Cairo University
Faculty of Science
Department of Mathematics

Stochastic Processes

Stat 301

Mostafa SABRI

Preface

These lecture notes are aimed at undergraduate students. Though a prior knowledge of measure theory is helpful, it is not necessary, everything needed is summarized in the first chapter. I have taught this course twice for students with no knowledge of measure theory without any difficulty. I have chosen not to avoid measure theory altogether because I feel it is quite essential that students get used to this spirit early on.

There is room for enlarging the contents if the students are very clever. For this I have added guidelines in the last chapter, any of these topics can be taught more properly if desired. The sections called "further results" in various chapters are optional and could be expanded if the time allows.

These are lecture notes and bear no claim for originality, sometimes I very closely followed some books; the purpose of the notes is the choice of the material and sometimes adding certain remarks, mixing from here and there and so forth.

I am not a specialist of this field and welcome any comments or simpler proofs at mmsabri@sci.cu.edu.eg.

Mostafa Sabri

Contents

Chapter 1

Review of Probability Theory

In this chapter we briefly review some concepts of probability theory that the student learned in an earlier course, see e.g. [15]. We also take the opportunity to reformulate some definitions more precisely using a new language that will be very useful for the future chapters.

1.1 Reminders

A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying the following properties :

(1) Ω is a non-empty set which we call the *sample space*. It represents the outcomes of an experiment.

(2) $\mathcal{F}$ is a family of *events*. Each event is represented by a subset $A \subset \Omega$. The family of events must satisfy that :
— $\emptyset, \Omega \in \mathcal{F}$. These represent the *impossible* and *sure* events, respectively.
— $A \in \mathcal{F} \implies A^c \in \mathcal{F}$. In other words, if A is an event, then A^c is also an event, called the complementary event.
— $A_1, A_2, \cdots \in \mathcal{F} \implies \cup_n A_n \in \mathcal{F}$. In other words, if A_i are events, we can speak of the event "A_1 or A_2 or..."

(3) $\mathbb{P} : \mathcal{F} \to [0,1]$ is called *a probability measure*. It measures the probability of events in $\mathcal{F}$. It satisfies that
— $\mathbb{P}(\Omega) = 1$, $\mathbb{P}(\emptyset) = 0$,
— If $A_1, A_2, \dots$ are pairwise disjoint, then $\mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n)$.

Example 1.1. The experiment of throwing a dice is modeled by the triple $(\Omega, \mathcal{F}, \mathbb{P})$ given by $\Omega = \{1, \dots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, the power set of Ω, and $\mathbb{P}(A) = \frac{N(A)}{6}$, where $N(A)$ is the number of points in A. Recall that the power set of Ω is the family of all subsets of Ω.

This example is typical of experiments with finitely many outcomes :

Lemma 1.2. *If Ω is a finite set with N elements, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(A) = \frac{N(A)}{N}$, then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.*

See [15, Chapter 2]. The sample space however may be infinite :

Example 1.3. Consider $\Omega = \mathbb{R}$ and $\mathbb{P}([a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2}\, dx$. Then $\mathbb{P}$ is a probability measure on Ω.

We were a bit sloppy in the last example. What is the family $\mathcal{F}$ in this case ? We discuss this in the next section.

1.2 A new language

Definition 1.4. Let Ω be a non-empty set. A *σ-algebra $\mathcal{F}$ on Ω* is a family of events of Ω. Thus $\mathcal{F}$ must satisfy the following :

— $\emptyset \in \mathcal{F}$,

— $A \in \mathcal{F} \implies A^c \in \mathcal{F}$,

— $A_1, A_2, \cdots \in \mathcal{F} \implies \cup_n A_n \in \mathcal{F}$.

Events $A \in \mathcal{F}$ are called *measurable sets*.

The terminology "algebra" comes from the fact that the 2nd and 3rd conditions mean that $\mathcal{F}$ is closed under the operations of taking the complement and taking the union, much like an additive group is closed under addition. The symbol σ comes from allowing (countably) infinite unions.

We now answer what family $\mathcal{F}$ should be considered on $\mathbb{R}$ as in Example 1.3.

Definition 1.5. Let $\Omega = \mathbb{R}$. We always equip $\mathbb{R}$ with the family $\mathcal{F} = \mathcal{B}(\mathbb{R})$ of *Borel sets*. This is defined as the smallest σ-algebra containing all intervals of $\mathbb{R}$.

It is natural to ask why we didn't just consider $\mathcal{F} = \mathcal{P}(\mathbb{R})$ as in Lemma 1.2. This is indeed a legitimate σ-algebra after all. The problem is that the most natural probability measures on $\mathbb{R}$, such as the one given in Example 1.3, cannot be defined on the whole family $\mathcal{P}(\mathbb{R})$. This family is too big; we can define $\mathbb{P}(A)$ if A is an interval, a union of intervals and so forth. But some sets in $\mathcal{P}(\mathbb{R})$ are too complicated, so we restrict the domain of $\mathbb{P}$ to the smaller family $\mathcal{B}(\mathbb{R})$ of Borel sets. The student will learn more about this in a course on measure theory.

Definition 1.6. We equip $\Omega = [a,b]$ with the family $\mathcal{B}([a,b])$ of Borel subsets of $[a,b]$. This is the smallest σ-algebra containing all sub-intervals of $[a,b]$.

If a measure is defined on the Borel sets, it suffices to specify its action on the intervals $[a, b]$.

Definition 1.7. The *Lebesgue measure* $\text{Leb} : \mathcal{B}(\mathbb{R}) \to [0, \infty)$ satisfies

$$\text{Leb}([a, b]) = b - a$$

for any $[a, b] \in \mathcal{B}(\mathbb{R})$. In other words, Leb generalizes the concept of the length of an interval to arbitrary $B \in \mathcal{B}(\mathbb{R})$.

We mention that Leb can be defined on the larger domain of *Lebesgue measurable* sets, but this is not important for our course. The student will learn in measure theory how to construct Leb, it takes effort, here we just use it.

Example 1.8. 1. Let $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$ and $\mathbb{P} = \text{Leb}$. Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

2. More generally, let $\Omega = [a, b]$, $\mathcal{F} = \mathcal{B}([a, b])$ and $\mathbb{P} = \frac{1}{b-a}\text{Leb}$. Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

3. Let $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ and $\mathbb{P}([a, b]) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2} \, dx$. Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

4. More generally, if $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ and f is any piecewise continuous, nonnegative function with $\int_{\mathbb{R}} f(x) \, dx = 1$, then letting $\mathbb{P}([a, b]) = \int_a^b f(x) \, dx$, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Definition 1.9. We say that a property holds *almost surely* (summarized as "a.s.") if it holds with probability one. In other words, there exists $A \in \mathcal{F}$ such that $\mathbb{P}(A) = 1$, and for any $\omega \in A$, the property is satisfied. When $\mathbb{P} = \text{Leb}$, we also use the terminology *almost everywhere* (summarized as "'a.e."). For example, if $\Omega = [0, 1]$ and $\mathbb{P} = \text{Leb}$, then $f(x) = \frac{1}{x}$ is finite a.e. Indeed, is is finite on $(0, 1]$ and only explodes at 0. Since $\mathbb{P}((0, 1]) = 1$, it is finite a.e. (or a.s., it's the same thing).

We conclude this section by recalling some results studied in [15, Chap. 2] :

Lemma 1.10. *Suppose $A_1 \subset A_2 \subset \ldots$ is an increasing sequence of events. Then $\mathbb{P}(\cup_n A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$.*

Suppose $B_1 \supset B_2 \supset \ldots$ is a decreasing sequence of events. Then $\mathbb{P}(\cap_n B_n) = \lim_{n\to\infty} \mathbb{P}(B_n)$.

Lemma 1.11 (First Borel-Cantelli lemma)**.** *Suppose $A_1, A_2, \ldots$ is an infinite sequence of events with $\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty$. Then with probability one, only finitely many of the events $A_1, A_2, \ldots$ occur.*

In symbols, $\mathbb{P}(\cap_n \cup_{k\geq n} A_k) = 0$.

1.3 Random variables

Recall that if $X : \Omega \to \mathbb{R}$ and $B \subseteq \mathbb{R}$, then

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \subseteq \Omega$$

is the *inverse image of* B. We often use the intuitive notation

$$\{X \in B\} := X^{-1}(B)$$

which simply describes the event that "the values of X belong to B".

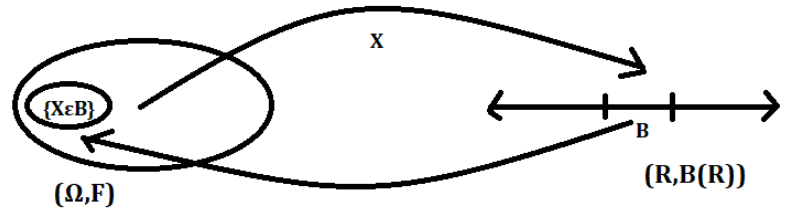


Figure 1.1 – A map $X : \Omega \to \mathbb{R}$, a Borel set $B \subset \mathbb{R}$ and its inverse image $\{X \in B\} \subseteq \Omega$.

We now recall the central object of study in probability theory.

Definition 1.12. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that $X : \Omega \to \mathbb{R}$ is *$\mathcal{F}$-measurable* if $X^{-1}(B) \in \mathcal{F}$ for any $B \in \mathcal{B}(\mathbb{R})$.

We say that X is a *random variable* if X is $\mathcal{F}$-measurable.

Thus, a random variable is simply a map $X : \Omega \to \mathbb{R}$ such that the inverse image of any measurable set is measurable.

Example 1.13. 1. Any constant function $X = c$ is a random variable. In fact $X^{-1}(B) = \Omega \in \mathcal{F}$ if $c \in B$ and $X^{-1}(B) = \emptyset \in \mathcal{F}$ if $c \notin B$.

2. Define the *indicator function* of a set A by

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

If $A \in \mathcal{F}$, then $\mathbf{1}_A$ is a random variable. In fact,

$$\mathbf{1}_A^{-1}(B) = \begin{cases} A & \text{if } 1 \in B \text{ and } 0 \notin B, \\ A^c & \text{if } 1 \notin B \text{ and } 0 \in B, \\ \Omega & \text{if } 0, 1 \in B, \\ \emptyset & \text{if } 0, 1 \notin B. \end{cases}$$

All the sets on the RHS are in $\mathcal{F}$ so $\mathbf{1}_A$ is a random variable.

3. Endow Ω with the power set $\mathcal{F} = \mathcal{P}(\Omega)$. Then any $X : \Omega \to \mathbb{R}$ is a random variable.
4. Endow Ω with $\mathcal{F} = \{\Omega, \emptyset\}$. Then only the constant functions are random variables. Indeed, if a random variable X took at least two distinct values c_1, c_2, there would be two non-empty sets $X^{-1}(\{c_1\})$ and $X^{-1}(\{c_2\})$ in $\mathcal{F}$.

Lemma 1.14. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $X : \Omega \to \mathbb{R}$ satisfies $X^{-1}(B) \in \mathcal{F}$ for any interval $B = (a, \infty)$, then X is a random variable.*

Proof. Admitted, see e.g. [17, Lemma A.10]. □

Definition 1.15. A *Borel function* $\phi : \mathbb{R} \to \mathbb{R}$ is a function satisfying $\phi^{-1}(B) \in \mathcal{B}(\mathbb{R})$ for any $B \in \mathcal{B}(\mathbb{R})$.

Such ϕ is a random variable on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ for any probability $\mathbb{P}$ on $\mathbb{R}$.

Lemma 1.16. *If $X : \mathbb{R} \to \mathbb{R}$ is piecewise continuous, then X is a Borel function.*

Proof. Admitted. □

A very important concept for our course is the following.

Definition 1.17. Let Ω be a non-empty set and $X : \Omega \to \mathbb{R}$ a map. The *σ-algebra generated by X*, denoted by $\sigma(X)$, consists of all sets of the form $X^{-1}(B)$, $B \in \mathcal{B}(\mathbb{R})$.

Clearly any $X : \Omega \to \mathbb{R}$ is $\sigma(X)$-measurable. More generally,

Lemma 1.18. *Let $X : \Omega \to \mathbb{R}$ be a map and $\phi : \mathbb{R} \to \mathbb{R}$ be a Borel function. Then $\phi \circ X =: \phi(X)$ is $\sigma(X)$-measurable.*

Proof. Let $B \in \mathcal{B}(\mathbb{R})$. We have

$$X^{-1}(\phi^{-1}(B)) = \{\omega : X(\omega) \in \phi^{-1}(B)\} = \{\omega : \phi(X(\omega)) \in B\} = (\phi \circ X)^{-1}(B)\,.$$

Since ϕ is Borel, $\phi^{-1}(B) \in \mathcal{B}(\mathbb{R})$. So $X^{-1}(\phi^{-1}(B)) \in \sigma(X)$ by definition of $\sigma(X)$. Thus, $(\phi \circ X)^{-1}(B) \in \sigma(X)$. □

Quite remarkably, the converse is also true.

Lemma 1.19 (Doob-Dynkin)**.** *Let Y be a random variable and suppose $Z : \Omega \to \mathbb{R}$ is $\sigma(Y)$-measurable. Then $Z = \phi(Y)$ for some Borel function $\phi : \mathbb{R} \to \mathbb{R}$.*

More generally, a random variable Z is $\sigma(X_1, \ldots, X_k)$-measurable iff there is a Borel $\phi : \mathbb{R}^k \to \mathbb{R}$ such that $Z = \phi(X_1, \ldots, X_k)$, i.e. $Z(\omega) = \phi((X_1(\omega), \ldots, X_k(\omega)))$.

Proof. Admitted. The proof of this very useful result is actually not really difficult, see [9, Lemma 1.13] and [3, Theorem 20.1]. □

We can generalize Definition 1.17 to families of random variables.

Definition 1.20. Let $(X_\alpha)_{\alpha\in I}$ be a family of maps $X_\alpha : \Omega \to \mathbb{R}$. The *$\sigma$-algebra generated by* $(X_\alpha)_{\alpha\in I}$ is defined to be the smallest σ-algebra containing all sets of the form $X_\alpha^{-1}(B)$, $B \in \mathcal{B}(\mathbb{R})$, $\alpha \in I$.

Note that the index I above may be finite, countable or uncountable.

Definition 1.21. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that $f : \Omega \to \mathbb{R}$ is *integrable* if $\int_\Omega |f| \, \mathrm{d}\,\mathbb{P} < \infty$. We denote the space of integrable functions by $L^1(\Omega, \mathcal{F}, \mathbb{P})$.

Similarly, f is said to be *square integrable* if $\int_\Omega |f|^2 \, \mathrm{d}\,\mathbb{P} < \infty$. The space of such functions is denoted by $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Lemma 1.22. *We have $L^2(\Omega, \mathcal{F}, \mathbb{P}) \subset L^1(\Omega, \mathcal{F}, \mathbb{P})$.*

Proof. Admitted, follows immediately from the Cauchy-Schwarz inequality the student will learn in another course. □

This concludes the "new language" concerning random variables. The rest of this chapter consists of reminders of Stat 201.

Definition 1.23. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. We define the *distribution of* X to be the probability measure on $\mathbb{R}$ defined by

$$P_X(B) = \mathbb{P}(X \in B) \, .$$

Recall Figure 1.1 for an illustration. We often describe this by saying that X *pushes forward* the probability measure $\mathbb{P}$ from Ω to $\mathbb{R}$.

Definition 1.24. We say that $X : \Omega \to \mathbb{R}$ is a *discrete* random variable if X only takes countably many values $\{\alpha_1, \alpha_2, \ldots\}$. In this case,

$$P_X(B) = \sum_{\alpha_k \in B} P_X(\{\alpha_k\}) = \sum_{\alpha_k \in B} \mathbb{P}(X = \alpha_k) \, .$$

We say that $X : \Omega \to \mathbb{R}$ is a *continuous* random variable[1] if there exists a nonnegative integrable function p_X such that for any interval $I \subseteq \mathbb{R}$,

$$P_X(I) = \int_I p_X(t) \, \mathrm{d}t = \mathbb{P}(X \in I) \, .$$

In this case we call p_X the *density* of X.

1. The more correct terminology is to say that such random variables are absolutely continuous.

Definition 1.25. Let $X, Y : \Omega \to \mathbb{R}$ be two random variables. Then the *joint distribution* of X and Y is the probability measure $P_{X,Y}$ on $\mathbb{R}^2$ defined by

$$P_{X,Y}(I \times J) = \mathbb{P}(X \in I \text{ and } Y \in J)$$

for any intervals $I, J \subseteq \mathbb{R}$.

As before, if X, Y are discrete, $\operatorname{Ran} X = \{\alpha_i\}$ and $\operatorname{Ran} Y = \{\beta_j\}$, then

$$P_{X,Y}(B) = \sum_{(\alpha_i,\beta_j)\in B} P_{X,Y}(\{(\alpha_i,\beta_j)\}) = \sum_{(\alpha_i,\beta_j)\in B} \mathbb{P}(X = \alpha_i,\ Y = \beta_j)$$

for $B \subseteq \mathbb{R}^2$. Similarly, if X, Y are continuous, then

$$P_{X,Y}(B) = \iint_B p_{X,Y}(s,t)\,\mathrm{d}s\mathrm{d}t\,.$$

In this case, $p_{X,Y}$ is a nonnegative integrable function on $\mathbb{R}^2$ which is called the *joint density* of X and Y.

Definition 1.26. Let X be an integrable random variable. We define the *expectation* of X by

$$\mathbb{E}(X) = \int_\Omega X\,\mathrm{d}\,\mathbb{P}\ .$$

If X is square integrable, we define the variance

$$\operatorname{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2\,.$$

Example 1.27. Recall the indicator function of Example 1.13. We have

$$\mathbb{E}(\mathbf{1}_A) = \mathbb{P}(A)$$

for any $A \in \mathcal{F}$. More generally, if $X = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i}$ with $A_i \in \mathcal{F}$, we have

$$\mathbb{E}(X) = \sum_{i=1}^n \alpha_i\,\mathbb{P}(A_i)\,.$$

We now mention a very important result.

Theorem 1.28 (Change of variables formula)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable, $\varphi : \mathbb{R} \to \mathbb{R}$ be a Borel function and suppose $\varphi(X)$ is integrable. Then*

$$\mathbb{E}(\varphi(X)) = \int_{-\infty}^{\infty} \varphi(x)\,\mathrm{d}P_X(x) = \begin{cases} \sum_j \varphi(\alpha_j) P_X(\{\alpha_j\}) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} \varphi(x) p_X(x)\,\mathrm{d}x & \text{if } X \text{ is continuous.} \end{cases}$$

Similarly, if X, Y are random variables with joint distribution $P_{X,Y}$, then for any Borel $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$, we have

$$\begin{aligned} \mathbb{E}(\varphi(X,Y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x,y)\,\mathrm{d}P_{X,Y}(x,y) \\ &= \begin{cases} \sum_{j,k} \varphi(\alpha_j,\beta_k) P_{X,Y}(\{\alpha_j,\beta_k\}) & \text{if } X,Y \text{ are discrete,} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x,y) p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y & \text{if } X,Y \text{ are continuous.} \end{cases} \end{aligned} \tag{1.1}$$

Proof. Admitted. See e.g. [17, Theorem A.31]. □

We conclude this section with one last definition and lemma.

Definition 1.29. Let X be a random variable. We define the *distribution function of* X to be the map $F_X : \mathbb{R} \to [0,1]$ defined by

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

Lemma 1.30. *The distribution function satisfies the following properties :*

(a) *F_X is non-decreasing.*

(b) *F_X is right-continuous.*

(c) *We have*

$$\lim_{x\to-\infty} F_X(x) = 0 \qquad \textit{and} \qquad \lim_{x\to\infty} F_X(x) = 1\,.$$

(d) *If X has a density p_X, and if p_X is continuous at x, then $\frac{\mathrm{d}}{\mathrm{d}x}F_X(x) = p_X(x)$.*

(e) *If X is a discrete random variable with values $\{\alpha_i\}_i$, then F_X is constant on each interval $(s,t]$ not containing any α_i. F_X has jumps of size $\mathbb{P}(X = \alpha_i)$ at each α_i.*

Proof (can be skipped). (a) If $x < y$ then $F_X(y) - F_X(x) = \mathbb{P}(X \leq y) - \mathbb{P}(X \leq x) = \mathbb{P}(X \in (x,y]) \geq 0$, so $F_X(y) \geq F_X(x)$.

(b) Let $x_n \to x^+$. We may assume $x_1 \geq x_2 \geq \dots$. We should prove that $F(x_n) \to F(x)$. Since $x_{n+1} \leq x_n$, if $B_n = \{X \leq x_n\}$, then B_n are decreasing, $B_1 \supseteq B_2 \supseteq \dots$. Moreover, $\cap_n B_n = \{X \leq x\}$. Indeed, if $X \leq x$ then $X \leq x_n$ for all n since $x_n \geq x$. Conversely, if $X \leq x_n$ for all n, taking $n \to \infty$ gives $X \leq x$. By Lemma 1.10 we get $\mathbb{P}(X \leq x) = \lim_{n\to\infty} \mathbb{P}(B_n)$. So $F_X(x) = \lim_{n\to\infty} F_X(x_n)$.

(c) Let $A_n = \{X \leq n\}$ and $B_n = \{X \leq -n\}$. Clearly A_n is increasing and B_n is decreasing with $n \geq 0$. Moreover, $\Omega = \cup_n A_n$ and $\emptyset = \cap_n B_n$. Indeed, this simply says that for any $\omega \in \Omega$ we can find n such that $-n < X(\omega) \leq n$. Using Lemma 1.10, we get

$$1 = \mathbb{P}(\Omega) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} F_X(n) \quad \text{and} \quad 0 = \mathbb{P}(\emptyset) = \lim_{n\to\infty} F_X(-n)\,.$$

Since F_X is non-decreasing, we have $\lim_{x\to\infty} F_X(x) = \lim_{n\to\infty} F_X(n)$ and $\lim_{x\to-\infty} F_X(x) = \lim_{n\to-\infty} F_X(n)$. This is by the Sandwich theorem since $F_X(n-1) \leq F_X(x) \leq F_X(n)$ for $n = \lfloor x \rfloor$. This proves (c).

(d) If X has density p_X then $F_X(x) = \int_{-\infty}^{x} p_X(t)\,\mathrm{d}t$. If moreover p_X is continuous at x, then by calculus, we know $F_X'(x) = p_X(x)$.

(e) If $(s,t]$ has no α_i, then for $x, y \in (s,t]$, $x < y$, we have $F_X(y) - F_X(x) = \mathbb{P}(X \in (x,y]) = 0$, so $F_X(x) = F_X(y)$ and F_X is constant on $(s,t]$. Next, we should prove that $\lim_{x\to\alpha_i^+} F_X(x) - \lim_{x\to\alpha_i^-} F_X(x) = \mathbb{P}(X = \alpha_i)$. For this, we have $\lim_{x\to\alpha_i^+} F_X(x) = F_X(\alpha_i)$ by right-continuity. As in (b), we can show that if $y_n \to x^-$, $y_1 < y_2 < \dots$, then $\{X < x\} = \cup_n \{X \le y_n\}$, implying $\lim_{x\to x^-} F_X(x) = \mathbb{P}(X < x)$. We conclude that $\lim_{x\to\alpha_i^+} F_X(x) - \lim_{x\to\alpha_i^-} F_X(x) = F_X(\alpha_i) - \mathbb{P}(X < \alpha_i) = \mathbb{P}(X = \alpha_i)$. □

1.4 Conditional probability and independence

Definition 1.31. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $A, B \in \mathcal{F}$ and $\mathbb{P}(B) \neq 0$. We define the *conditional probability of A given B* by

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

We recall the important

Theorem 1.32 (Total probability formula)**.** *If $B_1, B_2, \dots$ are pairwise disjoint events with $\cup_j B_j = \Omega$, then*

$$\mathbb{P}(A) = \sum_j \mathbb{P}(A \mid B_j)\, \mathbb{P}(B_j).$$

The student should revise [15, Chapter 3] for the proof and more comments, examples on conditional probability. By knowing that B has occurred, B essentially becomes the new sample space and is endowed $\mathbb{P}_B(A) := \mathbb{P}(A \mid B)$.

Definition 1.33. We say that two events A, B are *independent* if

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

We say $A_1, \dots, A_n \in \mathcal{F}$ are *independent* if for any $J \subseteq \{1, \dots, n\}$, we have

$$\mathbb{P}(\cap_{j\in J} A_j) = \prod_{j\in J} \mathbb{P}(A_j).$$

We say that an infinite sequence $A_1, A_2, \dots$ of events are independent if $A_1, \dots, A_n$ are independent for any n.

Lemma 1.34. *Let $\mathbb{P}(B) \neq 0$. Then $A, B \in \mathcal{F}$ are independent iff $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.*

This is easily checked; see [15].

Definition 1.35. We say that the random variables $X_1, \dots, X_n$ are *independent* if for any $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$, the events $\{X_1 \in B_1\}, \dots, \{X_n \in B_n\}$ are independent.

An infinite sequence of random variables is independent if any finite subset of this family is independent.

Theorem 1.36. *If $X_1, \dots, X_n$ are independent integrable random variables then*

$$\mathbb{E}(X_1 \cdots X_n) = \mathbb{E}(X_1) \cdots \mathbb{E}(X_n)\,.$$

Theorem 1.37. *If X, Y are independent, then their joint distribution is a product measure : we have $\mathrm{d}P_{X,Y}(x,y) = \mathrm{d}P_X(x)\mathrm{d}P_Y(y)$.*

In particular, if X, Y are independent, then in (1.1), we get $P_{X,Y}(\{\alpha_j, \beta_k\}) = P_X(\{\alpha_j\})P_Y(\{\beta_k\})$ and $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ in the discrete and continuous cases, respectively. A similar result holds for independent $X_1, \dots, X_n$.

Definition 1.38. The σ-algebras $\mathcal{G}_1, \dots, \mathcal{G}_n \subset \mathcal{F}$ are said to be *independent* if any n events $A_1 \in \mathcal{G}_1, \dots, A_n \in \mathcal{G}_n$ are independent.

A random variable X is *independent* of a σ-algebra $\mathcal{G} \subset \mathcal{F}$ if $\sigma(X)$ and $\mathcal{G}$ are independent.

Comparing Definitions 1.35 and 1.38, we notice that X, Y are independent iff $\sigma(X)$ and $\sigma(Y)$ are independent.

1.5 Exercises

1. Let A be an event. Find $\sigma(\mathbf{1}_A)$.
2. Let A, B be events. Find $\sigma(\mathbf{1}_A + 2\,\mathbf{1}_B)$.
3. If X takes only n values $\{\alpha_k\}_{k=1}^n$, find $\sigma(X)$.
4. Let $A \in \mathcal{F}$. Is A independent of $\Omega, \emptyset$?
5. Show that X and $\mathcal{G}$ are independent iff X and $\mathbf{1}_A$ are independent for any $A \in \mathcal{G}$.

Chapter 2

Conditional expectation

The concept of conditional expectation, in full generality, is not entirely intuitive. This is why we first discuss various important special cases and show how to perform computations. We follow [5] closely in this chapter, simply adding a few remarks and results, and occasionally solving slightly differently.

In the future chapters, we will need the concept of $\mathbb{E}(X \mid \mathcal{G})$, where $\mathcal{G} \subset \mathcal{F}$ is a σ-algebra generated by a set of random variables $(X_\alpha)_{\alpha \in I}$ for some index set I.

2.1 Conditioning on an event

Definition 2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X be an integrable random variable and $B \in \mathcal{F}$ such that $\mathbb{P}(B) \neq 0$. We define the *conditional expectation of X given B* to be the number

$$\mathbb{E}(X \mid B) = \frac{1}{\mathbb{P}(B)} \int_B X \, \mathrm{d}\, \mathbb{P} \, . \tag{2.1}$$

This is simply the expectation over the probability space $(B, \mathcal{F}, \mathbb{P}_B)$, where $\mathbb{P}_B(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(A \mid B)$.

In practice, if we have a random variable, the grades of students in a forthcoming exam for example, and we are trying to make a prediction on its most likely value, then if we have no information at all, our best guess is to take $\mathbb{E}(X) = \frac{1}{101} \sum_{k=0}^{100} k = \frac{(100)(101)}{2(100)} = 50$. Here we assumed the exam is over 100 and took a mean over all possible 101 grades. On the other hand, if we know that an event B has occurred, for example, we are dealing with a very good class, then our guess should be modified. For example, we can take the mean over the 26 grades from 75 to 100. This is $\frac{1}{26} \sum_{k=75}^{100} k = \frac{1}{26} (\sum_{k=0}^{100} k - \sum_{k=0}^{74} k) = \frac{1}{26} (\frac{(100)(101)}{2} - \frac{(74)(75)}{2}) = \frac{2275}{26} = 87.5$. This is exactly what (2.1) is about : in the

first case the sample space is $\Omega = \{0, \ldots, 100\}$ and $\mathbb{P}(k) = \frac{k}{101}$, in the second case $B = \{75, \ldots, 100\}$ and $\mathbb{P}_B(k) = \frac{k}{26} = \frac{100}{26} \cdot \frac{k}{100} = \frac{1}{\mathbb{P}(B)} \cdot \mathbb{P}(k)$.

Example 2.2. Three coins of values 10p, 20p and 50p are tossed. Let X be the total value given by the coins that land heads up. What is the expected value of X, given that two coins have landed heads up ?

Solution. The sample space Ω consists of 8 points ABC with $A, B, C \in \{H, T\}$, each $\mathbb{P}(ABC) = \frac{1}{8}$. Let B be the event that two coins have landed heads up. Then $B = \{HHT, HTH, THH\}$. We have $\mathbb{P}(B) = \frac{3}{8}$. Thus,

$$\mathbb{E}(X \mid B) = \frac{1}{\mathbb{P}(B)} \sum_{\alpha_k \in B} X(\alpha_k)\, \mathbb{P}(\alpha_k) = \frac{1}{\frac{3}{8}} \cdot \frac{X(HHT) + X(HTH) + X(THH)}{8}$$
$$= \frac{30 + 60 + 70}{3} \approx 53.33\,.$$

Example 2.3. Find $\mathbb{E}(X \mid \Omega)$.

Solution. $\mathbb{E}(X \mid \Omega) = \frac{1}{\mathbb{P}(\Omega)} \int_\Omega X \,\mathrm{d}\,\mathbb{P} = \int_\Omega X \,\mathrm{d}\,\mathbb{P} = \mathbb{E}(X)$.

2.2 Conditioning on a discrete random variable

We now move on to an important generalization.

