent and that $\frac{Y}{1-X}$ is B–distributed with parameters (λ_2, λ). ∘

Theorem 2.1.4. *Assume that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Let $\mathcal{Z}, \mathbb{H})$ be a measurable space and let $t : \mathcal{X} \to \mathcal{Z}$ be a $\mathbb{E}-\mathbb{H}$–measurable map. Define $Z = t(X)$. Then the conditional distribution $(Q_{x,z})_{(x,z)\in\mathcal{X}\times\mathcal{Z}}$ of Y given (X, Z) is given by*

$$Q_{x,z} = P_x \qquad \text{for all } x \in \mathcal{X}, z \in \mathcal{Z} \tag{2.2}$$

Note: This is a situation where it is quite clear that conditional distributions are not uniquely determined. The variable (X, Z) has not values in the entire $\mathcal{X} \times \mathcal{Z}$ but only on the graph of t, meaning the set of points

$$\{(x, y) \in \mathcal{X} \times \mathcal{Z} \,:\, z = t(x)\}$$

Then $Q_{x,z}$ could be defined as any probability measure outside the graph, if only some measurability conditions are fulfilled. Hence the the Markov kernel defined in (2.2) is not the only possible conditional distribution of Y given (X, Z) – it is simply a convenient choice.

Proof. It is easily argued that a $(\mathcal{X} \times \mathcal{Z}, \mathbb{E} \otimes \mathbb{H})$–Markov kernel $(Q_{x,z})_{(x,z)\in\mathcal{X}\times\mathcal{Z}}$ on $(\mathcal{Y}, \mathbb{K})$ is defined by (2.2). For $A \in \mathbb{E}$, $B \in \mathbb{K}$ and $C \in \mathbb{H}$ we have

$$\begin{aligned}
\int_{A\times C} Q_{x,z}(B)\, \mathrm{d}(X,Z)(P)(x,z) &= \int 1_{A\times C}(x,z) Q_{x,z}(B)\, \mathrm{d}((\mathrm{id}, t) \circ X)(P)(x,z) \\
&= \int 1_{A\times C} \circ (\mathrm{id}, t)(x)\, Q_{(\mathrm{id},t)(x)}(B)\, \mathrm{d}X(P)(x) \\
&= \int 1_{A\cap t^{-1}(C)}(x) P_x(B)\, \mathrm{d}X(P)(x)\,.
\end{aligned}$$

Since $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X, the last integral can be identified as

$$\begin{aligned}
P(X \in A \cap t^{-1}(C), Y \in B) &= P(X \in A, Z \in C, Y \in B) \\
&= P((X, Z) \in A \times C, Y \in B)\,.
\end{aligned}$$

By fixing B and letting $A \times C$ vary it is obtained (by uniqueness of measures) that

$$\int_G Q_{x,z}(B)\,\mathrm{d}(X,Z)(P)(x,z) = P((X,Z) \in G, Y \in B)$$

for all $G \in \mathbb{E}\otimes\mathbb{H}$ and all $B \in \mathbb{K}$. Hence it is concluded that $(Q_{x,z})_{(x,z)\in\mathcal{X}\times\mathcal{Z}}$ is the conditional distribution of Y given (X,Z). □

Theorem 2.1.5. *Let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of Y given X. Let $(\mathcal{Z},\mathbb{H})$ be a measurable space and let $t : \mathcal{X} \to \mathcal{Z}$ be an $\mathbb{E}-\mathbb{H}$–measurable map. Define $Z = t(X)$.*
If an $(\mathcal{Z},\mathbb{H})$–Markov kernel $(Q_z)_{z\in\mathcal{Z}}$ on $(\mathcal{Y}\,\mathbb{K})$ exists such that

$$P_x = Q_{t(x)} \qquad \text{for all } x \in \mathcal{X}\,,$$

then $(Q_z)_{z\in\mathcal{Z}}$ is the conditional distribution of Y given Z

A more relaxed formulation of this is, that if the conditional distribution of Y given X only depends on X through $t(X)$, then this is also the conditional distribution of Y given $t(X)$.

Proof. Let $C \in \mathbb{H}$ and $B \in \mathbb{K}$. According to the change–variable–theorem we have

$$\begin{aligned} P(Z \in C, Y \in B) &= P(X \in t^{-1}(C), Y \in B) \\ &= \int 1_{t^{-1}(C)}(x) P_x(B)\,\mathrm{d}X(P)(x) \\ &= \int 1_C \circ t(x) Q_{t(x)}(B)\,\mathrm{d}X(P)(x) \\ &= \int 1_C(z) Q_z(B)\,\mathrm{d}(t\circ X)(P)(z) \\ &= \int_C Q_z(B)\,\mathrm{d}Z(P)(z)\,. \end{aligned}$$

Hence $(Q_z)_{z\in\mathcal{Z}}$ is the conditional distribution of Y given Z. □

Theorem 2.1.6. *Let Z be a random variable with values in $(\mathcal{Z},\mathbb{H})$. Assume that $(Q_{x,y})_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$ is the conditional distribution of Z given (X,Y), and assume that $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. Then $(R_x)_{x\in\mathcal{X}}$ is the conditional distribution of Z given X, where*

$$R_x(C) = \int Q_{x,y}(C)\,\mathrm{d}P_x(y)$$

for $C \in \mathbb{H}$.

Proof. For fixed $x \in \mathcal{X}$ we note that the reduced family $(Q_{x,y})_{y\in\mathcal{Y}}$ is a $(\mathcal{Y}, \mathbb{K})$–Markov kernel on $(\mathcal{Z}, \mathbb{H})$. And R_x is the mixture (the first coordinate of the integrated measure) of this Markov kernel with respect to P_x. In particular we see that R_x is a probability measure on $(\mathcal{Z}, \mathbb{H})$.

Choose $C \in \mathbb{H}$. Since $(Q_{x,y})_{x,y)\in\mathcal{X}\times\mathcal{Y}}$ is a Markov kernel on $(\mathcal{Z}, \mathbb{H})$ we have that

$$(x, y) \mapsto Q_{x,y}(C)$$

is an $\mathbb{E} \otimes \mathbb{K}$–measurable, non–negative map. Hence

$$x \mapsto R_x(C) = \int Q_{x,y}(C)\, \mathrm{d}P_x(y)$$

is $\mathbb{E}$–measurable, which means that $(R_x)_{x\in\mathcal{X}}$ is an $(\mathcal{X}, \mathbb{E})$–Markov kernel on $(\mathcal{Z}, \mathbb{H})$.

Finally let $A \in \mathbb{E}$ and $C \in \mathbb{H}$. Then the extended Tonelli'sTheorem yields that

$$\begin{aligned} P(X \in A, Z \in C) &= P(X \in A, Y \in \mathcal{Y}, Z \in C) \\ &= \int_{A\times\mathcal{Y}} Q_{x,y}(C)\, \mathrm{d}(X, Y)(P)(x, y) \\ &= \iint 1_{A\times\mathcal{Y}}(x, y) Q_{x,y}(C)\, \mathrm{d}P_x(y)\, \mathrm{d}X(P)(x) \\ &= \int 1_A(x) R_x(C)\, \mathrm{d}X(P)(x) \end{aligned}$$

Thereby it follows that $(R_x)_{x\in\mathcal{X}}$ is the conditional distribution of Z given X. □

The two next Theorems can be considered as one bi–implication saying that if X_1 and X_2 are independent, then independence between (X_1, Y_1) and (X_2, Y_2) can be expressed as a property of the conditional distribution of (Y_1, Y_2) given (X_1, X_2).

Theorem 2.1.7. *Assume that the variables (X_1, Y_1) and (X_2, Y_2) are independent, and let $(P^i_x)_{x\in\mathcal{X}_i}$ be the conditional distribution of Y_i given X_i, $i = 1, 2$. Then $(Q_{x_1,x_2})_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2}$ is the conditional distribution of (Y_1, Y_2) given (X_1, X_2), where*

$$Q_{x_1,x_2} = P^1_{x_1} \otimes P^2_{x_2}\,.$$

Proof. For each $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$ we obviously have that Q_{x_1,x_2} is a probability measure on $(\mathcal{Y}_1 \times \mathcal{Y}_2, \mathbb{K}_1 \otimes \mathbb{K}_2)$. For a measurable product set $B_1 \times B_2$, where $B_1 \in \mathbb{K}_1$ and $B_2 \in \mathbb{K}_2$ we have

$$Q_{x_1,x_2}(B_1 \times B_2) = P^1_{x_1}(B_1) P^2_{x_2}(B_2)\,.$$

And since both $(x_1, x_2) \mapsto P^1_{x_1}(B_1)$ and $(x_1, x_2) \mapsto P^2_{x_2}(B_2)$ are $\mathbb{E}_1 \otimes \mathbb{E}_2$–measurable, we must have that $Q_{x_1,x_2}(B_1 \times B_2)$ is measurable as well. According to Lemma 1.1.3 we must have, that $Q_{x_1,x_2}(G)$ is $\mathbb{E}_1 \otimes \mathbb{E}_2$–measurable for all $G \in \mathbb{K}_1 \otimes \mathbb{K}_2$. Hence $(Q_{x_1,x_2})_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2}$ is a Markov kernel.

Now let $A_i \in \mathbb{E}_i$ and $B_i \in \mathbb{K}_i$ for $i = 1, 2$. The assumption that (X_1, Y_1) and (X_2, Y_2) are independent will in particular make X_1 and X_2 independent, such that

$$\begin{aligned}\int_{A_1\times A_2} & Q_{x_1,x_2}(B_1 \times B_2)\,\mathrm{d}(X_1, X_2)(P)(x_1, x_2) \\ &= \int 1_{A_1}(x_1)1_{A_2}(x_2)P^1_{x_1}(B_1)P^2_{x_2}(B_2)\,\mathrm{d}(X_1(P) \otimes X_2(P))(x_1, x_2)\end{aligned}$$

Using Tonelli gives

$$\begin{aligned}\int_{A_1\times A_2} & Q_{x_1,x_2}(B_1 \times B_2)\,\mathrm{d}(X_1, X_2)(P)(x_1, x_2) \\ &= \left(\int 1_{A_1}(x_1)P^1_{x_1}(B_1)\,\mathrm{d}X_1(P)(x_1)\right)\left(\int 1_{A_2}(x_2)P^2_{x_2}(B_2)\,\mathrm{d}X_2(P)(x_2)\right) \\ &= P(X_1 \in A_1, Y_1 \in B_1)\,P(X_2 \in A_2, Y_2 \in B_2) \\ &= P((X_1, Y_1) \in A_1 \times B_1, (X_2, Y_2) \in A_2 \times B_2) \\ &= P((X_1, X_2) \in A_1 \times A_2, (Y_1, Y_2) \in B_1 \times B_2)\end{aligned}$$

From having A_1 and A_2 fixed while varying B_1 and B_2 it is seen (from uniqueness of measures) that

$$\begin{aligned}\int_{A_1\times A_2} & Q_{x_1,x_2}(G_2)\,\mathrm{d}(X_1, X_2)(P)(x_1, x_2) \\ &= P((X_1, X_2) \in A_1 \times A_2, (Y_1, Y_2) \in G_2)\end{aligned}$$

If we conversely fix G_2 in this expression and let A_1 and A_2 vary, then we obtain

$$\int_{G_1} Q_{x_1,x_2}(G_2)\,\mathrm{d}(X_1, X_2)(P)(x_1, x_2) = P((X_1, X_2) \in G_1, (Y_1, Y_2) \in G_2)$$

for all $G_1 \in \mathbb{E}_1 \otimes \mathbb{E}_2$ and $G_2 \in \mathbb{K}_1 \otimes \mathbb{K}_2$. Hence $(Q_{x_1,x_2})_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2}$ is the conditional distribution of (Y_1, Y_2) given (X_1, X_2). □

Theorem 2.1.8. *Let the variables X_1 and X_2 be independent and let $(Q_{x_1,x_2})_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2}$ be the conditional distribution of (Y_1, Y_2) given (X_1, X_2). Assume that each Q_{x_1,x_2} factorises on the form*

$$Q_{x_1,x_2} = P^1_{x_1} \otimes P^2_{x_2}$$

for two families $(P^1_{x_1})_{x_1\in\mathcal{X}_1}$ and $(P^2_{x_2})_{x_2\in\mathcal{X}_2}$ of probability measures on $(\mathcal{Y}_1, \mathbb{K}_1)$ and $(\mathcal{Y}_2, \mathbb{K}_2)$ respectively.

Then $(P^1_{x_1})_{x_1\in\mathcal{X}_1}$ is the conditional distribution of Y_1 given X_1 and $(P^2_{x_2})_{x_2\in\mathcal{X}_2}$ is the conditional distribution of Y_2 given X_2. Furthermore (X_1, Y_1) and (X_2, Y_2) are independent.

Proof. Firstly, we will argue that $(P^1_{x_1})_{x_1\in\mathcal{X}_1}$ is an $(\mathcal{X}_1, \mathbb{E}_1)$–Markov kernel on $(\mathcal{Y}_1, \mathbb{K}_1)$. Let $B_1 \in \mathbb{K}_1$ and choose a fixed $x_2 \in \mathcal{X}_2$. Then

$$P^1_{x_1}(B_1) = P^1_{x_1}(B_1)P^2_{x_2}(\mathcal{Y}_2) = Q_{x_1,x_2}(B_1 \times \mathcal{Y}_2)\,.$$

Since $(x_1, x_2) \mapsto Q_{x_1,x_2}(B_1 \times \mathcal{Y}_2)$ is $\mathbb{E}_1 \otimes \mathbb{E}_2$–measurable and the inclusion map $x_1 \mapsto (x_1, x_2)$ is $\mathbb{E}_1 - \mathbb{E}_1 \otimes \mathbb{E}_2$ –measurable, we can conclude that $x_1 \mapsto P^1_{x_1}(B_1)$ is $\mathbb{E}_1$–measurable. Similarly $(P^2_{x_2})_{x_2\in\mathcal{X}_2}$ is an $(\mathcal{X}_2, \mathbb{E}_2)$-Markov kernel on $(\mathcal{Y}_2, \mathbb{K}_2)$.

For $A_1 \in \mathbb{E}_1$ and $B_1 \in \mathbb{K}_1$ we have

$$\begin{aligned}
P(X_1 \in A_1, Y_1 \in B_1) &= P((X_1, X_2) \in A_1 \times \mathcal{X}_2, (Y_1, Y_2) \in B_1 \times \mathcal{Y}_2) \\
&= \int_{A_1\times\mathcal{X}_2} Q_{x_1,x_2}(B_1 \times \mathcal{Y}_2)\,\mathrm{d}(X_1, X_2)(P)(x_1, x_2) \\
&= \int 1_{A_1}(x_1)P^1_{x_1}(B_1)\,\mathrm{d}(X_1(P) \otimes X_2(P))(x_1, x_2) \\
&= \int_{A_1} P^1_{x_1}(B_1)\,\mathrm{d}X_1(P)(x_1)\,,
\end{aligned}$$

where we have used Tonelli's Theorem. Hence $(P^1_{x_1})_{x_1\in\mathcal{X}_1}$ is the conditional distribution of Y_1 given X_1. Similarly, it is seen that $(P^2_{x_2})_{x_2\in\mathcal{X}_2}$ is the conditional distribution of Y_2 given X_2.

For $A_i \in \mathbb{E}_i$ and $B_i \in \mathbb{K}_i$, $i = 1, 2$, we have

$$\begin{aligned}
&P((X_1, Y_1) \in A_1 \times B_1, (X_2, Y_2) \in A_2 \times B_2) \\
&\quad= P((X_1, X_2) \in A_1 \times A_2, (Y_1, Y_2) \in B_1 \times B_2) \\
&\quad= \int_{A_1\times A_2} Q_{x_1,x_2}(B_1 \times B_2)\,\mathrm{d}(X_1, X_2)(P)(x_1, x_2) \\
&\quad= \int 1_{A_1}(x_1)1_{A_2}(x_2)P^1_{x_1}(B_1)P^2_{x_2}(B_2)\,\mathrm{d}(X_1(P) \otimes X_2(P))(x_1, x_2) \\
&\quad= \left(\int_{A_1} P^1_{x_1}(B_1)\,\mathrm{d}X_1(P)(x_1)\right)\left(\int_{A_2} P^2_{x_2}(B_2)\,\mathrm{d}X_2(P)(x_2)\right) \\
&\quad= P(X_1 \in A_1, Y_1 \in B_1)P(X_2 \in A_2, Y_2 \in B_2) \\
&\quad= P((X_1, Y_1) \in A_1 \times B_1)P((X_2, Y_2) \in A_2 \times B_2)
\end{aligned}$$

From fixing A_1 and B_1 and letting A_2 and B_2 vary we obtain (by using the uniqueness of

measures)

$$P((X_1,Y_1)\in A_1\times B_1,(X_2,Y_2)\in G_2)$$
$$= P((X_1,Y_1)\in A_1\times B_1)P((X_2,Y_2)\in G_2)$$

for all $A_1\in\mathbb{E}_1$, $B_1\in\mathbb{K}_1$ and $G_2\in\mathbb{E}_2\otimes\mathbb{K}_2$. And by fixing G_2 and letting A_1 and B_1 vary, we obtain that

$$P((X_1,\,Y_1)\in G_1,(X_2,Y_2)\in G_2) = P((X_1,Y_1)\in G_1)P((X_2,Y_2)\in G_2)$$

for all $G_1\in\mathbb{E}_1\otimes\mathbb{K}_1$ and $G_2\in\mathbb{E}_2\otimes\mathbb{K}_2$. Thereby it is seen that (X_1,Y_1) and (X_2,Y_2) are independent. □

2.2 Conditional moments

Recall (See appendix) that if X and Y are random variables defined on $(\Omega,\mathbb{F},P)$ with values in $(\mathcal{X},\mathbb{E})$ and $(\mathbb{R},\mathbb{B})$ respectively, and furthermore $E|Y|<\infty$, then there exists a random variable $E(Y|X)$ satisfying 1)–3) in Theorem A.2.6

1) $E(Y\mid X)$ is $\sigma(X)$-measurable
2) $E|E(Y\mid X)|<\infty$
3) For all $A\in\mathbb{E}$ it holds that

$$\int_{(X\in A)} E(Y\mid X)\mathrm{d}P = \int_{(X\in A)} Y\mathrm{d}P\,,$$

and that the fact that $E(Y|X)$ is $\sigma(X)$–measurable is equivalent to the existence of a measurable map $\phi:(\mathcal{X},\mathbb{E})\to(\mathbb{R},\mathbb{B})$ such that

$$E(Y|X)=\phi(X)\quad P \text{ almost surely}\,.$$

We call $\phi(x)$ the conditional expectation of Y given $X=x$ and write

$$\phi(x)=E(Y|X=x)\,.$$

We have the following important (and notationally comfortable) result saying, that the conditional expectation $E(Y|X=x)$ can be calculated as the expectation in the conditional distribution

Theorem 2.2.1. *Assume that X and Y are random variables defined on $(\Omega, \mathbb{F}, P)$ and with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathbb{R}, \mathbb{B})$ respectively. Let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of Y given X.*

If $E|Y| < \infty$, then $X(P)(A_0) = 1$, where

$$A_0 = \{x \in \mathcal{X} \,:\, P_x \text{ has finite first order moment}\}\,.$$

Define for $x \in A_0$

$$E(Y \,|\, X = x) = \int y \,\mathrm{d}P_x(y)$$

Then the function $x \mapsto E(Y \,|\, X = x)$ is a conditional expectation of Y given X.

Note: The last result above could be understood as follows: Define the function $\phi : \mathcal{X} \to \mathbb{R}$

$$\phi(x) = 1_{A_0}(x) \int y \,\mathrm{d}P_x(y)$$

Then the random variable $\phi(X)$ satisfies 1)–3) in Theorem A.2.6, such that it is a conditional expectation of Y given X.

Proof. Consider the function $f : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ given by $f(x, y) = y$. Since

$$\int |f(x,y)| \,\mathrm{d}(X,Y)(P)(x,y) = \int |f(X,Y)| \,dP = E|Y| < \infty\,,$$

it follows from the extended Fubini, that

$$\int |y| \,\mathrm{d}P_x(y) = \int |f(x,y)| \,\mathrm{d}P_x(y) < \infty$$

for $X(P)$ almost all $x \in \mathcal{X}$, such that $X(P)(A_0) = 1$.

Define $\phi : \mathcal{X} \to \mathbb{R}$ by

$$\phi(x) = \begin{cases} \int y \,\mathrm{d}P_x(y)\,, & x \in A_0 \\ 0\,, & x \notin A_0 \end{cases}$$

Then according to the extended Fubini, we have that ϕ is $\mathbb{E} - \mathbb{B}$–measurable. We will argue, that $\phi(X)$ is a conditional expectation of Y given X by verifying the conditions 1)–3). Since ϕ is measurable we have that $\phi(X)$ is $\sigma(X)$–measurable. Furthermore (using the change-of-

variable theorem, Theorem A.1.8)

$$\begin{aligned}
E|\phi(X)| &= \int |\phi(X)|\, dP \\
&= \int |\phi(x)|\, \mathrm{d}X(P)(x) \\
&= \int_{A_0} |\int y\, \mathrm{d}P_x(y)|\, \mathrm{d}X(P)(x) \\
&\leq \int \left(\int |y| \mathrm{d}P_x(y) \right) \mathrm{d}X(P)(x) \\
&= \int |y|\, \mathrm{d}(X,Y)(P)(x,y) < \infty
\end{aligned}$$

In the last equality we have used the extended Tonelli. This shows, that 2) is satisfied for $\phi(X)$. Finally we have for $A \in \mathbb{E}$

$$\begin{aligned}
\int_{(X \in A)} \phi(X)\, dP &= \int_A \phi(x)\, \mathrm{d}X(P)(x) \\
&= \int_{A \cap A_0} \int y\, \mathrm{d}P_x(y)\, \mathrm{d}X(P)(x) \\
&= \iint 1_{A \cap A_0}(x) y\, \mathrm{d}P_x(y)\, \mathrm{d}X(P)(x) \\
&= \int 1_{A \cap A_0}(x) y\, \mathrm{d}(X,Y)(P)(x,y) \\
&= \int 1_{A \cap A_0}(X) Y\, dP \\
&= \int_{(X \in A \cap A_0)} Y\, dP \\
&= \int_{(X \in A)} Y\, dP
\end{aligned}$$

In the fourth equality we have used the extended Fubini's theorem. This shows that also condition 3) is fulfilled, such that $\phi(X)$ is a conditional expectation of Y given X, so we in particular have

$$E(Y \mid X = x) = \phi(x) = \int y\, \mathrm{d}P_x(y)$$

for $x \in A_0$. □

In the framework of theorem 2.2.1, where $E|Y| < \infty$, we have that

$$\int y\, \mathrm{d}P_x(y) = E(Y \mid X = x)$$

Notation: Since $\phi(x) = 1_{A_0}(x) \int y \, \mathrm{d}P_x(y)$ is a measurable function of x according to the extended Fubini, then it makes sense to consider the random variable $\phi(X)$ and write

$$E(Y|X) = \phi(X)$$

As noted in the comment after Theorem 2.2.1 this equals a version of the "ordinary" conditional expectation of Y given X.

The following result, that is already known for conditional expectations, can be shown using the conditional distribution results from the proof of Theorem 2.2.1:

Theorem 2.2.2. *Assume that X and Y are random variables defined on $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathbb{R}, \mathbb{B})$ respectively. If $E|Y| < \infty$, then $E|E(Y \mid X)| < \infty$ and*

$$E((E(Y \mid X)) = EY$$

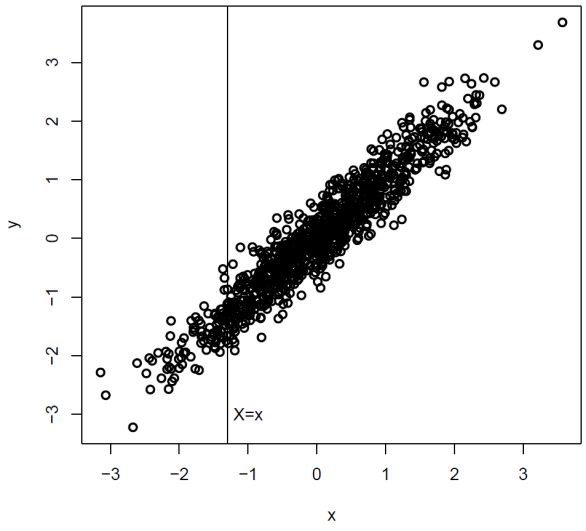

Figure 2.1: 1000 points simulated from the distribution from (X, Y). EY describes the center of all points projected to the y–axis (here this is ≈ 0). $E(Y \mid X = x)$ describes the mean of the points, that has first coordinate x (here this mean will be ≈ -1.5)

Proof. In the proof of 2.2.1 we saw that $\phi(X) = E(Y \mid X)$ is integrable, and that

$$\int_{(X \in A)} \phi(X) \, dP = \int_{(X \in A)} Y \, dP$$

Simply let $A = \mathcal{X}$. Then

$$E(E(Y \mid X)) = \int \phi(X) = \int Y \, dP = EY$$

□

Theorem 2.2.3. *Assume that X and Y are random variables defined $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Suppose that the conditional distribution $(P_x)_{x\in\mathcal{X}}$ of Y given X exists. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable function and define $Z = \phi(X, Y)$. Assume that $E|Z| < \infty$. Then*

$$E(Z \,|\, X = x) = \int \phi(x, y) \, \mathrm{d}P_x(y)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

Proof. According to Theorem 2.1.1 we have that the conditional distribution of Z given X is given by the Markov kernel $(\tilde{P}_x)_{x\in\mathcal{X}}$, where

$$\tilde{P}_x = (\phi \circ i_x)(P_x)$$

Then according to Theorem 2.2.1 we have for $X(P)$ almost all $x \in \mathcal{X}$

$$E(Z \,|\, X = x) = \int z \, \mathrm{d}\tilde{P}_x(z) = \int (\phi \circ i_x)(y) \, \mathrm{d}P_x(y) = \int \phi(x, y) \, \mathrm{d}P_x(y)$$

□

Corollary 2.2.4. *Assume that X and Y are independent random variables defined $(\Omega, \mathbb{F}, P)$ with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable function and define $Z = \phi(X, Y)$. Assume that $E|\phi(X, Y)| < \infty$. Then*

$$E(\phi(X, Y) \,|\, X = x) = \int \phi(x, y) \, \mathrm{d}X(P)(y)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

Proof. We have that $(X(P))_{x\in\mathcal{X}}$ is the conditional distribution of Y given X. □

We can also use Theorem 2.2.3 to show the following result, that is well–known from the framework of conditional expectations:

Corollary 2.2.5. *Assume that Y and Z are real valued random variables with $E|Y| < \infty$ and $E|Z| < \infty$. Let X be a random variable with values in $(\mathcal{X}, \mathbb{E})$. Then*

$$E(Y + Z \,|\, X = x) = E(Y \,|\, X = x) + E(Z \,|\, X = x)$$

for $X(P)$ almost all $x \in \mathcal{X}$.

Proof. Let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of (Y,Z) given X. Then

$$E(Y+Z\,|\,X=x)=\int (y+z)\,\mathrm{d}P_x(y,z)$$

for $X(P)$ almost all $x\in\mathcal{X}$. And

$$E(Y\,|\,X=x)=\int y\,\mathrm{d}P_x(y,z)\qquad E(Z\,|\,X=x)=\int z\,\mathrm{d}P_x(y,z)$$

for $X(P)$ almost all $x\in\mathcal{X}$. □

If Y is real valued with $EY^2<\infty$, and $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of Y given X, then we can define the conditional variance of Y given $X=x$ by

$$V(Y\,|\,X=x)=\int y^2\,\mathrm{d}P_x(y)-\left(\int y\,\mathrm{d}P_x(y)\right)^2$$

which will be well–defined for $X(P)$ almost all $x\in\mathcal{X}$. Letting $V(Y\,|\,X)$ be the composition of X and $x\mapsto V(Y\,|\,X=x)$ gives

$$V(Y\,|\,X)=E(Y^2\,|\,X)-E(Y\,|\,X)^2$$

Theorem 2.2.6. *Let X and Y be random variables defined on $(\Omega,\mathbb{F},P)$ with values in $(\mathcal{X},\mathbb{E})$ and $(\mathbb{R},\mathbb{B})$ respectively. If $EY^2<\infty$, then*

$$VY=E\big(V(Y\,|\,X)\big)+V\big(E(Y\,|\,X)\big)\,.$$

Proof.

$$\begin{aligned}E\big(V(Y\,|\,X)\big)+V\big(E(Y\,|\,X)\big)&=E\big(E(Y^2\,|\,X)-E(Y\,|\,X)^2\big)\\&\quad+E\big(E(Y\,|\,X)^2\big)-\big(E\big(E(Y\,|\,X)\big)\big)^2\\&=E\big(E(Y^2\,|\,X)\big)-\big(E\big(E(Y\,|\,X)\big)\big)^2\end{aligned}$$

□

Example 2.2.7. Consider Figure 2.2. The variance VY measures how much the projection of all points onto the y–axis varies around their center. $V(Y\,|\,X=x)$ measures how much the part of the points, that have first coordinate x varies around their center. The two expressions are normally not particularly close. In this example VY is rather big, while $V(Y\,|\,X=x)$ is small for *all* x. ∘

Example 2.2.8. In example 2.1.2 we studied the situation where

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_{r+s}\left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right),$$

and we found that

$$X_1 \,|\, X_2 = x \sim \mathcal{N}_r(\xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x - \xi_2)\,,\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

If we assume that X_1 is one–dimensional, we have defined $E(X_1 \mid X_2 = x)$ and $V(X_1 \mid X_2 = x)$. Since conditional expectations and conditional variances are calculated as expectations and variances in the conditional distributions, we have

$$\begin{aligned} E(X_1 \mid X_2) &= \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x - \xi_2) \\ V(X_1 \mid X_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

Note that the conditional variance does not depend on x but is different from $V(X_1)$. ∘

Confusing conditional variances and ordinary variances is a quite common mistake – and that may lead to substantial problems.

2.3 Exercises

Exercise 2.1. Assume that X is uniformly distributed on $(0, 1)$ and that the conditional distribution of Y given $X = x$ is a binomial distribution with parameters (n, x) We could say that Y has a binomial distribution with fixed length n and random probability parameter.

(1) What are the possible values of Y? Argue that $E|Y| < \infty$.

(2) Find $E(Y \,|\, X = x)$ and $E(Y \,|\, X)$.

(3) Find EY.

(4) Find $P(Y = k)$ for all k being a possible value of Y. What is the marginal distribution of Y?

∘

Exercise 2.2. Let X and Y be random variables with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ respectively. Assume that (P_x) is the conditional distribution of Y given X. Let

$$A_0 = \{x \in \mathcal{X} \mid \int |y| \, P_x(dy) < \infty\}$$

and assume that $X(P)(A_0) = 1$. Define

$$\phi(x) = 1_{A_0}(x) \int |y| \, P_x(dy) \, .$$

Show that $E\phi(X) = E|Y|$ and conclude that

$$E\phi(X) < \infty \quad \text{if and only if} \quad E|Y| < \infty$$

(Hint). ∘

Exercise 2.3. Assume that X has the exponential distribution with mean 1, and assume that the conditional distribution of Y given $X = x$ is a Poisson distribution with parameter x. We could say that Y is Poisson distributed with random parameter.

(1) Use Exercise 2.2 to argue that $E|Y| < \infty$.

(2) Find $E(Y \mid X = x)$ and $E(Y \mid X)$.

(3) Find EY.

(4) Find $P(Y = k)$ for all k being a possible value of Y. What is the marginal distribution of Y?

∘

Exercise 2.4. Let X and Y be independent random variables that both have the uniform distribution on $(0, 1)$. Define $Z = XY$.

(1) Find the conditional distribution of Z given X.

(2) What are the possible values of Z? Argue that $E|Z| < \infty$.

(3) Find $E(Z \mid X)$ and use this to find EZ.

(4) Find EZ without using conditional distributions.

∘

Exercise 2.5. Assume that $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables such that $E|X_1| < \infty$. Assume that N is a random variables with values in $\mathbb{N}$ such that $EN < \infty$. Assume that N and $(X_1, X_2, \ldots)$ are independent (we consider $(X_1, X_2, \ldots)$ as a random variable with values in $(\mathbb{R}^\infty, \mathbb{B}^\infty)$). Define the random variable Y by

$$Y = \sum_{k=1}^{N} Y_k$$

(1) Show that the conditional distribution $(P_n)_{n\in\mathbb{N}}$ of Y given N is determined such that P_n is the distribution of $\sum_{k=1}^{n} Y_k$. Argue similarly that the conditional distribution $(Q_n)_{n\in\mathbb{N}}$ of $\sum_{k=1}^{N} |Y_k|$ given N is determined such that Q_n is the distribution of $\sum_{k=1}^{n} |Y_k|$.

(2) Show that

$$\int |y| \, Q_n(dy) = n|E_1|$$

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

(3) Use (2) and Exercise 2.2 to obtain that

$$E\left(\sum_{k=1}^{N} |Y_k|\right) = ENE|Y_1| < \infty$$

(4) Show that $E(Y \mid N = n) = nEY_1$ and that $EY = ENEY_1$.

∘

Exercise 2.6. Let f and g be densities for distributions on $[0, \infty)$. Assume that there exists a constant $c > 0$ such that

$$f(x) \leq cg(x) \qquad \text{for all } x \in [0, \infty)$$

Think of a situation where we want to simulate random variables with a distribution that has density f, but where f is so complicated that this is not straightforward to do directly. Suppose on the other hand that g is a simple well–known density that we actually *can* simulate from. An algorithm to produce a random variable X with density f is the **acceptance–rejection algorithm**:

(i) Generate Y with density g and U uniform on $(0,1)$ such that $Y \perp\!\!\!\perp U$

(ii) If $U \leq f(Y)/(cg(Y))$, let $X = Y$. Otherwise return to (i)

The idea of this exercise is to show that X generated in the algorithm above actually has density f.

So let Y have density g and let U be uniform on $(0,1)$. Assume that Y and U are independent. Define the random variable

$$Z = \begin{cases} 1\,, & U \leq \frac{f(Y)}{cg(Y)} \\ 0\,, & U > \frac{f(Y)}{cg(Y)} \end{cases}$$

(1) Show that $P(Z = 1) = \frac{1}{c}$ (Hint).

(2) Show that $P(Y \in B \,|\, Z = 1) = \int_B f(x)\,dx$ for all $B \in \mathbb{B}$.

(3) Conclude that the algorithm produces a variable X with density f, and discuss which value of c we should choose.

∘

Exercise 2.7. Think of a situation where we want to estimate the value z that is given by

$$z = EZ$$

for some real valued random variable Z. Let $Z_1, Z_2, \ldots, Z_n$ be independent replications of Z. Then

$$\hat{z}_n^1 = \frac{1}{n}\sum_{k=1}^{n} Z_k$$

is an estimator for z.

(1) Show that $\hat{z}_n^1$ is unbiased

$$E\hat{z}_n^1 = z$$

and find the variance $V\hat{z}_n^1$.

A method to improve the estimator could be finding some random variable X and consider the new variable $E(Z \,|\, X)$. Now let $(Z_1, X_1), \ldots, (Z_n, X_n)$ be independent replications of (Z, X), and define the estimator

$$\hat{z}_n^2 = \frac{1}{n}\sum_{k=1}^{n} E(Z_k \,|\, X_k)$$

(2) Show that $\hat{z}_n^2$ is unbiased and that

$$V\hat{z}_n^2 \leq V\hat{z}_n^1$$

Apparently this method will improve the estimator no matter which variable X we choose. But of course some choices may be more clever than others.

(3) What happens, if we use $X = 1$ (or some other constant), and why is this a bad idea anyway?

We will consider two specific examples of variables Z. In both examples we shall just let $n = 1$, since increasing values of n simply makes both variances smaller by a factor $1/n$, and thereby does not change anything in the comparison.

In the first example we shall find estimators for the very well–known value π (although we already know π much more accurately than we will ever be able to estimate, the example serves as a very good illustration of what is going on). Let

$$Z = 4 \cdot 1_{(U_1^2+U^2\leq 1)}\,,$$

where U_1 and U_2 are independent and both uniform on $(0, 1)$. Define the first estimator $\hat{z}_1 = Z$.

(4) Show that $E\hat{z}_1 = \pi$ (Hint).

Define the estimator $\hat{z}_2$ by

$$\hat{z}_2 = E(Z \,|\, U_1)$$

(5) Show that $\hat{z}_2 = 4\sqrt{1 - U_1^2}$.

(6) Try to simulate 10000 replications of both $\hat{z}_1$ and $\hat{z}_2$. Compare the variances – and also compare with the theoretical variance of $\hat{z}_1$.

In the next example, the estimation has some real practical use. Assume that X_1 and X_2 are independent and has a distribution ν. Assume that $X_1, X_2 \geq 0$ and that ν has density f with

respect to the Lebesgue measure. Furthermore think of a situation, where the distribution of $S = X_1 + X_2$ is complicated to calculate. We are interested in estimating

$$z(s) = P(S > x)$$

especially for large values of x.

The simple estimator will in this framework be

$$\hat{z}_1(x) = 1_{(X_1+X_2>x)}$$

The problem is, that if x is very large, then it is very rare that this estimator is non–zero. Even if we make many replications. Instead we shall try to construct an estimator using conditional expectations.

Firstly, we try something similar to above. Define

$$\hat{z}_2(x) = P(S > x \,|\, X_1)$$

(7) Show that

$$\hat{z}_2(x) = \bar{F}(x - X_1)\,,$$

where $\bar{F}$ is the *survival function* for ν:

$$\bar{F}(x) = \nu((x, \infty))\,.$$

Let

$$X_{(1)} = \min\{X_1, X_2\} \qquad \text{and} \qquad X_{(2)} = \max\{X_1, X_2\}$$

(8) Show that the conditional distribution of $X_{(2)}$ given $X_{(1)}$ is determined by the Markov kernel $(P_y)_{y\geq 0}$, where

$$P_y(B) = \frac{\nu(B \cap (y, \infty))}{\nu((y, \infty))}$$

(Hint).

