

Stochastic Processes

Lasse Leskelä
Aalto University

December 1, 2020

Contents

1	Markov chains and stochastic models	5
1.1	Markov property	5
1.2	Transition matrix and transition diagram	6
1.3	Transient distributions	11
1.4	Many-step transition probabilities	12
1.5	Path probabilities	13
1.6	Occupancy of states	14
1.7	Simulation of Markov chains	15
2	Markov chains in the long run	18
2.1	Invariant and limiting distributions	18
2.2	Connectivity	21
2.3	Invariant distribution of an irreducible chain	22
2.4	Periodicity	23
2.5	Invariant distribution of an irreducible aperiodic chain	24
3	Markov additive processes	25
3.1	Definitions	25
3.2	Behaviour in finite time horizon	26
3.3	Ergodicity	28
3.4	Long-term behaviour	30
3.5	Remarks	31
4	Passage times and hitting probabilities	32
4.1	Passage times	32
4.2	Hitting probabilities	36
4.3	Gambler's ruin	39
5	General Markov chains and random walks	42
5.1	Infinite vectors and matrices	42
5.2	Markov chains	43
5.3	Long-term behaviour	44
5.4	Convergence theorem	46
5.5	Reversibility	48
5.6	Random walk on the nonnegative integers	49

6	Branching processes	53
6.1	Transition matrix	53
6.2	Generating functions	54
6.3	Expected population size	55
6.4	Extinction probability	56
6.5	Sure extinction	59
7	Random point patterns and counting processes	60
7.1	Random point pattern	60
7.2	Counting measure and counting process	60
7.3	Independent scattering	61
7.4	Poisson process	65
7.5	Constructing independently scattered point patterns	65
8	Poisson processes and renewal processes	68
8.1	Poisson process defined as a stochastic process	68
8.2	Superposed Poisson processes	69
8.3	Compound Poisson process	70
8.4	Thinned Poisson process	72
8.5	Renewal processes	74
9	Continuous-time Markov chains in finite time horizon	79
9.1	Markov property	79
9.2	Transition matrices	80
9.3	Generator matrix	83
9.4	Transition semigroup generators	86
10	Analysis of Markov jump processes	90
10.1	Jump rates and jump probabilities	90
10.2	Determining the generator matrix	91
10.3	Memoryless races	93
10.4	Constructing Markov chain models	95
10.5	Invariant distributions	97
10.6	Convergence	98
11	Martingales and information processes	100
11.1	Conditional expectation with respect to information	100
11.1.1	Definition for finite-state random variables	100
11.1.2	Rules	102
11.1.3	General definition	104
11.2	Martingales	105
11.3	Properties of martingales	108
11.4	Long-term behavior of martingales	109
11.4.1	Martingales and Markov chains	110

12 Stopped martingales and optional times	113
12.1 Gambling with unit bets	113
12.2 Doubling strategy	114
12.3 Adaptive betting	115
12.4 Optional times	118
12.5 Stopped martingales	120
12.6 Optional stopping theorem	121
A Suomi–English dictionary	124

Prologue

These lecture notes contain material for MS-C2111 Stochastic Processes at Aalto University, 2018–2020. The lectures notes have been translated from a corresponding Finnish version, originally written in 2015. Warmest thanks go to Kalle Kytölä, Aleksi Karrila, Joonas Karjalainen, Hoa Ngo, Jarno Ruokokoski, Olli Huopio, Maryam Kiashemshaki, Veli Kuuranne, Joonas Juvonen, Akseli Mäkinen, Vili Nieminen, Martti Ranta, Erkkä Tahvanainen, Emmi Vaara, and Juri Voloskin for their corrections and helpful comments for improving the text. Especially, Aleksi Karrila has kindly written several TikZ codes for transition diagram plots. The notes will be updated frequently during autumn 2020. All comments and suggestions are most welcome.

Chapter 1

Markov chains and stochastic models

1.1 Markov property

A finite-state Markov chain is a random process which moves from state x to state y with probability $P(x, y)$, independently of its past states. The *state space* (*tilajoukko*) is denoted by S , and the collection of transition probabilities $P = \{P(x, y) : x, y \in S\}$ is called the *transition matrix* (*siirtymämatriisi*). The transition matrix is a square matrix with rows and columns indexed by states $x, y \in S$. Being probabilities, the entries of the transition matrix satisfy

$$0 \leq P(x, y) \leq 1, \quad x, y \in S,$$

and because the chain certainly moves to some state, the row sums are equal to

$$\sum_{y \in S} P(x, y) = 1, \quad x \in S.$$

More precisely, an S -valued random sequence (X_0, X_1, X_2, \dots) defined on a probability space (Ω, \mathbb{P}) is a *Markov chain* (*Markov-ketju*) with state space S and transition matrix P if

$$\mathbb{P}(X_{t+1} = y \mid X_t = x, H_{t-}) = P(x, y) \quad (1.1)$$

for all $x, y \in S$, all $t \geq 0$, and all events $H_{t-} = \{X_0 = x_0, \dots, X_{t-1} = x_{t-1}\}$ such that $\mathbb{P}(X_t = x, H_{t-}) > 0$. The next state of a Markov chain depends on its past history only via its current state, and previous states do not have any statistical relevance when predicting the future. Equation (1.1) is named *Markov property* (*Markov-ominaisuus*) after a Russian mathematician Andrey Markov (1856–1922). The Markov property can be defined analogously also for random processes with continuous time parameter and infinite state spaces. The class of general Markov processes includes several important stochastic models such as Poisson processes, Brownian motions, which will be discussed later.

The following fundamental result tells that the past history H_{t-} may be ignored in formula (1.1). The proof can be skipped at a first reading.

Theorem 1.1. For any finite-state Markov chain with transition probability matrix P ,

$$\mathbb{P}(X_{t+1} = y | X_t = x) = P(x, y) \quad (1.2)$$

for any $t \geq 0$ and any $x, y \in S$ such that $\mathbb{P}(X_t = x) > 0$.

Proof. Let us denote the joint probability mass function of the random variables X_0, \dots, X_t as

$$f_t(x_0, \dots, x_{t-1}, x_t) = \mathbb{P}(X_0 = x_0, \dots, X_{t-1} = x_{t-1}, X_t = x_t).$$

Then the conditional probability of the event $X_{t+1} = y$ given $X_t = x$ and $H_{t-} = \{X_0 = x_0, \dots, X_{t-1} = x_{t-1}\}$ can be written as

$$\begin{aligned} \mathbb{P}(X_{t+1} = y | X_t = x, H_{t-}) &= \frac{\mathbb{P}(X_{t+1} = y, X_t = x, H_{t-})}{\mathbb{P}(X_t = x, H_{t-})} \\ &= \frac{f_{t+1}(x_0, \dots, x_{t-1}, x, y)}{f_t(x_0, \dots, x_{t-1}, x)}, \end{aligned}$$

and the Markov property (1.1) can be rephrased as

$$\frac{f_{t+1}(x_0, \dots, x_{t-1}, x, y)}{f_t(x_0, \dots, x_{t-1}, x)} = P(x, y).$$

By multiplying both sides of the above equation by $f_t(x_0, \dots, x_{t-1}, x)$, and then summing both sides over all possible past states, we find that

$$\sum_{x_0, \dots, x_{t-1} \in S} f_{t+1}(x_0, \dots, x_{t-1}, x, y) = \sum_{x_0, \dots, x_{t-1} \in S} f_t(x_0, \dots, x_{t-1}, x) P(x, y). \quad (1.3)$$

By the law of total probability, the left side of (1.3) equals $\mathbb{P}(X_t = x, X_{t+1} = y)$ and the right side equals $\mathbb{P}(X_t = x)P(x, y)$. Hence we see that

$$\mathbb{P}(X_t = x, X_{t+1} = y) = \mathbb{P}(X_t = x)P(x, y),$$

and the claim follows by dividing both sides above by $\mathbb{P}(X_t = x)$. \square

1.2 Transition matrix and transition diagram

The structure of a Markov chain is usually best illustrated by a transition diagram. The *transition diagram* (*siirtymäkaavio*) of a transition matrix P and a corresponding Markov chain is a directed graph with node set being the state space and link set consisting of ordered node pairs (x, y) such that $P(x, y) > 0$. The transition diagram is usually viewed as a weighted graph by setting the weight of a link to be the corresponding transition probability. Let us next investigate three examples which can be modeled using a Markov chain.

Example 1.2 (Weather model). The summer weather of day $t = 0, 1, \dots$ in Espoo can be modeled using a random sequence in state space $S = \{1, 2\}$, where state 1 = 'cloudy' and 2 = 'sunny'. It is assumed that a cloudy day is followed by a sunny day with probability $p = 0.2$, and that a sunny day is followed by a cloudy day with probability $q = 0.5$, independently of the past days. The state of the weather model can be represented as a Markov chain (X_0, X_1, \dots) with transition matrix

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}$$

and transition matrix described in Figure 1.1.

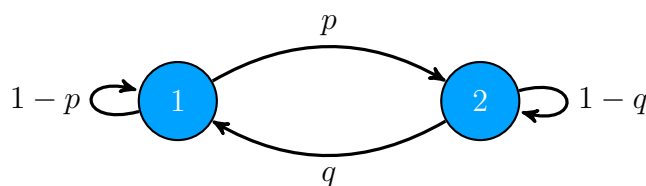


Figure 1.1: Transition diagram of the weather model.

Let us assume that Monday (day $t = 0$) is cloudy. Then the weather model predicts Tuesday to be cloudy with probability $1-p$ and sunny with probability p , so that

$$\mathbb{P}(X_1 = 1 | X_0 = 1) = 1-p \quad \text{ja} \quad \mathbb{P}(X_1 = 2 | X_0 = 1) = p.$$

The probability that it is cloudy also on Wednesday is obtained by conditioning on the possible states of Tuesday's weather according to

$$\begin{aligned} \mathbb{P}(X_2 = 1 | X_0 = 1) &= \mathbb{P}(X_1 = 1 | X_0 = 1)\mathbb{P}(X_2 = 1 | X_1 = 1, X_0 = 1) \\ &\quad + \mathbb{P}(X_1 = 2 | X_0 = 1)\mathbb{P}(X_2 = 1 | X_1 = 2, X_0 = 1) \\ &= (1-p)^2 + pq. \end{aligned}$$

Therefore, Wednesday is predicted to be a cloudy day with probability $(1-p)^2 + pq = 0.740$. ■

The following, more complicated example is typical in applications related to industrial engineering and management. More examples of similar kind are available for example in the book [Kul16].

Example 1.3 (Inventory model). Katiskakauppa.com Oyj sells laptops in a store which is open Mon–Sat during 10–18. The inventory is managed using the following policy. Every Saturday at 18:00 a sales clerk computes the number of laptops in stock. If this number is less than two, sufficiently many new laptops are ordered so that next Monday morning there will five laptops in stock. The demand for new laptops during a week is predicted to be Poisson distributed with mean $\lambda = 3.5$. Customers finding an empty stock at an instant of purchase

go to buy their laptops elsewhere. Develop a Markov chain to model the state of the inventory.

Let X_t be a random variable describing the number of laptops in stock on Monday 10:00 during week $t = 0, 1, \dots$. Denote by D_t a random variable modeling the demand of laptops during the corresponding week. Then the number of laptops in stock in the end of week t equals $\max(X_t - D_t, 0)$. If $X_t - D_t \geq 2$, then no laptops are ordered during the weekend and hence $X_{t+1} = X_t - D_t$. Otherwise a new order is placed and $X_{t+1} = 5$. Therefore

$$X_{t+1} = \begin{cases} X_t - D_t, & \text{if } X_t - D_t \geq 2, \\ 5, & \text{else.} \end{cases}$$

Hence the state space of the random process (X_0, X_1, \dots) is $S = \{2, 3, 4, 5\}$. If we assume that the demand for new laptops during a week is independent of the demands of other weeks, then it follows that (X_0, X_1, \dots) is a Markov chain.

Let us next determine the transition probabilities $P(i, j)$. Consider first the case $i = 2$ and $j = 2$ which corresponds to the event that the number of laptops in stock is 2 in the beginning and in the end of a week t . This event takes place if and only if the demand during week t equals $D_t = 0$. Because the demand during week t is independent of past demands (and hence also on the past inventory states), it follows that

$$\begin{aligned} P(2, 2) &= \mathbb{P}(X_{t+1} = 2 \mid X_t = 2, H_{t-}) \\ &= \mathbb{P}(D_t = 0 \mid X_t = 2, H_{t-}) \\ &= \mathbb{P}(D_t = 0) \\ &= e^{-\lambda} \end{aligned}$$

for all events $H_{t-} = \{X_0 = x_0, \dots, X_{t-1} = x_{t-1}\}$. Indeed, a transition from any state i to a state $j \in \{2, 3, 4\}$ corresponds to an event $D_t = i - j$, and hence

$$\begin{aligned} P(i, j) &= \mathbb{P}(X_{t+1} = j \mid X_t = i, X_{t-1}, \dots, X_0) \\ &= \mathbb{P}(X_t - D_t = j \mid X_t = i, X_{t-1}, \dots, X_0) \\ &= \mathbb{P}(i - D_t = j \mid X_t = i, X_{t-1}, \dots, X_0) \\ &= \mathbb{P}(D_t = i - j) \end{aligned}$$

for all $i \in \{2, 3, 4, 5\}$ and $j \in \{2, 3, 4\}$. Because D_t is $\text{Poi}(\lambda)$ -distributed, we know that

$$\mathbb{P}(D_t = k) = \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!}, & k \geq 0, \\ 0, & k < 0. \end{cases} \quad (1.4)$$

From these formulas we can compute the transition probabilities $P(i, j)$ for columns $j = 2, 3, 4$. Let us next determine the entries for $j = 5$. If $i \in \{2, 3, 4\}$, such a transition corresponds to replenishing the stock by ordering new laptops,

that is, $X_t - D_t \leq 1$. Hence

$$\begin{aligned}
 P(i, 5) &= \mathbb{P}(X_{t+1} = 5 \mid X_t = i, X_{t-1}, \dots, X_0) \\
 &= \mathbb{P}(X_t - D_t \leq 1 \mid X_t = i, X_{t-1}, \dots, X_0) \\
 &= \mathbb{P}(i - D_t \leq 1 \mid X_t = i, X_{t-1}, \dots, X_0) \\
 &= \mathbb{P}(D_t \geq i - 1)
 \end{aligned}$$

for all $i \in \{2, 3, 4\}$. Finally we need the value $P(5, 5)$. A transition from state $i = 5$ to state $j = 5$ occurs in two cases: either there is no demand during week t , or the demand is 4 or more. Therefore,

$$\begin{aligned}
 P(5, 5) &= \mathbb{P}(X_{t+1} = 5 \mid X_t = 5, X_{t-1}, \dots, X_0) \\
 &= \mathbb{P}(D_t = 0) + \mathbb{P}(D_t \geq 4).
 \end{aligned}$$

By computing the probabilities of D_t from the Poisson distribution (1.4), we may write the transition probability matrix as

$$P = \begin{bmatrix} 0.03 & 0 & 0 & 0.97 \\ 0.11 & 0.03 & 0 & 0.86 \\ 0.18 & 0.11 & 0.03 & 0.68 \\ 0.22 & 0.18 & 0.11 & 0.49 \end{bmatrix}.$$

Note that the rows and columns of P are indexed using the set $S = \{2, 3, 4, 5\}$. The corresponding transition diagram is plotted in Figure 1.2.

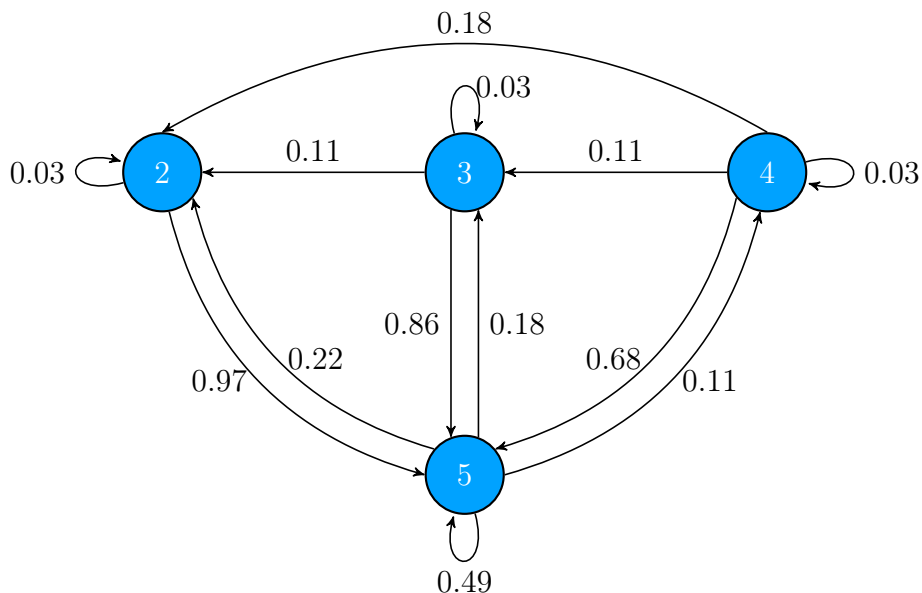


Figure 1.2: Transition diagram of the inventory model.

■

```

# R-code for computing the transition matrix
la <- 3.5
P <- matrix(0,4,4)
rownames(P) <- 2:5
colnames(P) <- 2:5
P[, "2"] <- dpois(0:3, la)
P[, "3"] <- dpois(0:3-1, la)
P[, "4"] <- dpois(0:3-2, la)
P["2", "5"] <- 1 - ppois(0, la)
P["3", "5"] <- 1 - ppois(1, la)
P["4", "5"] <- 1 - ppois(2, la)
P["5", "5"] <- dpois(0, la) + 1-ppois(3, la)

```

Markov chains encountered in applications in technology and science can have huge state spaces. The state space of the following example contains billions of nodes and grows all the time.

Example 1.4 (Web page ranking). A web search for a given search string usually matches thousands of web pages, so an important question is how to select the most relevant matches to display for the user. The founders of Google developed for this purpose an algorithm called the PageRank [BP98] which is defined as follows.

Consider a directed graph where the nodes consists of all web pages in the world, and links correspond to hyperlinks between the pages. Denote the set of nodes by S , and define the adjacency matrix of the graph as a square matrix G with entries

$$G(x, y) = \begin{cases} 1, & \text{if there is a link from } x \text{ to } y, \\ 0, & \text{else.} \end{cases}$$

Then define a transition matrix on state space S by the formula¹

$$P(x, y) = c \frac{1}{n} + (1 - c) \frac{G(x, y)}{\sum_{y' \in S} G(x, y')},$$

where n is the number of nodes and constant $c \in [0, 1]$ is called a damping factor. The *PageRank* $\pi(x)$ of node x is the probability that a Markov chain with transition matrix P is found in state x after long time ($t \rightarrow \infty$). Whether or not this definition makes sense is not at all trivial. Later we will learn to recognize when such a limiting probability is well defined, and we also learn to compute the probability.

The Markov chain of the PageRank algorithm can be interpreted as a surfer browsing the web by randomly selecting hyperlinks. At times the surfer gets bored and restarts the browsing by selecting a web pages uniformly at random. The damping factor can be interpreted as the probability of the surfer getting bored. ■

¹The formula is valid for graphs where the outdegree $\sum_{y'} G(x, y')$ of every node x is nonzero. When this condition is not met (for example the real web graph), the algorithm needs to be modified, for example by first removing all nodes with zero outdegree.

1.3 Transient distributions

The transient distributions of a Markov chain describe the behavior of the chain in a bounded time horizon. The *distribution* (*jakauma*) of a Markov chain (X_0, X_1, \dots) at time t is the probability distribution of the random variable X_t and is denoted by

$$\mu_t(x) = \mathbb{P}(X_t = x), \quad x \in S.$$

The distribution μ_0 is called the *initial distribution* (*alkujakauma*) of the chain.

The probability that the chain is in state y at time instant $t + 1$ can be computed by conditioning on the state at time instant t according to

$$\mathbb{P}(X_{t+1} = y) = \sum_{x \in S} \mathbb{P}(X_t = x) \mathbb{P}(X_{t+1} = y | X_t = x).$$

By applying (1.2), the above equation can be written as

$$\mu_{t+1}(y) = \sum_{x \in S} \mu_t(x) P(x, y).$$

When the distributions μ_t and μ_{t+1} are interpreted as row vectors indexed by the state space S , we may express the above equation briefly as

$$\mu_{t+1} = \mu_t P. \tag{1.5}$$

This observation leads to the following important result.

Theorem 1.5. *The distribution of a Markov chain at an arbitrary time instant $t = 0, 1, 2, \dots$ can be computed from the initial distribution using the formula*

$$\mu_t = \mu_0 P^t, \tag{1.6}$$

where P^t is the t -th power of the transition matrix P .

Proof. The claim is obviously true for $t = 0$ because P^0 is by definition the identity matrix. If the claim is true for some time instant $t \geq 0$, then by equation (1.5) and the associativity of matrix multiplication, it follows that

$$\mu_{t+1} = \mu_t P = (\mu_0 P^t) P = \mu_0 (P^t P) = \mu_0 P^{t+1},$$

and hence the claim also holds for time instant $t + 1$. According to the induction principle, the claim is valid for all $t \geq 0$. \square

Example 1.6 (Weather model). Let us predict the weather in Otaniemi using the Markov chain in Example 1.2. Assume that it is cloudy on Monday (day $t = 0$). What is the probability that Wednesday is cloudy in Otaniemi? What about Saturday?

The initial distribution corresponding to the nonrandom initial state $X_0 = 1$ equals the Dirac distribution at state 1 which can be written as a row vector

$\mu_0 = [1, 0]$. According to (1.6), the weather distribution of Wednesday can be computed using the formula $\mu_2 = \mu_0 P^2$, so that

$$[\mu_2(1), \mu_2(2)] = [1, 0] \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}^2 = [0.740, 0.260].$$

Hence Wednesday is cloudy with probability 0.740, which is the same number that was found by the manual computation in Example 1.2. Analogously, the distribution of the weather on Saturday can be obtained as $\mu_5 = \mu_0 P^5$, so that,

$$[\mu_5(1), \mu_5(2)] = [1, 0] \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}^5 = [0.715, 0.285].$$

The latter matrix product can be computed using R as

```
P = matrix(c(0.8,0.2,0.5,0.5), nrow=2, byrow=TRUE)
mu0 = c(1,0)
mu5 = mu0*%*(P%^~5)
```

■

1.4 Many-step transition probabilities

The entry $P(x, y)$ of the transition matrix tells the probability of moving from state x to state y during one time step. The following result tells that ...

Theorem 1.7. *The probability that a Markov chain moves from state x to state y during t time steps can be computed using the transition matrix P by the formula*

$$\mathbb{P}(X_t = y \mid X_0 = x) = P^t(x, y), \quad (1.7)$$

where $P^t(x, y)$ is the entry of the t -th power of the transition matrix corresponding to row x and column y .

Proof. The claim is true at time instant $t = 0$ because the identity matrix $I = P^0$ satisfies $P^0(x, y) = \delta_x(y)$.

Assume next that the claim is true for some time instant $t \geq 0$. Then by conditioning on the possible states of X_t , and applying the Markov property (1.1) we find that

$$\begin{aligned} \mathbb{P}(X_{t+1} = y \mid X_0 = x) &= \sum_{x'} \mathbb{P}(X_t = x' \mid X_0 = x) \mathbb{P}(X_{t+1} = y \mid X_t = x', X_0 = x) \\ &= \sum_{x'} P^t(x, x') \mathbb{P}(X_{t+1} = y \mid X_t = x', X_0 = x) \\ &= \sum_{x'} P^t(x, x') P(x', y) \\ &= P^{t+1}(x, y). \end{aligned}$$

Hence the claim is also true for time instant $t+1$, and by the induction principle it holds for all time instants $t \geq 0$. □

Example 1.8 (Weather model). Onninen family has booked a two-day holiday package worth 1900 EUR to a Scottish paradise island. A travel agent offers an insurance at a price of 300 EUR which gives your money back if both days are cloudy. The weather at the destination today is sunny, and the first travel day is after 14 days. Should the Onninen family buy the insurance, when we assume that the weather at the destination follows the Markov chain in Example 1.2?

We use the weather model to compute the probability $\mathbb{P}(X_{14} = 1, X_{15} = 1)$ that both days are cloudy. By conditioning on the state X_{14} and applying the initial condition $X_0 = 2$, we find using (1.7) that

$$\begin{aligned} \mathbb{P}(X_{14} = 1, X_{15} = 1) &= \mathbb{P}(X_{14} = 1) \mathbb{P}(X_{15} = 1 \mid X_{14} = 1) \\ &= \mathbb{P}(X_{14} = 1 \mid X_0 = 2) \mathbb{P}(X_{15} = 1 \mid X_{14} = 1) \\ &= P^{14}(2, 1)P(1, 1) \\ &= 0.571. \end{aligned}$$

The expected net cost of the holiday using the travel insurance is hence $300 + (1 - 0.571) \times 1900 = 1151$ EUR, so that travel insurance is a good investment. ■

1.5 Path probabilities

The initial distribution and the transition matrix of a Markov chain determine the probabilities all possible finite trajectories. The following result tells how these can be computed.

Theorem 1.9. *For any finite-state Markov chain with transition probability matrix P and any $t \geq 1$,*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_t = x_t) = \mu_0(x_0)P(x_0, x_1) \cdots P(x_{t-1}, x_t), \quad (1.8)$$

where $\mu_0(x_0) = \mathbb{P}(X_0 = x_0)$ is the distribution of the initial state X_0 .

Proof. Equality (1.8) is true for $t = 1$ because

$$\mathbb{P}(X_0 = x_0, X_1 = x_1) = \mathbb{P}(X_0 = x_0) \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) = \mu_0(x_0)P(x_0, x_1).$$

To proceed by induction, assume that (1.8) is true for some $t \geq 1$, and denote by $B_t = \{X_0 = x_0, \dots, X_t = x_t\}$ the event that the path of the chain up to time t equals a particular list of states (x_0, \dots, x_t) . Then by noting that $B_{t+1} = B_t \cap \{X_{t+1} = x_{t+1}\}$, we find that

$$\mathbb{P}(B_{t+1}) = \mathbb{P}(B_t) \mathbb{P}(B_{t+1} \mid B_t) = \mathbb{P}(B_t) \mathbb{P}(X_{t+1} = x_{t+1} \mid B_t).$$

Furthermore, the Markov property (1.1) implies that

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid B_t) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t, B_{t-1}) = P(x_t, x_{t+1}).$$

By combining these two equations and then applying the induction assumption, it now follows that

$$\begin{aligned}\mathbb{P}(B_{t+1}) &= \mathbb{P}(B_t) \mathbb{P}(X_{t+1} = x_{t+1} \mid B_t) \\ &= \mathbb{P}(B_t) P(x_t, x_{t+1}) \\ &= \mu_0(x_0) P(x_0, x_1) \cdots P(x_{t-1}, x_t) P(x_t, x_{t+1}),\end{aligned}$$

and therefore (1.8) also holds for time instant $t + 1$. \square

1.6 Occupancy of states

To analyze frequencies of states we employ the following notations. The *indicator* (*indikaattori*) of event A is a binary random variable $1(A)$ such that²

$$1(A) = \begin{cases} 1, & \text{if event } A \text{ occurs,} \\ 0, & \text{else.} \end{cases}$$

The *frequency* (*esiintyvyyys*) of state y among the first t states of the chain is a random integer

$$N_t(y) = \sum_{s=0}^{t-1} 1(X_s = y), \quad (1.9)$$

which tells how many times y occurs in a path (X_0, \dots, X_{t-1}) realized by the Markov chain. The expected frequency of state y for initial state x is defined by

$$M_t(x, y) = \mathbb{E}(N_t(y) \mid X_0 = x).$$

The square matrix M_t with rows and columns indexed by the states $x, y \in S$ is called the *occupancy matrix* (*esiintyvyyysmatriisi*) of the first t states of the chain.