Definition 2.4. Let X be an integrable random variable and Y be a discrete random variable taking distinct values $\{y_1, y_2, \ldots\}$. We define the *conditional expectation of X given Y* to be the random variable

$$\mathbb{E}(X \mid Y)(\omega) = \mathbb{E}(X \mid \{Y = y_n\}) \quad \text{if } \omega \in \{Y = y_n\}\,.$$

For example, suppose Y takes only 4 values as in Figure 2.1 :

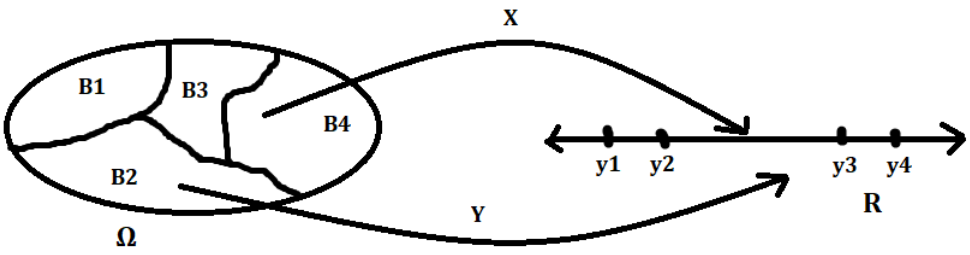


Figure 2.1 – Here Y takes values y_1, y_2, y_3, y_4. This induces a partition of Ω into four sets $B_k = \{Y = y_k\} = Y^{-1}(y_k)$.

The fact that Y has occurred means that one of the values $\{y_1, \dots, y_4\}$ has appeared. We do not know which, so we study each case. In Section 2.1 we were conditioning over a fixed set, so the result was a fixed number. Here we are conditioning over a random set $B_k = \{Y = y_k\}$, $k = 1, \dots, 4$, so we have a random variable. The definition says that the most important information we extract from Y is not its values $\{y_1, y_2, \dots\}$, but the partition it induces on Ω via its inverse images $Y^{-1}(\{y_k\})$.

Remark 2.5. The definition implies that if Y takes n values, then $\mathbb{E}(X \mid Y)$ also takes n values (which may not be distinct).

Example 2.6. Three coins of values 10p, 20p, 50p are tossed. Let X be the total value given by the coins that land heads up. Let Y be the total value given by the first two coins that land heads up. Find $\mathbb{E}(X \mid Y)$.

Solution. Y takes values $0, 10, 20, 30$. In fact $\{Y = 0\} = \{TTT, TTH\}$, $\{Y = 10\} = \{HTT, HTH\}$, $\{Y = 20\} = \{THT, THH\}$ and $\{Y = 30\} = \{HHT, HHH\}$. Each of these events has probability $\frac{2}{8}$. We have

$$\mathbb{E}(X \mid Y = 0) = \frac{1}{\frac{2}{8}} \cdot \frac{X(TTT) + X(TTH)}{8} = \frac{0 + 50}{2} = 25\,.$$

Similarly, $\mathbb{E}(X \mid Y = 10) = \frac{10+60}{2} = 35$, $\mathbb{E}(X \mid Y = 20) = \frac{20+70}{2} = 45$ and $\mathbb{E}(X \mid Y = 30) = \frac{30+80}{2} = 55$. Thus,

$$\mathbb{E}(X \mid Y)(\omega) = \begin{cases} 25 & \text{if } Y(\omega) = 0, \\ 35 & \text{if } Y(\omega) = 10, \\ 45 & \text{if } Y(\omega) = 20, \\ 55 & \text{if } Y(\omega) = 30. \end{cases}$$

Example 2.7. Let $\Omega = [0, 1]$, $\mathbb{P} = \text{Leb}$,

$$X(t) = t^2 \qquad \text{and} \qquad Y(t) = \begin{cases} 1 & \text{if } t \in [0, \frac{1}{3}], \\ 2 & \text{if } t \in (\frac{1}{3}, \frac{2}{3}), \\ 0 & \text{if } t \in (\frac{2}{3}, 1]. \end{cases}$$

Find $\mathbb{E}(X \mid Y)$.

Solution. Since Y is discrete, we have for $t \in [0, \frac{1}{3}]$,

$$\mathbb{E}(X \mid Y)(t) = \mathbb{E}(X \mid Y = 1) = \frac{1}{\mathbb{P}([0, \frac{1}{3}])} \int_0^{1/3} t^2 \, \mathrm{d}t = \frac{1}{\frac{1}{3}} \cdot \frac{1}{3} \cdot \frac{1}{27} = \frac{1}{27}\,.$$

For $t \in (\frac{1}{3}, \frac{2}{3})$,

$$\mathbb{E}(X \mid Y)(t) = \mathbb{E}(X \mid Y = 2) = \frac{1}{\mathbb{P}((\frac{1}{3}, \frac{2}{3}))} \int_{1/3}^{2/3} t^2 \, \mathrm{d}t = \frac{1}{\frac{1}{3}} \cdot \frac{1}{3} \cdot \frac{7}{27} = \frac{7}{27}.$$

For $t \in [\frac{2}{3}, 1]$,

$$\mathbb{E}(X \mid Y)(t) = \mathbb{E}(X \mid Y = 0) = \frac{1}{\mathbb{P}([\frac{2}{3}, 1])} \int_{2/3}^{1} t^2 \, \mathrm{d}t = \frac{1}{\frac{1}{3}} \cdot \frac{1}{3} \cdot \frac{19}{27} = \frac{19}{27}.$$

This completes the calculation of $\mathbb{E}(X \mid Y)$ for all $t \in \Omega$.

We make two remarks on Example 2.7 : first, the value of $\mathbb{E}(X \mid Y)$ does not depend on the values of Y. More precisely, if Y took any three distinct values y_1, y_2, y_3 on the same intervals $[0, \frac{1}{3}]$, $(\frac{1}{3}, \frac{2}{3})$, $[\frac{2}{3}, 1]$, then $\mathbb{E}(X \mid Y)$ would be the same. Secondly, note that if we take the average of the three values of $\mathbb{E}(X \mid Y)$, we get $\frac{\frac{1}{27} + \frac{7}{27} + \frac{19}{27}}{3} = \frac{27}{27 \cdot 3} = \frac{1}{3}$. This is also the value of $\mathbb{E}(X) = \int_0^1 t^2 \, \mathrm{d}t = \frac{1}{3}$. This property is a general fact, see Proposition 2.1.(2) with $A = \Omega$.

Example 2.8. Suppose Y is constant. Find $\mathbb{E}(X \mid Y)$.

Solution. Say $Y = c$. Then for any $\omega \in \Omega = \{Y = c\}$, we have $\mathbb{E}(X \mid Y)(\omega) = \frac{1}{\mathbb{P}(Y=c)} \int_{\{Y=c\}} X \, \mathrm{d}\mathbb{P} = \frac{1}{\mathbb{P}(\Omega)} \int_\Omega X \, \mathrm{d}\mathbb{P} = \mathbb{E}(X)$.

Example 2.8 says that the information provided by a constant random variable is completely useless. It has not induced any interesting partition of Ω.

Proposition 2.1. *Let X be an integrable random variable and Y be a discrete random variable. Then*

(1) *$\mathbb{E}(X \mid Y)$ is $\sigma(Y)$-measurable,*

(2) *For any $A \in \sigma(Y)$, we have*

$$\int_A \mathbb{E}(X \mid Y) \, \mathrm{d}\mathbb{P} = \int_A X \, \mathrm{d}\mathbb{P} \, .$$

Proof. Assume Y takes the distinct values $y_1, y_2, \ldots$ Let $Z = \mathbb{E}(X \mid Y)$. Then Z is a discrete random variable with values $z_1, z_2, \ldots$ on the sets $\{Y = y_1\}$, $\{Y = y_2\}, \ldots$, respectively.

Given a Borel set $B \subset \mathbb{R}$, we have

$$Z^{-1}(B) = \bigcup_{z_j \in B} Z^{-1}(\{z_j\}) = \bigcup_{z_j \in B} \{\omega : Z(\omega) = z_j\} = \bigcup_{z_j \in B} \{\omega : Y(\omega) = y_j\} \in \sigma(Y)$$

since it is a union of events in $\sigma(Y)$. Thus, Z is $\sigma(Y)$-measurable. Next,

$$\begin{aligned} \int_{\{Y=y_j\}} \mathbb{E}(X \mid Y) \, \mathrm{d}\mathbb{P} &= \int_{\{Y=y_j\}} \mathbb{E}(X \mid Y = y_j) \, \mathrm{d}\mathbb{P} \\ &= \mathbb{E}(X \mid Y = y_j) \cdot \mathbb{P}(Y = y_j) = \int_{\{Y=y_j\}} X \, \mathrm{d}\mathbb{P} \, . \end{aligned} \tag{2.1}$$

But we saw that any $A = Z^{-1}(B) \in \sigma(Y)$ is a union of events of the form $\{Y = y_j\}$. These events are pairwise disjoint since Y cannot take two values at the same time. Thus, using (2.1),

$$\int_A \mathbb{E}(X \mid Y)\,\mathrm{d}\,\mathbb{P} = \int_{\cup_j \{Y=y_j\}} \mathbb{E}(X \mid Y)\,\mathrm{d}\,\mathbb{P} = \sum_j \int_{\{Y=y_j\}} \mathbb{E}(X \mid Y)\,\mathrm{d}\,\mathbb{P}$$
$$= \sum_j \int_{\{Y=y_j\}} X\,\mathrm{d}\,\mathbb{P} = \int_{\cup_j \{Y=y_j\}} X\,\mathrm{d}\,\mathbb{P} = \int_A X\,\mathrm{d}\,\mathbb{P}$$

as required. □

2.3 Conditioning on an arbitrary random variable

2.3.1 Basic properties

Defining $\mathbb{E}(X \mid Y)$ when Y is arbitrary is not straightforward. We know from Proposition 2.1 that if Y is discrete, then $\mathbb{E}(X \mid Y)$ satisfies two properties. We now use them to actually define the concept for general Y :

Definition 2.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let X be an integrable random variable and Y be an arbitrary random variable. We define the *conditional expectation of X given Y* to be the random variable $\mathbb{E}(X \mid Y)$ such that

(1) $\mathbb{E}(X \mid Y)$ is $\sigma(Y)$-measurable,

(2) For any $A \in \sigma(Y)$,
$$\int_A \mathbb{E}(X \mid Y)\,\mathrm{d}P = \int_A X\,\mathrm{d}P\,.$$

Unlike the discrete case, here $\mathbb{E}(X \mid Y)$ is defined *implicitly*. We do not give a formula $\mathbb{E}(X \mid Y)(\omega) = (...)$, instead we say that $\mathbb{E}(X \mid Y)$ is a random variable satisfying some properties. So we should check that there exists indeed a random variable satisfying this, and that it is essentially unique.

Before doing so however we give a few remarks to explain the conditions. The first condition says that $\mathbb{E}(X \mid Y)$ should vary according to the "information" i.e. σ-algebra provided by Y. For example, we will see later on that it implies that if Y is constant on a region, then $\mathbb{E}(X \mid Y)$ must also be constant on that region. Similarly, if Y has some symmetries, they will be inherited by $\mathbb{E}(X \mid Y)$.

Condition (2) generalizes the idea that $\mathbb{E}(X \mid Y)$ should be a kind of normalized average. In the discrete case we defined $\mathbb{E}(X \mid Y)(\omega) = \frac{1}{\mathbb{P}(Y=y_j)} \int_{\{Y=y_j\}} X\,\mathrm{d}P$ if $Y(\omega) = y_j$. Here we cannot do the same because $\mathbb{P}(Y = y_j)$ can be zero. However we still have this idea of normalized average, but *in mean :* we cannot say

that for each ω we have an average, instead, if we take the mean over A of $\mathbb{E}(X \mid Y)$, then this gives the mean of the normalized average, namely $\int_A X \, \mathrm{d}P$.

Theorem 2.10 (Existence)**.** *The random variable $\mathbb{E}(X \mid Y)$ exists.*

Moreover, $\mathbb{E}(X \mid Y) \geq 0$ if $X \geq 0$.

Proof. We first assume $X \geq 0$. On the σ-algebra $\mathcal{G} = \sigma(Y)$, we define the measure $\nu(A) = \int_A X \, \mathrm{d}P$ for $A \in \mathcal{G}$. We notice that if $P(A) = 0$ then $\nu(A) = 0$. It follows from the Radon-Nikodym theorem that there exists $f \geq 0$ which is $\mathcal{G}$-measurable such that $\nu(A) = \int_A f \, \mathrm{d}P$ (the student will learn the Radon-Nikodym theorem in a measure theory course, see e.g. [17, Theorem A.38]). We thus take $\mathbb{E}(X \mid Y) := f$ in this case.

In general if $X = X_+ - X_-$ with $X_\pm \geq 0$ the positive and (minus) negative parts of X, then applying the above to X_+ and X_-, we find f_+ and f_- such that $\int_A X_\pm \, \mathrm{d}P = \int_A f_\pm \, \mathrm{d}P$. Taking $\mathbb{E}(X \mid Y) := f_+ - f_-$ we get that $\mathbb{E}(X \mid Y)$ is $\mathcal{G}$-measurable and $\int_A \mathbb{E}(X \mid Y) \, \mathrm{d}P = \int_A (X_+ - X_-) \mathrm{d}P = \int_A X \, \mathrm{d}P$ as required. □

Uniqueness also follows from the Radon-Nikodym theorem (see [17, Theorem A.38]). In fact the above proof simply says that $\mathbb{E}(X \mid Y) = \frac{\mathrm{d}\nu}{\mathrm{d}\mathbb{P}}$ on the space $(\Omega, \mathcal{G})$. However we can prove it by hand to make things more self-contained, this is not difficult. We first have the following result.

Lemma 2.11. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. If Z is $\mathcal{G}$-measurable and for any $A \in \mathcal{G}$ we have $\int_A Z \, \mathrm{d}\,\mathbb{P} = 0$, then $Z = 0$ a.s.*

Proof. Let $\varepsilon > 0$. Then $\{Z \leq -\varepsilon\} \in \mathcal{G}$ because Z is $\mathcal{G}$-measurable. So we have by hypothesis $\int_{\{Z \leq -\varepsilon\}} Z \, \mathrm{d}\,\mathbb{P} = 0$. But

$$\int_{\{Z \leq -\varepsilon\}} Z \, \mathrm{d}\,\mathbb{P} \leq \int_{\{Z \leq -\varepsilon\}} (-\varepsilon) \, \mathrm{d}\,\mathbb{P} = -\varepsilon \, \mathbb{P}(Z \leq -\varepsilon) \, .$$

It follows that $0 \leq -\varepsilon \, \mathbb{P}(Z \leq -\varepsilon)$. This can only happen if $\mathbb{P}(Z \leq -\varepsilon) = 0$.

Similarly, we see that $\mathbb{P}(Z \geq \varepsilon) = 0$.

We deduce that $\mathbb{P}(-\varepsilon < Z < \varepsilon) = 1$ (just use $\mathbb{P}(B^c) = 1 - \mathbb{P}(B)$).

In particular, $\mathbb{P}(\frac{-1}{n} < Z < \frac{1}{n}) = 1$ for any n.

Let $A_n = \{\frac{-1}{n} < Z < \frac{1}{n}\}$. Then $\{Z = 0\} = \cap_n A_n$. Indeed, if $Z = 0$ then $\frac{-1}{n} < Z < \frac{1}{n}$ for any n. Conversely, if $\frac{-1}{n} < Z < \frac{1}{n}$ for any n, then taking $n \to \infty$ gives $Z = 0$. Moreover, the sets A_n are clearly decreasing. It follows that $\mathbb{P}(Z = 0) = \lim_{n \to \infty} \mathbb{P}(A_n) = 1$. Thus, $Z = 0$ a.s. as asserted. □

Corollary 2.12 (Uniqueness)**.** *The random variable $\mathbb{E}(X \mid Y)$ is unique a.s.*

Proof. Suppose two random variables Z_1, Z_2 satisfy the assumptions of Definition 2.9 and let $\mathcal{G} = \sigma(Y)$. Then $Z_1 - Z_2$ is $\mathcal{G}$-measurable and $\int_A Z_1 \,\mathrm{d}\,\mathbb{P} = \int_A X \,\mathrm{d}\,\mathbb{P} = \int_A Z_2 \,\mathrm{d}\,\mathbb{P}$ for any $A \in \mathcal{G}$. Applying Lemma 2.11 to $Z = Z_1 - Z_2$, we thus get $Z = 0$ a.s., i.e. $Z_1 = Z_2$ a.s. □

We may now give some further properties of $\mathbb{E}(X \mid Y)$ to better understand it.

Lemma 2.13. (1) *There exists a Borel function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}(X \mid Y) = \phi(Y)$.*

(2) *If Y is constant on a set B, then $\mathbb{E}(X \mid Y)$ is also constant on B.*

Moreover, if $\mathbb{P}(Y = c) > 0$, then $\mathbb{E}(X \mid Y)(\omega) = \mathbb{E}(X \mid Y = c)$ for all ω with $Y(\omega) = c$.

(3) *If Y is symmetric then $\mathbb{E}(X \mid Y)$ has the same symmetry.*

As an example of (3), if Y is even, then $\mathbb{E}(X \mid Y)$ is even.

Proof. (1) Since $\mathbb{E}(X \mid Y)$ is $\sigma(Y)$-measurable, we may find a Borel function ϕ such that $\mathbb{E}(X \mid Y) = \phi(Y)$ by the Doob-Dynkin theorem.

(2) If $Y(\omega) = c$ for $\omega \in B$, then $\mathbb{E}(X \mid Y)(\omega) = \phi(Y(\omega)) = \phi(c)$ for $\omega \in B$. Hence, $\mathbb{E}(X \mid Y)$ is also constant on B.

To find the value of the constant, we use the second condition. The set $\{Y = c\} \in \sigma(Y)$, so we have $\int_{\{Y=c\}} \mathbb{E}(X \mid Y) \,\mathrm{d}\,\mathbb{P} = \int_{\{Y=c\}} X \,\mathrm{d}\,\mathbb{P}$. But $\int_{\{Y=c\}} \mathbb{E}(X \mid Y) \,\mathrm{d}\,\mathbb{P} = \int_{\{Y=c\}} \phi(Y) \,\mathrm{d}\,\mathbb{P} = \phi(c)\,\mathbb{P}(Y = c)$. Thus, $\phi(c) = \frac{1}{\mathbb{P}(Y=c)} \int_{\{Y=c\}} X \,\mathrm{d}\,\mathbb{P} = \mathbb{E}(X \mid Y = c)$.

(3) If Y is symmetric, there exists some f such that $Y(f(\omega)) = Y(\omega)$. For example, $\Omega = \mathbb{R}$ and $f(\omega) = -\omega$ means that Y is even. Now using (1), $\mathbb{E}(X \mid Y)(f(\omega)) = \phi(Y(f(\omega))) = \phi(Y(\omega)) = \mathbb{E}(X \mid Y)(\omega)$. So $\mathbb{E}(X \mid Y)$ has indeed the same symmetry. □

Remark 2.14. Recall that $\int_A \mathbb{E}(X \mid Y) \,\mathrm{d}\,\mathbb{P} = \int_A X \,\mathrm{d}\,\mathbb{P}$ for any $A \in \sigma(Y)$. One may be tempted to apply Lemma 2.11 to $Z = \mathbb{E}(X \mid Y) - X$ to deduce that $Z = 0$, i.e. $\mathbb{E}(X \mid Y) = X$. This doesn't work because X is in general not $\sigma(Y)$-measurable, so neither is Z and we cannot apply the lemma. This remark and Lemma 2.13 show that the first condition of Definition 2.9 is a very important one.

2.3.2 Examples

Example 2.15. Let $\Omega = [0,1]$, $\mathbb{P} = \text{Leb}$,

$$X(t) = f(t) \qquad \text{and} \qquad Y(t) = t$$

for $t \in \Omega$, for some Borel function f. Find $\mathbb{E}(X \mid Y)$.

Solution. We notice that X is $\sigma(Y)$-measurable because $X(t) = f(t) = f(Y(t))$, i.e. $X = f(Y)$. It follows that $\mathbb{E}(X \mid Y) = X$. Indeed, we just saw X satisfies the first condition of Definition 2.9, and it trivially satisfies the second condition $\int_A X \,\mathrm{d}\,\mathbb{P} = \int_A X \,\mathrm{d}\,\mathbb{P}$.

The previous example can be interpreted as follows : by revealing all values $Y(t) = t$, we have obtained an exact information on how X looks like : we have $\mathbb{E}(X \mid Y) = X$. In the following examples the information provided by Y is less precise, so the mean quantity $\mathbb{E}(X \mid Y)$ becomes a bit different.

Example 2.16. Let $\Omega = [0,1]$, $\mathbb{P} = \text{Leb}$,

$$X(t) = 2t^2 \qquad \text{and} \qquad Y(t) = \begin{cases} 2 & \text{if } t \in [0, \frac{1}{2}), \\ t & \text{if } t \in [\frac{1}{2}, 1]. \end{cases}$$

Find $\mathbb{E}(X \mid Y)$.

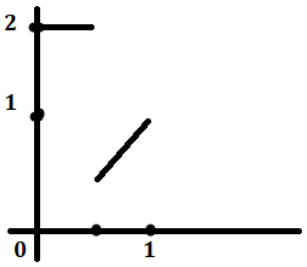


Figure 2.2 – Graph of $Y(t)$ in Example 2.16.

Solution. We always start by studying Y. Since Y is constant $Y = 2$ on $[0, \frac{1}{2})$, we know from Lemma 2.13 that for any $t \in [0, \frac{1}{2})$, we have

$$\mathbb{E}(X \mid Y)(t) = \mathbb{E}(X \mid Y = 2) = \frac{1}{P([0, \frac{1}{2}))} \int_{[0, \frac{1}{2})} X \,\mathrm{d}\,\mathbb{P}$$
$$= \frac{1}{1/2} \int_0^{1/2} 2t^2 \,\mathrm{d}t = \frac{4}{3 \cdot 8} = \frac{1}{6}.$$

It remains to find $\mathbb{E}(X \mid Y)$ for $t \in [\frac{1}{2}, 1]$. Let $B \subset [\frac{1}{2}, 1]$ be a Borel set. Then $\int_{Y^{-1}(B)} \mathbb{E}(X \mid Y) \,\mathrm{d}\,\mathbb{P} = \int_{Y^{-1}(B)} X \,\mathrm{d}\,\mathbb{P}$. But here $Y^{-1}(B) = \{t : Y(t) \in B\} = B$. So

$\int_B \mathbb{E}(X \mid Y)\,\mathrm{d}\,\mathbb{P} = \int_B X\,\mathrm{d}\,\mathbb{P}$. Since $B \subset [\frac{1}{2}, 1]$ is arbitrary, we get $\mathbb{E}(X \mid Y)(t) = X(t)$ for all $t \in [\frac{1}{2}, 1]$. Conclusion :

$$\mathbb{E}(X \mid Y)(t) = \begin{cases} \frac{1}{6} & \text{if } t \in [0, \frac{1}{2}), \\ 2t^2 & \text{if } t \in [\frac{1}{2}, 1]. \end{cases} \tag{2.1}$$

Example 2.17. Let $\Omega = [0, 1]$, $\mathbb{P} = \text{Leb}$,

$$X(t) = \mathrm{e}^t \qquad \text{and} \qquad Y(t) = 1 - |2t - 1|\,.$$

Find $\mathbb{E}(X \mid Y)$.

Solution. We notice that $Y(t) = \begin{cases} 2t & \text{if } 0 \le t \le \frac{1}{2} \\ 2 - 2t & \text{if } \frac{1}{2} \le t \le 1. \end{cases}$

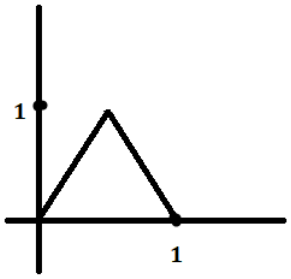


Figure 2.3 – Graph of $Y(t)$ in Example 2.17.

By drawing the graph, we notice the symmetry $Y(1 - t) = Y(t)$. So we know from Lemma 2.13 that $\mathbb{E}(X \mid Y)$ should have the same symmetry.

To find $\mathbb{E}(X \mid Y)$ we assume Z is a random variable satisfying both conditions of Definition 2.9. Then $Z = \phi(Y)$ for some ϕ and $\int_{Y^{-1}(B)} Z\,\mathrm{d}\,\mathbb{P} = \int_{Y^{-1}(B)} X\,\mathrm{d}\,\mathbb{P}$.

Now from the graph we see that for $B \subset [0, 1]$, we have

$$Y^{-1}(B) = \{t : Y(t) \in B\} = \{t : 2t \in B \text{ or } 2 - 2t \in B\} = \frac{B}{2} \bigcup 1 - \frac{B}{2}\,,$$

where $\alpha B = \{\alpha b : b \in B\}$ and $r \pm B = \{r \pm b : b \in B\}$. The sets $\frac{B}{2}$ and $1 - \frac{B}{2}$ are essentially disjoint (they may only meet at the point $t = \frac{1}{2}$), so $\int_{Y^{-1}(B)} = \int_{\frac{B}{2} \cup (1 - \frac{B}{2})} = \int_{\frac{B}{2}} + \int_{1 - \frac{B}{2}}$.

Now

$$\begin{aligned} \int_{Y^{-1}(B)} Z\,\mathrm{d}\,\mathbb{P} &= \int_{\frac{B}{2}} \phi(Y(t))\,\mathrm{d}t + \int_{1 - \frac{B}{2}} \phi(Y(t))\,\mathrm{d}t \\ &= \int_{\frac{B}{2}} \phi(Y(t))\,\mathrm{d}t + \int_{\frac{B}{2}} \phi(Y(1 - t))\,\mathrm{d}t \end{aligned}$$

We now use that $Y(1-t) = Y(t)$ to deduce that $\int_{Y^{-1}(B)} Z \,\mathrm{d}\,\mathbb{P} = 2\int_{\frac{B}{2}} \phi(Y(t))\,\mathrm{d}t$.

Similarly, $\int_{Y^{-1}(B)} X \,\mathrm{d}\,\mathbb{P} = \int_{\frac{B}{2}} X(t)\,\mathrm{d}t + \int_{\frac{B}{2}} X(1-t)\,\mathrm{d}t$. Here X is in contrast not symmetric.

We thus showed that for any Borel $B \subset [0,1]$, we have $\int_{\frac{B}{2}} 2\phi(Y(t))\,\mathrm{d}t = \int_{\frac{B}{2}} (X(t) + X(1-t))\,\mathrm{d}t$. By varying $B \subset [0,1]$, we deduce that $\int_C 2\phi(Y(t))\,\mathrm{d}t = \int_C (X(t) + X(1-t))\,\mathrm{d}t$ for any Borel $C \subset [0, \frac{1}{2}]$. It follows that $2\phi(Y(t)) = (X(t) + X(1-t))$ for a.e. $t \in [0, \frac{1}{2}]$. In other words, $\phi(Y(t)) = \frac{X(t)+X(1-t)}{2}$. For $t \in [\frac{1}{2}, 1]$, we have $\phi(Y(t)) = \phi(Y(1-t)) = \frac{X(1-t)+X(t)}{2}$.

Conclusion : we showed that if Z is a random variable satisfying both conditions of Definition 2.9, we must have $Z(t) = \frac{X(t)+X(1-t)}{2}$ for all $t \in \Omega$. Thus,

$$\mathbb{E}(X \mid Y)(t) = \frac{X(t) + X(1-t)}{2} = \frac{\mathrm{e}^t + \mathrm{e}^{1-t}}{2}.$$

Example 2.18. Let Ω be the unit square $[0,1] \times [0,1]$ and $\mathbb{P} = \mathrm{Leb}$ (this just measures the area of sets $B \subset [0,1]^2$). Suppose X and Y are random variables on Ω with joint density

$$p_{X,Y}(x,y) = \begin{cases} x + y & \text{if } x, y \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

Show that $\mathbb{E}(X \mid Y) = \frac{2+3Y}{3+6Y}$.

Solution. The random variable $\frac{2+3Y}{3+6Y}$ is $\sigma(Y)$-measurable since it has the form $\phi(Y)$. It remains to check the second condition, $\int_{\{Y \in B\}} \mathbb{E}(X \mid Y) \,\mathrm{d}\,\mathbb{P} = \int_{\{Y \in B\}} X \,\mathrm{d}\,\mathbb{P}$. Recall the *change of variables formula :*

$$\mathbb{E}(h(X,Y)) = \int_{\mathbb{R}^2} h(x,y)\,\mathrm{d}P_X(x,y) = \int_{\mathbb{R}^2} h(x,y) p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y\,.$$

This implies $\int_{\{Y \in B\}} X \,\mathrm{d}\,\mathbb{P} = \int_\Omega X \cdot \mathbf{1}_{\{Y \in B\}} \,\mathrm{d}\,\mathbb{P} = \int_0^1 \int_0^1 x \cdot \mathbf{1}_{\{y \in B\}} \cdot (x+y)\,\mathrm{d}x\mathrm{d}y = \int_B \int_0^1 x(x+y)\,\mathrm{d}x\mathrm{d}y = \int_B (\frac{1}{3} + \frac{y}{2})\,\mathrm{d}y$.

Similarly, $\int_{\{Y \in B\}} \phi(Y) \,\mathrm{d}\,\mathbb{P} = \int_\Omega \phi(Y) \cdot \mathbf{1}_{\{Y \in B\}} \,\mathrm{d}\,\mathbb{P} = \int_B \int_0^1 \phi(y)(x+y)\,\mathrm{d}x\mathrm{d}y = \int_B \phi(y)(\frac{1}{2} + y)\,\mathrm{d}y$.

We finally replace $\phi(y) = \frac{2+3y}{3+6y}$. Then we find indeed the two results coincide : we get $\int_{\{Y \in B\}} \phi(Y) \,\mathrm{d}\,\mathbb{P} = \int_{\{Y \in B\}} X \,\mathrm{d}\,\mathbb{P}$. This completes the proof.

2.4 Conditioning on a sigma algebra

We note that Definition 2.9 of $\mathbb{E}(X \mid Y)$ only depends on Y via $\sigma(Y)$. Indeed,

Lemma 2.19. *If $\sigma(Y) = \sigma(Y')$ then $\mathbb{E}(X \mid Y) = \mathbb{E}(X \mid Y')$ a.s.*

Proof. Let $\mathcal{G} = \sigma(Y) = \sigma(Y')$. Then $\int_A \mathbb{E}(X \mid Y)\,\mathrm{d}\,\mathbb{P} = \int_A X\,\mathrm{d}\,\mathbb{P} = \int_A \mathbb{E}(X \mid Y')\,\mathrm{d}\,\mathbb{P}$ for any $A \in \mathcal{G}$. Applying Lemma 2.11 to $Z = \mathbb{E}(X \mid Y) - \mathbb{E}(X \mid Y')$, we get $Z = 0$ a.s. □

This suggests we generalize Definition 2.9 to arbitrary σ-algebras, not just $\sigma(Y)$, as follows.

Definition 2.20. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let X be an integrable random variable and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. We define the *conditional expectation of X given $\mathcal{G}$* to be the random variable $\mathbb{E}(X \mid \mathcal{G})$ such that

(1) $\mathbb{E}(X \mid \mathcal{G})$ is $\mathcal{G}$-measurable,

(2) For any $A \in \mathcal{G}$,
$$\int_A \mathbb{E}(X \mid \mathcal{G})\,\mathrm{d}\,\mathbb{P} = \int_A X\,\mathrm{d}\,\mathbb{P}\ .$$

Comparing with Definition 2.9 we see that $\mathbb{E}(X \mid Y) = \mathbb{E}(X \mid \sigma(Y))$.