We now define a conditional estimator by

$$\hat{z}_3(x) = P(S > x \,|\, X_{(1)})$$

(9) Show that

$$\hat{z}_3(x) = \frac{\bar{F}\big(\max\{x - X_{(1)}, X_{(1)}\}\big)}{\bar{F}(X_{(1)})}\,,$$

Now assume that ν is the Weibull distribution with shape parameter 0.5. Then the density f is given by

$$0.5x^{-0.5}e^{-x^{0.5}}$$

for $x > 0$. And $\bar{F}$ is

$$\bar{F}(x) = e^{-x^{0.5}}$$

(10) Simulate 10000 replications of the three estimators (with e.g. $x = 20$ and $x = 50$) and compare the variances.

∘

Exercise 2.8. Let X be a real valued random variable with $E|X| < \infty$.

(1) Show that the conditional distribution of X given X is given by the Markov kernel $(\delta_x)_{x\in\mathcal{X}}$, where δ_x is the Dirac Measure in x:

$$\delta_x(B) = \begin{cases} 1\,, & x \in B \\ 0\,, & x \notin B \end{cases}$$

(2) Show that $E(X \mid X = x) = x$ and $E(X \mid X) = X$.

(3) Assume that Y is another real valued random variable with $E|Y| < \infty$ and $E|XY| < \infty$. Show that $E(XY \mid X = x) = xE(Y \mid X = x)$ and $E(XY \mid X) = XE(Y \mid X)$.

∘

Exercise 2.9. Let W be the set $(0, 1)^2$. Assume that we generate N points in W in the following way:

- Let N be Poisson distributed with parameter λ.
- Let $(U_1^1, U_1^2), (U_2^1, U_2^2), \ldots, (U_N^1, U_N^2)$ be independent and identically distributed such that U_k^1 and U_k^2 are independent and uniformly distributed on $(0, 1)$. This makes each (U_k^1, U_k^2) uniformly distributed on W.

In this exercise we will show that the collection of points $(U_1^1, U_1^2), \ldots, (U_N^1, U_N^2)$ in W is a *Poisson process* on W: Define for a subset $A \subseteq W$ the random variable $N(A)$ to be the

number of points in A:

$$N(A) = \sum_{k=1}^{N} 1_{(U_1^k, U_2^k) \in A)}$$

- $N(A)$ is Poisson distributed with parameter $\lambda m_2(A)$, where $m_2(A)$ is the area (2–dimensional Lebesgue measure) of A.
- For disjoint sets $A_1, \ldots, A_m$ the variables $N(A_1), \ldots, N(A_m)$ are independent.

The result will follow by finding conditional distributions given N

(1) Show that for U_1 and U_2 independent and uniformly distributed on $(0, 1)$ and A some subset of W, then

$$P((U_1, U_2) \in A) = m_2(A)$$

(2) Let $A_1, \ldots, A_m$ be disjoint subsets of W such that $\bigcup_{j=1}^{m} A_j = W$. Argue that the conditional distribution of $(N(A_1), \ldots, N(A_m))$ given $N = n$ is a polynomial distribution with length n and probability parameters $(m_2(A_1), \ldots, m_2(A_m))$ (Hint).

(3) Show that $N(A_1), \ldots, N(A_m)$ are independent and that each $N(A_j)$ is Poisson distributed with parameter $\lambda m_2(A_j)$ (Hint).

Now assume that $k : W \to [0, 1]$ is a measurable function that is bounded by 1. Define for each subset A of W the number

$$K(A) = \int_A k(x, y)\, m_2(dx, dy)$$

(4) Give a suggestion for how to obtain a collection of points $(V_1^1, V_1^2), \ldots, (V_M^1, V_M^2)$ in W, such that for each subset A of W we have that the number of points in A

$$M(A) = \sum_{k=1}^{M} 1_{((V_j^1, V_j^2) \in A)}$$

is Poisson distributed with parameter $\lambda K(A)$ (Hint).

∘

Exercise 2.10. Assume that (X, Y) is a real valued random vector, such that $E|Y| < \infty$. Assume that the random vector $(X, \tilde{Y})$ has the same distribution as (X, Y), where $\tilde{Y}$ is another real valued random variable..

(1) Show that $E(Y \mid X) = E(\tilde{Y} \mid X)$ a.s.

Now assume that $X_1, \ldots, X_n$ are independent and identically distributed with $E|X_1| < \infty$. Define $S_n = X_1 + \cdots + X_n$.

(2) Argue that (X_1, S_n) has the same distribution as (X_k, S_n) for all $k = 1, \ldots, n$.

(3) Show that $E(X_1 \mid S_n) = S_n/n$ (Hint).

$\circ$

Chapter 3

Conditional independence

In this chapter we will work on a general probability space $(\Omega, \mathbb{F}, P)$. All events occurring will silently be assumed to be $\mathbb{F}$-measurable, all σ-algebras occurring will silently be assumed to be subalgebras of $\mathbb{F}$, and all stochastic variables $X : (\Omega, \mathbb{F}) \to (\mathcal{X}, \mathbb{E})$ will silently be assumed to be $\mathbb{F} - \mathbb{E}$ measurable.

The general convention is that random variables with names like X or X_i or variations thereof have values in a generic space $(\mathcal{X}, \mathbb{E})$, unless it is explicitly stated that they are real valued (or integer valued or whatever). Similarly, variables with names like Y or Z will have values in $(\mathcal{Y}, \mathbb{K})$ and $(\mathcal{Z}, \mathbb{G})$ respectively.

Recall that $(\mathcal{X}, \mathbb{E})$ is a **Borel space** if it is in bijective, bimeasurable correspondence with $(\mathbb{R}, \mathbb{B})$ or a subspace of this. Such a correspondence enables us to replace $\mathcal{X}$ with $\mathbb{R}$, whenever there is an advantage in that. It turns out that every sensible space has this property, unless it its very, very huge (non-separable metric spaces, with the σ-algebra generated by the open sets, say).

The above generic $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$-spaces are always assumed to be Borel spaces.

3.1 Conditional probabilities given a σ–algebra

So far we have considered conditional expectations $E(Y \mid X)$, where we condition on a random variable X. This is an integrable and $\sigma(X)$–measurable random variable satisfying

$$\int_{(X \in A)} E(Y \mid X)\, dP = \int_{(X \in A)} Y\, dP$$

for all $A \in \mathcal{X}$. We have furthermore seen that this variable can be calculated as the mean in the conditional distribution of Y given X.

A natural generalisation of this concept is conditional expectations given a σ–algebra: For a real valued random variable Y satisfying that $E|Y| < \infty$ and a σ-algebra $\mathbb{H}$, the **conditional expectation** $E(Y \mid \mathbb{H})$ of Y given $\mathbb{H}$ is any $\mathbb{H}$-measurable and integrable random variable satisfying the integral conditions

$$\int_H E(Y \mid \mathbb{H})\, dP = \int_H Y\, dP \qquad \text{for all } H \in \mathbb{H}\,. \tag{3.1}$$

as defined in Definition A.2.2.

We shall be concerned with the **conditional probability** of an event A given $\mathbb{H}$ – This is defined similar to $P(A \mid X)$: We simply use the conditional expectation of the indicator 1_A, that is

$$P(A \mid \mathbb{H}) = E(1_A \mid \mathbb{H})\,.$$

The integrability condition (3.1) will in this case take the form

$$\int_H P(A \mid \mathbb{H})\, dP = P(A \cap H) \qquad \text{for all } H \in \mathbb{H}\,. \tag{3.2}$$

We will make frequent use of the monotonicity property of conditional expectations, that make sure that

$$0 \leq P(A \mid \mathbb{H}) \leq 1 \qquad \text{a.s.}$$

and even that

$$A \subseteq B \quad \Rightarrow \quad P(A \mid \mathbb{H}) \leq P(B \mid \mathbb{H}) \quad \text{a.s}$$

Furthermore, the double conditioning theorem (Theorem A.2.5) says in this context that

$$E\Big(P(A \mid \mathbb{H}) \mid \mathbb{G}\Big) = P(A \mid \mathbb{G}) \qquad \text{a.s}$$

whenever the two σ-algebras $\mathbb{G}$ and $\mathbb{H}$ satisfies that $\mathbb{G} \subseteq \mathbb{H}$.

3.2 Conditionally independent events

Definition 3.2.1. *Two events A and B are are* **conditionally independent** *given a σ-algebra $\mathbb{H}$, if*

$$P(A \cap B \mid \mathbb{H}) = P(A \mid \mathbb{H})\, P(B \mid \mathbb{H}) \qquad a.s. \tag{3.3}$$

Symbolically, we will write $A \perp\!\!\!\perp B \mid \mathbb{H}$ if (3.3) is satisfied.

Speaking colloquially, we will frequently say that A and B are independent given $\mathbb{H}$ if (3.3) is satisfied - repeated use of the word *conditionally* makes the sentences sound tedious.

Please note that conditional independence represents an intricate relation between the two events and the σ-algebra. The σ-algebra $\mathbb{H}$ is really an integral part of the definition. Whether A and B are conditionally independent or not, depends crucially on which σ-algebra we are conditioning.

If $\mathbb{H} \subseteq \mathbb{G}$ are two σ-algebras, it is completely possible that two events A and B are independent given $\mathbb{H}$, while they are not independent given the finer σ-algebra $\mathbb{G}$. But it is equally possible that A and B are independent given $\mathbb{G}$, while they are not independent given the coarser σ-algebra $\mathbb{H}$. Changing the σ-algebra on which we are conditioning is usually a very challenging task - and indeed a task which is at the core of Markov Chain Theory.

Example 3.2.2. Recall that a σ-algebra $\mathbb{H}$ is a **trivial** if every event in $\mathbb{H}$ has probability 0 or 1. The most obvious trivial σ-algebra is

$$\mathbb{H} = \{\emptyset, \Omega\}\,,$$

but there are plenty of other trivial algebras arising all over probability theory - tail algebras, symmetric algebras, invariant σ-algebras in ergodic theory and what not. If $\mathbb{H}$ is trivial, we observe that

$$P(A \mid \mathbb{H}) = P(A) \qquad \text{a.s.}$$

for any event A, since the relation

$$\int_H P(A)\, dP = P(A \cap H)\,,$$

is satisfied for all $\mathbb{H}$-sets H, both those of probability 0 (where there is nothing to prove) and those of probability 1 (where there is also nothing to prove). Hence (3.3) translates to

$$P(A \cap B) = P(A)\, P(B)\,. \tag{3.4}$$

A priori the formula has an a.s.-qualifier, but as it is a relation between deterministic numbers, it is either true or false, with no probability involved.

Hence we see that conditional independence of two events given a trivial σ-algebra is simply classical independence of the events. ∘

Example 3.2.3. If C is yet another event, and if $\mathbb{H}$ is the σ-algebra generated by that event,

$$\mathbb{H} = \{\emptyset, C, C^c, \Omega\},$$

then it is readily checked that

$$P(A \mid \mathbb{H}) = \begin{cases} \dfrac{P(A \cap C)}{P(C)} & \text{on } C \\ \dfrac{P(A \cap C^c)}{P(C^c)} & \text{on } C^c \end{cases} \qquad \text{a.s}$$

for any event A. If we suppose that $\mathbb{H}$ is non-trivial, meaning that $P(C) \in (0,1)$, we see that (3.3) translates to the two conditions

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

$$\frac{P(A \cap B \cap C^c)}{P(C^c)} = \frac{P(A \cap C^c)}{P(C^c)} \frac{P(B \cap C^c)}{P(C^c)}.$$

These two conditions cannot be deduced from each other, and they are not related to (3.4). For instance, the probability table

C			C^c		
	B	B^c		B	B^c
A	$\frac{2}{18}$	$\frac{1}{18}$	A	$\frac{2}{18}$	$\frac{4}{18}$
A^c	$\frac{4}{18}$	$\frac{2}{18}$	A^c	$\frac{1}{18}$	$\frac{2}{18}$

corresponds to a situation where $A \perp\!\!\!\perp B \mid \mathbb{H}$ but where A and B are dependent, as can readily be checked.

On the other hand, the probability table

C			C^c		
	B	B^c		B	B^c
A	$\frac{1}{12}$	$\frac{2}{12}$	A	$\frac{2}{12}$	$\frac{1}{12}$
A^c	$\frac{2}{12}$	$\frac{1}{12}$	A^c	$\frac{1}{12}$	$\frac{2}{12}$

corresponds to a situation where A and B are independent, but where they are **not** independent given $\mathbb{H}$. ∘

Example 3.2.4. If we have a finite partition $\mathbb{D}$ of Ω,

$$\mathbb{D} = \{D_1, \ldots, D_n\}$$

where the **atoms** of $\mathbb{D}$ (the D_i-sets) are pairwise disjoint and unite to the whole of Ω, the σ-algebra generated by $\mathbb{D}$ is the family of all unions,

$$\mathbb{H} = \left\{ \bigcup_{i \in I} D_i \mid I \subseteq \{1, \ldots, n\} \right\}.$$

If we let

$$\mathbb{D}^* = \{D \in \mathbb{D} \mid P(D) > 0\},$$

it is easily checked that

$$P(A \mid \mathbb{H}) = \sum_{D \in \mathbb{D}^*} \frac{P(A \cap D)}{P(D)} 1_D \qquad \text{a.s.}$$

for any event A. In this setting, condtion (3.3) translates into

$$\frac{P(A \cap B \cap D)}{P(D)} = \frac{P(A \cap D)}{P(D)} \frac{P(B \cap D)}{P(D)} \qquad \text{for all } D \in \mathbb{D}^*.$$

Again, whether this holds or not is very sensitive to the specific atoms. If an atom is divided into two, there is no telling if A and B are independent on each of the two subatoms, just because we know if they are independent on the original atom. And similarly, if two atoms are coalesced, we may loose or create conditional independence, as the case may be. ∘

3.3 Conditionally independent σ-algebras

Definition 3.3.1. *Two classes of events, $\mathcal{A}$ and $\mathcal{B}$, are conditionally independent given a σ-algebra $\mathbb{H}$ if*

$$A \perp\!\!\!\perp B \mid \mathbb{H} \qquad \textit{for all } A \in \mathcal{A}, B \in \mathcal{B}. \tag{3.5}$$

Symbolically, we will write $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H}$ if (3.5) is satisfied.

We will almost exclusively use this concept in situations where the two classes of events are σ-algebras, but it is nice to be allowed to formulate things in a slightly broader fashion. We may for instance see that it typically is enough to check (3.5) on two generators of the σ-algebras under consideration:

Lemma 3.3.2. *Let $\mathcal{A}$ and $\mathcal{B}$ be two classes of events, both stable under formation of intersections. Then*

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H} \qquad \Rightarrow \qquad \sigma(\mathcal{A}) \perp\!\!\!\perp \sigma(\mathcal{B}) \mid \mathbb{H}\,.$$

Proof. A prototypical application of Dynkins lemma. For each set $F \in \mathbb{F}$ we consider the class

$$\mathcal{C}_F = \{E \in \mathbb{F} \mid F \perp\!\!\!\perp E \mid \mathbb{H}\}\,,$$

and we observe that this is a Dynkin class. If we take $A \in \mathcal{A}$, we know that $\mathcal{B} \subseteq \mathcal{C}_A$. Using Dynkins lemma, we see that $\sigma(\mathcal{B}) \subseteq \mathcal{C}_A$. On the other hand, conditional independence of two events is a property that is symmetric in the two events, so we can reformulate this fact as $\mathcal{A} \subseteq \mathcal{C}_B$ for any set $B \in \sigma(\mathcal{B})$. Using Dynkins lemma again establishes that $\sigma(\mathcal{A}) \subseteq \mathcal{C}_B$ for any set $B \in \sigma(\mathcal{B})$. And though this may look awkward, it is in fact the property we are after. □

Conditional independence of classes of events is of course just as sensitive to the exact choice of the σ-algebra on which we are conditioning, as conditional independence of events were. In fact, if

$$\mathbb{A} = \{\emptyset, A, A^c, \Omega\}\,, \qquad \mathbb{B} = \{\emptyset, B, B^c, \Omega\}\,,$$

then $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$ if and only if $A \perp\!\!\!\perp B \mid \mathbb{H}$, as is readily seen from lemma 3.3.2. So the counterexamples to any kind of simple behaviour under change of the conditioning algebra given in section 3.2 also apply in this setting.

Example 3.3.3. Assume that $\mathbb{H}$ is a trivial σ–algebra. Then we saw in Example 3.2.2, that two sets A and B are conditionally independent given $\mathbb{H}$, if and only if they are truly independent. This translates directly into conditional independence of classes of events: If $\mathbb{H}$ is trivial, then two classes $\mathcal{A}$ and $\mathcal{B}$ satisfies

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H} \qquad \Leftrightarrow \qquad \mathcal{A} \perp\!\!\!\perp \mathcal{B}$$

Assume conversely that $\mathbb{H} = \mathbb{F}$. Then for all $F \in \mathbb{F}$ we have

$$P(F \mid \mathbb{F}) = 1_F \qquad \text{a.s.}\,,$$

since 1_F is $\mathbb{F}$–measurable. Hence it is seen that for any choice of $\mathcal{A}$ and $\mathcal{B}$ we have with $A \in \mathcal{A}$ and $B \in \mathcal{B}$ that

$$P(A \mid \mathbb{F})P(B \mid \mathbb{F}) = 1_A \cdot 1_B = 1_{A \cap B} = P(A \cap B \mid \mathbb{F}) \quad \text{a.s.}\,,$$

so we conclude that $\mathcal{A}$ and $\mathcal{B}$ are always conditionally independent given $\mathbb{F}$. ∘

Example 3.3.4. Assume that $\mathcal{A}$, $\mathcal{B}$ and $\mathbb{H}$ are independent. Then with $A \in \mathcal{A}$ we observe

$$P(A \,|\, \mathbb{H}) = P(A) \qquad \text{a.s.}$$

since for $H \in \mathbb{H}$ the relation

$$\int_H P(A)\, dP = P(A)P(H) = P(A \cap H)$$

is satisfied. Then – using the independence between $\mathcal{A}$ and $\mathcal{B}$ – we obtain

$$P(A \,|\, \mathbb{H})P(B \,|\, \mathbb{H}) = P(A) \cdot P(B) = P(A \cap B) = P(A \cap B \,|\, \mathbb{H}) \quad \text{a.s.}\,.$$

In the last equality we used that $A \cap B \perp\!\!\!\perp \mathbb{H}$ since both A and B are independent of $\mathbb{H}$. We conclude that $\mathcal{A}$ and $\mathcal{B}$ are independent given $\mathbb{H}$ as well. ∘

Lemma 3.3.5 (Reduction)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two classes of events, and let $\mathcal{A}' \subseteq \mathcal{A}$ be a subclass. Then*

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H} \qquad \Rightarrow \qquad \mathcal{A}' \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H}\,.$$

Proof. This is a quite trivial observation, which hardly deserves to be called a lemma. The statement on the right hand side involves fewer events than the statement on the left hand side, so the implication is obvious. □

Theorem 3.3.6. *Let $\mathbb{A}, \mathbb{B}$ and $\mathbb{H}$ be three σ-algebras. Suppose that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$. If X is an $\mathbb{A}$-measurable real valued random variable, and if Y is a $\mathbb{B}$-measurable real valued random variable, such that $E|X| < \infty$, $E|Y| < \infty$ and $E|XY| < \infty$, then it holds that*

$$E(XY \mid \mathbb{H}) = E(X \mid \mathbb{H})\, E(Y \mid \mathbb{H}) \qquad \textit{a.e.}$$

Proof. A prototypical extension result. We know the theorem to be true for indicator variables. Hence it is true for simple variables. The monotone convergence theorem for conditional expectations will show it is true for non-negative variables, and a final handwaving will dismiss the problems of positive and negative parts. □

Conditional independence is by its very definition symmetric in the two events, or more general, in the two classes of events. Rather surprisingly, it turns out that the most fruitful way of working with the concept is through an asymmetric formulation:

Theorem 3.3.7. *Let $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{H}$ be σ-algebras. It holds that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$ if and only if*

$$P(A \mid \mathbb{B} \vee \mathbb{H}) = P(A \mid \mathbb{H}) \qquad a.s \tag{3.6}$$

for every event $A \in \mathbb{A}$.

In the theorem $\mathbb{B} \vee \mathbb{H}$ denotes the smallest σ–algebra that contains both $\mathbb{B}$ and $\mathbb{H}$. This σ–algebra must be generated by the $\cap$–stable generating system given by

$$\{B \cap H \,:\, B \in \mathbb{B}, H \in \mathbb{H}\}$$

Proof. Notice that for any three events $A \in \mathbb{A}$, $B \in \mathbb{B}$ and $H \in \mathbb{H}$ we have that

$$\begin{aligned} \int_{B\cap H} P(A \mid \mathbb{H})\, dP &= \int_H 1_B\, P(A \mid \mathbb{H})\, dP = \int_H E\Big(1_B\, P(A \mid \mathbb{H}) \mid \mathbb{H}\Big)\, dP \\ &= \int_H P(A \mid \mathbb{H})\, P(B \mid \mathbb{H})\, dP \end{aligned} \quad , \tag{3.7}$$

In the second equality we have used the integration property from the definition of conditional expectations. In the third equality we have used the following calculation rule from conditional expectations: If X is $\mathbb{H}$–measurable, then $E(XY \mid \mathbb{H}) = XE(Y \mid \mathbb{H})$ (we also need to assume that $E|X| < \infty$, $E|Y| < \infty$ and $E|XY| < \infty$ such that the conditional expectations are well defined). Suppose that $\mathbb{A}$ and $\mathbb{B}$ are conditionally independent given $\mathbb{H}$. Then we can work the above line of equations one step further to see that

$$\int_{B\cap H} P(A \mid \mathbb{H})\, dP = \int_H P(A \cap B \mid \mathbb{H})\, dP = P(A \cap B \cap H)\,.$$

Since the events of the form $B \cap H$ is a generator for the σ-algebra $\mathbb{B} \vee \mathbb{H}$ that is stable under formation of intersections, and as $P(A \mid \mathbb{H})$ is $\mathbb{H}$-measurable, and thereby in particular $\mathbb{B} \vee \mathbb{H}$–measurable, we conclude that $P(A \mid \mathbb{H})$ indeed does satisfy all conditions for being the conditional probability of A given $\mathbb{B} \vee \mathbb{H}$. And hence (3.6) holds.

For the opposite implication, we may utilise (3.6) on the starting end of (3.7), and obtain that

$$\int_H P(A \mid \mathbb{H})\, P(B \mid \mathbb{H})\, dP = \int_{H\cap B} P(A \mid \mathbb{B} \vee \mathbb{H})\, dP = P(A \cap B \cap H)\,.$$

As $P(A \mid \mathbb{H})\, P(B \mid \mathbb{H})$ is indeed $\mathbb{H}$-measurable, we see that it satisfies all conditions for being the conditional probability of $A \cap B$ given $\mathbb{H}$. And hence A and B are conditionally independent given $\mathbb{H}$. □

The asymmetric condition (3.6) is usually paraphrased by saying that there is no extra information in $\mathbb{B}$ for making predictions on the occurrence of an $\mathbb{A}$-set, when we already have access to the information in $\mathbb{H}$. All the information in $\mathbb{B}$, useful for that prediction, is already contained in $\mathbb{H}$. The symmetry between $\mathbb{A}$ and $\mathbb{B}$ is not clearly visible here, but somehow it is still there.

Corollary 3.3.8. *Let $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{H}$ be σ-algebras. If $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$ then it holds for any $\mathbb{A}$-measurable real random variable X such that $E|X| < \infty$ that*

$$E(X \mid \mathbb{B} \vee \mathbb{H}) = E(X \mid \mathbb{H}) \qquad a.e \tag{3.8}$$

Proof. Follows from 3.3.7 by the same extension technique, that was used to prove theorem 3.3.6. □

Example 3.3.9. Assume that $\mathbb{A}$ and $\mathbb{H}$ are σ–algebras. We clearly have $\mathbb{H} \vee \mathbb{H} = \mathbb{H}$, such that

$$P(A \mid \mathbb{H} \vee \mathbb{H}) = P(A \mid \mathbb{H})$$

Then it follows that $\mathbb{A} \perp\!\!\!\perp \mathbb{H} \mid \mathbb{H}$. The same argument applies to deduce that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$ whenever $\mathbb{B} \subseteq \mathbb{H}$. ∘

3.4 Shifting information around

The asymmetric approach to the defining conditional independence from Theorem 3.3.7 can be explored further: In fact, when we already condition on $\mathbb{H}$ it makes no difference adding sets from $\mathcal{H}$ to the sets from $\mathbb{B}$.

Theorem 3.4.1. *Let $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{H}$ be σ-algebras.*

$$\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H} \quad \Rightarrow \quad \mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{H}) \mid \mathbb{H}\,.$$

Proof. Take $A \in \mathbb{A}$. We have that

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

where the first equality is true for trivial reason (we are conditioning on the same σ-algebra), and the second equality is true due to the conditional independence of $\mathbb{A}$ and $\mathbb{B}$ given $\mathbb{H}$. But now conditional independence of $\mathbb{A}$ and $\mathbb{B} \vee \mathbb{H}$ given $\mathbb{H}$ follows from theorem 3.3.7. □

Theorem 3.4.2. *Let $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{H}$ be σ-algebras. Suppose that $\mathbb{G}$ is yet another σ-algebra, satisfying that $\mathbb{H} \subseteq \mathbb{G} \subseteq \mathbb{H} \vee \mathbb{B}$. Then it holds that*

$$\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H} \quad \Rightarrow \quad \mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{G}\,.$$

Proof. Take $A \in \mathbb{A}$. By repeated conditioning we have that

$$\begin{aligned} P(A \mid \mathbb{B} \vee \mathbb{G}) &= E\Big(P(A \mid \mathbb{B} \vee \mathbb{G} \vee \mathbb{H}) \mid \mathbb{B} \vee \mathbb{G}\Big) = E\Big(P(A \mid \mathbb{B} \vee \mathbb{H}) \mid \mathbb{B} \vee \mathbb{G}\Big) \\ &= E\Big(P(A \mid \mathbb{H}) \mid \mathbb{B} \vee \mathbb{G}\Big) = P(A \mid \mathbb{H}) \end{aligned}$$

as $P(A \mid \mathbb{H})$ is itself $\mathbb{H}$-measurable, and thus $\mathbb{G}$-measurable, and in particular $\mathbb{B} \vee \mathbb{G}$-measurable. But by the exact same argument we have that

$$P(A \mid \mathbb{G}) = E\Big(P(A \mid \mathbb{H} \vee \mathbb{B}) \mid \mathbb{G}\Big) = E\Big(P(A \mid \mathbb{H}) \mid \mathbb{G}\Big) = P(A \mid \mathbb{H})\,.$$

And thus in particular $P(A \mid \mathbb{B} \vee \mathbb{G}) = P(A \mid \mathbb{G})$, which establishes conditional independence of $\mathbb{A}$ and $\mathbb{B}$ given $\mathbb{G}$. □

Theorem 3.4.3. *Let $\mathbb{A}$, $\mathbb{B}$, $\mathbb{G}$ and $\mathbb{H}$ be σ-algebras. It holds that*

$$\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H} \text{ and } \mathbb{A} \perp\!\!\!\perp \mathbb{G} \mid \mathbb{B} \vee \mathbb{H} \quad \Rightarrow \quad \mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{G}) \mid \mathbb{H}\,.$$

Proof. Take $A \in \mathbb{A}$. It holds that

$$P(A \mid (\mathbb{B} \vee \mathbb{G}) \vee \mathbb{H}) = P(A \mid \mathbb{B} \vee \mathbb{H}) = P(A \mid \mathbb{H})\,.$$

The first equality is due to conditional independence of $\mathbb{A}$ and $\mathbb{G}$ given $\mathbb{B} \vee \mathbb{H}$, the second is due to conditional independence of $\mathbb{A}$ and $\mathbb{B}$ given $\mathbb{H}$. The combination of course gives that $\mathbb{A}$ and $\mathbb{B} \vee \mathbb{G}$ are independent given $\mathbb{H}$. □

Example 3.4.4. None of the theorems so far will tell us how to throw information away in the conditioning algebra, while retaining conditional independence. But the theorems can in certain situations be combined to that effect.

Suppose that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{G} \vee \mathbb{H}$. Theorem 3.4.3 tells us that if we furthermore know that

$$\mathbb{A} \perp\!\!\!\perp \mathbb{G} \mid \mathbb{H}$$

then $\mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{G}) \mid \mathbb{H}$. But we can throw events away in the classes that are conditionally independent for free, so it actually follows that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$. By symmetry, we can also get rid of $\mathbb{G}$ if we know it is conditionally independent of $\mathbb{B}$ given $\mathbb{H}$. ∘

3.5 Conditionally independent random variables

In many cases we have σ-algebras generated by random variables. We will make no distinction between the random variable X and the σ-algebra $\sigma(X)$ generated by X, and we will write things like

$$X \perp\!\!\!\perp Y \mid Z \qquad \text{instead of} \quad \sigma(X) \perp\!\!\!\perp \sigma(Y) \mid \sigma(Z)$$

without notification.

Example 3.5.1. Assume that the random variables X, Y and Z are independent. then of course the corresponding σ–algebras $\sigma(X)$, $\sigma(Y)$ and $\sigma(Z)$ are independent, such that Example 3.3.4 gives

$$\sigma(X) \perp\!\!\!\perp \sigma(Y) \mid \sigma(Z)$$

which corresponds to saying

$$X \perp\!\!\!\perp Y \mid Z$$

∘

Example 3.5.2. Consider a normal distribution in three dimensions, where the one–dimensional marginals are standard normals,

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \beta \\ \rho & 1 & \beta \\ \beta & \beta & 1 \end{pmatrix} \right). \tag{3.9}$$

Here we have taken the two correlations involving Z to be identical, to keep the problem simple.

Independence of X and Y is controlled be the parameter ρ. If $\rho = 0$ they are independent, if $\rho > 0$ they are positively correlated and if $\rho < 0$ they are negatively correlated.

The conditional distribution of X and Y given $Z = z$ is

$$\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \beta \\ \beta \end{pmatrix}(z-0)\,,\ \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \begin{pmatrix} \beta \\ \beta \end{pmatrix}(\beta\,\beta)\right)$$
$$= \mathcal{N}\left(\begin{pmatrix} \beta\, z \\ \beta\, z \end{pmatrix}, \begin{pmatrix} 1-\beta^2 & \rho-\beta^2 \\ \rho-\beta^2 & 1-\beta^2 \end{pmatrix}\right).$$

Note that the variance does not depend on the specific value of z. Hence we can conclude that X and Y are conditionally independent given Z if

$$\rho - \beta^2 = 0\,.$$

More precisely, the sign of $\rho - \beta^2$ controls the direction of the conditional correlation between X and Y given Z.

In figure 3.1 we have illustrated this phenomenon. In the (ρ, β)-plane we have found the domain which corresponds to legal covariance-matrices (all three eigenvalues being non-negative). It is seen that this domain is divided into three: a part which corresponds to negative marginal correlation **and** negative conditional correlation between X and Y. A part which corresponds to positive marginal correlation but negative conditional correlation. And a third part which corresponds to positive marginal and conditional correlation. If we did not employ the restriction that the two Z-correlations should be equal, we could of course have a fourth domain, corresponding to negative marginal but positive conditional correlation.

In this context, the message is that marginal correlations and conditional correlations are two very different things, and in particular that marginal independence and conditional independence are unrelated phenomena. ∘

Theorem 3.5.3. *Let the conditional distributions of Y and Z given X be respectively $(P_x)_{x\in\mathcal{X}}$ and $(Q_x)_{x\in\mathcal{X}}$. Define*

$$R_x = P_x \otimes Q_x\ , L_{x,z} = P_x\,.$$

If $Y \perp\!\!\!\perp Z \mid X$, then $(R_x)_{x\in\mathcal{X}}$ is the conditional distribution of (Y, Z) given X, and $(L_{x,z})_{(x,z)\in\mathcal{X}\times\mathcal{Z}}$ is the conditional distribution of Y given (X, Z).

Proof. It is easily checked that $(R_x)_{x\in\mathcal{X}}$ is a $\mathcal{X}$-kernel on $\mathcal{Y}\times\mathcal{Z}$. To check the integral

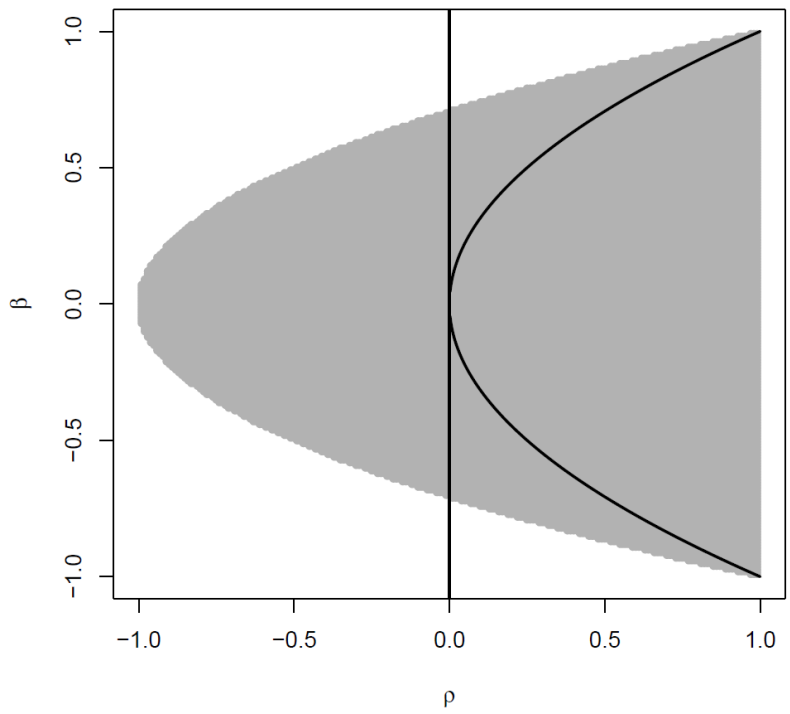

Figure 3.1: *Marginal independence and conditional independence in normal distributions of type (3.9). The shaded area contains the (ρ, β)-values for which the normal distribution exists. The vertical line corresponds to marginal independence of X and Y (positive correlation is on the right hand side). The parabolic curve corresponds to conditional independence of X and Y given Z (positive conditional correlation is in the interior of the parabola). Note the domain where there is positive marginal correlation but negative conditional correlation.*

condition, we write

$$\begin{aligned}\int_A R_x(B \times C)\, X(P)(dx) &= \int_A P_x(B) Q_x(C)\, X(P)(dx) \\ &= \int_{(X \in A)} P(Y \in B \mid X) P(Z \in C \mid X)\, dP \\ &= \int_{(X \in A)} P(Y \in B, Z \in C \mid X)\, dP \\ &= P(X \in A, Y \in B, Z \in C)\end{aligned}$$

In the second equality above we use that the function $x \mapsto P_x(B)$ is a version of the conditional probability $x \mapsto P(B \mid X = x)$, and similarly $Q_x(C)$ can be replaced by $P(C \mid X = x)$. The

third equality is an application of conditional independence, and the fourth is simply the definition of conditional probabilities. Standard arguments extend these computations from product sets $B \times C$ to general measurable subsets of $\mathcal{Y} \times \mathcal{Z}$. Hence $(R_x)_{x \in \mathcal{X}}$ **is** the conditional distribution of (Y, Z) given X.

The proof of the second half of the theorem proceeds in exactly the same way, utilising the asymmetric formulation of conditional independence instead of the definition:

$$\begin{aligned}\int_{A \times C} L_{x,z}(B)\,(X,Z)(P)(dx,dz) &= \int_{A \times C} P_x(B)\,(X,Z)(P)(dx,dz) \\ &= \int_{(X \in A, Z \in C)} P(Y \in B \mid X)\,dP \\ &= \int_{(X \in A, Z \in C)} P(Y \in B \mid X, Z)\,dP \\ &= P(X \in A, Y \in B, Z \in C)\end{aligned}$$

□

There are converses of both halves of this theorem. We choose to formulate them separately. They may come in handy under various circumstances, but in general we will go to quite some length in order to circumvent any use of them.

Theorem 3.5.4. *Suppose that the conditional distribution $(R_x)_{x \in \mathcal{X}}$ of (Y, Z) given X has product structure of the form*

$$R_x = P_x \otimes Q_x \qquad \text{for all } \; x \in \mathcal{X}$$

for two families $(P_x)_{x \in \mathcal{X}}$ and $(Q_x)_{x \in \mathcal{X}}$ of probability measures on $\mathcal{Y}$ and $\mathcal{Z}$ respectively. Then both these families are Markov kernels, they are the conditional distributions of Y given X and of Z given X respectively, and it holds that $Y \perp\!\!\!\perp Z \mid X$.

Proof. The first two statements are trivially checked. The statement about conditional independence follows from

$$\begin{aligned}P(Y \in B, Z \in C \mid X = x) &= R_x(B \times C) \\ &= P_x(B) Q_x(C) \\ &= P(Y \in B \mid X = x) P(Z \in C \mid X = x)\end{aligned}$$

for all $x \in \mathcal{X}$. Such that also

$$P(Y \in B, Z \in C \mid X) = P(Y \in B \mid X)P(Z \in C \mid X)$$

Note: this is not just a.s., but for all $\omega \in \Omega$! □

Theorem 3.5.5. *Suppose that the conditional distribution $(L_{x,z})_{(x,z)\in\mathcal{X}\times\mathcal{Z}}$ of Y given (X, Z) has the structure*

$$L_{x,z} = P_x \qquad \text{for all } x \in \mathcal{X}, z \in \mathcal{Z}$$

for some family $(P_x)_{x\in\mathcal{X}}$ of probability measures on $\mathcal{Y}$. Then this family is a Markov kernel, it is the conditional distribution of Y given X, and it holds that $Y \perp\!\!\!\perp Z \mid X$.