Theorem 1.10. *The occupancy matrix of a Markov chain can be computed using the transition matrix P by*

$$M_t = \sum_{s=0}^{t-1} P^s. \quad (1.10)$$

Proof. Observe first that the expectation of the indicator variable of an arbitrary event A equals

$$\begin{aligned}\mathbb{E}1(A) &= 0 \times \mathbb{P}(1(A) = 0) + 1 \times \mathbb{P}(1(A) = 1) \\ &= \mathbb{P}(1(A) = 1) \\ &= \mathbb{P}(A).\end{aligned}$$

²More precisely, $1(A)$ is a function from the underlying probability space Ω to the set $\{0, 1\}$ which maps ω to 1 if and only if $\omega \in A$.

Hence by formula (1.9) and linearity of the expectation, it follows that

$$\mathbb{E}_x N_t(y) = \mathbb{E}_x \sum_{s=0}^{t-1} 1(X_s = y) = \sum_{s=0}^{t-1} \mathbb{E}_x 1(X_s = y) = \sum_{s=0}^{t-1} \mathbb{P}_x(X_s = y).$$

Because $\mathbb{P}_x(X_s = y) = P^s(x, y)$ due to (1.7), this implies that

$$M_t(x, y) = \mathbb{E}_x N_t(y) = \sum_{s=0}^{t-1} P^s(x, y),$$

which is an entry-by-entry representation of the matrix equation (1.10). \square

Example 1.11 (Weather model). Predict the expected number of cloudy days during a week starting with a sunny day, using the model of Example 1.2.

The requested quantity is the entry $M_7(2, 1)$ of the occupancy matrix M_7 . By applying (1.10) we find that

$$M_7 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix} + \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}^2 + \cdots + \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}^6 = \begin{bmatrix} 5.408 & 1.592 \\ 3.980 & 3.020 \end{bmatrix}.$$

According to the prediction, the expected number of cloudy days is hence 3.980. The above sum of matrix powers can be computed using R as

```
# R-code for computing an occupancy
library(expm)
P = matrix(c(0.8,0.2,0.5,0.5), nrow=2, byrow=TRUE)
M <- Reduce('+', lapply(0:6, function(s) P^s))
```

■

1.7 Simulation of Markov chains

A Markov chain with a given transition matrix can be simulated as follows. First we need to find a random variable U with state space S' and a deterministic function $f : S \times S' \rightarrow S$ such that

$$\mathbb{P}(f(x, U) = y) = P(x, y) \quad \text{for all } x, y \in S. \quad (1.11)$$

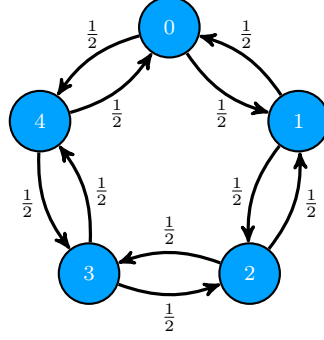
A pair (f, U) satisfying (1.11) is called a *stochastic representation* (*stokastinen esitys*) of the transition matrix P . Then a Markov chain with transition matrix P can be simulated recursively using formula

$$X_{t+1} = f(X_t, U_{t+1}), \quad t = 0, 1, \dots,$$

where random variables U_1, U_2, \dots are mutually independent, independent of X_0 , and have the same distribution as U . Verifying that the resulting random sequence (X_0, X_1, \dots) satisfies the Markov property (1.1) is left as an exercise to the active reader.

Example 1.12 (Random walk on a ring). Consider a cycle graph with node set $S = \{0, \dots, 4\}$. Let (X_0, X_1, \dots) be a symmetric random walk which moves one step clockwise and one step counterclockwise on S with probabilities $1/2$. The transition matrix of the resulting Markov chain is

$$P = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \end{bmatrix}$$



Define a function $f : S \times \{-1, +1\} \rightarrow S$ by

$$f(x, u) = x + u \pmod{5},$$

and let U be a uniformly distributed random variable in $\{-1, +1\}$. Then the pair (f, U) constitutes a stochastic representation of P . The random walk on the ring can hence be simulated using independent coin flips U_1, U_2, \dots where the result of the t -th coin $U_t \in \{-1, +1\}$ tells whether the chain moves counterclockwise ($U_t = -1$) or clockwise ($U_t = +1$). ■

Theorem 1.13. *Every transition matrix P on a finite state space S admits a stochastic representation (f, U) where U is a random number uniformly distributed on the continuous interval $(0, 1)$.*

Proof. Let us label the state space according to $S = \{x_1, \dots, x_n\}$, and let us denote the partial row sums of the transition matrix by

$$q_{i,j} = \sum_{r=1}^j P(x_i, x_r), \quad i, j = 1, \dots, n.$$

We will also set $q_{i,0} = 0$ and define a function $f : S \times (0, 1) \rightarrow S$ by formula

$$f(x_i, u) = x_j, \quad \text{when } q_{i,j-1} < u \leq q_{i,j}.$$

Then if U is a uniformly distributed random number on the continuous interval $(0, 1)$, it follows that

$$\mathbb{P}(f(x_i, U) = x_j) = \mathbb{P}(q_{i,j-1} < U \leq q_{i,j}) = q_{i,j} - q_{i,j-1} = P(x_i, x_j).$$

Because the above equation holds for all states x_i and x_j we conclude that (f, U) is a stochastic representation of P . □

Stochastic representations are not unique. To see why, it suffices to note that the random variable $1 - U$ is uniformly distributed on $(0, 1)$ whenever U has the same property. Therefore, if (f, U) is a stochastic representation of P of the form in Theorem 1.13, then so is the pair (g, U) with $g(x, u) = f(x, 1 - u)$. Indeed, it is not hard to verify that there are infinitely many stochastic representations for any transition matrix. Moreover, Theorem 1.13 is valid for arbitrary measurable state spaces. When the state space is countably infinite, the same proof as above can easily be generalized. When the state space is uncountably infinite, deeper results of measure theory are needed, see for example [Kal02].

Chapter 2

Markov chains in the long run

2.1 Invariant and limiting distributions

In the previous chapter we learned to compute the transient distributions μ_t of a Markov chain with initial distribution μ_0 using the formula $\mu_t = \mu_0 P^t$. When looking at a long time horizon, it is natural to ask the following questions:

1. Do the transient distributions admit a *limiting distribution* (*rajaajakauma*) $\lim_{t \rightarrow \infty} \mu_t$ as t grows larger and larger?
2. If a limiting distribution exists, does it depend on the initial distribution, or is it unique?
3. If a limiting distribution exists, how can it be computed?

Answering the first two questions requires careful analysis and sufficient structural assumptions. The third question is easier, so we will treat it first.

A probability distribution $\pi = (\pi(x) : x \in S)$ is called an *invariant distribution* (*tasapainojakauma*) of a transition matrix P and a corresponding Markov chain, if it satisfies the balance equations

$$\sum_{x \in S} \pi(x) P(x, y) = \pi(y), \quad y \in S, \quad (2.1)$$

or in matrix form (with π interpreted as a row vector),

$$\pi P = \pi.$$

If a Markov chain is started with initial distribution $\mu_0 = \pi$, we find by using Theorem 1.5 and the associativity of matrix multiplication that

$$\mu_t = \pi P^t = (\pi P) P^{t-1} = \pi P^{t-1} = \dots = \pi P = \pi.$$

Hence for a Markov chain with a random initial state distributed according to an invariant distribution, the distribution of X_t remains invariant for all time instants $t = 0, 1, 2, \dots$

The following result tells that if a Markov chain has a limiting distribution, it can be determined as a solution of the linear system of equations (2.1).

Theorem 2.1. *If π is a limiting distribution of a finite-state Markov chain, then it is also an invariant distribution.*

Proof. By the associativity of matrix multiplication we see that

$$\mu_{t+1} = \mu_0 P^{t+1} = (\mu_0 P^t) P = \mu_t P,$$

which can be written entry-by-entry as

$$\mu_{t+1}(y) = \sum_{x \in S} \mu_t(x) P(x, y).$$

If we assume that $\mu_t(x) \rightarrow \pi(x)$ for every $x \in S$, we see by taking limits on both sides of the above equation that

$$\pi(y) = \lim_{t \rightarrow \infty} \mu_{t+1}(y) = \sum_{x \in S} \lim_{t \rightarrow \infty} \mu_t(x) P(x, y) = \sum_{x \in S} \pi(x) P(x, y).$$

Hence the balance equation (2.1) is valid. Moreover, because μ_t is a probability distribution, we know that

$$\sum_{x \in S} \mu_t(x) = 1$$

for all t . By taking limits on both sides of the above equation as $t \rightarrow \infty$ we see that $\sum_{x \in S} \pi(x) = 1$, so that π is a probability distribution on S . \square

Esimerkki 2.2 (Brand loyalty). A smartphone market is dominated by three manufacturers. When buying a new phone, a customer chooses to buy a phone from the same manufacturer i as the previous one with probability β_i , and otherwise the customer randomly chooses one of the other manufacturers. Assume that $\beta_1 = 0.8$, $\beta_2 = 0.6$, and $\beta_3 = 0.4$, and that all smartphones have the same lifetime regardless of the manufacturer. Will the market shares of the different manufacturers stabilize in the long run?

Let us model the manufacturer of a typical customer's phone after the t -th purchase instant by a Markov chain (X_0, X_1, \dots) with state space $S = \{1, 2, 3\}$ and transition matrix

$$P = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}.$$

We can compute powers of P using a computer:

$$P^2 = \begin{bmatrix} 0.69 & 0.17 & 0.14 \\ 0.34 & 0.44 & 0.22 \\ 0.42 & 0.33 & 0.25 \end{bmatrix}, \dots,$$

$$P^{10} = \begin{bmatrix} 0.5471287 & 0.2715017 & 0.1813696 \\ 0.5430034 & 0.2745217 & 0.1824748 \\ 0.5441087 & 0.2737123 & 0.1821790 \end{bmatrix}, \dots,$$

$$P^{20} = \begin{bmatrix} 0.5454610 & 0.2727226 & 0.1818165 \\ 0.5454452 & 0.2727341 & 0.1818207 \\ 0.5454494 & 0.2727310 & 0.1818196 \end{bmatrix}.$$

The above computations indicate that after 20 phone purchases, an initial customer of manufacturer i is a customer of manufacturer 1 with probability $P^{20}(i, 1) \approx 0.545$. Because the rows of P^{20} are approximately equal, the effect of initial state $i = 1, 2, 3$ becomes negligible over time. Hence it appears that the market shares stabilize towards a limiting distribution

$$[0.5454545, 0.2727273, 0.1818182].$$

The balance equations $\pi P = \pi$ and $\sum_{x=1}^3 \pi(x) = 1$ for transition matrix P can be written as

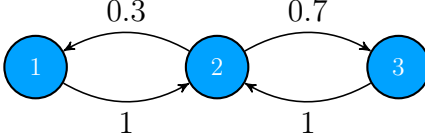
$$\begin{aligned} 0.8\pi(1) + 0.2\pi(2) + 0.3\pi(3) &= \pi(1) \\ 0.1\pi(1) + 0.6\pi(2) + 0.3\pi(3) &= \pi(2) \\ 0.1\pi(1) + 0.2\pi(2) + 0.4\pi(3) &= \pi(3) \\ \pi(1) + \pi(2) + \pi(3) &= 1. \end{aligned}$$

The unique solution of the above system of linear equations is

$$\pi = \left[\frac{6}{11}, \frac{3}{11}, \frac{2}{11} \right] \approx [0.5454545, 0.2727273, 0.1818182],$$

which is close to the numerically found limiting distribution, as it should according to Theorem 2.1. ■

Example 2.3 (Chain with no limiting distribution). Consider a Markov chain on state space $S = \{1, 2, 3\}$ with initial state $X_0 = 1$, and transition matrix

$$P = \begin{bmatrix} 0.0 & 1 & 0.0 \\ 0.3 & 0 & 0.7 \\ 0 & 1 & 0 \end{bmatrix}.$$


By computing powers of P we see that

$$\begin{aligned} P^2 &= \begin{bmatrix} 0.3 & 0 & 0.7 \\ 0 & 1 & 0 \\ 0.3 & 0 & 0.7 \end{bmatrix}, & P^3 &= \begin{bmatrix} 0 & 1 & 0 \\ 0.3 & 0 & 0.7 \\ 0 & 1 & 0 \end{bmatrix}, \\ P^4 &= \begin{bmatrix} 0.3 & 0 & 0.7 \\ 0 & 1 & 0 \\ 0.3 & 0 & 0.7 \end{bmatrix}, & P^5 &= \begin{bmatrix} 0 & 1 & 0 \\ 0.3 & 0 & 0.7 \\ 0 & 1 & 0 \end{bmatrix}, \end{aligned}$$

from which we observe that

$$P^t = \begin{cases} P, & t = 1, 3, 5, \dots, \\ P^2 & t = 2, 4, 6, \dots \end{cases}$$

The distribution μ_t of the chain with nonrandom initial state $X_0 = 1$ (corresponding to initial distribution $\mu_0 = [1, 0, 0]$) hence satisfies

$$\mu_t = \mu_0 P^t = \begin{cases} [0, 1, 0] & \text{for } t = 1, 3, 5, \dots, \\ [0.3, 0, 0.7] & \text{for } t = 2, 4, 6, \dots \end{cases}$$

Such a chain has no limiting distribution. However, a direct computation shows that $\pi = [0.15, 0.50, 0.35]$ is an invariant distribution for the chain. ■

Example 2.4 (Chain with many limiting distributions). Consider a Markov chain on state space $S = \{1, 2, 3, 4\}$ with transition matrix

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0.1 & 0.8 & 0.1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

A direct computation reveals that

$$\mu_0 P^t = \begin{cases} [0.5, 0.5, 0, 0] & \text{for all } t \geq 1 \text{ if } \mu_0 = [1, 0, 0, 0], \\ [0, 0, 0, 1] & \text{for all } t \geq 1 \text{ if } \mu_0 = [0, 0, 0, 1]. \end{cases}$$

This Markov chain can hence have several limiting distributions, depending on the initial state. As a consequence (Theorem 2.1), both $\pi^{(1,2)} = [0.5, 0.5, 0, 0]$ and $\pi^{(4)} = [0, 0, 0, 1]$ are invariant distributions of P . By linearity, one can verify that every probability distribution of the form

$$\pi = \alpha \pi^{(1,2)} + (1 - \alpha) \pi^{(4)}, \quad 0 \leq \alpha \leq 1,$$

is an invariant distribution of P . ■

2.2 Connectivity

Given a transition matrix P , we denote $x \rightsquigarrow y$, if the corresponding transition diagram of contains a path from x to y . Here we allow paths of length zero, so that $x \rightsquigarrow x$. A transition matrix P and a corresponding Markov chain is called *irreducible* (*yhtenäinen*), if $x \rightsquigarrow y$ for all $x, y \in S$. In graph theoretical terms, a Markov chain is irreducible if and only if its transition diagram is a strongly connected directed graph.

Example 2.5 (Irreducible Markov chains). The following Markov chains are irreducible:

- Weather model (Example 1.2)
- Inventory model (Example 1.3)

- Brand loyalty (Example 2.2) ■

The structure of Markov chains which are not irreducible can be analyzed by defining a symmetric relation by denoting $x \leftrightarrow y$ if $x \rightsquigarrow y$ and $y \rightsquigarrow x$. This equivalence relation partitions the state space into equivalence classes $C(x) = \{y \in S : y \leftrightarrow x\}$, called the *components* (*komponentti*) of P . An irreducible chain has only one component which contains all states of the state space. A component is called *absorbing* (*absorboiva*) if the chain cannot exit the component, otherwise the component is called *transient* (*väistyvä*).

Example 2.6 (Reducible Markov chain). The chain in Example 2.4 is not irreducible because the chain cannot move away from state 4. The transition diagram of this chain can have three components $C(1) = C(2) = \{1, 2\}$, $C(3) = \{3\}$, and $C(4) = \{4\}$. The components $\{1, 2\}$ and $\{4\}$ are absorbing, and the component $\{3\}$ is transient. ■

Theorem 2.7. *A transition matrix P is irreducible if and only if for all $x, y \in S$ there exists an integer $t \geq 1$ such that $P^t(x, y) > 0$.*

Proof. Assume first that P is irreducible and select some states $x \neq y$. Then the transition diagram contains a path $x = x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_t = y$, so that

$$P(x_0, x_1)P(x_1, x_2) \cdots P(x_{t-1}, x_t) > 0.$$

As a consequence,

$$\begin{aligned} P^t(x, y) &= \mathbb{P}(X_t = y \mid X_0 = x) \\ &= \mathbb{P}(X_t = x_t \mid X_0 = x_0) \\ &\geq \mathbb{P}(X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1 \mid X_0 = x_0) \\ &= P(x_0, x_1)P(x_1, x_2) \cdots P(x_{t-1}, x_t) \\ &> 0. \end{aligned}$$

To prove the converse statement, assume that $P^t(x, y) > 0$ for some integer $t \geq 1$. Then $\mathbb{P}(X_t = y \mid X_0 = x) > 0$, so that a Markov chain starting at x can be located in state y after t time instants. This is only possible if the transition diagram contains a path of length t from x to y , so that $x \rightsquigarrow y$. □

2.3 Invariant distribution of an irreducible chain

Theorem 2.8. *Every irreducible transition matrix on a finite state space has a unique invariant distribution.*

A clear and detailed proof of Theorem 2.8 is presented in [LPW08, Sec 1.5], and here we only describe the main ideas of the proof. The existence of an invariant distribution can be shown by verifying that

$$\pi(x) = \frac{1}{\mathbb{E}(\tau_x^+ \mid X_0 = x)} \tag{2.2}$$

is a probability distribution which satisfies the balance equations (2.1), where the random variable

$$\tau_x^+ = \min\{t \geq 1 : X_t = x\}$$

denotes the *positive passage time* (*positiivinen kulku-aika*) of the Markov chain to state x . For an irreducible chain on a finite state space one can prove that the chain surely visits all states of the state space, and hence τ_x^+ is a well-defined random integer.

Formula (2.2) can be interpreted as follows. The invariant probability $\pi(x)$ corresponds to the relative proportion of time instants that the Markov chain is observed in state x . This quantity is inversely proportional to the expected length of the time intervals between consecutive visits in state x . In practice, the invariant distribution usually cannot be computed from (2.2). Instead, the invariant distribution is obtained by solving the balance equation $\pi = \pi P$.

The uniqueness of the invariant distribution can be justified by first verifying that for an irreducible transition matrix, all column vectors solving $Ph = h$ are of the form $h = [c, c, \dots, c]^T$, so that the null space of $P - I$ is one-dimensional. Using basic facts of linear algebra one can conclude from this that also the linear space of (row vector) solutions to $\mu(P - I) = 0$ has dimension one. This space contains at most one solution satisfying the normalization constraint $\sum_x \mu(x) = 1$. Hence an irreducible transition matrix P may have at most one invariant distribution.

2.4 Periodicity

The *period* (*jakso*) of state x for a Markov chain moving according to transition matrix P is the greatest common denominator of the time instants at which the chain started at x may return to its initial state. The set of possible return times can be written as

$$\mathcal{T}_x = \{t \geq 1 : P^t(x, x) > 0\},$$

so that the period of x is the largest positive integer which divides all elements of \mathcal{T}_x . The period is not defined for states for which the set of possible return times is empty.

Usually the period of a state is easy to determine from the transition diagram. If the lengths of all cycles starting and ending at x are multiples of some integer d , and if d is the largest such integer, then d is the period of x . A transition matrix P and a corresponding Markov chain is *aperiodic* (*jaksoton*) if every state has period 1.

Example 2.9 (Aperiodic Markov chains). The following Markov chains are aperiodic (convince yourself that this really is the case):

- Weather model (Example 1.2)
- Inventory model (Example 1.3)

- Brand loyalty model (Example 2.2)

The PageRank chain (Example 1.4) is aperiodic whenever the damping factor c is nonzero. ■

Example 2.10 (Periodic chain). The Markov chain in Example 2.3 is periodic with every state having period 2. ■

2.5 Invariant distribution of an irreducible aperiodic chain

The following important result summarizes the basic theory of Markov chains and explains why nearly all Markov chains on finite state spaces settle into a statistical equilibrium in the long run.

Theorem 2.11. *Every irreducible and aperiodic Markov chain on a finite state space admits a unique limiting distribution which also equals the unique invariant distribution of the chain, and can be determined by solving the balance equations $\pi P = \pi$ and $\sum_x \pi(x) = 1$.*

If (X_0, X_1, X_2, \dots) is a Markov chain satisfying the conditions of Theorem 2.11 and X_∞ is a random variable distributed according to the invariant distribution π , then the result of the above theorem is usually expressed as

$$X_t \xrightarrow{d} X_\infty,$$

which means that the random sequence (X_0, X_1, \dots) *converges in distribution* (*suppenee jakaumaltaan*) towards random variable X_∞ . This notion of convergence can be defined for probability distributions on general topological spaces. In case of a finite or countably infinite state space this means that the probability mass functions μ_t of the random variables X_t converge pointwise to π . Let us emphasize that the realizations of the random sequence (X_0, X_1, \dots) do *not* in general converge to any fixed point in S . Instead, the limit describes a statistical equilibrium where the chain will settle in the long run.

The existence of the limit in Theorem 2.11 can be proved using methods of matrix analysis, or by applying stochastic couplings. Students majoring in mathematics are recommended to have a look at [LPW08, Sec 4–5], where both proof techniques are explained in detail. The fact that the limiting distribution is also an invariant distribution follows from Theorem 2.1.

Chapter 3

Markov additive processes

3.1 Definitions

In many applications we need to analyse sums of random numbers which depend on the realised trajectory of a Markov chain. Examples include cumulative rewards in reinforcement learning, revenues and costs in financial models and technological systems, and frequencies related to statistical models. Markov additive processes provide a rich modeling framework for such applications and admit powerful numerical formulas based on linear algebra.

A random sequence $(X_0, V_0), (X_1, V_1), \dots$ is called a *Markov additive process* (*Markov-additiivinen prosessi*) if (X_0, X_1, \dots) is a Markov chain and (V_0, V_1, \dots) is a real-valued random process which can be represented as

$$V_t = \phi(X_0, U_0) + \dots + \phi(X_{t-1}, U_{t-1}) \quad (3.1)$$

for some deterministic function ϕ and some independent and identically distributed random variables U_0, U_1, \dots such that U_t is independent of (X_0, \dots, X_t) for all $t \geq 0$. For $t = 0$, the empty sum above is defined to be $V_0 = 0$. Here (X_t) is called the *Markov component* and (V_t) the *additive component* of the Markov additive process.

Example 3.1 (November rain). A simple model of November weather in Espoo consists of a Markov chain (X_0, X_1, \dots) with state space $\{-30, -29, \dots, 30\}$ modeling the daily temperature, and a sequence of random variables U_0, U_1, \dots with two possible values: 0 = “dry” and 1 = “rain”. The number V_t of snowy days among the first t days of the month can be expressed using (3.1) with

$$\phi(x, u) = \begin{cases} 1, & \text{if } x \leq -1 \text{ and } u = 1, \\ 0, & \text{else.} \end{cases}$$

If the rain indicators U_0, U_1, \dots are mutually independent, identically distributed, and independent of the daily temperatures, then $(X_0, V_0), (X_1, V_1), \dots$ is a Markov additive process. ■

3.2 Behaviour in finite time horizon

The following result tells how the expectation

$$g_t(x) = \mathbb{E}(V_t | X_0 = x)$$

related to a Markov additive process (X_t, V_t) defined by (3.1) can be computed using the transition matrix of the underlying Markov chain and the function $v : S \rightarrow \mathbb{R}$ defined by

$$v(x) = \mathbb{E}\phi(x, U_0). \quad (3.2)$$

We usually consider the above functions g_t and v as *column vectors* indexed by the states. In this case the result below can be written in matrix form as

$$g_t = \sum_{s=0}^{t-1} P^s v, \quad (3.3)$$

which also equals $M_t v$ where M_t is the occupancy matrix appearing in (1.10).

Theorem 3.2. *For a Markov additive process in which the Markov component (X_0, X_1, \dots) has transition matrix P and finite state space S ,*

$$\mathbb{E}(V_t | X_0 = x) = \sum_{s=0}^{t-1} \sum_{y \in S} P^s(x, y) v(y).$$

Proof. The Markov property of (X_t) implies (U_t can be treated below as if it were deterministic because it is independent of the Markov chain) that

$$\mathbb{E}(\phi(X_t, U_t) | X_t = y, X_0 = x) = \mathbb{E}(\phi(X_t, U_t) | X_t = y) = \mathbb{E}\phi(y, U_t).$$

Because U_t has the same distribution as U_0 , we get $\mathbb{E}\phi(y, U_t) = \mathbb{E}\phi(y, U_0) = v(y)$, and hence

$$\mathbb{E}(\phi(X_t, U_t) | X_t = y, X_0 = x) = v(y).$$

As a consequence, by conditioning on the possible values of X_t , it follows that

$$\begin{aligned} \mathbb{E}(\phi(X_t, U_t) | X_0 = x) &= \sum_{y \in S} \mathbb{P}(X_t = y | X_0 = x) \mathbb{E}(\phi(X_t, U_t) | X_t = y, X_0 = x) \\ &= \sum_{y \in S} P^t(x, y) v(y). \end{aligned}$$

By linearity of the expectation, it hence follows by (3.1) that

$$\mathbb{E}(V_t | X_0 = x) = \sum_{s=0}^{t-1} \mathbb{E}(\phi(X_s, U_s) | X_0 = x) = \sum_{s=0}^{t-1} \sum_{y \in S} P^s(x, y) v(y).$$

□

Example 3.3 (Inventory model). Recall the inventory model of Example 1.3. Assume that the store buys laptops for 590 EUR and sells them for 790 EUR. The storage cost per week is 50 EUR for every laptop in stock at the beginning of a week. Determine the expected net revenue from ten forthcoming weeks, when in the beginning of the first week there are five laptops in stock.