Theorem 2.21. *The random variable $\mathbb{E}(X \mid \mathcal{G})$ exists and is unique a.s.*

Proof. The proof is exactly the same as Theorem 2.10 and Corollary 2.12 by replacing $\sigma(Y)$ by $\mathcal{G}$. □

Definition 2.22. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. We define the *conditional probability of an event $A \in \mathcal{F}$ given $\mathcal{G}$* to be the random variable
$$\mathbb{P}(A \mid \mathcal{G}) = \mathbb{E}(\mathbf{1}_A \mid \mathcal{G})\,.$$

Example 2.23. Suppose $\mathcal{G} = \{\emptyset, \Omega\}$. Find $\mathbb{E}(X \mid \mathcal{G})$.

Solution. We need $Z = \mathbb{E}(X \mid \mathcal{G})$ to be $\mathcal{G}$-measurable. So for any Borel $B \subset \mathbb{R}$, we need $Z^{-1}(B) \in \{\emptyset, \Omega\}$. This means that $\{\omega : Z(\omega) \in B\}$ is either all of Ω or empty. This implies $Z(\omega) = c$ is a constant function. Indeed, if Z took two distinct values, there would be at least two distinct non-empty sets $Z^{-1}(\{c_1\})$ and $Z^{-1}(\{c_2\})$ in $\mathcal{G}$.

But Z satisfies $\int_\Omega Z\,\mathrm{d}\,\mathbb{P} = \int_\Omega X\,\mathrm{d}\,\mathbb{P}$. So $c = c\,\mathbb{P}(\Omega) = \int_\Omega Z\,\mathrm{d}\,\mathbb{P} = \int_\Omega X\,\mathrm{d}\,\mathbb{P} = \mathbb{E}(X)$. This shows that $\mathbb{E}(X \mid \mathcal{G}) = c = \mathbb{E}(X)$.

Example 2.23 tells us that the information provided by $\mathcal{G} = \{\emptyset, \Omega\}$ is completely useless. We have $\mathbb{E}(X \mid \mathcal{G}) = \mathbb{E}(X)$, so knowing $\mathcal{G}$ did not provide any additional information over just taking the mean of X over the whole space.

Example 2.24. Suppose X is $\mathcal{G}$-measurable. Find $\mathbb{E}(X \mid \mathcal{G})$.

Solution. If X is $\mathcal{G}$-measurable then $\mathbb{E}(X \mid \mathcal{G}) = X$. Indeed, X clearly satisfies both conditions of Definition 2.20 in this case.

Example 2.24 deals with the opposite extreme situation : if we know that X is $\mathcal{G}$-measurable, then the information provided by $\mathcal{G}$ is all we need to completely determine X : we have $\mathbb{E}(X \mid \mathcal{G}) = X$.

2.5 Properties of conditional expectation

Lemma 2.25. *The following properties hold true a.s. :*

(1) $\mathbb{E}(aX + bY \mid \mathcal{G}) = a\,\mathbb{E}(X \mid \mathcal{G}) + b\,\mathbb{E}(Y \mid \mathcal{G})$,

(2) $\mathbb{E}(\mathbb{E}(X \mid \mathcal{G})) = \mathbb{E}(X)$,

(3) $\mathbb{E}(XY \mid \mathcal{G}) = X\,\mathbb{E}(Y \mid \mathcal{G})$ *if X is $\mathcal{G}$-measurable,*

(4) $\mathbb{E}(X \mid \mathcal{G}) = \mathbb{E}(X)$ *if X is independent of $\mathcal{G}$,*

(5) $\mathbb{E}(\mathbb{E}(X \mid \mathcal{G}) \mid \mathcal{H}) = \mathbb{E}(X \mid \mathcal{H})$ *if $\mathcal{H} \subset \mathcal{G}$.*

Proof. (1) $a\,\mathbb{E}(X \mid \mathcal{G}) + b\,\mathbb{E}(Y \mid \mathcal{G})$ is $\mathcal{G}$-measurable. Moreover, given $A \in \mathcal{G}$,

$$\begin{aligned}\int_A [a\,\mathbb{E}(X \mid \mathcal{G}) + b\,\mathbb{E}(Y \mid \mathcal{G})]\,\mathrm{d}\,\mathbb{P} &= a\int_A \mathbb{E}(X \mid \mathcal{G})\,\mathrm{d}\,\mathbb{P} + b\int_A \mathbb{E}(Y \mid \mathcal{G})\,\mathrm{d}\,\mathbb{P} \\ &= a\int_A X\,\mathrm{d}\,\mathbb{P} + b\int Y\,\mathrm{d}\,\mathbb{P} = \int_A (aX + bY)\,\mathrm{d}\,\mathbb{P}\ .\end{aligned}$$

We thus have $a\,\mathbb{E}(X \mid \mathcal{G}) + b\,\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}(aX + bY \mid \mathcal{G})$ by uniqueness.

(2) $\mathbb{E}(\mathbb{E}(X \mid \mathcal{G})) = \int_\Omega \mathbb{E}(X \mid \mathcal{G})\,\mathrm{d}\,\mathbb{P} = \int_\Omega X\,\mathrm{d}\,\mathbb{P} = \mathbb{E}(X)$ by the 2nd condition of Definition 2.20 with $A = \Omega$.

(3) $X\,\mathbb{E}(Y \mid \mathcal{G})$ is $\mathcal{G}$-measurable, as a product of $\mathcal{G}$-measurable functions. To check the second condition, we first consider $X = \mathbf{1}_B$, $B \in \mathcal{G}$. In this case,

$$\int_A \mathbf{1}_B\,\mathbb{E}(Y \mid \mathcal{G})\,\mathrm{d}\,\mathbb{P} = \int_{A\cap B} \mathbb{E}(Y \mid \mathcal{G})\,\mathrm{d}\,\mathbb{P} = \int_{A\cap B} Y\,\mathrm{d}\,\mathbb{P} = \int_A \mathbf{1}_B Y\,\mathrm{d}\,\mathbb{P}$$

for any $A \in \mathcal{G}$. Thus, $\mathbf{1}_B\,\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}(\mathbf{1}_B Y \mid \mathcal{G})$ by uniqueness.

Using (1), this implies that $X\,\mathbb{E}(Y \mid \mathcal{G}) = \mathbb{E}(XY \mid \mathcal{G})$ for any X of the form $X = \sum_{j=1}^m a_j \mathbf{1}_{B_j}$ with $B_j \in \mathcal{G}$ (these are called *simple functions*). Such functions are dense in $L^1(\Omega, \mathcal{G}, \mathbb{P})$, so we get (3) for any integrable X (admitted).

(4) The constant function $\mathbb{E}(X)$ is $\mathcal{G}$-measurable. If X is independent of $\mathcal{G}$ then X and $\mathbf{1}_A$ are independent for any $A \in \mathcal{G}$. Hence,

$$\int_A \mathbb{E}(X)\,\mathrm{d}\,\mathbb{P} = \mathbb{E}(X)\,\mathbb{P}(A) = \mathbb{E}(X)\,\mathbb{E}(\mathbf{1}_A) = \mathbb{E}(X\mathbf{1}_A) = \int_A X\,\mathrm{d}\,\mathbb{P}\ .$$

Thus, $\mathbb{E}(X) = \mathbb{E}(X \mid \mathcal{G})$ by uniqueness.

(5) $\mathbb{E}(X \mid \mathcal{H})$ is $\mathcal{H}$-measurable. Moreover,

$$\int_A \mathbb{E}(X \mid \mathcal{H}) \, \mathrm{d}\,\mathbb{P} = \int_A X \, \mathrm{d}\,\mathbb{P} = \int_A \mathbb{E}(X \mid \mathcal{G}) \, \mathrm{d}\,\mathbb{P}$$

for any $A \in \mathcal{H} \subset \mathcal{G}$. Thus, $\mathbb{E}(X \mid \mathcal{H}) = \mathbb{E}(\mathbb{E}(X \mid \mathcal{G}) \mid \mathcal{H})$ by uniqueness. □

To give one more general property, we recall the concept of a *convex function*. We say that $\varphi : \mathbb{R} \to \mathbb{R}$ is *convex* if for any $\lambda \in [0, 1]$,

$$\varphi(\lambda x + (1 - \lambda)y) \leq \lambda\varphi(x) + (1 - \lambda)\varphi(y) \, .$$

This means that the graph of φ lies below the line segment between $\varphi(x)$ and $\varphi(y)$. For example, x^2 and e^x are convex. In general if φ is twice differentiable, it is equivalent to ask that $\varphi'' \geq 0$.

Proposition 2.2 (Jensen's inequality)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex and X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\varphi(X)$ is also integrable. Then*

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X)) \, .$$

More generally, if $\mathcal{G} \subset \mathcal{F}$ is a σ-algebra then

$$\varphi(\mathbb{E}(X \mid \mathcal{G})) \leq \mathbb{E}(\varphi(X) \mid \mathcal{G}) \qquad a.s.$$

Proof. Admitted. See e.g. [2, Theorem 1.7,Theorem 2.10]. □

We conclude with a property beyond the scope of this course, the student will understand its meaning after taking a course on functional analysis. We put it here for completeness (and it is easy to prove).

Proposition 2.3. *If $X \in L^2(\mathcal{F})$ and $\mathcal{G} \subset \mathcal{F}$, then $\mathbb{E}(X \mid \mathcal{G})$ is the orthogonal projection of X onto the subspace $L^2(\mathcal{G})$ of $L^2(\mathcal{F})$.*

Proof. Admitted. See e.g. [5, Solution 2.17]. □

2.6 Further examples

Example 2.26. Safeya made a Feteer Meshaltet for her two daughters. Eating more than half of it will give an indigestion to anyone. While she is having tea with the neighbors, the older daughter helps herself to a piece of the pie. Then the younger daughter takes a piece of what is left from her sister.

Assume the daughters choose each piece uniformly at random over what is available. What is the expected size of what is left of the pie after both daughters ate, given that neither daughter gets an indigestion ?

Solution. If $[0,1]$ represents the whole pie, then the first daughter chooses her piece $x \in [0,1]$ using the uniform density $p_X(x) = 1$. Then the second daughter chooses her piece y uniformly from the remaining $[0, 1-x]$, hence with density $p_Y(y) = \frac{1}{1-x}$. It follows that the sample space is

$$\Omega = \{(x,y) : 0 \le x \le 1,\, 0 \le y \le 1-x\}$$

and we may take as joint density

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) = \frac{1}{1-x}$$

if $(x,y) \in \Omega$ and zero otherwise.

The event "neither daughter gets an indigestion" is $A = \{(x,y) : x, y < \frac{1}{2}\}$.

What is left after both daughters ate is $Z(x,y) = 1 - x - y$. We should thus compute $\mathbb{E}(Z \mid A)$. We have

$$\mathbb{P}(A) = \int_A p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = \int_0^{1/2}\int_0^{1/2} \frac{1}{1-x}\,\mathrm{d}x\mathrm{d}y = \int_0^{1/2} \ln 2 = \ln\sqrt{2}\,.$$

Thus,

$$\begin{aligned}\mathbb{E}(Z \mid A) &= \frac{1}{\mathbb{P}(A)}\int_A (1-x-y)p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \frac{1}{\ln\sqrt{2}}\int_0^{1/2}\int_0^{1/2}\frac{1-x-y}{1-x}\,\mathrm{d}x\mathrm{d}y = \frac{1}{\ln\sqrt{2}}\int_0^{1/2}\int_0^{1/2}\Big(1-\frac{y}{1-x}\Big)\,\mathrm{d}x\mathrm{d}y \\ &= \frac{1}{\ln\sqrt{2}}\int_0^{1/2}\Big(\frac{1}{2} - y\ln 2\Big)\,\mathrm{d}y = \frac{\frac{1}{4}-\frac{1}{8}\ln 2}{\frac{1}{2}\ln 2} = \frac{1-\ln\sqrt{2}}{\ln 4}\,.\end{aligned}$$

Example 2.27. Let X, Y be integrable random variables with joint density $p_{X,Y}(x,y)$. Show that

$$\mathbb{E}(X \mid Y) = \frac{\int_{\mathbb{R}} x p_{X,Y}(x,Y)\,\mathrm{d}x}{\int_{\mathbb{R}} p_{X,Y}(x,Y)\,\mathrm{d}x} \qquad a.s.$$

Solution. This generalizes Example 2.18. As before we use the change of variables formula to see that

$$\begin{aligned}\int_{\{Y\in B\}} X\,\mathrm{d}\,\mathbb{P} = \int_\Omega X\mathbf{1}_{\{Y\in B\}}\,\mathrm{d}\,\mathbb{P} &= \int_{\mathbb{R}}\int_{\mathbb{R}} x\mathbf{1}_{\{y\in B\}}\cdot p_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_B \Big(\int_{\mathbb{R}} x p_{X,Y}(x,y)\,\mathrm{d}x\Big)\mathrm{d}y\end{aligned}$$

and if ϕ is such that $\mathbb{E}(X \mid Y) = \phi(Y)$ (using Doob-Dynkin), we have

$$\int_{\{Y\in B\}} \phi(Y)\,\mathrm{d}\,\mathbb{P} = \int_\Omega \phi(Y)\mathbf{1}_{\{Y\in B\}}\,\mathrm{d}\,\mathbb{P} = \int_B \phi(y)\Big(\int_{\mathbb{R}} p_{X,Y}(x,y)\,\mathrm{d}x\Big)\mathrm{d}y\,.$$

But $\int_{\{Y\in B\}} X\,\mathrm{d}\,\mathbb{P} = \int_{\{Y\in B\}} \phi(Y)\,\mathrm{d}\,\mathbb{P}$. Since $B \in \mathcal{B}(\mathbb{R})$ is arbitrary, we deduce that $\int_{\mathbb{R}} x p_{X,Y}(x,y)\,\mathrm{d}x = \phi(y)\int_{\mathbb{R}} p_{X,Y}(x,y)\,\mathrm{d}x$ a.s. Thus, $\phi(y) = \frac{\int_{\mathbb{R}} x p_{X,Y}(x,y)\,\mathrm{d}x}{\int_{\mathbb{R}} p_{X,Y}(x,y)\,\mathrm{d}x}$ a.s. Since $\mathbb{E}(X \mid Y) = \phi(Y)$, this proves the result.

2.7 Further results

We now give a very useful property for applications :

Theorem 2.28. *If X, Y are independent and f is a Borel function, then*

$$\mathbb{E}(f(X,Y) \mid Y) = \mathbb{E}(f(\cdot, Y)) \qquad a.s., \tag{2.1}$$

where $\mathbb{E}(f(\cdot, Y))$ means taking the average w.r.t. X only while Y is fixed. We often denote this by $\mathbb{E}_X(f)$.

More generally, if $X_1, \ldots, X_n$ are independent then

$$\mathbb{E}(f(X_1, \ldots, X_n) \mid X_{i_1}, \ldots, X_{i_k}) = \mathbb{E}_{X_{j_1}, \ldots, X_{j_{n-k}}}(f)\,,$$

where $\{j_1, \ldots, j_{n-k}\} = \{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$.

Proof. $\mathbb{E}(f(\cdot, Y))$, as a function of Y, is $\sigma(Y)$-measurable. Next, if X and Y have distributions P_X and P_Y, respectively, their joint distribution is the product measure $\mathrm{d}P_{X,Y}(x,y) = \mathrm{d}P_X(x)\mathrm{d}P_Y(y)$ by independence. By change of variables we have $\int_{Y \in B} f(X,Y)\,\mathrm{d}\,\mathbb{P} = \int_B \int_{\mathbb{R}} f(x,y)\,\mathrm{d}P_X(x)\mathrm{d}P_Y(y)$. On the other hand, $\mathbb{E}(X \mid Y) = \phi(Y)$ by Doob-Dynkin and $\int_{Y \in B} \phi(Y)\,\mathrm{d}\,\mathbb{P} = \int_B \int_{\mathbb{R}} \phi(y)\,\mathrm{d}P_X(x)\,\mathrm{d}P_Y(y) = \int_B \phi(y)\,\mathrm{d}P_Y(y)$. If we take $\phi(y) = \mathbb{E}(f(\cdot, y)) = \int_{\mathbb{R}} f(x,y)\,\mathrm{d}P_X(x)$, we find that $\int_{\{Y \in B\}} \phi(Y)\,\mathrm{d}\,\mathbb{P} = \int_B \int_{\mathbb{R}} f(x,y)\,\mathrm{d}P_X(x)\mathrm{d}P_Y(y) = \int_{\{Y \in B\}} f(X,Y)\,\mathrm{d}\,\mathbb{P}$. Thus, the RHS of (2.1) satisfies both conditions of Definition 2.9.

The general case is similar using the multivariable Doob-Dynkin result, replacing $\{Y \in B\}$ by the event "$\{X_{i_1} \in B_1\}, \ldots, \{X_{i_k} \in B_k\}$". □

To conclude this chapter we briefly mention a few words on the topic of *regular conditional probabilities*. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Recall that if $B \in \mathcal{F}$, $\mathbb{P}(B) \neq 0$, then $\mathbb{P}(\cdot \mid B)$ is a probability measure (the easy proof was done in Stat 201, the student can check this again as an exercise). What about $\mathbb{P}(\cdot \mid \mathcal{G})$ for $\mathcal{G} \subset \mathcal{F}$, a sub-$\sigma$-algebra ? This is a random variable. Is it a random probability measure ? Not necessarily. One easily sees that $0 \leq \mathbb{P}(A \mid \mathcal{G}) \leq 1$ a.s. and $\mathbb{P}(\cup_n A_n \mid \mathcal{G}) = \sum_n \mathbb{P}(A_n \mid \mathcal{G})$ a.s. by Lemma 2.25(1) and monotone convergence. This "a.s." however is a bit problematic here because the bad null set can depend on the sequence (A_n).

It is natural to ask if one can choose a "regular version" of $\mathbb{P}(\cdot \mid \mathcal{G})(\omega)$ which is a probability measure on Ω. This is generally possible, except on rather bad probability spaces. More precisely, if Ω is a separable, complete metric space (known as Polish spaces, e.g. $\mathbb{R}^n$ or a closed subset of it) and $\mathcal{F}$ is its family of

Borel subsets, then it is possible to find such regular conditional probabilities, see e.g. [6, Theorem 10.2.2].

A very useful application of this theory is the *disintegration theorem*. This says that if Ω (again assumed to be Polish) is a product space $\Omega = C_1 \times C_2$, with product σ-algebra $\mathcal{F}_1 \otimes \mathcal{F}_2$, and if $\mathcal{F}_2$ is the family of Borel subsets of C_2, then for any probability $\mathbb{P}$ on Ω, we can find *conditional distributions* $\mathbb{P}_x$ such that

$$\mathbb{P}(A \times B) = \int_A \mathbb{P}_x(B)\, \mathrm{d}\mu(x)\,. \tag{2.2}$$

Here, if $\pi_1 : \Omega \to C_1$, $\pi_1(x,y) = x$, then $\mu = \mathbb{P} \circ \pi_1^{-1}$ is the image measure of $\mathbb{P}$ on C_1. Such a decomposition is trivially true if $\mathbb{P}$ is a product measure $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$, in this case $\mathbb{P}(A \times B) = \mathbb{P}_1(A)\,\mathbb{P}_2(B)$ so we may take $\mathbb{P}_x = \mathbb{P}_2$ and $\mu = \mathbb{P}_1$. The power of (2.2) is that it holds for any $\mathbb{P}$. This result is often used when studying random variables which are not independent (for independent random variables, product measures suffice). The theorem also guarantees the following properties : for each x, $\mathbb{P}_x$ is a probability measure on $(C_2, \mathcal{F}_2)$ and for each $B \in \mathcal{F}_2$, the map $x \mapsto \mathbb{P}_x(B)$ is $\mathcal{F}_1$-measurable. Formula (2.2) can be generalized to integrate any measurable subsets (not just rectangles) via sections, and in fact $\int g\, \mathrm{d}\,\mathbb{P} = \iint g(x,y)\, \mathrm{d}\,\mathbb{P}_x(y) \mathrm{d}\mu(x)$. Again see [6, Theorems 10.2.1-10.2.2] for a proof.

2.8 Exercises

1. Let A, B be events. Find $\mathbb{E}(\mathbf{1}_A \mid B)$, assuming $0 < \mathbb{P}(B)$.
2. Let A, B be events. Find $\mathbb{E}(\mathbf{1}_A \mid \mathbf{1}_B)$, assuming $0 < \mathbb{P}(B) < 1$.
3. Study Example 2.6 for four coins and an adequate Y.
4. Let $\Omega = [0,1]^2$, $\mathbb{P} = \mathrm{Leb}$. If X, Y are random variables with joint density

$$p_{X,Y}(x,y) = \begin{cases} \frac{3}{2}(x^2+y^2) & \text{if } x,y \in [0,1], \\ 0 & \text{otherwise,} \end{cases}$$

   find $\mathbb{E}(X \mid Y)$.

5. Let Ω be the unit disc $\{(x,y) : x^2 + y^2 \leq 1\}$ with the Lebesgue measure normalized so that $\mathbb{P}(\Omega) = 1$. Since $\mathrm{Area}(\Omega) = \pi 1^2 = \pi$, this means that

$$\mathbb{P}(A) = \frac{1}{\pi} \iint_A \mathrm{d}x\, \mathrm{d}y$$

   for Borel $A \subseteq \Omega$.

Suppose X, Y are the projections onto the x and y axes, respectively :

$$X(x,y) = x\,, \qquad Y(x,y) = y$$

for any $(x,y) \in \Omega$. Find $\mathbb{E}(X^2 \mid Y)$.

6. Let $\Omega = [0,1]$, $\mathbb{P} = \text{Leb}$,

$$X(t) = 2t^2\,, \qquad Y(t) = \begin{cases} 2t & \text{if } 0 \le t < \frac{1}{2}, \\ 2t-1 & \text{if } \frac{1}{2} \le t < 1. \end{cases}$$

Find $\mathbb{E}(X \mid Y)$.

7. Let $\Omega = [0,1]$, $\mathbb{P} = \text{Leb}$. Let $Y : \Omega \to \mathbb{R}$ satisfy $Y(t) = Y(1-t)$ and suppose Y is injective when restricted to $[0, \frac{1}{2}]$. Find $\mathbb{E}(X \mid Y)$.

8. Prove that if $B \in \mathcal{G}$, then $\mathbb{E}(\mathbb{E}(X \mid \mathcal{G}) \mid B) = \mathbb{E}(X \mid B)$.

9. Let $A \in \mathcal{F}$. Suppose $\mathcal{G} = \{\emptyset, A, A^c, \Omega\}$. Find $\mathbb{E}(X \mid \mathcal{G})$.

Chapter 3

Markov Chains

3.1 Introduction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *stochastic process* is a measurable function $X(t, \omega)$ from $\Lambda \times \Omega \to \mathbb{R}$. Here Λ represents a time set. Usually $\Lambda = \mathbb{N}$ and we speak of processes in *discrete time*, or $\Lambda = [0, \infty)$ and we speak of processes in *continuous time*.

It is customary to regard $X(t, \omega)$ as a family of random variables $X_t(\omega)$ which are indexed by time. The map $t \mapsto X_t(\omega)$ represents the path taken by a fixed sample as time goes on. The map $\omega \mapsto X_t(\omega)$ represents the random variable at time t. Roughly speaking, $\{X_t(\omega)\}$ represents a system that evolves randomly with time. The X_t represents our observation at time t, such observations can be taken at discrete times (taking photos) or continuous times (taking a video).

In the present chapter and the next one, we assume we have *discrete* random variables. All random variables $(X_t)_{t \in \Lambda}$ take value in some countable set $\{\varepsilon_1, \varepsilon_2, \ldots\}$ which we call the *states* of the system. We call $S = \{\varepsilon_1, \varepsilon_2, \ldots\}$ the *state space*. The system jumps from state to state randomly as time goes on.

In this chapter we take $\Lambda = \mathbb{N}$ and speak of *Markov chains*. We shall study transition probabilities, recurrent vs transient states, limiting probabilities, and of course illustrate these concepts on examples. In the next chapter we shall take $\Lambda = [0, \infty)$ and speak of *continuous Markov processes*. Continuous Markov processes include the fundamental *Poisson process* and the *Birth/Death process*, which brings us to the doors of *Queueing Theory*.

What if the random variables $X_t(\omega)$ are *continuous* random variables instead of discrete ? In that case the theory becomes more complicated; the most important example of such a process is the *Brownian motion*.

3.2 Markov chains : Basic properties

A Markov chain is a family of discrete random variables $\{X_n\}_{n\in\mathbb{N}}$ with values in some set $\{\varepsilon_1, \varepsilon_2, \ldots\}$. These can be interpreted as describing a physical system that evolves randomly. The system starts at some initial state ε_i, which is picked at random by X_0, then we track the consecutive transitions $X_0 \to X_1 \to X_2 \to \ldots$ at times $n = 0, 1, 2, \ldots$. At each step, X_n picks a state ε_j at random. We denote

$$p_i^0 = \mathbb{P}(X_0 = \varepsilon_i)$$

the probability that the system is at state ε_i at time $n = 0$ and make the following fundamental hypothesis :

$$\mathbb{P}(X_{n+1} = \varepsilon \mid X_0, X_1, \ldots, X_n) = \mathbb{P}(X_{n+1} = \varepsilon \mid X_n)\,. \tag{3.1}$$

Property (3.1) says that the state of the system at time $n+1$ only depends on the state at time n. We do not care how the system evolved from 0 to n; such additional information on the path does not bring any improvement to predicting the transition from n to $n+1$. For example, if we know that at time $n = 1$ the system was at ε_k and we want to know how likely is the system to be in ε_m at time $n = 2$, then knowing that the system at time $n = 0$ started at ε_1 or ε_2 doesn't help our prediction. We are dealing with a system that *lacks memory* : its choice to jump to the next state is only affected by where it stands at present, not how it actually reached this state. [1]

Property (3.1) is called the *Markov property*.

Some Markov chains also satisfy

$$\mathbb{P}(X_{n+1} = \varepsilon_j \mid X_n = \varepsilon_i) = \mathbb{P}(X_1 = \varepsilon_j \mid X_0 = \varepsilon_i) \qquad \text{for all } n\,. \tag{3.2}$$

Property (3.2) says that the probability that the system goes from ε_i to ε_j does not change with time. We say the Markov chain is *time homogeneous*.

Given a time-homogeneous Markov chain, we call

$$p_{i,j} = \mathbb{P}(X_{n+1} = \varepsilon_j \mid X_n = \varepsilon_i)$$

the *transition probabilities*. They are independent of n by time homogeneity.

1. A more formal way to understand this is to use Doob-Dynkin : we have $\mathbb{P}(A_{n+1} \mid X_0, X_1, \ldots, X_n) = \phi(X_0, \ldots, X_n)$ and $\mathbb{P}(A_{n+1} \mid X_n) = \psi(X_n)$ for some Borel ϕ, ψ. The Markov property says that $\phi(X_0, \ldots, X_n) = \psi(X_n)$, i.e. the probability does not depend on the data $X_0, \ldots, X_{n-1}$; if we vary them we still get $\psi(X_n)$ anyway.

Let

$$p_{i,j}(n) = \mathbb{P}(X_n = \varepsilon_j \mid X_0 = \varepsilon_i)$$

be the probability that the system goes from ε_i to ε_j in n steps.

A matrix $(a_{i,j})$ is called *stochastic* if $a_{i,j} \geq 0$ and $\sum_j a_{i,j} = 1$ for any i. In other words, each row sum is 1.

Lemma 3.1. *The matrix $P(n) = (p_{i,j}(n))$ is stochastic.*

Proof. We have for any i, $\sum_j p_{i,j}(n) = \sum_j \mathbb{P}(X_n = \varepsilon_j \mid X_0 = \varepsilon_i)$. The events $\{X_n = \varepsilon_j\}$ are mutually exclusive for different j, so we get $\sum_j p_{i,j}(n) = \mathbb{P}(X_n = \varepsilon_1 \text{ or } \varepsilon_2 \text{ or } \cdots \mid X_0 = \varepsilon_i) = 1$ (this is just the probability that the system moved from ε_i to some state ε after n steps). Here we used that $\mathbb{P}(\cdot \mid X_0 = \varepsilon_i)$ is a probability measure. □

Proposition 3.1 (Kolmogorov-Chapman Equations)**.** *We have*

$$p_{i,j}(n+m) = \sum_k p_{i,k}(n) p_{k,j}(m)\,.$$

Proof. The events $\{X_m = \varepsilon_k\}$ are mutually exclusive for different k, and their union is Ω. Hence,

$$\begin{aligned}\mathbb{P}(X_{n+m} = \varepsilon_j \cap X_0 = \varepsilon_i) &= \sum_k \mathbb{P}(X_{n+m} = \varepsilon_j \cap X_0 = \varepsilon_i \cap X_m = \varepsilon_k) \\ &= \sum_k \mathbb{P}(X_{n+m} = \varepsilon_j \mid X_0 = \varepsilon_i, X_m = \varepsilon_k)\, \mathbb{P}(X_m = \varepsilon_k \cap X_0 = \varepsilon_i)\,.\end{aligned}$$

Using the Markov property, we deduce that

$$\mathbb{P}(X_{n+m} = \varepsilon_j \cap X_0 = \varepsilon_i) = \sum_k \mathbb{P}(X_{n+m} = \varepsilon_j \mid X_m = \varepsilon_k)\, \mathbb{P}(X_m = \varepsilon_k \cap X_0 = \varepsilon_i)\,.$$

Dividing by $\mathbb{P}(X_0 = \varepsilon_i)$ and using time homogeneity we get $p_{i,j}(n+m) = \sum_k p_{i,j}(n) p_{i,j}(m)$. □

Corollary 3.2. *Denote $P = P(1)$. Then $P(n) = P^n$.*

Proof. The quantity $\sum_k p_{i,k}(n) p_{k,j}(m)$ is by definition the (i,j) entry of the matrix product $P(n)P(m)$. So the Kolmogorov-Chapman equations imply that $P(n+m) = P(n)P(m)$. It follows that $P(n) = P(1)^n$. □

3.3 Examples

Throughout the chapter we shall illustrate the concepts using three examples. [2]

2. From now on we largely follow [13, Chapter 7], adding details and results when necessary.

3.3.1 The book pile problem

Consider a pile of m books lying on a table on top of one another. The books are numbered (referring for example the different volumes of a manga series), the pile is initially in some random order. The book on top thus has the number i_1, the next one down has number i_2, and the book at the bottom has number i_m. The state of this "system" is thus described by an ordered tuple $(i_1, \dots, i_m)$, which are just the set of permutations of the numbers $\{1, \dots, m\}$.

We consider the following process. A boy picks the volume k of the series with a probability p_k, reads it, then puts it on top of the pile. Then repeat. Show that this gives a Markov chain and find the transition probabilities.

Solution. The state space $S = \{(i_1, \dots, i_m)\}$ is the set of permutations of $\{1, \dots, m\}$, so $|S| = m!$ At time $n = 0$ the system is in state $(i_1, \dots, i_m)$. At time $n = 1$, the system may be in the same state, this occurs if the boy chose the volume i_1. This happens with probability p_{i_1}. Otherwise the boy chose a different volume, say i_k. Then the state at time $n = 1$ becomes $(i_k, i_1, \dots, i_{k-1}, i_{k+1}, \dots, i_m)$. This occurs with probability p_{i_k}. We thus see the transitions probabilities are

$$p_{(i_1,\dots,i_m),(j_1,\dots,j_m)} = \begin{cases} p_{i_1} & \text{if } (j_1, \dots, j_m) = (i_1, \dots, i_m), \\ p_{i_k} & \text{if } (j_1, \dots, j_m) = (i_k, i_1, \dots, i_{k-1}, i_{k+1}, \dots, i_m), \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Time homogeneity is clear : the above transition probabilities do not change with time and only depend on the states $(i_1, \dots, i_m)$ and $(j_1, \dots, j_m)$.