Proof. The first statement is trivially checked. The second statement follows from Theorem 2.1.4. The statement about conditional independence follows from the asymmetric characterization, since

$$P(Y \in B \mid X = x) = P_x(B) = L_{x,z}(B) = P(Y \in B \mid X = x, Z = z)$$

for all x and z, such that

$$P(Y \in B \mid X) = P(Y \in B \mid X, Z)\,.$$

□

Note: Apparently the fact that $Y \perp\!\!\!\perp Z \mid X$ is a statement about the joint distribution of (X, Y, Z). Then three other variables X', Y' and Z' that have the same joint distribution, will satisfy the same conditional independence relation as the original triple.

Finally comes a rather deep result, that indeed will be useful when trying to understand Markov chains

Theorem 3.5.6. *Let X and Y be random variables with values in $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. There exists a map $\phi : \mathcal{X} \times (0, 1) \to \mathcal{Y}$, which is $\mathbb{E} \otimes \mathbb{B}_{(0,1)} - \mathbb{K}$ measurable, with the following property: if X' is a random variable with the same distribution as X, U is a real valued random variable, independent of X and uniformly distributed on $(0, 1)$, and if we let*

$$Y' = \phi(X', U)$$

then (X', Y') has the same distribution as (X, Y).

Proof. Due to the underlying assumption that the spaces involved are Borel spaces, we may assume that $(\mathcal{Y}, \mathbb{K}) = (\mathbb{R}, \mathbb{B})$. Let $(P_x)_{x\in\mathcal{X}}$ be the conditional distribution of Y given X.

We know that the conditional distribution of U given X is very degenerate:

$$Q_x = \nu \qquad \text{for all } x \in \mathcal{X}\,,$$

where ν is the uniform distribution on $(0, 1)$. By the substitution theorem, the conditional distribution of Y' given X' is

$$R_x = \phi \circ i_x(Q_x) = \phi \circ i_x(\nu)\,.$$

The proof is complete, once we show how to choose ϕ such that $R_x = P_x$ for every x, as the joint distribution is uniquely determined from one marginal distribution and the conditional distribution of the remaining marginal given the first.

The deep claim is not so much that it is possible to choose ϕ is such a way that

$$\phi \circ i_x(\nu) = P_x \qquad \text{for all } x \in \mathcal{X}\,. \tag{3.10}$$

For if we let F_x be the distribution function corresponding to P_x, and if we let q_x be a quantile function for F_x, it is well known that $q_x(\nu) = P_x$. So we may let

$$\phi(x, u) = q_x(u)\,,$$

and (3.10) will be satisfied bona fide.

What **is** a deep claim is that the construction can be carried out in a way that guarantees ϕ to be measurable. There is a choice involved, in the sense that quantile functions are not unique, and even though the individual quantile functions are increasing, and thus necessarily measurable, the various choices may destroy joint measurability.

The key is to get rid of the choices, and find an operationally defined quantile function. A nice one is

$$q_x(p) = \inf\{y \in \mathbb{R} \mid F_x(y) > p\} \qquad \text{for all } x \in \mathcal{X}, p \in (0, 1)\,.$$

The idea is to single out the largest possible p-quantile whenever there is a choice. Let us prove that this is in fact a quantile function:

For fixed x and p, we have that

$$\{y \in \mathbb{R} \mid F_x(y) > p\} = \begin{cases} (y_0, \infty) \\ [y_0, \infty) \end{cases}\,,$$

for some $y_0 \in \mathbb{R}$. Whether we have the open or the halfclosed interval, depends on the specifics of the situation, but in both cases we see that $q_x(p) = y_0$. For each n we have that $y_0 + \frac{1}{n} > y_0$, and thus

$$F_x\left(y_0 + \frac{1}{n}\right) > p\,.$$

Using right continuity of F_x, we can conclude that

$$F_x(y_0) \geq p\ .$$

Similarly, $y_0 - \frac{1}{n} < y_0$, and so

$$F_x\left(y_0 - \frac{1}{n}\right) \leq p\,.$$

Using monotonicity of F_x, we can conclude that

$$F_x(y_0-) \leq p\ .$$

Together these inequalities show that y_0 is a p-quantile for F_x.

As for measurability, an elementary argument shows that

$$\{(x,p) \mid q_x(p) < z\} = \bigcup_{w<z, w\in\mathbb{Q}} \{(x,p) \mid F_x(w) > p\}\,. \tag{3.11}$$

For any fixed w, the map

$$x \mapsto F_x(w) = P_x\Big((-\infty, w]\Big)$$

is measurable, as $(P_x)_{x\in\mathcal{X}}$ is a Markov kernel. Hence $(x,p) \mapsto (F_x(w), p)$ is measurable, and thus

$$\{(x,p) \mid F_x(w) > p\} = \{(x,p) \mid F_x(w) - p > 0\}$$

is measurable set. The fact that the right hand side of (3.11) is a countable union, shows that the left hand side is a measurable set. □

The point of theorem 3.5.6 is that we may think of as any pair of variables as generated in a two-step procedure, where the generation of the second variable can be accomplished by mixing the first variable with random noise. It is the way that the mixing is carried out, that determines the joint distribution.

The **update function** ϕ is not at all unique. There are literally uncountably many ways to choose it. In certain cases it matters which one we use, in most cases it is irrelevant. However, in typical applications there is a specific update function that almost forces itself upon us.

3.6 Exercises

Exercise 3.1. Let Y and Z be real valued random variables such that $EY^2 < \infty$ and $EZ^2 < \infty$. Let X be a random variable with values in the measurable space $(\mathcal{X}, \mathbb{E})$. Define the **conditional covariance** between Y and Z given X by

$$Cov(Y, Z \mid X) = E(YZ \mid X) - E(Y \mid X)E(Z \mid X)$$

(1) Show that

$$Cov(Y, Z) = E(Cov(Y, Z \mid X)) + Cov(E(Y \mid X), E(Z \mid X))$$

(2) Assume that $Y \perp\!\!\!\perp Z \mid X$. Show that $Cov(Y, Z \mid X) = 0$ a.s. (Hint).

Now assume that X is a real valued random variable with $EX^2 < \infty$, and assume that Y_1 and Y_2 are two other random variables with the same conditional distribution $(P_x)_{x\in\mathbb{R}}$ given X, where

$$P_x = \mathcal{N}(x, 1)$$

Assume that $Y_1 \perp\!\!\!\perp Y_2 \mid X$.

(3) Show that $EY_1^2 = EY_2^2 < \infty$ (Hint).

(4) Show that $Cov(Y_1, Y_2) = V(X)$ (Hint).

∘

Exercise 3.2. Assume that X_1 and X_2 are real valued random variables. Let $(P_x)_{x\in\mathbb{R}}$ be the conditional distribution of X_1 given $X_1 + X_2$.

Define for each $x \in \mathbb{R}$ and $B \in \mathbb{B}$ the set

$$x - B = \{x - y : y \in B\}$$

and define the collection of measures $(Q_x)_{x\in\mathbb{R}}$ by

$$Q_x(B) = P_x(x - B)$$

(1) Show that $(Q_x)_{x\in\mathbb{R}}$ is the conditional distribution of X_2 given $X_1 + X_2$.

Define for each $x \in \mathbb{R}$ the measure R_x on $(\mathbb{R}^2, \mathbb{B}^2)$ by

$$R_x(A \times B) = P_x(A \cap (x - B))$$

for $A, B \in \mathbb{B}$.

(2) Show that $(R_x)_{x\in\mathbb{R}}$ is the conditional distribution of (X_1, X_2) given $X_1 + X_2$ (Hint).

(3) Assume that $X_1 \perp\!\!\!\perp X_2 \mid X_1 + X_2$. Show that for all $x \in \mathbb{R}$ it holds that $P_x(A) \in \{0, 1\}$ for all $A \in \mathbb{B}$. Conclude that $P_x = \delta_{\phi(x)}$, where $\phi(x)$ is some real number dependent on x (Hint).

(4) Show that the function ϕ from (3) is measurable (Hint).

(5) Show that if $X_1 \perp\!\!\!\perp X_2 \mid X_1 + X_2$, then there exists measurable functions ϕ_1 and ϕ_2, such that

$$X_1 = \phi_1(X_1 + X_2) \quad \text{a.s.,} \quad \text{and} \quad X_2 = \phi_2(X_1 + X_2) \quad \text{a.s.}$$

(Hint).

(6) Give an example of real random variables X_1, X_2 and X_3, where

$$X_1 \perp\!\!\!\perp X_2 \mid X_1 + X_2 + X_3\,.$$

(Hint).

∘

Exercise 3.3. In this exercise we shall find an update function (as in theorem 3.5.6) in the situation, where both $\mathcal{X}$ and $\mathcal{Y}$ are finite.

(1) Assume that $p_1, \ldots, p_n \in (0, 1)$ with $p_1 + \cdots + p_n = 1$. Define for $k = 1, \ldots, n$

$$q_k = p_1 + \ldots + p_k$$

Let U have the uniform distribution on $(0, 1)$. Define the random variable Y by

$$Y = \sum_{k=1}^{n} 1_{(Y \le q_k)}$$

Show that $P(Y = k) = p_k$ for each $k = 1, \ldots, n$.

(2) Assume that X is a random variable with values in $\{1, \ldots, m\}$ and that Y is a random variable with values in $\{1, \ldots, n\}$. Find a measurable function

$$\phi : \{1, \ldots, m\} \times (0, 1) \to \{1, \ldots, n\}$$

such that if X' has the same distribution as X and if U is uniform on $(0, 1)$ and independent of X', then (X', Y') has the same distribution as (X, Y), where $Y' = \phi(X', U)$ (Hint).

$\circ$

Exercise 3.4. Assume that X is a real random variable and that $(P_x)_{x \in \mathbb{R}}$ is the conditional distribution of Y given X, where P_x is the exponential distribution with mean x.

Find a measurable function

$$\phi : \mathbb{R} \times (0, 1) \to \mathbb{R}$$

such that if X' has the same distribution as X and if U is uniform on $(0, 1)$ and independent of X', then (X', Y') has the same distribution as (X, Y), where $Y' = \phi(X', U)$ (Hint). $\circ$

Chapter 4

Markov chains

Also in this chapter we will work on a general probability space $(\Omega, \mathbb{F}, P)$ and all events occurring will be assumed to be $\mathbb{F}$–measurable and other σ–algebras will be assumed to be sub σ–algebras of $\mathbb{F}$. Furthermore all random variables (typically sequences $X_0, X_1, X_2, \ldots$) will be defined on this probability space and usually have values in the Borel space $(\mathcal{X}, \mathbb{E})$, unless it is explicitly stated that they a real valued.

4.1 The fundamental Markov property

Definition 4.1.1. *A sequence $X_0, X_1, X_2, \ldots$ of random variables with values in a common space $(\mathcal{X}, \mathbb{E})$ is a* **Markov chain** *if*

$$X_{n+1} \perp\!\!\!\perp (X_0, X_1, \ldots, X_{n-1}) \mid X_n \qquad \text{for } n = 1, 2, \ldots \tag{4.1}$$

We refer to (4.1) as the **fundamental Markov property**. In colloquial terms, we say that the immediate future - represented by X_{n+1} - is independent of the entire past given the present.

For a Markov chain $X_0, X_1, \ldots$ the one-step conditional distributions are of paramount importance as we shall see from the following immediate result

Theorem 4.1.2. *Assume that $X_0, X_1, \ldots$ is a Markov chain with values in $(\mathcal{X}, \mathbb{E})$. Let for each $n \in \mathbb{N}$ the Markov kernel $(P_{n,x})_{x\in\mathcal{X}}$ be the conditional distributions of X_{n+1} given X_n. Then the Markov kernel given by*

$$(x_0, x_1, \ldots, x_n) \mapsto P_{n,x_n}$$

is in fact the conditional distribution of X_{n+1} given $(X_0, X_1, \ldots, X_n)$.

Proof. This follows directly from Theorem 3.5.3 and the conditional independence from the definition of $X_0, X_1, \ldots$ being a Markov chain. □

We will call $(P_{n,x})_{x\in\mathcal{X}}$ the **one–step transition probabilities**. That the Markov kernel $(P_{n-1,x_n})_{(x_0,\ldots,x_{n-1})\in\mathcal{X}^n}$ is the conditional distribution of X_n given $(X_0, X_1, \ldots, X_{n-1})$ gives

$$\begin{aligned} &P(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) \\ &\qquad = \int_{A_0\times\ldots\times A_{n-1}} P_{n-1,x_{n-1}}(A_n)\,\mathrm{d}(X_0, X_1, \ldots, X_{n-1})(P)(x_0, x_1, \ldots, x_{n-1})\,. \end{aligned}$$

But utilising that $(P_{n-2,x})_{x\in\mathcal{X}}$ by a slight change of the index set can be considered the conditional distribution of X_{n-1} given $(X_0, X_1, \ldots, X_{n-2})$, we can by the extended Tonelli theorem write the above integral as a double integral:

$$\begin{aligned} &P(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) \\ &\quad = \int_{A_0\times\ldots\times A_{n-2}} \int_{A_{n-1}} P_{n-1,x_{n-1}}(A_n)\,\mathrm{d}P_{n-2,x_{n-2}}(x_{n-1})\,\mathrm{d}(X_0, \ldots, X_{n-2})(P)(x_0, \ldots, x_{n-2})\,. \end{aligned}$$

And of course this process can be carried on, until we have the probability expressed as a n-fold integral:

$$\begin{aligned} &P(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) \\ &\qquad = \int_{A_0} \int_{A_1} \ldots \int_{A_{n-1}} P_{n-1,x_{n-1}}(A_n)\,\mathrm{d}P_{n-2,x_{n-2}}(x_{n-1}) \ldots \mathrm{d}P_{0,x_0}(x_1)\,\mathrm{d}X_0(P)(dx_0)\,. \end{aligned}$$

In order to be slightly more specific, and avoid the indexing circus and the dots, an example of such a statement is

$$\begin{aligned} &P(X_0 \in A, X_1 \in B, X_2 \in C, X_3 \in D) \\ &\qquad = \int_A \int_B \int_C P_{2,z}(D)\,\mathrm{d}P_{1,y}(z)\,\mathrm{d}P_{0,x}(y)\,\mathrm{d}X_0(P)(x)\,. \end{aligned}$$

We can think of a Markov chain $X_0, X_1, \ldots$ as a random variable with values in the sequence space $(\mathcal{X}^\infty, \mathbb{E}^\infty)$. So far we have assumed that a Markov chain $(X_0, X_1, \ldots)$ was given and

used this to express the finite dimensional distributions of $(X_0, \ldots, X_n)$ using the distribution of X_0 and all the transition probabilities $(P_{n,x})_{x\in\mathcal{X}}$. Suppose conversely that we are given a probability measure μ on $(\mathcal{X}, \mathbb{E})$ and Markov kernels $(P_{n,x})_{x\in\mathcal{X}}$ for all $n \in \mathbb{N}$ for each $n \in \mathbb{N}$ then we want to know whether a Markov chain $(X_0, X_1, \ldots)$ exists with $X_0(P) = \mu$ and such that $(P_{n,x})_{x\in\mathcal{X}}$ is the conditional distribution of X_{n+1} given X_n. We will construct the probability measure on $(\mathcal{X}^\infty, \mathbb{E}^\infty)$ that is the distribution of this Markov chain. Firstly, we can construct a probability measure $\mathcal{P}_\mu^n$ on $(\mathcal{X}^{n+1}, \mathbb{E}^{n+1})$ by defining

$$\begin{aligned} &P_\mu^n(A_0 \times A_1 \times \cdots \times A_n) \\ &\qquad = \int_{A_0} \int_{A_1} \cdots \int_{A_{n-1}} P_{n-1,x_{n-1}}(A_n)\, \mathrm{d}P_{n-2,x_{n-2}}(x_{n-1}) \ldots \mathrm{d}P_{0,x_0}(x_1)\, \mathrm{d}\mu(x_0)\,. \end{aligned}$$

These probability measures can be "collected" to a probability on $(\mathcal{X}^\infty, \mathbb{E}^\infty)$

Theorem 4.1.3. *Given a probability measure μ on $(\mathcal{X}, \mathbb{E})$ and Markov kernels $(P_{n,x})_{x\in\mathcal{X}}$ for all $n \in \mathbb{N}$ there exists a uniquely determined probability measure $\mathcal{P}_\mu$ on $(\mathcal{X}^\infty, \mathbb{E}^\infty)$ that satisfies*

$$\mathcal{P}_\mu(B_{n+1} \times E^\infty) = \mathcal{P}_\mu^n(B_{n+1})$$

for all $B_{n+1} \in \mathbb{E}^{n+1}$. Any process $(X_0, X_1, \ldots)$ with this distribution is a Markov chain with $X_0(P) = \mu$ and $(P_{n,x})_{x\in\mathcal{X}}$ as the conditional distribution of X_{n+1} given X_n.

Proof. The existence and uniqueness of the probability measure is a direct application of Kolmogorov's consistency theorem (we shall not go into any details, but according to the consistency theorem it suffices that the measures $\mathcal{P}_\mu^n$ satisfies

$$\mathcal{P}_\mu^{n+1}(B_{n+1} \times E) = \mathcal{P}_\mu^n(B_{n+1})$$

for all $n \in \mathbb{N}_0$ and all $B_{n+1} \in \mathbb{E}^{n+1}$, which is easily seen to be true.)

Suppose that $(X_0, X_1, \ldots)$ has distribution $\mathcal{P}_\mu$ such that $(X_1, \ldots, X_n)$ has distribution $\mathcal{P}_\mu^n$. Then doing the calculations from above backwards shows that

$$\begin{aligned} &P(X_0 \in A_0, X_1 \in A_1, \ldots, X_{n+1} \in A_{n+1}) \\ &\qquad = \int_{A_0 \times \ldots \times A_n} P_{n,x_n}(A_{n+1})\, \mathrm{d}(X_0, X_1, \ldots, X_n)(P)(x_0, x_1, \ldots, x_n)\,. \end{aligned}$$

which by definition gives that $(P_{n,x_n})_{(x_1,\ldots,x_n)\in\mathcal{X}}$ is the conditional distribution of X_{n+1} given $(X_0, \ldots, X_n)$. Since this only depends on x_n, we conclude from Theorem 3.5.5 that

$$X_{n+1} \perp\!\!\!\perp (X_0, \ldots, X_{n-1}) \mid X_n$$

and that $(P_{n,x})_{x\in\mathcal{X}}$ as the conditional distribution of X_{n+1} given X_n. □

Recall that it is always possible to find an underlying probability space $(\Omega, \mathbb{F}, P)$ and a random variable defined on Ω, such that this variable has a given probability measure as its distribution!!

Example 4.1.4. Assume that $\mathcal{X}$ is finite – for convenience assume that $\mathcal{X} = \{1, \ldots, m\}$ – and let $X_0, X_1, \ldots$ be a Markov chain on $\mathcal{X}$. Assume furthermore that the transition probabilities $(P_x)_{x \in \mathcal{X}}$ are independent of n. Then $X_0, X_1, \ldots$ is a so-called time homogeneous Markov chain (this concept will be discussed in a later section) on the discrete state space $\mathcal{X}$. We have that $(P_x)_{x \in \mathcal{X}}$ is determined by the point probabilities

$$p_{ij} = P(X_{n+1} = j \mid X_n = i)$$

with $i, j \in \{1, \ldots, m\}$. We shall call this collection of probabilities (p_{ij}) the **transition matrix** for the Markov chain, and write it as

$$\hat{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}.$$

In this example it is also possible to express the probability measure P_μ^n, since it is completely determined by the one–point probabilities:

$$\begin{aligned} \mathcal{P}_\mu^n(\{x_0, x_1, \ldots, x_n\}) &= P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) \\ &= \mu(\{x_0\}) \prod_{k=1}^{n} p_{x_{k-1}, x_k} \end{aligned}$$

∘

We have learned how to find the finite-dimensional distributions of a Markov chain through multiple integrals involving the one-step transition kernels. Believe it or not, this horrible characterisation is usually taken as the definition of a Markov chain!

It seems plausible to most people that this property generalises certain facts about Markov Chains on a discrete space. But nobody has the slightest clue on how to check if it is satisfied for a concrete Markov chain. The literature abounds with statements that this or that collection of random variables form a Markov chain, but there is never a proof - the Markov property is taken as selfevident, even when it clearly is not. The problem is that no one will even know where to start, if they have to check that the finite-dimensional marginal distributions have an integral representation of the specified form.... It is way too

complicated to be checkable in any practical sense. And hence the common conspiracy in the literature: if everybody keeps quite, nobody will notice the problem.

As we shall see, definition 4.1.1 can in fact be checked in a number of non-trivial situations, and so it represents a definite progress - we do not have to rely on divine insight when we claim processes to be Markovian.

Theorem 4.1.5. *If $X_0, X_1, X_2, \ldots$ is a Markov Chain, it holds that*

$$(X_n, X_{n+1}, \ldots) \perp\!\!\!\perp (X_0, X_2, \ldots, X_n) \mid X_n \qquad \text{for all } n = 1, 2, \ldots$$

Proof. We show by induction on k that

$$(X_n, X_{n+1}, \ldots, X_{n+k}) \perp\!\!\!\perp (X_0, X_1, \ldots, X_n) \mid X_n \tag{4.2}$$

As the algebra

$$\bigcup_{k=1}^{\infty} \sigma(X_n, \ldots, X_{n+k})$$

is a generator for $\sigma(X_n, X_{n+1}, \ldots)$, stable under intersections, the extension of the result from the 'finite horizon future' to the 'infinite horizon future' follows from lemma 3.3.2.

To show (4.2) we observe that the statement for $k = 1$ is the very definition of the Markov chain (and for $k = 0$ it is downright triviality).

We know that

$$X_{n+k+1} \perp\!\!\!\perp (X_0, \ldots, X_n, X_{n+1}, \ldots, X_{n+k}) \mid X_{n+k}\,.$$

By shifting information to the right hand σ–algebra to the conditioning σ–algebra, we obtain that

$$X_{n+k+1} \perp\!\!\!\perp (X_0, \ldots, X_n, X_{n+1}, \ldots, X_{n+k}) \mid (X_n, \ldots X_{n+k})\,.$$

If we by induction assume that the property (4.2) is true for k, we have that combine via theorem 3.4.3 to obtain that

$$(X_n, X_{n+1}, \ldots, X_{n+k}, X_{n+k+1}) \perp\!\!\!\perp (X_0, X_1, \ldots, X_n) \mid X_n\,.$$

□

We usually refer to theorem 4.1.5 as the **general Markov property** - or simply as the Markov property. Colloquially speaking, the σ-algebra generated by $(X_n, X_{n+1}, \ldots)$ represents 'the future', and so the Markov property says that the future is independent of the past, given the present. What we have just proved is that if the immediate future only depends upon the past via the present at all times, then the general future will also depend upon the past via the present. Variations of the theme is clearly possible, for instance that

$$(X_0, X_1, \ldots, X_m) \perp\!\!\!\perp (X_n, X_{n+1}, \ldots) \mid (X_m, \ldots, X_n)\,.$$

whenever $m < n$. This follows from shifting information around as we just did, followed by a reduction.

A formulation of the Markov property that is sometimes useful, and in fact by some authors is taken as the definition of a Markov Chain, is the following: if $X_0, X_1, \ldots$ is a Markov chain, and if $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$ is a bounded, measurable function, then for any n it holds that

$$E\Big(f(X_n, X_{n+1}, \ldots) \mid X_0, X_1, \ldots, X_n\Big) = E\Big(f(X_n, X_{n+1}, \ldots) \mid X_n\Big) \qquad \text{a.s}$$

This follows from combining theorem 4.1.5 and corollary 3.3.8. It is a nice property to have, and it is very flexible to work with. Used on functions like

$$(x_1, x_2, \ldots) \mapsto 1_B(x_2)$$

it gives the fundamental Markovian property as a consequence. But considered as a definition, it has the same basic flaw as the definition via multiple integrals: nobody has a clue on how to check if it is satisfied in concrete examples.

Theorem 4.1.6. *Let $Y_1, Y_2, \ldots$ be independent variables, which we for simplicity assume have values in the same space $(\mathcal{Y}, \mathbb{K})$. Furthermore, let $\phi_n : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ be a sequence of measurable maps.*

Let X_1 be yet another variable, independent of the Y's, and define

$$X_n = \phi_n(X_{n-1}, Y_n) \qquad \textit{for} \quad n = 1, 2, \ldots \tag{4.3}$$

The process $X_0, X_1, \ldots$ is a Markov chain.

Proof. Due to independence, we have that

$$Y_{n+1} \perp\!\!\!\perp (X_0, Y_1, \ldots, Y_n)\,.$$

Which we might formulate as

$$Y_{n+1} \perp\!\!\!\perp (X_0, Y_1, \ldots, Y_n) \mid \{\emptyset, \Omega\}\,.$$

As X_n is deterministically given by $(X_0, Y_1, \ldots, Y_n)$, it is of course measurable with respect to the σ-algebra generated by these variables. And hence we may float it to the conditioning side,

$$Y_{n+1} \perp\!\!\!\perp (X_0, Y_1, \ldots, Y_{n-1}) \mid X_n \,.$$

From there it may float back to the leftmost algebra, giving

$$(X_n, Y_{n+1}) \perp\!\!\!\perp (X_0, Y_1, \ldots, Y_n) \mid X_n \,.$$

Now, X_{n+1} is (X_n, Y_{n+1})-measurable, and $X_0, X_1, \ldots, X_n$ are all $(X_0, Y_1, \ldots, Y_n)$-measurable. So by diminishing, we obtain that

$$X_{n+1} \perp\!\!\!\perp (X_0, X_1, \ldots, X_{n-1}) \mid X_n$$

as desired. □

We usually refer to (4.3) as an **update scheme** for the Markov proces, and we refer to the Y-process as the **underlying error variables** or noise variables.

Theorem 4.1.7. *Let $X_0, X_1, \ldots$ be a Markov chain. There are update functions*

$$\phi_n : \mathcal{X} \times (0,1) \to \mathcal{X} \qquad \textit{for all } n = 1, 2, \ldots$$

with the following property: if $U_1, U_2, \ldots$ are a sequence of independent standard uniformly distributed stochastic variables, and if X_1' is independent of the U's, and has the same distribution as X_1, then the update scheme

$$X_n' = \phi_n(X_{n-1}', U_n) \qquad \textit{for } n = 1, 2, \ldots$$

produces a proces $X_0', X_1', X_2', \ldots$ with the same distribution as the original proces $X_0, X_1, X_2, \ldots$.

Proof. We may represent the original chain by its initial distribution (the distribution of X_0) and each of its one-step transition kernels $(P_{n,x})_{x \in \mathcal{X}}$. From these building blocks, we can build up the finite dimensional distributions of the process, and hence the joint distribution of all the entire process.

Each of the one-step transition kernels has an update function ϕ_n according to theorem 3.5.6. Using these in the update scheme above will produce a Markov chain with the same onestep transition kernels and then same initial distribution as the original chain, and hence the same overall distribution. □

So from at distributional point of view, we may always assume that a Markov chains is given by an update scheme - if a specific process, we happen to study, is not in update form, we can replace it by another process which is in update form, and which is indistinguishable from the first from a probabilistic point of view. The caveat is that the update functions are not in any way unique, and it may not be easy to produce update functions that make any sense intuitively.

The representations of Markov chains via update schemes is necessary for simulation purposes: a computer program that simulates a Markov chain must almost inevitably have form of an update scheme. But the idea also has a number of purely probabilistic applications.

Example 4.1.8. We will show in an exercise how a discrete state space Markov chains can be constructed using an update scheme. ◦

Example 4.1.9. The **random walk**, based on an iid. innovation sequence $X_1, X_2, \ldots$, is by definition the stochastic process $S_0, S_1, S_2, \ldots$ given by

$$S_n = \sum_{i=1}^{n} X_i \,,$$

with the convention that $S_0 = 0$. This is a Markov chain with update scheme

$$S_n = S_{n-1} + X_n \,.$$

It is typically assumed that the innovations have mean zero, but random walks with positive (or negative) drift (meaning that the innovations have a non–zero mean) are study objects in their own right. ◦

Example 4.1.10. The **reflecting random walk**, based on an iid. innovation sequence $X_1, X_2, \ldots$, is by definition the Markov chain with update scheme

$$T_0 = 0 \,, \quad T_n = (T_{n-1} + X_n)^+ \,.$$

A random walk with negative drift is frequently studied through the corresponding reflecting random walk, which exhibits the 'upwards excursions' of the random walk ◦

Example 4.1.11. The classical AR(1)-process on the real axis is given by the update scheme

$$X_{n+1} = \rho X_n + \epsilon_{n+1}$$

where the ϵ's are independent and identically distributed. As a first choice, the errors are typically normally distributed with mean zero. But other choices are clearly possible. We

also have to specify the distribution of X_0 in order to specify the joint distribution of the process.

In a sense, the behaviour of the AR(1)-process is not very dependent on the specific choice of error distribution or initial distribution. The key is the magnitude of ρ. If $|\rho| < 1$, the process will behave in a stable and quite predictable way. If $|\rho| > 1$ the process will on the other hand explode. If $\rho = 1$ we are back in the random walk case. And if $\rho = -1$, we are essential also back in the random walk case, even though it becomes slightly more complicated to formulate the results. We will return to this classification time and time again during the course. ∘

Example 4.1.12. There is a straight forward generalisation of the AR(1) process to $\mathbb{R}^k$ via the update scheme

$$X_n = RX_{n-1} + \epsilon_n$$

Here R is a $k \times k$ matrix, and the ϵ's are an iid. sequence of $\mathbb{R}^k$-valued stochastic variables - a typical choice is to make the errors $\mathcal{N}(0, \Sigma)$-distributed, where Σ is some legal variance matrix.

It is rather complicated to describe the long time behaviour of the chain, but at a first description it will depend on the eigenvalues of R. If all the eigenvalues are smaller than one in modulus, the matrix represents a linear map that contracts everything to 0. And this contraction in so dominating, that it even governs the stochastic behaviour. If some of the eigenvalues are outside of the complex unit circle, things become more complicate. The corresponding eigen–directions will be 'directions of explosion', and they will in a sense govern the stochastic behaviour, unless the error distribution is so singular, that the process will never have a non-zero component in an exploding direction.

Hence there is a very delicate interplay between the deterministic behaviour of the underlying linear map, and the measure–theoretic singularities of the error distribution. At first sight it would seem like a mathematical game to explore this interplay - it does not seem to be relevant from a modelling point of view. But it actually pops up in many places, and the problem must be considered seriously, ∘

Example 4.1.13. The AR(2)-process on the real axis is given by the update scheme

$$X_{n+1} = \alpha X_n + \beta X_{n-1} + \epsilon_{n+1}$$

where the ϵ's are independent and identically distributed, typically normal with mean zero. As it stands, this update scheme does **not** give rise to a Markov process, because it does

not just depend on the present observation, but also a **lagged** observation. Furthermore, we need both X_0 and X_1 in order to be able to run the update mechanism.

But a slight rearrangement will in fact give a Markov chain. If we **stack** the process, and consider the process in $\mathbb{R}^2$ given by

$$Y_n = \begin{pmatrix} X_n \\ X_{n-1} \end{pmatrix},$$

we see that the Y-proces fits into the update scheme

$$\begin{aligned} Y_n &= \begin{pmatrix} \alpha X_{n-1} + \beta X_{n-2} + \epsilon_n \\ X_{n-1} \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_{n-1} \\ X_{n-2} \end{pmatrix} + \begin{pmatrix} \epsilon_n \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \alpha & \beta \\ 1 & 0 \end{pmatrix} Y_{n-1} + \begin{pmatrix} \epsilon_n \\ 0 \end{pmatrix} \end{aligned}$$

This shows that the AR(2)-process is just an AR(1)-process in disguise, and hence it is 'practically Markovian'. The price we pay for this simplification is however, that the errors in the AR(1) updating scheme are quite degenerate - they are essentially one-dimensional. This perhaps sheds some light on the remarks as to why it is necessary to study AR(1)-process in full generality, even with singular error distributions. ∘

Example 4.1.14. Consider independent, identically distributed non-negative real random variables $Y_1, Y_2, \ldots$, and think of them as representing **waiting times** between events. The occurrence of the n'th event is thus happening at time

$$S_n = \sum_{i=1}^{n} Y_i \,.$$

The corresponding **renewal process** is the continuous time process, which for each time point indicates how many events that have occurred,

$$N_t = \sup\{n \,:\, S_n \leq t\}$$

Renewal processes are very important in many branches of probability, in particular in Markov chain theory.

We will mainly be interested in the case where all the waiting times are integers, and this we assume from now on. Hence the natural discrete time renewal process is

$$N_n = \sup\{k \,:\, S_k \leq n\} \qquad \text{for } n = 0, 1, 2, \ldots$$

Note that the renewal process itself is **not** Markovian. Not in general, at least. If we consider the case where

$$P(Y_i = 2) = P(Y_i = 3) = \frac{1}{2} \qquad \text{for } \; i = 1, 2, \ldots,$$

it is clear that exactly one event has taken place at time 3, that is $N_3 = 1$. This makes the σ-algebra generated by N_3 trivial, and so the Markov property is ruled out, if we show that N_2 and N_4 are not independent. But the joint distribution of these two variables is given by the table

	$N_4 = 1$	$N_4 = 2$
$N_2 = 0$	$\frac{1}{2}$	0
$N_2 = 1$	$\frac{1}{4}$	$\frac{1}{4}$

and this table does not give independence. If we think about it for at moment, the lack of Markovianess of renewal processes is rather evident: N_{n+1} will either be equal to N_n or equal to $N_n + 1$, the latter case corresponding to an event occuring at time $n + 1$. When we try to predict whether an event will occure at time $n + 1$, the relevant knowledge is not how many events that have occurred at time n, but rather the exact time of the last event - an information hidden deeper in the past.

But there are other processes, associated to the renewal proces, that do posses the Markov property. One such proces is the **forward recurrence time chain**, $V_1, V_2, \ldots$ given by

$$V_n = \inf\{S_k - n \, : \, k \text{ such that } S_k > n\}$$

For any timepoint n, the value of V_n is the waiting time until the next event. If $V_n \geq 2$, there is no event taking place at time $n + 1$, and so $V_{n+1} = V_n - 1$. But if $V_n = 1$, there is an event taking place at time $n + 1$, and the value of V_{n+1} will be the length of the waiting period until the next event. Hence it is very easy to calculate the one-step probabilities:

$$\tilde{P} = \begin{pmatrix} \nu_1 & \nu_2 & \nu_3 & \ldots \\ 1 & 0 & 0 & \ldots \\ 0 & 1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where $\nu_1, \nu_2, \ldots$ form the point masses of the waiting time distribution ν. But the relevance of the one-step probabilities are not clear, unless we know that the forward recurrence time chain is a Markov chain. And while that is true, a rigorous demonstration is not trivial. In a later example we will establish Markovianess as a consequence of the so-called strong Markov property for the underlying random walk.

A related process is the **backward recurrence time chain** , $B_0, B_1, \ldots$ given by

$$B_n = \inf\{n - S_k \,:\, S_k \leq n\} \qquad \text{for } n = 0, 1, 2, \ldots$$

For any time point n, the value of B_n is the time that has occurred since the last event. If an event is taking place at time n, the value of B_n is 0. Otherwise, we have the simple relation $B_n = B_{n-1} + 1$. Also in this case it is easy to compute the one-step probabilities,

$$\tilde{P} = \begin{pmatrix} \mu_0 & 1 & 0 & 0 & \ldots \\ \mu_1 & 0 & 1 & 0 & \ldots \\ \mu_2 & 0 & 0 & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where

$$\mu_k = P(Y_1 = k + 1 \mid Y_1 > k)\,.$$

The relevance of this matrix, though, will only become clear once it is established that the backward recurrence time chain is Markovian. ∘

Example 4.1.15. If $X_0, X_1, \ldots$ is a Markov chain, and if $f : \mathcal{X} \to \mathbb{Y}$ is a measurable function, we may consider the process $Y_0, Y_1, \ldots$ given by

$$Y_n = f(X_n) \qquad \text{for } n = 0, 1, 2, \ldots$$

It is an important problem to find out if the Y-process is Markovian as well. While reduction easily gives that

$$Y_{n+1} \perp\!\!\!\perp (Y_0, \ldots, Y_{n-1}) \mid X_n\,,$$

there is no telling when we can shrink the conditioning algebra from $\sigma(X_n)$ to $\sigma(Y_n)$. The prominence of this problem arises, of course, from the fact that Y-process is usually **not** Markovian. It is actually rather difficult to find examples where the Markov property is preserved, but non-Markovianess is usually a mess to establish.

