

3 Stochastic Processes and Markov Chains (†)

The measure-theoretic probability in the previous chapter covers almost all the definitions required. Occasionally, however, we make use of infinite sequences of random variables and for these one requires just a little more machinery. The purpose of this chapter is to describe this machinery. We expect most readers will skip this chapter on the first reading, perhaps referring to it when necessary.

A basic problem in connection to infinite sequences of random variables asks whether they exist at all if we start putting constraints on their joint distribution. The simplest of these constraints asks for that the sequence should be independent, meaning that any finite subcollection of the random variables in the sequence should be independent. We may also ask for that each random variable share the same distribution. The first theorem we cite answers the question positively. A consequence of this theorem is that we can write, for example, “let X_1, X_2, \dots be an infinite sequence of independent standard Gaussian random variables” and be comfortable knowing there exists a probability space on which these random variables can be defined. To state the theorem we need the concept of **Borel spaces**, which we introduce next.

Let λ be the Lebesgue measure on $([0, 1], \mathfrak{B}([0, 1]))$. Two measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ are said to be **isomorphic** if there exists a bijective function $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that f is \mathcal{F}/\mathcal{G} -measurable and f^{-1} is \mathcal{G}/\mathcal{F} -measurable. A **Borel space** is a measurable space $(\mathcal{X}, \mathcal{F})$ that is isomorphic to $(A, \mathfrak{B}(A))$ with $A \in \mathfrak{B}(\mathbb{R})$ a Borel measurable subset of the of the reals. The spaces in which the random elements live are all Borel. These include \mathbb{R}^n for any $n \in \mathbb{N}^+$ and its measurable subsets.

THEOREM 3.1 *Let μ be a probability measure on a Borel measurable space \mathcal{S} . Then there exists a sequence of independent random elements X_1, X_2, \dots on $([0, 1], \mathfrak{B}([0, 1]), \lambda)$ such that the law $\lambda_{X_t} = \mu$ for all t .*

We give a sketch of the proof because, although it is not really relevant for the material in this book, it illustrates the general picture and dispels some of the mystic about what is really going on. Exercise 3.1 asks you to provide the missing steps from the proof.

Proof sketch of Theorem 3.1 For simplicity we consider only the case that $\mathcal{S} = ([0, 1], \mathfrak{B}([0, 1]))$ and μ is the Lebesgue measure. For any $x \in [0, 1]$ let $F_1(x), F_2(x), \dots$ be the binary expansion of x , which is the unique binary-valued

infinite sequence such that

$$x = \sum_{t=1}^{\infty} F_t(x)2^{-t}.$$

We can view F_1, F_2, \dots as (binary valued) random variables over the probability space $([0, 1], \mathfrak{B}([0, 1]), \lambda)$. Viewed as such, a direct calculation shows that F_1, F_2, \dots are independent. From this we can create an infinite sequence of uniform random variables by reversing the process. To do this we rearrange the $(F_t)_{t=1}^{\infty}$ sequence into a grid. For example:

$$\begin{array}{l} F_1, F_2, F_4, F_7, \dots \\ F_3, F_5, F_8, \dots \\ F_6, F_9, \dots \\ F_{10}, \dots \\ \vdots \end{array}$$

Letting $X_{m,t}$ be the t th entry in the m th row of this grid, we define $X_m = \sum_{t=1}^{\infty} 2^{-t} X_{m,t}$ and again one can easily check that with this choice the sequence X_1, X_2, \dots is independent and $\lambda_{X_t} = \mu$ is uniform for each t . \square

3.1 Stochastic processes

Let \mathcal{T} be an arbitrary set. A **stochastic process** on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a collection of random variables $\{X_t : t \in \mathcal{T}\}$. In this book \mathcal{T} will always be countable and so in the following we restrict ourselves to $\mathcal{T} = \mathbb{N}$. The first theorem is not the most general, but suffices for our purposes and is more easily stated than more generic alternatives. It is also a theorem that guarantees the existence of an infinite sequence of random variables.