Denote by V_t the net revenue (sales income minus storage costs) from the first t weeks. The number of laptops in stock X_t in the beginning of week t is a Markov chain with state space $S = \{2, 3, 4, 5\}$ with initial state $X_0 = 5$. Now consider a week t starting with X_t laptops in stock. Then the storage costs (EUR) for the week equal $50X_t$, and the number of sold laptops equals $\min(X_t, D_t)$ where D_t is the demand of week t . Because the weekly demands are mutually independent and identically distributed, and D_t is independent of (X_0, \dots, X_t) , it follows that (X_t, V_t) is a Markov additive process with representation

$$V_t = \sum_{s=0}^{t-1} \phi(X_s, D_s)$$

where

$$\phi(x, u) = (790 - 590) \min(x, u) - 50x.$$

To compute the expectation of V_t using Theorem 3.2, we need to compute the function $v(x) = \mathbb{E}\phi(x, D_0)$. Because the demands are Poisson distributed with mean $\lambda = 3.5$, we see that the expected number of laptops sold during a week starting with x laptops in stock equals

$$\begin{aligned} \mathbb{E} \min(x, D_0) &= \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \min(x, k) \\ &= \sum_{k=0}^x e^{-\lambda} \frac{\lambda^k}{k!} k + \left(1 - \sum_{k=0}^x e^{-\lambda} \frac{\lambda^k}{k!}\right) x \\ &= x - \sum_{k=0}^x e^{-\lambda} \frac{\lambda^k}{k!} (x - k), \end{aligned}$$

and hence

$$v(x) = (790 - 590) \left(x - \sum_{k=0}^x e^{-\lambda} \frac{\lambda^k}{k!} (x - k) \right) - 50x.$$

By evaluating formula (3.3) using a computer program we find that (recall that column vectors are indexed by the states $x = 2, 3, 4, 5$)

$$v = \begin{bmatrix} 266.78 \\ 352.61 \\ 395.29 \\ 400.20 \end{bmatrix} \quad \text{and} \quad g_{10} = \begin{bmatrix} 3627.24 \\ 3704.00 \\ 3735.81 \\ 3735.00 \end{bmatrix}.$$

Hence the expected net revenue from next ten weeks is 3735 EUR. Note that the expected net revenue would be 0.81 EUR higher if there would initially be 4 instead of 5 laptops in stock. This is in contrast with one-week expected revenues $g_1(x) = v(x)$ satisfying $g_1(4) < g_1(5)$, and indicates that actions which maximise one-week outcomes may not be optimal for longer time horizons. ■

```
# R-code for computing the function v(x)
v <- numeric(4)
for (x in 2:5) {
  k <- 0:x
  v[x-1] <- (790-590)*(x - sum((x-k)*dpois(k,la))) - 50*x
}

# R-code for computing the function g(x)
library(expm)
M <- Reduce('+', lapply(0:9, function(s) P%~%s))
g <- M%*%v
```

3.3 Ergodicity

So far we have learned that the distribution of an irreducible and aperiodic Markov chain converges to the unique invariant distribution π of the chain. The following result provides an alternative interpretation for the invariant distribution which tells that a long-term time average of a random sequence $\phi(X_0), \phi(X_1), \dots$ is close to the mathematical expectation of the invariant distribution. Such a phenomenon is called an *ergodic* (*ergodinen*) property. Note that periodicity is not an issue in the statement below because the time averages smoothen out periodic effects present in the model.

Theorem 3.4. *For any irreducible Markov chain with a finite state space S and for any function $\phi : S \rightarrow \mathbb{R}$,*

$$\frac{1}{t} \sum_{s=0}^{t-1} \phi(X_s) \rightarrow \sum_{y \in S} \pi(y) \phi(y) \quad \text{as } t \rightarrow \infty$$

with probability one, regardless of the initial state of the chain.

The above result can be proved by fixing some initial state x and keeping track of successive visits of the chain to x . By the Markov property, the paths between successive visits are stochastically independent, and Theorem 3.4 can be proved by applying a strong law of large numbers [LPW08, Sec 4.7].

As an important consequence, we obtain the following result regarding empirical relative frequencies. The *empirical relative frequency* (*empiirinen suhteellinen esiintyvyyys*) of state y among the first t states of a stochastic process (X_0, X_1, \dots) is defined by

$$\hat{\pi}_t(y) = \frac{N_t(y)}{t},$$

where $N_t(y) = \sum_{s=0}^{t-1} 1(X_s = y)$ is the corresponding absolute frequency. Note that $\hat{\pi}_t(y)$ is a random number determined by the realised trajectory of (X_0, \dots, X_{t-1}) .

The following result confirms that the value of the invariant distribution $\pi(y)$ can be interpreted as the long-term relative limiting frequency of time instants that the chain spends in state y .

Theorem 3.5. *The relative frequencies of an irreducible Markov chain with a finite state space S satisfy*

$$\lim_{t \rightarrow \infty} \hat{\pi}_t(y) = \pi(y) \quad (3.4)$$

with probability one, regardless of the initial state of the chain. Moreover, the occupancy matrix of the chain satisfies

$$\lim_{t \rightarrow \infty} \frac{M_t(x, y)}{t} \rightarrow \pi(y) \quad \text{for all } x, y \in S. \quad (3.5)$$

Proof. Fix a state y , and define a function $\phi : S \rightarrow \mathbb{R}$ by

$$\phi(x) = 1(x = y) = \begin{cases} 1, & \text{if } x = y, \\ 0, & \text{else.} \end{cases}$$

Then the frequency $N_t(y)$ of state y can be written as

$$N_t(y) = \sum_{s=0}^{t-1} \phi(X_s).$$

By applying Theorem 3.4 we conclude that

$$\lim_{t \rightarrow \infty} \hat{\pi}_t(y) = \lim_{t \rightarrow \infty} \frac{N_t(y)}{t} = \sum_{x \in S} \pi(x) \phi(x) = \pi(y)$$

with probability one, regardless of the initial state.

Moreover, the relative frequency of state y is bounded by

$$0 \leq \hat{\pi}_t(y) \leq 1$$

with probability one for all t . By taking the limit $t \rightarrow \infty$ inside an expectation¹ and applying (3.4), it follows that

$$\lim_{t \rightarrow \infty} \frac{M_t(x, y)}{t} = \lim_{t \rightarrow \infty} \mathbb{E} \left(\hat{\pi}_t(y) \mid X_0 = x \right) = \mathbb{E} \left(\lim_{t \rightarrow \infty} \hat{\pi}_t(y) \mid X_0 = x \right) = \pi(y).$$

□

¹This is allowed for bounded random sequences due to Lebesgue's dominated convergence theorem, which is a topic of the course MS-E1600 Probability theory.

3.4 Long-term behaviour

For a Markov additive process (X_t, V_t) , the process V_t usually does not converge to a statistical equilibrium even if the underlying Markov chain (X_t) does so. Rather V_t might tend to infinity or minus infinity in the long run. Therefore, it makes sense to analyse the long-term growth rates V_t/t . The following result tells that under mild regularity conditions, the expected growth rate

$$\frac{g_t(x)}{t} = \mathbb{E} \left(\frac{V_t}{t} \mid X_0 = x \right)$$

has a limit as $t \rightarrow \infty$ which does not depend on the initial state $X_0 = x$.

Theorem 3.6. *For a Markov additive process (X_t, V_t) in which the Markov component (X_t) is irreducible on a finite state space S ,*

$$\lim_{t \rightarrow \infty} \frac{g_t(x)}{t} = \sum_{y \in S} \pi(y)v(y).$$

for all $x \in S$.

Proof. By Theorem 3.2 we see that

$$g_t(x) = \sum_{y \in S} M_t(x, y)v(y).$$

Therefore, by (3.5),

$$\lim_{t \rightarrow \infty} \frac{g_t(x)}{t} = \sum_{y \in S} \left(\lim_{t \rightarrow \infty} \frac{M_t(x, y)}{t} \right) v(y) = \sum_{y \in S} \pi(y)v(y).$$

□

Example 3.7 (Inventory model). Let us continue the analysis of Example 3.3. What is the long-term expected revenue rate?

Because the Markov chain (X_t) is irreducible, it has a unique invariant distribution π which can be solved from the balance equations $\pi P = \pi$ and $\sum_x \pi(x) = 1$. By applying Theorem 3.6 we conclude that the long-term expected revenue rate equals

$$\lim_{t \rightarrow \infty} \frac{g_t(x)}{t} = \sum_{y \in S} \pi(y)v(y)$$

which does not depend on the initial state x of the inventory. By computing the numerical values, we find that the expected long-term revenue rate equals 371.29 EUR per week. This corresponds to approximately 3713 EUR revenue rate per a 10-week period, and is quite close to the expected cumulative revenues computed in Example 3.3 which depend on the initial state. ■

3.5 Remarks

The theory of Markov additive processes can be generalised into continuous time and general uncountable state spaces. Also, Theorem 3.6 can be generalised to a form where convergence takes place with probability one. Asmussen's book [Asm03] provides details.

Chapter 4

Passage times and hitting probabilities

4.1 Passage times

The *passage time* (*kulkuaiika*) of a random process (X_0, X_1, \dots) into set A is defined by

$$T_A = \min\{t \geq 0 : X_t \in A\},$$

with the notational convention that $T_A = \infty$ if the process never visits A . The passage time is hence a random variable which takes on values in the extended set of integers $\{0, 1, 2, \dots\} \cup \{\infty\}$. The *expected passage time* (*odotettu kulkuaiika*) into set A for a Markov chain starting at state x is denoted by

$$k_A(x) = \mathbb{E}(T_A | X_0 = x).$$

Theorem 4.1. *The expected passage times $(k_A(x) : x \in S)$ form the smallest nonnegative solution to the system of equations*

$$\begin{aligned} f(x) &= 1 + \sum_{y \notin A} P(x, y)f(y), & x \notin A, \\ f(x) &= 0, & x \in A. \end{aligned} \tag{4.1}$$

From the harmonic analysis point of view, the system of equations (4.1) corresponds to a Poisson equation on $B = A^c$

$$Df(x) = -1, \quad x \in B, \tag{4.2}$$

with boundary condition

$$f(x) = 0, \quad x \in \partial B,$$

where $B = A^c$, $\partial B = A$, and the linear map $D : f \mapsto Pf - f$ is called the *drift matrix* (*virtausmatriisi*) of the Markov chain. The smallest nonnegative solution can be found by first setting $f_0(x) = 0$ for all x and the recursively computing

$$f_{n+1}(x) = \begin{cases} 1 + \sum_{y \notin A} P(x, y)f_n(y), & x \notin A, \\ 0, & x \in A. \end{cases}$$

Then it is possible to prove that f_0, f_1, f_2, \dots forms a nondecreasing sequence of functions with pointwise limit $f(x) = \lim_{n \rightarrow \infty} f_n(x)$. The limit f takes on values in the extended number set $[0, \infty]$ and is the smallest nonnegative solution of (4.1). Verifying these statements is a good exercise for a mathematically oriented reader. A good exercise for a programming oriented reader is to implement an algorithm which computes the above limit numerically.

Before proving Theorem 4.1 let us consider the following example where the result can be applied.

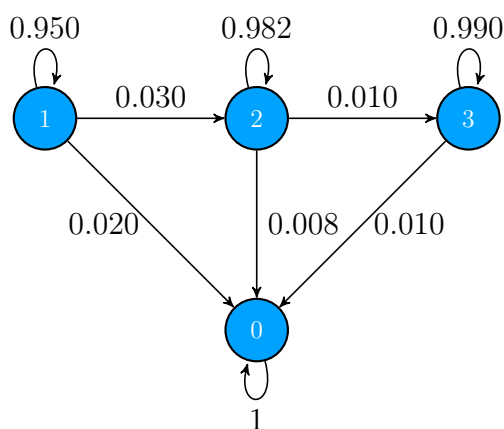
Example 4.2 (Human resource management). Kalvonväentäjät Oy is management consulting company which has 100 employees divided into three salary categories: 1 = 'junior', 2 = 'senior' ja 3 = 'partner'.

An employee holding a junior position in the beginning of a month gets promoted to senior with probability 0.030, leaves the company with probability 0.020, and otherwise continues in the same position in the beginning of next month. Similarly, a senior gets promoted to a partner with probability 0.010, leaves the company with probability 0.008, and otherwise continues in the same position. A partner leaves the company with probability 0.010. What is the expected duration that a newly recruited employee remains in the company? How long is a freshly promoted partner expected to serve in the company?

We will assume that all promotions and exits occur independently of the states of the previous months. The career development of an employee can then be modeled using a Markov chain on state space $\{0, 1, 2, 3\}$ where state 0 means that the employee has left the company, with transition matrix

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.020 & 0.950 & 0.030 & 0 \\ 0.008 & 0 & 0.982 & 0.010 \\ 0.010 & 0 & 0 & 0.990 \end{bmatrix}. \quad (4.3)$$

State 0 is absorbing and the other states are transient, as is clearly visible from the transition diagram below:



The time (in months) in service for a newly recruited junior is the passage time of the Markov chain from state 1 into state 0. The expectation of this

random integer equals $k_A(1)$ with $A = \{0\}$. According to Theorem 4.1, the expected passage times solve the equations

$$f(x) = 1 + \sum_{y=1}^3 P(x, y)f(y), \quad x = 1, 2, 3,$$

which now can be written as

$$\begin{aligned} f(1) &= 1 + 0.950 f(1) + 0.030 f(2) \\ f(2) &= 1 + 0.982 f(2) + 0.010 f(3) \\ f(3) &= 1 + 0.990 f(3). \end{aligned}$$

These can be solved by first setting

$$f(3) = \frac{1}{1 - 0.990} = 100,$$

then

$$f(2) = \frac{1 + 0.010 f(3)}{1 - 0.982} = 111.11,$$

and finally

$$f(1) = \frac{1 + 0.030 f(2)}{1 - 0.950} = 86.67.$$

Because we found only one nonnegative solution $f = [f(1), f(2), f(3)]$ to the above equations, the above solution provides the expected passage time according to Theorem 4.1, so that $k_A = f$. Hence a freshly hired junior is expected to serve in the company for 86.67 months ≈ 7.2 years, and a freshly promoted partner is expected to serve in the company for 100 months ≈ 8.3 years. ■

Proof of Theorem 4.1. Let us first verify that the numbers $k_A(x)$ satisfy equations (4.1). We will do this by applying first-step analysis, that is, by conditioning on the possible states of the first state. When the initial state $x \in A$, we surely have $T_A = 0$, so that $k_A(x) = 0$. Assume next that $x \notin A$. Then by conditioning on X_1 we find that

$$k_A(x) = \sum_{y \in S} P(x, y) \mathbb{E}(T_A | X_1 = y, X_0 = x). \quad (4.4)$$

When $x \notin A$,

$$T_A = \min\{t \geq 1 : X_t \in A\} = 1 + \min\{t \geq 0 : X_{t+1} \in A\},$$

so that by applying the Markov property we may conclude that

$$\mathbb{E}(T_A | X_1 = y, X_0 = x) = 1 + \mathbb{E}(T_A | X_0 = y) = 1 + k_A(y).$$

By combining the above observation with formula (4.4) we see that

$$\begin{aligned} k_A(x) &= \sum_{y \in S} P(x, y)(1 + k_A(y)) \\ &= \sum_{y \in S} P(x, y) + \sum_{y \in S} P(x, y)k_A(y). \end{aligned}$$

The uppermost equality in (4.1) follows from this after recalling that the rows sums of P equal one, and $k_A(y) = 0$ for $y \in A$.

Let us next verify that $(k_A(x) : x \in S)$ is the smallest nonnegative solution. Assume that $(f(x) : x \in S)$ some nonnegative solution of (4.1). Then we need to verify that

$$f(x) \geq k_A(x) \tag{4.5}$$

for all x . Obviously (4.5) holds for all $x \in A$, because then $f(x) = k_A(x) = 0$. Assume next that $x \notin A$. Then

$$\begin{aligned} f(x) &= 1 + \sum_{y \notin A} P(x, y)f(y) \\ &= 1 + \sum_{y \notin A} P(x, y) \left(1 + \sum_{z \notin A} P(y, z)f(z) \right) \\ &= 1 + \sum_{y \notin A} P(x, y) + \sum_{y \notin A} \sum_{z \notin A} P(x, y)P(y, z)f(z). \end{aligned}$$

Because¹ $\mathbb{P}_x(T_A \geq 1) = 1$ and

$$\sum_{y \notin A} P(x, y) = \mathbb{P}_x(T_A \geq 2),$$

the above equation can be written as

$$f(x) = \mathbb{P}_x(T_A \geq 1) + \mathbb{P}_x(T_A \geq 2) + \sum_{y \notin A} \sum_{z \notin A} P(x, y)P(y, z)f(z).$$

By repeating the same argument several times in a row we find that

$$\begin{aligned} f(x) &= \mathbb{P}_x(T_A \geq 1) + \cdots + \mathbb{P}_x(T_A \geq t) \\ &\quad + \sum_{y_1 \notin A} \cdots \sum_{y_t \notin A} P(x, y_1)P(y_1, y_2) \cdots P(y_{t-1}, y_t)f(y_t). \end{aligned}$$

Because $f \geq 0$, this implies that

$$f(x) \geq \mathbb{P}_x(T_A \geq 1) + \cdots + \mathbb{P}_x(T_A \geq t)$$

¹For convenience we denote by \mathbb{P}_x and \mathbb{E}_x conditional probabilities and expectation given $X_0 = x$.

for all integers $t \geq 1$. Hence by taking $t \rightarrow \infty$ and applying Lemma 4.3 below, we find that

$$f(x) \geq \sum_{t=1}^{\infty} \mathbb{P}_x(T_A \geq t) = \mathbb{E}_x T_A = k_A(x).$$

□

Lemma 4.3. *Any random variable X taking on values in the set $\mathbb{Z}_+ \cup \{\infty\} = \{0, 1, 2, \dots, \infty\}$ satisfies*

$$\mathbb{E}X = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x). \quad (4.6)$$

Proof. If $\mathbb{P}(X = \infty) = 0$, then by changing the summing order of the nonnegative sums we see that

$$\sum_{x=1}^{\infty} \mathbb{P}(X \geq x) = \sum_{x=1}^{\infty} \sum_{y=x}^{\infty} \mathbb{P}(X = y) = \sum_{y=1}^{\infty} \sum_{x=1}^y \mathbb{P}(X = y) = \sum_{y=1}^{\infty} y \mathbb{P}(X = y) = \mathbb{E}X.$$

On the other hand, if $\mathbb{P}(X = \infty) > 0$, then

$$\sum_{x=1}^{\infty} \mathbb{P}(X \geq x) \geq \sum_{x=1}^{\infty} \mathbb{P}(X = \infty) = \infty.$$

Because $\mathbb{E}X = \infty$ whenever $\mathbb{P}(X = \infty) > 0$, the claim is also true when $\mathbb{P}(X = \infty) > 0$. □

4.2 Hitting probabilities

Consider a Markov chain on a finite state space S with transition matrix P . Select a nonempty set of states $A \subset S$. An irreducible chain will surely visit every state, but a reducible chain might not. What is the probability that a chain starting at x eventually visits A ? Let us denote this probability by

$$h_A(x) = \mathbb{P}(X_t \in A \text{ for some } t \geq 0 \mid X_0 = x). \quad (4.7)$$

This is called the *hitting probability* (*osumatodennäköisyys*) of the set A from initial state x .

Theorem 4.4. *The vector of hitting probabilities $h_A = (h_A(x) : x \in S)$ is the smallest nonnegative solution to the system of equations*

$$\begin{aligned} f(x) &= \sum_{y \in S} P(x, y) f(y), & x \notin A, \\ f(x) &= 1, & x \in A. \end{aligned} \quad (4.8)$$

Similarly as with expected passage times also the above system of equations can be interpreted in harmonic analytic terms as a Poisson equation

$$Df(x) = 0, \quad x \in B, \quad (4.9)$$

with boundary condition

$$f(x) = 1, \quad x \in \partial B,$$

when we denote $D = P - I$, $B = A^c$ and $\partial B = A$. The Poisson equation (4.9) with the right side being zero is in general called a Laplace equation. Before proving the theorem, let us see how it can be applied.

Example 4.5 (Human resource management). Consider the company describe in Example 4.2. What is the probability that a freshly recruited new employee eventually becomes a partner in the company?

This answer is the hitting probability $h_A(1)$ of the set $A = \{3\}$ from initial state $X_0 = 1$. The system of equations (4.8) is now of the form

$$f(x) = \sum_{y=0}^3 P(x, y)f(y), \quad x = 0, 1, 2,$$

$$f(3) = 1,$$

and for the transition matrix in (4.3) this corresponds to the equations

$$f(0) = f(0),$$

$$f(1) = 0.020 f(0) + 0.950 f(1) + 0.030 f(2),$$

$$f(2) = 0.008 f(0) + 0.982 f(2) + 0.010 f(3),$$

$$f(3) = 1.$$

Because there is no access from state 0 to state 3, we know that $f(0) = 0$. In light of this we may solve the other equations to obtain $f = [0, 0.333, 0.556, 1]$. It is not hard to verify that this f is the smallest nonnegative solution to the system of equations. By Theorem 4.4, this solution equals $f = h_A$. Hence the probability that a freshly recruited junior eventually becomes a partner equals $f(1) = k_A(1) = 0.333$. Note that the entries of f do not sum into one, even though they are probabilities. (Not all vectors of probabilities represent probability distributions.) ■

Proof of Theorem 4.4. This proof follows the same line of thought as the proof of Theorem 4.1. Let us first verify that the hitting probabilities satisfy the equations (4.8). Again we denote conditional probabilities given $X_0 = x$ by \mathbb{P}_x . Then $h_A(x) = \mathbb{P}_x(T_A < \infty)$, where T_A is the passage time of the chain into set A . If the initial state $x \in A$, then the chain surely visits A , so that $h_A(x) = 1$. Assume next that $x \notin A$. Then by applying the Markov property we may conclude that

$$\begin{aligned} \mathbb{P}_x(T_A < \infty | X_1 = y) &= \mathbb{P}(T_A < \infty | X_1 = y, X_0 = x) \\ &= \mathbb{P}(T_A < \infty | X_1 = y) \\ &= h_A(y), \end{aligned}$$

so that

$$\begin{aligned}
h_A(x) &= \mathbb{P}_x(T_A < \infty) \\
&= \sum_{y \in S} \mathbb{P}_x(X_1 = y) \mathbb{P}_x(T_A < \infty \mid X_1 = y) \\
&= \sum_{y \in S} P(x, y) h_A(y).
\end{aligned}$$

Hence $(h_A(x) : x \in S)$ is a nonnegative solution to (4.8).

Assume next that $f = (f(x) : x \in S)$ is some nonnegative solution to (4.8) and let us show that then $f(x) \geq h_A(x)$ for all x . Now obviously $f(x) = h_A(x) = 1$ for all $x \in A$. If $x \notin A$, then

$$\begin{aligned}
f(x) &= \sum_{y \in S} P(x, y) f(y) \\
&= \sum_{y \in A} P(x, y) + \sum_{y \notin A} P(x, y) f(y) \\
&= \mathbb{P}_x(X_1 \in A) + \sum_{y \notin A} P(x, y) f(y).
\end{aligned}$$

By substituting the formula of $f(y)$ to the right side above we see that

$$\begin{aligned}
f(x) &= \mathbb{P}_x(X_1 \in A) + \sum_{y \notin A} P(x, y) \left(\sum_{z \in A} P(y, z) + \sum_{z \notin A} P(y, z) f(z) \right) \\
&= \mathbb{P}_x(X_1 \in A) + \mathbb{P}_x(X_1 \notin A, X_2 \in A) + \sum_{y \notin A} \sum_{z \notin A} P(x, y) P(y, z) f(z) \\
&= \mathbb{P}_x(T_A = 1) + \mathbb{P}_x(T_A = 2) + \sum_{y \notin A} \sum_{z \notin A} P(x, y) P(y, z) f(z).
\end{aligned}$$

By iterating this argument we find that

$$\begin{aligned}
f(x) &= \mathbb{P}_x(T_A = 1) + \cdots + \mathbb{P}_x(T_A = t) \\
&\quad + \sum_{y_1 \notin A} \cdots \sum_{y_t \notin A} P(x, y_1) P(y_1, y_2) \cdots P(y_{t-1}, y_t) f(y_t).
\end{aligned}$$

Because $f \geq 0$, this implies that

$$f(x) \geq \mathbb{P}_x(T_A = 1) + \cdots + \mathbb{P}_x(T_A = t)$$

for all integers $t \geq 1$, so by taking $t \rightarrow \infty$ above we conclude that

$$f(x) \geq \sum_{t=1}^{\infty} \mathbb{P}_x(T_A = t) = \mathbb{P}_x(T_A < \infty) = h_A(x).$$

□

4.3 Gambler's ruin

Consider a random walk on state space $S = \{0, 1, \dots, M\}$ which moves up with probability q and down with probability $1 - q$, and gets absorbed at the boundary states 0 and M . This is a Markov chain with transition probabilities $P(x, x + 1) = q$ and $P(x, x - 1) = 1 - q$ for $1 \leq x \leq M - 1$, together with $P(0, 0) = 1$ and $P(M, M) = 1$, and all other transition probabilities being zero, see Figure 4.1.

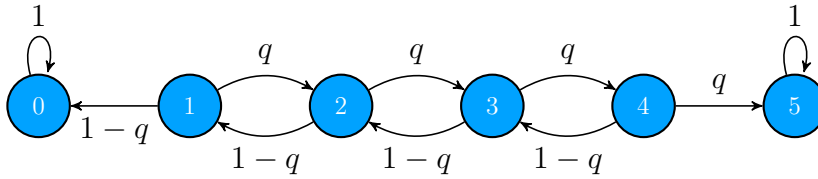


Figure 4.1: Transition diagram of a random walk with $M = 5$.

In a gambling context, the associated Markov chain X_t represents the wealth of a gambler after t rounds in a game where the gambler wins 1 EUR with probability q and loses 1 EUR with probability $1 - q$. The game stops if the wealth hits the value M (gambler's target) or the value 0 (gambler's money is all gone). A basic question here is to determine the probability of the gambler hitting the target, given that the initial wealth equals x . That is, we wish to compute the probability

$$h(x) = \mathbb{P}(X_t = M \text{ for some } t \geq 0 \mid X_0 = x).$$

Because the chain surely eventually hits either 0 or M , we see that the probability of the gambler's eventual ruin equals $1 - h(x)$.

The probability $h(x)$ equals the hitting probability $h_A(x)$ defined in (4.7) for the singleton set $A = \{M\}$. Hence by Theorem 4.4 the function $h(x)$ is the minimal nonnegative solution to the system of equations (4.8) which in this take the form

$$\begin{aligned} h(0) &= h(0), \\ h(x) &= (1 - q)h(x - 1) + qh(x + 1), \quad 0 < x < M, \\ h(M) &= 1. \end{aligned}$$

The first equation above tells us nothing, but the problem formulation makes it clear that $h(0) = 0$. Hence we are left with finding the minimal nonnegative solution to the equation

$$h(x) = (1 - q)h(x - 1) + qh(x + 1) \tag{4.10}$$

for $0 < x < M$, with boundary conditions $h(0) = 0$ and $h(M) = 1$.