To construct an explicit example, we may let the X-process be an asymmetric random walk on three points, say with one-step transition matrix

$$\tilde{P} = \begin{pmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{pmatrix}$$

The initial distribution can be taken as the equidistribution. This process is a random movement on the corners of a triangle. When $p \neq \frac{1}{2}$, the process has a preoccupation for

steps with a specific orientation. If p is close to one, the X-process will move $1 \mapsto 2 \mapsto 3 \mapsto 1 \mapsto 2 \mapsto \ldots$, if p is close to 0 the X-process will move the other way around. As selftransitions are not possible, the variable (X_0, X_1, X_2) has only twelve non-zero pointmasses,

$$\begin{array}{cccc|cccc} 1 & 2 & 1 & p(1-p)/3 & 2 & 3 & 1 & p^2/3 \\ 1 & 2 & 3 & p^2/3 & 2 & 3 & 2 & p(1-p)/3 \\ 1 & 3 & 1 & p(1-p)/3 & 3 & 1 & 2 & p^2/3 \\ 1 & 3 & 2 & (1-p)^2/3 & 3 & 1 & 3 & p(1-p)/3 \\ 2 & 1 & 2 & p(1-p)/3 & 3 & 2 & 1 & (1-p)^2/3 \\ 2 & 1 & 3 & (1-p)^2/3 & 3 & 2 & 3 & p(1-p)/3 \end{array}$$

The transformation we will consider is $f : \{1,2,3\} \to \{1,2\}$ given by

$$f(1) = 1\,, f(2) = 2\,, f(3) = 2\,.$$

So the Y-process is identical to the X-process, except for the fact that the original states 2 and 3 are collapsed into one superstate, which for simplicity is called 2. The variable (Y_0, Y_1, Y_2) has five point masses (as 2-2 transitions are now perfectly legal, while 1-1 transitions are still forbidden),

$$\begin{array}{ccc|c} 1 & 2 & 1 & 2p(1-p)/3 \\ 1 & 2 & 2 & p^2/3 + (1-p)^2/3 \\ 2 & 1 & 2 & p^2/3 + 2p(1-p)/3 + (1-p)^2/3 \\ 2 & 2 & 1 & p^2/3 + (1-p)^2/3 \\ 2 & 2 & 2 & 2p(1-p)/3 \end{array}$$

If we stratify this probability table by Y_1, we get

$Y_1 = 1$

	$Y_2 = 1$	$Y_2 = 2$
$Y_0 = 1$	0	0
$Y_0 = 2$	0	$p^2/3 + 2p(1-p)/3 + (1-p)^2/3$

and

$Y_1 = 2$

	$Y_2 = 1$	$Y_2 = 2$
$Y_0 = 1$	$2p(1-p)/3$	$p^2/3 + (1-p)^2/3$
$Y_0 = 2$	$p^2/3 + (1-p)^2/3$	$2p(1-p)/3$

There is actually independence in the $Y_1 = 1$ table, even if it is of a somewhat degenerate form. But there is no independence in the $Y_1 = 2$ table, unless $p = \frac{1}{2}$. ∘

4.2 The strong Markov property

The Markov property formulates a relationship between the past, the present and the future, which is to hold for **all** values of 'the present', if a process is to be called a Markov chain. At least it has to hold for all deterministic values. But it turns out time and time again, that we need the Markov property to hold in extended situations, where the value of 'the present' is not known in advance, but has a certain stochastic element to it. As an example, we may consider 'the present' to be the first time, the process enters a certain subset of $\mathcal{X}$.

To introduce the relevant formalism, we focus on a fixed process $X_0, X_1, \ldots$ with values in some measurable space $(\mathcal{X}, \mathbb{E})$. The process may or may not be a Markov chain, presently that is not relevant. The process generates a **filtration**, a sequence of σ-algebras $\mathbb{F}_0 \subseteq \mathbb{F}_1 \subseteq \ldots$ given by

$$\mathbb{F}_n = \sigma(X_0, X_1, \ldots) \qquad \text{for } n = 0, 1, 2, \ldots$$

There is also a natural limit algebra $\mathbb{F}_\infty$, generated by all the variables $X_0, X_1, \ldots$ or - if we like - generated by the filtration. It may happen that $\mathbb{F}_\infty$ equals the fundamental σ-algebra $\mathbb{F}$, but typically this is not the case. All the random variables are of course **adapted** to the filtration, meaning that X_n is $\mathbb{F}_n$-measurable for each n.

A random variable τ with values in the countable set $\mathbb{N}_0^* = \{0, 1, 2, \ldots, \infty\}$ is called a **random time**. A random time is a **stopping time** with respect to the filtration $\mathbb{F}_0 \subseteq \mathbb{F}_1 \subseteq \ldots$ if it satisfies that

$$(\tau = n) \in \mathbb{F}_n \qquad \text{for } \ n = 0, 1, 2, \ldots$$

The stopping time condition means that for each n there is a measurable subset $B_n \subseteq \mathcal{X}^{n+1}$ such that

$$(\tau = n) = \Big((X_0, X_1, \ldots, X_n) \in B_n \Big) .$$

The implication is that we are able to read off from the values of $X_0, X_1, \ldots, X_n$ whether $\tau = n$ or not. By observing the X-process for some time, we know if τ has occurred or not.

Example 4.2.1. The most obvious example of a stopping time is a determinstic time. The 'random variable' $\tau = n$ satisfies the necessary condition, as is easily checked.

The second obvious example is the **first hitting time** of a set A, as in

$$\tau = \inf\{n = 0, 1, \ldots \,:\, X_n \in A\}$$

Observing that $\tau \wedge \sigma$ and $\tau \vee \sigma$ (minimum and maximum) for two stopping times τ and σ are themselves stopping times, we can construct a vast number of new stopping times. A typical

construction would be $\tau \wedge n$ for a fixed n. A similar construction would be the first hitting time for a set A **after** a given stopping time σ. As in

$$\tau = \inf\{\sigma + 1, \sigma + 2, \dots : X_{\sigma+j} \in A\}\,.$$

In Markov chain theory it is customary to discuss both the first hitting time of a set A, typically denoted by σ_A, and the **first return time** of A, typically denoted by τ_A which is the first hitting time after time 0. Unless the chain starts in A, the first hitting time and the first return time to A agree. But when there is a difference, the first return time is usually the most relevant. ∘

It is in principle allowed that a stopping time τ can obtain the value ∞. In martingale theory this is not only a sensible convention, but in fact a useful idea, that vastly simplifyes a number of formulations. But in Markov chain theory, infinite stopping times are a menace, and we will usually not allow them. We will focus on three types of stopping times: The **bounded** stopping times, which never take on values above a certain threshold known to us, The **finite** stopping times, which never take on the value ∞, but may take on arbitrarily large integral values. And the **almost surely finite** stopping times, which satisfy that

$$P(\tau < \infty) = 1\,.$$

We would really like all our stopping times to be finite - but that would exclude the first hitting times from considerations. Consider the waiting times until head comes up in a coin tossing experiment. With probability one, head comes up sooner or later. But there is a formal possibility that head never comes up, and we have to deal with this possibility in our formalism. We could cut the corresponding nullset out of the background probability space Ω, to ensure that head always comes up. But if we follow that route, we will have to do this kind of surgery on the background space whenever we introduce a new stopping time, and it becomes technically very unpleasant in the long run. It is much neater to allow that stopping times take on the value ∞ - as long as we sure this only happens on a nullset.

If $X_0, X_1, \dots$ is a process and τ is a corresponding stopping time, we introduce the symbol X_τ as the value of the proces at the random time τ. If τ is finite, the formal definition may be written as

$$X_\tau = \sum_{n=0}^{\infty} 1_{(\tau=n)}\, X_n\,.$$

From a strict point of view, this formula only makes sense if $\mathcal{X}$ is a vector space. But even in the general case it is a much more distinct way of expressing the definition than the case-by-case formula it covers.

But to a certain extend, the definition of X_τ breaks down if the stopping time can take on the value ∞ - even if this only happens with probability zero. In order to do something in that situation, we adopt the convention that whenever we introduce a new measurable space $(\mathcal{X}, \mathbb{E})$ on which stochastic variables may have values, we equip it with a standard variable X^* - on $\mathbb{R}^n$ we could let the standard variable have the deterministic value 0. We will assume that this standard variable is measurable with respect to $\mathbb{F}_\infty$ but no other details matter. Having introduced a standard variable, we may then define

$$X_\tau = \sum_{n=0}^{\infty} 1_{(\tau=n)} X_n + 1_{(\tau=\infty)} X^* .$$

If τ assumes the value ∞ with positive probability, the choice of standard variable is of course important for the behaviour of X_τ. But as long as τ is almost surely finite, the invention of the standard variable is a purely formal gimmick. Observe that X_τ becomes measurable with respect to $\mathbb{F}_\infty$:

$$\begin{aligned}(X_\tau \in A) &= \bigcup_{n=0}^{\infty} (X_\tau \in A) \cap (\tau = n) \quad \cup \quad (X_\tau \in A) \cap (\tau = \infty) \\ &= \bigcup_{n=0}^{\infty} (X_n \in A) \cap (\tau = n) \quad \cup \quad (X^* \in A) \cap (\tau = \infty)\end{aligned}$$

The only event in this composition that does not obviously satisfy the relevant measurability condition is $(\tau = \infty)$. But the complement $(\tau < \infty)$ is the union of event of the form $(\tau = n)$, and this establishes measurability.

Corresponding to a stopping time τ, we have a natural notion of 'the past', namely the σ-algebra

$$\mathbb{F}_\tau = \{F \in \mathbb{F} \mid F \cap (\tau = n) \in \mathbb{F}_n \text{ for all } n = 0, 1, \ldots\} .$$

Lemma 4.2.2. *Let $X_0, X_1, \ldots$ be a stochastic process, and let τ be an adapted stopping time. Then the variables τ and X_τ are both $\mathbb{F}_\tau$-measurable.*

Proof. Trivial manipulations. If we consider the event $(\tau = k)$, we have that

$$(\tau = k) \cap (\tau = n) = \begin{cases} (\tau = n) & \text{if } \ k = n \\ \emptyset & \text{if } \ k \neq n \end{cases}$$

In both cases we get that $(\tau = k) \cap (\tau = n) \in \mathbb{F}_n$. And hence $(\tau = k) \in \mathbb{F}_\tau$. This shows the measurability of τ.

Similarly, if we let A be a measurable subset of $\mathcal{X}$, we have that

$$(X_\tau \in A) \cap (\tau = n) = (X_n \in A) \cap (\tau = n) \in \mathbb{F}_n ,$$

so $(X_\tau \in A) \in \mathbb{F}_\tau$. □

Lemma 4.2.3. *Let $X_0, X_1, \ldots$ be a stochastic proces, and let τ and σ be two adapted stopping times. It holds that*

$$\sigma \leq \tau \qquad \Rightarrow \qquad \mathbb{F}_\sigma \subseteq \mathbb{F}_\tau \, .$$

Proof. This is well known. □

Lemma 4.2.4. *Let $X_0, X_1, \ldots$ be a Markov chain, with corresponding filtration $\mathbb{F}_0 \subseteq \mathbb{F}_1 \subseteq \ldots$. If Z and W are two bounded real variables, both $\mathbb{F}_n$-measurable, then it holds that*

$$E(Z \mid X_n) = E(W \mid X_n) \quad a.s. \quad \Rightarrow \quad E(Z \mid X_n, X_{n+1}) = E(W \mid X_n, X_{n+1}) \quad a.s. \, .$$

Proof. This is really a trivial consequence of the Markov property. The future variable X_{n+1} is independent of the past algebra $\mathbb{F}_n$, in particular of Z and W, given the present variable X_n. Referring to the asymmetric formulation of conditional independence in corollary 3.3.8, we obtain the string of equations

$$E(Z \mid X_n, X_{n+1}) = E(Z \mid X_n) = E(W \mid X_n) = E(W \mid X_n, X_{n+1}) \quad \text{a.s.}$$

□

Note the amusing fact that we are somehow using the Markov property backwards in this proof. The argument can be verbalised as saying that when we are attempting to 'predict the past', there is no information in knowing the future - only the present matters.

Lemma 4.2.5. *Let $X_0, X_1, \ldots$ be a process, with corresponding filtration $\mathbb{F}_0 \subseteq \mathbb{F}_1 \subseteq \ldots$, and let τ be an adapted stopping time. Let Z be a real valued random variable, measurable with respect to $\mathbb{F}_\tau$. For any $n < \infty$ it holds that $1_{(\tau=n)}Z$ is measurable with respect to $\mathbb{F}_n$.*

Proof. We simply observe that for any $B \in \mathbb{B}$ we have one of two situations, depending on whether B contains 0 or not:

$$\Big(1_{(\tau=n)}Z \in B\Big) = \begin{cases} (Z \in B) \cap (\tau = n) & 0 \notin B \\ (Z \in B) \cap (\tau = n) \quad \cup \quad (\tau \neq n) & 0 \in B \end{cases}$$

Since Z is assumed to be $\mathbb{F}_\tau$-measurable, $(Z \in B)$ will be an $\mathbb{F}_\tau$-set, and so $(Z \in B) \cap (\tau = n)$ will be an $\mathbb{F}_n$-set. Also the event $(\tau \neq n)$ is $\mathbb{F}_n$-measurable - its complement has the relevant measurability per definition. So in either case $\Big(1_{(\tau=n)} Z \in B\Big)$ is in $\mathbb{F}_n$. □

Lemma 4.2.6. *Let $X_0, X_1, \ldots$ be a process, with corresponding filtration $\mathbb{F}_0 \subseteq \mathbb{F}_1 \subseteq \ldots$, and let τ be an adapted stopping time. For any event F and any $n < \infty$ it holds that*

$$E\Big(1_{(\tau=n)} P(F \mid X_\tau, \tau) \mid X_n\Big) = P\Big(F \cap (\tau = n) \mid X_n\Big) \quad a.s \tag{4.4}$$

Proof. The claim that two conditional expectations with respect to X_n are the same, of course means that the two random variables integrate to the same thing, when integrated over $\sigma(X_n)$-events. Observe that

$$\begin{aligned}\int_{(X_n \in A)} 1_{(\tau=n)} P(F \mid X_\tau, \tau)\, dP &= \int_{(X_\tau \in A, \tau = n)} P(F \mid X_\tau, \tau)\, dP \\ &= P\Big(F \cap (X_\tau \in A, \tau = n)\Big),\end{aligned}$$

since the middle integral is oven an $\sigma(X_\tau, \tau)$-event. Similarly it holds that

$$\int_{(X_n \in A)} 1_{F \cap (\tau=n)}\, dP = P\Big(F \cap (X_n \in A) \cap (\tau = n)\Big) = P\Big(F \cap (X_\tau \in A, \tau = n)\Big).$$

□

Note that (4.4) may be formulated

$$E\Big(P(F \cap (\tau = n) \mid X_\tau, \tau) \mid X_n\Big) = P\Big(F \cap (\tau = n) \mid X_n\Big) \quad \text{a.s}$$

since $1_{(\tau=n)}$ is $\sigma(X_\tau, \tau)$-measurable. Hence we see that the statement is really about a double conditioning situation. The statement remains non-trivial, however, because the σ-algebras in question, $\mathbb{F}_n$ and $\sigma(X_\tau, \tau)$, are not nested. In fact, the statement is only true due to the specific nature of the event $F \cap (\tau = n)$ and its interplay with the two σ-algebras.

Theorem 4.2.7 (Strong Markov property)**.** *Let $X_0, X_1, \ldots$ be a Markov chain, with corresponding filtration $\mathbb{F}_0 \subseteq \mathbb{F}_1 \subseteq \ldots$. Let τ be an adapted stopping time, and assume that τ is almost surely finite. Then*

$$X_{\tau+1} \perp\!\!\!\perp \mathbb{F}_\tau \mid (\tau, X_\tau)$$

Proof. We prove that for any $F \in \mathbb{F}_\tau$ it holds that

$$P(F \mid X_\tau, X_{\tau+1}, \tau) = P(F \mid X_\tau, \tau) \quad \text{a.s} \tag{4.5}$$

which is another way of formulating the the-future-is-irrelvant-for-predicting-the-past phenomenon, we have previously encountered. The right hand side of (4.5) clearly has the measurability properties to be a version of the left hand side, so we only need to check that it has the right integrals over $\sigma(X_\tau, X_{\tau+1}, \tau)$-events. For finite n we see that

$$\int_{(\tau=n, X_\tau \in A, X_{\tau+1} \in B)} P(F \mid X_\tau, \tau)\, dP = \int_{(X_n \in A, X_{n+1} \in B)} 1_{(\tau=n)} P(F \mid X_\tau, \tau)\, dP$$

Combining the lemmas, we see that we can replace the integrand by $1_{(\tau=n)\cap F}$ to obtain

$$\begin{aligned}\int_{(\tau=n, X_\tau \in A, X_{\tau+1} \in B)} P(F \mid X_\tau, \tau)\, dP &= P\Big((X_n \in A, X_{n+1} \in B, \tau = n) \cap F\Big) \\ &= P\Big((X_\tau \in A, X_{\tau+1} \in B) \cap (\tau = n) \cap F\Big)\end{aligned}$$

It is trivially true that

$$\int_{(\tau=\infty, X_\tau \in A, X_{\tau+1} \in B)} P(F \mid X_\tau, \tau)\, dP = P\Big((X_\tau \in A, X_{\tau+1} \in B) \cap (\tau = \infty) \cap F\Big)$$

since both sides are zero, due to the assumption that τ is almost surely finite. The events of the form $(\tau = n, X_\tau \in A, X_{\tau+1} \in B)$ (including the events with $n = \infty$)) form a generator for $\sigma(X_\tau, X_{\tau+1}, \tau)$, stable under the formation of intersections. And hence it follows that

$$\int_G P(F \mid X_\tau, \tau)\, dP = P(G \cap F) \qquad \text{for all } G \in \mathbb{F}(X_\tau, X_{\tau+1}, \tau)\,.$$

That is, we have established (4.5). □

This 'strong Markov property' is slightly weaker than we would have liked. The immediate future only becomes independent of the past given the present if 'the present' includes a glance at the clock. One can construct examples which shows that in general the information of the random time cannot be dispensed of, see example 4.2.8. But in the next section we will go hunting for a condition, where all times look the same, and where there is no essential information in knowing the value of τ. In that framework we will be able to strengthen the conclusion in theorem 4.2.7 to obtain what is generally perceived as the strong Markov property in the literature.

Example 4.2.8. Consider an asymmetric random walk on three points with a direction

which oscillates back and forth. The one-step transition matrices are

$$P_{2n-1} = \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix}, \qquad P_{2n} = \begin{pmatrix} 0 & q & p \\ p & 0 & q \\ q & p & 0 \end{pmatrix}$$

where $p + q = 1$. As starting distribution, we can take the equidistribution on state 2 and 3. As stopping time we take the first hitting time of state 1. In that case $X_\tau = 1$ almost surely, and so there is no information contained in that variable.

If $X_{\tau+1} \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau$, this triviality implies that $X_{\tau+1}$ is unconditionally independent of $\mathbb{F}_\tau$. And as τ is $\mathbb{F}_\tau$-measurable, it in fact implies that $X_{\tau+1}$ is independent of τ. This is clearly false, because the conditional distribution of $X_{\tau+1}$ given τ will depend rather drastically on whether τ is odd or even - unless of course $p = \frac{1}{2}$.

The example demonstrates that we can not in all cases strengthen the general strong Markov property $X_{\tau+1} \perp\!\!\!\perp \mathbb{F}_\tau \mid (\tau, X_\tau)$ to the simpler and perhaps more attractive statement $X_{\tau+1} \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau$. ∘

4.3 Homogeneity

Virtually every single Markov chain we will consider, will have a further simplifying property called **time homogeneity**.

As it was shown in Theorem 4.1.3 the distribution of a Markov chain is given by the sequence of one-step transition probabilities and the initial distribution. The one-step probabilities is a sequence of kernels $(P_{n,x})_{x\in\mathcal{X}}$, where the n'th kernel is a version of the conditional distribution of X_{n+1} given X_n. There is a certain amount of choice involved in these kernels, and maybe it is possible to adjust these choices, so that a single Markov kernel can be used as the one-step transition kernel in every step. In that case we call the chain **time homogeneous**, and write

$$P \quad \stackrel{\mathcal{D}}{=} \quad X_{n+1} \mid X_n \quad \text{for} \quad n = 0, 1, 2, \ldots$$

Spelled out, the condition is that there is one Markov kernel that satisfies

$$P(X_n \in A, X_{n+1} \in B) = \int_A P_x(B)\, \mathrm{d}X_n(P)(x) \qquad \text{for all } A, B \text{ and } n\,.$$

Many Markov chains present themselves to us in a form, where the time homogeneity is

obvious. But if it is not obvious, time homogeneity is almost impossible to establish: there are many ways to pick the various 1-step transition kernels, and if these are not picked with the purpose of being equal, then they will surely differ.

The obvious example exhibiting the problems is the random walk, with symmetric ± 1 increments. It is a time homogenous Markov chain with transition matrix

$$p_{n\,m} = \begin{cases} \frac{1}{2} & \text{if } m = n+1 \\ \frac{1}{2} & \text{if } m = n-1 \\ 0 & \text{otherwise} \end{cases}$$

This is the obvious transition matrix that everybody will write down - before they start thinking. But if we somehow miss that, and just start calculating, there are lots of other choices. Usually we insist that the random walk starts in 0. If that is the case, the transition matrix for the time 2 to time 3 transition is only uniquely given from the states -2, 0 and 2. Similarly, the transition matrix for the time 3 to time 4 transition is only uniquely given from the states -3, -1, 1 and 3. Transitions from all other states are not determined at all. So if we pick the transition matrices one by one, it is quite unlikely that we will pick the same every time, unless we have some principle to guide us.

The reason why so many chains are blatantly time homogeneous, is because they arise via time homogeneous update schemes. That is, an update scheme of the form

$$X_{n+1} = \phi(X_n, Y_{n+1})$$

where $Y_1, Y_2, \ldots$ are independent **and identically distributed**. The error distribution does not vary with time, and the update function does not vary with time, hence the one-step transition probabilities do not vary with time.

In the opposite direction, it is also clear that if $(X_0, X_1, \ldots)$ is a Markov chain where all the one-step transitions kernels are the same, then there is an update scheme of the above sort generating the process.

The examples we gave of Markov chains with specified update schemes were all of this time homogeneous form. Actually, time-varying update schemes virtually never appears in applications. With one notable exception: **simulated annealing** which is an optimisation algorithm based on Markov chains.

However, for processes constructed on top of other processes, neither the Markov structure nor the time homogeneity may be immediately visible. We have shown that we may examine

the Markov property from first principles - but the random walk example above shows that we have to be very careful when we check for time homogeneity. We adopt the following slightly weaker definition:

Definition 4.3.1. *A Markov chain $X_0, X_1, \ldots$ is* **weakly time homogeneous** *is*

$$X_{\tau+1} \perp\!\!\!\perp \tau \mid X_\tau$$

for every adapted, almost surely finite stopping time τ.

This definition undoubtedly looks confusing. There is a nice linguistic catch in that homogeneity with this definition reflects that something is 'independent of time' in a stochastic sense. But apart from that, the definition may seem arbitrary. However, the definitions has its merits, as we shall see. And at least for Markov chains on finite spaces, it is possible to prove that weak time homogeneity implies strict homogeneity, as defined in terms of constant one-step transition kernels or constant update schemes.

In order to show an equivalent definition of weak homogeneity, we need the following technical results

Theorem 4.3.2. *Let X and Y be two random variables, with values in the same space $(\mathcal{X}, \mathbb{E})$. For any event A and any $\alpha > 0$ it holds that*

$$P\left(\Big|P(A \mid X) - P(A \mid Y)\Big| > \alpha\right) \leq \frac{16\, P(X \neq Y)}{\alpha}\,.$$

Proof. The key technical result we will have to prove is that

$$\left|\int_D P(A \mid X)\, dP - P(A \cap D)\right| \leq 2P(X \neq Y)\ , \tag{4.6}$$

for any event $D \in \sigma(X, Y)$. If we have this inequality at our disposal, we can use it on the event

$$D^+ = \Big(P(A \mid X) - P(A \mid X, Y) > \alpha\Big)$$

which is $\sigma(X, Y)$-measurable, to obtain that

$$\begin{aligned}\alpha P(D^+) &\leq \int_{D^+} P(A \mid X) - P(A \mid X, Y)\, dP = \int_{D^+} P(A \mid X)\, dP - P(A \cap D^+) \\ &\leq 2P(X \neq Y)\ .\end{aligned}$$

Using a similar argument in the other tail, we obtain that

$$P\left(\left|P(A \mid X) - P(A \mid X,Y)\right| > \alpha\right) \leq \frac{4\,P(X \neq Y)}{\alpha}\,.$$

And observing that the event

$$\left(\left|P(A \mid X) - P(A \mid Y)\right| > \alpha\right)$$

is a subset of

$$\left(\left|P(A \mid X) - P(A \mid X,Y)\right| > \frac{\alpha}{2}\right) \quad \cup \quad \left(\left|P(A \mid Y) - P(A \mid X,Y)\right| > \frac{\alpha}{2}\right)$$

the theorem is established.

To show (4.6), take $D \in \sigma(X,Y)$. We can assume that $D = ((X,Y) \in B)$ for some set $B \in \mathbb{E} \otimes \mathbb{E}$. Let

$$D^* = ((X,X) \in B)\,.$$

Clearly D^* is $\sigma(X)$-measurable, and

$$(D \triangle D^*) \subseteq (X \neq Y)\,,$$

where we have used the notation $A \triangle B = A \cup B \setminus A \cap B = A \setminus B \cup B \setminus A$. Now we have that

$$\begin{aligned}
&\left|\int_D P(A \mid X)\,dP - P(A \cap D)\right| \\
&\qquad \leq \left|\int_D P(A \mid X)\,dP - \int_{D^*} P(A \mid X)\,dP\right| + |P(A \cap D^*) - P(A \cap D)| \\
&\qquad \leq 2\,P(D \setminus D^*) + 2\,P(D^* \setminus D)
\end{aligned}$$

as desired. Here we have used that the integrand $P(A \mid X)$ is bounded by 1. □

Corollary 4.3.3. *Let $X, X_1, X_2, \ldots$ be a sequence of random variables. If*

$$P(X_n = X) \to 1 \qquad \textit{for} \;\; n \to \infty\,,$$

it holds for any event A that

$$P(A \mid X_n) \xrightarrow{P} P(A \mid X)\,.$$

Proof. It follows directly from theorem 4.3.2 that for any $\alpha > 0$,

$$P\left(\left|P(A \mid X_n) - P(A \mid X)\right| > \alpha\right) \leq \frac{16\,P(X_n \neq X)}{\alpha} \to 0 \qquad \text{for} \;\; n \to \infty\,,$$

which establishes convergence in probability. □

Now we can prove

Theorem 4.3.4. *If a Markov chain $X_0, X_1, \ldots$ satisfies that*

$$X_{\tau+1} \perp\!\!\!\perp \tau \mid X_\tau \tag{4.7}$$

for every bounded adapted stopping time τ, it is weakly time homogenous.

Proof. Let τ be an adapted and almost surely finite stopping time. Consider the event $(\tau = k)$, and let us first discuss the finite case, where $k < \infty$. Pick N so large that $k < N$, and consider the stopping time $\tau \wedge N$. Clearly $(\tau = k) = (\tau \wedge N = k)$. Using (4.7) on $\tau \wedge N$, we obtain that

$$\begin{aligned} P(\tau = k \mid X_{(\tau\wedge N)+1}, X_{\tau\wedge N}) &= P(\tau \wedge N = k \mid X_{(\tau\wedge N)+1}, X_{\tau\wedge N}) \\ &= P(\tau \wedge N = k \mid X_{\tau\wedge N}) \\ &= P(\tau = k \mid X_{\tau\wedge N}) \quad \text{a.s.} \end{aligned}$$

Letting N tend to infinity, we see that

$$P(X_{\tau\wedge N} = X_\tau) \to 1 \ , \quad P\Big((X_{(\tau\wedge N)+1}, X_{\tau\wedge N}) = (X_{\tau+1}, X_\tau)\Big) \to 1 \ ,$$

simply because $P(\tau \wedge N = \tau) = P(\tau \leq N) \to 1$ for $N \to \infty$. By corollary 4.3.3 it follows that

$$P(\tau = k \mid X_{\tau+1}, X_\tau) = P(\tau = k \mid X_\tau) \quad \text{a.s.}$$

To finish the proof, we have to consider the case $k = \infty$ as well. But as τ is almost surely finite,

$$P(\tau = \infty \mid X_{\tau+1}, X_\tau) = 0 = P(\tau = \infty \mid X_\tau) \quad \text{a.s.}$$

and we are done. □

Theorem 4.3.5. *A time homogeneous Markov chain $X_0, X_1, \ldots$ is weakly time homogeneous.*

Proof. We may assume that the Markov chain has update form,

$$X_{n+1} = \phi(X_n, U_{n+1})$$

for some fixed map ϕ, and a sequence of independent, standard uniformly distributed real stochastic variables $U_1, U_2, \ldots$. Let τ be a finite stopping time. Then

$$X_{\tau+1} = \sum_{n=0}^{\infty} 1_{(\tau=n)} X_{n+1} = \sum_{n=0}^{\infty} 1_{(\tau=n)} \phi(X_n, U_{n+1}) = \phi(X_\tau, \tilde{U}) \tag{4.8}$$

where we have introduced the variable

$$\tilde{U} = \sum_{n=0}^{\infty} 1_{(\tau=n)} U_{n+1} \,.$$

Take an event $F \in \mathbb{F}_\tau$. Then it holds that

$$P\Big((\tilde{U} \in A) \cap F\Big) = \sum_{n=0}^{\infty} P\Big((\tilde{U} \in A) \cap F \cap (\tau = n)\Big) = \sum_{n=0}^{\infty} P\Big((U_{n+1} \in A) \cap F \cap (\tau = n)\Big) \,.$$

Using that $F \cap (\tau = n)$ is $\mathbb{F}_n$-measurable, and that U_{n+1} is independent of $\mathbb{F}_n$, as this algebra is contained in $\sigma(X_0, U_1, \ldots, U_n)$, we get that

$$\begin{aligned} P\Big((\tilde{U} \in A) \cap F\Big) &= \sum_{n=0}^{\infty} P\Big(U_{n+1} \in A\Big) P\Big(F \cap (\tau = n)\Big) \\ &= P\Big(U_1 \in A\Big) \sum_{n=0}^{\infty} P\Big(F \cap (\tau = n)\Big) \\ &= P(U_1 \in A)\ P(F) \,. \end{aligned}$$

We can draw two consequences: For one thing, $\tilde{U}$ is standard uniformly distributed. But more important: we see that

$$\tilde{U} \perp\!\!\!\perp \mathbb{F}_\tau \,.$$

Observing that X_τ is $\mathbb{F}_\tau$-measurable, we may float information to the (trivial) conditioning side, and obtain that

$$\tilde{U} \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau \,.$$

We may float information back, and obtain

$$(\tilde{U}, X_\tau) \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau \,.$$

As $X_{\tau+1}$ according to (4.8) is $\sigma(X_\tau, \tilde{U})$-measurable, and ad τ is $\mathbb{F}_\tau$-measurable, it follows by reduction that

$$X_{\tau+1} \perp\!\!\!\perp \tau \mid X_\tau \,,$$

as desired. □

A nice consequence of the proof is that the conditional distribution of $X_{\tau+1}$ given X_τ is simply the same as the common conditional distribution of X_{n+1} given X_n, since it follows the same update rule with an error variable that is standard uniformly distributed.

Theorem 4.3.6 (Strong Markov property)**.** *Let $X_0, X_1, \ldots$ be a Markov chain, with corresponding filtration $\mathbb{F}_0 \subseteq \mathbb{F}_1 \subseteq \ldots$. Let τ be an adapted stopping time, and assume that τ is almost surely finite. If the chain is weakly time homogeneous, then*

$$X_{\tau+1} \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau$$

Proof. We combine weak homogeneity and theorem 4.2.7 via theorem 3.4.3, the result drops out for free. □

This version of the strong Markov property is perhaps the key property of time homogeneous Markov chains in any formulation.

Corollary 4.3.7. *Let $X_0, X_1, \ldots$ be a Markov chain, with corresponding filtration $\mathbb{F}_0 \subseteq \mathbb{F}_1 \subseteq \ldots$. Let τ be an adapted stopping time, and assume that τ is almost surely finite. If the chain is weakly time homogeneous, then*

$$(X_{\tau+1}, X_{\tau+2}, \ldots) \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau \tag{4.9}$$

Proof. We show by an induction argument that

$$(X_{\tau+1}, X_{\tau+2}, \ldots, X_{\tau+k}) \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau \tag{4.10}$$

for all values of k. The crux of the matter is that $\sigma = \tau + k$ is a stopping time. Hence

$$X_{\tau+k+1} \perp\!\!\!\perp \mathbb{F}_{\tau+k} \mid X_{\tau+k}$$

As the variables $X_\tau, X_{\tau+1}, \ldots, X_{\tau+k}$ are all $\mathbb{F}_{\tau+k}$-measurable, we can shift them to the conditiong algebra, and obtain

$$X_{\tau+k+1} \perp\!\!\!\perp \mathbb{F}_{\tau+k} \mid (X_\tau, X_{\tau+1}, \ldots, X_{\tau+k})$$

As $\tau + k \geq \tau$ we see that $\mathbb{F}_\tau \subseteq \mathbb{F}_{\tau+k}$, so by reduction it follows that

$$X_{\tau+k+1} \perp\!\!\!\perp \mathbb{F}_\tau \mid (X_\tau, X_{\tau+1}, \ldots, X_{\tau+k})$$

Combining with the inductive hypothesis (4.10) we obtain via theorem 3.4.3 that

$$(X_{\tau+1}, X_{\tau+2}, \ldots, X_{\tau+k+1}) \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau$$

as desired. □

Example 4.3.8. Let $X_0, X_1, \ldots$ be a weakly time homogeneous Markov chain with values in $\mathcal{X}$, let $x \in \mathcal{X}$ be a specified state, and let τ be an almost surely finite stopping time with the property that $X_\tau = x$ almost surely.

The obvious example where such a phenomenon occurs, is the case where $\mathcal{X}$ is finite, and where τ is the first (or the second or the k'the) hitting time for the state x. A slight amount of work will usually establish that τ is almost surely finite - though it need not always be the case, the reader is reminded of concepts like 'transience' and 'recurrence', that play important roles in the study of discrete (and as it turns out, also of general) Markov chains.

If we have such a situation where X_τ is constant, $\sigma(X_\tau)$ is a trivial algebra. Hence conditional independence given X_τ is the same as unconditional independence, and the strong Markov property thus implies that

$$X_{\tau+1}, X_{\tau+1}, \ldots \perp\!\!\!\perp \mathbb{F}_\tau \,.$$

This phenomenon is called **regeneration** - the future is independent of the past in such a time point. The importance of regeneration can not be overestimated. To a large extend, analysis of the asymptotic behaviour of Markov chains **is** analysis of regeneration.

To understand why regeneration is so important, assume that $\tau_1 < \tau_2 < \ldots$ is an increasing sequence of almost surely finite regenerations time points, for instance the first, the second, the third, ... hitting time of a specified state. Let $f : \mathcal{X} \to \mathbb{R}$ be a bounded measurable function. Then the **excursions**

$$\sum_{n=\tau_1+1}^{\tau_2} f(X_n) \,, \quad \sum_{n=\tau_2+1}^{\tau_3} f(X_n) \,, \quad \sum_{n=\tau_3+1}^{\tau_4} f(X_n) \,, \ldots$$

are independent random variables. Say, the the third is independent of the two former since it is given by the τ_3-future, while the two former are given by the τ_3-past. In most cases these variables are also identically distributed and have finite mean. And this gives the asymptotic result that

$$\frac{1}{k}\sum_{n=0}^{\tau_k} f(X_n) = \frac{1}{k}\sum_{n=1}^{\tau_1} f(X_n) + \frac{1}{k}\sum_{\ell=1}^{k-1}\sum_{n=\tau_\ell+1}^{\tau_{\ell+1}} f(X_n)$$

will converge almost surely to the mean of an individual excursion. A law of large number can be established from this observation, combined with the so called renewal theorem, which control the behaviour of the regeneration time points.

Also central limit theorems for Markov chains, and indeed laws of iterated logarithm, can be obtained from the decomposition of the chain into iid. excursions. ∘

Example 4.3.9. Let $Y_1, Y_2, \ldots$ be iid. stochastic variables with values in $\mathbb{N}$. They are interpreted as waiting times between events. The events themselves occur at times

$$S_n = \sum_{i=1}^{n} Y_i \qquad \text{for } n = 0, 1, \ldots$$

(though the event at time 0 is probably fake - it is purely conventional). We are interested in the associated forward recurrence time chain

$$V_n = \inf\{S_k - n \,:\, S_k > n\} \qquad \text{for } n = 0, 1, \ldots$$

and we intend to prove that this is a Markovian.

The key argument is an application of the strong Markov property of the underlying random walk $S_0, S_1, \ldots$, which of course is a time homogeneous Markov chain. For a given time point n we can define a random time

$$\tau = \inf\{m \in \mathbb{N} \,:\, S_m > n\}\,.$$

This is clearly a stopping with respect to the filtration generated by $S_0, S_1, \ldots$. As each waiting time Y_i is at least 1, we see that $\tau \leq n+1$ - so τ is in fact a bounded stopping time. Hence the strong Markov property shows that

$$(S_\tau, S_{\tau+1}, \ldots) \perp\!\!\!\perp \mathbb{F}_\tau \mid S_\tau\,.$$

But note that

$$V_n = S_\tau - n\,, \tag{4.11}$$

So the σ-algebra generated by V_n is the same as the σ-algebra generated by S_τ, and we have that

$$(S_\tau, S_{\tau+1}, \ldots) \perp\!\!\!\perp \mathbb{F}_\tau \mid V_n\,.$$

Hence the Markov property for the forward recurrence time chain will follow by reduction, if we can show that V_0, V_1, ... and V_{n-1} are measurable with respect to $\mathbb{F}_\tau$, and if V_{n+1} is measurable with respect to $\sigma(S_\tau, S_{\tau+1}, \ldots)$.