THEOREM 3.2 *For each $n \in \mathbb{N}^+$ let $(\Omega_n, \mathcal{F}_n)$ be a Borel space and μ_n be a measure on $(\Omega_1 \times \dots \times \Omega_n, \mathcal{F}_1 \otimes \dots \otimes \mathcal{F}_n)$ and assume that μ_n and μ_{n+1} are related through*

$$\mu_{n+1}(A \times \Omega_{n+1}) = \mu_n(A) \quad \text{for all } A \in \Omega_1 \otimes \dots \otimes \Omega_n. \quad (3.1)$$

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random elements X_1, X_2, \dots with $X_t : \Omega \rightarrow \Omega_t$ such that $\mathbb{P}_{X_1, \dots, X_n} = \mu_n$ for all n .



Sequences of measures $(\mu_n)_n$ satisfying Eq. (3.1) are called **projective**.

Theorem 3.1 follows immediately from Theorem 3.2. By assumption a random variable takes values in $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, which is Borel. Then let $\mu_n = \otimes_{t=1}^n \mu$ be the n -fold product measure of μ with itself. That this sequence of measures is projective is clear and the theorem does the rest.

3.2 Markov chains

A Markov chain is an infinite sequence of random elements X_1, X_2, \dots where the conditional distribution of X_{t+1} given X_1, \dots, X_t is the same as the conditional distribution of X_t given X_{t-1} . The sequence has the property that given the last element, the history is irrelevant to ‘predict’ the future. Such random sequences appear throughout probability theory and have many applications besides. The theory is much too rich to explain in detail, so we give the basics and point towards the literature for more details at the end. The focus here is mostly on the definition and existence of Markov chains.

Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ be measurable spaces. A **probability kernel** or **Markov kernel** between $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ is a function $K : X \times \mathcal{G} \rightarrow [0, 1]$ such that:

- (a) $K(x, \cdot)$ is a measure for all $x \in \mathcal{X}$.
- (b) $K(\cdot, A)$ is \mathcal{F} -measurable for all $A \in \mathcal{G}$.

The idea here is that K describes a stochastic transition. Having arrived at x , a process’s next state is sampled $Y \sim K(x, \cdot)$. If K_1 is a $(\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Y}, \mathcal{G})$ probability kernel and K_2 is a $(\mathcal{Y}, \mathcal{G}) \rightarrow (\mathcal{Z}, \mathcal{H})$ probability kernel, then the **product kernel** $K = K_1 \otimes K_2$ is the probability kernel from $(\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Z}, \mathcal{H})$ defined by

$$K(x, A) = \int_{\mathcal{Y}} K_2(y, A) K_1(x, dy), \quad (3.2)$$

for which an alternate notation is to write $K(x, dz) = \int_{\mathcal{Y}} K_2(y, dz) K_1(x, dy)$. Note the ‘ d ’ appearing inside the measure rather than outside: This is helpful as it shows what variable is being integrated over. Occasionally one sees the notation $K_x(A)$ rather than $K(x, A)$, in which case the notation $dK_x(y)$ would make more sense. The symbol ‘ dz ’ can be thought of indicating the potential to integrate over z : Multiplying both sides by the indicator of A , and integrating over A , we get (3.2).

The product kernel corresponds to taking one step using K_1 followed by a step from K_2 so that $Y \sim K_1(x, \cdot)$ and then $Z \sim K_2(Y, \cdot)$. The counterpart of Theorem 3.2 for Markov chains is known as the Ionescu Tulcea theorem, which we state next:

THEOREM 3.3 *For each $n \in \mathbb{N}^+$ let $(\Omega_n, \mathcal{F}_n)$ be a measurable space and K_n be a probability kernel from $\prod_{t=1}^{n-1} \Omega_t \rightarrow \Omega_n$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random elements X_1, X_2, \dots with $X_t : \Omega \rightarrow \Omega_t$ such that $\mathbb{P}_{X_1, \dots, X_n} = \bigotimes_{t=1}^n K_t$ for all $n \in \mathbb{N}^+$.*

A **homogeneous Markov chain** is a sequence of random elements X_1, X_2, \dots taking values in **state space** $\mathcal{S} = (\mathcal{X}, \mathcal{F})$ and with

$$\mathbb{E}[X_{t+1} \in \cdot \mid X_1, \dots, X_t] = \mathbb{E}[X_{t+1} \in \cdot \mid X_t] = \mu(X_t, \cdot) \quad \text{almost surely,}$$

where μ is a probability kernel from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{X}, \mathcal{F})$ and we assume that $\mathbb{E}[X_1 \in \cdot] = \mathbb{P}(X_1 \in \cdot) = \mu_0(\cdot)$ for some measure μ_0 on $(\mathcal{X}, \mathcal{F})$.