Let us first solve $h(x)$ in the asymmetric case where $q \in (0, 1)$ is such that $q \neq \frac{1}{2}$. Formula (4.10) is a second-order homogeneous linear difference equation

for which make the ansatz $h(x) = z^x$ for some real number $z > 0$. Substituting this leads to

$$z^x = (1 - q)z^{x-1} + qz^{x+1},$$

and dividing both sides by z^{x-1} yields the quadratic equation

$$qz^2 - z + (1 - q) = 0$$

which has two distinct roots $\alpha = \frac{1-q}{q}$ and $\beta = 1$. By the theory of linear difference equations, we know that all solutions to (4.10) are of the form

$$h(x) = c\alpha^x + d\beta^x$$

for some constants c and d . The boundary conditions $h(0) = 0$ and $h(M) = 1$ now become

$$\begin{aligned} c + d &= 0, \\ c\alpha^M + d &= 1, \end{aligned}$$

from which we solve $d = -c$ and $c = 1/(\alpha^M - 1)$, and obtain the solution

$$h(x) = \frac{\alpha^x - 1}{\alpha^M - 1}. \quad (4.11)$$

To obtain the solution of (4.10) in the symmetric case with $q = \frac{1}{2}$, we may inspect the how the solution of (4.11) behaves as a function of q as $q \rightarrow \frac{1}{2}$. In this case $\alpha = \frac{1-q}{q} \rightarrow 1$, and by l'Hôpital's rule, it follows that

$$\frac{\alpha^x - 1}{\alpha^M - 1} \rightarrow \frac{x}{M}, \quad \text{as } \alpha \rightarrow 1.$$

This solution can also be derived by making an ansatz of the form $h(x) = c + dx$ and solving c and d from the boundary conditions. We may now formulate our findings as follows.

Theorem 4.6. *The probability that a random walk on $\{0, 1, \dots, M\}$ described in Figure 4.1 started at x eventually hits M equals*

$$h(x) = \begin{cases} \frac{\left(\frac{1-q}{q}\right)^x - 1}{\left(\frac{1-q}{q}\right)^M - 1}, & q \neq \frac{1}{2}, \\ \frac{x}{M}, & q = \frac{1}{2}. \end{cases}$$

The main message of Theorem 4.6 is that when $q \leq \frac{1}{2}$, the probability of ever reaching a state M from an initial state x tends to zero as $M \rightarrow \infty$. As an application related to gambling, consider the following example.

Example 4.7 (Roulette). In a game of roulette where a bet of 1 EUR is placed on the ball falling into one of 18 red pockets out of 37 pockets, the probability of winning 1 EUR is $q = \frac{18}{37}$ and the probability of losing 1 EUR is $1 - q$. If a gambler targets to double his initial wealth x , then the probability $h(x)$ of successfully ending the game is obtained by applying Theorem 4.6 with $M = 2x$, see Table 4.1. ■

Initial wealth (EUR)	1	5	10	20	50
Success probability	0.4865	0.4328	0.3680	0.2533	0.0628

Table 4.1: Probability of successfully doubling the initial wealth in a game of roulette by betting 1 EUR on red.

Chapter 5

General Markov chains and random walks

5.1 Infinite vectors and matrices

We will now study random processes with values in a general countable (finite or countably infinite) state space S . The assumption that S is *countable* (*numeroituvu*) means that its elements can be numbered using positive integers according to $S = \{x_1, x_2, \dots\}$, or equivalently, there exists a surjection from the set of natural numbers onto S .

Example 5.1. The following sets can be shown to be countably infinite:

- The set of integers \mathbb{Z} and the set of rational numbers \mathbb{Q} .
- The set \mathbb{Z}^d of vectors (x_1, \dots, x_d) with integer coordinates.
- The set of finite strings composed of letters from a finite alphabet.

The following sets can be shown to be uncountably infinite:

- The set of real numbers \mathbb{R} and the set of complex numbers \mathbb{C} .
- The interval $[0, 1]$ of real numbers.
- The set of infinite binary sequences $x = (x_1, x_2, \dots)$ with $x_i \in \{0, 1\}$.

The sum of a nonnegative function f on a countably infinite space $S = \{x_1, x_2, \dots\}$ is defined by

$$\sum_{x \in S} f(x) = \sum_{i=1}^{\infty} f(x_i) = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i).$$

The theory of nonnegative sums tells that the value of the sum does not depend on how the elements of S are labelled. A probability distribution on S is a

function $\mu : S \rightarrow [0, 1]$ such that

$$\sum_{x \in S} \mu(x) = 1. \quad (5.1)$$

In the context of Markov chains, a standard way is to interpret a probability distribution $\mu = (\mu(x) : x \in S)$ as a row vector indexed by the states.

A *transition matrix* (*siirtymämatrissi*) is a function $P : S \times S \rightarrow [0, 1]$ such that

$$\sum_{y \in S} P(x, y) = 1 \quad \text{for all } x \in S,$$

which means that the row sums of the (infinite) square matrix P are one. Matrix multiplication with infinite matrices is defined in the same way as in the finite case. If μ is a probability distribution on S we define μP by the formula

$$\mu P(y) = \sum_{x \in S} \mu(x) P(x, y), \quad y \in S.$$

Then $\mu P(y) \geq 0$ for all $y \in S$. Moreover, by changing the order of summation (which is always allowed when the terms are nonnegative), we see that

$$\sum_{y \in S} \mu P(y) = \sum_{y \in S} \sum_{x \in S} \mu(x) P(x, y) = \sum_{x \in S} \mu(x) \left(\sum_{y \in S} P(x, y) \right) = 1,$$

so that μP is a probability distribution on S .

The matrix product $R = PQ$ of transition matrices $P, Q : S \times S \rightarrow [0, 1]$ is defined by

$$R(x, z) = \sum_{y \in S} P(x, y) Q(y, z), \quad x, z \in S.$$

Then $R(x, z) \geq 0$ for all x, z . By changing the order of summation we find that

$$\sum_{z \in S} R(x, z) = \sum_{z \in S} \sum_{y \in S} P(x, y) Q(y, z) = \sum_{y \in S} P(x, y) \sum_{z \in S} Q(y, z) = 1.$$

Hence the product of two transition matrices is again a transition matrix. Matrix powers are defined in the usual way as $P^0 = I$ and recursively $P^{t+1} = P^t P$ for $t \geq 0$, where the identity matrix $I : S \times S \rightarrow [0, 1]$ is given by

$$I(x, y) = \begin{cases} 1, & x = y, \\ 0, & x \neq y. \end{cases}$$

5.2 Markov chains

A Markov chain with transition matrix P on a countable state space S is an S -valued random sequence (X_0, X_1, \dots) defined on some probability space (Ω, \mathbb{P}) such that

$$\mathbb{P}(X_{t+1} = y \mid X_t = x, H_{t-}) = P(x, y)$$

for all $x, y \in S$, all $t \geq 0$, and all events $H_{t-} = \{X_0 = x_0, \dots, X_{t-1} = x_{t-1}\}$ such that $\mathbb{P}(X_t = x, H_{t-}) > 0$. This is precisely the same definition as (1.1) in Section 1. The only difference is that for countably infinite state spaces, the transition matrix P has infinitely many rows and columns. We can view the infinite transition matrix as a function which maps a pair of states (x, y) into the probability $P(x, y) = \mathbb{P}(X_{t+1} = y | X_t = x)$.

Theorem 5.2. *The distribution $\mu_t(x) = \mathbb{P}(X_t = x)$ of a Markov chain at time t can be computed using the initial distribution μ_0 and the transition matrix P as*

$$\mu_t = \mu_0 P^t, \quad (5.2)$$

where P^t on is the t -th power of P . Moreover,

$$\mathbb{P}(X_t = y | X_0 = x) = P^t(x, y).$$

Proof. The proofs of Theorems 1.5 and 1.7 work also for countably infinite state spaces. \square

5.3 Long-term behaviour

The long-term analysis of Markov chains on infinite state spaces has one fundamental difference compared to chains on finite spaces: *irreducibility does not guarantee the existence of an invariant distribution*. Every irreducible Markov chain in a finite state space visits all states infinitely often with probability one. In infinite spaces this may or may not be the case. To understand this, a key quantity is the probability

$$\rho(x, y) = \mathbb{P}(X_t = y \text{ for some } t \geq 1 | X_0 = x),$$

that a Markov chain started at state x visits state y at some future time instant. The quantity $\rho(x, x)$ is called the *return probability* (*paluutodennäköisyys*) of x . A state is called *recurrent* (*palautuva*) if it has return probability one, and *transient* (*väistynä*) otherwise.

Theorem 5.3. *If an irreducible Markov chain on a countable state space S has an invariant distribution π , then*

$$\pi(y) > 0 \quad \text{for all } y \in S, \quad (5.3)$$

all states are recurrent, and with probability one, the chain visits every state infinitely often, regardless of the initial state.

The proof of Theorem 5.3 utilizes the following auxiliary result.

Lemma 5.4. *If x is recurrent, then $\rho(y, x) = 1$ for all states y which are reachable from x .*

Proof. Let $t \geq 0$ be the length of the shortest path from x to y in the transition diagram of the chain. Then the transition diagram contains a t -hop path $x = x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_t = y$ which is such that x does not belong to $\{x_1, \dots, x_t\}$. By the Markov property, the probability that a chain started at x never returns to x is bounded by

$$1 - \rho(x, x) \geq P(x_0, x_1)P(x_1, x_2) \cdots P(x_{t-1}, x_t)(1 - \rho(y, x)).$$

Because $\rho(x, x) = 1$ and $P(x_0, x_1)P(x_1, x_2) \cdots P(x_{t-1}, x_t) > 0$, the above inequality implies that $\rho(y, x) = 1$. \square

Proof of Theorem 5.3. Let us first verify (5.3). Because $\sum_x \pi(x) = 1$, we can choose a state x_0 such that $\pi(x_0) > 0$. By irreducibility, the chain can move from x_0 to y via some path of length $t \geq 0$, so that $P^t(x_0, y) > 0$. Because $\pi P = \pi$, we also have $\pi P^t = \pi$, so that

$$\pi(y) = \sum_{x \in S} \pi(x)P^t(x, y) \geq \pi(x_0)P^t(x_0, y) > 0,$$

and hence (5.3) holds.

Let us study the event A_y that the chain visits state y , but only finitely many times. This event can be written as a disjoint union $A_y = \cup_{0 \leq t < \infty} A_{y,t}$, where

$$A_{y,t} = \{X_t = y, X_{t+1} \neq y, X_{t+2} \neq y, \dots\}$$

is the event that t is the last time instant at which the chain visits y . By Markov property, it follows that

$$\begin{aligned} \mathbb{P}(A_{y,t}) &= \mathbb{P}(X_t = y) \mathbb{P}(X_{t+1} \neq y, X_{t+2} \neq y, \dots \mid X_t = y) \\ &= \mathbb{P}(X_t = y) \mathbb{P}(X_1 \neq y, X_2 \neq y, \dots \mid X_0 = y) \\ &= \mathbb{P}(X_t = y)(1 - \rho(y, y)). \end{aligned} \tag{5.4}$$

The above equation holds for any initial distribution of the chain. Especially, if we denote by \mathbb{P}_π the distribution of the Markov chain corresponding to the initial distribution $\mu_0 = \pi$, then it follows that

$$\mathbb{P}_\pi(A_{y,t}) = \pi(y)(1 - \rho(y, y)),$$

and by summing both sides over t , we see that

$$\mathbb{P}_\pi(A_y) = \sum_{t=0}^{\infty} \mathbb{P}_\pi(A_{y,t}) = \sum_{t=0}^{\infty} \pi(y)(1 - \rho(y, y)).$$

Because terms of the sum on the right do not depend on t , we must have $\pi(y)(1 - \rho(y, y)) = 0$. Furthermore, by (5.3), $\pi(y) > 0$, so we conclude that $\rho(y, y) = 1$. Hence all states are recurrent.

Now let U_y be the event that the chain visits state y infinitely many times. The complement of this can be written as $U_y^c = A_y \cup B_y$ where B_y is the event

that the chain never visits y . Because $\rho(y, y) = 1$, equation (5.4) implies that $\mathbb{P}(A_{y,t}) = 0$ for all t , and therefore

$$\mathbb{P}(A_y) = \sum_{t=0}^{\infty} \mathbb{P}(A_{y,t}) = 0$$

regardless of the initial state of the chain. Now by Lemma 5.4, it follows that $\rho(x, y) = 1$ for all x, y . Therefore,

$$\begin{aligned} \mathbb{P}(B_y) &= \sum_{x \neq y} \mathbb{P}(X_0 = x) \mathbb{P}(B_y | X_0 = x) \\ &= \sum_{x \neq y} \mathbb{P}(X_0 = x) (1 - \rho(x, y)) \\ &= 0. \end{aligned}$$

Hence $\mathbb{P}(U_y^c) \leq \mathbb{P}(A_y) + \mathbb{P}(B_y) \leq 0$ implies $\mathbb{P}(U_y^c) = 0$. Finally, if U is the event that the chain visits every state infinitely often, then by the general union bound,

$$\mathbb{P}(U^c) = \mathbb{P}(\cup_y U_y^c) \leq \sum_y \mathbb{P}(U_y^c) = 0,$$

and we conclude that $\mathbb{P}(U) = 1$. □

5.4 Convergence theorem

Theorem 5.5. *Let (X_t) be an irreducible and aperiodic Markov chain, and assume that it has an invariant distribution π . Then the invariant distribution is unique and*

$$\lim_{t \rightarrow \infty} \mathbb{P}(X_t = y | X_0 = x) = \pi(y)$$

for all $x, y \in S$.

The above result can be rewritten as

$$P^t(x, y) \rightarrow \pi(y)$$

which in matrix terms means that each row of P^t converges to the row vector π entrywise. An equivalent statement is that $\mu_t \rightarrow \pi$ pointwise, regardless of the initial distribution μ_0 of the chain. One more equivalent (though not completely trivial) statement is that $\mu_t \rightarrow \pi$ in the total variation distance.

Proof. Let (X_t) and (Y_t) be independent Markov chains both having transition matrix P , and such that (X_t) has initial distribution μ and (Y_t) has initial distribution ν . Let

$$\tau = \min\{t \geq 0 : X_t = Y_t\}$$

be the first time instant (possibly ∞) at which the paths of the Markov chains meet each other. Observe next by conditioning on the possible values X_s that for any $s \leq t$,

$$\begin{aligned}\mathbb{P}(X_t = y, \tau = s) &= \sum_x \mathbb{P}(\tau = s, X_s = x, X_t = y) \\ &= \sum_x \mathbb{P}(\tau = s, X_s = x) \mathbb{P}(X_t = y | \tau = s, X_s = x).\end{aligned}$$

Observe next that whether or not $\tau = s$ occurs can be detected using a deterministic function of random vectors (X_0, \dots, X_s) and (Y_0, \dots, Y_s) , the latter being independent of (X_t) . Therefore, Markov property implies that

$$\mathbb{P}(X_t = y | \tau = s, X_s = x) = \mathbb{P}(X_t = y | X_s = x).$$

Furthermore, by the definition of τ , we see that

$$\mathbb{P}(\tau = s, X_s = x) = \mathbb{P}(\tau = s, Y_s = x).$$

Hence, by symmetry,

$$\begin{aligned}\mathbb{P}(X_t = y, \tau = s) &= \sum_x \mathbb{P}(\tau = s, X_s = x) \mathbb{P}(X_t = y | X_s = x) \\ &= \sum_x \mathbb{P}(\tau = s, Y_s = x) \mathbb{P}(Y_t = y | Y_s = x) \\ &= \mathbb{P}(Y_t = y, \tau = s).\end{aligned}$$

By summing the above equation over $s \leq t$, it follows that

$$\mathbb{P}(X_t = y, \tau \leq t) = \mathbb{P}(Y_t = y, \tau \leq t).$$

This implies that

$$\begin{aligned}\sum_y |\mathbb{P}(X_t = y) - \mathbb{P}(Y_t = y)| &= \sum_y |\mathbb{P}(X_t = y, \tau > t) - \mathbb{P}(Y_t = y, \tau > t)| \\ &\leq \sum_y \mathbb{P}(X_t = y, \tau > t) + \sum_y \mathbb{P}(Y_t = y, \tau > t) \\ &= 2\mathbb{P}(\tau > t).\end{aligned}$$

When (X_t) is started at x and (Y_t) is started at a random initial state distributed according to the invariant distribution π , this becomes

$$\sum_y |P^t(x, y) - \pi(y)| \leq 2\mathbb{P}(\tau > t).$$

To finish the proof, it suffices to show that $\mathbb{P}(\tau > t) \rightarrow 0$ as $t \rightarrow \infty$, which is equivalent to showing that $\mathbb{P}(\tau < \infty) = 1$. To do this, note that

$\{(X_t, Y_t) : t \geq 0\}$ is a Markov chain on the product space $S \times S$, with transition matrix \tilde{P} defined by

$$\tilde{P}((x_1, x_2), (y_1, y_2)) = P(x_1, y_1)P(x_2, y_2).$$

Furthermore, it is easy to verify that $\tilde{\pi}(x, y) = \pi(x)\pi(y)$ is an invariant distribution of \tilde{P} . It is also possible to show that \tilde{P} is irreducible (here we need the irreducibility and aperiodicity of P). In terms of the product chain (X_t, Y_t) , we see that τ is the first hitting time T_D of the product chain into the diagonal $D = \{(x, y) \in S \times S : x = y\}$, which is bounded from above by $T_D \leq T_{(x,x)}$ for any $x \in S$. By Theorem 5.3, $T_{(x,x)}$ is finite with probability one, and hence so is $\tau = T_D$. \square

5.5 Reversibility

A transition matrix P and a corresponding Markov chain is called *reversible* (*käänttyvä*) with respect to a probability distribution π if the following *detailed balance equations* (*pareittaiset tasapainoyhtälöt*)

$$\pi(x)P(x, y) = \pi(y)P(y, x) \tag{5.5}$$

are valid for all $x, y \in S$.

Theorem 5.6. *If P is reversible with respect to π , then π is an invariant distribution of P .*

Proof. If (5.5) holds, then for all $y \in S$,

$$\sum_{x \in S} \pi(x)P(x, y) = \sum_{x \in S} \pi(y)P(y, x) = \pi(y) \sum_{x \in S} P(y, x) = \pi(y).$$

Hence $\pi P = \pi$. \square

Reversibility can be interpreted as follows. Let (X_0, X_1, \dots) be a Markov chain with transition matrix P which is reversible with respect to π such that X_0 (and hence every X_t) is π -distributed. By applying the detailed balance equations (5.5) we then find that

$$\begin{aligned} \mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_t = x_t) &= \pi(x_0)P(x_0, x_1)P(x_1, x_2) \cdots P(x_{t-1}, x_t) \\ &= P(x_1, x_0)\pi(x_1)P(x_1, x_2) \cdots P(x_{t-1}, x_t) \\ &= \dots \\ &= P(x_1, x_0)P(x_2, x_1) \cdots P(x_t, x_{t-1})\pi(x_t) \\ &= \pi(x_t)P(x_t, x_{t-1}) \cdots P(x_1, x_0) \\ &= \mathbb{P}(X_t = x_0, X_{t-1} = x_1, \dots, X_0 = x_t). \end{aligned}$$

From this we may conclude that a π -reversible chain with initial distribution π appears statistically the same if observed backwards in time.

An important class of reversible Markov chains is discussed next. A *birth–death chain* (*syntymiskuoolemisketju*) is a Markov chain on a state space $S \subset \mathbb{Z}_+$ with a transition matrix such that $P(x, y) = 0$ for $|x - y| > 1$. Hence a birth–death can only move to its nearby states. Examples of birth–death chains include the gambler’s ruin (finite state space) and a random walk on \mathbb{Z}_+ , discussed soon.

Theorem 5.7. *If a birth–death chain has an invariant distribution π , then the chain is π -reversible.*

Proof. We need to verify that the detailed balance equation (5.5) holds for all $x, y \in S$. If $x = y$, then (5.5) is trivially true. The same conclusion is true also when $|x - y| > 1$ because in this case both sides of (5.5) are zero. Hence the only case that we need to investigate is the one where we assume that $x, y \in S$ are such that $y = x + 1$. In this case the balance equation $\pi = \pi P$ at v implies that

$$\pi(v) = \sum_u \pi(u)P(u, v)$$

and by summing over $v \in S$ such that $v \leq x$, we find that

$$\sum_{v \leq x} \pi(v) = \sum_u \pi(u) \sum_{v \leq x} P(u, v). \quad (5.6)$$

Now because the birth–death chain may only makes jumps of length zero or one,

$$\sum_{v \leq x} P(u, v) = \begin{cases} 1, & u \leq x - 1, \\ 1 - P(x, x + 1), & u = x, \\ P(x + 1, x), & u = x + 1, \\ 0, & u \geq x + 2. \end{cases}$$

Hence (5.6) can be written in the form

$$\sum_{v \leq x} \pi(v) = \sum_{u \leq x-1} \pi(u) + \pi(x)(1 - P(x, x + 1)) + \pi(x + 1)P(x + 1, x).$$

Now because $\sum_{v \leq x} \pi(v) = \pi(x) + \sum_{u \leq x-1} \pi(u)$, this implies

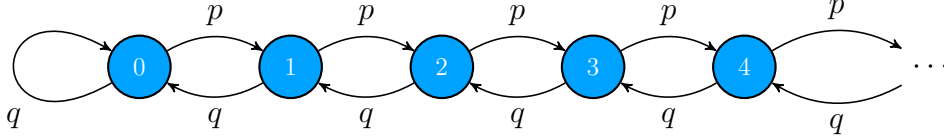
$$\pi(x)P(x, x + 1) = \pi(x + 1)P(x + 1, x),$$

so that (5.5) holds for $y = x + 1$. □

5.6 Random walk on the nonnegative integers

An irreducible and aperiodic Markov chain on a *finite* state space always has a unique invariant distribution π , and the distribution of X_t converges to π as $t \rightarrow \infty$ regardless of the initial state. In the context of infinite state spaces this does *not* hold in general.

A particle moves in the infinite set $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ so that at every time step the particle moves from state $x \geq 1$ to the right with probability p and to the left with probability $q = 1 - p$, independently of the past steps. With the boundary condition $P(0, 0) = q$, we get the transition diagram



and the infinite transition matrix

$$P = \begin{bmatrix} 1-p & p & 0 & \cdots & & & & \\ q & 0 & p & 0 & \cdots & & & \\ 0 & q & 0 & p & 0 & \cdots & & \\ 0 & 0 & q & 0 & p & 0 & \cdots & \\ 0 & 0 & 0 & q & 0 & p & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \quad (5.7)$$

From the transition diagram we see that the chain is irreducible for all $p \in (0, 1)$. In addition, $P(0, 0) > 0$ implies that chain is aperiodic.

Let us next study whether or not this random walk has an invariant distribution. The random walk is an instance of a birth–death chain, so that by Theorem 5.7, any possible invariant distribution π of P must satisfy the detailed balance equations (5.5) which in this case can be written as

$$\pi(x)P(x, x+1) = \pi(x+1)P(x+1, x), \quad x \geq 0,$$

or equivalently,

$$p\pi(x) = q\pi(x+1).$$

From this we find that $\pi(1) = \pi(0)\frac{p}{q}$ and $\pi(2) = \pi(0)(\frac{p}{q})^2$, and in general,

$$\pi(x) = \left(\frac{p}{q}\right)^x \pi(0), \quad x \geq 0.$$

For this to be a probability distribution, we need to have $\sum_x \pi(x) = 1$. If $p < q$, or equivalently $p < \frac{1}{2}$, this normalisation is possible by choosing $\pi(0) = 1 - \frac{p}{q}$. If $p \geq \frac{1}{2}$ this is not possible. We conclude that

- For $p < \frac{1}{2}$, the unique invariant distribution of the chain is the geometric distribution $\pi(x) = (1 - \frac{p}{q})(\frac{p}{q})^x$ on $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$.
- For $p \geq \frac{1}{2}$ the chain does *not* have an invariant distribution.

Case	Irreducible	Aperiodic	Recurrent	Invariant distribution
$p \in (0, \frac{1}{2})$	Yes	Yes	Yes	Yes (unique)
$p = \frac{1}{2}$	Yes	Yes	Yes	Does not exist
$p \in (\frac{1}{2}, 1)$	Yes	Yes	No	Does not exist

Table 5.1: Properties of the random walk on \mathbb{Z}_+ defined by (5.7).

Let us now investigate how the random walk behaves when $p \geq \frac{1}{2}$. We study the question whether or not the chain ever returns to state 0 after leaving it. The probability that the chain ever returns to 0 can be written as

$$\mathbb{P}_1(T_0 < \infty) = \lim_{M \rightarrow \infty} \mathbb{P}_1(T_0 < T_M)$$

where T_x denotes the first hitting time into state x , and \mathbb{P}_1 refers to the distribution of the random walk started at state 1. Now $\mathbb{P}_1(T_0 < T_M)$ also equals a gambler's ruin probability with initial wealth 1 and target wealth M , so that by Theorem 4.6,

$$\mathbb{P}_1(T_0 < T_M) = \begin{cases} 1 - \frac{\left(\frac{1-p}{p}\right)^1 - 1}{\left(\frac{1-p}{p}\right)^M - 1}, & p \neq \frac{1}{2}, \\ 1 - \frac{x}{M}, & p = \frac{1}{2}. \end{cases}$$

Hence the probability that the chain returns to 0 after leaving it equals

$$\mathbb{P}_1(T_0 < \infty) = \begin{cases} 1, & p < \frac{1}{2}, \\ 1, & p = \frac{1}{2}, \\ \frac{1-p}{p}, & p > \frac{1}{2}. \end{cases}$$

This means that the states of the chain are recurrent for $p \leq \frac{1}{2}$ and transient for $p > \frac{1}{2}$. The case $p = \frac{1}{2}$ is special in that although the chain eventually returns to every state, one can show that expected return time is infinite. Table 5.1 summarizes key properties of the random walk. Figure 5.1 describes paths of the random walk simulated using the code below.

```
# R-code for simulating a path of a random walk
T <- 1000 # Number of time steps
p <- 0.4 # Probability of moving right
X0 <- 0 # Initial state
X <- integer(T+1)
X[1] <- X0
for (t in 1:T) {
  X[i,t+1] <- max(X[i,t] + 2*rbinom(1,1,p)-1, 0)
}
```

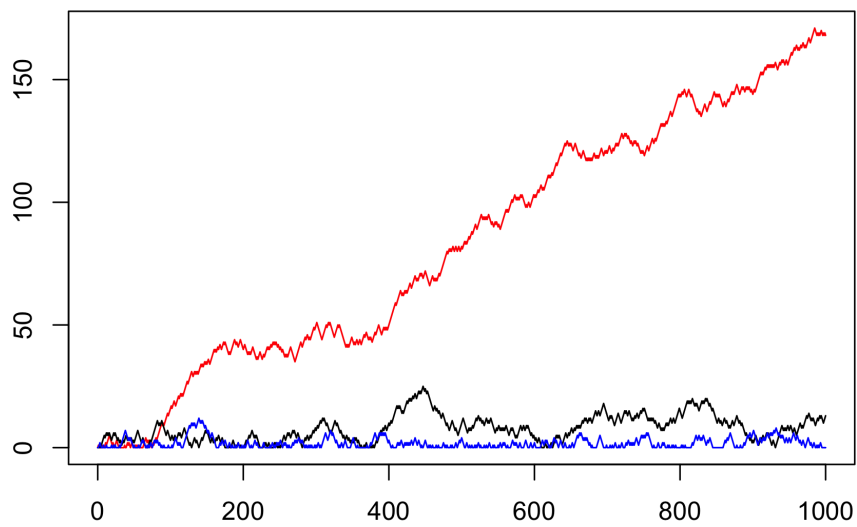


Figure 5.1: Simulated paths of the random walk on \mathbb{Z}_+ defined by (5.7) for $p = 0.4$ (blue), $p = 0.5$ (black), $p = 0.6$ (red).