The Markov property is also clear : the probability that the system jumps from $(i_1, \dots, i_m)$ to $(j_1, \dots, j_m)$ does not depend on how it reached $(i_1, \dots, i_m)$.

3.3.2 The optimal choice problem

We studied this example in Stat 201. Here we study it further as a stochastic process. Let us recall the problem.

A set of m suitors propose in succession to a fussy young lady. She wants the best possible partner but she doesn't know in advance who it is, she must examine the suitors one by one. We assume the following :

(1) The lady can accept the first suitor, or reject him in the hope of finding a better partner. Similarly for the next marriage proposals. Note that she doesn't know in advance the "quality" of the suitors that will come later.

(2) A rejected suitor will not propose again, so she loses him forever.

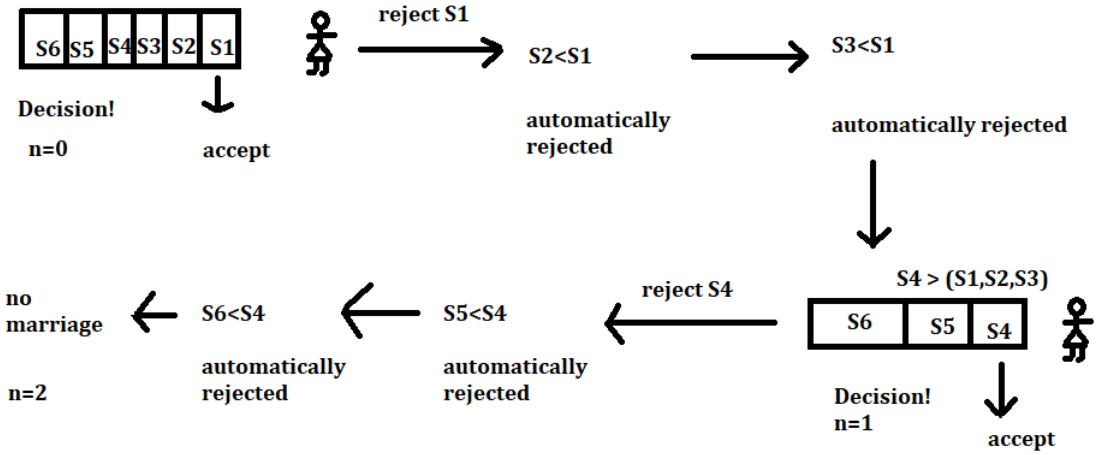


Figure 3.1 – Here $m = 6$, the S_k represent the suitors, S_4 being the best here. We have $X_0(\omega) = \varepsilon_1$, $X_1(\omega) = \varepsilon_4$, $X_2(\omega) = \varepsilon_7$.

(3) The lady never selects a suitor inferior to those previously rejected.

Note that rule (3) may cause the lady to never marry. This happens for example if the first suitor was in fact the best of them all, and she rejected him.

Show that this problem can be modeled as a Markov chain and find the transition probabilities.

Solution. Let S_k, $k = 1, \dots m$ be the suitors. They may appear to the lady in any order $(S_{i_1}, \dots, S_{i_m})$, so the *sample space* Ω is the set of permutations of this set and contains $m!$ elements. Be careful this is not the *state space*. The best of these suitors is a certain S_b whom the lady doesn't know in advance.

Since the lady rejects the suitors automatically if they are inferior to previously rejected ones, the only states that require consideration are

$$\varepsilon_k = \text{“the } k\text{-th inspected suitor is the best of the first } k \text{ suitors inspected”}$$

for $k = 1, \dots, m$. There is also an additional state :

$$\varepsilon_{m+1} = \text{“the best of all } m \text{ suitors has already been inspected and rejected.”}$$

The state space is $\{\varepsilon_1, \dots, \varepsilon_{m+1}\}$.

The times to decide only arise when moving from a state to a better one. A possible scenario is given in Figure 3.1. In formal terms, this represents the evolution of a fixed ω as time goes on, so we follow the sample path.

We generally have $m + 1$ random variables. If ω is such that the “quality” of suitors is arranged in increasing order then $X_k(\omega) = \varepsilon_k$. At the other extreme, we may have the suitors in decreasing order. In that case $X_0(\omega) = \varepsilon_1$, $X_1(\omega) = \varepsilon_{m+1}$

and we let $X_k(\omega) = \varepsilon_{m+1}$ for all $k \geq 1$. In general, if the best suitor S_b is being examined at time $n = \nu$, then $X_k(\omega) = \varepsilon_{m+1}$ for all $k > \nu$.

By construction, if $X_n(\omega) = \varepsilon_i$, then $X_{n+1}(\omega) = \varepsilon_j$ with

$$\begin{cases} j > i & \text{if } i \leq m, \\ j = i & \text{if } i = m+1. \end{cases}$$

This means that decisions are only made when moving from a state to a strictly better one (no two suitors are identical). On the other hand, if the system visits ε_{m+1}, it remains trapped in it forever.

We now give the transition probabilities. By the above, we have

$$p_{m+1,m+1} = 1 \qquad \text{and} \qquad p_{i,j} = 0 \quad \text{if } i \geq j \text{ and } j \leq m\,.$$

Next, let $i < j \leq m$. Then $p_{i,j} = \mathbb{P}(E_j \mid E_i) = \frac{\mathbb{P}(E_i \cap E_j)}{\mathbb{P}(E_i)}$, where $E_i = \{X_1 = \varepsilon_i\}$ and $E_j = \{X_2 = \varepsilon_j\}$. For E_i to occur, the i-th inspected suitor must be better than all those who preceded him. There are $i!$ ways to arrange the first i suitors, and there are $(i-1)!$ permutations of these such that their best is in the i-th position.[3] Hence, $\mathbb{P}(E_i) = \frac{(i-1)!}{i!} = \frac{1}{i}$.

Next, for E_iE_j to occur, the j-th suitor must be the best among those before him, and also the i-suitor must be the best of those before him. The i-th suitor must be the second-best among these first j suitors, since the j-th suitor was selected rightly afterwards. There are $(j-2)!$ ways to arrange the first j suitors such that the best is in the j-th position and the second-best is in the i-th position. Hence, $\mathbb{P}(E_iE_j) = \frac{(j-2)!}{j!} = \frac{1}{j(j-1)}$.

We conclude that $p_{i,j} = \frac{\mathbb{P}(E_i \cap E_j)}{\mathbb{P}(E_i)} = \frac{i}{(j-1)j}$.

Finally, if $i \leq m$, we find $p_{i,m+1} = \frac{\mathbb{P}(E_i \cap E_{m+1})}{\mathbb{P}(E_i)}$. For $E_i \cap E_{m+1}$ to occur, on the one hand, the i-th suitor must be the best among those before him, on the other hand, by continuing the process, we are told that the best of all suitors has been rejected. This means that the i-suitor must be the best of all suitors. There are $m!$ ways to arrange the m suitors, and $(m-1)!$ ways to put the best one in the i-th place. Hence, $\mathbb{P}(E_i \cap E_{m+1}) = \frac{(m-1)!}{m!} = \frac{1}{m}$. Thus, $p_{i,m+1} = \frac{i}{m}$.

3. It is also possible, but tedious, to find $\mathbb{P}(E_i)$ by counting all samples $\omega = (S_{k_1}, \ldots S_{k_m})$ such that $k_i > k_r$ for all $r < i$ (here one assumes without loss of generality that $S_1 < S_2 < \cdots < S_m$ in terms of "quality"). This number turns out to be $\frac{m!}{i}$, so $\mathbb{P}(E_i) = \frac{N(E_i)}{N} = \frac{m!/i}{m!} = \frac{1}{i}$. To see this, study the cases $k_i = i, i+1, \ldots, m$. You will need the relation $\sum_{k=0}^{n} \binom{k+r}{r} = \binom{r+n+1}{r+1}$ applied to $r = i-1$ and $n = m-i$.

Conclusion :

$$p_{i,j} = \begin{cases} 1 & \text{if } i = j = m+1, \\ 0 & \text{if } i \geq j \text{ and } j \leq m, \\ \frac{i}{(j-1)j} & \text{if } i < j \leq m, \\ \frac{i}{m} & \text{if } i \leq m \text{ and } j = m+1. \end{cases}$$

The Markov property is clear : only the last decision is necessary to examine the next suitor : if it is inferior to the last examined one, it is automatically rejected, if not, it can be accepted, and for this we do not require the information about past choices (because we only compare with the last examined one).

This chain is not time-homogeneous however. In fact, we must have $X_{m+1} = \varepsilon_{m+1}$, so $\mathbb{P}(X_{m+1} = \varepsilon_j \mid X_m = \varepsilon_i) = 0$ if $j < m+1$, while $p_{i,j} = \frac{i}{(j-1)j}$ if $i < j \leq m$.

3.3.3 One-dimensional random walks

Walking on the integers

Consider a particle moving randomly on $\mathbb{Z}$. At each step, it moves either one digit to the right, with probability p, or one to the left, with probability $q = 1 - p$. Let X_n be the position of the particle at time n. Show that X_n is a Markov chain and find the transition probabilities.

Solution. The state space is $\mathbb{Z}$. We have

$$p_{i,j} = \mathbb{P}(X_{n+1} = \varepsilon_j \mid X_n = \varepsilon_i) = \begin{cases} p & \text{if } j = i+1, \\ q & \text{if } j = i-1, \\ 0 & \text{otherwise.} \end{cases}$$

These transition probabilities do not change with time. The Markov property is clear as well; if the particle is currently at i, it will move to $i \pm 1$ with fixed probabilities $p_{i,i\pm1}$, regardless of how it actually reached the position i.

Climbing the Everest

A hiker is climbing the Everest, step by step. At each move, he may either take one step upwards, or fall down to the ground. Assume he moves upwards with probability p_i if at position i and falls down to zero with probability $q_i = 1 - p_i$. Show that the hiker's position is a Markov chain.

Solution. The mountain being very high, we may take $\mathbb{N}_0 = \{0, 1, \ldots\}$ to be

the state space. We have

$$p_{i,j} = \mathbb{P}(X_{n+1} = \varepsilon_j \mid X_n = \varepsilon_i) = \begin{cases} p_i & \text{if } j = i+1, \\ q_i & \text{if } j = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Again, these $p_{i,j}$ do not change with time n and the Markov property is clear; if at i, the hiker moves to $i+1$ or 0 with fixed probabilities $p_{i,i+1}, p_{i,0}$ regardless of how he reached i (and there is no real choice in reaching i, he must keep moving upward anyway, so such info is redundant).

3.4 Classification of states

Consider a Markov chain with states $\varepsilon_1, \varepsilon_2, \ldots$. Let us denote

$$u_n^{(i)} = p_{i,i}(n)\,,$$

$v_n^{(i)}$ = probability that the system starting at ε_i returns to ε_i for the first time after precisely n steps

for $n \geq 1$. We often remove the index i from the notation when ε_i is fixed and there is no confusion.

Lemma 3.3. *We have $u_n = \sum_{k=0}^{n} u_k v_{n-k}$ for any $n \geq 1$, where $u_0 := 1$ and $v_0 := 0$.*

Proof. Let B_k be the event "the system returns to ε_i for the first time after k steps", B_{n+1} the event "the system does not return at all to ε_i during the first n steps", $A = \{X_n = \varepsilon_i\}$ and $O = \{X_0 = \varepsilon_i\}$. Then the events $B_1, \ldots, B_n, B_{n+1}$ form a full set of mutually exclusive events, so

$$u_n = \mathbb{P}(A \mid O) = \frac{\mathbb{P}(A \cap O)}{\mathbb{P}(O)} = \sum_{k=1}^{n+1} \frac{\mathbb{P}(A \cap O \cap B_k)}{\mathbb{P}(O)}\,.$$

But $A \cap B_{n+1} = \emptyset$, so the term $k = n+1$ vanishes. So we get

$$u_n = \sum_{k=1}^{n} \frac{\mathbb{P}(A \cap O \cap B_k)}{\mathbb{P}(O \cap B_k)} \cdot \frac{\mathbb{P}(O \cap B_k)}{\mathbb{P}(O)} = \sum_{k=1}^{n} \mathbb{P}(A \mid O \cap B_k)\, \mathbb{P}(B_k \mid O)\,.$$

Clearly $\mathbb{P}(B_k \mid O) = v_k$. Next,

$$\mathbb{P}(A \mid O \cap B_k) = \mathbb{P}(X_n = \varepsilon_i \mid X_k = \varepsilon_i, X_j \neq \varepsilon_i \text{ for } 0 < j < k, X_0 = \varepsilon_i) = u_{n-k}$$

by the Markov property and time homogeneity. This proves the result. □

Let us introduce the generating functions

$$U(z) = \sum_{k=0}^{\infty} u_k z^k\,, \qquad V(z) = \sum_{k=0}^{\infty} v_k z^k\,, \qquad |z| < 1\,.$$

Corollary 3.4. $U(z) - u_0 = U(z)V(z)$.

Proof. The Cauchy product $U(z)V(z)$ is the series $\sum_{n=0}^{\infty} c_k z^k$, $c_n = \sum_{i=0}^{n} u_i v_{n-i}$. By Lemma 3.3, we get $c_n = u_n$ for all $n \geq 1$. For $n = 0$, we have $c_0 = u_0 v_0 = 0$. Since the zero-order term of $U(z) - u_0$ is also 0, this gives the corollary. □

Definition 3.5. Let $v^{(i)} = \sum_{n=0}^{\infty} v_n^{(i)}$ be the probability that the system sooner or later returns to ε_i. We say that ε_i is *recurrent* if $v^{(i)} = 1$ and *transient* if $v^{(i)} < 1$.

Theorem 3.6. *The state ε_i is recurrent if and only if $\sum_{n=0}^{\infty} u_n = \sum_{n=0}^{\infty} p_{i,i}(n) = \infty$.*

Proof. We first recall Abel's lemma : if $a_k \geq 0$ for all k, then $\lim_{x\to 1^-} \sum_{k=0}^{\infty} a_k x^k = \sum_{k=0}^{\infty} a_k$. This is simply an application of the monotone convergence theorem.

Let $\sum_n v_n = c \leq 1$. By Abel's lemma, $\lim_{z\to 1^-} V(z) = c$. So $\lim_{z\to 1^-} \frac{1}{1-V(z)} = \frac{1}{1-c}$, with the convention $\frac{1}{1-c} = \infty$ if $c = 1$. By Corollary 3.4, this is equivalent to $\lim_{z\to 1^-} U(z) = \frac{1}{1-c}$ (recall $u_0 = 1$). Using Abel's lemma again, this is equivalent to $\sum_n u_n = \frac{1}{1-c}$. Thus, $\sum_n u_n = \infty \iff c = 1$. □

Theorem 3.7. *If ε_i is recurrent, then with probability one, the system returns infinitely often to ε_i as the number of steps $n \to \infty$.*

If ε_i is transient, then with probability one, the system returns to ε_i only finitely often.

So if ε_i is transient, then after some steps, the system never returns to ε_i.

Proof. Let A_k be the event "there are at least k returns to ε_i as $n \to \infty$". Then

$$\mathbb{P}(A_1) = v\,.$$

If A_1 occurs, the system returns to ε_i after a number ν_1 of steps. By the Markov property, its subsequent behavior is the same as if it just started its motion. It follows that

$$\mathbb{P}(A_2 \mid A_1) = v\,.$$

Clearly $A_2 \subseteq A_1$. Therefore,

$$\mathbb{P}(A_2) = \mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_1) = v^2\,.$$

Similarly,

$$\mathbb{P}(A_k \mid A_{k-1}) = v\,, \qquad \mathbb{P}(A_k) = v^k\,. \tag{3.1}$$

If ε_i is transient then $v < 1$ and thus

$$\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \sum_{k=1}^{\infty} v^k = \frac{v}{1-v} < \infty\,.$$

By the first Borel-Cantelli lemma, this implies that with probability one, only finitely many of the events A_k occurs. So with probability one, the system returns to ε_i finitely often.

On the other hand, if ε_i is recurrent, then $v = 1$. This implies by (3.1) that $\mathbb{P}(A_k) = 1$ for all k. Let N be the number of times that the system returns to ε_i as $n \to \infty$. Then $A_k = \{N \geq k\}$. So $\mathbb{P}(N = \infty) = \mathbb{P}(\cap_k \{N \geq k\}) = \lim_{n\to\infty} \mathbb{P}(\{N \geq k\}) = \lim_{n\to\infty} \mathbb{P}(A_k) = 1$. □

Definition 3.8. We say that ε_j is *accessible from* ε_i if the probability that the system goes from ε_i to ε_j in some number of steps is positive. In other words, $p_{i,j}(M) > 0$ for some M which may depend on i, j.

Theorem 3.9. *If a state ε_j is accessible from a recurrent state ε_i, then ε_i is in turn accessible from ε_j, and ε_j is also recurrent.*

Proof. Suppose on the contrary that ε_i is not accessible from ε_j. Then the system goes from ε_i to ε_j with positive probability $p_{i,j}(M) = \alpha > 0$ for some M, and the system cannot return to ε_i afterwards. But then

$$\mathbb{P}(\text{system ever returns to } \varepsilon_i) \leq \mathbb{P}(\text{system does not visit } \varepsilon_j) = 1 - \alpha$$

which contradicts that ε_i is recurrent. Thus, ε_i is accessible from ε_j.

Say $p_{j,i}(N) = \beta > 0$ for some N. By Corollary 3.2,

$$P(n + M + N) = P(M)P(n)P(N) = P(N)P(n)P(M)$$

so we get

$$p_{i,i}(n + M + N) = \sum_{k,l} p_{i,k}(M)p_{k,l}(n)p_{l,i}(N) \geq p_{i,j}(M)p_{j,j}(n)p_{j,i}(N) = \alpha\beta p_{j,j}(n)\,,$$
$$p_{j,j}(n + M + N) = \sum_{k,l} p_{j,k}(N)p_{k,l}(n)p_{l,j}(M) \geq p_{j,i}(N)p_{i,i}(n)p_{i,j}(M) = \alpha\beta p_{i,i}(n)\,.$$

These inequalities show that the series $\sum_n p_{i,i}(n)$ and $\sum_n p_{j,j}(n)$ either both converge or both diverge. But $\sum_n p_{i,i}(n) = \infty$ by Theorem 3.6, since ε_i is recurrent. Thus, $\sum_n p_{j,j}(n) = \infty$ and ε_j is recurrent, using Theorem 3.6. □

Corollary 3.10. *If a Markov chain has only a finite number of states, each accessible from every other state, then all states are recurrent.*

Proof. Suppose we have m states, $S = \{\varepsilon_i\}_{i=1}^m$. If all are transient, then for each ε_i we can find Ω_i with $\mathbb{P}(\Omega_i) = 1$ such that if $\omega \in \Omega_i$, then $X_n(\omega) = \varepsilon_i$ for only finitely many n, say $X_n(\omega) \neq \varepsilon_i$ for $n > n_i(\omega)$. Let $\Omega_\star = \cap_{i=1}^m \Omega_i$. Then $\mathbb{P}(\Omega_\star) = 1$ and if $\omega \in \Omega_\star$ then $X_n(\omega) \neq \varepsilon_k$, $k = 1, \ldots, m$ for all $n > \max_{1 \leq i \leq m}(n_i(\omega))$, a contradiction since $X_n(\omega) \in S$. So at least one state, say ε_i, is recurrent. By hypothesis, all other states are accessible from ε_i. It follows from Theorem 3.9 that all states are recurrent. □

Example 3.11. Consider the book pile problem § 3.3.1. If each book i is picked with probability $p_i > 0$, then every state is accessible from any state. In fact, starting from $(i_1, \ldots, i_m)$, to reach $(j_1, \ldots, j_m)$, first pick j_m, then j_{m-1} and so on until j_1. This shows we can reach any state using at most m steps. The corollary implies that all states are recurrent.

If however $p_i = 0$ for some i, then any state of the form $(i_1, \ldots, i_m)$ with $i_1 = i$ is transient. Indeed, since $p_i = 0$, then in the first step, we will pick a book $j \neq i$. And so on, the book i will continue to move downward in the pile and will never come to the top again.

Example 3.12. Consider the optimal choice problem. Clearly, at step $m + 1$ we are already in step ε_{m+1}, in which we remain forever. This means that all states except ε_{m+1} are transient, while ε_{m+1} is recurrent.

Example 3.13. Consider the random walk on $\mathbb{Z}$ from § 3.3.3. Clearly, any state j is accessible from any state i using $|j - i|$ steps. Here however the state space is infinite so we cannot use Corollary 3.10 to study recurrence.

Instead, let us first show that

$$p_{i,j}(n) = \begin{cases} \binom{n}{\frac{n+j-i}{2}} p^{\frac{n+j-i}{2}} q^{\frac{n-j+i}{2}} & \text{if } |j-i| \leq n \text{ and } n+j-i \in 2\mathbb{N}_0 \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

where $2\mathbb{N}_0 = \{0, 2, 4, \ldots\}$. This simply means that the probability of going from i to j in n steps is the probability of having $\frac{n+j-i}{2}$ successes in a sequence of n Bernoulli trials, where success means taking a step to the right, thus following the Binomial distribution. In particular, $p_{i,i}(n)$ is the probability of taking $\frac{n}{2}$ steps to the right, which is intuitive. Let us prove (3.2) by induction on n.

For $n = 1$, we have $p_{i,i+1} = p = \binom{1}{1} p^1 q^0$, $p_{i,i-1} = q = \binom{1}{0} p^0 q^1$ and $p_{i,j} = 0$ for $j \neq i \pm 1$. This verifies the claim for $n = 1$.

Suppose the claim is true for n. Then by Kolmogorov-Chapman,

$$p_{i,j}(n+1) = \sum_k p_{i,k} p_{k,j}(n) = p \cdot p_{i+1,j}(n) + q \cdot p_{i-1,j}(n)\,. \tag{3.3}$$

This is zero if $|j-i| > n+1$ or $n+1+j-i \notin 2\mathbb{N}_0$. Indeed, the latter clearly implies the vanishing of the second term, and also implies that $n+j-i-1 = (n+j-i+1)-2 \notin 2\mathbb{N}_0$. So assume $|j-i| \le n+1$ and $n+1+j-i \in 2\mathbb{N}_0$.

If $j = i+n+1$, we get $p_{i,j}(n+1) = p \cdot \binom{n}{n} p^n q^0 + 0 = \binom{n+1}{n+1} p^{n+1} q^0$, which agrees with (3.2).

If $j = i-n-1$, we get $p_{i,j}(n+1) = 0 + q \cdot \binom{n}{0} p^0 q^n = \binom{n+1}{0} p^0 q^{n+1}$, which agrees with (3.2).

The cases $j = i \pm n$ are refused as they contradict $n+1+j-i \in 2\mathbb{N}_0$. So it remains to consider the cases $|j-i| \le n-1$. Here (3.3) yields

$$\begin{aligned} p_{i,j}(n+1) &= p \cdot \binom{n}{\frac{n+j-i-1}{2}} p^{\frac{n+j-i-1}{2}} q^{\frac{n-j+i+1}{2}} + q \cdot \binom{n}{\frac{n+j-i+1}{2}} p^{\frac{n+j-i+1}{2}} q^{\frac{n-j+i-1}{2}} \\ &= \left(\binom{n}{\frac{n+j-i-1}{2}} + \binom{n}{\frac{n+j-i+1}{2}} \right) p^{\frac{n+j-i+1}{2}} q^{\frac{n-j+i+1}{2}} = \binom{n+1}{\frac{n+1+j-i}{2}} p^{\frac{n+j-i+1}{2}} q^{\frac{n-j+i+1}{2}} \end{aligned}$$

which is (3.2) at $n+1$. Here we used that $\binom{n}{k-1} + \binom{n}{k} = \frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{k!(n-k)!} = \frac{k(n!)}{k!(n-k+1)!} + \frac{(n-k+1)n!}{k!(n-k+1)!} = \frac{(n+1)!}{k!(n-k+1)!} = \binom{n+1}{k}$. This proves (3.2) for all n.

We now address recurrence. By (3.2) and Stirling's formula, we have

$$p_{i,i}(2m) = \frac{(2m)!}{(m!)^2} p^m q^m \sim \frac{\sqrt{4\pi m}(2m)^{2m} \mathrm{e}^{-2m}}{(\sqrt{2\pi m}\, m^m \mathrm{e}^{-m})^2} p^m q^m = \frac{1}{\sqrt{\pi m}} (4pq)^m$$

for large m. Since $p_{i,i}(2m+1) = 0$, we deduce that

$$\sum_n p_{i,i}(n) < \infty \iff \sum_m p_{i,i}(2m) < \infty \iff \sum_m \frac{(4pq)^m}{\sqrt{\pi m}} < \infty\,.$$

But

$$4pq = (p+q)^2 - (p-q)^2 = 1 - (p-q)^2 \le 1\,,$$

with equality iff $p = q = \frac{1}{2}$. Thus, $\sum_n p_{i,i}(n) < \infty \iff p \neq q$.

We have shown that if $p \neq q$, then any state i is transient. This is intuitively clear, if $p > q$ for example, the walker has a bias to the right and will sooner or later abandon any state i. In contrast, if $p = q$, all states are recurrent, so any state is revisited infinitely often, because the walk is symmetric. [4]

Example 3.14. Consider the hiker problem of § 3.3.3. If $0 < p_i < 1$, then any state is accessible from any state, so by Theorem 3.9, the states are either all recurrent or all transient.

Suppose the system is initially at $i = 0$. Then the probability that it does not return to $i = 0$ after n steps is given by $p_0 p_1 \cdots p_{n-1}$, which is the probability of

4. Our argument implies that $v^{(i)} = 1$ if $p = q$ and $v^{(i)} < 1$ if $p \neq q$. A more precise calculation shows that $v^{(i)} = 1 - |p-q|$ in general. See [5, Proposition 5.5].

climbing $0 \to 1 \to \cdots \to n$. So the probability that the hiker never returns to $i = 0$, an event we can write as the intersection of all events "the hiker does not return to 0 after n steps", is given by $\lim_{n\to\infty} p_0 \cdots p_{n-1} = \prod_{i=0}^{\infty} p_i$.

If $\lim_{n\to\infty} p_0 \cdots p_{n-1} = 0$, then the hiker will surely return to $i = 0$ at some point, i.e. $v^{(0)} = 1$, so the state $i = 0$ is recurrent, and so are all other states. If the limit is nonzero, then $v^{(0)} = 1 - \lim_{n\to\infty} p_0 \cdots p_{n-1} < 1$, so $i = 0$ is transient, and so are all states.

Definition 3.15. Let (X_n) be a Markov chain and let T_i be the first time that X_n returns to ε_i, that is,

$$T_i = \inf\{n \geq 1 : X_n = \varepsilon_i\}.$$

Let $\mathbb{P}_i(A) = \mathbb{P}(A \mid X_0 = \varepsilon_i)$. Clearly, $\mathbb{P}_i(T_i = n) = v_n^{(i)}$, the probability that the system returns for the first time to ε_i after n steps, and $\mathbb{P}_i(T_i < \infty) = v^{(i)}$.

The *mean recurrence time at state* ε_i is the expected time m_i of first return :

$$m_i = \mathbb{E}_i(T_i) = \sum_{n=1}^{\infty} n\,\mathbb{P}_i(T_i = n) = \sum_{n=1}^{\infty} n v_n^{(i)}.$$

We say ε_i is *positive-recurrent* if $m_i < \infty$. We say that ε_i is *null-recurrent* if it is recurrent and $m_i = \infty$.

Lemma 3.16. *A positive-recurrent state is recurrent.*

If ε_i is transient then $m_i = \infty$.

Proof. If $\mathbb{E}_i(T_i) < \infty$ then $T_i < \infty$ a.s. so $\mathbb{P}_i(T_i < \infty) = 1$, hence $v^{(i)} = 1$.

By contraposition, if ε_i is transient then $m_i = \infty$. □

Example 3.17. Consider a Markov chain with transition matrix

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

with $p, q \in (0, 1)$. Show that both states are positive-recurrent.

Solution. To go from ε_1 to ε_1 for the first time in n steps, we must move from ε_1 to ε_2 in the first step, stay at ε_2 for $n-2$ steps, then move back to ε_1. Thus,

$$v_n^{(1)} = p_{1,2} p_{2,2}^{n-2} p_{2,1} = pq(1-q)^{n-2}.$$

Since $|1-q| < 1$, we get $\sum_n n v_n^{(1)} = \frac{pq}{1-q} \sum_n n(1-q)^{n-1} = \frac{pq}{1-q} \cdot (\sum_n x^n)'|_{x=1-q} = \frac{pq}{1-q} \cdot \frac{1}{q^2} < \infty$. Similarly, $\sum_n n v_n^{(2)} < \infty$.

Example 3.18. Consider the random walk on $\mathbb{Z}$ with $p = q = \frac{1}{2}$. Show that all states are null-recurrent.

Solution. We already showed the states are recurrent. For null-recurrence we need to calculate a bit more carefully. Consider the generating functions $U(x) = \sum_n u_n x^n$ and $V(x) = \sum_n v_n x^n$. We know $u_{2n} = p_{i,i}(2n) = \binom{2n}{n} p^n q^n$ and $u_{2n+1} = p_{i,i}(2n+1) = 0$. So[5] $U(x) = \sum_n \binom{2n}{n} \frac{x^{2n}}{4^n} = (1-x^2)^{-1/2}$. By Corollary 3.4, we deduce $V(x) = 1 - \frac{1}{U(x)} = 1 - \sqrt{1-x^2}$. So $V'(x) = \frac{x}{\sqrt{1-x^2}}$. But $V'(x) = (\sum_n v_n x^n)' = \sum_n n v_n x^{n-1}$. So By Abel's lemma, $\sum_n n v_n = \lim_{x \to 1^-} V'(x) = \lim_{x \to 1^-} \frac{x}{\sqrt{1-x^2}} = \infty$. Thus, all states are null-recurrent.

We now introduce a different classification of states.

Recall that if $m, n \in \mathbb{N}$, we say m is a divisor of n, denoted $m|n$, if $\frac{n}{m} \in \mathbb{N}$. If $A \subset \mathbb{N}$, we define $\gcd(A)$ to be the greatest common divisor of all $n \in A$. For example, $\gcd(8, 12) = 4$.

Definition 3.19. Let (X_n) be a Markov chain. Given a state ε_i, let

$$R_i = \{n \in \mathbb{N} : p_{i,i}(n) > 0\}.$$

We define the *period* of ε_i by $d_i = \gcd(R_i)$.

We say that ε_i is *periodic* if $d_i \geq 2$ and *aperiodic* if $d_i = 1$.