The past is easily dealt with. If we consider some $k = 0, 1, \ldots, n-1$, there is an associated stopping time

$$\sigma = \inf\{m \in \mathbb{N} \,:\, S_m > k\}\,,$$

Similarly to (4.11) we have that $V_k = S_\sigma$. As $\sigma \leq \tau$, it follows that $\mathbb{F}_\sigma \subseteq \mathbb{F}_\tau$, and as S_σ is $\mathbb{F}_\sigma$-measurable, it follows that V_k is $\mathbb{F}_\tau$-measurable as desired.

The future requires slightly more care. There are two cases, corresponding to whether an event occurs at time $n+1$ or not. That is corresponding to whether $S_\tau = n+1$ or $S_\tau > n+1$. In the latter case we will see the same event, looking forward from time n and $n+1$. But if there is an event at time $n+1$, this is the event we will see looking forward from time n, while we will see the next event when looking forward from time $n+1$. That event will be $S_{\tau+1}$. Combining these observations, we get

$$V_{n+1} = \begin{cases} S_\tau - (n-1) & \text{if } S_\tau > n+1 \\ S_{\tau+1} - (n+1) & \text{if } S_\tau = n+1 \end{cases}$$

It is evident from this formula, that V_{n+1} is measurable with respect to $\sigma(S_\tau, S_{\tau+1})$. And so the much more, it is measurable with respect to $\sigma(S_\tau, S_{\tau+1}, \ldots)$. ∘

4.4 An integration formula for a homogeneous Markov chain

Assume that $X_0, X_1, X_2, \ldots$ is a time homogeneous Markov chain. Let $(P_x)_{x\in\mathcal{X}}$ be the transition probability, and let μ be the distribution of X_0. Recall from the considerations before Theorem 4.1.3 that we derived an expression for the distribution of $(X_0, \ldots, X_n)$

$$\begin{aligned} &P(X_0 \in A_0, \ldots, X_n \in A_n) \\ &= \int_{A_0} \int_{A_1} \cdots \int_{A_{n-1}} P_{x_{n-1}}(A_n)\, \mathrm{d}P_{x_{n-2}}(x_{n-1}) \ldots \mathrm{d}P_{x_0}(x_1)\, \mathrm{d}X_0(P)(x_0) \end{aligned}$$

Note that this could be written as

$$\int_{A_0} \mathcal{P}^n_x(A_1 \times \cdots \times A_n)\, \mathrm{d}X_0(P)(x)\,,$$

where $\mathcal{P}^n_x$ is given by the integrals

$$\begin{aligned} &\mathcal{P}^n_x(A_1 \times \cdots \times A_n) \\ &= \int_{A_1} \cdots \int_{A_{n-1}} P_{x_{n-1}}(A_n)\, \mathrm{d}P_{x_{n-2}}(x_{n-1}) \ldots \mathrm{d}P_x(x_1) \end{aligned}$$

It is easily seen that $(\mathcal{P}^n_x)_{x\in\mathcal{X}}$ is a Markov kernel (on $(\mathcal{X}^n, \mathbb{E}^n)$). So we have

Theorem 4.4.1. *Let $X_0, X_1, X_2, \ldots$ be a time homogeneous Markov chain with one-step probabilities P. Then for all $n \in \mathbb{N}$*

$$(X_1, \ldots, X_n) \mid X_0 \quad \overset{\mathcal{D}}{=} \quad \mathcal{P}^n$$

Observe that if $X_0 \equiv x$ (similar to saying that $X_0(P) = \delta_x$) then we have that

$$P((X_1, \dots, X_n) \in B) = \int \mathcal{P}_y^n \, \mathrm{d}X_0(P)(y) = \mathcal{P}_x^n(B)\,,$$

so we can interpret $\mathcal{P}_x^n$ to be the marginal distribution of $(X_1, \dots, X_n)$ if the chain is started deterministically in x. Equivalently this can be expressed as $\mathcal{P}_x^n = \mathcal{P}_{\delta_x}^n$. This interpretation is very useful.

Since $(P_x)_{x\in\mathcal{X}}$ is the conditional distribution of X_{n+1} given X_n for all $n \in \mathbb{N}$ we almost immediately obtain

Theorem 4.4.2. *Let $X_0, X_1, X_2, \dots$ be a time homogeneous Markov chain with one-step probabilities P. Then for all $n \in \mathbb{N}$*

$$(X_{k+1}, \dots, X_{k+n}) \mid X_k \quad \stackrel{\mathcal{D}}{=} \quad \mathcal{P}^n \qquad \textit{for all } k \in \mathbb{N}$$

Proof. Simply write

$$\begin{aligned}
&P(X_k \in A_k, \dots, X_{k+n} \in A_{k+n}) \\
&= \int_{A_k} \int_{A_{k+1}} \dots \int_{A_{k+n-1}} P_{x_{k+n-1}}(A_{k+n}) \, \mathrm{d}P_{x_{k+n-2}}(x_{k+n-1}) \dots \mathrm{d}P_{x_k}(x_{k+1}) \, \mathrm{d}X_k(P)(x_k) \\
&= \int_{A_k} \mathcal{P}_x^n(A_{k+1} \times \dots \times A_{k+n}) \, \mathrm{d}X_k(P)(x)
\end{aligned}$$

which shows the result. □

4.5 The Chapmann-Kolmogorov equations

In this section, we will be discussing a number of Markov kernels on a fixed space $(\mathcal{X}, \mathbb{E})$. They will be denoted by generic symbols like P, Q and R. To be specific, when we write Q, we are talking about an $(\mathcal{X}, \mathbb{E})$-Markov kernel on $(\mathcal{X}, \mathbb{E})$, which in all its glory could be spelled out as $(Q_x)_{x\in\mathbb{X}}$.

Definition 4.5.1. *We define the* **composition** *of two Markov kernels P and Q on $\mathcal{X}$ as the new Markov kernel $P * Q$, given by*

$$(P * Q)_x(A) = \int P_y(A) \, \mathrm{d}Q_x(y) \qquad \textit{for all } A \in \mathbb{E}, x \in \mathcal{X}\ .$$

Of course it has to be checked that $P * Q$ really is a new Markov kernel, but that presents no difficulty. We could extend the definition to composition of kernels on different spaces: if Q is an $\mathcal{X}$-kernel on $\mathcal{Y}$, and P is a $\mathcal{Y}$-kernel on $\mathcal{Z}$, then $P * Q$ given by the above formula, is an $\mathcal{X}$-kernel on $\mathcal{Z}$. A certain amount of bookkeeping has to be developed in order to keep track on where the various kernels live, and we will not pursue this matter.

Lemma 4.5.2. *If P and Q are two Markov kernels on $\mathcal{X}$, then*

$$\int f(z)\,\mathrm{d}(P * Q)_x(z) = \iint f(z)\,\mathrm{d}P_y(z)\,\mathrm{d}Q_x(y)\,,$$

at least for all non-negative measurable functions $f : \mathcal{X} \to \mathbb{R}$, and all bounded, measurable functions.

Proof. An application of the extended Tonelli/Fubini theorem. Note that by definition, the integral formula is true for indicators $f(x) = 1_A(x)$. □

Lemma 4.5.3. *Let X, Y and Z be random variables with values in $(\mathcal{X}, \mathbb{E})$. Suppose*

1) Q is the conditional distribution of Y given X,

2) P is the conditional distribution of Z given Y,

3) $X \perp\!\!\!\perp Z \mid Y$.

Then $P * Q$ is the conditional distribution of Z given X.

Proof. A simple computation. Observe that the conditional independence gives that the conditional distribution of Z given (X, Y) is the Markov kernel $(x, y) \mapsto P_y$. Hence

$$\begin{aligned} P(X \in A, Z \in C) = P(X \in A, Y \in \mathcal{X}, Z \in C) &= \int_{A \times \mathcal{X}} P_y(C)\,\mathrm{d}(X, Y)(P)(x, y) \\ &= \int_A \int P_y(C)\,\mathrm{d}Q_x(y)\,\mathrm{d}X(P)(x) \\ &= \int_A (P * Q)_x(C)\,\mathrm{d}X(P)(x) \end{aligned}$$

□

Lemma 4.5.3 makes it evident that compositon of Markov kernels is not commutative. In general, $P * Q \neq Q * P$. But other simple properties hold:

Lemma 4.5.4. *Composition of Markov kernels on $(\mathcal{X}, \mathbb{E})$ is associative. That is, if P, Q and R are three Markov kernels, then*

$$\Big(P * Q\Big) * R = P * \Big(Q * R\Big)\,.$$

Proof. For any $x \in \mathcal{X}$ and $A \in \mathbb{E}$ we have

$$\begin{aligned}\Big(\Big(P * Q\Big) * R\Big)_x(A) &= \int (P * Q)_y(A)\,\mathrm{d}R_x(y) = \int\int P_z(A)\,\mathrm{d}Q_y(z)\,\mathrm{d}R_x(y)\\ &= \int P_z(A)\,\mathrm{d}(Q * R)_x(z) = \Big(P * \Big(Q * R\Big)\Big)_x(A)\,.\end{aligned}$$

□

The associativity lets us interpret long composition of Markov kernels without ambiguity - it does not really matter in which order, the compositions are carried out. In particular we can define powers of a Markov chain,

$$P^{*n} = \underbrace{P * P * \ldots * P}_{n \text{ factors}}$$

without worrying about if the compositions should be carried out from left to right or from right to left or from the middle and out. And we have immediately formulas like

$$P^{*\,(n+m)} = P^{*n} * P^{*m}\,. \tag{4.12}$$

We may even extend the notion of power to a 'power of zero', if we define $P^{*\,0}$ as the trivial Markov kernel, consisting of a onepoint measure in each point

$$P_x^{*0} = \delta_x$$

With this extension (4.12) holds for all $n, m \geq 0$.

Suppose now that $X_0, X_1, \ldots$ is a Markov chain on $(\mathcal{X}, \mathbb{E})$. Suppose it is time homogeneous in the classical sense: there is a single Markov kernel P that can act as one-step transition probability from time n to time $n+1$ for all values of n. Symbolically written:

$$P \quad \stackrel{\mathcal{D}}{=} \quad X_{n+1} \mid X_n \quad \text{for} \quad n = 0, 1, 2, \ldots$$

In this scenario we are able to write the entire transition structure of the chain in terms of composition powers of P. We obtain formulae like

$$P^{*k} \quad \stackrel{\mathcal{D}}{=} \quad X_{n+k} \mid X_n \quad \text{for} \quad n, k = 0, 1, 2, \ldots, \tag{4.13}$$

A formal proof of this statement is based on lemma 4.5.3, and proceeds via induction on k - the details are left to the reader. A combined use of (4.12) and (4.13) is prototypical in Markov chain theory, and is usually referred to as a use of the **Chapman-Kolmogorov equations**. You could say that the Chapman-Kolmogorov equations are not really equations, but a principle, that combines the power formula (4.12) (which is a trivial consequence of the associativity of composition of Markov kernels) with the interpretation of the kernels in the formula as specific conditional distributions. As in 'we can compute the conditional distribution of X_{n+k+m} given X_n by finding the conditional distribution of X_{n+k} given X_n and the conditional distribution of X_{n+k+m} give X_{n+k} and combine them via composition'.

Example 4.5.5. Let $X_0, X_1, X_2, \dots$ be a time homogeneous Markov chain with one-step transition probability P. For a fixed $k > 0$, the process $X_0, X_k, X_{2k}, X_{3k}, \dots$ is called the **k-skeleton** of the original chain. This is itself a Markov chain, as the original chain satisfies the Markov property

$$X_0, X_1, \dots, X_{nk-1} \perp\!\!\!\perp X_{nk+1}, X_{nk+2}, \dots \mid X_{nk}$$

and this can be reduced to

$$X_0, X_k, \dots, X_{(n-1)k} \perp\!\!\!\perp X_{(n+1)k} \mid X_{nk}\,.$$

Note also that the k-skeleton chain is time homogenous with one-step transition probaility $\hat{P}^k$ as follows from the Chapman-Kolmogorov equations. ∘

4.6 Stationary distributions

Recall that a process $X_0, X_1, X_2, \dots$ is called **stationary**, if

$$(X_1, X_2, X_3, \dots) \stackrel{\mathcal{D}}{=} (X_0, X_1, X_2, \dots)$$

and that we have the very useful result

Theorem 4.6.1. *A stochastic process $X_0, X_1, X_2, \dots$ is stationary if and only if*

$$(X_1, X_2, \dots, X_{n+1}) \stackrel{\mathcal{D}}{=} (X_0, X_1, \dots, X_n)$$

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

Recall that for a Markov chain $X_0, X_1, \dots$ with transition probability (P_x) and initial distribution μ, we find the marginal distribution of X_1 as

$$P(X_1 \in A) = P(X_0 \in \mathcal{X}, X_1 \in A) = \int P_x(A)\,\mathrm{d}X_0(P)(x) = \int P_x(A)\,\mathrm{d}\mu(x)$$

We have the following extremely simple condition for stationarity saying that it is only need that X_0 and X_1 has the same distribution

Theorem 4.6.2. *assume that $X_0, X_1, X_2, \ldots$ is a time homogeneous Markov chain with one-step transition probability P and initial distribution $\mu = X_0(P)$. Then the Markov chain is stationary, if*

$$\mu(A) = \int P_x(A)\,\mathrm{d}\mu(x) \tag{4.14}$$

for all $A \in \mathbb{E}$.

Proof. We can express the distribution of $(X_0, \ldots, X_n)$ as follows when using that $\mathcal{P}^n$ is the conditional distribution of $(X_1, \ldots, X_n)$ given X_0

$$P(X_0 \in A_0, \ldots, X_n \in A_n) = \int_{A_0} \mathcal{P}_x^n(A_1 \times \cdots A_n)\,\mathrm{d}X_0(P)(x)$$

Using that $\mathcal{P}^n$ is also the conditional distribution of $(X_2, \ldots, X_{n+1})$ given X_1 we obtain that the distribution of $(X_1, \ldots, X_{n+1})$ is given by

$$P((X_1 \in A_0, \ldots, X_{n+1} \in A_n) = \int_{A_0} \mathcal{P}_x^n(A_1 \times \cdots A_n)\,\mathrm{d}X_1(P)(x)$$

From this we see, that the process will be stationary, if $X_0(P) = X_1(P)$, which is similar to (4.14). That "only if" holds follows from letting $n = 1$ in Theorem 4.6.1, which gives that

$$X_0 \stackrel{\mathcal{D}}{=} X_1$$

is a necessity for stationarity. □

4.7 Exercises

Exercise 4.1. Assume that $X_0, X_1, X_2, \ldots$ is a Markov chain. Show that also $X_0, X_2, X_4, \ldots$ is a Markov chain (Hint). ∘

Exercise 4.2. Consider the one–dimensional AR(1)–process

$$X_{n+1} = \rho X_n + \epsilon_{n+1}\,,$$

where all $\epsilon_1, \epsilon_2, \ldots$ are iid with a $\mathcal{N}(0,1)$ distribution. Assume that X_0 is independent of all the ϵ's.

(1) Find the conditional distribution of X_{n+1} given X_n for all $n \in \mathbb{N}_0$

(2) Assume that $|\rho| < 1$, and assume that X_0 has a $\mathcal{N}(0, \sigma^2)$ distribution. Find σ^2 such that all X_n has the same distribution as X_0. Is this possible, if $|\rho| \leq 1$?

$\circ$

Exercise 4.3. In general a renewal process (as defined in Example 4.1.14) does not have the Markov property. In this exercise we shall see an example, where the "memoryless property" of a geometric distribution actually makes a renewal process Markovian.

Let $Z_1, Z_2, \ldots$ be independent and identically distributed Bernoulli variables with success probability $p \in (0, 1)$:

$$P(Z_n = 1) = p \qquad P(Z_n = 0)1 - p$$

Define the associated random walk

$$M_n = \sum_{k=1}^{n} Z_k$$

for $n = 0, 1, 2, \ldots$.

(1) Argue that $M_0, M_1, \ldots$ is a Markov chain.

Let $(\mathbb{F}_n)_{n \in \mathbb{N}_0}$ be the corresponding filtration:

$$\mathbb{F}_n = \sigma(M_0, \ldots, M_n)$$

Also define the random times

$$T_n = \inf\{m \in \mathbb{N}_0 \,:\, M_m \geq n\}$$

for $n = 0, 1, 2, \ldots$.

(2) Show that each T_n is a stopping time with respect to the filtration $(\mathbb{F}_n)_{n \in \mathbb{N}_0}$.

(3) Show that each T_n is almost surely finite (Hint).

(4) Argue that $T_0 < T_1 < T_2 < \cdots$ (not a deep result:-)).

Now define the "waiting times"

$$Y_n = T_n - T_{n-1}$$

for $n = 1, 2, \ldots$.

(5) Show that the sequence $Y_1, Y_2, \ldots$ are independent and identically distributed with common distribution

$$P(Y_n = k) = (1-p)^{k-1}p \quad \text{for } k = 1, 2, \ldots \tag{4.15}$$

(Hint).

(6) Define $N_0, N_1, N_2, \ldots$ to be the renewal process generated by $Y_1, Y_2, \ldots$: Let

$$S_n = \sum_{k=1}^{n} Y_k$$

and define

$$N_n = \sup\{k : S_k \leq n\}$$

Show that $N_n = M_n$ for all $n = 0, 1, 2, \ldots$ (Hint).

(7) Collect the results from (1)-(5) to a proof of the general statement: If $Y_1, Y_2, \ldots$ are independent and identically distributed with distribution as in (4.15), then the associated renewal process is a Markov chain.

∘

Exercise 4.4. Assume that $X_0, X_1, X_2, \ldots$ is a Markov chain. Assume that τ is an almost surely finite stopping time.

(1) Show that $\tau + k$ is a stopping time for each $k \in \mathbb{N}$ (Hint).

(2) Show that $\sigma(\tau) = \sigma(\tau + k)$ for each $k \in \mathbb{N}$.

(3) Show that the sequence $(X_{\tau+k}, \tau + k)_{k \in \mathbb{N}_0}$ is a Markov chain (Hint).

∘

Exercise 4.5. Assume that $X_0, X_1, X_2, \ldots$ is a time homogeneous Markov chain with values in $(\mathcal{X}, \mathbb{E})$. Assume that τ is an almost surely finite stopping time.

(1) Show that the sequence $X_\tau, X_{\tau+1}, X_{\tau+2}, \ldots$ is a Markov chain (Hint).

(2) Show that the Markov chain is time homogeneous with

$$X_{\tau+k+1} \mid X_{\tau+k} \quad \stackrel{\mathcal{D}}{=} \quad X_{k+1} \mid X_k$$

(Hint).

Now assume that $X_0 = x$ for some $x \in \mathcal{X}$. Furthermore let

$$\tau = \inf\{n \geq 1 \,:\, X_n = x\}$$

and assume that $P(\tau < \infty) = 1$.

(3) Argue that $X_0, X_1, X_2, \ldots$ and $X_\tau, X_{\tau+1}, X_{\tau+2}$ have the same distribution.

Define

$$N_x = \sum_{n=1}^{\infty} 1_{(X_n = x)}$$

(4) Show that $P(N_x = \infty) = 1$.

∘

Exercise 4.6. Assume that $(X_0^1, X_1^1, X_2^1, \ldots)$ and $(X_0^2, X_1^2, X_2^2, \ldots)$ are two independent time homogeneous Markov chains on $(\mathcal{X}, \mathbb{E})$ with the same transition probabilities $(P_x)_{x \in \mathcal{X}}$. Let

$$\tau = \inf\{n \in \mathbb{N}_0 \,:\, X_n^1 = X_n^2\}$$

and assume that $P(\tau < \infty) = 1$.

(1) Show that τ is a stopping time with respect to the (filtration generated by) the process $(X_n^1, X_n^2)_{n \in \mathbb{N}_0}$.

Define the process $X_0, X_1, X_2, \ldots$ by

$$X_n = \begin{cases} X_n^1 & n \leq \tau \\ X_n^2 & n > \tau \end{cases}$$

We can assume that the two Markov chains have the same update scheme:

$$X_{n+1}^1 = \phi(X_n^1, U_{n+1}^1)$$
$$X_{n+1}^2 = \phi(X_n^2, U_{n+1}^2)$$

where all $U_1^1, U_2^1, \ldots, U_1^2, U_2^2, \ldots$ are iid uniformly distributed. Define the new sequence $U_1, U_2, \ldots$ by

$$U_n = U_n^1 \cdot 1_{(\tau \geq n)} + U_n^2 \cdot 1_{(\tau < n)}$$

(2) Show that $U_1, U_2, \ldots$ are iid following the uniform distribution (Hint).

(3) Show that $X_0, X_1, X_2, \ldots$ is a time homogeneous Markov chain with transition probabilities $(P_x)_{x \in \mathcal{X}}$ (Hint).

$\circ$

Exercise 4.7. Assume that $X_0 X_1, X_2, \ldots$ is a time homogeneous Markov chain on $(\mathcal{X}, \mathbb{E})$. Let A be some measurable subset of $\mathcal{X}$ and define *the first hitting time* of A by

$$\tau_A = \inf\{n \in \mathbb{N}_0 \,:\, X_n \in A\}$$

Define σ_B to be the first time the process hits the set B after time τ_A

$$\sigma_B = \inf\{n > \tau_A \,:\, X_n \in B\}$$

(1) Show that

$$X_{\sigma_B} \perp\!\!\!\perp \mathbb{F}_{\tau_A} \mid X_{\tau_A}$$

(Hint).

Now let the set A be fixed and define the sequence of stopping times $\tau_1 < \tau_2 < \ldots$ by

$$\tau_1 = \tau_A$$

and recursively

$$\tau_{n+1} = \inf\{n > \tau_n \,:\, X_n \in A\}$$

(2) Show that $X_{\tau_1}, X_{\tau_2}, \ldots$ is a Markov chain.

(3) Show that $X_{\tau_1}, X_{\tau_2}, \ldots$ is time homogeneous (Hint).

$\circ$

Exercise 4.8. Assume that $X_0, X_1, X_2, \ldots$ is a time homogeneous Markov chain with transition probabilities $(P_x)_{x \in \mathcal{X}}$. Let τ be a first hitting time

$$\tau = \inf\{n \in \mathbb{N}_0 \,:\, X_n \in A\}$$

for some measurable set $A \subseteq \mathcal{X}$. Define the process $Y_0, Y_1, Y_2, \ldots$ by

$$Y_n = X_{\tau \wedge n}$$

for all $n \in \mathbb{N}_0$.

(1) Show that $Y_0, Y_1, Y_2, \ldots$ is a Markov chain (Hint).

(2) Show that $Y_0, Y_1, Y_2, \ldots$ is time homogeneous and find the transition probabilities (Hint).

$\circ$

Chapter 5

Ergodic theory for Markov chains on general state spaces

Consider a stochastic process $X_0, X_1, X_2, \ldots$ with values in some measurable space $(\mathcal{X}, \mathbb{E})$. Then we are often interested in knowing, when empirical means like

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k, X_{k+1}, X_{k+2}, \ldots)$$

converges, where $f : \mathcal{X}^\infty \to \mathbb{R}$ is some measurable function defined on the sequence space $\mathcal{X}^\infty$.

A well known simple result is the Strong Law of Large Numbers:

Theorem 5.0.1. *Assume that $X_0, X_1, X_2, \ldots$ are independent and identically distributed. Assume that $f : \mathcal{X} \to \mathbb{R}$ is a measurable function such that $E|f(X_0)| < \infty$. Then*

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \to Ef(X_0) \qquad a.s.$$

as $n \to \infty$.

A much more general result is The Ergodic Theorem. To formulate that result, we need to recall some definitions from the course VidSand1.

Let in the following $X_0, X_1, X_2, \ldots$ be a stochastic process with values in $(\mathcal{X}, \mathbb{E})$. Let $\mathcal{P}$ be the distribution of the sequence $(X_0, X_1, X_2, \ldots)$ – hence it is a probability measure on $(\mathcal{X}^\infty, \mathbb{E}^\infty)$. Let $S : \mathcal{X}^\infty \to \mathcal{X}^\infty$ be the **shift**

$$S(x_0, x_1, x_2, \ldots) = (x_1, x_2, x_3, \ldots)$$

Note that $X_0, X_1, X_2, \ldots$ is stationary, if

$$(X_0, X_1, X_2, \ldots) \quad \stackrel{\mathcal{D}}{=} \quad S(X_0, X_1, X_2, \ldots)$$

We define the **invariant σ–algebra** for S by

$$\mathcal{I} = \{A \in \mathbb{E}^\infty \, : \, S^{-1}(A) = A\}$$

Definition 5.0.2. *A stationary process $X_0, X_1, X_2, \ldots$ with distribution $\mathcal{P}$ is called **ergodic** if*

$$\mathcal{P}(A) \in \{0, 1\}$$

for all $A \in \mathcal{I}$.

For ergodic processes we have the Ergodic Theorem

Theorem 5.0.3 (Ergodic theorem)**.** *Let $X_0, X_1, X_2, \ldots$ be a stationary and ergodic process, and let $f : \mathcal{X}^\infty \to \mathbb{R}$ be a measurable map, such that*

$$E|f(X_0, X_1, X_2, \ldots)| < \infty \, .$$

Then

$$\frac{1}{n} \sum_{k=0}^{n-1} f(X_k, X_{k+1}, X_{k+2}, \ldots) \to Ef(X_0, X_1, X_2, \ldots) \qquad a.s.$$

as $n \to \infty$.

Note: The limit can be written as

$$Ef(X_0, X_1, X_2, \ldots) = \int f(x_0, x_1, x_2, \ldots) \, \mathrm{d}\mathcal{P}(x_0, x_1, x_2, \ldots)$$

which will be a useful notation later on.

A useful tool to check whether a process is ergodic is the concept of being **mixing**

Definition 5.0.4. *Let $X_0, X_1, X_2, \ldots$ be a stationary stochastic process with distribution $\mathcal{P}$. We say that S is mixing with respect to $\mathcal{P}$, if for all F and G in $\mathbb{E}^\infty$*

$$\lim_{n \to \infty} P(F \cap S^{-n}(G)) = P(F)P(G)$$

because we have the result

Theorem 5.0.5. *If S is mixing with respect to $\mathcal{P}$, then $X_0, X_1, X_2, \ldots$ is ergodic.*

The following Corollary gives a condition that ensures S being mixing

Corollary 5.0.6. *If it for all $m, k \in \mathbb{N}_0$ and all $A \in \mathbb{E}^{m+1}$ and $B \in \mathbb{E}^{k+1}$ holds that*

$$\begin{aligned}\lim_{N\to\infty} & P((X_0,\ldots,X_m) \in A, (X_N,\ldots,X_{N+k} \in B) \\ & = P((X_0,\ldots,X_m) \in A) P((X_0,\ldots,X_k) \in B)\,,\end{aligned}$$

then the shift S is mixing with respect to $\mathcal{P}$.

In this chapter we will find conditions that makes Markov chains ergodic, such that the Ergodic Theorem holds. In fact, we will obtain a much stronger result: The averages in the theorem may converge even in situations, where the process is not stationary!

5.1 Convergence of transition probabilities

We start by introducing a convergence concept for sequences of probability measures on $(\mathcal{X}, \mathbb{E})$.

Definition 5.1.1. *Let μ_0 and $\nu_0, \nu_1, \nu_2, \ldots$ be probability measures on the same measurable space $(\mathcal{X}, \mathbb{E})$. We say, that ν_n **converges to** μ_0 as $n \to \infty$ and write $\nu_n \to \mu_0$, provided*

$$\lim_{n\to\infty} \int h(y)\,\mathrm{d}\nu_n(y) = \int h(y)\,\mathrm{d}\mu_0(y) \tag{5.1}$$

for every bounded and measurable function h on $(\mathcal{X}, \mathbb{E})$.

This a very strong form of convergence, e.g. it follows for $h = 1_B$ that

$$\lim_{n\to\infty} \nu_n(B) = \mu_0(B)$$

for every $B \in \mathbb{E}$.

The standard form of weak convergence requires that (5.1) holds for all continuous and bounded functions. However, that will only make sense if there is a topology on the state space $\mathcal{X}$. But in that case the convergence above is stronger than weak convergence. If $\mathcal{X}$ is a finite space, then all bounded functions are continuous, and hence the two forms of convergence are equal.

Theorem 5.1.2. *Assume that $X_0, X_1, X_2, \ldots$ is a Markov chain with transition probabilities $(P_x)_{x\in\mathcal{X}}$. If there exists a probability measure μ_0 such that $P_x^{*n} \to \mu_0$ for every $x \in \mathcal{X}$, then μ_0 is a stationary initial distribution of $X_0, X_1, X_2, \ldots$, and it is the only stationary distribution.*

Proof. That μ_0 is a stationary distribution is seen as as follows: Let $A \in \mathbb{E}$. Then because of the Chapmann Kolmogorov equation

$$\mu_0(A) = \lim_{n\to\infty} P_y^{*(n+1)}(A) = \lim_{n\to\infty} \int P_x(A)\,\mathrm{d}P_y^{*n}(x) = \int P_x(A)\mathrm{d}\mu_0(x)$$

which due to Theorem 4.6.2 shows, that the Markov chain is stationary, if $X_0(P) = \mu_0$.

Now we show the uniqueness: Assume that μ is a stationary initial distribution. If we let $X_0(P) = \mu$ we obtain that $X_n(P) = X_0(P)$, so

$$\mu(A) = P(X_n \in A) = P(X_0 \in \mathcal{X}, X_n \in A) = \int P_x^{*n}(A)\,\mathrm{d}X_0(P)(x) = \int P_x^{*n}(A)\,\mathrm{d}\mu(x)$$

where we have used that $P^{*n} \overset{\mathcal{D}}{=} X_n \mid X_0$. But let $n \to \infty$ in the above equality. From dominated convergence (since all $0 \le P_x^{*n}(A) \le 1$) we obtain that

$$\mu(A) = \int \mu_0(A)\,\mathrm{d}\mu(x) = \mu_0(A)\,,$$

showing that $\mu = \mu_0$. □

Theorem 5.1.3. *Assume that $X_0, X_1, X_2, \ldots$ is a Markov chain with transition probabilities $(P_x)_{x\in\mathcal{X}}$, and assume that there exists a probability measure μ_0 such that $P_x^{*n} \to \mu_0$ for every $x \in \mathcal{X}$. Assume that $X_0(P) = \mu_0$ such that the Markov chain is stationary, and let $\mathcal{P}_{\mu_0}$ be the distribution of the Markov chain. Then it holds that the shift S is mixing with respect to $\mathcal{P}_{\mu_0}$, and in particular the Markov chain is ergodic.*

Proof. Assume that $X_0(P) = \mu_0$. Let $A \in \mathbb{E}^{m+1}$ and $B \in \mathbb{E}^k$ and consider

$$\begin{aligned}
&P((X_0,\ldots,X_m) \in A, (X_{N+1},\ldots,X_{N+k} \in B) \qquad (5.2)\\
=&P((X_0,\ldots,X_m) \in A, X_N \in \mathcal{X}, (X_{N+1},\ldots,X_{N+k}) \in B)\\
=&\int_{A\times\mathcal{X}} \mathcal{P}_{x_N}^k(B)\,\mathrm{d}(X_0,\ldots,X_m,X_N)(P)(x_0,\ldots,x_m,x_N)\\
=&\int_A \int \mathcal{P}_y^k(B)\,\mathrm{d}P_{x_m}^{*(N-m)}(y)\,\mathrm{d}(X_0,\ldots,X_m)(P)(x_0,\ldots,x_m)
\end{aligned}$$

In the second equality we have used that $\mathcal{P}^k \stackrel{\mathcal{D}}{=} (X_{N+1}, \ldots, X_{N+k}) \mid X_N$, and

$$(X_{N+1}, \ldots, X_{N+k}) \perp\!\!\!\perp (X_0, \ldots, X_m) \mid X_N \,.$$

In the third equality we used that $P^{*(N-m)} \stackrel{\mathcal{D}}{=} X_N \mid X_m$. For the inner integral we obtain

$$\begin{aligned}\int \mathcal{P}_y^k(B)\,\mathrm{d}P_{x_m}^{*(N-m)}(y) &\to \int \mathcal{P}_y^k(B)\,\mathrm{d}\mu_0(y) \\ &= P((X_1, \ldots, X_k) \in B) = P((X_0, \ldots, X_{k-1}) \in B)\end{aligned}$$

as $N \to \infty$. So from Dominated convergence we obtain that the probability in (5.2) has the limit

$$\begin{aligned}&\int_A P((X_0, \ldots, X_k) \in B)\mathrm{d}(X_0, \ldots, X_m)(P)(x_0, \ldots, x_m) \\ &= P((X_0, \ldots, X_m) \in A)P((X_0, \ldots, X_{k-1}) \in B)\end{aligned}$$

which according to Corollary 5.0.6 is precisely what is needed to say that S is mixing. □

5.2 Transition probabilities with densities

We will in the rest of this chapter make the assumption, that all the transition probabilities $(P_x)_{x\in\mathcal{X}}$ have densities with respect to some σ–finite measure ν on $(\mathcal{X}, \mathbb{E})$. We demand that for all $x \in \mathcal{X}$ it holds that

$$P_x(A) = \int_A k_x(y)\,\mathrm{d}\nu(y)\,,$$

where $(x, y) \mapsto k_x(y)$ is $(\mathcal{X}^2, \mathbb{E}^2) - (\mathbb{R}, \mathbb{B})$ measurable, and of course $y \mapsto k_x(y)$ is the density of P_x with respect to ν.

Then using the Chapmann Kolmogorov equation gives for $n > 1$ that

$$\begin{aligned}P_x^{*n}(A) &= (P * P^{*(n-1)})_x(A) \\ &= \int P_y(A)\,\mathrm{d}P_x^{*(n-1)}(y) \\ &= \int \left(\int_A k_y(z)\,\mathrm{d}\nu(z)\right)\mathrm{d}P_x^{*(n-1)}(y) \\ &= \int_A \left(\int k_y(z)\,\mathrm{d}P_x^{*(n-1)}dy)\right)\,\mathrm{d}\nu(z)\,,\end{aligned}$$

which shows that P_x^{*n} has density

$$k_x^n(y) = \int k_z(y)\mathrm{d}P_x^{(n-1)}(z)$$

with respect to ν. Using that $P_x^{*(n-1)}$ similarly have density $k_x^{(n-1)}$ gives

$$k_x^n(y) = \int k_x^{(n-1)}(z) k_z(y) \, \mathrm{d}\nu(z) \, .$$

Repeating the same arguments with n and k arbitrary gives

$$k_x^{(n+k)}(y) = \int k_x^{(n)}(z) \, k_z^{(k)}(y) \, \mathrm{d}\nu(z) \tag{5.3}$$

Now assume that $X_0, X_1, X_2, \ldots$ is a Markov chain with transition probabilities $(P_x)_{x \in \mathcal{X}}$ with densities as above. Assume furthermore that X_0 has distribution μ. Then the distribution of X_n is given by

$$\begin{aligned} P(X_n \in A) &= P(X_0 \in \mathcal{X}, X_n \in A) \\ &= \int P_x^{*n}(A) \, \mathrm{d}\mu(x) \\ &= \int \left(\int_A k_x^{(n)}(y) \, \mathrm{d}\nu(y) \right) \mathrm{d}\mu(x) \\ &= \int_A \left(\int k_x^{(n)}(y) \, \mathrm{d}\mu(x) \right) \mathrm{d}\nu(y) \end{aligned}$$

which shows that $X_n(P)$ has density $P^{*n}(\mu)$ with respect to ν, where

$$P^{*n}(\mu)(y) = \int k_x^{(n)}(y) \, \mathrm{d}\mu(x)$$

In the case, where μ has density f with respect to ν, we will use the notation $P^{*n}(f)$ for the density of $X_n(P)$. In that case we have

$$P^{*n}(f)(y) = \int k_x^{(n)}(y) f(x) \, \mathrm{d}\nu(x)$$

Using (5.3) gives the following version of the Chapmann Kolmogorov equation

$$\begin{aligned} P^{*(n+k)}(\mu)(y) &= \int k_x^{(n+k)}(y) \, \mathrm{d}\mu(x) \\ &= \iint k_x^{(n)}(z) \, k_z^{(k)}(y) \, \mathrm{d}\nu(z) \, \mathrm{d}\mu(x) \\ &= \int k_z^{(k)}(y) \underbrace{\left(\int k_x^{(n)}(z) \, \mathrm{d}\mu(x) \right)}_{= P^{*n}(\mu)(z)} \mathrm{d}\nu(z) \\ &= P^{*k}\Big(P^{*n}(\mu) \Big)(y) \end{aligned} \tag{5.4}$$

The equation (5.3) used again – now with $(n, k) = (n, 1)$ – gives

$$P^{*n}(k_x)(y) = \int k_z^{(n)}(y) \, k_x(z) \, \mathrm{d}\nu(z) = k_x^{(n+1)}(y) \tag{5.5}$$

5.3 Asymptotic stability

Let $L^1(\nu)$ be the vector space of all ν–integrable functions, and define

$$\mathcal{D} = \{ f \in L^1(\nu) \,:\, f \geq 0, \int f \, d\nu = 1 \}$$

to be the subset of $L^1(\nu)$ consisting of all the probability densities. On $L^1(\nu)$ we have the L^1 norm given by

$$\|f\| = \int |f| \, d\nu$$

for $f \in L^1(\nu)$.