Chapter 6

Branching processes

6.1 Transition matrix

A *branching process* (*haarautumisprosessi*) is a Markov chain (X_0, X_1, \dots) on state space $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ which models a population where each individual in generation t independently produces a random number of children, and these children form the next generation $t + 1$. The model is parametrised by an *offspring distribution* (*lisääntymisjakauma*) $p = (p(0), p(1), p(2), \dots)$ where the entry $p(k)$ equals the probability that an individual produces k children. The study of branching processes became popular after a question published by Francis Galton in 1873 which was later solved by Thomas Watson a couple of years later. This is why a branching process is often also called a *Galton–Watson process*. Branching processes are applied to several type of spreading phenomena. In epidemic modelling, the population refers to the infectious individuals, and producing children means transmitting a disease to others. In social sciences, the population may refer to people advocating an opinion, and producing children means communicating the opinion to others.

If there are $X_t = x$ individuals in generation t , then the size of generation $t + 1$ can be written as a sum

$$X_{t+1} = Y_1 + \dots + Y_x,$$

where Y_1, Y_2, \dots are independent p -distributed random integers. Hence the transition probability from state $x \geq 1$ to state $y \geq 0$ equals

$$P(x, y) = \mathbb{P}(Y_1 + \dots + Y_x = y). \quad (6.1)$$

If there are no individuals in generation t , then no children are born and hence also the next generation is empty. Therefore,

$$P(0, y) = \begin{cases} 1, & y = 0, \\ 0, & \text{else.} \end{cases} \quad (6.2)$$

State 0 is hence absorbing for the chain. When the chain enters 0, the population becomes extinct. Galton's question was:

What is the probability that a population eventually becomes extinct?

In other words, what is the hitting probability $\mathbb{P}(T_0 < \infty)$ of the chain into state zero?

6.2 Generating functions

After the offspring distribution p has been given, formulas (6.1)–(6.2) uniquely determine the entries of a infinite transition matrix P with rows and columns indexed by \mathbb{Z}_+ . The only problem is that computing numerical values of the entries of P can be difficult from (6.1). For example, to determine the entry $P(3, 7)$ requires computing the sum

$$P(3, 7) = \sum_{y_1} \sum_{y_2} \sum_{y_3} 1(y_1 + y_2 + y_3 = 7) p(y_1)p(y_2)p(y_3).$$

Generating functions provide a powerful tool for treating such formulas. The *probability generating function* (*todennäköisyydet generoiva funktio*) of a random integer $Y \in \mathbb{Z}_+$ distributed according $\mathbb{P}(Y = k) = p(k)$ is defined by

$$\phi_Y(s) = \mathbb{E}s^Y = \sum_{k=0}^{\infty} s^k p(k) \tag{6.3}$$

for those value of s for which the sum on the right converges. The probability generating function is always defined for $s \in [-1, 1]$. It is also defined for other values of s if the probabilities $p(k)$ vanish quickly enough for large values of k . The values of ϕ_Y on $[-1, 1]$ determine the probability distribution of Y uniquely, because the convergence radius of the power series on the right side of (6.3) is always at least 1, and therefore the above series can be differentiated infinitely many times term by term at every point in $(-1, 1)$. By differentiating ϕ_Y k times at zero we find that

$$\mathbb{P}(Y = k) = p(k) = \frac{\phi_Y^{(k)}(0)}{k!}, \quad k = 0, 1, 2, \dots$$

The key usefulness of generating functions is that they behave well for sums of independent random variables. Namely, if X and Y are independent \mathbb{Z}_+ -valued random integers, then

$$\phi_{X+Y}(s) = \mathbb{E}s^{X+Y} = \mathbb{E}s^X s^Y = \mathbb{E}s^X \mathbb{E}s^Y = \phi_X(s)\phi_Y(s).$$

The above formula readily extends to multiple independent summands. Especially, for any independent and identically distributed random integers $Y_1, Y_2, \dots, Y_n \geq 0$ we have

$$\phi_{Y_1+\dots+Y_n}(s) = \phi_{Y_1}(s)^n. \tag{6.4}$$

Hence for example the element $P(3, 7)$ of the transition matrix can be computed by writing $\phi_{Y_1}(s)^3$ as a power series and finding out the term corresponding to s^7 . This can also be done by differentiating $\phi_{Y_1}(s)^3$ seven times at zero and dividing the outcome by the factorial of 7.

The following result generalizes (6.4) to the case where also the number of summands is a random variable. (An empty sum $\sum_{k=1}^0 Y_k$ is defined as zero in the formula below.)

Theorem 6.1. *If N, Y_1, Y_2, \dots are independent \mathbb{Z}_+ -valued random numbers, and Y_1, Y_2, \dots are identically distributed, then the probability generating function of*

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

is obtained by $\phi_Z(s) = \phi_N(\phi_{Y_1}(s))$.

Proof. By conditioning on the possible values of N , and by applying independence and (6.4) we find that

$$\begin{aligned} \phi_Z(s) &= \sum_{n=0}^{\infty} \mathbb{P}(N = n) \mathbb{E} \left(s^{\sum_{k=1}^n Y_k} \mid N = n \right) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(N = n) \mathbb{E} \left(s^{\sum_{k=1}^n Y_k} \right) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(N = n) \phi_{Y_1}(s)^n \\ &= \phi_N(\phi_{Y_1}(s)). \end{aligned}$$

□

6.3 Expected population size

The following result helps to compute the expected population size as a function of time for a branching process where

$$m = \mathbb{E}(Y_1)$$

is the expected number of children produced by an individual. As a consequence, we see that the population size tends to zero when $m < 1$ and grows exponentially fast to infinity when $m > 1$.

Theorem 6.2. *The expected size of generation t in a branching process started with x individuals is*

$$\mathbb{E}(X_t \mid X_0 = x) = xm^t, \quad t = 0, 1, 2, \dots$$

Proof. By conditioning on the event $X_t = y$, we find that

$$\mathbb{E}(X_{t+1} | X_t = y) = \mathbb{E}\left(\sum_{k=1}^{X_t} Y_k \mid X_t = y\right) = \sum_{k=1}^y \mathbb{E}(Y_k | X_t = y) = my,$$

where $m = \mathbb{E}Y_1$. By multiplying both sides by $\mathbb{P}(X_t = y)$ and then summing over y this implies that

$$\mathbb{E}(X_{t+1}) = \sum_{y=0}^{\infty} \mathbb{E}(X_{t+1} | X_t = y)\mathbb{P}(X_t = y) = \sum_{y=0}^{\infty} my\mathbb{P}(X_t = y) = m\mathbb{E}(X_t).$$

The now claim follows by induction. \square

6.4 Extinction probability

Let us get back to Galton's question: What is the probability of eventual extinction? Observe first that the evolution of descendants of any particular individual behaves as a branching process started with initial state one, and that the branches of the initial individuals are mutually independent. Therefore, if the initial generation contains $x \geq 1$ individuals, then the probability of eventual extinction is the probability of all individual family lines becoming extinct, and this probability equals

$$\mathbb{P}(\text{extinction} \mid X_0 = x) = \eta^x,$$

where $\eta = \mathbb{P}_1(T_0 < \infty)$ is the extinction probability of a branching process with initial size $X_0 = 1$. Furthermore, the extinction probability η can be obtained as a fixed point of the probability generating function of the offspring distribution, as the following result confirms.

Theorem 6.3. *The extinction probability of a branching process starting with one individual is the smallest nonnegative solution of*

$$\phi_{Y_1}(s) = s.$$

Example 6.4. During its lifetime, each individual produces two children with probability a and no children otherwise. What is the probability that the family line of a particular individual eventually becomes extinct?

The probability generating function of the offspring distribution is $\phi(s) = (1 - a) + as^2$, so the fixed points of ϕ are the solutions of $as^2 - s + (1 - a) = 0$ given by

$$s = \frac{1 \pm \sqrt{1 - 4a(1 - a)}}{2a} = \frac{1 \pm \sqrt{(1 - 2a)^2}}{2a} = \begin{cases} (1 - a)/a, \\ 1. \end{cases}$$

By Theorem 6.3, the extinction probability is hence

$$\eta = \begin{cases} 1, & \text{when } a \leq 1/2, \\ \frac{1-a}{a}, & \text{when } a > 1/2. \end{cases}$$

■

To prove Theorem 6.3 we need an auxiliary result which tells how the probability generating function of X_t can be computed using the probability generating function of the offspring distribution $\phi(s) = \phi_{Y_1}(s)$.

Lemma 6.5. *For a branching process started at $X_0 = 1$, the probability generating function of X_t is given by*

$$\phi_{X_t}(s) = \underbrace{\phi \circ \phi \circ \cdots \circ \phi}_t(s).$$

Proof. By definition $X_0 = 1$, so that

$$\phi_{X_0}(s) = s.$$

The individuals of generation $t + 1$ are the children of individuals of generation t , so that the size of generation $t + 1$ can be represented as

$$X_{t+1} = \sum_{x=1}^{X_t} Y_{t,x},$$

where $Y_{t,1}, Y_{t,2}$ are mutually independent and p -distributed, and independent of X_t . By Theorem 6.1 we see that

$$\phi_{X_{t+1}}(s) = \phi_{X_t}(\phi(s)), \quad t = 0, 1, 2, \dots$$

By substituting $t = 0$ to the above formula we find that $\phi_{X_1}(s) = \phi(s)$. By substituting $t = 1$ we see that

$$\phi_{X_2}(s) = \phi_{X_1}(\phi(s)) = \phi(\phi(s)).$$

By continuing this way, that is, by applying induction, the claim follows. □

Proof of Theorem 6.3. (i) Let us first verify that η is a fixed point of $\phi(s) = \phi_{Y_1}(s)$. We can write η as

$$\eta = \mathbb{P} \left(\bigcup_{t=1}^{\infty} \{X_t = 0\} \right)$$

and note that by the continuity of probability measures,

$$\mathbb{P} \left(\bigcup_{t=1}^{\infty} \{X_t = 0\} \right) = \lim_{t \rightarrow \infty} \mathbb{P} \left(\bigcup_{s=1}^t \{X_s = 0\} \right).$$

The above continuity property follows from general probability axioms and is discussed on more detailed in probability theory courses. Next we observe that

$$\bigcup_{s=1}^t \{X_s = 0\} = \{X_t = 0\},$$

because state 0 is absorbing. Hence we may write

$$\eta = \lim_{t \rightarrow \infty} \mathbb{P} \left(\bigcup_{s=1}^t \{X_s = 0\} \right) = \lim_{t \rightarrow \infty} \eta_t,$$

where $\eta_t = \mathbb{P}(X_t = 0)$ is the probability of extinction by time t .

By applying probability generating functions, we may write

$$\mathbb{P}(X_t = 0) = \phi_{X_t}(0),$$

and with the help of Lemma 6.5,

$$\phi_{X_t}(0) = \phi(\phi_{X_{t-1}}(0)).$$

Therefore,

$$\eta_t = \phi(\eta_{t-1}) \tag{6.5}$$

for all $t \geq 1$. Because η and η_t are probabilities, they belong to the interval $[0, 1]$. Being a convergent power series, the function ϕ is continuous on $[0, 1]$, and hence

$$\eta = \lim_{t \rightarrow \infty} \eta_t = \lim_{t \rightarrow \infty} \phi(\eta_{t-1}) = \phi(\lim_{t \rightarrow \infty} \eta_{t-1}) = \phi(\eta).$$

Hence η is a fixed point of ϕ .

(ii) We will now show that η is the smallest fixed point of ϕ in $[0, 1]$. To do this, let us assume that $a \in [0, 1]$ as an arbitrary fixed point of ϕ . We will show that $\eta \leq a$. First, because ϕ is nondecreasing on $[0, 1]$, and X_1 distributed according to Y_1 , we see that

$$\eta_1 = \mathbb{P}(X_1 = 0) = \phi(0) \leq \phi(a) = a.$$

Therefore $\eta_1 \leq a$. On the other hand, by applying (6.5) and the monotonicity of ϕ ,

$$\eta_2 = \phi(\eta_1) \leq \phi(a) = a.$$

Hence also $\eta_2 \leq a$. By proceeding this way we may conclude that $\eta_t \leq a$ for all $t \geq 1$. Especially,

$$\eta = \lim_{t \rightarrow \infty} \eta_t \leq a.$$

□

6.5 Sure extinction

Let us finally derive the following fundamental result. Here $m = \mathbb{E}(Y_1)$ is the expected number of children for an individual. The result tells that a branching process cannot ever reach a statistical equilibrium with a sustainable nonzero population size. Namely, the only case where the population does not become eventually extinct is the one with $m > 1$, in which case the population grows to infinity exponentially fast according to Theorem 6.2. This is sometimes called a Malthusian property, after an English scholar Thomas Malthus (1766–1834).

Theorem 6.6. *For every branching process such that $X_0 = 1$ and $p(0) > 0$,*

- $\eta = 1$, for $m \leq 1$.
- $\eta \in (0, 1)$, for $m > 1$.

Proof. Let us first note that $\phi(1) = 1$. Furthermore, it can be shown that ϕ is convex on the interval $[0, 1]$. In addition, the left derivative of ϕ at point 1 satisfies $\phi'(1-) = m$. If $m \leq 1$, then by sketching a plot of ϕ on the interval $[0, 1]$ we see that ϕ does not have any fixed points $[0, 1)$. Hence the smallest fixed point of ϕ on $[0, 1]$ is $\eta = 1$.

If $m > 1$, then again by plotting ϕ on the interval $[0, 1]$ we see that ϕ has precisely one fixed point on $(0, 1)$. This fixed point is the smallest on $[0, 1]$, and hence $\eta \in (0, 1)$. Instead of sketching the plots, the proofs can be made rigorous by carefully inspecting Taylor expansions of ϕ around zero and around one. \square

Chapter 7

Random point patterns and counting processes

7.1 Random point pattern

A *random point pattern* (*satunnainen pistekuvio*) on an interval $S \subset \mathbb{R}$ is a locally finite¹ random subset of S , defined on some probability space (Ω, \mathbb{P}) . A random point pattern is hence a map $\omega \mapsto X(\omega)$ from Ω to the family of locally finite subsets of S . For clarity, and following the usual convention in stochastics, the symbol ω is omitted in what follows.

Example 7.1. Let U_1, \dots, U_n be independent and uniformly distributed random numbers on the interval² $(0, 1)$. Then the set $X = \{U_1, \dots, U_n\}$ is a random point pattern on $(0, 1)$. ■

Example 7.2. Let Z be a random integer which follows a Poisson distribution with mean $\lambda > 0$. Then the set $X = \{n \in \mathbb{Z}_+ : n \leq Z\}$ is a random point pattern on \mathbb{R}_+ . ■

Precisely speaking, in the definition of a random point pattern we need to require that the map $X : \Omega \rightarrow \mathcal{N}(S)$ is measurable with respect to the sigma-algebra on $\mathcal{N}(S)$ generated by the maps $B \mapsto |X \cap B|$, $B \subset S$ open, where $\mathcal{N}(S)$ is the family of all locally finite sets subsets of S . Such technical details are unimportant in the analysis here, and hence not treated further. For details, see for example the books [Kal02, SW08].

7.2 Counting measure and counting process

The *counting measure* (*laskurimitta*) of a random point pattern X on $S \subset \mathbb{R}$ is a random function

$$N(B) = |X \cap B|,$$

¹A subset X of an interval S is *locally finite* (*lokaalisti äärellinen*) if $X \cap K$ is finite whenever $K \subset S$ is closed and bounded.

² (a, b) refers to the open interval $a < x < b$.

which returns the point count of X restricted to set $B \subset S$.

Example 7.3. The counting measure of the random point pattern X in Example 7.1 can be written as

$$N(B) = \sum_{i=1}^n 1(U_i \in B), \quad B \subset (0, 1),$$

where the indicator of event $\{U_i \in B\}$ is defined by

$$1(U_i \in B) = \begin{cases} 1, & \text{if } U_i \in B, \\ 0, & \text{else.} \end{cases}$$

■

Time instants related to a random phenomenon under study can be modeled as random point patterns of \mathbb{R}_+ . In this case the point count on the interval $[0, t]$ is often briefly denote by

$$N(t) = N([0, t])$$

and the random function $t \mapsto N(t)$ is called the *counting process* (*laskuriprosessiksi*) of the point pattern X . The definition implies that the point count of X in an interval $(s, t]$ can be expressed as

$$|X \cap (s, t]| = N((s, t]) = N(t) - N(s).$$

7.3 Independent scattering

A random point pattern X is *independently scattered* (*riippumattomasti sironnut*) if the random variables $N(A_1), \dots, N(A_m)$ are independent whenever the sets A_1, \dots, A_m are disjoint. In this case information about the points of X within a set A is irrelevant when predicting how the point pattern behaves outside A . Independent scattering is indeed a very restrictive assumption, which only few point patterns satisfy.

Example 7.4. Is the point pattern $X = \{U_1, \dots, U_n\}$ of Example 7.1 independently scattered? By dividing the open unit interval into $A_1 = (0, 1/2]$ and $A_2 = (1/2, 1)$, we see that

$$\mathbb{P}(N(A_1) = 0) = \mathbb{P}(U_1 > 1/2, \dots, U_n > 1/2) = (1/2)^n.$$

On the other hand,

$$\mathbb{P}(N(A_1) = 0 \mid N(A_2) = n) = 1,$$

because by definition, the equation $N(A_1) + N(A_2) = n$ surely holds. This shows that X is not independently scattered. ■

The following important result characterizes how independent scattering, an intrinsically algebraic property, automatically yields a quantitative description of the distribution of point counts of the random point pattern. The result also underlines the central role of the Poisson distribution as a universal distribution describing point counts of independently scattered point patterns. A random point pattern X on \mathbb{R}_+ is *homogeneous* (*tasakoosteinen*) if its counting measure satisfies³

$$N(A+t) \stackrel{=st}{=} N(A)$$

for all $A \subset \mathbb{R}_+$ and all $t \geq 0$, where $A+t = \{a+t : a \in A\}$. The *intensity* (*intensiteetti*) of a homogeneous random point pattern is the expected point count $\mathbb{E}(N(0,1])$ on the unit interval $(0,1]$.

Theorem 7.5. *Let X be a homogeneous independently scattered random point pattern on \mathbb{R}_+ with intensity $0 < \lambda < \infty$. Then the point count of X in the interval $[0,t]$ is Poisson-distributed with mean λt , so that*

$$\mathbb{P}(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \dots$$

Proof. Denote by

$$v(t) = \mathbb{P}(N(0,t] = 0)$$

the probability that there are no points of X in the interval $(0,t]$. Because $N(0,s+t] = 0$ precisely when $N(0,s] = 0$ and $N(s,s+t] = 0$, we see that

$$\begin{aligned} v(s+t) &= \mathbb{P}(N(0,s+t] = 0) \\ &= \mathbb{P}(N(0,s] = 0, N(s,s+t] = 0) \\ &= \mathbb{P}(N(0,s] = 0) \mathbb{P}(N(s,s+t] = 0) \\ &= \mathbb{P}(N(0,s] = 0) \mathbb{P}(N(0,t] = 0) \\ &= v(s)v(t). \end{aligned}$$

Because v is a nonincreasing function, this implies (Exercise) that

$$v(t) = e^{-\alpha t} \tag{7.1}$$

for some $\alpha \geq 0$. Moreover, $\alpha > 0$, because in case $\alpha = 0$ the point pattern would be empty with probability one, which would be in conflict with the assumption $\lambda = \mathbb{E}(N(0,1]) > 0$. Analogously we may conclude that $\alpha < \infty$, because $\alpha = \infty$ would imply a conflict with the assumption $\lambda = \mathbb{E}(N(0,1]) < \infty$.

Let us next inspect the probability of $N(t) = k$ for some particular $t > 0$ and integer $k \geq 0$. Choose a large number $n \geq k$ and divide the interval $(0,t]$ into equally sized subintervals $I_{n,j} = (\frac{j-1}{n}t, \frac{j}{n}t]$, $j = 1, \dots, n$. Denote

$$\theta_j = 1(N(I_{n,j}) > 0) = \begin{cases} 1, & \text{if } N(I_{n,j}) > 0, \\ 0, & \text{else.} \end{cases}$$

³In these lecture notes $X \stackrel{=st}{=} Y$ means that X and Y are equal in distribution, that is, $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$ for all B .

Then $Z_n = \theta_1 + \dots + \theta_n$ is the number of subintervals which contains points of X . Denote by Ω_n the event that each subinterval contains at most one point. When the event Ω_n occurs, we have $N(t) = Z_n$, which implies that

$$\mathbb{P}(N(t) = k) = \mathbb{P}(Z_n = k) + \epsilon_n, \quad (7.2)$$

where

$$\epsilon_n = \mathbb{P}(N(t) = k, \Omega_n^c) - \mathbb{P}(Z_n = k, \Omega_n^c).$$

Because the indicator variables $\theta_1, \dots, \theta_n$ are independent (due to independent scattering) and each takes on value one with probability

$$q_n = 1 - v(t/n),$$

we find that Z_n follows the binomial $\text{Bin}(n, q_n)$ distribution.

$$\mathbb{P}(Z_n = k) = \binom{n}{k} q_n^k (1 - q_n)^{n-k}, \quad k = 0, \dots, n.$$

By equation (7.1) and l'Hôpital's rule we see that

$$nq_n = n(1 - e^{-\alpha t/n}) = \frac{1 - e^{-\alpha t/n}}{1/n} \rightarrow \alpha t$$

as $n \rightarrow \infty$. By the law of small numbers (Theorem 7.6) this allows to conclude that

$$\mathbb{P}(Z_n = k) \rightarrow e^{-\alpha t} \frac{(\alpha t)^k}{k!}, \quad \text{as } n \rightarrow \infty. \quad (7.3)$$

Because by Lemma 7.7, $|\epsilon_n| \leq 2\mathbb{P}(\Omega_n^c) \rightarrow 0$, and because the probability of the event $N(t) = k$ does not depend on n , we see from (7.2) and (7.3) that

$$\mathbb{P}(N(t) = k) = e^{-\alpha t} \frac{(\alpha t)^k}{k!}.$$

Therefore $N(t)$ is Poisson distributed with mean αt . Especially, $\mathbb{E}(N(t)) = \alpha t$ which shows that $\alpha = \lambda = \mathbb{E}(N(0, 1])$. \square

Lemma 7.6 (Law of small numbers). *Let Z_n be a $\text{Bin}(n, q_n)$ -distributed random integer, and assume that $nq_n \rightarrow \alpha \in (0, \infty)$ as $n \rightarrow \infty$. Then*

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n = k) = e^{-\alpha} \frac{\alpha^k}{k!} \quad \text{for all } k \geq 0.$$

Proof. By definition of the $\text{Bin}(n, q_n)$ distribution we find that

$$\begin{aligned} \mathbb{P}(Z_n = k) &= \frac{n!}{k!(n-k)!} (1 - q_n)^{n-k} q_n^k \\ &= \frac{n!}{n^k (n-k)!} \frac{1}{(1 - q_n)^k} \frac{(nq_n)^k}{k!} \left(1 - \frac{nq_n}{n}\right)^n. \end{aligned} \quad (7.4)$$

Let us analyze the right side of above equation as $n \rightarrow \infty$. The first term on the right side of (7.4) satisfies

$$\frac{n!}{n^k(n-k)!} = \frac{1}{n^k} \prod_{j=0}^{k-1} (n-j) = \prod_{j=0}^{k-1} (1-j/n) \rightarrow 1.$$

Because $q_n \rightarrow 0$, also the second term on the right side of (7.4) satisfies

$$\frac{1}{(1-q_n)^k} \rightarrow 1.$$

Furthermore, the assumption $nq_n \rightarrow \alpha$ implies that the third term on the right of (7.4) scales as

$$\frac{(nq_n)^k}{k!} \rightarrow \frac{\alpha^k}{k!}.$$

Hence the claim follows after verifying that

$$\lim_{n \rightarrow \infty} \left(1 - \frac{nq_n}{n}\right)^n = e^{-\alpha}. \quad (7.5)$$

The limit (7.5) can be justified as follows. Choose a small $\epsilon > 0$ and select n_0 so large that $\alpha - \epsilon \leq nq_n \leq \alpha + \epsilon$ for all $n \geq n_0$. Then for all $n \geq n_0$,

$$\left(1 - \frac{\alpha + \epsilon}{n}\right)^n \leq \left(1 - \frac{nq_n}{n}\right)^n \leq \left(1 - \frac{\alpha - \epsilon}{n}\right)^n.$$

By applying the formula $(1 + x/n)^n \rightarrow e^x$ (which is often taken as the definition of the exponential function) we see that the lower bound above converges to $e^{-\alpha-\epsilon}$ and the upper bound to $e^{-\alpha+\epsilon}$. Because the limiting bounds are valid for an arbitrarily small $\epsilon > 0$, equation (7.5) follows. \square

Lemma 7.7. *Let X be a random point pattern on an interval $S \subset \mathbb{R}$ with counting measure N . Let us divide the real axis into intervals $I_{n,j} = (\frac{j-1}{n}, \frac{j}{n}]$ of length $1/n$, indexed by $j \in \mathbb{Z}$. Then for any interval $A \subset S$ such that $\mathbb{E}(N(A)) < \infty$,*

$$\mathbb{P}\left(N(A \cap I_{n,j}) \leq 1 \text{ for all } j \in \mathbb{Z}\right) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Proof. Define the random number

$$D = \min\{|x - y| : x, y \in X \cap A, x \neq y\}$$

as the smallest interpoint distance of the point pattern restricted to A . When $D > 1/n$, then every pair of points in $X \cap A$ contains a gap of width $1/n$, so that every interval $I_{n,j}$ can contain at most one point of $X \cap A$. Therefore,

$$Z_n := \sup_j N(A \cap I_{n,j}) = \sup_j |X \cap A \cap I_{n,j}| \leq 1$$

on the event $D > 1/n$.

The assumption $\mathbb{E}(N(A)) < \infty$ implies that the set $X \cap A$ is finite with probability one. Hence $D > 0$ with probability one, and the above inequality shows that $\lim_{n \rightarrow \infty} 1(Z_n \leq 1) = 1$ with probability one. Now by applying Lebesgue's dominated convergence theorem to justify interchanging the limit and the expectation below, it follows that

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq 1) = \lim_{n \rightarrow \infty} \mathbb{E}\left(1(Z_n \leq 1)\right) = \mathbb{E}\left(\lim_{n \rightarrow \infty} 1(Z_n \leq 1)\right) = 1.$$

□

7.4 Poisson process

A random function $N : \mathbb{R}_+ \rightarrow \mathbb{Z}_+$ is a *Poisson process* (*Poisson-processi*) with intensity λ if

- $N(t) - N(s) \stackrel{\text{st}}{=} \text{Poi}(\lambda(t - s))$ for all $(s, t] \subset \mathbb{R}_+$.
- N has independent increments in the sense that

$$N(t_1) - N(s_1), \dots, N(t_n) - N(s_n)$$

are independent whenever $(s_1, t_1], \dots, (s_n, t_n] \subset \mathbb{R}_+$ are disjoint.