Example 3.20. Consider Figure 3.2 :

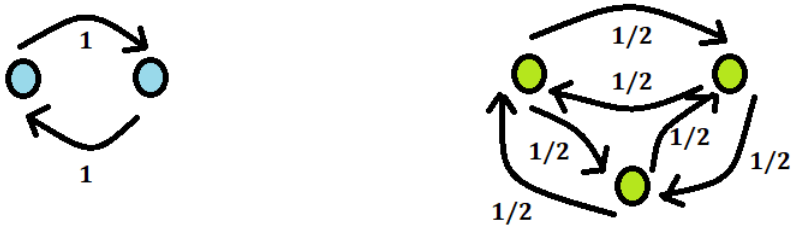


Figure 3.2 – States of the left Markov chain are periodic with period $d_i = 2$. States of the right Markov chain are aperiodic. From [10].

The left Markov chain has transition matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, the right one has transition matrix $\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$.

5. To calculate the sum, let $f(y) = (1-4y)^{-1/2}$. This is analytic so $f(y) = \sum_n \frac{1}{n!} f^{(n)}(0) y^n$. By induction $f^{(n)}(y) = \frac{(2n)!}{n!}(1-4y)^{-1/2-n}$. So $f(y) = \sum_n \binom{2n}{n} y^n$. Put $y = x^2/4$ to deduce the result.

For the left Markov chain, $p_{i,i}(n) > 0$ iff $n = 2, 4, 6, \ldots$. Clearly their gcd is $d_i = 2$. Both states are periodic.

For the right chain, we have $p_{i,i}(n) > 0$ iff $n = 2, 4, 6, \ldots$ or $n = 3, 6, 9, \ldots$. The gcd of all these n is $d_i = 1$. All states are aperiodic.

Definition 3.21. We say that states ε_i and ε_j *communicate* if ε_j is accessible from ε_i and ε_i is accessible from ε_j. We denote this by $\varepsilon_i \leftrightarrow \varepsilon_j$.

Lemma 3.22. *Communication is an equivalence relation.*

Proof. Clearly $p_{i,i}(0) = 1$ so $\varepsilon_i \leftrightarrow \varepsilon_i$. Obviously if $\varepsilon_i \leftrightarrow \varepsilon_j$ then $\varepsilon_j \leftrightarrow \varepsilon_i$. Finally if $\varepsilon_i \leftrightarrow \varepsilon_j$ and $\varepsilon_j \leftrightarrow \varepsilon_k$ then $p_{i,j}(M) > 0$, $p_{j,i}(M') > 0$, $p_{j,k}(N) > 0$ and $p_{k,j}(N') > 0$ for some M, M', N, N'. Then by Kolmogorov-Chapman, $p_{i,k}(M + N) = \sum_r p_{i,r}(M)p_{r,k}(N) \geq p_{i,j}(M)p_{j,k}(N) > 0$. Similarly, $p_{k,i}(M' + N') > 0$. □

Definition 3.23. Thus, communication partitions the state space into equivalence classes which we call *communicating classes*. We say we have a *class property* if once an element in the class satisfies the property, all elements in its class also satisfy the property.

Lemma 3.24. *Transience, recurrence, null-recurrence, positive-recurrence, periodicity are all class properties. Moreover, periodic states of the same class have the same period.*

Proof. Let $\varepsilon_i \leftrightarrow \varepsilon_j$. We showed in the proof of Theorem 3.9 that $\sum_n p_{i,i}(n)$ and $\sum_n p_{j,j}(n)$ either both converge or both diverge. So transience and recurrence are class properties.

Next we show $d_i \leq d_j$. It suffices to show that d_i is a divisor of R_j. So let $p_{j,j}(r) > 0$. Again in Theorem 3.9 we showed $p_{i,i}(n + M + N) \geq \alpha\beta p_{j,j}(n)$. So $p_{i,i}(r + M + N) > 0$, so $d_i | (r + M + N)$. Also, taking $n = 0$, we have that $p_{i,i}(M + N) \geq \alpha\beta > 0$, so $d_i | (M + N)$. Thus, $r + M + N = k_1 d_i$ and $M + N = k_2 d_i$, so $r = (k_1 - k_2)d_i$, that is, $d_i | r$. This shows that d_i is a divisor of R_j and $d_i \leq d_j$. Similarly $d_j \leq d_i$. This settles periodicity.

It remains to prove that null/positive recurrence are class properties. We omit this, see e.g. [12, Proposition 8.4.7]. □

Definition 3.25. Let (X_n) be a Markov chain with countable state space S. We say that $C \subseteq S$ is *closed* if once the chain enters C, it never leaves it. Thus, $\mathbb{P}(X_k \notin C \text{ for some } k \geq n \mid X_n \in C) = 0$.

We say that $C \subseteq S$ is *irreducible* if any two elements in C communicate.

We say the Markov chain is irreducible if any two states in S communicate.

Theorem 3.26. *Let (X_n) be a Markov chain with countable state space S. Then S can be partitioned as*

$$S = T \cup (\cup_j C_j)\,, \tag{3.4}$$

where T is the set of transient states and each C_j is a closed irreducible set of recurrent states.

Proof. Consider the partition of S into communicating classes. We know from Lemma 3.24 that transient and recurrent states cannot communicate, so the partition takes the form $S = (\cup_i T_i) \cup (\cup_j C_j)$ with T_i transient communicating classes and C_j recurrent communicating classes. Take $T = \cup_i T_i$. By definition each C_j is irreducible, so it remains to show each C_j is closed. Suppose $X_n \in C_j$, say $X_n = \varepsilon_k$. If $p_{k,l}(M) > 0$ for some $\varepsilon_l \in S$ and $M \geq 0$, then by Theorem 3.9, $\varepsilon_l \leftrightarrow \varepsilon_k$ and ε_l is recurrent. This means that $\varepsilon_l \in C_j$. Thus, if $\varepsilon_l \notin C_j$, we know $p_{k,l}(M) = 0$ for all M. This means that C_j is closed. □

Before moving on to limiting distributions, let us observe the following.

Lemma 3.27. *We also have $m_i = \sum_{n=1}^{\infty} \mathbb{P}_i(T_i \geq n)$.*

Proof. We have

$$\begin{aligned}\sum_{n\geq 1} \mathbb{P}_i(T_i \geq n) &= \sum_{n\geq 1}\sum_{k\geq n} \mathbb{P}_i(T_i = k) \\ &= \sum_{k\geq 1} \mathbb{P}_i(T_i = k) + \sum_{k\geq 2} \mathbb{P}_i(T_i = k) + \cdots = \sum_{m\geq 1} m\,\mathbb{P}_i(T_i = m) = \mathbb{E}_i(T_i)\,. \quad \square\end{aligned}$$

Proposition 3.2. *Let $C = C_j$ be one of the sets appearing in the partition* (3.4)*. If C is a finite set, then all its states are positive recurrent.*

Proof. Fix $\varepsilon_j \in C$. Denote $h_{i,j}^{(m)} = \mathbb{P}(X_k = \varepsilon_j \text{ for some } 1 \leq k \leq m \mid X_0 = \varepsilon_i)$ for $\varepsilon_i \in C$. Then $\lim_{m\to\infty} h_{i,j}^{(m)}$ is the probability that the system goes from ε_i to ε_j at some point, which is strictly positive since $\varepsilon_i \leftrightarrow \varepsilon_j$. Call the limit $v^{(i,j)} > 0$. Then there is some $m_{i,j}$ such that $h_{i,j}^{(m)} \geq \frac{v^{(i,j)}}{2}$ for $m \geq m_{i,j}$. Since C is finite, we may take $m_\star = \max_{i,j\in C} m_{i,j}$ and $\delta = \min_{i,j\in C} \frac{v^{(i,j)}}{2}$ to get $h_{i,j}^{(m_\star)} \geq \delta$ for all i.

Let $A_r = \{X_k \neq \varepsilon_j \ \forall 1 \leq k \leq r m_\star\}$ and $\mathbb{P}_i(A) = \mathbb{P}(A \mid X_0 = \varepsilon_i)$. We showed $\mathbb{P}_i(A_1) \leq 1 - \delta$ for any i. It follows that $\mathbb{P}(A_1 \mid X_0 \in K) \leq 1 - \delta$ for any $K \subseteq C$. Indeed, if $|K| = p$, then $\mathbb{P}(A_1 \mid X_0 \in K) = \frac{\mathbb{P}(A_1 \cap X_0 \in K)}{\mathbb{P}(X_0 \in K)} = \frac{\sum_{q=1}^{p} \mathbb{P}(A_1 \cap X_0 = \varepsilon_q)}{\sum_{q=1}^{p} \mathbb{P}(X_0 = \varepsilon_q)} = \frac{\sum_{q=1}^{p} \mathbb{P}(A_1 | X_0 = \varepsilon_q)\,\mathbb{P}(X_0 = \varepsilon_q)}{\sum_{q=1}^{p} \mathbb{P}(X_0 = \varepsilon_q)} \leq (1-\delta)\frac{\sum_{q=1}^{p} \mathbb{P}(X_0 = \varepsilon_q)}{\sum_{q=1}^{p} \mathbb{P}(X_0 = \varepsilon_q)} = (1-\delta)$.

Next,[6] $\mathbb{P}_i(A_2 \mid A_1) = \mathbb{P}(A_2 \mid X_{m_\star} \in C \setminus \{\varepsilon_j\}) = \mathbb{P}(A_1 \mid X_0 \in C \setminus \{\varepsilon_j\}) \leq (1-\delta)$ by the Markov property and time homogeneity (we also used that C is closed). Since $A_2 \subseteq A_1$, we get $\mathbb{P}_i(A_2) = \mathbb{P}_i(A_2 \mid A_1)\,\mathbb{P}_i(A_1) \leq (1-\delta)^2$. In general we see that $\mathbb{P}_i(A_r \mid A_{r-1}) \leq 1-\delta$ and $\mathbb{P}_i(A_r) \leq (1-\delta)^r$. If B_n is the event $\{X_k \neq \varepsilon_j \ \forall 1 \leq k \leq n\}$ then $B_n \subseteq A_{\lfloor n/m_\star \rfloor}$ because $n \geq \lfloor n/m_\star \rfloor m_\star$. Hence, $\mathbb{P}_i(B_n) \leq (1-\delta)^{\lfloor n/m_\star \rfloor}$ for any $\varepsilon_i \in C$. We conclude that

$$m_j = \sum_{n=0}^{\infty} \mathbb{P}_j(T_j \geq n+1) = \sum_{n=0}^{\infty} \mathbb{P}_j(B_n) \leq \sum_{n=0}^{\infty} (1-\delta)^{\lfloor n/m_\star \rfloor} = \frac{m_\star}{\delta} < \infty\,.$$

Since $\varepsilon_j \in C$ is arbitrary, all states in C are positive recurrent. □

Corollary 3.28. *On a finite state space, there are no null-recurrent states.*

Proof. Here, if $S = T \cup (\cup_j C_j)$, all C_j are finite sets, so all states in $(\cup_j C_j)$ are positive recurrent. □

Corollary 3.10 can also be strengthened as follows.

Corollary 3.29. *For an irreducible Markov chain on a finite state space, all states are positive recurrent.*

Proof. We know from Corollary 3.10 that all states are recurrent, so they must be positive recurrent by Corollary 3.28. □

3.5 Limiting distributions

Our aim is in this section is to study the asymptotic behavior of the chain as $n \to \infty$. We would like for example to understand the limiting behavior of $\lim_{n\to\infty} p_{i,j}(n)$. This refers to the probability that "in the end" we are in position ε_j, given that we started at ε_i. If the chain is well connected, one may suspect that such limiting behavior does not depend on where we started from, i.e. that the limit does not depend on ε_i. One can also guess that in the partition given in Theorem 3.26, the dynamics will only be interesting in the recurrent classes C_j. This is what we prove in the following lemma.

Lemma 3.30. *If ε_j is transient, then $\lim_{n\to\infty} p_{i,j}(n) = 0$ for any i.*

So "in the end" the chain will not be in ε_j no matter where we started.

6. Note that $\mathbb{P}_i(A \mid B) = \frac{\mathbb{P}_i(A\cap B)}{\mathbb{P}_i(B)} = \frac{\frac{\mathbb{P}(A\cap B\cap X_0=\varepsilon_i)}{\mathbb{P}(X_0=\varepsilon_i)}}{\frac{\mathbb{P}(B\cap X_0=\varepsilon_i)}{\mathbb{P}(X_0=\varepsilon_i)}} = \frac{\mathbb{P}(A\cap B\cap X_0=\varepsilon_i)}{\mathbb{P}(B\cap X_0=\varepsilon_i)} = \mathbb{P}(A \mid B\cap X_0=\varepsilon_i)$.

Proof. Let $v_n^{(i,j)}$ be the probability that we reach ε_j for the first time after n steps, if we start from ε_i. The same proof of Lemma 3.3 shows that

$$p_{i,j}(n) = \sum_{k=1}^{n} p_{j,j}(n-k) v_k^{(i,j)} .$$

Indeed, just replace ε_i by ε_j in the definition of A and B_k. We can write this sum as $p_{j,j}(n-1)v_1^{(i,j)} + \cdots + p_{j,j}(0)v_n^{(i,j)} = \sum_{k=0}^{n-1} p_{j,j}(k) v_{n-k}^{(i,j)}$. Letting $v_m^{(i,j)} = 0$ for $m \leq 0$, this is $\sum_{k=0}^{\infty} p_{j,j}(k) v_{n-k}^{(i,j)}$. It follows that

$$\sum_{n=0}^{\infty} p_{i,j}(n) = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} p_{j,j}(k) v_{n-k}^{(i,j)} = \sum_{k=0}^{\infty} p_{j,j}(k) \sum_{m=1}^{\infty} v_m^{(i,j)} \leq \sum_{k=0}^{\infty} p_{j,j}(k) .$$

As ε_j is transient, $\sum_k p_{j,j}(k) < \infty$, so $\sum_n p_{i,j}(n) < \infty$ and $\lim_{n\to\infty} p_{i,j}(n) = 0$. □

Let us introduce

$$p_j(n) = \mathbb{P}(X_n = \varepsilon_j) .$$

Lemma 3.31. *We have* $p_j(n) = \sum_k p_k^0 p_{k,j}(n) = \sum_k p_k(n-1) p_{k,j}$.

Proof. $\mathbb{P}(X_n = \varepsilon_j) = \sum_k \mathbb{P}(X_n = \varepsilon_j \mid X_0 = \varepsilon_k)\, \mathbb{P}(X_0 = \varepsilon_k)$.

Also, $\mathbb{P}(X_n = \varepsilon_j) = \sum_k \mathbb{P}(X_n = \varepsilon_j \mid X_{n-1} = \varepsilon_k)\, \mathbb{P}(X_{n-1} = \varepsilon_k)$. □

Note that $\sum_j p_j(n) = \mathbb{P}(X_n \in S) = 1$. If $p_j(n) \to \pi_j$, then we can hope that $\sum_j \pi_j = 1$ and $\pi_j \geq 0$. This leads us to think about probability measures on the state space S. Namely, suppose we endow each state ε_j a number $\mu_j \geq 0$ such that $\sum_j \mu_j = 1$, where the sum runs over all states of S. Then we define the probability measure $\mu = \sum_j \mu_j \delta_j$ by $\mu(A) = \sum_{\varepsilon_j \in A} \mu_j$ for $A \subseteq S$. Note $\mu(\{\varepsilon_j\}) = \mu_j$. Also, $\mu(S) = \sum_j \mu_j = 1$, so μ is indeed a probability measure on S.

Definition 3.32. We say that $\mu = \sum_j \mu_j \delta_j$ is an *invariant measure* for the Markov chain (X_n) if $\mu_j \geq 0$, $\sum_j \mu_j = 1$ and $\mu_j = \sum_i \mu_i p_{i,j}$.

The reason for the terminology "invariant" is that if P is the transition matrix and $\mu = (\mu_1, \mu_2, \dots)$, then we have $\mu P = \mu$, i.e. μ is invariant under right-multiplication by P. It also follows that $\mu \cdot P(n) = \mu P^n = \mu$.

μ is also called a "stationary distribution" for the following reason :

Lemma 3.33. *Suppose* $\mu = \sum_j \mu_j \delta_j$ *is an invariant measure. If we endow the state space with initial probabilities* $p_j^0 := \mu_j$, *then* $p_j(n) = p_j^0$ *for all* n.

In other words, $\mathbb{P}(X_0 = \varepsilon_j) = \mathbb{P}(X_n = \varepsilon_j)$ for all n, so the Markov chain becomes "stationary".

Proof. We have $p_j(n) = \sum_k p_k^0 p_{k,j}(n) = \sum_k \mu_k p_{k,j}(n) = (\mu \cdot P(n))_j = \mu_j = p_j^0$. □

Theorem 3.34. *Let (X_n) be a Markov chain with finitely many states $\varepsilon_1, \dots, \varepsilon_m$. Suppose*

$$\exists N \in \mathbb{N}\ \exists\, \delta > 0 :\ p_{i,j}(N) > \delta \quad \forall i, j\,. \tag{3.1}$$

Then we have convergence

$$\lim_{n\to\infty} p_{i,j}(n) = \pi_j\,, \qquad \lim_{n\to\infty} p_j(n) = \pi_j$$

for any i and any p_i^0, at exponential speed : we may find C, D such that

$$\max_i |p_{i,j}(n) - \pi_j| \le Ce^{-Dn}\,, \qquad |p_j(n) - \pi_j| \le Ce^{-Dn}\,.$$

Moreover, $\sum_j \pi_j \delta_j$ is an invariant measure on S, hence $\pi_j \ge 0$, $\sum_j \pi_j = 1$ and

$$\pi_j = \sum_{i=1}^m \pi_i p_{i,j}\,. \tag{3.2}$$

Note that condition (3.1) implies in particular that the Markov chain is irreducible. In fact irreducibility is the weaker condition that for any i, j there exists $N_{i,j}$ such that $p_{i,j}(N_{i,j}) > 0$. Here there is one N working for all states.

That convergence holds for any p_i^0 is worth mentioning because we saw in Lemma 3.33 that $p_i(n)$ not only converges, but is constant, for a special choice of p_i^0 if an invariant measure exists, which can happen under weaker conditions than those of Theorem 3.34.

Proof of Theorem 3.34. Let

$$r_j(n) = \min_i p_{i,j}(n)\,, \qquad R_j(n) = \max_i p_{i,j}(n)\,.$$

Then

$$r_j(n+1) = \min_i p_{i,j}(n+1) = \min_i \sum_{k=1}^m p_{i,k} p_{k,j}(n) \ge \min_i \sum_{k=1}^m p_{i,k} r_j(n) = r_j(n)\,,$$

$$R_j(n+1) = \max_i p_{i,j}(n+1) = \max_i \sum_{k=1}^m p_{i,k} p_{k,j}(n) \le \max_i \sum_{k=1}^m p_{i,k} R_j(n) = R_j(n)\,.$$

Thus, for any n,

$$r_j(1) \le r_j(2) \le \cdots \le r_j(n) \le R_j(n) \le \cdots \le R_j(2) \le R_j(1)\,. \tag{3.3}$$

Hence, $r_j(n)$ is monotone increasing, bounded above by $R_j(1)$, and $R_j(n)$ is monotone decreasing, bounded below by $r_j(1)$. Both sequences thus converge :

$$r_j(n) \to r_j \qquad \text{and} \qquad R_j(n) \to R_j \tag{3.4}$$

for some r_j, R_j. Next, for N as in (3.1), we have

$$\begin{aligned} p_{i,j}(n+N) &= \sum_k p_{i,k}(N) p_{k,j}(n) \\ &= \sum_k [p_{i,k}(N) - \delta p_{j,k}(n)] p_{k,j}(n) + \delta \sum_k p_{j,k}(n) p_{k,j}(n) \\ &= \sum_k [p_{i,k}(N) - \delta p_{j,k}(n)] p_{k,j}(n) + \delta p_{j,j}(2n)\,. \end{aligned} \tag{3.5}$$

Here $[p_{i,k}(N) - \delta p_{j,k}(n)] \geq \delta - \delta p_{j,k}(n) \geq 0$, since $p_{j,k}(n) \leq 1$. It follows from (3.5) that

$$p_{i,j}(n+N) \geq r_j(n) \sum_k [p_{i,k}(N) - \delta p_{j,k}(n)] + \delta p_{j,j}(2n) = r_j(n)(1-\delta) + \delta p_{j,j}(2n)\,.$$

Since i is arbitrary, this implies

$$r_j(n+N) \geq r_j(n)(1-\delta) + \delta p_{j,j}(2n)\,.$$

Similarly, (3.5) implies that

$$R_j(n+N) \leq (1-\delta) R_j(n) + \delta p_{j,j}(2n)\,.$$

Thus,

$$R_j(n+N) - r_j(n+N) \leq (1-\delta)(R_j(n) - r_j(n))\,. \tag{3.6}$$

Taking $d = 1 - \delta < 1$, it follows that

$$R_j(kN) - r_j(kN) \leq d^{k-1}(R_j(N) - r_j(N)) \leq d^{k-1}\,.$$

Taking $k \to \infty$ shows that $R_j(kN) - r_j(kN) \to 0$. But $R_j(kN) - r_j(kN) \to R_j - r_j$ by (3.4). Thus $R_j = r_j =: \pi_j$.

Since $R_j(n) \searrow$ and $r_j(n) \nearrow$, then $(R_j(n) - r_j(n)) \searrow$. If we write any $n = Np + m$, so that $p = \frac{n-m}{N} \geq \frac{n}{N} - 1$, we thus get

$$R_j(n) - r_j(n) \leq R_j(pN) - r_j(pN) \leq d^{p-1} \leq d^{\frac{n}{N}-2}\,.$$

But $p_{i,j}(n) - \pi_j \leq R_j(n) - \pi_j \leq R_j(n) - r_j(n)$ since $r_j(n) \nearrow \pi_j$. Similarly, $p_{i,j}(n) - \pi_j \geq r_j(n) - \pi_j \geq r_j(n) - R_j(n)$. Hence,

$$|p_{i,j}(n) - \pi_j| \leq R_j(n) - r_j(n) \leq d^{\frac{n}{N}-2}\,, \qquad i = 1, \ldots, m.$$

Next, given any initial distribution p_i^0, we have

$$\begin{aligned} |p_j(n) - \pi_j| &= \Big| \sum_{i=1}^m p_i^0 p_{i,j}(n) - \pi_j \Big| \\ &= \Big| \sum_{i=1}^m p_i^0 [p_{i,j}(n) - \pi_j] \Big| \leq d^{\frac{n}{N}-2} \sum_{i=1}^m p_i^0 = d^{\frac{n}{N}-2}\,. \end{aligned}$$

Taking $C = d^{-2}$, $D = \frac{-1}{N} \ln d$, we've proved convergence at exponential speed.

Finally, taking $n \to \infty$ in $\sum_{j=1}^{m} p_j(n) = 1$ gives $\sum_{j=1}^{m} \pi_j = 1$. Clearly $p_j(n) \geq 0$ also implies $\pi_j \geq 0$. Since $p_j(n) = \sum_{i=1}^{m} p_i(n-1)p_{i,j}$, taking $n \to \infty$ gives $\pi_j = \sum_{i=1}^{m} \pi_i p_{i,j}$ as required. □

Example 3.35. Consider the book pile problem. We showed in § 3.11 that if $p_i > 0$ for all i, then any state is accessible from any other state using at most m steps. This implies $p_{i,j}(m) > 0$ for any i, j. Since the number of states is finite, we may find $\delta > 0$ such that $p_{i,j}(m) \geq \delta$ for all i, j. We can thus apply Theorem 3.34 to deduce that $p_j(n) \to \pi_j$ for some π_j. To find π_j, we use the invariance equation.

Denote states by $(i_1, \ldots, i_m)$ and recall the transition probabilities (3.1). Then $\pi_j = \sum_i p_{i,j}\pi_i$ becomes

$$\pi_{(j_1,\ldots,j_m)} = p_{j_1} \sum_{(j'_1,\ldots,j'_m)} \pi_{(j'_1,\ldots,j'_m)}\,, \tag{3.7}$$

where $(j'_1, \ldots, j'_m)$ ranges over the m permutations of j_1

$$(j_1, j_2, \ldots, j_m)\,, (j_2, j_1, j_3, \ldots, j_m)\,, \ldots\,, (j_2, j_3, \ldots, j_m, j_1)$$

which give $(j_1, \ldots, j_m)$ when j_1 is moved on top.

Computing $\pi_{(i_1,\ldots,i_m)}$ for each state seems nontrivial.[7] It is easier to compute the limiting probability of just having the book i on top. This is a sum $\sum \pi_{(i,i_2,\ldots,i_m)}$ over $(m-1)!$ permutations $(i_2, \ldots, i_m)$. Since $\pi_{(i,i_2,\ldots,i_m)}$ satisfies (3.7), we may replace it by $p_i \sum_{(i'_1,\ldots,i'_m)} \pi_{(i'_1,\ldots,i'_m)}$, this sum running over m terms. The result is p_i multiplied by a sum over all $m \cdot (m-1)! = m!$ permutations $\pi_{(i'_1,\ldots,i'_m)}$, which just gives 1 since we have a probability measure. Thus, the limiting probability of finding book i on top is $p_i \cdot 1 = p_i$.

Example 3.36. Consider the random walk on $\mathbb{Z}$. We know $p_{i,i}(n) = 0$ for odd n. For even n, we showed in § 3.13 that $p_{i,i}(2m) \sim \frac{1}{\sqrt{\pi m}}(4pq)^m \to 0$ as $m \to \infty$, for any p, q. So $p_{i,i}(n)$ do not converge to a stationary measure, in fact $\sum_i \pi_i = 0$.

Example 3.37. Consider the hiker problem. If $\lim_{n\to\infty} p_0 p_1 \cdots p_n \neq 0$, then all states are transient, see § 3.14, so the limiting probability is just zero by Lemma 3.30.

So suppose $\lim_{n\to\infty} p_0 p_1 \cdots p_n = 0$. If an invariant measure μ exists, then $\mu_j = \sum_i \mu_i p_{i,j}$. This reduces to $\mu_j = \mu_{j-1} p_{j-1}$. Thus, $\mu_1 = \mu_0 p_0$, $\mu_2 = \mu_1 p_1 =$

7. A reasonable guess is that $\pi_{(i_1,\ldots,i_m)} = \frac{p_{i_1}\cdots p_{i_m}}{(1-p_{i_1})(1-p_{i_1}-p_{i_2})\cdots(1-p_{i_1}-\cdots-p_{i_{m-1}})}$. This seems to work for $m = 2, 3$, but seems tedious to check for general m.

$\mu_0 p_0 p_1$ and so on, $\mu_n = \mu_0 p_0 \cdots p_{n-1}$. The condition $\sum_j \mu_j = 1$ entails that $\sum_n \mu_0 p_0 \cdots p_{n-1} = 1$, thus $\mu_0 = \frac{1}{\sum_n p_0 \cdots p_{n-1}}$. If $\sum_n p_0 \cdots p_{n-1} = \infty$, then $\mu_0 = 0$ and all $\mu_n = 0$, contradicting the requirements.

We thus showed that if an invariant measure μ exists, then we must have $\sum_{n=0}^{\infty} p_0 \cdots p_{n-1} < \infty$, and μ must take the form

$$\mu_0 = \frac{1}{1 + p_0 + p_0 p_1 + \dots}, \qquad \mu_n = \frac{p_0 \cdots p_{n-1}}{1 + p_0 + p_0 p_1 + \dots} \quad n \geq 1 .$$

Conversely, if we define μ as above, then $\sum_n \mu_n = \mu_0 \sum_n p_0 \cdots p_{n-1} = 1$ and $\mu_n = \mu_{n-1} p_{n-1}$, so μ is an invariant measure.

We gave a criterion for the existence of an invariant measure, and gave its explicit form, however, we have not shown whether $\lim_{n\to\infty} p_{i,j}(n) = \mu_j$. This is why we denoted μ_j instead of π_j.

3.6 Further results

We have studied random walks on $\mathbb{Z}$ and shown in § 3.13 that the symmetric random walk is recurrent, while the non-symmetric walk is transient. One can similarly define the simple symmetric random walk on $\mathbb{Z}^d$, in which the particle jumps with equal probabilities to any of its $2d$ neighbors. Then Pólya proved the following fact in 1921 :

Theorem 3.38. *The simple symmetric random walk on $\mathbb{Z}^d$ is recurrent iff $d \leq 2$.*

See [16] for an extensive treatment of random walks.

Concerning limiting distributions, we have only given a good answer when the state space is finite. There are two parts to this question in general : first, does an invariant measure for the process exist ? Second, does it represent the limiting behavior of $p_{i,j}(n)$?

For the first question we have the powerful result :

Theorem 3.39. *Suppose (X_n) is irreducible and each state is positive recurrent. Then the Markov chain has a unique invariant measure given by $\mu_i = 1/m_i$ for each state ε_i, where m_i is the mean recurrence time.*

See [12, Theorem 8.4.1] for a proof. Recall that any irreducible Markov chain on a finite state space is positive recurrent by Corollary 3.29. So not only do we get the existence of an invariant measure without the need for (3.1), but we also get a quite explicit form for it.

But is (3.1) needed at all ? Essentially yes, if we want a positive answer to the second question, see Exercise 7. We have the following result in general :

Theorem 3.40. *If a Markov chain is irreducible and aperiodic and has an invariant measure* $\{\pi_j\}$, *then* $\lim_{n\to\infty} p_{i,j}(n) = \pi_j = \lim_{n\to\infty} p_j(n) =$ *for any* i *and* p_i^0.

See [12, Theorem 8.3.10, Corollary 8.3.11] for a proof. It is known that if the Markov chain is irreducible and aperiodic then for any i, j we can find $n_{i,j}$ such that $p_{i,j}(n) > 0$ for all $n \geq n_{i,j}$, see [12, Lemma 8.3.9]. If moreover the state space is finite, then by taking $N = \max_{i,j} n_{i,j}$, we get in particular that (3.1) is satisfied.

Summarizing, Theorem 3.34 holds more generally for any irreducible, aperiodic, positive recurrent Markov chain, whether S is finite or not. [8]

3.7 Exercises

1. A number from 1 to m is chosen at random, at each of the times $t = 1, 2, \ldots$. A system is said to be in the state ε_0 if no number has yet been chosen, and in the state ε_i if the largest number so far chosen is i. Show that the random process described by this model is a Markov chain. Find the corresponding transition probabilities $p_{i,j}$ $(i, j = 0, 1, \ldots, m)$.

2. In the preceding problem, which states are recurrent and which transient ?

3. Suppose $m = 4$ in Problem 1. Find the matrix $P(2) = (p_{i,j}(2))$, where $p_{i,j}(2)$ is the probability that the system will go from state ε_i to state ε_j in 2 steps.

4. An urn contains a total of N balls, some black and some white. Samples are drawn from the urn, m balls at a time ($m \leq N$). After drawing each sample, the black balls are returned to the urn, while the white balls are replaced by black balls and then returned to the urn. If the number of white balls in the urn is i, we say that the "system" is in the state ε_i. Prove that the random process described by this model is a Markov chain (imagine that samples are drawn at the times $n = 1, 2, \ldots$ and that the system has some initial probability distribution). Find the corresponding transition probabilities $p_{i,j}$ $(i, j = 0, 1, \ldots, N)$. Which states are recurrent and which are transient ?

5. In the preceding problem, let $N = 8$, $m = 4$, and suppose there are initially 5 white balls in the urn. What is the probability that no white balls are left after 2 drawings (of 4 balls each) ?