The word ‘homogeneous’ refers to the fact that the probability kernel does not change with time. Accordingly, sometimes one writes time-homogeneous instead of homogeneous. The reader can no doubt see how to define a Markov chain where μ depends on t , though doing so is purely cosmetic since the state-space can always be augmented to include a time component.

Note that if $\mu(x | \cdot) = \mu_0(\cdot)$ for all $x \in \mathcal{X}$, then Theorem 3.3 is yet another way to prove the existence of an infinite sequence of independent and identically distributed random variables. The basic questions in Markov chains resolve around understanding the evolution of X_t in terms of the probability kernel. For example, assuming that $\Omega_t = \Omega_1$ for all $t \in \mathbb{N}^+$, does the law of X_t converge to some fixed distribution as $t \rightarrow \infty$ and if so, how fast is this convergence? For now we make do with the definitions, but in the special case that \mathcal{X} is finite we will discuss some of these topics much later in Chapters 36 and 37.

3.3 Martingales and stopping times

Let X_1, X_2, \dots be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{F} = (\mathcal{F}_t)_{t=1}^n$ a filtration of \mathcal{F} and where we allow $n = \infty$. Recall that the sequence $(X_t)_{t=1}^n$ is \mathbb{F} -adapted if X_t is \mathcal{F}_t -measurable for all $1 \leq t \leq n$.

DEFINITION 3.1 A \mathbb{F} -adapted sequence of random variables $(X_t)_{t \in \mathbb{N}_+}$ is a \mathbb{F} -adapted **martingale** if

- (a) $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = X_{t-1}$ almost surely for all $t \in \{2, 3, \dots\}$.
- (b) X_t is integrable

If the equality is replaced with a less-than (greater-than), then we call $(X_t)_t$ a **supermartingale** (respectively, a **sub-martingale**).



The time index t need not run over \mathbb{N}^+ . Very often t starts at zero instead. Martingales can also be defined in continuous time, but we have no need for these here.

EXAMPLE 3.1 A gambler repeatedly throws a coin, winning a dollar for each heads and losing a dollar for each tails. Their total winnings over time is a martingale. To model this situation let Y_1, Y_2, \dots be a sequence of independent Rademacher distributions, which means that $\mathbb{P}(Y_t = 1) = \mathbb{P}(Y_t = -1) = 1/2$. The winnings after t rounds is $S_t = \sum_{s=1}^t Y_s$, which is a martingale adapted to the filtration $(\mathcal{F}_t)_{t=1}^\infty$ given by $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$. The definition of super/sub-martingales (the direction of inequality) can be remembered by remembering that the definition favors the casino, not the gambler.

DEFINITION 3.2 Let $\mathbb{F} = (\mathcal{F}_t)_{t \in \mathbb{N}}$ be a filtration. A random variable τ with values in $\mathbb{N} \cup \{\infty\}$ is a **stopping time** with respect to \mathbb{F} if $\mathbb{I}\{\tau \leq t\}$ is \mathcal{F}_t -measurable for all $t \in \mathbb{N}$. The σ -algebra at stopping time τ is

$$\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty : A \cap \{\tau \leq t\} \in \mathcal{F}_t \text{ for all } t\}.$$



We often just write τ is a \mathbb{F} -stopping time. The filtration is omitted when the underlying filtration is obvious from the context. This is also true for martingales.

A stopping time τ is a random variable that determines when a process stops and that only depends on information available at time τ , which means it cannot ‘peak into the future’ to determine stopping. Using the interpretation of σ -algebras encoding information, if $(\mathcal{F}_t)_t$ is thought of as the knowledge available at time t , \mathcal{F}_τ is the information available at the random time τ . Exercise 3.6 asks you to explore properties of stopped σ -algebras; amongst other things, it asks you to show that \mathcal{F}_τ is in fact a σ -algebra.