The above random function $t \mapsto N(t)$ is hence a continuous-time stochastic process with a countable state space \mathbb{Z}_+ . Theorem 7.5 can now be rephrased as follows.

Theorem 7.8. *The counting process $N(t) = N(0, t]$ of a homogeneous independently scattered random point pattern is a Poisson process with intensity $\lambda = \mathbb{E}(N(0, 1])$.*

7.5 Constructing independently scattered point patterns

Do independently scattered point patterns exist? Let us construct one. Define first the random numbers T_1, T_2, \dots by the formula

$$T_n = \tau_1 + \dots + \tau_n, \quad n \geq 1,$$

where τ_1, τ_2, \dots are independent and identically distributed positive random numbers. Figure 7.1 describes a so-constructed point patterns and a corresponding counting process.

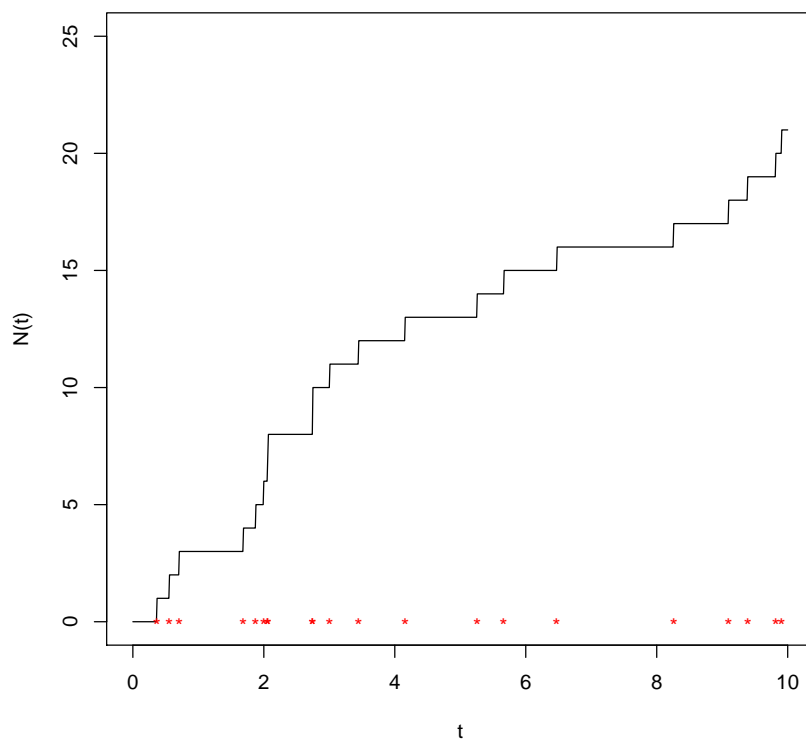


Figure 7.1: A point pattern simulated using the method in Theorem 7.9 and a corresponding Poisson process path on time interval $(0, 10]$.

Theorem 7.9. *If the interpoint distances τ_1, τ_2, \dots are exponentially distributed with rate parameter λ , then the point pattern $X = \{T_1, T_2, \dots\}$ is homogeneous and independently scattered, and the corresponding counting process*

$$N(t) = |X \cap (0, t]| = |\{k \geq 1 : T_k \leq t\}|$$

is a Poisson process with intensity λ .

Proof. See [Kal02, Proposition 12.15] for a detailed proof. □

Chapter 8

Poisson processes and renewal processes

8.1 Poisson process defined as a stochastic process

In the previous chapter we saw how a Poisson process naturally emerges as the counting process of a homogeneous and independently scattered point pattern. An alternative definition is the following. A random function $N : \mathbb{R}_+ \rightarrow \mathbb{Z}_+$ is a Poisson process with intensity $\lambda > 0$ if

- (i) $N(0) = 0$,
- (ii) $N(t) - N(s) \stackrel{\text{st}}{=} \text{Poi}(\lambda(t - s))$ for all $s < t$,
- (iii) N has independent increments in the sense that

$$\begin{aligned} & (s_1, t_1], \dots, (s_k, t_k] \text{ disjoint} \\ & \implies \\ & N(t_1) - N(s_1), \dots, N(t_k) - N(s_k) \text{ independent.} \end{aligned}$$

The paths of a Poisson process are piecewise constant, and grow with unit jumps at random time instants. Following the usual convention we impose the additional assumption that the paths of a Poisson process are right-continuous. Then the n -th jump instant of a Poisson process can be written as

$$T_n = \min\{t \geq 0 : N(t) = n\}, \quad n = 1, 2, \dots,$$

and the collection of jump instants $\{T_1, T_2, \dots\}$ forms a homogeneous and independently scattered random point pattern on $\mathbb{R}_+ : n$ with counting process $N(t)$, so that

$$N(t) = \sum_{i=1}^{\infty} 1(T_i \leq t).$$

The random variables T_1, T_2, \dots are often viewed as the events of Poisson process, and then the difference $N(t) - N(s)$ tells the number of events in the time interval $(s, t]$. With probability one, this number is the same for time intervals $[s, t]$ or (s, t) , because the probability of a Poisson process jumping at fixed nonrandom time instant is zero. This follows from the fact that distribution of T_n is continuous (T_n follows a gamma distribution with shape parameter n and rate parameter λ).

8.2 Superposed Poisson processes

The following theorem confirms the intuitively natural fact that by superposing several mutually independent Poisson processes we obtain a Poisson process. In the sum below the index set can be finite or countably infinite. In the latter case we need to assume that $\sum_j \lambda_j < \infty$.

Theorem 8.1. *If N_1, N_2, \dots are independent Poisson processes with intensities λ_j , then $N(t) = \sum_j N_j(t)$ is a Poisson process with intensity $\lambda = \sum_j \lambda_j$.*

The following auxiliary result is used to prove the theorem.

Lemma 8.2. *If $N_j =_{\text{st}} \text{Poi}(\lambda_j)$ are independent, then $\sum_j N_j =_{\text{st}} \text{Poi}(\sum_j \lambda_j)$.*

Proof. We will compute the probability generating function of N_j . This function at $z \in [0, 1]$ is obtained by

$$G_{N_j}(z) = \mathbb{E}(z^{N_j}) = \sum_{n=0}^{\infty} z^n \left(e^{-\lambda_j} \frac{\lambda_j^n}{n!} \right) = e^{-\lambda_j} e^{\lambda_j z} = e^{\lambda_j(z-1)}.$$

By independence, it follows that

$$G_{\sum_j N_j}(z) = \mathbb{E}(z^{\sum_j N_j}) = \prod_j \mathbb{E}(z^{N_j}) = \prod_j e^{\lambda_j(z-1)} = e^{\sum_j \lambda_j(z-1)}.$$

Because a probability generating function uniquely determines the distribution, $\sum_j N_j =_{\text{st}} \text{Poi}(\sum_j \lambda_j)$. \square

Proof of Theorem 8.1. Let us verify the three conditions in the definition (Section 8.1).

(i) Clearly $N(0) = \sum_j N_j(0) = 0$.

(ii) With the help of Lemma 8.2 we observe that $N(t) - N(s) = \sum_j (N_j(t) - N_j(s)) =_{\text{st}} \text{Poi}(\lambda(t-s))$, where $\lambda = \sum_j \lambda_j$.

(iii) Does N have independent increments? If time intervals $(s_1, t_1]$ and $(s_2, t_2]$ are disjoint, then

$$N_j(s_1, t_1] \perp\!\!\!\perp N_j(s_2, t_2] \quad \text{for all } j.$$

Because N_j are mutually independent, this allows to conclude that

$$\sum_j N_j(s_1, t_1] \perp\!\!\!\perp \sum_j N_j(s_2, t_2]$$

Hence the random integers $N(s_1, t_1]$ and $N(s_2, t_2]$ are independent. The above argument works in the same way also for multiple disjoint time intervals. Hence N has independent increments. \square

8.3 Compound Poisson process

A Poisson process $N(t)$ models the number of independently and uniformly scattered time instants during $[0, t]$. If the time instants are generated as a superposition of a several sparse event sequences, then the net counting process can be quite accurately modeled using a Poisson process. For example, this is the case for the traffic flow of cars on a large highway if the correlation effects due to traffic lights on inbound roads, the daily rhythm of the society (school start times, workday end times) are not too big.

In many random phenomena the time instants are often associated with other random variables that also need to be modeled. The following example describes one situation.

Example 8.3 (Traffic flow). The average flow of cars crossing the Helsinki–Espoo border on Länsiväylä during weekdays equals $\lambda = 40$ cars/min, and the average number of people per car is $m = 1.9$ with an estimated standard deviation of $\sigma = 1.2$. Model the flow of people traveling in cars across the city border as a stochastic process and derive a formula for the expectation and standard deviation for the flow of people crossing the border per hour. \blacksquare

We can add randomness to a random point pattern $X = \{T_1, T_2, \dots\}$ on \mathbb{R}_+ by defining

$$\tilde{X} = \{(T_1, Z_1), (T_2, Z_2), \dots\},$$

where Z_1, Z_2, \dots are random variables with values in some state space S . The resulting random point pattern \tilde{X} on $\mathbb{R}_+ \times S$ is called a *marked point pattern* (*merkitty pistekuvio*), and the random variables Z_1, Z_2, \dots are called the marks of the point pattern $\{T_1, T_2, \dots\}$. When the marks are real-valued, we may view Z_i as a reward (or cost) at time instant T_i . Then the net reward up to time t can be written as

$$S(t) = \sum_{i=1}^{\infty} Z_i 1(T_i \leq t),$$

or as

$$S(t) = \sum_{i=1}^{N(t)} Z_i, \tag{8.1}$$

where

$$N(t) = \sum_{i=1}^{\infty} 1(T_i \leq t),$$

denotes the counting process of the time instants $\{T_1, T_2, \dots\}$. When N is a Poisson process with intensity λ and the marks Z_1, Z_2, \dots are independent and identically distributed, and independent of N , then the stochastic process S defined by (8.1) is called a *compound Poisson process* (*yhdistetty Poisson-prosessi*).

Theorem 8.4. *A compound Poisson process has independent increments, and the mean and variance of a compound Poisson process at time t can be computed using the formulas*

$$\begin{aligned}\mathbb{E}(S(t)) &= \lambda mt, \\ \text{Var}(S(t)) &= \lambda(m^2 + \sigma^2)t,\end{aligned}$$

where $m = \mathbb{E}(Z_i)$ and $\sigma^2 = \text{Var}(Z_i)$.

Proof. The independence of increments is intuitively clear. Proving this rigorously can be done by carefully conditioning on event of the form $A_k = \{N_k(s_k) = m_k, N_k(t_k) = m_k + r_k\}$. The claims follow from Lemma 8.5 when we note that $N(t)$ is Poisson distributed with mean λt and hence $\mathbb{E}(N(t)) = \text{Var}(N(t)) = \lambda t$. \square

Lemma 8.5. *Let $S = \sum_{i=1}^N Z_i$, where Z_1, Z_2, \dots are identically distributed, and independent of each other and N .*

(i) *If N and Z_i have first moments, then $\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(Z_i)$.*

(ii) *If N and Z_i have second moments¹, then*

$$\text{Var}(S) = \mathbb{E}(N) \text{Var}(Z_i) + \text{Var}(N)(\mathbb{E}(Z_i))^2.$$

Proof. (i) Because the random variables Z_1, Z_2, \dots are independent of N , and $\mathbb{E}(Z_i)$ does depend on i , we find that

$$\mathbb{E}(S | N = n) = \mathbb{E}\left(\sum_{i=1}^n Z_i | N = n\right) = \mathbb{E}\left(\sum_{i=1}^n Z_i\right) = \sum_{i=1}^n \mathbb{E}(Z_i) = n\mathbb{E}(Z_i).$$

Therefore, by conditioning on the possible values of N we find that

$$\mathbb{E}(S) = \sum_{n \geq 0} \mathbb{P}(N = n) \mathbb{E}(S | N = n) = \sum_{n \geq 0} \mathbb{P}(N = n) n \mathbb{E}(Z_i) = \mathbb{E}(N) \mathbb{E}(Z_i).$$

(ii) The second equality can be proved by carefully manipulating the expression $\mathbb{E}((S - \mathbb{E}(S))^2 | N = n)$, and recalling that $\text{Var}(\sum_{i=1}^n Z_i) = \sum_{i=1}^n \text{Var}(Z_i)$ for mutually independent Z_i . Working through the details is a good exercise. \square

¹A random number Z has a finite second moment if $\mathbb{E}(Z^2) < \infty$. In this case it also has a finite first moment, because it can be proved that $\mathbb{E}(|Z|) \leq (\mathbb{E}(Z^2))^{1/2}$.

Example 8.6 (Traffic flow). The flow of cars in Example 8.3 can be modeled using a Poisson process with intensity $\lambda = 40$, when set the time unit as 1 min. To a car crossing the border at time instant T_i we attach a random variable Z_i which tells the number of people in the car. It is natural to assume that the random variables Z_1, Z_2, \dots are independent of each other and of the instants T_1, T_2, \dots . By doing so, the number of people who have crossed the border during $[0, t]$ can be represented as a compound Poisson process

$$S(t) = \sum_{i=1}^{N(t)} Z_i.$$

In this case we know that Z_i take values in $S = \{1, 2, \dots, 7\}$, and $\mathbb{E}(Z_i) = m$ and $\text{Var}(Z_i) = \sigma^2$ with $m = 1.9$ ja $\sigma = 1.2$.

By Theorem 8.4, at the time instant $t = 60$,

$$\mathbb{E}(S(t)) = \lambda m t = 40 \times 1.9 \times 60 = 4560$$

and

$$\text{Var}(S(t)) = \lambda(m^2 + \sigma^2)t = 40 \times (1.9^2 + 1.2^2) \times 60 = 12120.$$

The number of people $S(60)$ crossing the Helsinki–Espoo border hence has mean 4560 and standard deviation $\sqrt{12120} = 110.09$. Because the model is statistically shift invariant, the same conclusion holds for any time interval of 60 minutes. ■

8.4 Thinned Poisson process

In Section 8.2 we found that by superposing independent Poisson processes we obtain a new Poisson process. In this section we consider a corresponding reverse operation, splitting a Poisson process into several independent Poisson processes.

Example 8.7 (Thinned traffic flow). The average flow of cars crossing the Helsinki–Espoo border on Länsiväylä highway during weekdays equals $\lambda = 40$ cars/min. Of these cars, $p_1 = 30\%$ take the exit to Kehä I ring road, and rest continue west along Länsiväylä. Model statistically the flow of cars which continue west on Länsiväylä. What is the probability that during a particular minute, at most 20 cars continue on Länsiväylä, given that at least 30 cars exit to Kehä I? ■

Let us denote, for a time interval $[0, t]$,

- the total number of cars $N(t) = \sum_{i=1}^{\infty} 1(T_i \leq t)$,
- the number of cars exiting to Kehä I by $N_1(t) = \sum_{i=1}^{\infty} \theta_i 1(T_i \leq t)$,
- the number of cars continuing west by $N_2(t) = \sum_{i=1}^{\infty} (1 - \theta_i) 1(T_i \leq t)$,

where $\theta_i \in \{0, 1\}$ is the indicator variable for the event that the i -th car crossing the border takes the Kehä I exit. If we assume that $\theta_1, \theta_2, \dots$ are independent, the so-obtained counting process $N_1(t)$ is called a *thinned Poisson process* (*harvennettu Poisson-prosessi*) which is obtained by removing 70% of the events of the original Poisson process by independent sampling. Analogously, also $N_2(t)$ is a thinned Poisson process. The following result confirms that independently thinned Poisson processes are Poisson processes; and more strikingly, the thinned processes are mutually independent.

Theorem 8.8. *The thinnings $N_1(t) = \sum_{i=1}^{\infty} \theta_i 1(T_i \leq t)$ and $N_2(t) = \sum_{i=1}^{\infty} (1 - \theta_i) 1(T_i \leq t)$ of the Poisson process N are Poisson processes and mutually independent.*

Proof. Let us first verify that N_1 is a Poisson process. Obviously $N_1(0) = 0$. Let us next verify that $N_1(t)$ is Poisson distributed. The probability generating function of a $\text{Ber}(p_1)$ -distributed indicator variable θ_i is

$$G_{\theta_i}(z) = \mathbb{E}(z^{\theta_i}) = (1 - p_1) + p_1 z.$$

Because $N_1(t) = \sum_{i=1}^{N(t)} \theta_i$, we may apply Theorem 6.1, familiar from branching processes, according to which

$$G_{N_1(t)}(z) = G_{N(t)}(G_{\theta_i}(z)).$$

By applying this we see that

$$G_{N_1(t)}(z) = G_{N(t)}(G_{\theta_i}(z)) = e^{\lambda t(G_{\theta_i}(z)-1)} = e^{\lambda t p_1(z-1)},$$

which implies that $N_1(t) =_{\text{st}} \text{Poi}(\lambda p_1 t)$. In precisely the same way we can verify that $N_1(t) - N_1(s) =_{\text{st}} \text{Poi}(\lambda p_1(t - s))$. Moreover, because $N_1(t)$ is a compound Poisson process, it follows by Theorem 8.4 that N_1 has independent increments. Hence N_1 is a Poisson process with intensity λp_1 . In an analogous way we find that N_2 is a Poisson process with intensity $\lambda(1 - p_1)$.

Let us still verify why N_1 and N_2 are independent. The event $\{N_1(s, t] = j, N_2(s, t] = k\}$ occurs precisely when the interval $(s, t]$ contains $N(s, t] = j + k$ events, out of which to N_1 we select j events and to N_2 we select k events. Because the selections are done independently, we see by applying the binomial distribution, and noting $p_2 = 1 - p_1$, that

$$\begin{aligned} \mathbb{P}(N_1(t) = j, N_2(t) = k) &= \mathbb{P}(N(t) = j + k) \binom{j+k}{j} p_1^j (1 - p_1)^k \\ &= e^{-\lambda t} \frac{(\lambda t)^{j+k}}{(j+k)!} \binom{j+k}{j} p_1^j p_2^k \\ &= e^{-\lambda p_1 t} \frac{(\lambda p_1 t)^j}{j!} e^{-\lambda p_2 t} \frac{(\lambda p_2 t)^k}{k!} \\ &= \mathbb{P}(N_1(t) = j) \mathbb{P}(N_2(t) = k). \end{aligned}$$

Hence the random variables $N_1(t)$ and $N_2(t)$ are independent for every t . This argument can be generalised to show that the random vectors $(N_1(t_1), \dots, N_1(t_n))$ and $(N_2(t_1), \dots, N_2(t_n))$ are independent for arbitrary t_1, \dots, t_n , which corresponds to the independence of the processes N_1 and N_2 . \square

Example 8.9 (Thinned traffic flow). For the model of Example 8.7, it follows by Theorem 8.8 that the traffic flows corresponding cars continuing on Länsiväylä and exiting to Kehä I are mutually independent. Therefore the probability that during a particular minute, at most 20 cars continue on Länsiväylä, given that at least 30 cars exit to Kehä I equals

$$\mathbb{P}(N_2(1) \leq 20 \mid N_1(1) \geq 30) = \mathbb{P}(N_2(1) \leq 20).$$

Information about cars exiting to Kehä I in this setting has no relevance in predicting how many cars continue west on Länsiväylä. \blacksquare

The above independence is intuitively counterintuitive because by definition, $N_1(t) + N_2(t) = N(t)$ with probability one. The independence property is one of the magical properties of Poisson processes which are not valid in general for other counting processes. The result of Theorem 8.8 can be generalized to thinnings with more general random variables compared to coin flips.

Theorem 8.10. *If N is a Poisson process with intensity λ , and Z_1, Z_2, \dots are identically distributed, and independent of N and each other, then the thinned processes*

$$N_x(t) = \sum_{i=1}^{\infty} 1(Z_i = x)1(T_i \leq t), \quad x \in S,$$

are independent Poisson processes with intensities $\lambda_x = \lambda \mathbb{P}(Z_i = x)$.

8.5 Renewal processes

A fundamental and classical question related to random time events is the following.

Example 8.11 (Bus stop). Buses arrive at independent and identically distributed time intervals τ_1, τ_2, \dots . What is the expected waiting time for the next bus for a passenger who arrives randomly to the bus stop? \blacksquare

The above question appears natural but when we look at it carefully, it is not completely well specified because the meaning of “arrives randomly” is somewhat ambiguous. What is usually meant is that the passenger is assumed to arrive to the bus stop at a random time instant which is independent of the bus arrival times, and uniform somehow. But uniform distributions are only defined for bounded time intervals, and in the above question no such bound is given. We will study how this problem can be sensibly formulated in the context of renewal processes.

A *renewal process* (*uusiuutumisprosessi*) is the counting process

$$N(t) = \sum_{i=1}^{\infty} 1(T_i \leq t)$$

of a random point pattern $\{T_1, T_2, \dots\}$ on \mathbb{R}_+ defined by $T_n = \sum_{k=1}^n \tau_k$ where the interpoint distances $\tau_1, \tau_2, \dots \geq 0$ are independent and identically distributed². The probability distribution of the interpoint distances is called the *interevent distribution* (*väliaikajakautuma*) of the renewal process.

Example 8.12 (Poisson process). A renewal process with a memoryless interevent distribution $\text{Exp}(\lambda)$ is a Poisson process with intensity $\lambda > 0$. ■

Example 8.13 (Periodic event sequence). The counting process of the deterministic point pattern $\{h, 2h, 3h, \dots\}$ is a renewal process with interevent distribution being the Dirac distribution at h , so that $\mathbb{P}(\tau_k = h) = 1$ for all $k \geq 1$. ■

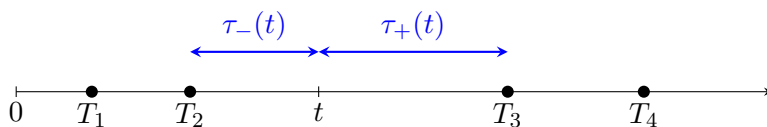


Figure 8.1: Backward and forward recurrence times.

Given a point pattern of time instants $\{T_1, T_2, \dots\}$ the distances from a reference time instant $t > 0$ to previous time instant $\tau_-(t) = t - T_{N(t)}$ is called the *backward recurrence time*, and the distance to the next time instant $\tau_+(t) = T_{N(t)+1} - t$ is called the *forward recurrence time*, see Figure 8.1. On the event $N(t) = 0$ we define $\tau_-(t) = \infty$. Then the interevent time seen from the reference point t equals

$$\tau_*(t) = \tau_-(t) + \tau_+(t).$$

On an infinite time interval $\mathbb{R}_+ = [0, \infty)$ we cannot choose a uniformly random point because no constant function satisfies the condition $\int_0^\infty f(u) du = 1$. However, we may still choose a uniformly random time point U_s from a long interval $[0, s]$ and then inspect what happens when $s \rightarrow \infty$. In this case the forward recurrence time $\tau_+(U_s)$ represents the waiting time until the next time instant for reference point selected uniformly at random from $[0, s]$. The corresponding cumulative density function equals

$$\mathbb{P}(\tau_+(U_s) \leq t) = \frac{1}{s} \int_0^s \mathbb{P}(\tau_+(u) \leq t) du.$$

The cumulative density functions of $\tau_-(U_s)$ and $\tau_*(U_s)$ can be written in a similar way. The following is a version of a general set of result known as a renewal theorem.

²A delayed renewal process can be defined as the counting process of $T_n = \tau_0 + \sum_{k=1}^n \tau_k$ where the initial delay τ_0 may have different distribution from the other interpoint distances.

Theorem 8.14. For a renewal process where the interpoint distances satisfy $\mathbb{P}(\tau_i > 0) = 1$ and $\mathbb{E}(\tau_i) \in (0, \infty)$,

$$\lim_{s \rightarrow \infty} \mathbb{P}(\tau_+(U_s) \leq t) = \lim_{s \rightarrow \infty} \mathbb{P}(\tau_-(U_s) \leq t) = F_+(t),$$

and

$$\lim_{s \rightarrow \infty} \mathbb{P}(\tau_*(U_s) \leq t) = F_*(t),$$

where the limiting cumulative distribution functions are defined by

$$F_+(t) = \frac{\mathbb{E}(\tau_i \wedge t)}{\mathbb{E}(\tau_i)} \quad \text{and} \quad F_*(t) = \frac{\mathbb{E}(\tau_i 1(\tau_i \leq t))}{\mathbb{E}(\tau_i)}. \quad (8.2)$$

The probability distribution F_+ in (8.2) is called the *stationary distribution* (*tasapainojakauma*) of the renewal process. By applying the equation $\tau_i \wedge t = \int_0^t 1(s < \tau_i) ds$ we may write

$$F_+(t) = \frac{\mathbb{E} \int_0^t 1(s < \tau_i) ds}{\mathbb{E}(\tau_i)} = \int_0^t \frac{\mathbb{P}(\tau_i > s)}{\mathbb{E}(\tau_i)} ds,$$

from which we see that the stationary distribution admits a density function

$$f_+(t) = \frac{\mathbb{P}(\tau_i > t)}{\mathbb{E}(\tau_i)}, \quad t \geq 0. \quad (8.3)$$

The probability distribution F_* in (8.2) is called a *size-biased* (*kokovinoutettu*) interevent distribution. If the interevent distribution of the renewal process has a density function f , then the size-biased interevent distribution has a density

$$f_*(t) = \frac{tf(t)}{\int_0^\infty sf(s) ds}, \quad t \geq 0. \quad (8.4)$$

The expectations of random variables τ_+ and τ_* distributed according to F_+ and F_* can be computed using the formulas

$$\mathbb{E}(\tau_+) = \frac{\mathbb{E}(\tau_i^2)}{2\mathbb{E}(\tau_i)} \quad \text{and} \quad \mathbb{E}(\tau_*) = \frac{\mathbb{E}(\tau_i^2)}{\mathbb{E}(\tau_i)}.$$

By applying the general inequality $\mathbb{E}(\tau_i) \leq (\mathbb{E}(\tau_i^2))^{1/2}$ we find that

$$\mathbb{E}(\tau_*) \geq \mathbb{E}(\tau_i). \quad (8.5)$$

This inequality is known as the *inspection paradox* (*tutkintaparadoksi*), and it tells that from the viewpoint of a randomly chosen reference point, the interevent times appear larger than what $\mathbb{E}(\tau_i)$ suggests. This is due to the fact that a randomly chosen reference point is likely to be located within a time interval which is larger than a typical time interval.

Heuristic proof of Theorem 8.14. We present an intuitive derivation of the stationary density (8.3) when the interevent distribution has density f . Here we view the forward recurrence time $\tau_+(t)$ (see Figure 8.1) as a random function of t . Then $t \mapsto \tau_+(t)$ is a continuous-time stochastic process with state space \mathbb{R}_+ , see Figure 8.2. Indeed, it can be shown that this is a Markov process with a well-defined limiting distribution. Let us assume that the limiting distribution has a density function f_+ , and consider how the process behaves in statistical equilibrium.