6. A particle moves randomly along the interval $[1, m]$, coming to rest only at the points with coordinates $x = 1, \ldots, m$. The particle's motion is described

8. Following Exercises 1-16 are copied from [13].

by a Markov chain such that

$$p_{1,2} = 1\,, \qquad p_{m,m-1} = 1\,,$$
$$p_{j,j+1} = p\,, \qquad p_{j,j-1} = q \qquad (j = 2, \dots, m-1),$$

with all other transition probabilities equal to zero. Which states are recurrent and which are transient ?

7. In the preceding problem, show that the limiting probabilities defined in Theorem 3.34 do not exist. In particular, show that the condition (3.1) does not hold for any N.

8. Consider the same kind of random walk as in Problem 6, but now suppose the nonzero transition probabilities are

$$p_{1,1} = q\,, \qquad p_{m,m} = p\,,$$
$$p_{j,j+1} = p\,, \qquad p_{j,j-1} = q \qquad (j = 1, \dots, m),$$

permitting the particle to stay at the points $x = 1$ and $x = m$. Which states are recurrent and which are transient ? Show that the limiting probabilities $\pi_1, \dots, \pi_m$ defined in Theorem 3.34 now exist.

9. In the preceding problem, calculate the limiting probabilities $\pi_1, \dots, \pi_m$.

10. Two marksmen A and B take turns shooting at a target. It is agreed that A will shoot after each hit, while B will shoot after each miss. Suppose A hits the target with probability $\alpha > 0$, while B hits the target with probability $\beta > 0$, and let n be the number of shots fired. What is the limiting probability of hitting the target as $n \to \infty$?

11. Suppose the condition (3.1) holds for a transition probability matrix whose column sums (as well as row sums) all equal unity. Find the limiting probabilities $\pi_1, \dots, \pi_m$.

12. Suppose m white balls and m black balls are mixed together and divided equally between two urns. A ball is then drawn at random from each urn and put into the other urn. Suppose this is done n times. If the number of white balls in a given urn is j, we say that the "system" is in the state ε_j (the number of white balls in the other urn is then $m - j$). Prove that the limiting probabilities $\pi_0, \dots, \pi_m$ defined in Theorem 3.34 exist, and calculate them.

13. Find the stationary distribution $\pi_1, \pi_2, \dots$ for the Markov chain whose only nonzero transition probabilities are

$$p_{j,1} = \frac{j}{j+1}, \qquad p_{j,j+1} = \frac{1}{j+1} \qquad (j = 1, 2, \dots)$$

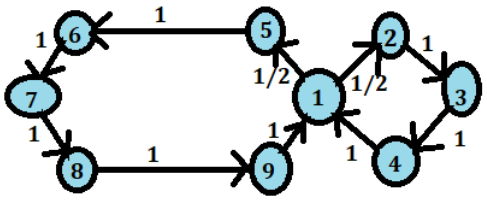
14. Two gamblers A and B repeatedly play a game such that A's probability of winning is p, while B's probability of winning is $q = 1 - p$. Each bet is a dollar, and the total capital of both players is m dollars. Find the probability of each player being ruined, given that A's initial capital is j dollars.

15. In the preceding problem, prove that if $p > q$, then A's probability of ruin increases if the stakes are doubled.

16. Prove that a gambler playing against an adversary with unlimited capital is certain to be ruined unless his probability of winning in each play of the game exceeds $\frac{1}{2}$.

17. Consider a Markov chain with transition matrix $P = \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. Draw a figure for this chain. Are the states periodic or aperiodic ?

18. Consider the following Markov chain :



Find the period d_1.

19. Consider the hiker problem. Give a criterion for the Markov chain to be positive recurrent. Is this Markov chain positive recurrent when $p_i = p \in (0, 1)$ for all i ?

20. Find the invariant measure for a Markov chain with $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Does $p_{ii}(n)$ converge to it as $n \to \infty$?

Chapter 4

Continuous Markov Processes

4.1 Definitions and basic properties

Consider a stochastic process $X_t(\omega)$, we assume each X_t is a discrete random variable taking values in some countable set $S = \{\varepsilon_1, \varepsilon_2, \ldots\}$ called the *state space*, which may be finite or infinite. We assume in this chapter that we have a continuous-time process, i.e. the parameter t varies in $[0, +\infty)$. The resulting process models a physical system that varies continuously in time, moving randomly from one state ε_i to another ε_j at random times. Consider the following assumptions :

(a) (Time homogeneity). For any states $\varepsilon_i, \varepsilon_j$ and times s, t, the transition probability

$$p_{ij}(t) = \mathbb{P}(X_{s+t} = \varepsilon_j \mid X_s = \varepsilon_i), \qquad i, j = 1, 2, \ldots \tag{4.1}$$

is independent of s. So it is equal to $\mathbb{P}(X_t = \varepsilon_j \mid X_0 = \varepsilon_i)$.

This is the probability that the system moves from ε_i to ε_j in time t.

(b) (Markov property). Let $\mathcal{F}_s = \sigma(X_r, r \leq s)$. Then

$$\mathbb{P}(X_t = \varepsilon \mid \mathcal{F}_s) = \mathbb{P}(X_t = \varepsilon \mid X_s)\,.$$

In other words, the additional information provided by the history of how the system reached X_s (which is encoded in X_r, $r < s$), is useless.

(c) The sample paths $t \mapsto X_t$ are right-continuous. That is, for any ω, $t \mapsto X_t(\omega)$ is right-continuous.

(d) The X_t make finitely many jumps in finite time intervals.

A stochastic process satisfying assumptions (a) and (b) is called a *continuous Markov process*. If the process also satisfies (c), it is called a *jump process*. Note

that in this case, for any t and ω, there is $\delta(t,\omega)$ such that $X_{t+s}(\omega) = X_t(\omega)$ for $s \in [t, t+\delta]$. Finally, if the process also satisfies (d), it is called a *regular jump process*. There exists continuous Markov processes which do not satisfy (c) and (d), but these are rather pathological. Roughly, what happens in this case is that when the system reaches some state, it immediately jumps to another state afterwards. So *we will always assume that (a),(b),(c),(d) hold*, which covers typical systems in real-life coming from biology, engineering and so on. A simple but important example is the *Poisson process* described later.

We define

$$p_j(t) = \mathbb{P}(X_t = \varepsilon_j), \qquad j = 1, 2, \dots \tag{4.2}$$

and denote

$$p_j^0 := p_j(0)\,. \tag{4.3}$$

Then as for Markov chains, we see that

$$p_j(s+t) = \sum_k p_k(s) p_{kj}(t), \qquad j = 1, 2, \dots \tag{4.4}$$

which says that to reach ε_j in time $s+t$, it is equivalent to reach an arbitrary ε_k in time s, then move from k to j in time t. Similarly, we have the *Kolmogorov-Chapman equation*

$$p_{ij}(s+t) = \sum_k p_{ik}(s) p_{kj}(t), \qquad i, j = 1, 2, \dots$$

If we define the transition matrix $P(t) = (p_{ij}(t))$, this says that

$$P(s+t) = P(s)P(t)\,.$$

Again $P(t)$ is a *stochastic matrix*, meaning that

$$\sum_j p_{ij}(t) = 1, \qquad \forall i = 1, 2, \dots \tag{4.5}$$

Also,

$$P(0) = \mathrm{Id}$$

since we clearly have $p_{ij}(0) = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases}$

We have the following continuity result :

Lemma 4.1. *For any i, j we have $\lim_{t \downarrow 0} p_{ij}(t) = p_{ij}(0)$. Consequently, $\lim_{t \downarrow 0} P(t) = \mathrm{Id}$. The $p_{ij}(t)$ are also continuous (both from right and left) at any $t > 0$.*

Proof. By right-continuity of sample paths (assumption (c)), if $X_0(\omega) = \varepsilon_i$, then there is δ_ω such that $X_t(\omega) = \varepsilon_i$ for all $t \leq \delta_\omega$ and thus $\mathbf{1}_{\{X_t=\varepsilon_j\}}(\omega) = \delta_{i,j} = \mathbf{1}_{\{X_0=\varepsilon_j\}}(\omega)$ for any $t \in [0, \delta_\omega]$. This means that for any ω, $\mathbf{1}_{\{X_t=\varepsilon_j\}}(\omega) \to \mathbf{1}_{\{X_0=\varepsilon_j\}}(\omega)$ as $t \downarrow 0$. But $p_{ij}(t) = \mathbb{P}(X_t = \varepsilon_j \mid X_0 = \varepsilon_i) = \mathbb{E}(\mathbf{1}_{\{X_t=\varepsilon_j\}} \mid X_0 = \varepsilon_i) = \frac{1}{\mathbb{P}(X_0=\varepsilon_i)} \int_{\{X_0=\varepsilon_i\}} \mathbf{1}_{\{X_t=\varepsilon_j\}}(\omega)\, \mathrm{d}\,\mathbb{P}(\omega)$. Taking $t \downarrow 0$ and applying the bounded convergence theorem, we see the limit is $\delta_{i,j}$.

To prove the second part, we show that for any $h \in \mathbb{R}$,

$$|p_{ij}(t+h) - p_{ij}(t)| \leq 1 - p_{ii}(|h|)\,. \tag{4.6}$$

Then continuity will follow from the first part. To prove this fact, first suppose $h > 0$ and use Kolmogorov-Chapman to get

$$p_{ij}(t+h) - p_{ij}(t) = \sum_k p_{ik}(h)p_{kj}(t) - p_{ij}(t) = \sum_{k\neq i} p_{ik}(h)p_{kj}(t) - p_{ij}(t)[1 - p_{ii}(h)]\,.$$

Thus,

$$\begin{aligned} -[1 - p_{ii}(h)] \leq -p_{ij}(t)[1 - p_{ii}(h)] &\leq p_{ij}(t+h) - p_{ij}(t) \\ &\leq \sum_{k\neq i} p_{ik}(h)p_{kj}(t) \leq \sum_{k\neq i} p_{ik}(h) \leq 1 - p_{ii}(h)\,. \end{aligned}$$

This shows that $|p_{ij}(t+h) - p_{ij}(t)| \leq 1 - p_{ii}(h)$. On the other hand,

$$|p_{ij}(t-h) - p_{ij}(t)| = |p_{ij}(t) - p_{ij}(t-h)| \leq 1 - p_{ii}(h)$$

by what we just proved. This completes the proof of (4.6). □

The historic motivation [1] for studying such processes came from Kolmogorov in 1931. Having understood the theory of Markov chains, he tried to build a theory continuous-time processes which would satisfy the Kolmogorov-Chapman equation. He found two kinds of continuous processes that arise in this fashion, depending on the behavior of the system in small time intervals :

— if we assume that in a small time interval there is an overwhelming probability that the state will remain unchanged; however, if it changes, the change may be radical, then we are led to jump processes,

— if we assume on the contrary that the system is under continuous change, and that changes in small time intervals are also small, then we are lead to *diffusion processes.* A fundamental example here is Brownian motion. The random variables X_t are no longer discrete.

The fact that jump processes remain unchanged *in small times* with an overwhelming probability is implied by Lemma 4.4, in fact such probability is like $1 - \lambda_i h + o(h)$ if h is a small time.

1. Source : wikipedia.

4.2 Jump times and sojourn times

Any jump process has a sequence of times $0 = \tau_0 < \tau_1 < \tau_2 < \dots$ and a sequence of states $s_n \in S$ such that

$$X_t = s_n \qquad \text{if } \tau_n \le t < \tau_{n+1}\,.$$

We call these the *jump times*. These are random times : for each ω we have corresponding times $\tau_n(\omega)$. If the process is regular, then $\tau_\infty := \lim_{n\to\infty} \tau_n = +\infty$. Indeed, $\tau_\infty = M < \infty$ would mean there are infinitely many jumps in $[0, M]$, contrary to our assumption.

We denote

$$\tau := \tau_1$$

the time of first jump.

Jump times are examples of *stopping times*. We will not develop this here.

We define the *sojourn time in* ε to be the time it takes to leave the state ε and move to some other state. To any sequence of jump times $\tau_1 < \tau_2 < \dots$, there corresponds a sequence of sojourn times τ_1, $\tau_2 - \tau_1$, $\tau_3 - \tau_2$, $\dots$ in the visited states.

It is sometimes useful to start observing the process at time t_0. In this case we denote by $\tau_n + t_0$ the n-th jump after t_0.

Lemma 4.2. *The set* $\{\tau + t_0 > t\} \cap \{X_{t_0} = \varepsilon\}$ *is an event.*

Proof. (The student can skip this proof.)

Assume $t_0 = 0$, the general case is similar.

To check measurability, note that $\{\tau > t\} \cap \{X_0 = \varepsilon\} = \cap_{0\le s\le t}\{X_s = \varepsilon\}$. Each event $\{X_s = \varepsilon\}$ is measurable since X_s is a random variable. Unfortunately we have an uncountable intersection here, so we cannot deduce measurability immediately. To avoid this problem, we write $\{\tau > t\} \cap \{X_0 = \varepsilon\} = \cap_{s\in[0,t]\cap\mathbb{Q}}\{X_s = \varepsilon\}$. If this equality is true, then we now have a countable intersection, hence measurable. We trivially have $\{\tau > t\} \cap \{X_0 = \varepsilon\} \subseteq \cap_{s\in[0,t]\cap\mathbb{Q}}\{X_s = \varepsilon\}$. Conversely, if $X_s(\omega) = \varepsilon$ for all $s \in [0, t] \cap \mathbb{Q}$, then given $s' \in [0, t]$, we use right continuity of sample paths (assumption (c)) to deduce that $X_{s'}(\omega) = \lim_{s_n\downarrow s'} X_{s_j}(\omega) = \varepsilon$, where s_n is a sequence in $[0, t] \cap \mathbb{Q}$ converging to s' from the right. This shows that $X_{s'}(\omega) = \varepsilon$ for all $s' \in [0, t]$ and thus $\tau(\omega) > t$ as required. □

Corollary 4.3. *The jump time* τ *is a random variable.*

Proof. In fact $\{\tau > t\} = \cup_j \{\tau > t\} \cap \{X_0 = \varepsilon_j\}$ is a countable union of measurable sets, hence measurable. This implies τ is measurable. □

The next step is to know how this important random variable is distributed.

Lemma 4.4. *Let X_t be a regular jump process, $\varepsilon_i \in S$ and $t_0 \geq 0$. Then there exists $\lambda_i \geq 0$ such that*

$$\mathbb{P}(\tau + t_0 > t \mid X_{t_0} = \varepsilon_i) = \mathrm{e}^{-\lambda_i t}\,.$$

Moreover, λ_i is independent of t_0.

This shows that the system has two extreme behaviors : either stay in ε_i forever ($\lambda_i = 0$), or leave ε_i quickly ($\lambda_i > 0$, in which case the probability of staying a long time in ε_i is exponentially small).

Proof. By time homogeneity,

$$\begin{aligned}\mathbb{P}(\tau + t_0 > t \mid X_{t_0} = \varepsilon_i) &= \mathbb{P}(X_{t_0+s} = \varepsilon_i \ \forall s \leq t \mid X_{t_0} = \varepsilon_i) \\ &= \mathbb{P}(X_s = \varepsilon_i \ \forall s \leq t \mid X_0 = \varepsilon_i) = \mathbb{P}(\tau > t \mid X_0 = \varepsilon_i)\end{aligned}$$

so we may assume that $t_0 = 0$.

Denote $\mathbb{P}^i(A) = \mathbb{P}(A \mid X_0 = \varepsilon_i)$ and let $\varphi(t) = \mathbb{P}^i(\tau > t)$. Note that $\varphi(0) = \mathbb{P}(X_0 = \varepsilon_i \mid X_0 = \varepsilon_i) = 1$.

Using time homogeneity, we may guess that

$$\mathbb{P}^i(\tau > s + t \mid \tau > s) = \varphi(t)\,.$$

To prove this, we write

$$\begin{aligned}\mathbb{P}^i(\tau > s + t \mid \tau > s) &= \mathbb{P}(X_u = \varepsilon_i \ \forall 0 \leq u \leq s + t \mid X_u = \varepsilon_i \ \forall 0 \leq u \leq s) \\ &= \mathbb{P}(X_u = \varepsilon_i \ \forall s < u \leq s + t \mid X_u = \varepsilon_i \ \forall 0 \leq u \leq s) \\ &= \mathbb{P}(X_u = \varepsilon_i \ \forall s < u \leq s + t \mid X_s = \varepsilon_i) \\ &= \mathbb{P}(X_{u+s} = \varepsilon_i \ \forall 0 < u \leq t \mid X_s = \varepsilon_i) \\ &= \mathbb{P}(X_u = \varepsilon_i \ \forall 0 < u \leq t \mid X_0 = \varepsilon_i) \\ &= \mathbb{P}^i(\tau > t) = \varphi(t)\,.\end{aligned}$$

In the third step we used the Markov property. In the fifth we used time-homogeneity.

Using the total probability formula, we deduce that

$$\begin{aligned}\mathbb{P}^i(\tau > s + t) &= \mathbb{P}^i(\tau > s + t \mid \tau > s)\,\mathbb{P}^i(\tau > s) + \mathbb{P}^i(\tau > s + t \mid \tau \leq s)\,\mathbb{P}^i(\tau \leq s) \\ &= \varphi(t)\varphi(s) + 0\,.\end{aligned}$$

Hence, φ satisfies the following properties :

(1) $\varphi(s+t) = \varphi(s)\varphi(t)$,

(2) φ is right-continuous. In fact, $\varphi(t) = 1 - \mathbb{P}^i(\tau \leq t)$. The second term is the distribution function of τ, which is right-continuous (see Chapter 1).

(3) φ is not identically zero, since $\varphi(0) = 1$.

The following lemma shows that $\varphi(t)$ must be an exponential function. □

Lemma 4.5. *Any function $\varphi : [0,\infty) \to [0,1]$ satisfying (1), (2), (3) takes the form $\varphi(t) = \mathrm{e}^{-\lambda t}$ for some $\lambda \geq 0$.*

Proof. We only sketch the proof of this auxiliary result[2]. It suffices to show that $\varphi(t) = \varphi(1)^t$, since we then take $\lambda = -\ln \varphi(1) \geq 0$. We prove this progressively. First, since $\varphi(1) = \varphi(\frac{1}{n} + \cdots + \frac{1}{n}) = \varphi(\frac{1}{n})^n$, we get $\varphi(\frac{1}{n}) = \varphi(1)^{1/n}$. This proves the relation for t of the form $\frac{1}{n}$. We then generalize this to t of the form $\frac{p}{q}$. Then generalize this to all $t \geq 0$ by approximating reals with rationals and using right-continuity. □

Corollary 4.6. *Under $\mathbb{P}^i$, the random variable τ has an exponential distribution with parameter λ_i. In particular, the mean sojourn time in ε_i is $\mathbb{E}^i(\tau) = \frac{1}{\lambda_i}$.*

Proof. In fact, for any $t \geq 0$, $\mathbb{P}^i(\tau \leq t) = 1 - \varphi(t) = 1 - \mathrm{e}^{-\lambda_i t} = \int_0^t \lambda_i \mathrm{e}^{-\lambda_i t}\,\mathrm{d}t$. Also, if $t < 0$ then $\mathbb{P}^i(\tau \leq t) = 0$ since $\tau \geq 0$. Thus, $\mathbb{P}^i(\tau \leq t) = \int_{-\infty}^t f(t)\,\mathrm{d}t$ with $f(t) = \mathbf{1}_{[0,\infty)}\lambda_i \mathrm{e}^{-\lambda_i t}$. This is the exponential distribution.

Next, $\mathbb{E}^i(\tau) = \int t f(t)\,\mathrm{d}t = \lambda_i \int_0^\infty t\mathrm{e}^{-\lambda_i t}\,\mathrm{d}t = -[t\mathrm{e}^{-\lambda_i t}]_0^\infty + \int_0^\infty \mathrm{e}^{-\lambda_i t}\,\mathrm{d}t = \frac{1}{\lambda_i}$. □

We did not define the jump time in case of Markov chains. Let us introduce it and compare its behavior with that of jump processes[3].

Lemma 4.7. *Let (X_n) be a Markov chain with transition probabilities p_{ij}. Define τ to be the first time $n \geq 1$ such that $X_n \neq X_0$, i.e. the time of first jump. Then for any initial state ε_i, under $\mathbb{P}^i$,*

(i) $\mathbb{P}^i(\tau > n) = p_{ii}^n$,

(ii) *the distribution of τ is geometric with parameter $1 - p_{ii}$,*

(iii) *the random variable X_τ is independent of τ and has distribution*

$$\mathbb{P}^i(X_\tau = \varepsilon_j) = \begin{cases} \frac{p_{ij}}{1-p_{ii}} & \text{if } j \neq i, \\ 0 & \text{otherwise.} \end{cases}$$

2. See e.g. Foata-Fuchs' "Calcul des probabilités" Chapter 14, Exercise 1 (which is solved).
3. The following two lemmas are taken from [11]

Proof. We have $\mathbb{P}^i(\tau > n) = \mathbb{P}^i(X_k = \varepsilon_i \ \forall 1 \le k \le n) = p_{ii}^n$. The last equation is proved by induction. It is clear for $n = 1$. If true for n, then

$$\begin{aligned}\mathbb{P}^i(X_k = \varepsilon_i \ \forall 1 \le k \le n+1) &= \frac{\mathbb{P}(\cap_{k=0}^{n+1}\{X_k = \varepsilon_i\})}{\mathbb{P}(X_0 = \varepsilon_i)} \\ &= \frac{\mathbb{P}(\cap_{k=0}^{n+1}\{X_k = \varepsilon_i\})}{\mathbb{P}(\cap_{k=0}^{n}\{X_k = \varepsilon_i\})} \cdot \frac{\mathbb{P}(\cap_{k=0}^{n}\{X_k = \varepsilon_i\})}{\mathbb{P}(X_0 = \varepsilon_i)} \\ &= \mathbb{P}(X_{n+1} = \varepsilon_i \mid \cap_{k=0}^{n} X_k = \varepsilon_i) \cdot p_{ii}^n = p_{ii}^{n+1}\end{aligned}$$

where we used the induction hypothesis in the third step and the Markov property in the last one. This proves (i).

For (ii), note that $\mathbb{P}^i(\tau = n) = \mathbb{P}^i(\tau > n-1) - \mathbb{P}^i(\tau > n) = p_{ii}^{n-1} - p_{ii}^n = p_{ii}^{n-1}(1 - p_{ii})$, which is precisely the geometric distribution with parameter $1 - p_{ii}$ on $\mathbb{N}^*$.

For (iii), clearly $\mathbb{P}^i(X_\tau = \varepsilon_i) = 0$. Let $j \neq i$. Then arguing as before, we get

$$\begin{aligned}\mathbb{P}^i(\tau = n+1 \text{ and } X_\tau = \varepsilon_j) &= \mathbb{P}^i(X_k = \varepsilon_i \ \forall k \le n \text{ and } X_{n+1} = \varepsilon_j) \\ &= p_{ii}^n p_{ij} = \mathbb{P}^i(\tau = n+1) \cdot \frac{p_{ij}}{1 - p_{ii}}\,. \end{aligned} \tag{4.1}$$

Let $A_n = \{\tau = n+1\}$. Then using (4.1),

$$\mathbb{P}^i(X_\tau = \varepsilon_j) = \sum_{n=0}^{\infty} \mathbb{P}^i(X_\tau = \varepsilon_j \cap A_n) = \frac{p_{ij}}{1 - p_{ii}} \sum_{n=0}^{\infty} \mathbb{P}^i(A_n) = \frac{p_{ij}}{1 - p_{ii}}\,.$$

Thus X_τ has the distribution given in (iii). Moreover, (4.1) shows that the distribution of (τ, X_τ) is a product measure. Hence, X_τ is independent of τ. □

Properties (i) and (ii) show the discrete analog of Lemma 4.4 and Corollary 4.6. It is natural to ask if (iii) also holds for continuous processes. The answer is yes.

Lemma 4.8. *Let (X_t) be a regular jump process. Then under $\mathbb{P}^i$, X_τ is independent of τ and has distribution*

$$\mathbb{P}^i(X_\tau = \varepsilon_j) = \lim_{n\to\infty} \frac{p_{ij}(2^{-n})}{1 - p_{ii}(2^{-n})}, \qquad j \neq i\,.$$

Proof. Fix $n \in \mathbb{N}^*$ and consider the discrete-time process $X_k^{(n)}(\omega) = X_{\frac{k}{2^n}}(\omega)$, $k = 0, 1, 2 \ldots$. Then $(X_k^{(n)})$ is a Markov chain with transition probabilities $p_{ij}(2^{-n})$. In particular, under $\mathbb{P}^i$, the jump time $\tau^{(n)}$ of this chain has the geometric distribution with parameter $1 - p_{ii}(2^{-n})$ by Lemma 4.7.

By right-continuity of sample paths, $\{\tau > t\} = \cap_n \{\tau^{(n)} > \lfloor 2^n t \rfloor\}$. This is proved as in Lemma 4.2 using the density of dyadic rationals (instead of general rationals). Namely, if $X_0 = \varepsilon_i$, we get by right-continuity that

$$\begin{aligned}\tau > t &\iff X_s = \varepsilon_i \ \forall s \le t \iff \forall n : X_{\frac{k}{2^n}} = \varepsilon_i \ \forall k \le \lfloor 2^n t \rfloor \\ &\iff \forall n : X_k^{(n)} = \varepsilon_i \ \forall k \le \lfloor 2^n t \rfloor \iff \forall n : \tau^{(n)} > \lfloor 2^n t \rfloor .\end{aligned}$$

We also note that the sets $\{\tau^{(n)} > \lfloor 2^n t \rfloor\}$ are decreasing. In fact, if $\tau^{(n+1)} > \lfloor 2^{n+1} t \rfloor$, then $X_s = \varepsilon_i$ for all $s = \frac{k}{2^{n+1}}$, $k \le 2^{n+1}$. In particular, $X_s = \varepsilon_i$ for all $s = \frac{j}{2^n}$, $j \le 2^n$ by taking $k = 2j$. Hence, $\tau^{(n)} > \lfloor 2^n t \rfloor$ as asserted.

Next, note that

— $\tau \le \frac{\tau^{(n)}}{2^n}$. In fact, by definition $X^{(n)}_{\tau^{(n)}} \ne X_0^{(n)}$, so $X_{\frac{\tau^{(n)}}{2^n}} \ne X_0$ and thus the first jump τ for the continuous process is $\le \frac{\tau^{(n)}}{2^n}$.

— Similarly, $\frac{\tau^{(n+1)}}{2^{n+1}} \le \frac{\tau^{(n)}}{2^n}$. In fact, as $X_{\frac{\tau^{(n)}}{2^n}} \ne X_0$, then the first time $k = \tau^{(n+1)}$ such that $X_{\frac{k}{2^{n+1}}} \ne X_0$ satisfies $k \le 2\tau^{(n)}$, since at $2\tau^{(n)}$ we already left X_0.

— Finally $\frac{\tau^{(n)}}{2^n} \to \tau$. In fact, $\frac{\tau^{(n)}-1}{2^n} \le \tau$. This is because $X^{(n)}_{\tau^{(n)}-1} = X_0$, so $X_{\frac{\tau^{(n)}-1}{2^n}} = X_0$ and thus the jump time τ for the continuous chain is $\ge \frac{\tau^{(n)}-1}{2^n}$. Combined with the first item this gives $|\tau - \frac{\tau^{(n)}}{2^n}| \le \frac{1}{2^n} \to 0$.

Thus, $\frac{\tau^{(n)}}{2^n} \downarrow \tau$. Using right-continuity, it follows that for any ω, $X_{\frac{\tau^{(n)}}{2^n}}(\omega) = X_\tau(\omega)$ for all n large enough.

We are finally ready to conclude. In fact, since $\{\tau > t\}$ is the intersection of decreasing sets $\{\tau^{(n)} > \lfloor 2^n t \rfloor\}$, we get by continuity of $\mathbb{P}$ that [4]

$$\begin{aligned}\mathbb{P}^i(X_\tau = \varepsilon_j \text{ and } \tau > t) &= \lim_{n\to\infty} \mathbb{P}^i(X_{\frac{\tau^{(n)}}{2^n}} = \varepsilon_j \text{ and } \tau^{(n)} > \lfloor 2^n t \rfloor) \\ &= \lim_{n\to\infty} \mathbb{P}^i(X_{\frac{\tau^{(n)}}{2^n}} = \varepsilon_j)\, \mathbb{P}^i(\tau^{(n)} \ge \lfloor 2^n t \rfloor) \\ &= \mathbb{P}^i(X_\tau = \varepsilon_j)\, \mathbb{P}^i(\tau > t) \qquad (4.2)\end{aligned}$$

where we used Lemma 4.7 in the second step. This shows independence. The distribution of X_τ also follows from Lemma 4.7 by right-continuity. □

4. Details : let $A_n = \{\tau^{(n)} > \lfloor 2^n t \rfloor\}$, $A = \{\tau > t\}$, $B_n = \{X_{\frac{\tau^{(n)}}{2^n}} = \varepsilon_j\}$, $B = \{X_\tau = \varepsilon_j\}$. As A_n are decreasing and $\cap A_n = A$, we have $\mathbb{P}^i(A) = \lim_n \mathbb{P}^i(A_n)$. Next, by right-continuity, for any n, there is some n_ω such that $X_{\frac{\tau^{(n)}}{2^n}}(\omega) = X_\tau(\omega)$ for all $n \ge n_\omega$, so $\mathbf{1}_{B_n}(\omega) = \mathbf{1}_B(\omega)$ for $n \ge n_\omega$, so $\mathbf{1}_{B_n} \to \mathbf{1}_B$ a.s. and thus $\mathbb{P}^i(B_n) = \mathbb{E}^i(\mathbf{1}_{B_n}) \to \mathbb{E}^i(\mathbf{1}_B) = \mathbb{P}^i(B)$ by bounded convergence. This proves the last equality in (4.2). The first one is quite similar. We have $\mathbf{1}_{A_n \cap B_n}(\omega) = \mathbf{1}_{A_n \cap B}(\omega)$ for $n \ge n_\omega$. If $\omega \in A \cap B$ then $\omega \in A_n \cap B$ for all n so $\mathbf{1}_{A_n \cap B}(\omega) = 1 = \mathbf{1}_{A \cap B}(\omega)$. Conversely, if $\omega \notin A \cap B$, then $\omega \notin A_m \cap B$ for some $m = m_\omega$, so $\omega \notin A_n \cap B$ for all $n \ge m_\omega$ since A_n are decreasing, so $\mathbf{1}_{A_n \cap B}(\omega) = 0 = \mathbf{1}_{A \cap B}(\omega)$ for all $n \ge m_\omega$. Combined, we have shown that $\mathbf{1}_{A_n \cap B_n} \to \mathbf{1}_{A \cap B}$ a.s. Conclude again by bounded convergence.

We conclude this section with the following fact :

Theorem 4.9 (Strong Markov Property)**.** *Let (X_t) be a regular jump process with transition matrix $P(t)$, and let τ be a stopping time with respect to (X_t). Then, given that $X_\tau = \varepsilon_k$,*

(i) *the process after τ is a regular jump process with transition matrix $P(t)$,*

(ii) *the process after τ and the process before τ are independent.*

We will only need this theorem in the case of $\tau = \tau_1$, the time of first jump.