Definition 5.3.1. *Transition probabilities $(P_x)_{x\in\mathcal{X}}$ that have densities $(k_x)_{x\in\mathcal{X}}$ with respect to ν are called **asymptotically stable** if there exists $f_0 \in \mathcal{D}$ such that*

$$\forall f \in \mathcal{D} \,:\, \lim_{n\to\infty} \|P^{*n}(f) - f_0\| = 0 \,. \tag{5.6}$$

Theorem 5.3.2. *Let $X_0, X_1, X_2, \ldots$ be a time homogeneous Markov chain with transition with transition probabilities $(P_x)_{x\in\mathcal{X}}$ that are asymptotically stable. Then the probability measure $\mu_0 = f_0 \cdot \nu$ with density f_0 with respect to ν is the only stationary initial distribution. Furthermore the shift S is mixing for the distribution $\mathcal{P}_{\mu_0}$ of the stationary chain. In particular the stationary chain is ergodic.*

Proof. Let $x \in \mathcal{X}$. Since $k_x^{(n)}(y) = P^{*(n-1)}(k_x)(y)$ according to (5.5) we have

$$\int \left| k_x^{(n)}(y) - f_0(y) \right| \, \mathrm{d}\nu(y) = \left\| P^{*(n-1)}(k_x) - f_0 \right\| \to 0$$

as $n \to \infty$, where the convergence follows from (5.6). Now let $h : \mathcal{X} \to \mathbb{R}$ be a bounded and measurable function with $|h| \leq c$. Then for all $x \in \mathcal{X}$ it holds that

$$\begin{aligned}
&\left| \int h(y) \, \mathrm{d}P_x^{*n}(y) - \int h(y) \, \mathrm{d}\mu_0(y) \right| \\
=& \left| \int h(y) \, k_x^{(n)}(y) \, \mathrm{d}\nu(y) - \int h(y) \, f_0(y) \mathrm{d}\nu(y) \right| \\
\leq& \int |h(y)| \left| k_x^{(n)}(y) - f_0(y) \right| \mathrm{d}\nu(y) \\
\leq& c \int \left| k_x^{(n)}(y) - f_0(y) \right| \mathrm{d}\nu(y) \quad \to 0
\end{aligned}$$

as $n \to \infty$. Hence we have shown that $P_x^{*n} \to \mu_0$ for all $x \in \mathcal{X}$. The theorem follows from the Theorems 5.1.2 and 5.1.3. □

Note: In the proof we only used that

$$\forall x \in \mathcal{X} \,:\, \lim_{n\to\infty} \left\| k_x^{(n)} - f_0 \right\| \to 0$$

but as we shall see in the following lemma, this is in fact equivalent to the assumption (5.6) giving asymptotic stability.

Lemma 5.3.3. *Consider the framework from Theorem 5.3.2. Let $\mathcal{P}(\mathcal{X}, \mathbb{E})$ be the set of all probability measures on $(\mathcal{X}, \mathbb{E})$. Then the following conditions (a), (b) and (c) are equivalent*

$$\begin{aligned}
&(a) \quad \forall x \in \mathcal{X} && : \quad \lim\nolimits_{n\to\infty} \left\| k_x^{(n)} - f_0 \right\| \to 0 \\
&(b) \quad \forall f \in \mathcal{D} && : \quad \lim\nolimits_{n\to\infty} \left\| P^{*n}(f) - f_0 \right\| = 0 \\
&(c) \quad \forall \mu \in \mathcal{P}(\mathcal{X}, \mathbb{E}) && : \quad \lim\nolimits_{n\to\infty} \left\| P^{*n}(\mu) - f_0 \right\| = 0
\end{aligned}$$

Proof. It is obvious that (c) $\Rightarrow$ (b) $\Rightarrow$ (a), so we only need to show that (a) $\Rightarrow$ (c). To show this, we firstly derive the following: Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$. Then

$$\begin{aligned}
\left\| P^{*n}(\mu) - f_0 \right\| &= \int \Big| \int k_x^{(n)}(y)\, \mathrm{d}\mu(x) - f_0(y) \Big| \mathrm{d}\nu(y) \\
&= \int \Big| \int k_x^{(n)}(y)\, \mu(dx) - \int f_0(y)\, \mathrm{d}\mu(x) \Big| \mathrm{d}\nu(y) \\
&\leq \int \Big(\int \left| k_x^{(n)}(y) - f_0(y) \right| \mathrm{d}\mu(x) \Big)\, \mathrm{d}\nu(y) \\
&= \int \Big(\int \left| k_x^{(n)}(y) - f_0(y) \right| \mathrm{d}\nu(y) \Big)\, \mathrm{d}\mu(x)\,,
\end{aligned}$$

and since

$$\int \left| k_x^{(n)}(y) - f_0(y) \right| \mathrm{d}\nu(y) \leq \int k_x^{(n)}(y) \mathrm{d}\nu(y) + \int f_0(y)\, \mathrm{d}\nu(y) = 2\,,$$

it follows from Dominated convergence and (a) that the double integral tends to 0 as $n \to \infty$. □

Note that the conditions (a), (b) and (c) are equivalent to $\mathcal{L}^1(\nu)$–convergence of the density of $X_n(P)$ to the density f_0, if

(a) $X_0 \equiv x$ for some $x \in \mathcal{X}$

(b) the distribution of X_0 has a density w.r.t. ν

(c) the distribution of X_0 is an arbitrary distribution

Condition (a) is the weakest of the three. Condition (b) will be the most convenient to work with, and condition (c) is the strongest implying the following result

Notation:
Recall that $\mathcal{P}_\mu$ denotes the distribution on $(\mathcal{X}^\infty, \mathbb{E}^\infty)$ of $X_0, X_1, X_2, \ldots$, when $X_0 \overset{\mathcal{D}}{=} \mu$.

We will also need a notation for the probability measure on $(\Omega, \mathbb{F})$ that gives X_0 this distribution: For a given probability measure μ on $(\mathcal{X}, \mathbb{E})$, we let P_μ denote the probability measure on $(\Omega, \mathbb{F})$, that makes $X_0 \overset{\mathcal{D}}{=} \mu$.

Theorem 5.3.4. *Assume that $X_0, X_1, X_2, \ldots$ is a time homogeneous Markov chain with transition probabilities $(P_x)_{x\in\mathcal{X}}$ that have densities with respect to ν and are asymptotically stable with $\mu_0 = f_0 \cdot \nu$ as the stationary distribution. then no matter which distribution μ that is chosen to be the initial distribution of X_0, then the following holds:*

If $f : \mathcal{X}^\infty \to \mathbb{R}$ is a measurable function, such that

$$\int |f(x_0, x_1, x_2, \ldots)| \, \mathrm{d}\mathcal{P}_{\mu_0}(x_0, x_1, x_2, \ldots) < \infty \,,$$

then

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k, X_{k+1}, X_{k+2}, \ldots) \to \int f(x_0, x_1, x_2, \ldots) \, \mathrm{d}\mathcal{P}_{\mu_0}(x_0, x_1, x_2, \ldots) \quad P_\mu\text{–a.s.}$$

as $n \to \infty$.

Proof. Recall that

$$\mathcal{P}^k = (X_{n+1}, \ldots, X_{n+k}) \mid X_n$$

for all $n \in \mathbb{N}$. Then we have for all $B_k \in \mathbb{E}^k$ that

$$\begin{aligned} P_\mu((X_{n+1}, \ldots, X_{n+k}) \in B_k) &= \int \mathcal{P}_x^k(B_k) \, \mathrm{d}X_n(P_\mu)(x) \\ &= \int \mathcal{P}_x^k(B_k) \, P^{*n}(\mu)(x) \, \mathrm{d}\nu(x) \,, \end{aligned}$$

where we have used, that X_n has density $P^{*n}(\mu)$ under P_μ. Also recall that $\mathcal{P}_x^k(B_k)$ can be considered as the probability of $P_{\delta_x}((X_1, \ldots, X_k) \in B_k)$ in a situation, where $X_0 \equiv x$ (hence the notation P_{δ_x}!!). Hence we can write

$$\mathcal{P}_x^k(B_k) = \mathcal{P}_{\delta_x}(\mathcal{X} \times B_k \times \mathcal{X}^\infty)$$

so we have

$$P_\mu((X_{n+1}, \ldots, X_{n+k}) \in B_k) = \int \mathcal{P}_{\delta_x}(\mathcal{X} \times B_k \times \mathcal{X}^\infty) \, P^{*n}(\mu)(x) \, \mathrm{d}\nu(x)$$

for all $B_k \in \mathbb{E}^k$. By standard extension arguments (since $\{B_k \times \mathcal{X}^\infty : B_k \in \mathbb{E}^k, k \in \mathbb{N}\}$ is an intersection stable generating class for $\mathbb{E}^\infty$) we have

$$P_\mu((X_{n+1}, X_{n+2}, \ldots) \in B) = \int \mathcal{P}_{\delta_x}(\mathcal{X} \times B)\, P^{*n}(\mu)(x)\, \mathrm{d}\nu(x)$$

for all $B \in \mathbb{E}^\infty$. Now define $H \in \mathbb{E}^\infty$ by

$$H = \left(\frac{1}{n} \sum_{k=0}^{n-1} f(X_k, X_{k+1}, X_{k+2}, \ldots) \to \int f(x_0, x_1, x_2, \ldots)\, \mathrm{d}\mathcal{P}_{\mu_0}(x_0, x_1, x_2, \ldots) \right)$$

Then H is in the tail σ–algebra for the sequence $X_0, X_1, X_2, \ldots$, so in fact, H has the form

$$H = \big((X_{n+1}, X_{n+2}, \ldots) \in D_n\big)$$

for every $n \in \mathbb{N}$ and some $D_n \in \mathbb{E}^\infty$. Thereby we have, that

$$\begin{aligned}
\big|P_\mu(H) - P_{\mu_0}(H)\big| &= \Big| \int \mathcal{P}_{\delta_x}(\mathcal{X} \times D_n)\, P^{*n}(\mu)(x)\, \mathrm{d}\nu(x) - \int \mathcal{P}_{\delta_x}(\mathcal{X} \times D_n)\, P^{*n}(f_0)(x)\, \mathrm{d}\nu(x) \Big| \\
&= \Big| \int \mathcal{P}_{\delta_x}(\mathcal{X} \times D_n)\, P^{*n}(\mu)(x)\, \mathrm{d}\nu(x) - \int \mathcal{P}_{\delta_x}(\mathcal{X} \times D_n)\, f_0(x)\, \mathrm{d}\nu(x) \Big| \\
&\leq \int \mathcal{P}_{\delta_x}(\mathcal{X} \times D_n) \big| P^{*n}(\mu)(x) - f_0(x) \big|\, \mathrm{d}\nu(x) \\
&\leq \int \big| P^{*n}(\mu)(x) - f_0(x) \big|\, \mathrm{d}\nu(x) \\
&= \big\| P^{*n}(\mu) - f_0 \big\| \\
&\to 0
\end{aligned}$$

as $n \to \infty$. Since we already know that $P_{\mu_0}(H) = 1$ according to Theorem 5.3.2, we must also have that $P_\mu(H) = 1$. □

The concept used above is $L^1(\nu)$–convergence of densities with respect to the σ–finite measure ν. If a sequence of densitites converges in $L^1(\nu)$–norm, there exists a subsequence converging ν–a.e. (almost everywhere). It is worth noting that suitably understood, the converse also holds

Lemma 5.3.5. *If $f, f_0, f_1, f_2, \ldots \in \mathcal{D}$ and $\lim_{n\to\infty} f_n(x) = f(x)$ ν–a.e., then $\lim_{n\to\infty} \|f_n - f\| = 0$.*

In other words, if a sequence converges pointwise a.e. there is also L^1–convergence.

Proof. Define $d_n = f - f_n$. Then $d_n^+ \leq f$ and from using dominated convergence it follows that

$$\lim_{n\to\infty} \int d_n^+ \, d\nu = \int \lim_{n\to\infty} d_n^+ \, d\nu = \int 0 \, d\nu = 0 \, .$$

Since $d_n^- = d_n^+ + f_n - f$ we have

$$\lim_{n\to\infty} \int d_n^- \, d\nu = \lim_{n\to\infty} \int d_n^+ \, d\nu + \lim_{n\to\infty} f_n \, d\nu - \int f \, d\nu = 0 + 1 - 1 = 0$$

(where it is used that f is a density) and therefore

$$\lim_{n\to\infty} \int |f - f_n| \, d\nu = \lim_{n\to\infty} \int d_n^+ \, d\nu + \lim_{n\to\infty} \int d_n^- \, d\nu = 0 + 0 = 0$$

□

Example 5.3.6. Consider the autoregressive process of order 1. Here $(\mathcal{X}, \mathbb{E}) = (\mathbb{R}, \mathbb{B})$, and for all $n \in \mathbb{N}_0$

$$X_{n+1} = \rho X_n + \epsilon_{n+1} \, ,$$

where $U_1, U_2, \ldots$ are independent and identically distributed and furthermore independent of X_0. We assume that each ϵ_n has a $\mathcal{N}(0,1)$ distribution. In an exercise it is shown that the conditional distribution $(P_x^{*n})_{x\in\mathbb{R}}$ of X_n given X_0 is given by

$$P_x^{*n} = \mathcal{N}(\rho^n x, \sigma_n^2) \, ,$$

where

$$\sigma_n^2 = \frac{1 - \rho^{2n}}{1 - \rho^2}$$

Hence the n–step transition densities are given by

$$k_x^{(n)}(y) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y - \rho^n x)^2}{2\sigma_n^2} \right)$$

If $|\rho| < 1$ we see that

$$\sigma_n^2 \to \sigma_0^2 := \frac{1}{1 - \rho^2}$$

and furthermore $\rho^n x \to 0$. So we have obtained that

$$k_x^{(n)}(y) \to \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{y^2}{2\sigma_0^2} \right)$$

for all x and y in $\mathbb{R}$. By a reference to Lemma 5.3.5 we see, that the transition probabilities are asymptotically stable with $\mathcal{N}(0, \sigma_0^2)$ as the stationary distribution. ∘

Example 5.3.7. Here we consider the ARCH(1)–process: Assume that the Markov chain $X_0, X_1, X_2, \ldots$ is given on the update form

$$X_{n+1} = \sqrt{\gamma + \alpha X_n^2}\epsilon_{n+1}\,,$$

where (again) $\epsilon_1, \epsilon_2, \ldots$ are independent and identically distributed and furthermore independent of X_0. Here we have the transition probabilities $(P_x)_{x\in\mathbb{R}}$, where

$$P_x = \mathcal{N}(0, \gamma + \alpha x^2)$$

corresponding to the densities

$$k_x(y) = \frac{1}{\sqrt{2\pi(\gamma + \alpha x^2)}} \exp\Big(-\frac{y^2}{2(\gamma + \alpha x^2)}\Big)$$

It is not possible to find the n–step transition densities and neither does one know the stationary initial distribution in the cases where it exists. We shall see later that the Markov chain is stable if (and only if) $\alpha < 3.56\ldots$. ∘

5.4 Minorisation

We still consider the setup, where the transition probabilities $(P_x)_{x\in\mathcal{X}}$ have densities $(k_x)_{x\in\mathcal{X}}$ with respect to a σ–finite measure ν.

Recall that we for a density $f \in \mathcal{D}$ found

$$P^{*n}(f)(y) = \int k_x^{(n)}(y) f(x)\,\mathrm{d}\nu(x) \tag{5.7}$$

to be the density of X_n if X_0 has density f. It makes sense to use this definition for all $f \in L^1(\nu)$ – we shall see in the following lemma, that the resulting function is well defined and that $P^{*n}(f) \in L^1(\nu)$ as well.

Lemma 5.4.1. *The definition in (5.7) defines a linear map $P^{*n} : L^1(\nu) \to L^1(\nu)$. Furthermore it holds that*

$$f \geq 0 \quad \Rightarrow \quad P^{*n}(f) \geq 0 \quad \text{and} \quad \|P^{*n}(f)\| = \|f\|\,.$$

For $f \in L^1(\nu)$ it holds that

$$\big(P^{*n}(f)\big)^+ \leq P^{*n}(f^+) \quad \text{and } \big(P^{*n}(f)\big)^- \leq P^{*n}(f^-)\,.$$

Moreover P^{*n} is a contraction

$$\|P^{*n}(f)\| \leq \|f\|\,.$$

Proof. It is obvious, that the map is linear, and that $f \geq 0$ implies that $P^{*n}f \geq 0$ (since k_x is non–negative). For $f \geq 0$ we have

$$\|P^{*n}(f)\| = \int P^{*n}(f)(y)\,\mathrm{d}\nu(y) = \int \Big(\int k_x^{(n)}(y) f(x)\,\mathrm{d}\nu(x)\Big)\mathrm{d}\nu(y)\,,$$

which using Fubini's theorem

$$= \int f(x)\Big(\int k_x^{(n)}(y)\,\mathrm{d}\nu(y)\Big)\mathrm{d}\nu(x) = \int f(x)\,\mathrm{d}\nu(x) = \|f\|\,.$$

From this we see in particular that the definition of $P^{*n}(f)$ must make sense for all $f \in L^1(\nu)$, since the integrals involved are finite when integrating $|f|$. Since the map is linear and positive (maps non–negative functions to non–negative functions), it must be increasing such that (since $f^+ \geq f$) $P^{*n}(f^+) \geq P^{*n}(f)$. But since always $P^{*n}(f^+) \geq 0$ we derive

$$P^{*n}(f^+) \geq \big(P^{*n}(f)\big)^+$$

The argument for the negative part is similar. Then

$$\begin{aligned} |P^{*n}(f)| &= \big(P^{*n}(f)\big)^+ + \big(P^{*n}(f)\big)^- \\ &\leq \big(P^{*n}(f^+)\big) + \big(P^{*n}(f^-)\big) = P^{*n}(f^+ + f^-) = P^{*n}(|f|) \end{aligned}$$

so

$$\|P^{*n}(f)\| \leq \|P^{*n}(|f|)\| = \|f\|$$

□

Definition 5.4.2. *A non–negative integrable function $h \in L^1(\nu)$ is said to be a* ***minorant*** *for the transition probabilities if $\|h\| > 0$ and there for all $f \in \mathcal{D}$ exists a sequence of non–negative functions $\epsilon_n(f)$ in $L^1(\nu)$ such that*

$$P^{*n}(f) \geq h - \epsilon_n(f) \tag{5.8}$$

and

$$\lim_{n\to\infty} \|\epsilon_n(f)\| = \int \epsilon_n(f)\,d\nu = 0 \tag{5.9}$$

Note:
If it holds that $P^{*n}(f) - h \geq 0$ from some step onward, then we have that h is a minorant, and we can simply let $\epsilon_n(f) = 0$ from this step.

The inequality (5.8) is equivalent to saying that

$$\epsilon_n(f) \geq (h - P^{*n}(f))^+ = (P^{*n}(f) - h)^-$$

so we could of course define $\epsilon_n = (P^{*n}(f) - h)^-$ and have that (5.9) is satisfied precisely when $\|(P^{*n}(f) - h)^-\| \to 0$. However the definition using ϵ_n–functions is simpler to work with. The inequality (5.8) may be written as

$$|P^{*n}(f) - h| \leq P^{*n}(f) - h + 2\epsilon_n(f)$$

so

$$\begin{aligned}\|P^{*n}(f) - h\| &\leq \int P^{*n}(f) - h + 2\epsilon_n(f)\, d\nu \\ &= 1 - \|h\| + 2\|\epsilon_n(f)\|\end{aligned}$$

Letting $n \to \infty$ on the right hand side gives

$$\limsup_{n\to\infty} \|P^{*n}(f) - h\| \leq 1 - \|h\| \tag{5.10}$$

Think of a Markov chain $X_0, X_1, X_2, \ldots$, where $X_0 \stackrel{\mathcal{D}}{=} \mu$ and where the transition probabilities $(P_x)_{x\in\mathcal{X}}$ have a minorant. Then we have the following, using $P^{*n}(\mu) = P^{*(n-1)}(P^{*1}(\mu))$, where $P^{*1}(\mu)$ is a density

$$\liminf_{n\to\infty} P_\mu(X_n \in B) = \liminf_{n\to\infty} \int_B P^{*n}(\mu)\, d\nu = \int_B P^{*(n-1)}(P^{*1}(\mu))\, d\nu \geq \int_B h\, d\nu$$

Since $\int h\, d\nu > 0$, this shows that the probability mass cannot "vanish" when $n \to \infty$, if a minorant exists. Hence (part of) the intuition of the existence of a minorant is that the distribution of the variables cannot keep changing – the probability mass is (partly) fixed by the function h, so there are limits for how much it can "move around".

Theorem 5.4.3. *The transition probabilities $(P_x)_{x\in\mathcal{X}}$ for a Markov chain are asymptotically stable if and only if there exists a minorant*

Proof. First assume that the transition probabilities are stable

$$\forall f \in \mathcal{D} \,:\, \lim_{n\to\infty} \|P^{*n}(f) - f_0\| = 0\,.$$

In particular we have for a given density $f \in \mathcal{D}$ that

$$\|(P^{*n}(f) - f_0)^-\| \leq \|P^{*n}(f) - f_0\| \to 0$$

which shows that the stationary density f_0 is a minorant.

Assume conversely, that a minorant h exists. From (5.10) it follows that $0 < \|h\| \leq 1$. It is also seen from (5.10), that it suffices to show that there exists a minorant h with $\|h\| = 1$. The idea in the proof will be to find a maximal minorant. let

$$c = \sup\{\, \|h\| \,:\, h \text{ is a minorant}\}\,.$$

Then $0 < c \leq 1$ and we can choose a sequence of minorants $h_1, h_2, \ldots$ so $\|h_m\| \to c$ for $m \to \infty$. Now we show that the maximum $h_1 \vee h_2$ is again a minorant if both h_1 and h_2 are minorants: We have

$$P^{*n}(f) \geq h_1 - \epsilon_{1,n}(f) \quad \text{and} \quad P^{*n}(f) \geq h_2 - \epsilon_{2,n}(f)$$

Then also

$$P^{*n}(f) \geq h_1 \vee h_2 - \epsilon_{1,n}(f) \vee \epsilon_{2,n}(f)\,,$$

where $\|\epsilon_{1,n}(f) \vee \epsilon_{2,n}(f)\| \leq \|\epsilon_{1,n}(f) + \epsilon_{2,n}(f)\| \leq \|\epsilon_{1,n}(f)\| + \|\epsilon_{2,n}(f)\| \to 0$. And this shows, that $h_1 \vee h_2$ is a minorant.

Then we can assume that the sequence $h_1, h_2, \ldots$ is increasing, $h_1 \leq h_2 \leq \cdots$, and hence the limit $h_0 = \lim_{m\to\infty} h_n$ is well defined. Furthermore we see from monotone convergence that

$$\|h_0\| = \int h_0 \, d\nu = \lim_{m\to\infty} \int h_m \, d\nu = \lim_{m\to\infty} \|h_m\| = c$$

and similarly we have that $\lim_{m\to\infty} \|h_0 - h_m\| = 0$. Furthermore we must have, that h_0 is a minorant as well, since

$$P^{*n}(f) - h_0 = P^{*n}(f) - h_m + h_m - h_0 \geq -\epsilon_{m,n}(f) - |h_m - h_0|\,,$$

giving $\left(P^{*n}(f) - h_0\right)^- \leq \epsilon_{m,n}(f) + |h_m - h_0|$ such that

$$\limsup_{n\to\infty} \left\|\left(P^{*n}(f) - h_0\right)\right\| \leq \|h_m - h_0\|\,.$$

And since this is true for all $m \in \mathbb{N}$, and the right hand side have limit 0, when $m \to \infty$, we must have that

$$\left\|\left(P^{*n}(f) - h_0\right)\right\| \to 0$$

as requested.

Now let h be another minorant. Then also $h \vee h_0$ is a minorant, and since

$$c \geq \int h \vee h_0 \, d\nu \geq \int h_0 \, d\nu = c$$

we conclude that $h \leq h_0$ ν–a.e.

From the inequality $P^{*n}(f) \geq h_0 - \epsilon_n(f)$ it follows (using that P^{*1} is increasing) that

$$P^{*(n+1)}(f) = P^{*n}(P^{*n}(f)) \geq P^{*1}(h_0) - P^{*1}(\epsilon_n(f))\,,$$

and because $\|P^{*1}(\epsilon_n(f))\| = \|\epsilon_n(f)\| \to 0$, we see that $P^{*1}(h_0)$ is a minorant. Consequently $h_0 \geq P^{*1}(h_0) \quad \nu$–a.e., but since

$$\int h_0 \, d\nu = \|h_0\| = \|P^{*1}(h_0)\| = \int P^{*1}(h_0) \, d\nu$$

we conclude that in fact $h_0 = P^{*1}(h_0)$ ν–a.e. Defining $f_0 = h_0/c$ such that f_0 is a probability density, we see that also $f_0 = P^{*1}(f_0)$ ν–a.s. showing that f_0 is the density for a stationary initial distribution.

Now let h be some minorant. We want to show that the Markov chain is asymptotically stable. Let therefore $f \in \mathcal{D}$ and we want to show that $\|P^{*n}(f) - f_0\| \to 0$. Define $g = f - f_0$ and assume without loss of generality that $d = \|g\|/2 > 0$ (otherwise there would be nothing to show). Since $\int g \, d\nu = 1 - 1 = 0$ we must have that $\|g^+\| = \|g^-\| = d$. Now

$$\begin{aligned}
\|P^{*n}(f) - f_0\| &= \|P^{*n}(f) - P^{*n}(f_0)\| \\
&= \|P^{*n}(g)\| \\
&= \|d(P^{*n}(g^+/d) - h) - d(P^{*n}(g^-/d) - h)\| \\
&= d\|P^{*n}(g^+/d) - h\| + d\|P^{*n}(g^-/d) - h\| \,,
\end{aligned}$$

and because g^+/d, $g^-/d \in \mathcal{D}$ it follows from (5.10) that

$$\limsup_{n\to\infty} \|P^{*n}(f) - f_0\| \leq \limsup_{n\to\infty} d\Big(\|P^{*n}(g^+/d) - h\| + \|P^{*n}(g^-/d) - h\|\Big) \tag{5.11}$$

$$= 2d(1 - \|h\|) \tag{5.12}$$

$$= \|f - f_0\|(1 - \|h\|) \,. \tag{5.13}$$

By replacing f by $P^{*m}(f)$ we obtain for each $m = 1, 2, \ldots$ that

$$\begin{aligned}
\limsup_{n\to\infty} \|P^{*n}(f) - f_0\| &= \limsup_{n\to\infty} \|P^{*(n+m)}(f) - f_0\| \\
&= \limsup_{n\to\infty} \|P^{*n}(P^{*m}(f)) - f_0\| \\
&\leq \|P^{*m}(f) - f_0\|(1 - \|h\|) \,,
\end{aligned}$$

and by letting $m \to \infty$ it follows from (5.11) that

$$\limsup_{n\to\infty} \|P^{*n}(f) - f_0\| \leq \|f - f_0\|(1 - \|h\|)^2 \,.$$

Repeating this argument gives that

$$\limsup_{n\to\infty} \|P^{*n}(f) - f_0\| \leq \|f - f_0\|(1 - \|h\|)^k \,.$$

for each $k = 1, 2, \ldots$. And since $1 - \|h\| < 1$ we obtain

$$\lim_{n\to\infty} \|P^{*n}(f) - f_0\| = 0$$

from letting $k \to \infty$. □

For deciding whether a function is a minorant or not, the following result proves useful

Lemma 5.4.4. *Let $\mathcal{D}^*$ be a dense subset of $\mathcal{D}$. Then h is a minorant if for all $f \in \mathcal{D}^*$ there exists a sequence of non–negative functions $\epsilon_n(f)$ in $L^1(\nu)$ such that (5.8) and (5.9) are satisfied.*

Proof. For $f \in \mathcal{D}$ and $f^* \in \mathcal{D}^*$ we have

$$P^{*n}(f) - h = P^{*n}(f^*) - h + P^{*n}(f) - P^{*n}(f^*) \geq -\epsilon_n(f^*) + P^{*n}(f - f^*)\,,$$

and therefore $(P^{*n}(f) - h)^- \leq \epsilon_n(f^*) + |P^{*n}(f - f^*)|$ and consequently

$$\limsup_{n\to\infty} \|(P^{*n}(f) - h)^-\| \leq \limsup_{n\to\infty} \|P^{*n}(f - f^*)\| \leq \|f - f^*\|\,.$$

Since $\inf_{f^*\in\mathcal{D}} \|f - f^*\| = 0$ it follows that

$$\lim_{n\to\infty} \|(P^{*n}(f) - h)^-\| = 0$$

□

5.5 The drift criterion

In this section we will develop a very useful necessary condition for the existence of a minorant.

Theorem 5.5.1. *Let $V : (\mathcal{X}, \mathbb{E}) \to (\mathbb{R}, \mathbb{B})$ be a non–negative measurable function such that $0 \leq \alpha < 1$ and $0 \leq \beta < \infty$ exists with*

$$\int V(y)\,\mathrm{d}P_x(y) \leq \alpha V(x) + \beta \qquad \text{for all } x \in \mathcal{X} \tag{5.14}$$

Furthermore assume that there exists $m \in \mathbb{N}$ such that for some $r > \beta/(1-\alpha)+1$ the function

$$y \mapsto \inf_{\{x\,:\,V(x)\leq r\}} k_x^{(m)}(y)$$

is measurable, and that

$$\int \inf_{\{x\,:\,V(x)\leq r\}} k_x^{(m)}(y)\,\mathrm{d}\nu(y) > 0\,. \tag{5.15}$$

Then it holds that the transition probabilities are asymptotically stable.

The function V is called a **drift function** or a **Lyapounov function**. Typically the behaviour of V is such that $V(x) \to \infty$, when x approaches the boundary of $\mathcal{X}$, so the set $\{x \in \mathcal{X} : V(x) \leq r\}$ is bounded. Usually it is quite easy to verify condition (5.15), so the critical condition is (5.14). Since $\alpha < 1$ – informally phrased – the condition states that the Markov chain has a tendency to drift towards x's with low values of $V(x)$. The condition (5.14) may become more clear if we rewrite it using random variables:

$$E(V(X_{n+1}) \mid X_n = x) \leq \alpha V(x) + \beta$$

Proof. The idea in the proof is to determine a minorant, because then it will follow from Theorem 5.4.3 that the transition probabilities are asymptotically stable. Define $B = \{x \in \mathcal{X} : V(x) \leq r\}$ and

$$h(y) = \inf_{x \in B} k_x^{(m)}(y)\,.$$

Then for $f \in \mathcal{D}$

$$\begin{aligned} P^{*(m+n)}(f)(y) &= P^{*m}\big(P^{*n}(f)\big) \\ &= \int k_x^{(m)}(y) P^{*n}(f)(x)\,\mathrm{d}\nu(x) \\ &\geq h(y) \int_B P^{*n}(f)(x)\,\mathrm{d}\nu(x)\,. \end{aligned}$$

Let $\delta = 1 - \frac{1}{r}(\frac{\beta}{1-\alpha} + 1) > 0$. We shall show that

$$\int_B P^{*n}(f)(x)\,\mathrm{d}\nu(x) > \delta \tag{5.16}$$

for n sufficiently large (depending on f). Because then $P^{*(n+m)}(f)(y) \geq \delta h(y)$ for n sufficiently large, and it will follow that δh is a minorant.

Now

$$\begin{aligned} \int P^{*n}(f)(x)\mathrm{d}\nu(x) &= P_{f\cdot\nu}\big(X_n \in B\big) \\ &= P_{f\cdot\nu}\big(V(X_n) \leq r\big) \\ &= 1 - P_{f\cdot\nu}\big(V(X_n) > r\big) \end{aligned}$$

and by Markov's inequality

$$P_{f\cdot\nu}\big(V(X_n) > r\big) \leq \frac{1}{r} E_{f\cdot\nu}\big(V(X_{n+1})\big)\,,$$

where $E_{f\cdot\nu}$ denotes integration with respect to $P_{f\cdot\nu}$. To establish (5.16) it suffices to show that $E_{f\cdot\nu}\big(V(X_n)\big) \leq \frac{\beta}{1-\alpha} + 1$ for n sufficiently large. But

$$\begin{aligned}
E_{f\cdot\nu}\big(V(X_n)\big) &= \int V(x)\mathrm{d}X_n(P_{f\cdot\nu})(x) \\
&= \int V(y)\mathrm{d}(X_{n-1}, X_n)(P_{f\cdot\nu})(x,y) \\
&= \iint V(y)\,\mathrm{d}P_x(y)\,\mathrm{d}X_{n-1}(P_{f\cdot\nu})(x) \\
&\leq (\alpha V(x) + \beta)\mathrm{d}X_{n-1}(P_{f\cdot\nu})(x) \\
&= \alpha \int V(x)\mathrm{d}X_{n-1}(P_{f\cdot\nu})(x) + \beta \\
&\leq \cdots \\
&\leq \beta(1 + \alpha + \cdots + \alpha^{n-1}) + \alpha^n \int V(x)\mathrm{d}X_0(P_{f\cdot\nu})(x) \\
&= \beta(1 + \alpha + \cdots + \alpha^{n-1}) + \alpha^n \int V(x)f(x)\mathrm{d}\nu(x) \\
&\leq \frac{\beta}{1-\alpha} + \alpha^n \int V(x)f(x)\mathrm{d}\nu(x)
\end{aligned}$$

If $\int V(x)f(x)\,\mathrm{d}\nu(x) < \infty$, the last quantity is $\leq \frac{\beta}{1-\alpha} + 1$ for n sufficiently large.

The argument above works for $f \in \mathcal{D}$ such that $\int V(x)f(x)\mathrm{d}\nu(x) < \infty$, and we now complete the proof of the theorem by appealing to Lemma 5.4.4 and verifying that $\mathcal{D}^* = \{f \in \mathcal{D} : \int V(x)f(x)\,\mathrm{d}\nu(x) < \infty\}$ is dense in $\mathcal{D}$.

So let $f \in \mathcal{D}$. We want to find a sequence $(f_k)_{k\in\mathbb{N}}$ with each $f_k \in \mathcal{D}^*$ and with $\|f - f_k\| \to 0$. Define $B_k = \{x \in \mathcal{X} : V(x) \leq k\}$ and $c_k = \int_{B_k} f(x)\,\mathrm{d}\nu(x)$ for $k = 1, 2, \ldots$. Then $f_k = 1_{B_k} f/c_k \in \mathcal{D}^*$, and

$$\begin{aligned}
\|f - f_k\| &= \|f - 1_{B_k} f/c_k\| \\
&= (1/c_k - 1)\int_{B_k} f(x)\,\mathrm{d}\nu(x) + \int_{B_k^c} f(x)\,\mathrm{d}\nu(x) \\
&= (1/c_k - 1)c_k + (1 - c_k) \\
&= 2(1 - c_k) \to 0
\end{aligned}$$

as $k \to \infty$. □

Note: The assumption in the theorem that $y \mapsto \inf_{\{x\,:\,V(x)\leq r\}} k_x^{(m)}(y)$ should be measurable is not necessary. It is sufficient to complete the proof, that there exists a non–negative

measurable function $g : \mathcal{X} \to \mathbb{R}$ such that

$$\inf_{\{x\,:\,V(x)\leq r\}} k_x^{(m)}(y) \geq g(y)$$

for all $y \in \mathcal{X}$, and

$$\int g(y)\,\mathrm{d}\nu(y) > 0\,.$$

From Theorem 5.5.1 we can obtain that a given Markov chain has asymptotically stable transition probabilities, such that a stationary initial distribution exists and the averages from Theorem 5.3.4 converges (towards integrals with respect to the stationary distribution). But it says nothing about how the stationary distribution behaves. However, in order to use the convergence of such averages it is necessary to know whether integrals on the form

$$E_{f_0\cdot\nu}|f(X_0, X_1, X_2, \ldots)| = \int |f(x_0, x_1, x_2, \ldots)|\,\mathrm{d}\mathcal{P}_{f_0\cdot\nu}(x_0, x_1, x_2, \ldots)$$

are finite. For this the following corollary can be helpful.

Corollary 5.5.2. *Assume that the conditions from Theorem 5.5.1 are satisfied and let f_0 denote the density for the stationary initial distribution. Then*

$$E_{f_0\cdot\nu}(V(X_0)) = \int V(x)f_0(x)\,\mathrm{d}\nu(x) < \infty\,.$$

Proof. Let $f \in \mathcal{D}^*$ and $M > 0$. From the proof above we have

$$\begin{aligned}\int \big(V(x)\wedge M\big)P^{*n}(f)(x)\mathrm{d}\nu(x) &\leq \int V(x)P^{*n}(f)(x)\mathrm{d}\nu(x)\\ &= \int V(x)\,\mathrm{d}X_n(P_{f\cdot\nu})(x)\\ &\leq \frac{\beta}{1-\alpha} + \alpha^n \int V(x)f(x)\mathrm{d}\nu(x)\end{aligned}$$

and since the transition probabilities are asymptotically stable

$$\begin{aligned}&\Big|\int \big(V(x)\wedge M\big)P^{*n}(f)(x)\mathrm{d}\nu(x) - \int \big(V(x)\wedge M\big)f_0(x)\mathrm{d}\nu(x)\Big|\\ &\leq \int \big(V(x)\wedge M\big)\,\big|P^{*n}(f)(x) - f_0(x)\big|\,\mathrm{d}\nu(x)\\ &\leq \|P^{*n}(f) - f_0\| \to 0\end{aligned}$$

as $n \to \infty$. It follows that $\int \big(V(x)\wedge M\big)f_0(x)\mathrm{d}\nu(x) \leq \beta(1-\alpha)$ and letting $M \to \infty$ then yields that $\int V(x)f_0(x)\mathrm{d}\nu(x) \leq \beta(1-\alpha)$. □

Example 5.5.3. Continuation of Example 5.3.7. For the ARCH(1)–process we have

$$X_{n+1}^2 = (\gamma + \alpha X_n^2)\epsilon_{n+1}^2 \,.$$

We shall look for a drift function on the form

$$V(x) = |x|^{2\delta}$$

with $\delta > 0$. Since

$$X_{n+1}^{2\delta} = (\gamma + \alpha X_n^2)^\delta |\epsilon_{n+1}|^{2\delta}$$

the condition (5.14) means that (due to the Substitution Theorem 2.1.1, since $P_x \stackrel{\mathcal{D}}{=} \sqrt{\gamma + \alpha x^2}\epsilon$)

$$(\gamma + \alpha x^2)^\delta E(|\epsilon|^{2\delta}) \leq \alpha_0 |x|^{2\delta} + \beta$$

for some β, some $\alpha_0 < 1$, and where ϵ has a standard normal distribution. For small values of x, β takes care of this. The condition therefore becomes

$$\alpha^\delta E(|\epsilon|^{2\delta}) < 1 \,. \tag{5.17}$$

As a special case take $\delta = 1$. The condition is then satisfied if $\alpha < 1$ and from the corollary we see that the stationary initial distribution has finite second order moment.