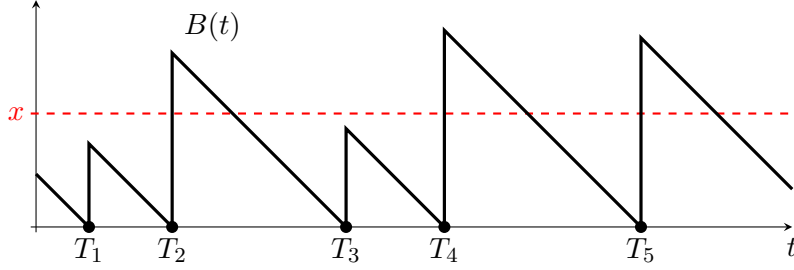


Figure 8.2: Forward recurrence time process $B(t)$.

Fix a level $x > 0$, consider the long-term expected rate of upcrossings of the level x . Such an upcrossing occurs during a short time interval $(t, t + h)$ if and only if $B(t) \in (0, h)$ and the next interevent time is larger than x . Therefore the upcrossing rate for large t is approximately

$$\mathbb{P}(0 < B(t) < h)\mathbb{P}(\tau_i > x) \approx f_+(0)h\mathbb{P}(\tau_i > x).$$

On the other hand, a downcrossing of level x occurs during a time interval $(t, t + h)$ if and only if $B(t) \in (x, x + h)$, and hence the downcrossing rate for large t is approximately

$$\mathbb{P}(x < B(t) < x + h) \approx f_+(x)h.$$

In a statistical equilibrium these rates should be equal, so we conclude that $f_+(x) = f_+(0)\mathbb{P}(\tau_i > x)$. By integrating over x we find that

$$1 = \int_0^\infty f_+(x) dx = f_+(0) \int_0^\infty \mathbb{P}(\tau_i > x) dx = f_+(0) \mathbb{E}(\tau_i),$$

so that $f_+(0) = 1/\mathbb{E}(\tau_i)$, and we obtain (8.3).

A rigorous proof of the full statement of Theorem 8.14 is based on general renewal theory arguments, see for example [Asm03, Luku V.4]. \square

Example 8.15 (Bus stop with exponential interarrivals). Assume the bus interarrival times in Example 8.11 are independent and $\text{Exp}(\lambda)$ -distributed with mean $\frac{1}{\lambda} = 10$ min. Then by (8.3) the stationary distribution of the renewal process has density

$$f_+(t) = \frac{\mathbb{P}(\tau_i > t)}{\mathbb{E}(\tau_i)} = \frac{e^{-\lambda t}}{1/\lambda} = \lambda e^{-\lambda t},$$

so that also the stationary distribution is $\text{Exp}(\lambda)$. Moreover, by (8.4) the size-biased interevent distribution has density

$$f_*(t) = \frac{tf(t)}{\int_0^\infty sf(s) ds} = \frac{\lambda te^{-\lambda t}}{1/\lambda} = \lambda^2 e^{-\lambda t},$$

which can be recognized as $\text{Gam}(2, \lambda)$ -distribution. By Theorem 8.14, a randomly arriving passenger experiences an expected waiting time of $\mathbb{E}(\tau_+) = \frac{1}{\lambda} = 10$ min. Moreover, the randomly arriving passenger observes that the expected time between the previous bus and the next bus is $\mathbb{E}(\tau_*) = 2/\lambda = 20$ min. ■

A $\text{Gam}(2, \lambda)$ -distributed random number τ_* discovered in Example 8.15 can also be represented as

$$\tau_* = \tau_- + \tau_+,$$

where τ_- ja τ_+ are independent and $\text{Exp}(\lambda)$ -distributed. This is natural because due to the memoryless property of exponential distributions, the distances from any reference point to the previous and next time instants of a Poisson process are mutually independent and $\text{Exp}(\lambda)$ -distributed. However, in general the backward and forward recurrence times are not independent.

Example 8.16 (Bus stop with periodic arrivals). Assume now that the bus interarrival times in Example 8.11 deterministic and all equal to $h = 10$ min. The counting process of the bus arrivals is then a renewal process with interevent distribution being the Dirac distribution at h . Then by (8.3) the stationary distribution of the renewal process has density

$$f_+(t) = \frac{\mathbb{P}(h > t)}{\mathbb{E}(h)} = \frac{1}{h}, \quad 0 < t < h,$$

which corresponding to the uniform distribution on $(0, h)$. The size-biased interevent distribution F_* does not have a density, but using (8.2) one can verify that $F_*(t) = 1(t < h)$ corresponds to the Dirac distribution at h . ■

Chapter 9

Continuous-time Markov chains in finite time horizon

9.1 Markov property

In general, a *stochastic process* (*stokastinen prosessi*) is a random function $X : T \rightarrow S$ with state space S and time range $T \subset \mathbb{R}$, defined on some measurable space with probability measure \mathbb{P} . A stochastic process $(X_t)_{t \in \mathbb{R}_+}$ with a countable state space S and time range $T = \mathbb{R}_+$ is called a *continuous-time Markov chain* (*jatkuva-aikainen Markov-ketju*) if it satisfies the *Markov property* (*Markov-ominaisuus*)

$$\mathbb{P}\left(X_u = y \mid X_t = x, (X_s)_{s \leq t} \in A\right) = \mathbb{P}\left(X_u = y \mid X_t = x\right) \quad (9.1)$$

for all states $x, y \in S$, all time indices $s \leq t \leq u$, and all measurable¹ sets of paths $A \subset S^{[0,t]}$ such that the conditioning events above have nonzero probability. The above definition means that information about past states $(X_s)_{s \leq t}$ of the chain is irrelevant for predicting a future state X_u , if we know the current state X_t . This Markov property extends [Kal02, Lemma 8.1] to joint distributions of several future states instead of just one. The *extended Markov property* (*laajennettu Markov-ominaisuus*) can be stated as

$$\mathbb{P}\left((X_u)_{u \geq t} \in B \mid X_t = x, (X_s)_{s \leq t} \in A\right) = \mathbb{P}\left((X_u)_{u \geq t} \in B \mid X_t = x\right) \quad (9.2)$$

for all states $x, y \in S$, all time indices $s \leq t \leq u$, and all measurable sets of paths $A \subset S^{[0,t]}$ and $B \subset S^{[t,\infty)}$ such that the conditioning events above have nonzero probability.

A continuous-time Markov chain is called *time-homogeneous* (*aikahomogeeninen*) if

$$\mathbb{P}\left(X_u = y \mid X_t = x\right) = \mathbb{P}\left(X_{u-t} = y \mid X_0 = x\right).$$

¹Here measurable refers to the product sigma-algebra on the space S^I of functions from I to S .

As for discrete-time chains earlier, all Markov chains are implicitly assumed to be time-homogeneous unless otherwise specified.

9.2 Transition matrices

The definition of a continuous-time Markov chain is close in spirit to the corresponding definition for discrete-time Markov chains, with one essential difference. Namely, now it no more suffices to keep track of transitions of unit-length time steps, but we also need to study transition probabilities for arbitrarily small (and large) time steps. As a consequence, instead of just one transition matrix, now need an infinite collection of transition matrices. The *t-step transition matrix* of a continuous-time Markov chain (X_t) is denoted by

$$P_t(x, y) = \mathbb{P}(X_t = y \mid X_0 = x).$$

The entries of the square matrix P_t are nonnegative, and the rows sums are one because

$$\sum_{y \in S} P_t(x, y) = \sum_{y \in S} \mathbb{P}(X_t = y \mid X_0 = x) = 1 \quad \text{for all } x \in S.$$

As in discrete time, the distribution of a continuous-time Markov chain is easily determined by the initial distribution and a suitable transition matrix.

Theorem 9.1. *The probability distribution $\mu_t(x) = \mathbb{P}(X_t = x)$ of a continuous-time Markov chain at time t is obtained from the initial distribution μ_0 and the t -step transition matrix P_t via $\mu_t = \mu_0 P_t$.*

Proof. By conditioning on the possible values of X_0 , we find that

$$\mathbb{P}(X_t = y) = \sum_{x \in S} \mathbb{P}(X_0 = x) \mathbb{P}(X_t = y \mid X_0 = x) = \sum_{x \in S} \mu_0(x) P_t(x, y)$$

□

The following result confirms a fundamental algebraic property of the transition matrices, stating that the collection $(P_t)_{t \geq 0}$ forms a *transition semigroup*.

Theorem 9.2. *The transition matrices of a continuous-time Markov chain satisfy $P_{s+t} = P_s P_t$ for all $s, t \geq 0$.*

Proof. By applying the definition of conditional probability, and the Markov

property at time instant s ,

$$\begin{aligned}
P_{s+t}(x, z) &= \mathbb{P}(X_{s+t} = z \mid X_0 = x) \\
&= \sum_{y \in S} \mathbb{P}(X_{s+t} = z, X_s = y \mid X_0 = x) \\
&= \sum_{y \in S} \mathbb{P}(X_s = y \mid X_0 = x) \mathbb{P}(X_{s+t} = z \mid X_s = y, X_0 = x) \\
&= \sum_{y \in S} \mathbb{P}(X_s = y \mid X_0 = x) \mathbb{P}(X_{s+t} = z \mid X_s = y) \\
&= \sum_{y \in S} P_s(x, y) P_t(y, z).
\end{aligned}$$

□

Example 9.3 (Satellite). A satellite that has been launched in space has a random operational time T which assumed to be $\text{Exp}(\mu)$ -distributed with mean $1/\mu = 10$ years. When the satellite breaks, it will not be repaired. Then the state of the satellite can be described as a stochastic process

$$X_t = \begin{cases} 1, & \text{if satellite is operational at time } t, \\ 0, & \text{else.} \end{cases}$$

We will now verify that (X_t) is a Markov chain. Given that $X_t = 1$ occurs, we know that the satellite is still operational at time t , and nothing has so far happened to the system. Therefore, by applying the memoryless property of exponential distributions, we see that for any event H_t determined by the past values $(X_s : s \leq t)$,

$$\begin{aligned}
\mathbb{P}(X_{t+h} = 1 \mid X_t = 1, H_t) &= \mathbb{P}(X_{t+h} = 1 \mid X_t = 1) \\
&= \mathbb{P}(T > t + h \mid T > t) \\
&= \mathbb{P}(T > h) \\
&= e^{-\mu h}.
\end{aligned}$$

By the law of total probability, the probability of the complementary event equals

$$\mathbb{P}(X_{t+h} = 0 \mid X_t = 1, H_t) = 1 - e^{-\mu h}.$$

Furthermore, because a broken satellite remains broken, we see that

$$\begin{aligned}
\mathbb{P}(X_{t+h} = 0 \mid X_t = 0, H_t) &= 1, \\
\mathbb{P}(X_{t+h} = 1 \mid X_t = 0, H_t) &= 0.
\end{aligned}$$

Together the above four equations show that (X_t) is a continuous-time Markov chain on state space $\{0, 1\}$, and its h -step transition matrix is

$$P_h = \begin{bmatrix} P_h(0, 0) & P_h(0, 1) \\ P_h(1, 0) & P_h(1, 1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 - e^{-\mu h} & e^{-\mu h} \end{bmatrix}, \quad h \geq 0.$$

■

Example 9.4 (Poisson process). Let $(N_t)_{t \in \mathbb{R}_+}$ be a Poisson process with intensity $\alpha > 0$. Because the random point pattern consisting of the jump instants of (N_t) is homogeneous and independently scattered, it follows that for any event H_t determined by the values of the Poisson process up to time $[0, t]$,

$$\begin{aligned} \mathbb{P}(N_{t+h} = j \mid N_t = i, H_t) &= \mathbb{P}(N_{t+h} - N_t = j - i \mid N_t = i, H_t) \\ &= \mathbb{P}(N_{t+h} - N_t = j - i) \\ &= \mathbb{P}(N_h - N_0 = j - i) \\ &= \mathbb{P}(N_h = j - i). \end{aligned}$$

Because the random variable N_h is Poisson distributed with mean αh , it follows that (N_t) is a continuous-time Markov process on state space \mathbb{Z}_+ with h -step transition matrix

$$P_h(i, j) = \begin{cases} e^{-\alpha h} \frac{(\alpha h)^{j-i}}{(j-i)!}, & j \geq i, \\ 0, & \text{else.} \end{cases}$$

■

Example 9.5 (Poisson modulated chain). A *Poisson modulated chain* is a random process $(X_t)_{t \in \mathbb{R}_+}$ of the form

$$X_t = Y_{N(t)},$$

where $(Y_n)_{n \in \mathbb{Z}_+}$ is a discrete-time Markov chain on state space S with transition matrix P and $(N(t))_{t \in \mathbb{R}_+}$ is a Poisson process with intensity λ which is independent of $(Y_n)_{n \in \mathbb{Z}_+}$. If we denote the jump instants of the Poisson process by T_1, T_2, \dots , then we see that

$$X_t = \begin{cases} Y_0, & 0 \leq t < T_1, \\ Y_1, & T_1 \leq t < T_2, \\ Y_2, & T_2 \leq t < T_3, \end{cases}$$

and so on, see Figure 9.1. Because the interevent times $T_n - T_{n-1}$ are independent, and $\text{Exp}(\lambda)$ -distributed, it possible to show using the memoryless property of exponential distributions that (X_t) is a continuous-time Markov chain on state space S .

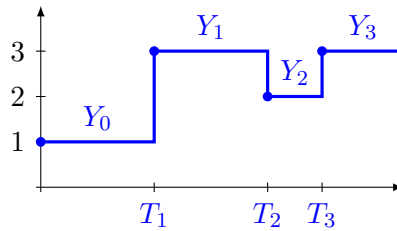


Figure 9.1: Path of a Poisson modulated chain.

The t -step transition matrices of (X_t) can be computed using powers of the underlying discrete-time transition matrix P , because conditioning on the number $N(t)$ shows that

$$\begin{aligned}\mathbb{P}(X_t = y \mid X_0 = x) &= \sum_{n=0}^{\infty} \mathbb{P}(N(t) = n) \mathbb{P}(Y_n = y \mid Y_0 = x) \\ &= \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} P^n(x, y)\end{aligned}$$

and hence

$$P_t(x, y) = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} P^n(x, y). \quad (9.3)$$

■

Poisson modulated chains in Example 9.5 provide a rich and versatile class of continuous-time Markov chains. For example, the process of Example 9.3 is a Poisson modulated chain with $(N(t))$ being a Poisson process with intensity μ , and (Y_n) being a discrete-time Markov chain on $\{0, 1\}$ with transition matrix

$$P = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

Similarly, any Poisson process (recall Example 9.4) can be seen as a special instance of a Poisson modulated chain where $Y_n = n$ is a Markov chain on \mathbb{Z}_+ which deterministically moves one step up at every discrete time step. We will later see that indeed *all* continuous-time Markov chains with bounded total jump rates can be represented as Poisson modulated chains (Section 10.2).

9.3 Generator matrix

For discrete-time Markov chains, the multi-step transition matrices are given as matrix powers $P_t = P^t$ for $t = 0, 1, 2, \dots$ where $P = P_1$ is the one-step transition matrix. In this sense the one-step transition matrix P_1 generates the full transition semigroup $(P_t)_{t \in \mathbb{Z}_+}$ of the discrete-time Markov chain, and makes analysing discrete-time chain computationally convenient using numerical linear algebra. This leads ourselves to ask the following fundamental question:

Is it possible to generate the transition semigroup $(P_t)_{t \in \mathbb{R}_+}$ of a continuous-time Markov chain using just one matrix?

To see why this might be possible, observe that the semigroup property (Theorem 9.2) implies that $P_{nt} = P_t^n$ for any $t \geq 0$ and any integer $n \geq 0$. Hence if we knew the transition matrices for a small time interval $t \in (0, \epsilon)$, then we would be able to compute the transition matrix P_t for *every* $t \geq 0$ via the formula

$$P_t = P_{n \cdot (t/n)} = P_{t/n}^n$$

after choosing an integer n to be large enough so that $t/n \in (0, \epsilon)$. This means that we only need to know the transition matrices for t arbitrarily close to zero. However, this reasoning does not yet reveal whether or not there exists a single matrix which generates the full semigroup. A natural candidate would be the entrywise limit $\lim_{t \rightarrow 0} P_t$. This does not work because the limit P_0 equals the identity matrix and hence contains no information about the behaviour of the Markov chain.

To see what might work, let us investigate what formula (9.3) in Example (9.5) suggests. Note first that the *matrix exponential* of a square matrix A is defined as a square matrix

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!}$$

so that the (x, y) of e^A equals

$$e^A(x, y) = \lim_{N \rightarrow \infty} \sum_{n=0}^N \frac{A^n}{n!}(x, y).$$

The limit on the right converges for every finite matrix A , and also for all suitably bounded countably infinite transition matrices. Now (9.3) tells that the t -step transition matrix of a Poisson modulated chain (X_t) with underlying discrete-time transition matrix P and clock rate λ can be written as

$$P_t = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} P^n = e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t P)^n}{n!} = e^{-\lambda t} e^{\lambda t P},$$

with the understanding that each entry of the square matrix on left equals the entry of the square matrix on the right. By noting that $e^{-\lambda t} I = e^{-\lambda t I}$ and applying the formula $e^A e^B = e^{A+B}$ which is valid when $AB = BA$, we find that

$$P_t = e^{-\lambda t} e^{\lambda t P} = e^{-\lambda t I} e^{\lambda t P} = e^{\lambda t (P - I)} = e^{tQ}$$

where

$$Q = \lambda(P - I). \tag{9.4}$$

We conclude that the transition semigroup $(P_t)_{t \in \mathbb{R}_+}$ is completely determined in terms of single matrix Q via the formula $P_t = e^{tQ}$. A consequence of the above is that if we differentiate the square matrix P_t (entry-by-entry) with respect to t , then (assuming that we can bring the derivative inside the infinite sum)

$$\frac{d}{dt} P_t = \sum_{n=0}^{\infty} \frac{d}{dt} \frac{(tQ)^n}{n!} = \sum_{n=0}^{\infty} \frac{d}{dt} \frac{t^n Q^n}{n!} = \sum_{n=1}^{\infty} \frac{t^{n-1}}{(n-1)!} Q^n = \sum_{n=0}^{\infty} \frac{1}{n!} t^n Q^{n+1}.$$

This implies Kolmogorov's backward equation

$$\frac{d}{dt} P_t = Q P_t,$$

and by taking $t = 0$ and noting that $P_0 = I$, it follows that

$$Q = \left[\frac{d}{dt} P_t \right]_{t=0}.$$

The above derivation motivates the following definition. The *generator matrix* of a transition semigroup $(P_t)_{t \in \mathbb{R}_+}$ and a corresponding continuous-time Markov chain is defined as the square matrix

$$Q = \left[\frac{d}{dt} P_t \right]_{t=0},$$

provided that the entries of the right side are well defined as

$$\lim_{h \rightarrow 0^+} \frac{P_h(x, y) - I(x, y)}{h}.$$

When the state space is finite, the steps in the above derivation can be justified rigorously, and the following theorem can be proved. The statement of the theorem holds also for sufficiently regular continuous-time Markov chains on countably infinite spaces, but not for all.

Theorem 9.6. *For any transition semigroup $(P_t)_{t \in \mathbb{R}_+}$ of a continuous-time Markov chain on a finite state space, the generator matrix Q exists, satisfies Kolmogorov's backward and forward differential equations*

$$\frac{d}{dt} P_t = Q P_t, \quad \frac{d}{dt} P_t = P_t Q,$$

and determines the transition matrices of the chain via

$$P_t = \sum_{n=0}^{\infty} \frac{t^n Q^n}{n!}.$$

Example 9.7 (Poisson modulated chain). Let $X_t = Y_{N(t)}$ be a Poisson modulated chain as in Example 9.5 where $(Y_n)_{n \in \mathbb{Z}_+}$ is a discrete-time Markov chain on state space S with transition matrix P and $(N(t))_{t \in \mathbb{R}_+}$ is a Poisson process with intensity λ which is independent of $(Y_n)_{n \in \mathbb{Z}_+}$. Then the generator matrix of $(X_t)_{t \in \mathbb{R}_+}$ is given by formula (9.4) as

$$Q = \lambda(P - I).$$

■

Example 9.8 (Satellite). Consider the $\{0, 1\}$ -valued continuous-time Markov chain describing the state of a satellite in Example 9.3. By noting that this is a Poisson modulated chain with $(N(t))$ being a Poisson process with intensity μ , and (Y_n) being a discrete-time Markov chain on $\{0, 1\}$ with transition matrix

$$P = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},$$

we see by (9.4) that the generator matrix equals

$$Q = \begin{bmatrix} 0 & 0 \\ \mu & -\mu \end{bmatrix}.$$

Example 9.9 (Poisson process). Similarly, any Poisson process (recall Example 9.4) of rate λ can be seen as a special instance of a Poisson modulated chain where $Y_n = n$ is a Markov chain on \mathbb{Z}_+ which deterministically moves one step up at every discrete time step. Hence by (9.4) the generator matrix of the Poisson process equals

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & 0 & \cdots \\ \vdots & & & \ddots & \ddots & \ddots \end{bmatrix}$$

In the above examples, each generator matrix has zero row sums and non-negative offdiagonal entries. This is a general fact. Indeed it follows from

$$Q(x, y) = \lim_{h \rightarrow 0^+} \frac{P_h(x, y) - I(x, y)}{h},$$

that the row sums of Q must be zero, because P_h and I have unit row sums. The above formula also implies that the offdiagonal entries with $x \neq y$ satisfy

$$Q(x, y) = \lim_{h \rightarrow 0^+} \frac{P_h(x, y)}{h},$$

and are hence nonnegative. Hence it also follows that the diagonal entries of Q are given by

$$Q(x, x) = -\sum_{y \neq x} Q(x, y).$$

9.4 Transition semigroup generators

We discuss square matrices $A : S \times S \rightarrow \mathbb{R}$ with rows and columns index by a countable (finite or countably infinite) state space S . Such a matrix is called *bounded* if there exists a finite constant c such that

$$\sum_{y \in S} |A(x, y)| \leq c \quad \text{for all } x \in S,$$

and the smallest such upper bound is denoted² by

$$\|A\| = \sup_{x \in S} \sum_{y \in S} |A(x, y)|.$$

²The *supremum* of a nonempty set of real numbers A is defined as the smallest (possibly ∞) upper bound of A . The supremum of A is denoted $\sup A$, and the supremum of a set of numbers $a(x)$ indexed by $x \in S$ by $\sup_{x \in S} a(x)$. For finite sets $\sup A = \max A$.

In matrix terms this means that the row sums of the absolute values of A are bounded from above. All square matrices on a finite state space are bounded. On the other hand, on an infinite state space there exist unbounded matrices with bounded entries (for example, the matrix with all entries equal to one). By definition, every transition matrix P on S is bounded because

$$\|P\| = \max_{x \in S} \sum_{y \in S} |P(x, y)| = \max_{x \in S} \sum_{y \in S} P(x, y) = 1.$$

Theorem 9.10. *The map $A \mapsto \|A\|$ is a norm in the sense that for all matrices A, B and all constants c ,*

- (i) $\|A\| \geq 0$, where equality holds if and only if $A = 0$,
- (ii) $\|A + B\| \leq \|A\| + \|B\|$,
- (iii) $\|cA\| = |c|\|A\|$.

The map also satisfies

- (iv) $|A(x, y)| \leq \|A\|$ for all x, y ,
- (v) $\|AB\| \leq \|A\|\|B\|$.

Proof. (i) is clear. (ii) and (iii) can be verified (details as an exercise) by noting that $\sup_x (a(x) + b(x)) \leq \sup_x a(x) + \sup_x b(x)$, and $\sup_x ca(x) \leq |c| \sup_x |a(x)|$. (iv) is clear. To verify (v), note that for all x ,

$$\begin{aligned} \sum_z |AB(x, z)| &= \sum_z \left| \sum_y A(x, y)B(y, z) \right| \\ &\leq \sum_z \sum_y |A(x, y)| |B(y, z)| \\ &= \sum_y |A(x, y)| \sum_z |B(y, z)| \\ &\leq \sum_y |A(x, y)| \|B\| \\ &\leq \|A\| \|B\| \end{aligned}$$

so that $\|AB\| = \sup_x \sum_z |AB(x, z)| \leq \|A\| \|B\|$. \square

A *generator matrix* on a countable state space S is a function $Q : S \times S \rightarrow \mathbb{R}$ such that $Q(x, y) \geq 0$ for all $x \neq y$, $\sum_{y \neq x} Q(x, y) < \infty$ for all x , and $\sum_y Q(x, y) = 0$ for all x . As usual, such a matrix is considered a finite or infinite square matrix.

Theorem 9.11. *For any bounded generator matrix Q , the matrix exponential $P_t = e^{tQ}$ is well defined for all $t \geq 0$, and the collection $(P_t)_{t \in \mathbb{R}_+}$ is a transition semigroup on S , which solves the differential equations*

$$\frac{d}{dt} P_t = Q P_t \quad \text{and} \quad \frac{d}{dt} P_t = P_t Q.$$

Proof. For $s, t \geq 0$, note that because the matrices sQ and tQ commute, it follows that

$$P_s P_t = e^{sQ} e^{tQ} = e^{sQ+tQ} = e^{(s+t)Q} = P_{s+t}$$

Observe next that

$$\frac{P_{t+h} - P_t}{h} = \frac{1}{h} \left(\sum_{n=0}^{\infty} \frac{(t+h)^n}{n!} Q^n - \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n \right) = \sum_{n=1}^{\infty} \left(\frac{(t+h)^n - t^n}{h} \right) \frac{Q^n}{n!}.$$

By applying the binomial theorem it is possible to verify that for $|h| \leq 1$,

$$\left\| \left(\frac{(t+h)^n - t^n}{h} \right) \frac{Q^n}{n!} \right\| = \left| \frac{(t+h)^n - t^n}{h} \right| \left\| \frac{Q^n}{n!} \right\| \leq (t+1)^n \frac{\|Q\|^n}{n!}.$$

Because the right side above is summable with respect to n , we may take the limit with respect to $h \rightarrow 0$ inside the sum to conclude with the help of l'Hôpital's rule that

$$\lim_{h \rightarrow 0} \frac{P_{t+h} - P_t}{h} = \sum_{n=1}^{\infty} \lim_{h \rightarrow 0} \left(\frac{(t+h)^n - t^n}{h} \right) \frac{Q^n}{n!} = \sum_{n=1}^{\infty} n t^{n-1} \frac{Q^n}{n!} = \sum_{n=0}^{\infty} t^n \frac{Q^{n+1}}{n!}.$$

Hence the entrywise matrix derivative

$$\frac{d}{dt} P_t = \sum_{n=0}^{\infty} t^n \frac{Q^{n+1}}{n!} \tag{9.5}$$

exists at every $t \geq 0$ (as a right-sided derivative for $t = 0$).