Proof. Admitted, see [4, Theorem 8.4.1]. □

4.3 The Kolmogorov equations

The aim of this section is to derive a system of differential equations that a typical jump process must satisfy. This can be used to calculate $p_{ij}(t)$ in simple situations. We begin by studying the infinitesimal behavior of the transition probabilities.

If h is a small parameter, recall that we say $f(h) = o(h)$ if $f(h)$ is negligible in front of h, that is, $\frac{f(h)}{h} \to 0$ as $h \to 0$. For example, the Taylor-Young formula tells us that $\mathrm{e}^h = 1 + h + \frac{h^2}{2!} + \cdots + \frac{h^n}{n!} + o(h^n)$, for any n.

Be careful that $o(h) \pm o(h) = o(h)$ and $C \cdot o(h) = o(h)$.

Now define

$$\rho_{ij} := \mathbb{P}(X_\tau = \varepsilon_j \mid X_0 = \varepsilon_i)$$

and denote

$$\lambda_{ij} := \begin{cases} \lambda_i \rho_{ij} & j \neq i, \\ -\lambda_i & j = i, \end{cases} \tag{4.1}$$

where $\lambda_i \geq 0$ is the sojourn parameter appearing in Lemma 4.4. Note that

$$\sum_{j \neq i} \lambda_{ij} = \lambda_i \sum_{j \neq i} \rho_{ij} = \lambda_i \,. \tag{4.2}$$

For simplicity, we shall assume in this section that

$$\sup_j \lambda_j < \infty \,. \tag{4.3}$$

Recall that if $\lambda_i = 0$, then the system stays forever in ε_i. A very big λ_i would mean on the contrary that the system spends very little time in ε_i. Assumption (4.3) can be interpreted as saying that there is some $\delta > 0$ such that for any visited state ε_i, the system spends some time $t \geq \delta$ in ε_i.

Lemma 4.10. *If a regular jump process satisfies* (4.3)*, we have in infinitesimal times h,*

$$\begin{aligned} 1 - p_{ii}(h) &= \lambda_i h + o(h) \qquad i = 1, 2, \dots, \\ p_{ij}(h) &= \lambda_{ij} h + o(h) \qquad j \neq i,\ i, j = 1, 2, \dots . \end{aligned} \tag{4.4}$$

The lemma implies that $p_{ij}(t)$ is differentiable at zero (from the right), with

$$p'_{ii}(0) = -\lambda_i = \lambda_{ii} \qquad \text{and} \qquad p'_{ij}(0) = \lambda_{ij} \quad \text{for } j \neq i\,.$$

This result is actually true without assumption (4.3), see [4, Chapter 8] and [1] for a very different argument. Our proof here follows [14] instead, but with gaps filled from [4]. The above result is even true without assuming we have a jump process, i.e. only assuming a continuous Markov process without assumptions (c),(d). However in this case $p'_{ii}(0)$ may be infinite. See [4, Chapter 8] for details.

Proof. By Lemma 4.4, we know that

$$\mathbb{P}^i(\text{at least one jump by time } h) = \mathbb{P}^i(\tau \leq h) = 1 - \mathrm{e}^{-\lambda_i h} = \lambda_i h + o(h)\,. \tag{4.5}$$

and by Lemma 4.8,

$$\mathbb{P}^i(\tau \leq h \text{ and } X_\tau = \varepsilon_k) = (1 - \mathrm{e}^{-\lambda_i h})\rho_{ik} = \lambda_i \rho_{ik} h + o(h)\,. \tag{4.6}$$

Thus,

$$\begin{aligned} \mathbb{P}^i(\text{at least two jumps by time } h) &\leq \sum_{k \neq i} \mathbb{P}^i(\tau_1 \leq h \text{ and } X_{\tau_1} = \varepsilon_k \text{ and } \tau_2 - \tau_1 \leq h) \\ &= \sum_{k \neq i} \mathbb{P}^i(\tau_1 \leq h \text{ and } \tau_2 - \tau_1 \leq h \mid X_{\tau_1} = \varepsilon_k)\rho_{ik} \\ &= \sum_{k \neq i} \rho_{ik}\, \mathbb{P}^i(\tau_1 \leq h \mid X_{\tau_1} = \varepsilon_k)\, \mathbb{P}^i(\tau_2 - \tau_1 \leq h \mid X_{\tau_1} = \varepsilon_k) \\ &= \sum_{k \neq i} \rho_{ik}\, \mathbb{P}^i(\tau \leq h)\, \mathbb{P}^k(\tau \leq h) \\ &= \sum_{k \neq i} (1 - \mathrm{e}^{-\lambda_i h})\rho_{ik}(1 - \mathrm{e}^{-\lambda_k h}) = o(h) \end{aligned} \tag{4.7}$$

where we used Theorem 4.9 in the third and fourth lines, Lemma 4.8 in the fourth line, and the last equality holds using (4.3) and $\sum_k \rho_{ik} = 1$.

On the other hand,

$$\mathbb{P}^i(\text{no jump by time } h) = \mathbb{P}^i(\tau > h) = \mathrm{e}^{-\lambda_i h} = 1 - \lambda_i h + o(h)\,. \tag{4.8}$$

We are ready to prove the lemma. We have

$$p_{ii}(h) = \mathbb{P}^i(\text{no jump by time } h) + \mathbb{P}^i(\text{at least two jumps by time } h \text{ and } X_h = \varepsilon_i)$$

Using (4.8) and (4.7), deduce that $p_{ii}(h) = 1 - \lambda_i h + o(h)$.
Similarly, for $j \neq i$,

$$\begin{aligned} p_{ij}(h) &= \mathbb{P}^i(\text{exactly one jump by time } h \text{ and } X_h = \varepsilon_j) \\ &\quad + \mathbb{P}^i(\text{at least two jumps by time } h \text{ and } X_h = \varepsilon_j) \\ &= \mathbb{P}^i(\tau \leq h \text{ and } X_s = \varepsilon_j \ \forall s \in [\tau, h]) + o(h) \\ &= \mathbb{P}^i(\tau \leq h \text{ and } X_\tau = \varepsilon_j) \\ &\quad - \mathbb{P}^i(\tau \leq h \text{ and } X_\tau = \varepsilon_j \text{ and } \exists s \in [\tau, h] : X_s \neq \varepsilon_j) + o(h) \\ &= \lambda_i \rho_{ij} + o(h) \end{aligned}$$

where we used (4.7) several times and (4.6) in the end. □

Theorem 4.11 (Kolmogorov equation)**.** *Let (X_t) be a jump process satisfying* (4.3) *and define λ_{ij} as in* (4.1)*. Then the transition probabilities $p_{ij}(t)$ satisfy two systems of linear differential equations : the* forward Kolmogorov equations

$$p'_{ij}(t) = \sum_k p_{ik}(t)\lambda_{kj}\,, \qquad i, j = 1, 2, \ldots \tag{4.9}$$

and the backward Kolmogorov equations

$$p'_{ij}(t) = \sum_k \lambda_{ik} p_{kj}(t)\,, \qquad i, j = 1, 2, \ldots \tag{4.10}$$

subject to initial conditions $p_{ij}(0) = \delta_{ij}$, $i, j = 1, 2, \ldots$.

In matrix form, if $\Lambda = (\lambda_{ij})$, then the equations read as $P'(t) = P(t)\Lambda$ and $P'(t) = \Lambda P(t)$, respectively. Roughly speaking, this says that $P(t) = \mathrm{e}^{t\Lambda}$. The precise terminology is that Λ is the *infinitesimal generator of the semigroup* $P(t)$.

Again, the theorem is partly true without assumption (4.3). Moreover, the backward equation holds under weaker conditions than the forward equation. See [1, 4] for details.

Proof. By Kolmogorov-Chapman we have for $h > 0$,

$$p_{ij}(t + h) = \sum_k p_{ik}(t) p_{kj}(h) = \sum_k p_{ik}(h) p_{kj}(t)\,. \tag{4.11}$$

Using the first expansion we get

$$\frac{p_{ij}(t + h) - p_{ij}(t)}{h} = p_{ij}(t) \cdot \frac{p_{jj}(h) - 1}{h} + \sum_{k \neq j} p_{ik}(t) \frac{p_{kj}(h)}{h}\,.$$

As $h \downarrow 0$, the first term goes to $-p_{ij}(t)\lambda_j = p_{ij}(t)\lambda_{jj}$, and each term in the sum goes to $p_{ik}(t)\lambda_{kj}$, by Lemma 4.10. To obtain (4.9), it remains to justify the interchange of $\lim_{h\downarrow 0}$ and $\sum_{k\neq j}$. For this, note that for $k \neq j$,

$$\frac{p_{kj}(h)}{h} = \frac{\mathbb{P}(X_h = \varepsilon_j \mid X_0 = \varepsilon_k)}{h} \leq \frac{\mathbb{P}(\tau \leq h \mid X_0 = \varepsilon_k)}{h} = \frac{1 - \mathrm{e}^{-\lambda_k h}}{h} \leq \lambda_k$$

using[5] that $1 - \mathrm{e}^{-x} \leq x$ for $x \geq 0$ and Lemma 4.4. But by (4.3), $\lambda_k \leq C$ for all k. Thus, $p_{ik}(t)\frac{p_{kj}(h)}{h} \leq Cp_{ik}(t)$, and $\sum_k p_{ik}(t) = 1 < \infty$. It follows by Lebesgue's dominated convergence theorem that we can interchange series and limit.

This almost proves (4.9) (see the end of the proof). We now turn to the second expansion in (4.11). This time we get

$$\frac{p_{ij}(t+h) - p_{ij}(t)}{h} = \frac{p_{ii}(h) - 1}{h} \cdot p_{ij}(t) + \sum_{k\neq i} \frac{p_{ik}(h)}{h} p_{kj}(t)\,.$$

Again we have the desired limit (4.10) if interchange of limits is justified. The above argument says that $\frac{p_{ik}(h)}{h} \leq \lambda_i$. Unfortunately there is no reason for $\sum_k p_{kj}(t)$ to converge. So we argue as follows. Since

$$\sum_{k\neq i} \frac{p_{ik}(h)}{h} p_{kj}(t) \geq \sum_{\substack{k\neq i \\ k\leq n}} \frac{p_{ik}(h)}{h} p_{kj}(t)$$

for any n, taking $h \downarrow 0$ followed by $n \to \infty$ we get

$$\liminf_{h\downarrow 0} \sum_{k\neq i} \frac{p_{ik}(h)}{h} p_{kj}(t) \geq \sum_{k\neq i} \lambda_{ik} p_{kj}(t)\,.$$

On the other hand, for $n > i$,

$$\begin{aligned}\sum_{k\neq i} \frac{p_{ik}(h)}{h} p_{kj}(t) &\leq \sum_{\substack{k\neq i \\ k\leq n}} \frac{p_{ik}(h)}{h} p_{kj}(t) + \sum_{k>n} \frac{p_{ik}(h)}{h} \\ &= \sum_{\substack{k\neq i \\ k\leq n}} \frac{p_{ik}(h)}{h} p_{kj}(t) + \frac{1 - p_{ii}(h)}{h} - \sum_{\substack{k\neq i \\ k\leq n}} \frac{p_{ik}(h)}{h}\end{aligned}$$

so taking $h \downarrow 0$ followed by $n \to \infty$ gives

$$\limsup_{h\downarrow 0} \sum_{k\neq i} \frac{p_{ik}(h)}{h} p_{kj}(t) \leq \sum_{k\neq i} \lambda_{ik} p_{kj}(t) + \lambda_i - \sum_{k\neq i} \lambda_{ik}\,.$$

Using (4.2), we deduce that $\limsup_{h\downarrow 0} \sum_{k\neq i} \frac{p_{ik}(h)}{h} p_{kj}(t) \leq \sum_{k\neq i} \lambda_{ik} p_{kj}(t)$. This completes the proof of the interchange.

5. To prove this, define $f(x) = x + \mathrm{e}^{-x} - 1$. Then $f(0) = 0$ and $f'(x) = 1 - \mathrm{e}^{-x} \geq 0$ if $x \geq 0$. Consequently f increases, so $f(x) \geq f(0) = 0$ for $x \geq 0$ and thus $x \geq 1 - \mathrm{e}^{-x}$.

We have almost proved the theorem. Actually we showed that $p_{ij}(t)$ has *right-derivatives* given by (4.9), (4.10), since we must take $h > 0$ in the above arguments. But the RHS of (4.9), (4.10) is continuous in t by Lemma 4.1 (this time interchange in (4.10) is easy using (4.2)). Any function with continuous right-derivatives is differentiable (this is not difficult to prove, see [1, Lemma 1.7.2]). This completes the proof of the theorem. □

4.4 Poisson and related processes

4.4.1 The Poisson Process

We now discuss a simple yet important jump process : the Poisson process. There are many equivalent ways to define it. Here we follow [13] and [8, Section 3.6].

Suppose that certain events occur randomly in time, for example telephone calls at a call center. Let X_t be the number of events that occur in $[0, t]$.

What is the distribution of X_t ?

To answer this question, we make the following assumptions :

a) X_t is the number of events in $[0, t]$. In particular $X_0 = 0$ (no events when starting observation).

b) The number of events occurring in disjoint time intervals are independent. In other words, for any $t_1 < t_2 < \dots$, the random variables X_{t_1}, $X_{t_2} - X_{t_1}$, $X_{t_3} - X_{t_2}, \dots$ are independent.

c) The flow of events is stationary, i.e. the distribution of the number of events only depends on the length of the time interval. Thus, $X_{t_n} - X_{t_{n-1}}$ has the same distribution as $X_{t_n - t_{n-1}}$ for each n.

A process satisfying these assumptions is called a *Poisson process*. The state space in this case is $\mathbb{N} = \{0, 1, 2, \dots\}$.

Lemma 4.12. *The Poisson process is a regular jump process.*

Proof. Denote $N(a, b] = X_b - X_a$ the number of events occurring in $(a, b]$.

Let $s > 0$, denote $C = \{X_{s_1} = i_1, \dots, X_{s_k} = i_k\}$, where $s_k \leq s$. Then

$$\begin{aligned}\mathbb{P}(X_{t+s} = j \mid X_s = i, C) &= \frac{\mathbb{P}(N[0,t+s] = j \text{ and } N[0,s] = i \text{ and } C)}{\mathbb{P}(X_s = i, C)} \\ &= \frac{\mathbb{P}(N(s,s+t] = j-i \text{ and } N[0,s] = i \text{ and } C)}{\mathbb{P}(X_s = i \text{ and } C)} \\ &= \frac{\mathbb{P}(N(s,s+t] = j-i)\,\mathbb{P}(N[0,s] = i \text{ and } C)}{\mathbb{P}(X_s = i \text{ and } C)} \\ &= \mathbb{P}(N(s,s+t] = j-i) = \mathbb{P}(N(0,t] = j-i)\end{aligned}$$

where we used b) in the third step and c) in the last one. This shows both time homogeneity (independence of s) and Markov property (information C is irrelevant). We also deduce that

$$p_{ij}(t) = p_{0,j-i}(t) \tag{4.1}$$

since $\mathbb{P}(N(0,t] = j-i) = \mathbb{P}(X_t = j-i \mid X_0 = 0)$, as we know $X_0 = 0$.

The process has right-continuous sample paths. In fact, if the number of events in $[0,t]$ is n, then even if some event occurs right after t, say at time $t+\delta$, then we still have[6] $X_{t+\frac{\delta}{2}} = X_t = n$.

Finally the number of events in a finite time interval $[0,M]$ is $X_M \in \mathbb{N}$ which is finite by definition. □

This process has additional properties : X_t increases with time, and it has jumps of size one (when an event occurs, we pass from i to $i+1$).

To find the distribution of X_t, we shall use the Kolmogorov equations[7]

We first calculate λ_i. Recall it is related to the sojourn time in state i. So suppose at time t_0 the process is at i. Let $\varphi_i(t) = \mathbb{P}(\tau + t_0 > t \mid X_{t_0} = i)$. Then

$$\begin{aligned}\varphi_i(t+s) &= \mathbb{P}(X_u = i \ \forall u \in (t_0, t_0+t+s] \mid X_{t_0} = i) \\ &= \frac{\mathbb{P}(N(t_0, t_0+t+s] = 0 \text{ and } N[0,t_0] = i)}{\mathbb{P}(N[0,t_0] = i)} \\ &= \mathbb{P}(N(t_0, t_0+t+s] = 0) = \mathbb{P}(N(0,t+s] = 0)\end{aligned}$$

But

$$\begin{aligned}\mathbb{P}(N(0,t+s] = 0) &= \mathbb{P}(N(0,t] = 0 \text{ and } N(t,t+s] = 0) \\ &= \mathbb{P}(N(0,t] = 0)\,\mathbb{P}(N(0,s] = 0)\,.\end{aligned}$$

6. Technically right-continuity holds because we count events in $[0,t]$. If we counted in $[0,t)$, then it could be that the number there is n, while the number in $[0,t]$ is $n+1$ (if a new event occurs at t), which would violate right-continuity.

7. An alternative route to find the distribution is to use generating functions, see [13, p.74]. Yet another derivation appears in [8, Section 3.6]

Hence, $\varphi_i(t+s) = \varphi_i(t)\varphi_i(s)$ is multiplicative and actually independent of i, since $\varphi_i(t) = \mathbb{P}(N(0,t] = 0)$. It follows that there exists $\lambda \geq 0$ such that $\varphi_i(t) = \mathrm{e}^{-\lambda t}$. Comparing with Lemma 4.4, it follows that $\lambda_i = \lambda$ for all i.

Next, recall that ρ_{ij} is the probability that if the process is at i, after jumping for the first time it arrives at j. Clearly $\rho_{ij} = 0$ if $j \neq i+1$ and $\rho_{i,i+1} = 1$. It follows from (4.1) that

$$\lambda_{ij} = \begin{cases} \lambda & \text{if } j = i+1, \\ -\lambda & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

In view of (4.1) we denote

$$g_k(t) := p_{0k}(t)\,, \qquad k = 0, 1, 2, \ldots$$

Then the forward Kolmogorov equations (4.9) take the form

$$\begin{aligned} g_0'(t) &= \sum_k p_{0k}(t)\lambda_{k0} = -\lambda g_0(t)\,, \\ g_j'(t) &= \sum_k p_{0k}(t)\lambda_{kj} = \lambda g_{j-1}(t) - \lambda g_j(t)\,, \qquad j = 1, 2, \ldots \end{aligned}$$

To solve this system, consider

$$f_j(t) := \mathrm{e}^{\lambda t} g_j(t)\,.$$

Then

$$\begin{aligned} f_0'(t) &= \lambda f_0(t) + \mathrm{e}^{\lambda t} g_0'(t) = \lambda f_0(t) - \lambda \mathrm{e}^{\lambda t} g_0(t) = 0\,, \\ f_j'(t) &= \lambda f_j(t) + \mathrm{e}^{\lambda t} g_j'(t) \\ &= \lambda f_j(t) + \mathrm{e}^{\lambda t}(\lambda g_{j-1}(t) - \lambda g_j(t)) = \lambda f_{j-1}(t) \qquad j = 1, 2, \ldots \end{aligned} \tag{4.3}$$

Moreover

$$f_0(0) = 1 \qquad \text{and} \qquad f_j(0) = 0\,, \qquad j = 1, 2, \ldots$$

since $p_{ij}(0) = \delta_{ij}$. The solution of (4.3) subject to these initial conditions is clearly

$$f_0(t) = 1\,, \qquad f_1(t) = \lambda t\,, \ldots\,, \qquad f_n(t) = \frac{(\lambda t)^n}{n!}\,, \ldots$$

Thus, $p_{0j}(t) = \frac{(\lambda t)^j}{j!}\mathrm{e}^{-\lambda t}$. Recalling a), we thus get

$$\mathbb{P}(X_t = j) = \frac{(\lambda t)^j}{j!}\mathrm{e}^{-\lambda t}\,, \qquad j = 0, 1, 2, \ldots \tag{4.4}$$

We thus answered the question : the distribution of the number of events follows the Poisson distribution (hence the name of the process).

As previously mentioned, there are other equivalent ways to define this process. A common way is actually to start with (4.4), along with a), b) and prove c) instead. The definition presented here is in some sense more qualitative. Yet another approach, more "constructive", is to consider a sequence of i.i.d. random variables $\eta_1, \eta_2, \ldots$, each having an exponential distribution, put $\xi_n = \sum_{k=1}^{n} \eta_k$ and let $X_t := \max\{n : t \geq \xi_n\}$. Then X_t is again the Poisson process. This definition is interpreted as follows : the sequence (η_k) can be thought of as the times between the emissions of radioactive particles[8], ξ_n is then the time of the n-th emission, and X_t is the number of emissions up to time $t \geq 0$. From here one can prove b), c) and (4.4). See e.g. [5, Chapter 6] for this approach.

4.4.2 Pure Birth Process

We may generalize the Poisson process by considering a regular jump process with the same state space $S = \mathbb{N}$, but with transitions rates

$$\lambda_{ij} = \begin{cases} \lambda_i & \text{if } j = i+1, \\ -\lambda_i & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

instead of (4.2). This gives rise to the *pure birth process*. The only difference with Poisson is that in pure birth, the rate of leaving the state can depend on the state.

4.4.3 Birth/Death Process

A pure birth process always jumps from i to $i+1$ in a transition. We may generalize this by also allowing jumps to $i-1$. This gives rise to the *birth/death process*. We thus want the process to have transition rates (4.5). As this is an important process in queueing theory, we pause to discuss how it arises in practice.

Consider a state space $S = \{0, 1, \ldots, m\}$ or $S = \mathbb{N}$, where state i means we have i individuals in the population. Fix *birth rates* $b_i \geq 0$ and *death rates* $d_i \geq 0$.

If $X_t = i$, in the next jump we may have a birth or a death in the population. It is reasonable to assume the birth and death events are independent. On the other hand, if we want the process to be Markovian, the sojourn time in i must be exponentially distributed by Lemma 4.2. Hence, if $X_t = i$, we

8. which should be exponentially distributed, as they are sojourn times.

run two independent exponential clocks B_i and D_i with rates b_i and d_i, respectively. If $B_i < D_i$ (meaning that a birth occurred first), the process jumps to $i+1$ at time $t + B_i$. If $D_i < B_i$, the process jumps to $i-1$ at time $t + D_i$. It follows from the hypotheses that the joint density of B_i and D_i is given by $f_{B_i,D_i}(s,t) = (b_i \mathrm{e}^{-b_i s}\mathbf{1}_{[0,\infty)}(s)) \cdot (d_i \mathrm{e}^{-d_i t}\mathbf{1}_{[0,\infty)}(t))$. Hence,

$$\begin{aligned}\rho_{i,i+1} = \mathbb{P}(B_i < D_i) &= \iint \mathbf{1}_{\{s<t\}}(s,t) f_{B_i,D_i}(s,t)\,\mathrm{d}s\mathrm{d}t = b_i d_i \int_0^\infty \int_0^t \mathrm{e}^{-b_i s}\,\mathrm{d}s \mathrm{e}^{-d_i t}\,\mathrm{d}t \\ &= d_i \int_0^\infty (1-\mathrm{e}^{-b_i t})\mathrm{e}^{-d_i t}\,\mathrm{d}t = d_i\Big(\frac{1}{d_i} - \frac{1}{b_i+d_i}\Big) = \frac{b_i}{b_i+d_i}\,.\end{aligned}$$

Similarly [9], $\rho_{i,i-1} = \frac{d_i}{b_i+d_i}$. On the other hand, the sojourn time in i is found from

$$\mathbb{P}(\text{no birth and no death in time } s \mid X_t = i) = \mathrm{e}^{-b_i s}\mathrm{e}^{-d_i s} = \mathrm{e}^{-(b_i+d_i)s}\,.$$

Thus, the sojourn parameter here is $\lambda_i = (b_i + d_i)$. It follows from (4.1) that the transition rates are given by

$$\lambda_{ij} = \begin{cases} b_i & \text{if } j = i+1, \\ d_i & \text{if } j = i-1, \\ -(b_i+d_i) & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

For future calculations, we remind the reader that b_i and d_i are by definition the parameters such that

(4.6)

$$\mathbb{P}(\text{no birth in time } s \mid X_t = i) = \mathrm{e}^{-b_i s}\,, \quad \mathbb{P}(\text{no death in time } s \mid X_t = i) = \mathrm{e}^{-d_i s}\,.$$

This completes the description of the process. The preceding derivation is not completely rigorous, for example we haven't shown that the process we described is indeed a regular jump process. We refer the reader to [4, Section 9.1.2] for full proofs.

Let us study two examples with finite state space :

(1) Suppose that telephone calls arrive at random times to a call center. Suppose the call center has only one employee. Then the service system has two possible states : either being free ε_0, or busy ε_1. If the service is free, it

9. Note that we ignored the event $B_i = D_i$. This is because it has probability zero. In fact, intuitively, $\mathbb{P}(B_i = D_i) = \iint \mathbf{1}_{\{s=t\}}(s,t) f_{B_i,D_i}(s,t)\,\mathrm{d}s\mathrm{d}t = b_i d_i \int_0^\infty \int_t^t \mathrm{e}^{-b_i s}\,\mathrm{d}s \mathrm{e}^{-d_i t}\,\mathrm{d}t = 0$. If that doesn't seem convincing, calculate $\mathbb{P}(|B_i - D_i| < 1/n)$ the same way and take $n \to \infty$, noting these events are decreasing.

accepts calls, otherwise calls are rejected [10]. We may regard this as a birth-death process if we assume both the times between random calls and the random service times (i.e. time to finish ongoing phone call with the agent) are independent and exponentially distributed with rates $\lambda > 0$ and $\mu > 0$, as described above. In this special framework (4.1) reduces to two systems :

$$\lambda_{0j} = \begin{cases} \lambda & \text{if } j = 1, \\ -\lambda, & \text{if } j = 0, \end{cases} \qquad \lambda_{1j} = \begin{cases} \mu & \text{if } j = 0, \\ -\mu & \text{if } j = 1. \end{cases}$$

Indeed, $\mathbb{P}(\text{no call in time } t) = \mathrm{e}^{-\lambda t}$, $\mathbb{P}(\text{remain busy for time } t) = \mathrm{e}^{-\mu t}$, $\rho_{01} = \rho_{10} = 1$. The forward Kolmogorov equation (4.9) thus take the form

$$\begin{aligned} p'_{00}(t) &= p_{00}(t)\lambda_{00} + p_{01}(t)\lambda_{10} = -\lambda p_{00}(t) + \mu[1 - p_{00}(t)]\,, \\ p'_{11}(t) &= p_{10}(t)\lambda_{01} + p_{11}(t)\lambda_{11} = \lambda[1 - p_{11}(t)] - \mu p_{11}(t)\,. \end{aligned}$$

In other words,

$$\begin{aligned} p'_{00}(t) + (\lambda + \mu)p_{00}(t) &= \mu \\ p'_{11}(t) + (\lambda + \mu)p_{11}(t) &= \lambda\,. \end{aligned}$$

With the initial condition $p_{00}(0) = p_{11}(0) = 1$, the solution is [11]

$$\begin{aligned} p_{00}(t) &= \Big(1 - \frac{\mu}{\lambda + \mu}\Big)\mathrm{e}^{-(\lambda+\mu)t} + \frac{\mu}{\lambda + \mu}\,, \\ p_{11}(t) &= \Big(1 - \frac{\lambda}{\lambda + \mu}\Big)\mathrm{e}^{-(\lambda+\mu)t} + \frac{\lambda}{\lambda + \mu}\,. \end{aligned} \tag{4.7}$$

Finally, we may deduce $p_{01}(t)$ and $p_{10}(t)$ using

$$p_{01}(t) = 1 - p_{00}(t) \qquad \text{and} \qquad p_{10}(t) = 1 - p_{11}(t)\,.$$

(2) Consider a call center again, but this time suppose we have m employees to receive calls. The system then has $m + 1$ possible states $\{\varepsilon_0, \ldots, \varepsilon_m\}$, where ε_j means that j employees are busy. We assume the random arrival time is exponentially distributed with parameter $\lambda > 0$. We also assume the time it takes for any employee to service a call is exponentially distributed with parameter $\mu > 0$, and assume all random times are independent. We claim

$$\lambda_{0j} = \begin{cases} \lambda & \text{if } j = 1, \\ -\lambda, & \text{if } j = 0, \\ 0 & \text{otherwise,} \end{cases} \qquad \lambda_{mj} = \begin{cases} m\mu & \text{if } j = m - 1, \\ -m\mu & \text{if } j = m, \\ 0 & \text{otherwise,} \end{cases} \tag{4.8}$$

10. so no call is put on hold.

11. As the student learned in ODE course, the solution of $af'(t) + bf(t) + c = 0$ is $f(t) = A\mathrm{e}^{-bt/a} - \frac{c}{b}$. Also, $y' = -\lambda y + g$ has solution $y(t) = y(0)\mathrm{e}^{-\lambda t} + \mathrm{e}^{-\lambda t}\int_0^t \mathrm{e}^{\lambda s} g(s)\,\mathrm{d}s$.

and for $1 \leq i \leq m-1$,

$$\lambda_{ij} = \begin{cases} \lambda & \text{if } j = i+1, \\ i\mu & \text{if } j = i-1 \\ -(\lambda + i\mu) & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

In fact, to pass $i \to i+1$, a birth (call) must occur, and $\mathbb{P}(\text{no birth in time } t) = e^{-\lambda t}$ for any $i = 0, \ldots, m-1$, showing that $b_i = \lambda$ by (4.6). Similarly, a transition from i to $i-1$ occurs iff one of the i busy employees finishes an ongoing call. By independence, $\mathbb{P}(\text{none of the } i \text{ employees finishes the call}) = (e^{-\mu t})^i$. Thus, $d_i = i\mu$ for $i = 1, \ldots, m$. This gives the rates.

4.4.4 Compound Poisson Process

We conclude by mentioning very briefly another generalization of the Poisson process. This time we want to retain the two homogeneity features of the Poisson process : the fact that the sojourn parameter is independent of the state, and the fact that the transition from i to j in time t only depends on $|j-i|$, see (4.1). However, we now allow transitions to any states, not just $i \to i+1$.

Consider the state space $S = \mathbb{Z}$. Choose any $(\alpha_k)_{k\in\mathbb{Z}}$ such that $\alpha_k \geq 0$ for all k, $\sum_k \alpha_k = 1$ and $\alpha_0 < 1$. Fix $\lambda > 0$. Then the *compound Poisson process* is a regular jump process with transition rates

$$\lambda_{ij} = \begin{cases} \lambda\alpha_{j-i} & \text{if } j \neq i, \\ -\lambda(1-\alpha_0) & \text{if } j = i. \end{cases}$$

The sojourn parameter is thus $\lambda(1-\alpha_0)$ for all i. The ordinary Poisson process is the special case $\alpha_1 = 1$ and $\alpha_k = 0$ otherwise. See [7, Section XII.2] for details.

4.5 Limiting probabilities. Erlang's Formula

In this section we would like to understand the behavior of $p_{ij}(t)$ as $t \to +\infty$. The results are similar to the case of Markov chains, after establishing a few lemmas. As an application, we will find the limiting distributions of the call center with m employees, which is known as Erlang's formula, and is actually used in real-life.