Because $E((\alpha\epsilon^2)^0) = 1$, the inequality (5.17) may be written

$$\begin{aligned} &E\Big(\frac{\exp(\delta \log(\alpha\epsilon^2) - \exp(0 \log(\alpha\epsilon^2)}{\delta}\Big) \\ =&\frac{1}{\delta}\big(E(\exp(\delta \log(\alpha\epsilon^2)) - E(\exp(0 \log(\alpha\epsilon^2))\big) < 0 \end{aligned}$$

If we let $\delta \to 0$ we obtain the condition $E(\log(\alpha\epsilon^2)) < 0$ which is equivalent to

$$\alpha < \exp(-E(\log(\epsilon^2))) = 3.56\ldots$$

Hence this condition ensures that the transition probabilities are asymptotically stable. ∘

5.6 Exercises

Exercise 5.1. Assume that $X_0, X_1, X_2, \ldots$ is a time homogeneous Markov chain on $\mathcal{X}, \mathbb{E})$ with transition probabilities $(P_x)_{x\in\mathcal{X}}$. Let $x \in \mathcal{X}$ and define the stopping time

$$\tau = \inf\{n \in \mathbb{N} \,:\, X_n = x\} \,.$$

Let $X_0 \equiv x$ and assume that $P(\tau < \infty) = 1$. Assume furthermore that $E\tau < \infty$. Define for each set $A \in \mathbb{E}$

$$\mu(A) = \frac{1}{E\tau} E\left(\sum_{n=0}^{\tau-1} 1_{(X_n \in A)}\right)$$

(1) Show that μ is a probability measure on $(\mathcal{X}, \mathbb{E})$

(2) Show that for all non–negative measurable functions $f : \mathcal{X} \to \mathbb{R}$ it holds that

$$\int f \, d\mu = \frac{1}{E\tau}\left(\sum_{n=0}^{\tau-1} f(X_n)\right)$$

(Hint).

(3) Show that for $A \in \mathbb{E}$

$$\begin{aligned}\int P_x(A) \, \mathrm{d}\mu(x) &= \frac{1}{E\tau} \sum_{n=0}^{\infty} E\big(P_{X_n}(A) 1_{(\tau > n)}\big) \\ &= \frac{1}{E\tau} \sum_{n=0}^{\infty} \int_{(\tau > n)} P(X_{n+1} \in A \mid \mathbb{F}_n) \, dP \,,\end{aligned}$$

where $\mathbb{F}_n = \sigma(X_0, \ldots, X_n)$ (Hint).

(4) Obtain that

$$\int P_x(A) \, \mathrm{d}\mu(x) = \frac{1}{E\tau} E\left(\sum_{n=0}^{\tau-1} 1_{(X_{n+1} \in A)}\right)$$

(Hint).

(5) Show that

$$E(1_{(X_\tau \in A)}) = E(1_{(X_0 \in A)})$$

and use this to obtain

$$\int P_x(A) \, \mathrm{d}\mu(x) = \frac{1}{E\tau} E\left(\sum_{n=0}^{\tau-1} 1_{(X_n \in A)}\right) .$$

Conclude that μ is a stationary initial distribution for the Markov chain.

Now define $\tau_1 = \tau$ and recursively

$$\tau_N = \inf\{n > \tau_{N-1} \,:\, X_n = x\}$$

Let furthermore $f : \mathcal{X} \to \mathbb{R}$ be bounded and measurable.

(6) Argue that the real valued variables

$$\sum_{n=0}^{\tau_1-1} f(X_n), \quad \sum_{n=\tau_1}^{\tau_2-1} f(X_n), \quad \sum_{n=\tau_2}^{\tau_3-1} f(X_n), \dots$$

are independent and identically distributed (you do not have to give detailed arguments).

Also argue that $\tau_1, \tau_2-\tau_1, \tau_3-\tau_2, \dots$ are independent and identically distributed (Hint).

(7) Show that

$$\frac{1}{\tau_N} \sum_{n=0}^{\tau_N-1} f(X_n) \to \int f \, d\mu \qquad \text{a.s.}$$

as $n \to \infty$ (Hint).

∘

Exercise 5.2. Reconsider the one–dimensional AR(1)–process on $(\mathbb{R}, \mathbb{B})$ from 4.2

$$X_{n+1} = \rho X_n + \epsilon_{n+1} \,,$$

where all $\epsilon_1, \epsilon_2, \dots$ are iid with a $\mathcal{N}(0,1)$ distribution. Assume that X_0 is independent of all the ϵ's.

(1) Find the conditional distribution $(P_x^{*n})_{x\in\mathbb{R}}$ of X_n given X_0 for all $n \in \mathbb{N}_0$ (Hint).

(2) Assume that $|\rho| < 1$. Show that P_x^{*n} converges to μ_0 for all $x \in \mathbb{R}$, where

$$\mu_0 = \mathcal{N}\Big(0, \frac{1}{1-\rho^2}\Big)$$

(Hint).

∘

Exercise 5.3. Consider the AR(1) model from 5.2 and Example 5.3.6, but now assume that $|\rho| > 1$. From 5.2 and the example it is known that the transition densities are given by

$$k_x^{(n)}(y) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y-\rho^n x)^2}{2\sigma_n^2}\right)$$

with respect to the Lebesgue measure λ.

(1) Show that for all $x, y \in \mathbb{R}$

$$k_x^{(n)}(y) \to 0$$

as $n \to \infty$.

(2) Show that there does **not** exist a stationary initial distribution for the Markov chain (Hint).

$\circ$

Exercise 5.4. Let $X_0, X_1, X_2, \ldots$ be a time homogeneous Markov chain on $(\mathcal{X}, \mathbb{E})$ with transition probabilities $(P_x)_{x \in \mathcal{X}}$ and initial distribution μ. Assume that each P_x has density k_x with respect to the σ–finite measure ν. Furthermore assume that μ has density f with respect to ν.

Recall that we in section 4.4 defined $(\mathcal{P}_x^n)_{x \in \mathcal{X}}$ to be the conditional distribution of $(X_1, \ldots, X_n)$ given X_1.

(1) Show that $\mathcal{P}_x^n$ has density $k_{n,x}(x_1, \ldots, x_n)$ with respect to $\nu^{\otimes n}$, where

$$k_{n,x}(x_1, \ldots, x_n) = \prod_{i=1}^{n} k_{x_{i-1}}(x_i)$$

(Hint).

(2) Show that the distribution of $(X_0, \ldots, X_n)$ has density

$$h_n(x_0, \ldots, x_n) = f(x_0) k_{n,x_0}(x_1, \ldots, x_n)$$

with respect to $\nu^{\otimes n+1}$.

If the transition densities $(k_x)_{x \in \mathcal{X}} = (k_x^\theta)_{x \in \mathcal{X}}$ depends on some unknown parameter $\theta \in \Theta$, we can use the density h_n from (2) to write the likelihood function, when observing $(X_0, \ldots, X_n)$. However, in Markov chain models it is quite often only the transition densities that are specified and not the initial density f. In such cases, one will typically use the conditional likelihood given $X_0 = x$ as in question (1). Especially in situations with many observations it can be argued that not very much information is thrown away by considering the first observation as known.

Now consider the AR(1) model from 5.2, where $|\rho| < 1$.

(3) Find the conditional likelihood function for $(X_1, \ldots, X_n)$ given $X_0 = x$.

(4) Show that the conditional maximum likelihood estimate for ρ given $X_0 = x$ is

$$\hat{\rho}_n = \frac{\sum_{i=0}^{n-1} X_i X_{i+1}}{\sum_{i=0}^{n-1} X_i^2}$$

(Hint).

(5) Show that no matter what is the distribution of X_0, then the estimate from (4) is strongly consistent:

$$\hat{\rho}_n \to \rho \quad \text{a.s.}$$

as $n \to \infty$ (Hint).

∘

Exercise 5.5. Let $X_0, X_1, X_2, \ldots$ be a real valued time homogeneous Markov chain given by the update scheme

$$X_{n+1} = \phi(X_n, U_{n+1}),$$

where all $U_1, U_2, \ldots$ are independent and identically distributed with common density f with respect to the Lebesgue measure λ on $(\mathbb{R}, \mathbb{B})$. Assume furthermore that $(U_1, U_2, \ldots)$ is independent of X_0, and that $\phi : \mathbb{R}^2 \to \mathbb{R}$ is measurable, such that

- $y \mapsto \phi(x, y)$ is bijective and continuously differentiable for all $x \in \mathbb{R}$.
- $\delta_y \phi(x, y) \neq 0$ for all $(x, y) \in \mathbb{R}^2$, where $\delta_y \phi(x, y)$ denotes the derivative with respect to y.

Show that the transition probabilities $(P_x)_{x \in \mathcal{X}}$ for the Markov chain have densities $(k_x)_{x \in \mathcal{X}}$ with respect to the Lebesgue measure, where

$$k_x(y) = \frac{1}{|\delta_y \phi(x, \phi(x, \cdot)^{-1}(y))|} f(\phi(x, \cdot)^{-1}(y))$$

(Hint). ∘

Exercise 5.6. Assume that $\xi : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to (0, \infty)$ are continuous functions, and that $(U_n)_{n \in \mathbb{N}}$ is a sequence of independent and identically distributed random variables with finite second order moment, such that $EU_1 = 0$ and $VU_1 = 1$. Define recursively

$$X_{n+1} = \xi(X_n) + \sigma(X_n) U_{n+1},$$

where X_0 is a real random variable, that is independent of $(U_n)_{n \in \mathbb{N}}$ and has distribution μ.

(1) Argue that $X_0, X_1, X_2, \ldots$ is a Markov chain.

(2) Assume that for some $n \in \mathbb{N}$ it holds that

$$E\xi(X_n)^2 < \infty \quad \text{and} \quad E\sigma(X_n)^2 < \infty$$

Show that $EX_{n+1}^2 < \infty$ and find $E(X_{n+1} \mid X_n = x)$ and $V(X_{n+1} \mid X_n = x)$.

Now assume that the distribution of U_1 (and all U_n) has density f with respect to the Lebesgue measure.

(3) Find the density for the transition probabilities.

Assume that the density $f : \mathbb{R} \to \mathbb{R}$ is continuous and strictly positive.

(4) Assume that

$$\limsup_{|x|\to\infty} \frac{\xi(x)^2 + \sigma(x)^2}{x^2} < 1 \tag{5.18}$$

Show that the transition probabilities are asymptotically stable using $V(x) = x^2$ as a drift function (Hint).

(5) Assume that $EX_0^2 < \infty$. Show that $EX_n^2 < \infty$ for all $n \in \mathbb{N}$ (Hint).

∘

Exercise 5.7. This exercise gives an example of the Markov chains from 5.6. So assume that $U_1, U_2, \ldots$ are independent and identically distributed with $EU_1 = 0$ and $VU_1 = 1$. Assume that U_1 has a continuous and strictly positive density f with respect to λ. Define $\sigma(x) \equiv 1$ and

$$\xi(x) = (\alpha_1 x + \beta_1)1_{(-\infty,\gamma)}(x) + (\alpha_2 x + \beta_2)1_{[\gamma,\infty)}(x)\,,$$

and assume that $\alpha_1\gamma + \beta_1 = \alpha_2\gamma + \beta_2$.

(1) Draw $\xi(x)$ as a function of x, and argue that it is continuous.

(2) Let $X_0, X_1, X_2, \ldots$ be a Markov chain given by

$$X_{n+1} = \xi(X_n) + \sigma(X_n)U_{n+1}\,,$$

where X_0 is independent of $(U_n)_{n\in\mathbb{N}}$. Find conditions in terms of $\alpha_1, \beta_1, \alpha_2, \beta_2$ such that the transition probabilities are asymptotically stable.

(3) Try simulating $X_0, \ldots, X_{10000}$ in a situation, where $X_0 = 0$, $U_1 \sim \mathcal{N}(0,1)$, $-\alpha_1 = \alpha_2 = \frac{1}{2}$, and $\beta_1 = \beta_2 = \gamma = 0$. Plot $(n, X_n)_{n\in\mathbb{N}_0}$ and $(X_n, X_{n+1})_{n\in\mathbb{N}_0}$ and comment the two plots.

$\circ$

Exercise 5.8. Assume that $f : [0,\infty) \to (0,\infty)$ is a continuous and strictly positive density on the interval $[0,\infty)$. Let $U_1, U_2, \ldots$ be a sequence of independent and identically distributed real non–negative variables with common density f. Assume that there exists $\beta > 0$ such that

$$\int_0^\infty e^{\beta x} f(x)\, dx < \infty\,.$$

Assume that X_0 is independent of $(U_n)_{n\in\mathbb{N}}$ and define the Markov chain $X_0, X_1, X_2, \ldots$ by

$$X_{n+1} = |X_n - U_{n+1}|$$

for $n \in \mathbb{N}_0$.

(1) Show that the transition probabilities $(P_x)_{x\in\mathbb{R}}$ have densities $(k_x)_{x\in\mathbb{R}}$ given by

$$k_x(y) = f(x+y) + f(x-y)1_{(y<x)}\,.$$

(Hint).

(2) Show that

$$\int_0^\infty e^{\beta y} k_x(y)\, dy = e^{-\beta x}\int_x^\infty e^{\beta y} f(y)\, dy + \int_0^x e^{-\beta y} f(y)\, dy$$

(3) Show that the transition probabilities are asymptotically stable and that there exists a uniquely determined stationary density f_0 (Hint).

(4) Show that if X_0 has the stationary distribution, then $EX_0^n < \infty$ for all $n \in \mathbb{N}$ (Hint).

(5) Let X_0 have the stationary distribution. Show that

$$EX_0 = \frac{EU_1^2}{2EU_1}$$

(Hint).

(6) Assume that $f(x) = e^{-x}$. Show that in this case also $f_0(x) = e^{-x}$ (Hint).

∘

Exercise 5.9. Let $X_0, X_1, X_2, \ldots$ be a Markov chain on $(\mathcal{X}, \mathbb{E})$ with transition probabilities $(P_x)_{x\in\mathcal{X}}$ such that each P_x has density k_x with respect to a σ–finite measure ν. Assume that for some m it holds that $y \mapsto \inf_{\{x\in\mathcal{X}\}} k^{(m)}(y)$ is measurable with

$$\int \inf_{\{x\in\mathcal{X}\}} k^{(m)}(y)\, \mathrm{d}\nu(y) > 0\,.$$

Show that the transition probabilities are asymptotically stable (Hint). ∘

Exercise 5.10. The Gibb's sampler.
Let (X, Y) be a vector of random variables with values in $(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$. Assume that (X, Y) has density f_0 with respect to $\nu_1 \otimes \nu_2$.

Think of a situation, where f_0 is rather complicated and difficult to simulate from (or maybe we do not even know it). Instead we know the densities for the conditional distributions of $X \mid Y$ and $Y \mid X$. These densities will be denoted $(f^1_y)_{y\in\mathcal{X}}$ and $(f^2_x)_{x\in\mathcal{X}}$ respectively.

The idea is to generate a Markov chain $(Z_n)_{n\in\mathbb{N}_0}$ consisting of vectors $Z_n = (X_n, Y_n)$ that have asymptotically stable transition densities with f_0 as the stationary distribution.

We define $(Z_n)_{n\in\mathbb{N}_0}$ as follows:

(a) Let $Z_0 = (X_0, Y_0) = (x, y)$ for some $(x, y) \in \mathcal{X}^2$.

(b) Given $Z_n = (X_n, Y_n) = (x_n, y_n)$ draw X_{n+1} from the distribution with density $f^1_{y_n}$.

(c) Given $X_{n+1} = x_{n+1}$ draw Y_{n+1} from the distribution with density $f^2_{x_{n+1}}$.

(d) Define $Z_{n+1} = (X_{n+1}, Y_{n+1})$

(e) Add 1 to n and go back to (b).

The above successive definition of $Z_0, Z_1, Z_2, \ldots$ could of course be written more formally as: For each $n \in \mathbb{N}_0$ we have

- $X_{n+1} \perp\!\!\!\perp (Z_0, \ldots, Z_n) \mid Y_n$
- With $(Q^1_y)_{y\in\mathcal{Y}} = X_{n+1} \mid Y_n$ each Q^1_y has density f^1_y.

- $Y_{n+1} \perp\!\!\!\perp (Z_0, \ldots, Z_n) \mid X_{n+1}$
- With $(Q_x^2)_{x \in \mathcal{X}} = Y_{n+1} \mid X_{n+1}$ each Q_x^2 has density f_x^2.

(1) Argue that $Z_0, Z_1, Z_2, \ldots$ is a Markov chain.

Let $(P_{x,y})_{(x,y) \in \mathcal{X} \times \mathcal{Y}}$ be the transition probabilities of this Markov chain.

(2) Show that the P_{x_1,y_1} has density $k_{(x_1,y_1)}$ with respect to $\nu_1 \otimes \nu_2$, where

$$k_{(x_1,y_1)}(x_2, y_2) = f_{y_1}^1(x_2) f_{x_2}^2(y_2)$$

(Hint).

(3) Show that f_0 is a stationary density for the transitions (Hint).

Now assume that $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, $\nu_1 = \lambda$ (restricted to $[0,1]$), and that ν_2 is the counting measure τ on $\{0,1\}$. Furthermore assume that the conditional distribution of X_{n+1} given Y_n is given by

$$Q_y^1 \stackrel{\mathcal{D}}{=} B(y+1, 2-y)$$

meaning that

$$f_{y_1}^1(x_2) = \frac{1}{B(y_1+1, 2-y_1)} x_2^{y_1} (1-x_2)^{1-y_2}$$

and that the conditional distribution of Y_{n+1} given $X_{n+1} = x_{n+1}$ is a Binomial distribution with parameters

$$\Big(1, \frac{x_{n+1}a}{x_{n+1}a + (1 - x_{n+1}b)}\Big)$$

for some given non–negative constants a and b.

(4) Show that the Markov chain $Z_0, Z_1, Z_2, \ldots$ is asymptotically stable (Hint).

$\circ$

Chapter 6

An introduction to Bayesian networks

6.1 Introduction

Consider an n–dimensional vector of random variables $\mathbb{X} = (X_1, \ldots, X_n)$, where each X_i e.g. have values in some Borel space $(\mathcal{X}, \mathbb{E})$.

A very simple model for the distribution of $\mathbb{X}$ could be to assume that $X_1, \ldots, X_n$ are independent. Then we have

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$$

so in order to determine the distribution of $\mathbb{X}$, we only need to determine the n marginal distributions. The problem is, that this model very often is far too simple; It is necessary to allow for some dependence structure in the simultaneous distribution.

If we take this to the other extreme, we could simply use the model, where $\mathbb{X}$ is allowed to have any distribution on $(\mathcal{X}^n, \mathbb{E}^n)$. Then we will have to take care of all dependencies between all combinations of variables. If n is very large, this will be far too complex to handle.

Example 6.1.1. Consider the variables $X_1, \ldots, X_n$ in the very simple case, where $\mathcal{X} = \{0, 1\}$. If the variables are independent only the n parameters $p_1, \ldots, p_n$ will be necessary to determine the distribution of $(X_1, \ldots, X_n)$

$$P(X_k = 1) = p_k$$

If on the other hand, there is dependence then all the 2^n probabilities

$$P(X_1 = a_1, \ldots, X_n = a_n)$$

will have to be decided, with $a_k \in \{0, 1\}$ for $k = 1, \ldots, n$. ◦

Example 6.1.2. In a microarray experiment the expression levels of a large number of genes are measured simultaneously, and it is expected that there will be correlations between the levels of the different genes. These correlation structures are of great interest, and can lead to a deeper biological understanding.

Consider a gene expression data set produced from an experiment, where the expression levels of p genes are recorded for n independent samples. This leads to a data set $\{x^1, \ldots, x^n\}$ where each vector $x^k = (x_1^k, \ldots, x_n^k)$ is an independent observation of the vector $(X_1, \ldots, X_p)$. The difficulty in the analysis of data like this is that typically p is much larger than n. ◦

For large values of n a solution could be to assume that there is still some dependence in the model, but in such a way that there are not too many dependencies to take care of.

As we shall see in the following example it often makes sense to assume some structure in a set of observations.

Example 6.1.3. Imagine the situation, when you go to the university in the morning. It may happen that you do not arrive in time for the first lecture.... And this can have various reasons. It could be that your alarm did not work this morning, such that you overslept, and it is also possible that the bus was late. All of this will probably have influence on whether you were at the university in time.

In Figure 6.1 it is indicated which dependence structures that seems reasonable. For example it does not make sense to believe that there is dependence between the bus arrival and your alarm. Furthermore it is reasonable that the alarm only has influence on the arrival time via the knowledge of whether you overslept: If we know that you did **not** oversleep, then it will not be fair to use the alarm as an excuse for being late. ◦

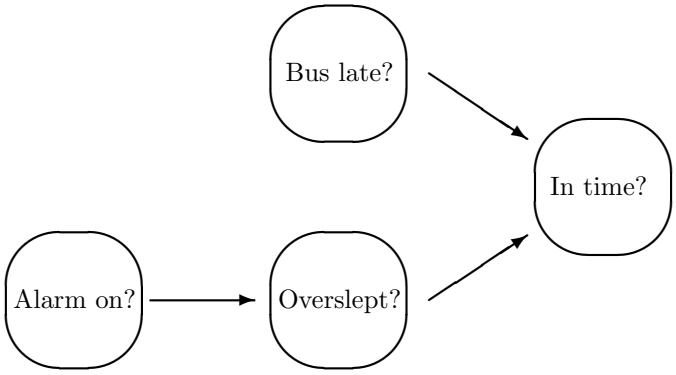

Figure 6.1: *A scheme of your morning situation.*

The purpose of this chapter is to give a theoretical description of dependence structures like the one in Example 6.1.3.

6.2 Directed graphs

Definition 6.2.1. *A **directed graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a set of nodes $\mathcal{V} = (v_1, \ldots, v_n)$ and a set of edges $\mathcal{E}$. Each edge is a directed connection between to elements $v_i, v_j \in \mathcal{V}$, where $i \neq j$. Such a directed connection will be denoted $v_i \to v_j$.*

Definition 6.2.2. *We say that $v_1, \ldots, v_k$ form a **path** in the directed graph $\mathcal{G}$, if $v_i \to v_{i+1}$ for every $i = 1, \ldots, k-1$. We write $v_i \rightsquigarrow v_j$ if there exists a path from v_i to v_j.*

Definition 6.2.3. *A **cycle** in a directed graph $\mathcal{G}$ is a path $v_1, \ldots, v_k$, where $v_1 = v_k$. A graph is **acyclic** if it contains no cycles.*

The concept of a directed acyclic graph (DAG) will be the basis in this chapter. This will be the graphical representation that underlies Bayesian networks.

Definition 6.2.4. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed acyclic graph, and let B be a subset of $\mathcal{V}$. The **subgraph** $\mathcal{G}_B = (B, \mathcal{E}_B)$ is the graph, that only consists of the nodes in B and only with the edges from $\mathcal{E}$ that are connections between elements in B.*

Another useful notion is that of an ordering of the nodes in a directed graph that is consistent with the directions of the edges.

Definition 6.2.5. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed acyclic graph. An **ordering** of the nodes $v_1, \ldots, v_n$ is an ordering relative to $\mathcal{G}$: If $v_i \to v_j$, then $i < j$.*

It is always possible to find an ordering in a directed acyclic graph!

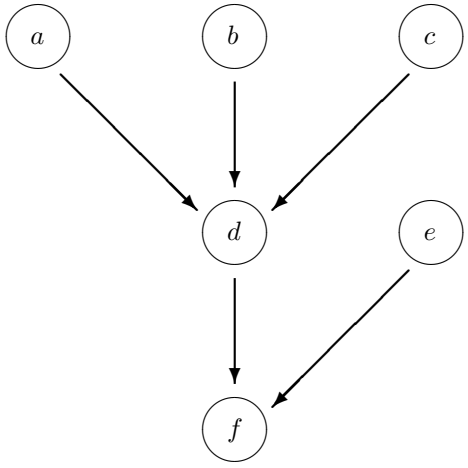

Figure 6.2: *An example of a directed acyclic graph with 6 nodes and 5 edges. We e.g. see that $a \rightsquigarrow f$, and that $\{a, b, c\}$ is the set of parents of d.*

We shall also for a node $v \in \mathcal{V}$ define the parents, descendants, and non–descendants of v.

Definition 6.2.6. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed acyclic graph, and let $v_i \in \mathcal{V}$. Then we say that*

(i) v_j *is a **parent** of* v_i*, if* $v_j \to v_i$*.*

$$P_{v_i} = \{v_j \in \mathcal{V} \,:\, v_j \to v_i\}$$

denotes the set of all the parents of v_i*.*

(ii) v_j *is a **child** of* v_i*, if* $v_i \to v_j$*. We let*

$$C_{v_i} = \{v_j \in \mathcal{V} \,:\, v_i \to v_j\}$$

denote the set of all children of v_i*.*

(iii) v_k *is an **ancestor** of* v_i*, if there exists a path, such that* $v_k \rightsquigarrow v_i$*. We let*

$$A_{v_i} = \{v_k \in \mathcal{V} \,:\, v_k \rightsquigarrow v_i\} \cup \{v_i\}$$

denote the set of all ancestors of v_i.

Let B be some subset of $\mathcal{V}$. Then we let A_B denote the subset of $\mathcal{V}$ that contains B and all ancestors of elements in B. Then

$$A_B = \cup_{v \in B} A_v$$

*(iv) v_k is a **descendant** of v_i, if there exists a path, such that $v_i \rightsquigarrow v_k$. We let*

$$D_{v_i} = \{v_k \in \mathcal{V} : v_i \rightsquigarrow v_k\}$$

denote the set of all descendants of v_i. Furthermore we let ND_{v_i} denote the set of all non-descendants of v_i and define it by

$$ND_{v_i} = \mathcal{V} \setminus \{D_{v_i} \cup \{i\}\}$$

6.3 Moral graphs and separation

Definition 6.3.1. *An **undirected graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a set of nodes $\mathcal{V} = (v_1, \ldots, v_n)$ and a set of edges $\mathcal{E}$. Each edge is a connection between two elements $v_i, v_j \in \mathcal{V}$, where $i \neq j$. Such an undirected connection will be denoted $v_i - v_j$.*

Definition 6.3.2. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed acyclic graph. The **moral graph** $\mathcal{G}^m$ of $\mathcal{G}$ is an undirected graph with the same nodes as $\mathcal{G}$, but where $v_i - v_j$ in $\mathcal{G}^m$ if either they are connected in $\mathcal{G}$ or they share a child.*

In the anachronistically named moral graph, all parents are "married"!

Definition 6.3.3. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, and let S be some subset of $\mathcal{V}$.*

*Let $v_i, v_j \in \mathcal{V} \setminus S$. We say that S **separates** v_i and v_j, if all paths from v_i to v_j intersects S.*

Assume that A, B and S are disjoint subsets of $\mathcal{V}$. We say that S separates A and B, if it separates all $v_i \in A$ and $v_j \in B$.

Example 6.3.4.

Consider the graph $\mathcal{G}$ that is depicted in Figure 6.3. The graph is acyclic, and the parent and child sets of each node can be found. For example $P_4 = \{1, 2\}$ and $C_4 = \{5\}$. Similarly $A_4 = \{1, 2\}$, $D_4 = \{5\}$ and $ND_4 = \{1, 2, 3\}$. The corresponding moral graph is also shown in this figure. ◦

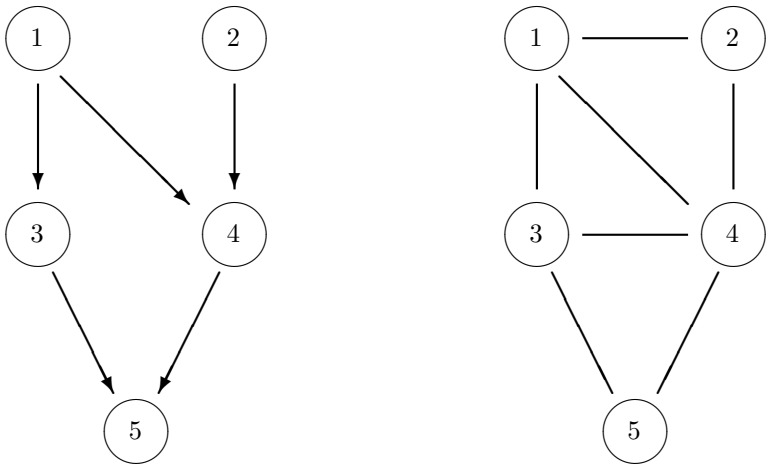

Figure 6.3: *The graphs discussed in Example 6.3.4. On the left is an example of a directed acyclic graph, and on the right the moral graph.*

6.4 Bayesian networks

Consider a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $V = \{v_1, \ldots, v_n\}$ and consider a random vector $\mathbb{X}$ of length n, that is indexed by the nodes in the graph

$$\mathbb{X} = (X_{v_1}, \ldots, X_{v_n}) .$$

In the following we shall use the notation $P_v(\mathbb{X}) = \{X_u : u \in P_v\}$ and similarly for $C_v(\mathbb{X})$, $A_v(\mathbb{X})$ $D_v(\mathbb{X})$ and $ND_v(\mathbb{X})$.

Assume that each of the variables X_v has values in the Borel space $(\mathcal{X}, \mathbb{E})$. We let $\mathcal{P} = \mathbb{X}(P)$ denote the distribution of $\mathbb{X}$ – hence $\mathcal{P}$ is a probability measure on $(\mathcal{X}^n, \mathbb{E}^n)$.

We have the following definition of a Bayesian network

Definition 6.4.1. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG containing n nodes, and let $\mathbb{X} = (X_{v_1}, \ldots, X_{v_n})$ be a vector of random variables with values in $(\mathcal{X}, \mathbb{E})$ indexed by $\mathcal{V}$. let $\mathcal{P}$ be the distribution of $\mathbb{X}$.*

The triplet $(\mathcal{G}, \mathbb{X}, \mathcal{P})$ is called a Bayesian network, if for each $v \in \mathcal{V}$

$$X_v \perp\!\!\!\perp ND_v(\mathbb{X}) \mid P_v(\mathbb{X}) .$$

Example 6.4.2. Assume that $\mathcal{X}$ is indexed by the graph in Figure 6.2, and assume that $(\mathcal{G}, \mathbb{X}, \mathcal{P})$ is a Bayesian network. Then we e.g. have the conditional independence

$$X_f \perp\!\!\!\perp (X_a, X_b, X_c) \mid (X_d, X_e)$$

and the true independence

$$X_a \perp\!\!\!\perp X_b \perp\!\!\!\perp X_c$$

∘

Example 6.4.3. Assume that $X_0, X_1, X_2, \ldots$ is a Markov chain, and consider $\mathbb{X} = (X_0, \ldots, X_n)$ for some n. Consider the very simple graph $\mathcal{G}$ given by

$$0 \to 1 \to 2 \to \cdots \to n$$

then for each $k \in \{0, \ldots, n\}$ we have that $P_k = \{k-1\}$, $ND_k = \{0, \ldots, k-1\}$ and we therefore have

$$X_k \perp\!\!\!\perp ND_k(\mathbb{X}) \mid P_k(\mathbb{X})$$

Hence it is seen that $(\mathcal{G}, \mathbb{X}, \mathbb{X}(P))$ is a Bayesian network. ∘

In the rest of this chapter we shall (for our own convenience and without loss of generality) assume that $\mathcal{V} = \{1, \ldots, n\}$ and that this an ordering of the elements in $\mathcal{V}$.

We shall need a (little...) more notation. Let $n(i)$ be the number of elements in P_i for each $i \in \mathcal{V}$, and let

$$(P_{i,x})_{x \in \mathcal{X}^{n(i)}}$$

denote the conditional distribution of X_i given $P_i(\mathbb{X})$. If P_i is empty, we simply use the unconditional distribution of X_i. Similarly to the definition of $P_i(\mathbb{X})$ we shall consider $P_i(\mathbf{x})$ for the vector $\mathbf{x} = (x_1, \ldots, x_n)$. So we have

$$P_i(\mathbf{x}) = \{x_j \,:\, j \in P_i\}$$

Now we can formulate and prove

Theorem 6.4.4. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with $\mathcal{V} = \{1, \ldots, n\}$ (for convenience. Let $\mathbb{X} = (X_1, \ldots, X_n)$ be a random vector indexed by $\mathcal{V}$. Assume that $1, \ldots, n$ is an ordering of the nodes in $\mathcal{V}$.*

Then $(\mathcal{G}, \mathbb{X}, \mathcal{P})$ is a Bayesian network if and only if the distribution $\mathcal{P}$ of $\mathbb{X}$ is given by the conditional distributions

$$\left\{(P_{i,x})_{x \in \mathcal{X}^{n(i)}} \,:\, i \in \mathcal{V}\right\}$$

in the following way

$$\begin{aligned}&\mathcal{P}(A_1 \times \cdots \times A_n)\\ &= \int_{A_1} \cdots \int_{A_{n-1}} P_{n,P_n(\mathbf{x})}(A_n)\, P_{n-1,P_{n-1}(\mathbf{x})}(dx_{n-1}) \cdots P_{2,P_2(\mathbf{x})}(dx_2) P_1(dx_1)\end{aligned}$$

Proof. Let for each $i \in \{1, \ldots, n\}$ $\left(Q_{i,(x_1,\ldots,x_{i-1})}\right)$ be the conditional distribution of X_i given $(X_1, \ldots, X_{i-1})$. We let Q_1 be the marginal distribution of X_1. With this notation we can always write

$$\begin{aligned}&\mathcal{P}(A_1 \times \cdots \times A_n)\\ &= \int_{A_1} \cdots \int_{A_{n-1}} Q_{n,(x_1,\ldots,x_{n-1})}(A_n)\, Q_{n-1,(x_1,\ldots,x_{n-2})}(dx_{n-1}) \cdots Q_{2,x_1}(dx_2) Q_1(dx_1)\end{aligned}$$

Now the integral representation of $\mathcal{P}(A_1 \times \cdots \times A_n)$ in the theorem follows by noticing that for each i we have the conditional independence

$$X_i \perp\!\!\!\perp (X_1, \ldots, X_{i-1}) \mid P_i(\mathbb{X})\,,$$

such that

$$Q_{i,(x_1,\ldots,x_{i-1})} = P_{i,P_i(\mathbf{x})}\,.$$

Assume conversely that the distribution $\mathcal{P}$ has the integral form. Then it is seen that for each $i \in \{1, \ldots, n\}$ the conditional distribution of X_i given $(X_1, \ldots, X_{i-1})$ is given by the Markov kernel (with a slight change of the index–set)

$$(P_{i,x})_{x \in \mathcal{X}^{n(i)}} \tag{6.1}$$

From this we see for each i

$$X_i \perp\!\!\!\perp (X_1, \ldots, X_{i-1}) \mid P_i(\mathbb{X})$$

Now let i be fixed, and let $v_1, \ldots, v_m$ be the elements in the non–descendants of i that are not among $\{1, \ldots, i-1\}$:

$$\{v_1, \ldots, v_m\} = ND_i \setminus \{1, \ldots, i-1\}$$

We furthermore assume that $v_1, \ldots, v_m$ are ordered according to the ordering of $\mathcal{V}$. We want to show that

$$X_i \perp\!\!\!\perp (X_1, \ldots, X_{i-1}, X_{v_1}, \ldots, X_{v_m}) \mid P_i(\mathbb{X}) \tag{6.2}$$

and this will be shown using induction over $k = 0, \ldots, m$. As the induction start for $k = 0$ we have the conditional independence in (6.1). So now assume that for some $k \in \{0, \ldots, m\}$ we have

$$X_i \perp\!\!\!\perp (X_1, \ldots, X_{i-1}, X_{v_1}, \ldots, X_{v_k}) \mid P_i(\mathbb{X}) \tag{6.3}$$

For $X_{v_{k+1}}$ we have

$$X_{v_{k+1}} \perp\!\!\!\perp (X_1, \ldots, X_{v_{k+1}-1}) \mid P_{v_{k+1}}(\mathbb{X})$$

Note that because of the ordering, we must have $\{1, \ldots, i-1, v_1, \ldots, v_k\} \subseteq \{1, \ldots, v_{k+1}-1\}$. So we can move this information to the conditioning side

$$X_{v_{k+1}} \perp\!\!\!\perp (X_1, \ldots, X_{v_{k+1}-1}) \mid P_{v_{k+1}}(\mathbb{X}), X_1, \ldots, X_{i-1}, X_{v_1}, \ldots, X_{v_k}$$

Also note that since v_{k+1} is among the non–descendants of i, then also the elements in $P_{v_{k+1}}$ must be among the non–descendants. Hence $P_{v_{k+1}} \subseteq \{1, \ldots, i-1, v_1, \ldots, v_k\}$, so we actually have

$$X_{v_{k+1}} \perp\!\!\!\perp (X_1, \ldots, X_{v_{k+1}-1}) \mid X_1, \ldots, X_{i-1}, X_{v_1}, \ldots, X_{v_k}$$

which by reduction gives

$$X_{v_{k+1}} \perp\!\!\!\perp X_i \mid X_1, \ldots, X_{i-1}, X_{v_1}, \ldots, X_{v_k}$$

Hence (since obviously $P_i \subseteq \{1, \ldots, i-1\}$) we have from Theorem 3.4.3 that

$$X_i \perp\!\!\!\perp (X_1, \ldots, X_{i-1}, X_{v_1}, \ldots, X_{v_{k+1}}) \mid P_i(\mathbb{X})$$

So the desired result (6.2) follows by induction. □

The integral form from Theorem 6.4.4 becomes much more clear in the situation, where the distribution $\mathcal{P}$ has density.