EXAMPLE 3.2 In the gambler example, the first time when the gambler’s winnings hits 100 is a stopping time: $\tau = \min\{t : S_t = 100\}$. On the other hand, $\tau = \min\{t : S_{t+1} = -1\}$ is not a stopping time because $\mathbb{I}\{\tau = t\}$ is not \mathcal{F}_t -measurable.

A fundamental result of Doob shows that the expectation of martingales do not change when they are randomly stopped, as long as there is no peeking into the future.

THEOREM 3.4 (Doob’s optional stopping) *Let $\mathbb{F} = (\mathcal{F}_t)_{t \in \mathbb{N}}$ be a filtration and $(X_t)_{t \in \mathbb{N}}$ be an \mathbb{F} -adapted martingale and τ an \mathbb{F} -stopping time such that at least one of the following holds:*

- (a) *There exists an $n \in \mathbb{N}$ such that $\mathbb{P}(\tau > n) = 0$.*
- (b) *$\mathbb{E}[\tau] < \infty$ and there exists a constant $c \in \mathbb{R}$ such that for all $t \in \mathbb{N}$, $\mathbb{E}[|X_{t+1} - X_t| \mid \mathcal{F}_t] \leq c$ almost surely on the event that $\tau > t$.*
- (c) *There exists a constant c such that $|X_{t \wedge \tau}| \leq c$ almost surely for all $t \in \mathbb{N}$.*

Then X_τ is almost-surely well defined and $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$. Furthermore, when (X_t) is a super/sub-martingale rather than a martingale, then equality is replaced with less/greater-than, respectively.

One application of Doob’s optional stopping theorem is a useful and apriori surprising generalization of Markov’s inequality to nonnegative supermartingales.


THEOREM 3.5 (Maximal inequality) *Let $(X_t)_{t \in \mathbb{N}}$ be a supermartingale with $X_t \geq 0$ almost surely for all t . Then*

$$\mathbb{P}\left(\sup_{t \in \mathbb{N}} X_t \geq \varepsilon\right) \leq \frac{\mathbb{E}[X_1]}{\varepsilon}.$$

Proof Let A_n be the event that $\sup_{t \leq n} X_t \geq \varepsilon$ and $\tau = (n + 1) \wedge \min\{t \leq n : X_t \geq \varepsilon\}$, where the minimum of an empty set is assumed to be infinite so that $\tau = n + 1$ if $X_t < \varepsilon$ for all $t \in [n]$. Clearly τ is a stopping time and $\mathbb{P}(\tau \leq n + 1) = 1$. Then by Theorem 3.4 and elementary calculation,

$$\mathbb{E}[X_1] \geq \mathbb{E}[X_\tau] \geq \mathbb{E}[X_\tau \mathbb{I}\{\tau \leq n\}] \geq \mathbb{E}[\varepsilon \mathbb{I}\{\tau \leq n\}] = \varepsilon \mathbb{P}(\tau \leq n) = \varepsilon \mathbb{P}(A_n),$$

where the second inequality uses the definition of the stopping time and the nonnegativity of the supermartingale. Rearranging shows that $\mathbb{P}(A_n) \leq \mathbb{E}[X_1]/\varepsilon$ for all $n \in \mathbb{N}$. Since $A_1 \subseteq A_2 \subseteq \dots$ it follows that $\mathbb{P}(\sup_{t \in \mathbb{N}} X_t \geq \varepsilon) = \mathbb{P}(\cup_{n \in \mathbb{N}} A_n) \leq \mathbb{E}[X_1]/\varepsilon$. \square

 Markov's inequality (which we will cover in the next chapter) combined with the definition of a supermartingale shows that $\mathbb{P}(X_n \geq \varepsilon) \leq \mathbb{E}[X_1]/\varepsilon$. In fact, in the above we have effectively applied Markov's inequality to the random variable X_τ . The maximal inequality is a strict improvement by replacing X_n with $\sup_{t \in \mathbb{N}} X_t$ at no cost whatsoever.