As in Markov chains, we say that ε_j is *accessible from* ε_i if there is $s > 0$ such that $p_{ij}(s) > 0$. This s may depend on i, j.

Lemma 4.13. *If $\lambda_{ij} \neq 0$ then ε_j is accessible from ε_i.*

Moreover, accessibility is a transitive relation.

Proof. Assume on the contrary that $p_{ij}(s) = 0$ for all s. Then $0 = p'_{ij}(0) = \lambda_{ij}$, contradicting the hypothesis.

On the other hand, if ε_j is accessible from ε_i and ε_k is accessible from ε_j then there are s_1, s_2 such that $p_{ij}(s_1) > 0$ and $p_{jk}(s_2) > 0$, so by Kolmogorov-Chapman, $p_{ik}(s_1 + s_2) \geq p_{ij}(s_1)p_{jk}(s_2) > 0$. □

We now have the following :

Lemma 4.14. *Let (X_t) be a regular jump process. If $p_{ij}(t_0) > 0$ for some t_0, then $p_{ij}(t) > 0$ for all $t \geq t_0$.*

In particular, if X_t has finitely many states, each accessible from any other state, then $p_{ij}(t) > 0$ for all i, j and $t > 0$.

Proof. We have $p_{ii}(0) = 1$ and $p_{ij}(t)$ are continuous by Lemma 4.1, so $p_{ii}(h) > 0$ for small h. Using Kolmogorov-Chapman, we deduce that $p_{ii}(t) \geq (p_{ii}(t/n))^n > 0$, by taking n large enough. Next, suppose $p_{ij}(t_0) > 0$. By Kolmogorov-Chapman again, we find that $p_{ij}(t_0 + s) \geq p_{ij}(t_0)p_{jj}(s) > 0$ using the first part. This proves the first claim.

Now suppose we have m states, fix i, j and $t > 0$. We must find $t_0 \leq t$ such that $p_{ij}(t_0) > 0$. Since j is accessible from i, we know there is some t_{ij} such that $p_{ij}(t_{ij}) > 0$. Now choose n such that $n \geq \frac{mt_{ij}}{t}$ and consider the Markov chain with the same states $\varepsilon_1, \ldots, \varepsilon_m$, with transition probability $p_{ij}^{(n)} = p_{ij}(\frac{t_{ij}}{n})$. Then $p_{ij}^{(n)}(r) = \sum_{k_1,\ldots,k_{r-1}} p_{ik_1}^{(n)} p_{k_1k_2}^{(n)} \cdots p_{k_{r-1}j}^{(n)} = p_{ij}(\frac{rt_{ij}}{n})$. But $p_{ij}(t_{ij}) > 0$, so $p_{ij}^{(n)}(n) > 0$ and hence ε_j is accessible from ε_i in the chain. But if there is a road from i to j in the chain, and if the total number of states is m, then we can also reach j from i in a number of steps $k \leq m$. Thus, $p_{ij}^{(n)}(k) > 0$ for some $k \leq m$. It follows that $p_{ij}(\frac{kt_{ij}}{n}) > 0$. Since $\frac{kt_{ij}}{n} \leq \frac{mt_{ij}}{n} \leq t$, the proof is complete. □

We are finally ready to give the main result of this section. Recall definitions (4.1), (4.2) and (4.3).

Theorem 4.15. *Let X_t be a regular jump process with finitely many states $\varepsilon_1, \ldots, \varepsilon_m$, each accessible from any other state. Then for any state ε_j, there exists π_j such that*

$$\lim_{t\to\infty} p_j(t) = \pi_j \qquad \text{and} \qquad \lim_{t\to\infty} p_{ij}(t) = \pi_j$$

for any i and p_i^0. Moreover,

(i) *the speed of convergence is exponentially fast,*

(ii) $\pi_j \geq 0$ *and* $\sum_{j=1}^m \pi_j = 1$,

(iii) $\pi_j = \sum_{i=1}^m \pi_i p_{ij}(t)$,

(iv) $\sum_{i=1}^m \pi_i \lambda_{ij} = 0$.

Note that in contrast to the Markov chains version, we have not assumed the existence of N such that $p_{ij}(N) \geq \delta > 0$ for all i, j. This is because this condition is automatically satisfied here. In fact, Lemma 4.14 implies that for any $t > 0$ there is $\delta_t > 0$ such that $p_{ij}(t) \geq \delta_t > 0$ for any $i, j = 1, \ldots, m$.

Proof. As in the case of Markov chains, we define

$$r_j(t) = \min_i p_{ij}(t) \qquad \text{and} \qquad R_j(t) = \max_i p_{ij}(t)$$

and use Kolmogorov-Chapman to show that $r_j(t)$ increases with t while $R_j(t)$ decreases. Moreover, we know by Lemma 4.14 that there is $\delta > 0$ such that $\min_{i,j} p_{ij}(1) \geq \delta$. We easily see that $R_j(1) - r_j(1) \leq 1 - \delta$, and we show as before that $r_j(t+1) \geq r_j(t)(1-\delta) + p_{jj}(2t)$ while $R_j(t+1) \leq (1-\delta)R_j(t) + \delta p_{jj}(2t)$, so that $R_j(t+1) - r_j(t+1) \leq (1-\delta)(R_j(t) - r_j(t))$. It follows that $R_j(t) - r_j(t) \leq (1-\delta)^{\lfloor t \rfloor} \leq (1-\delta)^{t-1}$. Hence, $r_j(t)$ and $R_j(t)$ converge to the same limit π_j. We deduce the above results up to (iii) as before using (4.4) and (4.5). Finally (iv) follows from (iii) by differentiation since $p'_{ij}(0) = \lambda_{ij}$. □

Example 4.16. Consider the call center with m employees. Using Lemma 4.13, (4.8) and (4.9), we see that every state is accessible from any other state. We may apply Theorem 4.15 to deduce the existence of limiting probabilities π_j. To calculate them, we solve the system of equations $\sum_{i=0}^m \pi_i \lambda_{ij} = 0$. In view of (4.8) and (4.9), this reduces to

$$\begin{aligned} -\lambda\pi_0 + \mu\pi_1 &= 0\,, \\ \lambda\pi_{j-1} - (\lambda + j\mu)\pi_j + (j+1)\mu\pi_{j+1} &= 0\,, \qquad j = 1, \ldots, m-1\,, \\ \lambda\pi_{m-1} - m\mu\pi_m &= 0\,. \end{aligned}$$

Solving this system yields $\pi_j = \frac{1}{j!}(\frac{\lambda}{\mu})^j \pi_0$. Since $\sum_{j=0}^m \pi_j = 1$, we get

$$\pi_j = \frac{\frac{1}{j!}(\frac{\lambda}{\mu})^j}{\sum_{j=0}^m \frac{1}{j!}(\frac{\lambda}{\mu})^j}\,, \qquad j = 0, \ldots, m \tag{4.1}$$

which is known as *Erlang's formula.*

Suppose $S = \{\varepsilon_1, \varepsilon_2, \ldots\}$ and $\{\mu_k\}_{k\geq 1}$ is a sequence such that $\mu_k \geq 0$ and $\sum_{k\geq 1} \mu_k = 1$. Then we say that (μ_k) is a *stationary distribution* if for all $t > 0$, the row vector $\mu = (\mu_1, \mu_2, \ldots)$ satisfies the matrix equation

$$\mu P(t) = \mu\,.$$

In this case, we call the probability measure $\mu = \sum_k \mu_k \delta_{\varepsilon_k}$ an *invariant measure* for the jump process. The reason for the name "invariant measure" is the above invariance relation which says that the measure is in some sense preserved by $P(t)$. The reason for "stationary distribution" is that if we put $p_j^0 := \mu_j$ as initial distribution, we get $p_j(t) = p_j^0$ for all $t > 0$. In fact, using (4.4), $p_j(t) = \sum_k p_k^0 p_{kj}(t) = \sum_k \mu_k p_{kj}(t) = \mu_j = p_j^0$.

Theorem 4.15 says that $\{\pi_j\}$ is an invariant measure for the jump process. [12]

Remark 4.17. To conclude this chapter, note that compared to Markov chains, we have not discussed here the classification of states. This can be done by considering an *embedded Markov chain*. If τ_n is the sequence of jump times, define $X_n := X_{\tau_n}$, $n = 0, 1, 2, \ldots$. If $\tau_n = \infty$ for some n, we let $X_\infty := c$, where c is an arbitrary element not in S. The strong Markov property implies that such (X_n) is a Markov chain. We then say that a state ε_i if recurrent for (X_t) iff it is recurrent for (X_n) in the sense of the previous chapter, otherwise we say it is transient. We can also define positive recurrence for (X_t) by $\mathbb{E}_i(T_i) < \infty$, where T_i is the first time the continuous process returns to ε_i (here we don't need the embedded chain). We refer to [4, Chapter 8.5] for details.

4.6 Exercises

1. Suppose each alpha particle emitted by a sample of radium has probability p of being recorded by a Geiger counter. What is the probability of exactly n particles being recorded in t seconds ?
2. A man has two telephones on his desk, one receiving calls with density λ_1, the other with density λ_2. What is the probability of exactly n calls being received in t seconds ?
3. Given a Poisson process with density λ, let $X(t)$ be the number of events occurring in time t. Find the correlation coefficient of the random variables $X(t)$ and $X(t+s)$, where $s > 0$.
4. Show that (4.7) leads to Erlang's formula (4.1) for $m = 1$.
5. The arrival of customers at the complaint desk of a department store is described by a Poisson process with density λ. Suppose each clerk takes a random time τ to handle a complaint, where τ has an exponential distribution with parameter μ, and suppose a customer leaves whenever he finds

12. Exercises 1-10 are copied from Rozanov [13, Ch.8].

all the clerks busy. How many clerks are needed to make the probability of customers leaving unserved less than 0.015 if $\lambda = \mu$?

6. A single repairman services m automatic machines, which normally do not require his attention. Each machine, has probability $\lambda\Delta t + o(\Delta t)$ of breaking down in a small time interval Δt. The time required to repair each machine is exponentially distributed with parameter μ. Find the limiting probability of exactly j machines being out of order.

7. In the preceding problem, find the average number of machines awaiting the repairman's attention.

8. Solve Problem 6 for the case of r repairmen, where $1 < r < m$.

9. An electric power line serves m identical machines, each operating independently of the others. Suppose that in a small interval of time Δt each machine has probability $\lambda\Delta t + o(\Delta t)$ of being turned on and probability $\mu\Delta t + o(\Delta t)$ of being turned off. Find the limiting probability π_j of exactly j machines being on.

10. Show that the answer to the preceding problem is just what one would expect by an elementary argument if $\lambda = \mu$.

11. Let (X_t) be a continuous-time Markov process. Let $0 < t_1 < t_2 < \cdots < t_k$. Show that

$$\mathbb{P}^i(X_{t_1} = \varepsilon_{j_1}, \dots X_{t_k} = \varepsilon_{j_k}) = p_{ij_1}(t_1)p_{j_1j_2}(t_2 - t_1)\cdots p_{j_{k-1}j_k}(t_k - t_{k-1})\,.$$

12. Let (X_t) be a regular jump process with jump times τ_n. Use the strong Markov property to show that

$$\begin{aligned}\mathbb{P}^i(X_{\tau_1} = \varepsilon_{i_1}, \dots, X_{\tau_n} = \varepsilon_{i_n}, \tau_1 - \tau_0 > a_1, \dots, \tau_n - \tau_{n-1} > a_n)& \\ = \mathrm{e}^{-\lambda_i a_1}\rho_{ii_1}\mathrm{e}^{-\lambda_{i_1}a_2}\rho_{i_1i_2}\cdots\mathrm{e}^{-\lambda_{i_{n-1}}a_n}\rho_{i_{n-1}i_n}\,,&\end{aligned}$$

where $\rho_{ij} = \mathbb{P}^i(X_\tau = \varepsilon_j)$ and λ_k are the sojourn parameters in ε_k.

13. Let (X_t) be a pure birth process on $\mathbb{N}$ with birth rates $b_i \geq 0$. Show that

$$p_{ii}(t) = \mathrm{e}^{-b_it}\,, \qquad p_{ij}(t) = b_{j-1}\mathrm{e}^{-b_jt}\int_0^t \mathrm{e}^{b_js}p_{i,j-1}(s)\,\mathrm{d}s \qquad j > i$$

Deduce $p_{0n}(t)$ recursively by calculating $p_{i,i+1}(t)$. You will have to analyze two cases : when $b_{i+1} = b_i$ and when $b_{i+1} \neq b_i$.

(Hint : The solution of $y' = -\lambda y + g$ is $y(t) = y(0)\mathrm{e}^{-\lambda t} + \mathrm{e}^{-\lambda t}\int_0^t \mathrm{e}^{\lambda s}g(s)\,\mathrm{d}s$.)

14. Show that $\mu = (\mu_k)$ is a stationary distribution for a regular jump process iff $\mu_k \geq 0$, $\sum_k \mu_k = 1$ and, as a matrix product,

$$\mu\Lambda = 0\,.$$

In other words $\sum_i \mu_i \lambda_{ij} = 0$ for any j.

15. Let (X_t) be a birth/death process with birth rates $b_i \geq 0$ and death rates $d_i > 0$. Show that the process has a stationary distribution μ iff

$$C = \sum_{j\geq 1} \frac{b_0 \cdots b_{j-1}}{d_1 \cdots d_j} < \infty\,,$$

in which case

$$\mu_0 = \frac{1}{C+1} \qquad \text{and} \qquad \mu_j = \frac{b_0 \cdots b_{j-1}}{d_1 \cdots d_j}\mu_0\,.$$

(Hint : re-write the system of equations as $d_{j+1}\mu_{j+1} - b_j\mu_j = d_j\mu_j - b_{j-1}\mu_{j-1}$*).*

16. Show that a pure birth process with birth rates $b_i > 0$ for all i has no stationary distribution.

17. Show that if (X_t) is a pure birth process and $b_{i_0} = 0$ for some i_0, then (X_t) has a stationary distribution.

18. Consider a call center with infinitely many employees. Assume that the inter-arrival times have rate λ, and that the rate of service per employee is μ. Let X_t be the number of customers in the queue. Show that this process has a stationary distribution given by $\nu_j = \frac{\lambda^j}{\mu^j j!}\mathrm{e}^{-\lambda/\mu}$.

Chapter 5

Additional Topics

In this last chapter we briefly mention several important processes not discussed in the main course. The sections can be read independently. There are books devoted to each of these sections, our aim here is just to give a glimpse of the topic so the student vaguely knows what it is about if (s)he encounters the concept later on.

5.1 Branching processes

Consider a group of particles, each randomly producing particles of the same type. We assume

(i) the probability that a given particle transforms into k particles after time t has passed is $p_k(t)$, $k = 0, 1, 2, \ldots$, and $p_k(t)$ is the same for all particles,

(ii) the different particles behave independently of one another.

Such a process is called a *branching process*. Examples include nuclear chain reactions, survival of populations, etc. Note that $p_0(t)$ is the probability that one particle transforms into 0 particle in time t, i.e. that the particle disappears.

Let X_t be the total number of particles at time t. Then X_t is a regular jump process. In fact, $p_{ij}(t) = \mathbb{P}(X_{s+t} = j \mid X_s = i)$ is the probability that a population of i particles produces $j - i$ particles in the time interval $(s, t + s]$, and is clearly the same as $\mathbb{P}(X_t = j \mid X_0 = i)$, since condition (i) above only require a time t to pass, not necessarily the time from 0 to t. This shows time homogeneity. The Markov property also intuitively holds : to pass from i to j particles, it doesn't matter how the system reached the state i, it is only important to find the probability that each particle produces the right amount of particles so the total becomes j. Right continuity here is similar to the case of the Poisson process.

This process has a finite number of jumps (particle production) in finite time intervals (this is clear in concrete populations or physical problems, and should be assumed otherwise).

Suppose $X_0 = k$, i.e. we initially have k particles. Then we may express $X_t = X_t^{(1)} + \cdots + X_t^{(k)}$, where $X_t^{(j)}$ is the number of particles produced by the j-th particle after time t has passed. By (i) and (ii), we know the $X_t^{(j)}$ are i.i.d., with $\mathbb{P}(X_t^{(j)} = n) = p_n(t)$. Note that in terms of transition probabilities, we have $p_n(t) = p_{1,n}(t)$, the probability that one particle produces n particles in time t. Being a regular jump process, we thus have $p_n(h) = p_{1,n}(h) = \lambda_{1,n}h + o(h)$ for $n \neq 1$ and $p_1(h) = p_{1,1}(h) = 1 + \lambda_{1,1}h + o(h)$ for small h. We henceforth use the shorthand notation $\lambda_n := \lambda_{1,n}$ for $n \neq 1$ and $\lambda := -\lambda_{1,1}$.

If $p_{k,n}(t)$ is the probability that k particles produce n particles in time t, then the backward Kolmogorov equation takes the form $p_n'(t) = \sum_{k=0}^{\infty} \lambda_k p_{k,n}(t)$.

An important question in branching processes is the *extinction probability*. If we originally have a single particle, then the probability that all particles disappear in time t is $p_0(t)$. If we originally have k particles, then by independence this probability becomes $(p_0(t))^k$. In concrete problems we can often find the transition rates λ_k and the question is to deduce $p_0(t)$. This can be done by applying the above Kolmogorov equation to the generating function $F(t,z) = \sum_{n=0}^{\infty} p_n(t)z^n$. It turns out that the extinction probability is intimately related to the function $f(x) = \sum_{k=0}^{\infty} \lambda_k x^k$. Using the Kolmogorov equations, one can show that $\lim_{t\to\infty} p_0(t) = \alpha$, where α is the smaller root of the equation $f(x) = 0$. This is the probability that the particles eventually die out. See [13, Appendix 3] for details. On the contrary, it may happen that the population explodes and we have infinitely many particles as $t \to \infty$. This probability can also be computed by analyzing $f(x)$, see [13].

In this section we discussed branching processes in continuous time. We can also study branching processes in discrete time. In this case we have a Markov chain instead. See [5, p.96-100] for analogous questions.

5.2 Martingales

Consider a stochastic process $X_1, X_2, \ldots$ in discrete time.

As the time n increases, so does our knowledge about what happened in the past. This is modeled by *filtrations*. A filtration on Ω is simply a sequence of σ-algebras $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$. Here $\mathcal{F}_n$ represents our knowledge at time n, which increases as time passes.

We say that $X_1, X_2, \ldots$ is *adapted* to $\mathcal{F}_1, \mathcal{F}_2, \ldots$ if X_n is $\mathcal{F}_n$-measurable. As the $\mathcal{F}_n$ are increasing, X_k are also $\mathcal{F}_n$-measurable for $k < n$. This implies that $\mathcal{F}_n$ contains all that can be learned from $X_1, \ldots, X_n$. We can take for example $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ to have the X_j adapted to $\mathcal{F}_j$, and this is the smallest possible choice of an adapted filtration (in general $\mathcal{F}_n$ contains even more info).

Martingales first appeared in the context of gambling. They have important applications, for example in financial mathematics. We say that $X_1, X_2, \ldots$ is a martingale with respect to a filtration $\mathcal{F}_1, \mathcal{F}_2, \ldots$ if

(i) X_n is integrable for each $n = 1, 2, \ldots,$

(ii) $X_1, X_2, \ldots$ is adapted to $\mathcal{F}_1, \mathcal{F}_2, \ldots,$

(iii) $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$ a.s. for each $n = 1, 2, \ldots.$

For example, if a person plays a sequence of games of chance, let Y_n be the winning/losses in game n. Then the total fortune after n games is $X_n = Y_1 + \cdots + Y_n$. Take $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$. If n games have been played so far, the total knowledge of the player is $\mathcal{F}_n$. The game is *fair* if $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$. This means that we expect our fortune at step $n+1$ to be on average the same as our fortune at time n. We see that this corresponds to a martingale. It can happen that the game is *favorable* to the player. In that case $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \geq X_n$ and we speak of a *submartingale*. It can also happen that the game is *unfavorable* to the player. In this case $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq X_n$ and we speak of a *supermartingale*.

If we regard the Y_n to be the winning/losses per unit stake (say for betting one dollar), then the player can try to improve his winnings by using a *gambling strategy*. This is done by varying the stakes. The player decides to bet a stake α_n at game n based on the outcomes of the first $n-1$ games. We should hence assume α_n is $\mathcal{F}_{n-1}$-measurable. We thus define a gambling strategy with respect to a filtration $\mathcal{F}_1, \mathcal{F}_2, \ldots$ to be a sequence $\alpha_1, \alpha_2, \ldots$ of random variables such that α_n is $\mathcal{F}_{n-1}$ measurable, where $\mathcal{F}_0 := \{\emptyset, \Omega\}$.

By following a strategy, the total earnings at time n now become $X_n = \alpha_1 Y_1 + \cdots + \alpha_n Y_n$. An interesting result says that a fair game will always remain fair, no matter what gambling strategy is used. Similarly, it is impossible to turn an unfavorable game into a favorable one by using gambling strategies $\alpha_n \geq 0$. So the player can never beat the casino.

Lastly, we mention the important concept of *stopping times*. While playing the game of chance, the player usually has the option to quit at any time. This can be deterministic if the player decides from the start to play a fixed number of games, for example five. Then $\tau = 5$. But usually this time is random, based on his knowledge at time n. We thus define a stopping time with respect

to a filtration $\mathcal{F}_1, \mathcal{F}_2, \ldots$ to be a random variable with values $\{1, 2, \ldots\} \cup \{\infty\}$, such that the event $\{\tau = n\} \in \mathcal{F}_n$ for each $n = 1, 2, \ldots$. This means that our knowledge $\mathcal{F}_n$ at time n tells us whether we should quit after n games, i.e. $\tau = n$.

For more on martingales, the student can read for example [5, Chapters 3,4].

Stopping times appear very often in modern probability theory. They include the *first hitting time*, defined as $\tau_B = \min\{n \geq 0 : X_n \in B\}$ for $B \in \mathcal{F}$. This is the first time the random variable hits B. In the context of Markov chains, we also define the *return time to* ε_i by $\tau_i = \min\{n \geq 1 : X_n = \varepsilon_i\}$, where it is assumed $X_0 = \varepsilon_i$. This is also a stopping time. Finally, we can also define stopping times for continuous processes. Here we require that $\{\tau \leq t\} \in \sigma(X_s, s \leq t)$. Each jump time τ_n is a stopping time for a regular jump process (X_t).

5.3 Brownian Motion

Brownian motion [1] is a phenomenon discovered by the botanist Robert Brown in 1827. Brown found while looking through a microscope that the pollen of a certain plant, when immersed in water, moves quite erratically. The rough path followed by the pollen was explained as being a consequence of the frequent collisions with water molecules. See Figure 5.1. Such a motion was subsequently studied mathematically by Wiener in the nineties as a random process.

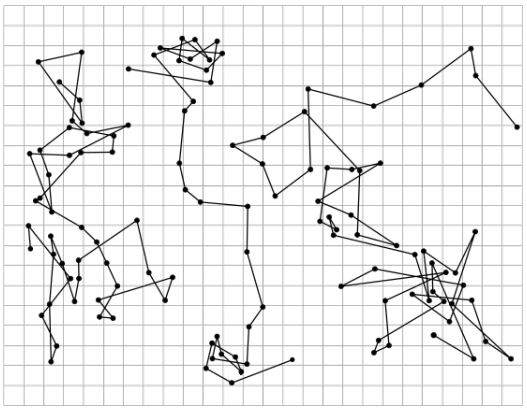


Figure 5.1 – Motion of particles under the microscope, from the book of Jean Perrin, *les atomes* - courtesy of wikipedia.

For simplicity we only consider dimension $d = 1$ here. We start by considering a random walk as follows.

(1) The particle moves only along the x axis.

1. We follow here [13] and [8].

(2) The particle moves only at the times $t = n\Delta t$, $n = 0, 1, 2, \ldots$.

(3) If the particle is at position x at time t, then the particle moves either to $x + \Delta x$ or $x - \Delta x$ with equal probability $\frac{1}{2}$, at time $t + \Delta t$.

Let X_t be the position of the particle at time t. We assume $X_0 = 0$.

We now ask : what is the distribution of X_t as $\Delta t \to 0$ and $\Delta x \to 0$? This means that instead of hopping, the particle will move continuous with time.

We notice that for $t = n\Delta t$, we have

$$X_t = \Delta x(Y_1 + \cdots + Y_n)\,,$$

where

$$Y_k = \begin{cases} 1 & \text{if the k-th jump is to the right,} \\ -1 & \text{if the k-th jump is to the left.} \end{cases}$$

As the successive jumps are independent, the random variables Y_k are independent and have the same distribution $\frac{1}{2}(\delta_{-1} + \delta_1)$. We have

$$\mathbb{E}(Y_k) = (-1)\,\mathbb{P}(Y_k = -1) + 1\,\mathbb{P}(Y_k = 1) = \frac{-1}{2} + \frac{1}{2} = 0\,,$$

also

$$\mathrm{Var}(Y_k) = \mathbb{E}(Y_k^2) - (\mathbb{E}(Y_k))^2 = (-1)^2\,\mathbb{P}(Y_k = -1) + (1)^2\,\mathbb{P}(Y_k = 1) - 0 = 1$$

for any k. As the random variables are independent, we get

$$\mathbb{E}(X_t) = 0 \qquad \text{and} \qquad \mathrm{Var}(X_t) = (\Delta x)^2 n = (\Delta x)^2 (t/\Delta t)\,.$$

We now consider the limiting process as $\Delta x \to 0$ and $\Delta t \to 0$. We should choose the speed of convergence wisely. For example, if we take $\Delta x = \Delta t \to 0$, we get that $\mathbb{E}(X_t) \to 0$ and $\mathrm{Var}(X_t) \to 0$. This implies the limiting process has zero mean and variance and is thus zero almost everywhere. So we take instead $\Delta x = \sqrt{\Delta t} \to 0$. Then $\mathbb{E}(X_t) \to 0$ and $\mathrm{Var}(X_t) \to t$. We denote, a bit abusively, the limiting process also by X_t. Then

Proposition 5.1. *The limiting process X_t has a normal distribution $N(0, \sqrt{t})$.*

Proof. We may regard the walk as a sequence of Bernoulli trials where moving to the right means a success and moving to the left means a failure. Let $t = n\Delta t$. If S_n is the number of successes in the first n steps, then we recall the quantity S_n^* appearing in the De Moivre-Laplace central limit theorem :

$$S_n^* = \frac{S_n - np}{\sqrt{npq}} = \frac{S_n - \frac{n}{2}}{\sqrt{n\frac{1}{4}}} = \frac{2S_n - n}{\sqrt{n}}\,.$$

But to reach X_t, we take S_n steps to the right and $n - S_n$ to the left. Thus,

$$X_t = S_n\Delta x - (n - S_n)\Delta x = (2S_n - n)\Delta x = S_n^*\sqrt{n}\Delta x = S_n^*\sqrt{t/\Delta t}\Delta x = S_n^*\sqrt{t}\,.$$

Taking $n \to \infty$, which means $\Delta t \to 0$, we get that X_t converges in distribution to $N(0, \sqrt{t})$, by the De Moivre-Laplace central limit theorem. □

It is this limiting process which we call *Brownian motion*. The preceding discussion implies that it satisfies the following properties :

(a) (X_t) has *independent increments*, i.e. if $0 < t_1 < \cdots < t_n$ then $X_{t_1}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_n-1}$ are independent.

To see this, note that if $t_k = n_k\Delta t$, then $X_{t_k} - X_{t_{k-1}} = (\Delta x) \sum_{i=n_{k-1}+1}^{n_k} Y_i$. Since all Y_i are independent, the finite sums $X_{t_1}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{n-1}$ are independent. Such independence is retained in the limit.

(b) (X_t) has *stationary increments*, i.e. $X_{t+s} - X_t$ has the same distribution as X_s.

To see this, note that if $t = n\Delta t$ and $s = m\Delta t$, then $X_{t+s} - X_t = \sum_{i=n+1}^{n+m} Y_i$ is a sum of m random variables which have the same distribution as $\sum_{i=1}^{m} Y_i$ since all Y_i are identically distributed.

(c) The trajectory followed by X_t appears to be non-differentiable. Indeed, since we take $\Delta x = \sqrt{\Delta t}$, then $\frac{\Delta x}{\Delta t} = \frac{1}{\sqrt{\Delta t}} \to \infty$, showing the trajectory is not differentiable. See Figure 5.2.

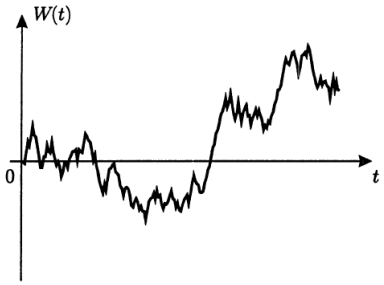


Figure 5.2 – A typical Brownian motion path. From [5].

The preceding results motivate the abstract definition of a Brownian motion to be a stochastic process $(X_t)_{t\geq 0}$ satisfying the following properties : $X_0 = 0$, (X_t) has independent and stationary increments, for any $t > 0$, X_t has distribution $N(0, \sqrt{t})$. With this abstract definition we can forget the random walk and derive a very rich theory solely relying on these assumptions. We refer the student to [5] and [8] for some basic results, the deeper theory is taught in graduate school.

Bibliography

[1] W. Anderson. *Continuous-Time Markov Chains. An Applications-Oriented Approach*. Springer, 1991.

[2] R. Bhattacharya and E.C. Waymire. *A Basic Course in Probability Theory*. Universitext. Springer, second edition, 2016.

[3] P. Billingsley. *Probability and measure*. Wiley, third edition, 1995.

[4] P. Brémaud. *Markov Chains, Gibbs Fields, Monte Carlo Simulation and Queues*. Springer, 1999.

[5] Z. Brzeźniak and T. Zastawniak. *Basic Stochastic Processes. A Course Through Exercises*. Springer, 2005.

[6] R. M. Dudley. *Real Analysis and Probability*. CUP, 2004.

[7] W. Feller. *An Introduction to Probability Theory and its Applications. Volume I*. Wiley, third edition, 1968.

[8] D. Foata and A. Fuchs. *Processus Stochastiques. Processus de poisson, chaîne de Markov et martingales*. Dunod, 2004.

[9] O. Kallenberg. *Foundations of Modern Probability*. Springer, second edition, 2002.

[10] A. Klenke. *Probability theory. A comprehensive course*. Universitext. Springer, London, second edition, 2014.

[11] S. Lalley. *Continuous-Time Markov Chains. Lecture notes.*

[12] J.S. Rosenthal. *A First Look at Rigorous Probability Theory*. World Scientific, second edition, 2006.

[13] Y. A. Rozanov. *Probability theory: A concise course*. Dover Publications, Inc., New York, english edition, 1977. Translated from the Russian and edited by Richard A. Silverman.

[14] Y. A. Rozanov. *Probability Theory, Random Processes and Mathematical Statistics*. Mathematics and Its Applications. Springer, 1995.

[15] M. Sabri. *Probability Theory. Stat 201. Lecture notes.*

[16] F. Spitzer. *Principles of Random Walk*. Graduate Texts in Mathematics. Springer, second edition, 2001.

[17] G. Teschl. *Mathematical methods in quantum mechanics*, volume 157 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2014. With applications to Schrödinger operators.