Theorem 6.4.5. *Assume that the distribution of $(X_1, \ldots, X_n)$ has density f with respect to ν^{*n}, where ν is a σ–finite measure on $(\mathcal{X}, \mathbb{E})$. Then $(\mathcal{G}, \mathbb{X}, \mathcal{P})$ is a Bayesian network if and only if the density factorises*

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid P_i(\mathbf{x}))\,,$$

where each $f(x_i \mid P_i(\mathbf{x}))$ is the density of the conditional distribution of X_i given $P_i(\mathbb{X})$.

Proof. This result follows immediately from Theorem 6.4.4, since we know from Chapter 1, that all the conditional densities exists and that a factorisation as above determines the simultaneous distribution uniquely. □

Of course we can use update schemes in order to describe how X_i depends on it's parents $P_i(\mathbb{X})$ more concretely. We have results very similar to Theorem 4.1.6 and Theorem 4.1.7.

Theorem 6.4.6. *Assume that* $\mathbb{X} = (X_1, \ldots, X_n)$ *is defined recursively such that*

$$\begin{aligned} X_1 &= \phi_1(U_1) \\ X_2 &= \phi_2(X_1, U_2) \\ X_3 &= \phi_3(X_1, X_2, U_3) \\ \ldots &\quad \ldots \\ X_n &= \phi_n(X_1, \ldots, X_{n-1}, U_n) \end{aligned}$$

where $U_1, \ldots, U_n$ *are independent. Define a graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$*, where* $j \notin P_i$ *if* $j > i$*, and if* $j < i$ *then* $j \in P_i$ *if and only if* $\phi_i(x_1, \ldots, x_{i-1})$ *depends on* x_j*. Then* $(\mathcal{G}, \mathbb{X}, \mathcal{P})$ *is a Bayesian network, and* $1, \ldots, n$ *is an ordering of the notes in* $\mathcal{V}$*.*

Proof. Similar to the proof of Theorem 4.1.6 - but with heavier notation. □

The most simple form of the update scheme is a linear model. This is explained in the following example.

Example 6.4.7. Assume that $\mathbb{X} = (X_1, \ldots, X_n)$ is defined recursively such that

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \gamma_{2,1} X_1 + \epsilon_2 \\ X_3 &= \gamma_{3,1} X_1 + \gamma_{3,2} X_2 + \epsilon_3 \\ \ldots &\quad \ldots \\ X_n &= \gamma_{n,1} X_1 + \cdots + \gamma_{n,n-1} X_{n-1} + \epsilon_n \end{aligned}$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid. A Bayesian network over $\mathbb{X}$ then consists of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\gamma_{i,j} \neq 0$ if and only if $j \in P_i$.

A simple and useful model would be obtained by assuming that $\epsilon_i \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, n$. In that case Theorem 6.4.5 can be applied to derive the simultaneous density of $\mathbb{X}$. ∘

Theorem 6.4.8. *Assume* $(\mathcal{G}, \mathbb{X}, \mathcal{P})$ *is a Bayesian network, and assume that* $1, \ldots, n$ *is an ordering of the notes in* $\mathcal{V}$*. Then there exists update functions* $\phi_1, \ldots, \phi_n$ *with* $\phi_i : \mathcal{X}^{n(i)} \times (0, 1) \to \mathcal{X}$*, such that if* $U_1, \ldots, U_n$ *are independent standard uniformly distributed, and*

$X'_1, \ldots, X'_n$ are defined by

$$\begin{aligned} X_1 &= \phi_1(U_1) \\ X_2 &= \phi_2(X_1, U_2) \\ X_3 &= \phi_3(X_1, X_2, U_3) \\ \ldots &\qquad \ldots \\ X_n &= \phi_n(X_1, \ldots, X_{n-1}, U_n) \end{aligned}$$

then $(X'_1, \ldots, X'_n)$ has the same distribution as $(X_1, \ldots, X_n)$.

Proof. Similar to the proof of Theorem 4.1.7. □

6.5 Global Markov property

We have the following result, that can be seen as a generalisation of the conditional independence that appears in the definition of a Bayesian network.

Theorem 6.5.1. *Assume that $(\mathcal{G}, \mathbb{X}, \mathcal{P})$ is a Bayesian network, and let A, B and S be disjoint sets in $\mathcal{V}$, such that S separates A and B in the moral graph $\left(\mathcal{G}_{A_{A\cup B\cup S}}\right)^m$ of the sub-graph containing all ancestors of A, B and S. Then*

$$X_A \perp\!\!\!\perp X_B \mid X_S \,,$$

where e.g. X_A denotes $\{X_i : i \in A\}$.

Example 6.5.2. In Figure 6.4 a representation of a Bayesian network can be seen. It can be seen, that S separates A and B in the moral graph $\left(\mathcal{G}_{A_{A\cup B\cup S}}\right)^m$, so we have that

$$X_A \perp\!\!\!\perp X_B \mid X_S \,,$$

∘

Most of the work in the proof is in obtaining the following result, which is formulated as a separate result in order to ease notation.

Lemma 6.5.3. *Assume that $(\mathcal{G}, \mathbb{X}, \mathcal{P})$ is a Bayesian network, and let A, B and S be disjoint sets in $\mathcal{V}$, such that $\mathcal{V} = A \cup B \cup S$. Assume furthermore that S separates A and B in the moral graph $\mathcal{G}^m$. Then*

$$X_A \perp\!\!\!\perp X_B \mid X_S \,,$$

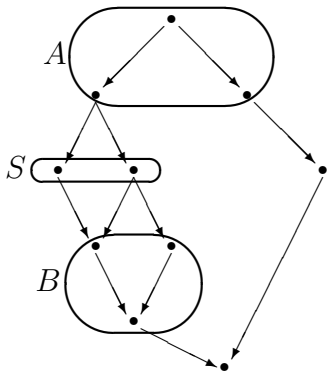

Figure 6.4: *A representation of a Bayesian network, where A and B are conditionally independent given S.*

Proof. We still assume that $\mathcal{V} = \{1, \ldots, n\}$ are ordered. Define for each $k \in \{1, \ldots, n\}$

$$A_k = A \cap \{1, \ldots, k\}$$
$$B_k = B \cap \{1, \ldots, k\}$$
$$S_k = S \cap \{1, \ldots, k\}$$

Let k_0 be the smallest number among $\{1, \ldots, n\}$, where both A_{k_0} and B_{k_0} are non–empty. Assume that $k_0 \in A$. Then both B_{k_0} and S_{k_0} are parts of the non–descendants of k_0, so

$$X_{k_0} \perp\!\!\!\perp X_{B_{k_0}}, X_{S_{k_0}} \mid P_{k_0}(\mathbb{X})$$

We can move $X_{S_{k_0}}$ to the conditioning side and use reduction

$$X_{k_0} \perp\!\!\!\perp X_{B_{k_0}} \mid P_{k_0}(\mathbb{X}), X_{S_{k_0}}$$

and since the parents of k_0 necessarily must be in S_{k_0}, we have

$$X_{k_0} \perp\!\!\!\perp X_{B_{k_0}} \mid X_{S_{k_0}}$$

We now proceed by induction over $k = k_0, \ldots, n$. So assume that for some k

$$X_{A_k} \perp\!\!\!\perp X_{B_k} \mid X_{S_k} \tag{6.4}$$

and consider $k + 1$. There are three different scenarios

(a) $k + 1 \in A$

(b) $k+1 \in B$

(c) $k+1 \in S$

and we would like to argue that in any case, it holds that

$$X_{A_{k+1}} \perp\!\!\!\perp X_{B_{k+1}} \mid X_{S_{k+1}} \tag{6.5}$$

The arguments for (a) and (b) will be identical (because of symmetry), so we shall only consider the scenarios (a) and (c).

First assume that $k+1 \in A$. Since all of the subsets A_k, B_k and S_k are among the non–descendants of $k+1$ we have

$$X_{k+1} \perp\!\!\!\perp X_{A_k}, X_{B_k}, X_{S_k} \mid P_{k+1}(\mathbb{X})$$

and we move X_{A_k} and X_{S_k} to the conditioning side and use reduction

$$X_{k+1} \perp\!\!\!\perp X_{B_k} \mid P_{k+1}(\mathbb{X}), X_{A_k}, X_{S_k}$$

Furthermore we see that the parents of $k+1$ are in $S_k \cup A_k$, so the result is simply

$$X_{k+1} \perp\!\!\!\perp X_{B_k} \mid X_{A_k}, X_{S_k} \tag{6.6}$$

When combining (6.4) and (6.6) it follows from Theorem 3.18, that

$$X_{A_k}, X_{k+1} \perp\!\!\!\perp X_{B_k} \mid X_{S_k}$$

which is the same as (6.5).

Now assume instead that $k+1 \in S$. Then it is not possible that $k+1$ has parents in both A and B. So assume that $P_{k+1} \subseteq A \cup S$ (the proof in the B–case will be similar). As before we must have

$$X_{k+1} \perp\!\!\!\perp X_{A_k}, X_{B_k}, X_{S_k} \mid P_k(\mathbb{X})$$

and thereby also

$$X_{k+1} \perp\!\!\!\perp X_{B_k} \mid P_{k+1}(\mathbb{X}), X_{A_k}, X_{S_k}$$

such that (since $P_{k+1} \subseteq A_k, S_k$)

$$X_{k+1} \perp\!\!\!\perp X_{B_k} \mid X_{A_k}, X_{S_k}$$

Together with (6.4) this gives according to Theorem 3.18 that

$$X_{A_k}, X_{k+1} \perp\!\!\!\perp X_{B_k} \mid X_{S_k}$$

Now we move X_{k+1} to the conditioning side and use reduction. Then

$$X_{A_k} \perp\!\!\!\perp X_{B_k} \mid X_{S_k}, X_{k+1}$$

which is the same as (6.5) in the situation, where $k+1 \in S$. □

Proof of Theorem 6.5.1. Let $\mathcal{V}'$ be the subset of $\mathcal{V}$ that contains all the nodes in $A_{A \cup B \cup S}$, and let $\mathcal{E}'$ be all the edges in $\mathcal{E}$ that are connections between elements in $\mathcal{V}'$. Let $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ be the corresponding DAG. Then we must have that

$$(\mathcal{G}', X_{\mathcal{V}'}, X_{\mathcal{V}'}(P))$$

is a Bayesian network. We now define A' to be all the nodes in $\mathcal{V}'$ that are not separated from A by S. Let B' be all the nodes in $\mathcal{V}'$ that are not in A' or S. Then obviously, we have that A', B' and S are disjoint with $A \subseteq A'$, $B \subseteq B'$ and such that A' and B' are separated by S.

Then it follows from Lemma 6.5.3 that

$$X_{A'} \perp\!\!\!\perp X_{B'} \mid X_S$$

such that it by reduction follows that

$$X_A \perp\!\!\!\perp X_B \mid X_S \,.$$

□

Appendix A

Supplementary material

A.1 Measurable spaces

In this section we recall some of the main definitions and results from measure theory that are used throughout the book. Consider a set $\mathcal{X}$ and let $\mathbb{E}$ be a collection of subsets of $\mathcal{X}$.

Definition A.1.1. *We say that $\mathbb{E}$ is a σ–algebra on $\mathcal{X}$, if it holds that*

- $\mathcal{X} \in \mathbb{E}$
- *If $A \in \mathbb{E}$, then $A^c \in \mathbb{E}$*
- *If $A_1, A_2, \ldots \in \mathbb{E}$, then $\cup_{n=1}^{\infty} A_n \in \mathbb{E}$*

If $\mathcal{X}$ is some set, and $\mathbb{E}$ is a σ–algebra on $\mathcal{X}$, then we say that the pair $(\mathcal{X}, \mathbb{E})$ is a measurable space. If $\mathbb{D}$ is a collection of subsets of $\mathcal{X}$, then we define $\sigma(\mathbb{D})$ to be the smallest σ–algebra on $\mathcal{X}$ that contains $\mathbb{D}$. For a σ–algebra $\mathbb{E}$ on $\mathcal{X}$ and a collection $\mathbb{H}$ of subsets of $\mathcal{X}$, we say that $\mathbb{H}$ is a generating system for $\mathbb{E}$, if $\mathbb{E} = \sigma(\mathbb{H})$.

If it for some collection $\mathbb{H}$ of subsets of $\mathcal{X}$ holds for all $A, B \in \mathbb{H}$ that $A \cap B \in \mathbb{H}$, then we say that $\mathbb{H}$ is stable under finite intersections.

Definition A.1.2. *We say that $\mathbb{H}$ is a Dynkin class on $\mathcal{X}$, if it holds that*

1) $\mathcal{X} \in \mathbb{H}$*,*

2) If $A, B \in \mathbb{H}$ *with* $A \subseteq B$*, then* $B \setminus A \in \mathbb{H}$

3) If $A_1, A_2, \ldots \in \mathbb{H}$ *with* $A_1 \subseteq A_2 \subseteq \ldots$*, then* $\cup_{n=1}^{\infty} A_n \in \mathbb{H}$

We have

Theorem A.1.3 (Dynkin's lemma)**.** *Let* $\mathbb{D} \subseteq \mathbb{H} \subseteq \mathbb{E}$ *be collections of subsets of* $\mathcal{X}$*. assume that* $\mathbb{E} = \sigma(\mathbb{D})$ *and that* $\mathbb{D}$ *is stable under finite intersections. If furthermore* $\mathbb{H}$ *is a Dynkin class, then* $\mathbb{H} = \mathbb{E}$*.*

Definition A.1.4. *Let* $(\mathcal{X}, \mathbb{E})$ *be a measurable space. We say that a function* $\mu : \mathbb{H} \to [0, \infty]$ *is a measure (on* $(\mathcal{X}, \mathbb{H})$*), if*

1) $\mu(\emptyset) = 0$

2) If $A_1, A_2, \ldots \in \mathbb{H}$ *are pairwise disjoint sets, then* $\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$

We say that a measure μ on $(\mathcal{X}, \mathbb{E})$ is a probability measure, if $\mu(\mathcal{X}) = 1$. In the affirmative we call $(\mathcal{X}, \mathbb{E}, \mu)$ a probability space.

Theorem A.1.5 (Uniqueness theorem for probability measures)**.** *Let* μ *and* ν *be two probability measures on* $(\mathcal{X}, \mathbb{E})$*. Let* $\mathbb{H}$ *be a generating system for* $\mathbb{E}$ *which is stable under finite intersection. If* $\mu(A) = \nu(A)$ *for all* $A \in \mathbb{H}$*, then* $\mu(A) = \nu(A)$ *for all* $A \in \mathbb{E}$*.*

Let $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ be two measurable spaces. Then we can consider the product space $(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$. Here the product σ–algebra $\mathbb{E} \otimes \mathbb{K}$ is generated by the system of all product sets

$$\mathbb{D} = \{A \times B \,:\, A \in \mathbb{E},\, B \in \mathbb{K}\}$$

Note that $\mathbb{D}$ is stable under intersections. If λ and $\tilde{\lambda}$ are two measures on $(\mathcal{X} \times \mathcal{Y}, \mathbb{E} \otimes \mathbb{K})$ that are equal on product sets

$$\lambda(A \times B) = \tilde{\lambda}(A \times B)$$

for all $A \in \mathbb{E}$ and $B \in \mathbb{K}$, then according to Theorem A.1.5 we have $\lambda = \tilde{\lambda}$.

Let μ be a measure on $(\mathcal{X}, \mathbb{E})$ and μ be a measure on $(\mathcal{Y}, \mathbb{K})$. Then $\mu \otimes \nu$ denotes the uniquely determined measure defined by $(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$.

Theorem A.1.6 (Tonelli's theorem)**.** *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and ν be probability measure on $(\mathcal{Y}, \mathbb{K})$, and assume that f is nonnegative and $\mathbb{E} \otimes \mathbb{K}$ measurable. Then*

$$\int f(x,y)\,\mathrm{d}(\mu \otimes \nu)(x,y) = \iint f(x,y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x).$$

Theorem A.1.7 (Fubini's theorem)**.** *Let μ be a probability measure on $(\mathcal{X}, \mathbb{E})$ and ν be probability measure on $(\mathcal{Y}, \mathbb{K})$, and assume that f is $\mathbb{E} \otimes \mathbb{K}$ measurable and $\mu \otimes \nu$ integrable. Then $y \mapsto f(x,y)$ is integrable with respect to ν for μ-almost all x, the set where this is the case is measurable, and it holds that*

$$\int f(x,y)\,\mathrm{d}(\mu \otimes \nu)(x,y) = \iint f(x,y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x).$$

We will also need the following abstract change-of-variable theorem

Theorem A.1.8. *Let mu be a measure on $(\mathcal{X}, \mathbb{E})$ and let $(\mathcal{Y}, \mathbb{K})$ be some other measurable space. Let $t : XX \to YY$ be measurable, and let $f : YY \to \mathbb{R}$ be Borel measurable. Then f is $t(\mu)$-integrable if and only if $f \circ t$ is μ-integrable, and in the affirmative, it holds that $\int f\,\mathrm{d}t(\mu) = \int f \circ t\,\mathrm{d}\mu$.*

Let again $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$ be two measurable spaces. Define the inclusion map $i_x : \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ by

$$i_x(y) = (x,y) \qquad \text{for } y \in \mathcal{X}\,.$$

Then i_x is $\mathbb{K} - \mathbb{E} \otimes \mathbb{K}$–measurable for each fixed $y \in \mathcal{Y}$. For $G \in \mathbb{E} \otimes \mathbb{K}$ define

$$G^x = \{y \in \mathcal{Y} \ : \ (x,y) \in G\} = i_x^{-1}(G)$$

Note that G^x is $\mathbb{K}$–measurable due to the measurability of i_x.

A.2 Random variables and conditional expectations

Assume that $(\Omega, \mathbb{F}, P)$ is a probability space and $(\mathcal{X}, \mathbb{E})$ is some measurable space. We say that $X : \Omega \to \mathcal{X}$ is a random variable on $(\Omega, \mathbb{F})$ with values in $(\mathcal{X}, \mathbb{E})$, if it $\mathbb{F} - \mathbb{E}$–measurable. That is

$$X^{-1}(A) = (X \in A) \in \mathbb{F}$$

for all $A \in \mathbb{E}$. For a random variable X on $(\Omega, \mathbb{F})$ with values in $(\mathcal{X}, \mathbb{E})$ we define $\sigma(X)$ to be the smallest σ–algebra that makes X measurable. Then $\sigma(X)$ is the sub σ–algebra of $\mathbb{F}$ given by

$$\sigma(X) = \{(X \in A) \, : \, A \in \mathbb{E}\}$$

We have the following extremely useful result

Theorem A.2.1. *Assume that X is a random variable with values in $(\mathcal{X}, \mathbb{E})$ and that Z is a real–valued random variable.Then Z is $\sigma(X)$–measurable if and only if there exists a measurable function $\phi : (\mathcal{X}, \mathbb{E}) \to (\mathbb{R}, \mathbb{B})$ such that*

$$Z = \phi \circ X$$

Let $\mathbb{D}$ be a sub σ–algebra of $\mathbb{F}$.

Definition A.2.2. *Let X be a real random variable defined on $(\Omega, \mathbb{F}, P)$ with $E|X| < \infty$. A conditional expectation of X given $\mathbb{D}$ is a real random variable denoted $E(X \mid \mathbb{D})$ that satisfies*

1) $E(X \mid \mathbb{D})$ is $\mathbb{D}$–measurable

2) $E|E(X \mid \mathbb{D})| < \infty$

3) For all $D \in \mathbb{D}$ it holds

$$\int_D E(X \mid \mathbb{D}) \mathrm{d}P = \int_D X \mathrm{d}P$$

We uniqueness of conditional expectations

Theorem A.2.3. *(1) If U and $\widetilde{U}$ are both conditional expectations of X given $\mathbb{D}$, then $U = \widetilde{U}$ a.s.*

(2) If U is a conditional expectation of X given $\mathbb{D}$ and $\widetilde{U}$ is $\mathbb{D}$-measurable with $\widetilde{U} = U$ a.s. then $\widetilde{U}$ is also a conditional expectation of X given $\mathbb{D}$.

and existence

Theorem A.2.4. *If X is a real random variable with $E|X| < \infty$, then there exists a conditional expectation of X given $\mathbb{D}$.*

Furthermore we have a series of nice properties. Let X, X_n and Y be real random variables, all of which are integrable.

Theorem A.2.5. *(1) If $X = c$ a.s., where $c \in \mathbb{R}$ is a constant, then $E(X|\mathbb{D}) = c$ a.s.*

(2) For $\alpha, \beta \in \mathbb{R}$ it holds that

$$E(\alpha X + \beta Y|\mathbb{D}) = \alpha E(X|\mathbb{D}) + \beta E(Y|\mathbb{D}) \text{ a.s.}$$

(3) If $X \geq 0$ a.s. then $E(X|\mathbb{D}) \geq 0$ a.s. If $Y \geq X$ a.s. then $E(Y|\mathbb{D}) \geq E(X|\mathbb{D})$ a.s.

(4) If $\mathbb{D} \subseteq \mathbb{D}$ are sub σ-algebras of $\mathbb{F}$ then

$$E(X|\mathbb{D}) = E[E(X|\mathbb{E})|\mathbb{D}] = E[E(X|\mathbb{D})|\mathbb{E}] \text{ a.s.}$$

(5) If $\sigma(X)$ and $\mathbb{D}$ are independent then

$$E(X|\mathbb{D}) = EX \text{ a.s.}$$

(6) If X is $\mathbb{D}$-measurable then

$$E(X|\mathbb{D}) = X \text{ a.s.}$$

(7) If it holds for all $n \in \mathbb{N}$ that $X_n \geq 0$ a.s. and $X_{n+1} \geq X_n$ a.s. with $\lim X_n = X$ a.s., then

$$\lim_{n\to\infty} E(X_n|\mathbb{D}) = E(X|\mathbb{D}) \text{ a.s.}$$

(8) If X is $\mathbb{D}$-measurable and $E|XY| < \infty$, then

$$E(XY|\mathbb{D}) = X\, E(Y|\mathbb{D}) \text{ a.s.}$$

(9) If $f : \mathbb{R} \to \mathbb{R}$ is a measurable function that is convex on an interval I, such that $P(X \in I) = 1$ and $E|f(X)| < \infty$, then it holds that

$$f\big(E(X|\mathbb{D})\big) \leq E\big(f(X)|\mathbb{D})\big) \text{ a.s.}$$

Now assume that X is a random variable with values in $(\mathcal{X}, \mathbb{E})$ and that Y is a real random variable with $E|Y| < \infty$. If we are looking for the conditional expectation of Y given $\mathbb{D} = \sigma(X)$, then we write $E(Y \mid X)$ rather than $E(Y \mid \sigma(X))$. The resulting random variable is referred to as the conditional expectation of Y given X. We have

Theorem A.2.6. *Let Y be a real random variable with $E|Y| < \infty$, and assume that X is a random variable with values in $(\mathcal{X}, \mathbb{E})$. Then the conditional expectation $E(Y \mid X)$ of Y given Y is characterised by*

1) $E(Y \mid X)$ is $\sigma(X)$-measurable

2) $E|E(Y \mid X)| < \infty$

3) For all $A \in \mathbb{E}$ it holds that

$$\int_{(X \in A)} E(Y \mid X) \mathrm{d}P = \int_{(X \in A)} Y \mathrm{d}P$$

According to Theorem A.2.1 there exists a measurable map $\phi : \mathcal{X} \to \mathbb{R}$ such that

$$E(Y \mid X) = \phi(X)$$

Appendix B

Hints for exercises

B.1 Hints for chapter 1

Hints for exercise 1.1. Use Theorem 1.3. Show that the conditional distribution is a hypergeometric distribution and find the parameters. ∘

Hints for exercise 1.3.

(1) Use the extended Tonelli to calculate an integral with respect to $(X, Y)(P)$ as a double integral.

(2) P_x has density

$$f_x(y) = \frac{1}{x} e^{-y/x} \qquad \text{for } y > 0$$

with respect to the Lebesgue measure. Use question (1) to find EY.

∘

Hints for exercise 1.10.

(2) Recall that according to Theorem 1.5.1 there exists a $\mathbb{E}-\mathbb{B}$–measurable function ϕ_F with values in $[0, 1]$ that satisfies (1.4) in the notes. Use this function in the construction.

◦

B.2 Hints for chapter 2

Hints for exercise 2.2. Get inspiration from the proof of Theorem 2.2.1! ◦

Hints for exercise 2.6.

(1) Write

$$P(Z=1) = P(U \leq \frac{f(Y)}{cg(Y)}) = E(P(U \leq \frac{f(Y)}{cg(Y)}|Y)) = ...$$

◦

Hints for exercise 2.7.

(4) Use that (U_1, U_2) is uniform on $(0,1)^2$, such that for $A \subseteq (0,1)^2$ the probability $P((U_1, U_2) \in A)$ is simply the area of A.

(8) You should show that

$$\int_A P_y(B)\, X_{(1)}(P)(dy) = P(X_{(1)} \in A, X_{(2)} \in B)$$

Use the change–of–variable theorem to obtain

$$\begin{aligned}\int_A P_{x_1}(B)\, X_{(1)}(P)(dx_1) = &\int_{(X_1 \in A)} 1_{(X_1 < X_2)} \frac{\nu(B \cap (X_1, \infty))}{\nu((X_1, \infty))} dP \\ &+ \int_{(X_2 \in A)} 1_{(X_2 < X_1)} \frac{\nu(B \cap (X_2, \infty))}{\nu((X_2, \infty))} dP\end{aligned}$$

Use the change–of–variable theorem again (integrate with respect to $(X_1, X_2)(P)$) and use Tonelli (since $X_1 \perp\!\!\!\perp X_2$).

◦

Hints for exercise 2.9.

(2) Simply use that the points $(U_1^1, U_1^2), \ldots, (U_N^1, U_N^2)$ are independent and e.g. have probability $m_2(A_j)$ to be in A_j.

(3) Show the (unconditional) probability

$$P(N(A_1) = n_1, \ldots, N(A_m) = n_m) = \prod_{j=1}^{m} \frac{(\lambda m_2(A_j))^{n_j}}{n_j!} e^{-\lambda m_2(A_j)}$$

(4) Use the points $(U_1^1, U_1^2), \ldots, (U_N^1, U_N^2)$ and then remove some of them with a probability that depends on the value of $k(U_k^1, U_k^2)$. Get inspiration from Exercise 2.6...

∘

Hints for exercise 2.10.

(3) Use (1) and (2) to calculate $E(S_n \mid S_n)$.

∘

B.3 Hints for chapter 3

Hints for exercise 3.1.

(2) Use Theorem 3.3.6.

(3) Use e.g. the extended Tonelli.

(4) Use (1) and (2).

∘

Hints for exercise 3.2.

(2) Write

$$P(X_1 \in A, X_2 \in B, X_1 + X_2 \in C) = P(X_1 \in A, (X_1 + X_2) - X_1 \in B, X_1 + X_2 \in C)$$

and write it as an integral with respect to $(X_1, X_1 + X_2)$. Use the extended Tonelli.

(3) For the first result, write $B = x - A$ and use Theorem 3.5.3 to find an alternative expression for R_x. For the second result, you can e.g. consider the distribution function for P_x and realise that it only has the values 0 and 1, such that it must have exactly one jump (and this must be of size 1).

(4) Calculate $E(X_1 \mid X_1 + X_2)$.

(5) You should use $\phi_1 = \phi$. Calculate the probability $P(X_1 = \phi(X_1 + X_2))$ as an integral with respect to (X_1, X_1+X_2) and use the extended Tonelli. You now have an expression for P_x...

(6) Choose X_3 such that $X_1 + X_2 + X_3$ is particularly simple.

∘

Hints for exercise 3.3.

(2) Let for $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$

$$p_{ij} = P(Y = j \mid X = i)$$

These conditional probabilities are simply the point probabilities in the conditional distribution of Y given X. Use (1) for each value of i.

∘

Hints for exercise 3.4. Recall that if U is uniform on $(0, 1)$, then $-\beta \log(1 - U)$ has an exponential distribution. ∘

B.4 Hints for chapter 4

Hints for exercise 4.1. Use Theorem 4.1.5 and reduction. ∘

Hints for exercise 4.3.

(3) The Strong Law of Large Numbers may be useful: If $X_1, X_2, \ldots$ are iid with $E|X_1| < \infty$,

then

$$\frac{1}{n}\sum_{k=1}^{n} X_k \to EX_1 \quad P\text{–a.s.}$$

(5) Rewrite $P(Y_1 = m_1, \ldots, Y_n = m_n)$ to an event concerning the Z–variables.

(6) Show that both of the processes $N_0, N_1, \ldots$ and $M_0, M_1, \ldots$ only have jumps at the times $T_1, T_2, \ldots$.

∘

Hints for exercise 4.4.

(1) Rewrite the set $(\tau + k = n) = (\tau = n - k)$.

(3) You should show that for all $n \in \mathbb{N}_0$

$$\big(X_{\tau+n+1}, \tau + n + 1\big) \perp\!\!\!\perp \big(X_\tau, \tau, X_{\tau+1}, \tau + 1, \ldots, X_{\tau+n}, \tau + n\big) \mid X_{\tau+n}, \tau + n$$

Use Theorem 4.2.7 with the stopping time $\tau + n$, and move some information around...

∘

Hints for exercise 4.5.

(1) Use Theorem 4.3.6 for the stopping times $\tau + k$.

(2) See the remark after the proof of Theorem 4.3.5.

∘

Hints for exercise 4.6.

(2) Show that $P(U_1 \in A_1, \ldots, U_n \in A_n) = P(U_1 \in A_1) \cdots P(U_n \in A_n)$ for all $n \in \mathbb{N}$. Divide according to the value of τ, and use that e.g. U_{k+1}^1 is independent of $(U_1 \in A_1, \ldots, U_k \in A_k) \cap (\tau = k)$.

(3) Realise that $X_{n+1} = \phi(X_n, U_{n+1})$.

∘

Hints for exercise 4.7.

(1) Write

$$X_{\sigma_B} = \sum_{k=0}^{\infty} X_{\tau_A+k} 1_{(\sigma_B=\tau_A+k)}$$

and write $(\sigma_B = \tau_A + k)$ as a set involving $X_{\tau_A+1}, \ldots, X_{\tau_A+k}$.

(3) Argue that

$$(X_1, X_2, \ldots) \mid X_0 \quad \stackrel{\mathcal{D}}{=} \quad (X_{\tau_k+1}, X_{\tau_k+2}, \ldots) \mid X_{\tau_k}$$

and that for some function ψ, we have

$$X_{\tau_1} = \psi(X_1, X_2, \ldots) \quad \text{and} \quad X_{\tau_{k+1}} = \psi(X_{\tau_k+1}, X_{\tau_k+2}, \ldots)$$

∘

Hints for exercise 4.8.

(1) You should show that

$$X_{\tau\wedge(n+1)} \perp\!\!\!\perp (X_{\tau\wedge 0}, \ldots, X_{\tau\wedge(n-1)}) \mid X_{\tau\wedge n}$$

Use the strong Markov property to obtain

$$(X_{\tau\wedge n}, X_{\tau\wedge n+1}) \perp\!\!\!\perp \mathbb{F}_{\tau\wedge n} \mid X_{\tau\wedge n}$$

Firstly, use that $(X_{\tau\wedge 0}, \ldots, X_{\tau\wedge(n-1)})$ is $\mathbb{F}_{\tau\wedge n}$–measurable. Then write

$$X_{\tau\wedge(n+1)} = X_{\tau\wedge n+1} 1_{(\tau>n)} + X_{\tau\wedge n} 1_{(\tau\leq n)} \tag{B.1}$$

Obtain e.g. $1_{(\tau\leq n)}$ as a function of $X_{\tau\wedge n}$.

(2) Use (B.1) and the substitution theorem. Show that the conditional distribution of Y_{n+1} given Y_n is $(Q_x)_{x\in\mathcal{X}}$, where

$$Q_x = \begin{cases} \delta_x & x \in A \\ P_x & x \notin A \end{cases}$$

∘

B.5 Hints for chapter 5

Hints for exercise 5.1.

(2) Standard approximation argument using indicator functions.

(3) For the last equality recall/use that $P_{X_n}(A)$ is a version of $P(X_{n+1} \in A \mid X_n)$. Use the Markov property.

(4) Hint: Use that $(\tau > n)$ is $\mathbb{F}_n$–measurable.

(6) See example 4.3.8.

(7) Use The Strong Law of Large Numbers (applied to both $\frac{1}{N}\sum_{k=0}^{\tau_N-1} f(X_n)$ and τ_N/N).

∘

Hints for exercise 5.2.

(1) Show and use that

$$X_n = \rho^n X_0 + \rho^{n-1}\epsilon_1 + \rho^{n-2}\epsilon_2 + \cdots + \rho\epsilon_{n-1} + \epsilon_n$$

(2) If f is a bounded measurable function and $\nu = \mathcal{N}(\mu, \sigma^2)$, then

$$\int f\, d\nu = \int f(x) \frac{1}{\sqrt{2\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2 \Big) \mathrm{d}\lambda(x)\,,$$

where λ is the Lebesgue measure. Use Dominated convergence and that

$$\sum_{n=0}^{\infty} \rho^{2n} = \frac{1}{1-\rho^2}\,.$$

∘

Hints for exercise 5.3.

(2) Assume for contradiction that a stationary initial distribution μ exists. Find $N \in \mathbb{N}$

such that $P(X_0 \in [-N, N]) > 0$. Then for all $n \in \mathbb{N}$

$$\begin{aligned} 0 < P(X_0 \in [-N, N]) &= P(X_n \in [-N, N]) \\ &= \cdots \\ &= \int_{\mathbb{R}} \int_{[-N,N]} k_x^{(n)}(y) \, \mathrm{d}\lambda(y) \mathrm{d}\mu(x) \, . \end{aligned}$$

Let $n \to \infty$ and get 0.

∘

Hints for exercise 5.4.

(1) Simply use the integral representation of $\mathcal{P}_x^n(A_1 \times \cdots \times A_n)$.

(4) Consider the conditional likelihood function as a function of ρ and find maximum.

(5) In Example 5.3.6 (and almost also in Exercise 5.2) it was shown that the transition densities are asymptotically stable and that $\mathcal{N}(0, 1/(1-\rho^2))$ is the stationary initial distribution. Let f_0 be the density for this distribution. Use Theorem 5.3.4 to obtain that e.g.

$$\frac{1}{n} \sum_{i=0}^{n-1} X_i X_{i+1} \to E_{f_0 \cdot \lambda}(X_0 X_1) \quad \text{a.s.} \, ,$$

where $E_{f_0 \cdot \lambda}$ means expectation in a model, where $X_0 \sim \mathcal{N}(0, 1/(1-\rho^2))$.

∘

Hints for exercise 5.5. A standard argument using transformation of densities. ∘

Hints for exercise 5.6.

(4) Use that there must exist $\alpha < 1$ and $K < \infty$ such that

$$\frac{\xi(x)^2 + \sigma(x)^2}{x^2} < \alpha \qquad \text{for } |x| > K$$

Also use that for all compact sets B we must (?) have

$$\inf_{(x,y) \in [-r,r] \times B} k_x(y) > 0 \, ,$$

since $(x, y) \to k_x(y)$ is continuous.

(5) Use induction and (2).

∘

Hints for exercise 5.8.

(1) You could show that

$$\int_0^y k_x(y)\,dx = P(|x - U_1| \leq y)\,.$$

Use substitution...

(3) Use (2) and the drift function $V(x) = e^{\beta x}$. Let $\alpha = \int_0^\infty e^{-\beta y} f(y)\,dy$.

(4) Use Corollary 5.5.2 and that for all $n \in \mathbb{N}$ exists K_n such that $x^n \leq e^{\beta x} + K_n$.

(5) Use that $EX_0^2 = EX_1^2$.

(6) Let X_0 have density e^{-x} and find the density of X_1.

∘

Hints for exercise 5.9. Consider a constant drift function. ∘

Hints for exercise 5.10.

(2) You should simply show, that

$$\begin{aligned} &P(X_1 \in A_1, Y_1 \in B_1, X_2 \in A_2, Y_2 \in B_2) \\ &= \int_{A_1 \times B_1} \Big(\int_{A_2 \times B_2} k_{(x_1,y_1)}(x_2, y_2)\,\mathrm{d}(\nu_1 \otimes \nu_2)(x_2, y_2) \Big)\,\mathrm{d}(X_1, Y_1)(P)(x_1, y_1) \end{aligned}$$

(3) You should show that

$$f_0(x_2, y_2) = P^{*1}(f_0) = \iint k_{(x_1,y_1)}(x_2, y_2) f_0(x_1, y_1) \mathrm{d}\nu_1(x_1)\,\mathrm{d}\nu_2(y_1)$$

For this, first show that

$$\int f_0(x_1, y_1) \mathrm{d}\nu_1(x_1) f^1_{y_1}(x_2) = f_0(y_1, x_2)$$

(Recall, that f_0 is the density for (X, Y). It can be helpful to invent the notation $f_{0,1}$ for the marginal density of X and $f_{0,2}$ for the marginal density of Y).

(4) Find the transition density $k_{(x_1,y_1)}(x_2, y_2)$ (this will not depend on x_1). Use 5.9, since infimum over $(x_1, y_1) \in [0, 1] \times \{0, 1\}$ is extremely simple.

$\circ$

Index