3.4 Notes

- 1 Some authors include in the definition of a stopping time τ that $\mathbb{P}(\tau < \infty) = 1$ and call random times without this property **Markov times**. We do *not* adopt this convention and allow stopping times to be infinite with nonzero probability. Stopping times are also called **optional times**.
- 2 There are several notations for probability kernels depending on the application. The following are commonly seen and equivalent: $K(x, A) = K(A | x) = K_x(A)$. For example, in statistics a parametric family is often given by $\{\mathbb{P}_\theta : \theta \in \Theta\}$ where Θ is the parameter space and \mathbb{P}_θ is a measure on some measurable space (Ω, \mathcal{F}) . This notation is often more convenient than writing $\mathbb{P}(\theta, \cdot)$. In Bayesian statistics the posterior is a probability kernel from the observation space to the parameter space and this is often written as $\mathbb{P}(\theta | X)$.
- 3 There is some disagreement about whether or not a Markov chain on an uncountable state space should instead be called a **Markov process**. In this book we use Markov chain for arbitrary state spaces and discrete time. When time is continuous (which it never is in this book), there is general agreement that 'process' is more appropriate. For a little more history on this see the preface of the book by [Meyn and Tweedie \[2012\]](#).

3.5 Bibliographic remarks

There are many places to find the construction of a stochastic process. Like before we recommend [Kallenberg \[2002\]](#) for readers who want to refresh their

memory and Billingsley [2008] for a more detailed account. For Markov chains the recent book by Levin and Peres [2017] provides a wonderful introduction. After reading that you might like the tome by Meyn and Tweedie [2012]. A proof of Theorem 3.1 is given Theorem 3.19 in the book by Kallenberg [2002]. Theorem 3.2 is credited to Percy John Daniell by Kallenberg [2002] (see Aldrich 2007). More general versions of this theorem exist. Readers looking for these should look up **Kolmogorov's extension theorem** [Kallenberg, 2002, Thm 6.16]. The theorem of Ionescu Tulcea (Theorem 3.3) is attributed to him [Tulcea, 1949–50] with a modern proof in the book by [Kallenberg, 2002, Thm 6.17]. There are lots of minor variants of the optional stopping theorem, most of which can be found in any probability book featuring martingales. The most historically notable source is by the man himself [Doob, 1953]. A more modern book that also gives the maximal inequalities is the book on optimal stopping by Peskir and Shiryaev [2006].

3.6 Exercises

3.1 Fill in the details of Theorem 3.1:

- Prove that F_1, F_2, \dots are $\mathcal{S} \rightarrow (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ random variables.
- In what follows equip \mathcal{S} with $\mathbb{P} = \lambda$, the uniform probability measure, Show that for any $t \geq 1$, F_t is uniformly distributed: $\mathbb{P}(F_t = 0) = \mathbb{P}(F_t = 1) = 0.5$.
- Show that F_1, F_2, \dots are independent.
- Show that $(X_{m,t})_{t \geq 1}$ is also an independent sequence of Bernoulli random variables, that are uniformly distributed.
- Show that $X_t = \sum_{t \geq 1} X_{m,t} 2^{-t}$ is uniformly distributed on $[0, 1]$.

3.2 Let X_1, X_2, \dots be an infinite sequence of independent Rademacher random variables and $S_t = \sum_{s=1}^t X_s 2^{s-1}$.

- Show that S_1, S_2, \dots is a martingale.
- Let $\tau = \min\{t : S_t = 1\}$ and show that $\mathbb{P}(\tau < \infty) = 1$.
- What is $\mathbb{E}[S_\tau]$?
- Explain why this does not contradict Doob's optional stopping theorem.

3.3 Give an example of a martingale S_1, S_2, \dots and stopping time τ such that

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_{\tau \wedge n}] \neq \mathbb{E}[X_\tau].$$

3.4 Show that Theorem 3.5 does not hold in general for supermartingales if the assumption that it be nonnegative is dropped.

3.5 Let τ_1, τ_2, \dots be an almost surely increasing sequence of \mathbb{F} -stopping times on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{F} = (\mathcal{F}_t)$, which means that $\tau_1(\omega) \leq \tau_2(\omega) \leq \dots$ almost surely. Prove that $\tau(\omega) = \lim_{n \rightarrow \infty} \tau_n(\omega)$ is a \mathbb{F} -stopping time.