Next, it appears clear that

$$\sum_{n=0}^{\infty} t^n \frac{Q^{n+1}}{n!} = Q \sum_{n=0}^{\infty} t^n \frac{Q^n}{n!}, \tag{9.6}$$

but for infinite matrices some care must be taken to justify the interchange of two infinite sums, the one with respect to n displayed above, and the other hidden inside the matrix product. We compute the matrix entry of the left side for row x and column y as

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^{n+1}(x, y) &= \sum_{n=0}^{\infty} \frac{t^n}{n!} \sum_{z \in S} Q(x, z) Q^n(z, y) \\ &= \sum_{z \in S} Q(x, z) \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n(z, y) \\ &= \sum_{z \in S} Q(x, z) e^{tQ}(z, y) \\ &= (Q e^{tQ})(x, y), \end{aligned}$$

where the change of the summation order is justified because

$$\sum_{n=0}^{\infty} \sum_{z \in S} \frac{t^n}{n!} |Q(x, z)Q^n(z, y)| \leq \sum_{n=0}^{\infty} \sum_{z \in S} \frac{t^n}{n!} |Q(x, z)| \|Q\|^n \leq \|Q\| \sum_{n=0}^{\infty} \frac{t^n}{n!} \|Q\|^n$$

Because the above equation holds for all x and y , we see that the entrywise matrix equation (9.6) is indeed true. By combining (9.5) and (9.6), we obtain $\frac{d}{dt}P_t = QP_t$. A similar reasoning can be used to verify $\frac{d}{dt}P_t = P_tQ$. \square

Chapter 10

Analysis of Markov jump processes

10.1 Jump rates and jump probabilities

The *first jump instant* (*ensimmäinen hyppyhetki*) of a continuous-time Markov chain is denoted by

$$T = \min\{t \geq 0 : X_t \neq X_0\}, \quad (10.1)$$

where we set $T = \infty$ if (X_t) never leaves its initial state.¹ The extended random number $T \in [0, \infty]$ tells when the Markov chain first exits its initial state. The *jump rate* (*hyppyvauhti*) of the chain in state x is

$$\lambda(x) = \frac{1}{\mathbb{E}(T \mid X_0 = x)},$$

and we set $\lambda(x) = 0$ when the denominator is infinite.

The following result tells that a continuous-time Markov chain spends an exponentially distributed random time in every state it visits. Here we interpret an exponential distribution $\text{Exp}(0)$ with rate zero as the distribution of a random variable which is infinite with probability one.

Theorem 10.1. *The first jump instant T of a continuous-time Markov chain (X_t) started at state x is exponentially distributed with rate parameter $\lambda(x)$.*

Proof. By applying an extended Markov property (9.2) we can verify that

$$\begin{aligned} \mathbb{P}(T > t + h \mid T > t) &= \mathbb{P}(X_u = x \forall u \in [t, t + h] \mid X_s = x \forall s \in [0, t]) \\ &= \mathbb{P}(X_u = x \forall u \in [t, t + h] \mid X_t = x) \\ &= \mathbb{P}(X_u = x \forall u \in [0, h] \mid X_0 = x) \\ &= \mathbb{P}(T > h). \end{aligned}$$

This means that the distribution of T is memoryless, so that the tail distribution function $\phi(t) = \mathbb{P}(T > t)$ satisfies $\phi(t + h) = \phi(t)\phi(h)$ for all $t, h \geq 0$. Because ϕ

¹We follow the usual convention that the paths of all continuous-time processes are right-continuous. This means that the state of a process at a jump instant is the state where the process jumps at the time.

is nonincreasing, it follows by the theory of Cauchy's functional equations that ϕ must be of the form

$$\phi(t) = e^{-\lambda t}$$

for some $\lambda \in [0, \infty)$. In case $\lambda > 0$, this shows that the random variable T is $\text{Exp}(\lambda)$ -distributed. In case $\lambda = 0$, it follows that

$$\mathbb{P}(T = \infty) = \lim_{n \rightarrow \infty} \mathbb{P}(T > n) = 1,$$

which corresponds to an exponential distribution with rate parameter zero. \square

Using Theorem 10.1 we may characterise the behaviour of a continuous-time Markov chain over time. A chain starting at state x :

- spends a random $\text{Exp}(\lambda(x))$ -distributed time in state x ,
- jumps from x to state y with probability $P_*(x, y)$,
- spends a random $\text{Exp}(\lambda(y))$ -distributed time in state y ,
- jumps from y to state z with probability $P_*(y, z)$,
- ...

The chain evolves as above as long as it visits states with a nonzero jump rate. If the chain hits a state with jump rate zero, it remains stuck there.

The number $P_*(x, y)$ tells the probability at which the chain enters y when it leaves x . By the Markov property, the new state is selected independently of its past trajectory. The square matrix P_* with rows and columns indexed by the states $x, y \in S$ is called the *jump probability matrix* (*hyppytodennäköisyysmatriisi*). Because $P_*(x, y) \geq 0$ for all $x, y \in S$, and

$$\sum_{y \in S} P_*(x, y) = 1 \quad \text{for all } x \in S,$$

we see that P_* is a transition matrix on S . In addition, the diagonal entries of P_* are zero, because the chain changes state on a jump instant. For states with jump rate $\lambda(x) = 0$, it is usual to define the jump rates as $P_*(x, y) = 1(x = y)$, although these entries have no effect on the behaviour of the chain.

10.2 Determining the generator matrix

Continuous-time Markov chains were described in Section 10.1 using jump rates $\lambda(x)$ and jump probabilities $P_*(x, y)$. This description appears quite different from the definition in the previous chapter which discusses t -step transition probabilities $P_t(x, y)$ and generator matrices. In this section we see how the transition semigroup and the generator matrix can be determined from jump rates and jump probabilities.

Overclocking (*ylikellottaminen*) is a technique where we first generate a background Poisson process of intensity α such that $\alpha \geq \lambda(x)$ for all x . The jump instants of this Poisson process are used to trigger all possible transitions of the Markov chain. However, the background Poisson process generates too many jump instants. The effect of this can be compensated by allowing the Markov chain to stay put at some of the triggering events. Let us define a matrix \hat{P} by

$$\hat{P}(x, y) = \frac{\lambda(x)}{\alpha} P_*(x, y) + \left(1 - \frac{\lambda(x)}{\alpha}\right) I(x, y), \quad (10.2)$$

where I is the identity matrix on the state space S . The entries of \hat{P} are nonnegative and the row sums are one, so that \hat{P} is a transition matrix. This matrix represents a discrete-time Markov chain where at every time step we flip a coin, and with probability $\frac{\lambda(x)}{\alpha}$ we move according to transition matrix P_* , and with probability $1 - \frac{\lambda(x)}{\alpha}$ we move according to transition matrix I (which means that we do not move anywhere). Now let us define

$$X_t = Y_{N(t)}, \quad t \in \mathbb{R}_+,$$

where (Y_0, Y_1, \dots) is a discrete-time Markov chain with transition matrix \hat{P} , which is independent of the underlying Poisson process $N(t)$. Then we saw in Example 9.5 that (X_t) is a continuous-time Markov chain (a Poisson modulated chain) with t -step transition matrices

$$P_t = \sum_{n=0}^{\infty} e^{-\alpha t} \frac{(\alpha t)^n \hat{P}^n}{n!}, \quad t \in \mathbb{R}_+. \quad (10.3)$$

The above formula determines the transition matrices of the chain. However, it is slightly inconvenient because it involves the auxiliary parameter α with no physical meaning. The following result provides a more convenient description.

Theorem 10.2. *For any continuous-time Markov chain with bounded jump rates $\lambda(x)$ and arbitrary jump probabilities $P_*(x, y)$, the generator matrix Q equals*

$$Q(x, y) = \begin{cases} \lambda(x)P_*(x, y), & x \neq y, \\ -\lambda(x), & x = y, \end{cases} \quad (10.4)$$

and the t -step transition matrices are given by $P_t = e^{tQ}$.

Proof. By (10.3), and applying the formula $e^{-\alpha t I} e^{\alpha t \hat{P}} = e^{-\alpha t I + \alpha t \hat{P}}$ (which is valid because the matrices $-\alpha t I$ and $\alpha t \hat{P}$ commute), the t -step transition matrix P_t can be written as

$$P_t = e^{-\alpha t} e^{\alpha t \hat{P}} = e^{-\alpha t} I e^{\alpha t \hat{P}} = e^{-\alpha t I} e^{\alpha t \hat{P}} = e^{\alpha t \hat{P} - \alpha t I} = e^{tQ},$$

where

$$Q = \alpha(\hat{P} - I).$$

The matrix Q is the generator matrix of the chain (as seen earlier). By using the definition of \hat{P} in (10.2) we see that

$$\begin{aligned} Q(x, y) &= \alpha(\hat{P}(x, y) - I(x, y)) \\ &= \alpha\left(\frac{\lambda(x)}{\alpha}P_*(x, y) + \left(1 - \frac{\lambda(x)}{\alpha}\right)I(x, y) - I(x, y)\right) \\ &= \lambda(x)P_*(x, y) - \lambda(x)I(x, y). \end{aligned}$$

□

Formula (10.4) also implies a simple way to determine the jump rates $\lambda(x)$ and jump probabilities $P_*(x, y)$ from the generator matrix Q . Namely, because P_* has zero diagonal and unit rows sums, it follows that

$$\lambda(x) = \sum_{y \neq x} Q(x, y) \quad \text{and} \quad P_*(x, y) = \frac{Q(x, y)}{\sum_{y \neq x} Q(x, y)}.$$

For a state x where the total jump rate $\lambda(x) = 0$, the above formula for P_* is not well defined, but this is naturally the case because such x is an absorbing state. For such x , one can define P_* to be an arbitrary transition matrix, for example I , without affecting the statistical behaviour of the chain.

10.3 Memoryless races

To prepare ourselves to construct continuous-time Markov chains for more complicated models, in this section we will analyse some features of independent exponential distributions.

A set of competitors labeled $i \in I$ participate in a race. The time of competitor i equals T_i and is exponentially distributed with rate parameter λ_i . We assume that the times of the competitors are independent. The winning time of the race equals

$$T_{\min} = \min_{i \in I} T_i$$

and the label of the winner is

$$I_{\min} = \arg \min_{i \in I} T_i.$$

Being independent random numbers with a continuous distribution, the times are distinct from each other with probability one, so that the winner of the race is uniquely defined. The following, slightly counterintuitive, result tells that information about who wins the race tells nothing about the winning time. This magical property does not hold in general for other distributions beside the exponential.

Theorem 10.3. If $\lambda = \sum_{i \in I} \lambda_i < \infty$ (e.g. when I is finite), then T_{\min} is $\text{Exp}(\lambda)$ -distributed with rate parameter λ , and I_{\min} is distributed according to

$$\mathbb{P}(I_{\min} = i) = \frac{\lambda_i}{\lambda}, \quad i \in I.$$

Moreover, T_{\min} and I_{\min} are independent.

Proof. Let us first determine the distribution of the winning time. Because

$$\mathbb{P}(T_{\min} > t) = \mathbb{P}(T_i > t \text{ for all } i \in I) = \prod_{i \in I} e^{-\lambda_i t} = e^{-\lambda t},$$

we may conclude that $T_{\min} =_{\text{st}} \text{Exp}(\lambda)$.

Competitor i wins the race precisely when $T_i < T'$, where random number $T' = \min_{j \neq i} T_j$ tells the best time among the rivals of i . By the previous part, we see that $T' =_{\text{st}} \text{Exp}(\lambda')$ with $\lambda' = \sum_{j \neq i} \lambda_j$. Because T_i and T' are independent from each other, we see that

$$\mathbb{P}(T_{\min} > t, I_{\min} = i) = \mathbb{P}(T_i > t, T_i < T').$$

By writing the probability on the right as

$$\mathbb{P}(T_i > t, T_i < T') = \mathbb{E}h(T_i, T'),$$

where $h(t_i, t') = 1(t_i > t)1(t_i < t')$, we find that by applying the independence of T_i and T' that

$$\begin{aligned} \mathbb{P}(T_i > t, T_i < T') &= \int_0^\infty \int_0^\infty h(t_i, t') \lambda_i e^{-\lambda_i t_i} \lambda' e^{-\lambda' t'} dt_i dt' \\ &= \int_0^\infty \int_0^\infty 1(t_i > t) 1(t_i < t') \lambda_i e^{-\lambda_i t_i} \lambda' e^{-\lambda' t'} dt_i dt' \\ &= \int_0^\infty 1(t_i > t) \lambda_i e^{-\lambda_i t_i} \left(\int_{t_i}^\infty \lambda' e^{-\lambda' t'} dt' \right) dt_i \\ &= \int_0^\infty 1(t_i > t) \lambda_i e^{-\lambda_i t_i} e^{-\lambda' t_i} dt_i \\ &= \int_t^\infty \lambda_i e^{-\lambda t_i} dt_i \\ &= \frac{\lambda_i}{\lambda} e^{-\lambda t}. \end{aligned}$$

From this we conclude that

$$\mathbb{P}(T_{\min} > t, I_{\min} = i) = \mathbb{P}(T_{\min} > t) \frac{\lambda_i}{\lambda}.$$

By substituting $t = 0$ to the above formula, we see that $\mathbb{P}(I_{\min} = i) = \frac{\lambda_i}{\lambda}$. We can rewrite the above formula as

$$\mathbb{P}(T_{\min} > t, I_{\min} = i) = \mathbb{P}(T_{\min} > t) \mathbb{P}(I_{\min} = i),$$

from which we obtain the independence of T_{\min} and I_{\min} . □

10.4 Constructing Markov chain models

To illustrate how continuous-time Markov chain models are constructed in practice, we will carefully study the following example. This illustrates basic principles which also work for much more complicated situations.

Example 10.4 (Two machines). A factory has two machines, and each of them remains operational for an expected time of $1/\lambda = 40$ weeks, independently of each other. When a machine breaks down, its repair takes on average $1/\mu = 2$ weeks. All operation times and repair times are assumed to be mutually independent and exponentially distributed. We denote

$$X_t = \text{Number of broken machines at time } t.$$

We will analyse the process by inspecting how the first jump takes place. Let us denote the first jump time by $T = \min\{t \geq 0 : X_t \neq X_0\}$, and the state after the first jump by X_T .

In state 0, both machines are operational. Denote by L_i the remaining operational time of machine $i = 1, 2$. Then the waiting time until the next jump equals $T = \min(L_1, L_2)$, and by Theorem 10.3 we see that $T \stackrel{\text{st}}{=} \text{Exp}(2\lambda)$. At the time of jump, one of the machines gets repaired, and the remaining operational time of the other machine, by the memoryless property of the exponential distribution, still follows the $\text{Exp}(\lambda)$ -distribution as if had just been repaired. Hence from time T onwards, the process (X_t) behaves just as if it were started afresh in state 1.

In state 1, one machine is operational and the other one broken. Let us label the machines so that the operational machine has label 1 and the broken machine label 2. Then the remaining operational time L_1 of machine 1 is $\text{Exp}(\lambda)$ -distributed, and the remaining repair time of the broken machine is $\text{Exp}(\mu)$ -distributed. What happens next depends on whether $L_1 < M_2$ or $L_1 > M_2$ (the probability of equality is zero).

- If $L_1 < M_2$, the machine 1 breaks down before machine 2 gets repaired, so that the system moves to state 2. When this happens, then after the breakdown event the remaining repair time of machine 2 is still $\text{Exp}(\mu)$ -distributed, by the memoryless property. Hence after the breakdown, the system behaves as if it were just started in state 2.
- If $L_1 > M_2$, then machine 2 gets repaired before machine 1 breaks down. Then the system moves to state 0, and afterwards the system behaves just as if it were freshly started in state 0.

Which of the above alternatives realises corresponds to a race of two memoryless runners. By Theorem 10.3, the winning time of the race $T = \min(L_1, M_2)$ is $\text{Exp}(\lambda + \mu)$ -distributed, and independently of T , the process moves to state 0 with probability $\mu/(\lambda + \mu)$ and into state 2 with probability $\lambda/(\lambda + \mu)$.

In state 2, both machines are broken. The repair times are independent and $\text{Exp}(\mu)$ -distributed, so that the first jump instant is $\text{Exp}(2\mu)$ -distributed, and

at this time instant the process jumps into state 1. As above, after the jump instant, the process evolves just as if it were freshly started in state 1.

By gathering together the above observations we may conclude that (X_t) is a continuous-time Markov chain on state space $S = \{0, 1, 2\}$ with jump rates $\lambda(0) = 2\lambda$, $\lambda(1) = \lambda + \mu$, $\lambda(2) = 2\mu$, and jump probabilities given by

$$P_* = \begin{bmatrix} P_*(0,0) & P_*(0,1) & P_*(0,2) \\ P_*(1,0) & P_*(1,1) & P_*(1,2) \\ P_*(2,0) & P_*(2,1) & P_*(2,2) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{\mu}{\lambda+\mu} & 0 & \frac{\lambda}{\lambda+\mu} \\ 0 & 1 & 0 \end{bmatrix}.$$

Hence by Theorem 10.2, the generator matrix equals

$$Q = \begin{bmatrix} -2\lambda & 2\lambda & 0 \\ \mu & -(\lambda + \mu) & \lambda \\ 0 & 2\mu & -2\mu \end{bmatrix} = \begin{bmatrix} -0.050 & 0.050 & 0 \\ 0.500 & -0.525 & 0.025 \\ 0 & 1.000 & -1.000 \end{bmatrix}.$$

The transition diagram of the chain is in Figure 10.1.

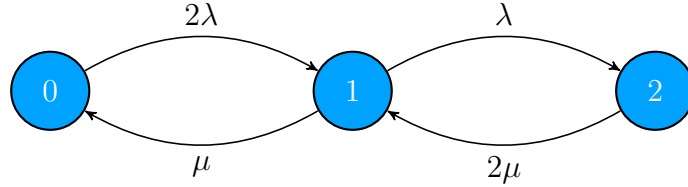


Figure 10.1: Two machines.

The 3-step transition matrix which tells the probabilities of states after three weeks can be computed (with the help of a computer) as

$$P_3 = e^{3Q} = \begin{bmatrix} 0.9259028 & 0.07267122 & 0.001425934 \\ 0.7267122 & 0.26404492 & 0.009242859 \\ 0.5703737 & 0.36971437 & 0.059911911 \end{bmatrix}.$$

Recalling that the states are indexed by $S = \{0, 1, 2\}$, we get the 3-week transition probabilities corresponding to initial state 0 (both machines operating) from the first row of P_3 . Alternatively, we use the initial distribution $\delta_0 = [1, 0, 0]$ being the Dirac measure at state 0, to compute

$$\begin{aligned} \delta_0 P_3 &= [1 \ 0 \ 0] \begin{bmatrix} 0.9259028 & 0.07267122 & 0.001425934 \\ 0.7267122 & 0.26404492 & 0.009242859 \\ 0.5703737 & 0.36971437 & 0.059911911 \end{bmatrix} \\ &= [0.9259028 \ 0.07267122 \ 0.001425934]. \end{aligned}$$

Hence both machines are operating 3 weeks after today with probability $P_3(0, 0) = 0.926$. The matrix exponential can be computed in R or Python as follows:

```

# R code
library(expm)
la <- 1/40
mu <- 1/2
t <- 3
Q <- matrix(0,3,3)
Q[1,] <- c(-2*la,2*la,0)
Q[2,] <- c(mu,-(1a+mu),1a)
Q[3,] <- c(0,2*mu,-2*mu)
P3 <- expm(t*Q)

# Python code
import numpy as np
from scipy.linalg import expm
la = 1.0/40
mu = 1.0/2
t = 3.0
Q = np.array([
    [-2*la, 2*la, 0],
    [mu, -1a-mu, 1a],
    [0, 2*mu, -2*mu]])
P3 = expm(t*Q)

```



10.5 Invariant distributions

A probability distribution π is an *invariant distribution* (*tasapainojakauma*) of a transition semigroup $(P_t)_{t \in \mathbb{R}_+}$ and a corresponding continuous-time Markov chain if $\pi P_t = \pi$ for all $t \geq 0$. In this case the distribution of X_t for a Markov chain started with a π -distributed initial state does not change over time. This means that the chain remains in statistical equilibrium.

Theorem 10.5. *The following are equivalent for a continuous-time Markov chain with bounded jump rates, and for any probability distribution π :*

- (i) π is an invariant distribution of the chain.
- (ii) $\pi Q = 0$, where Q is the generator matrix of the chain.
- (iii) $\pi \hat{P} = \pi$, where \hat{P} is a transition matrix defined by overclocking.

Because the row sums of Q are zero, the balance equation $\pi Q = 0$ can be written as

$$\sum_{x:x \neq y} \pi(x)Q(x,y) = \pi(y) \sum_{z:z \neq y} Q(y,z),$$

where the left side describes the long-term average rate of jumps into state y , and the right side the corresponding rate of out from y .

Proof. (i) \implies (ii). By differentiating the formula $\pi P_t(y) = \sum_x \pi(x)P_t(x,y)$ term by term with respect to t , we see by Kolmogorov's backward equation $\frac{d}{dt}P_t = QP_t$ that

$$\frac{d}{dt}(\pi P_t) = \pi \frac{d}{dt}P_t = \pi(QP_t) = (\pi Q)P_t. \quad (10.5)$$

If π is invariant, this implies that $0 = (\pi Q)P_t$. By substituting $t = 0$, we see that $0 = \pi Q P_0 = \pi Q I = \pi Q$.

(ii) \implies (i). If $\pi Q = 0$, formula (10.5) shows that $\frac{d}{dt}(\pi P_t) = 0$, so that $t \mapsto \pi P_t$ is constant over time. Therefore $\pi P_t = \pi P_0 = \pi$ for all $t \geq 0$, that is, π is invariant.

(ii) \implies (iii). The overclocked transition matrix \hat{P} is defined by (10.2), where $\alpha > 0$ satisfies $\alpha \geq \lambda(x)$ for all x . The definition directly shows that

$$\alpha(\hat{P} - I) = Q.$$

When $\pi Q = 0$, this implies that $\pi(\hat{P} - I) = 0$ so that $\pi\hat{P} = \pi$. By applying the above formula we can also verify the converse implication (iii) \implies (ii). \square

Example 10.6 (Two machines). The balance equations $\pi Q = 0$ for the generator matrix in Example 10.4 can be written as

$$\begin{aligned} -\pi(0)2\lambda + \pi(1)\mu + \pi(2) \cdot 0 &= 0, \\ \pi(0)2\lambda - \pi(1)(\lambda + \mu) + \pi(2)2\mu &= 0 \\ \pi(0) \cdot 0 + \pi(1)\lambda - \pi(2)2\mu &= 0 \end{aligned}$$

Together with the normalising equation $\pi(0) + \pi(1) + \pi(2) = 1$, we can solve the equilibrium distribution as

$$\pi = [p^2 \quad 2p(1-p) \quad (1-p)^2]$$

where $p = \frac{\mu}{\lambda + \mu}$. By substituting $p = 0.952381$, we get the solution

$$\pi = [0.907029478 \quad 0.090702948 \quad 0.002267574].$$

■

10.6 Convergence

Irreducibility for continuous-time Markov chains is defined in the same way as in discrete time. Recall that the transition diagram of a generator matrix Q and a corresponding continuous-time Markov chain is a directed graph with nodes being the states, and links being the ordered node (x, y) for which $Q(x, y) > 0$. A generator matrix Q and a corresponding Markov chain is *irreducible* if its transition diagram is strongly connected in the sense that for any distinct nodes x and y there exists a path from x to y in the transition diagram. For continuous-time Markov chains we never need to worry about periodicity issues, because all continuous-time chains are automatically aperiodic.

Theorem 10.7. *Any irreducible continuous-time Markov chain on a finite state space has a unique invariant distribution.*

Proof. When the state space is finite, the jump rates are always bounded. Hence we can choose $\alpha > 0$ so that $\alpha \geq \lambda(x)$ for all x , and define the overclocked transition matrix \hat{P} by (10.2). Indeed we can choose α is large that $\alpha > \lambda(x)$ for all x . In this case $\hat{P}(x, x) > 0$ for all x so that \hat{P} is aperiodic. Moreover, because the original continuous-time chain is irreducible, then one can verify that so is \hat{P} . Then an earlier theorem concerning discrete-time chains tells that \hat{P} has a unique invariant distribution π . Then Theorem 10.5 tells that this π is also the unique invariant distribution of the continuous-time chain. \square

Theorem 10.8. *Any irreducible continuous-time Markov chain has at most one invariant distribution. If π is an invariant distribution of such a chain, then*

$$\lim_{t \rightarrow \infty} P_t(x, y) = \pi(y) \quad \text{for every } x \in S.$$

Proof. See for example [Dur12, Theorem 4.4]. □

Theorem 10.9. *For any irreducible continuous-time Markov chain $(X_t)_{t \in \mathbb{R}_+}$ and any function $f : S \rightarrow \mathbb{R}$, as $t \rightarrow \infty$,*

$$\frac{1}{t} \int_0^t f(X_s) ds \rightarrow \sum_{x \in S} f(x) \pi(x)$$

with probability one, regardless of the initial state of the chain.

Proof. The result can be proved using overclocking to reduce the analysis to the discrete-time case. □

Chapter 11

Martingales and information processes

Martingales (*martingaalit*) are random processes for which the best prediction of a future value given past information is the present value. Martingales are central in economics because in efficient markets the prices of publicly exchangeable assets naturally satisfy the martingale property. Martingale theory also provides a powerful theoretical tool for studying the pathwise convergence of random sequences, stochastic integrals, and randomised algorithms.

11.1 Conditional expectation with respect to information

11.1.1 Definition for finite-state random variables

Conditional expectations can be treated from different points of view. We will first assume that X and Y are random numbers with values in a finite set. Then

$$\mathbb{E}(Y) = \sum_y y \mathbb{P}(Y = y)$$

is by definition the expected value of Y from an external observer's point of view, and

$$\mathbb{E}(Y | X = x) = \sum_y y \mathbb{P}(Y = y | X = x)$$

is the expected value of Y from the point of view of an insider who knows that the outcome of X equals x . In advanced prediction models it is useful to define a conditional expectation

$$\mathbb{E}(Y | X),$$

which corresponds to the value of Y expected by an insider who knows X , *from the viewpoint of an external observer who does not know what the insider knows*. The value of $\mathbb{E}(Y | X)$ depends on the realisation of X , so that $\mathbb{E}(Y | X)$ is a

random variable. On the other hand, all randomness related to $\mathbb{E}(Y | X)$ is due to the randomness of X , so that

$$\mathbb{E}(Y | X) = h(X)$$

for some deterministic function h . The value of this function at x equals the insider's expected value for Y on the event $\{X = x\}$, that is,

$$h(x) = \mathbb{E}(Y | X = x) = \sum_y y \mathbb{P}(Y = y | X = x). \quad (11.1)$$

Care must be taken with notations when working with conditional expectations. For example, we cannot substitute a random variable X to both sides of (11.1) because in general

$$h(X) \neq \mathbb{E}(Y) = \mathbb{E}(Y | X = X).$$

A correct way to interpret $h(X)$ is to first define a deterministic function h using formula (11.1) and then define a random variable $h(X)$ as a mapping $\omega \mapsto h(X(\omega))$ from the underlying probability space (Ω, \mathbb{P}) to the real numbers. The apparent conflict due to replacing x by random variable X in (11.1) is caused by the shorthand notation for the event $\{X = x\}$. When we recall that the event $\{X = x\}$ is a subset $\{\omega' \in \Omega : X(\omega') = x\}$ of the reference space Ω , then the realisation of $h(X)$ at point ω can be written as

$$h(X(\omega)) = \mathbb{E}(Y | \{\omega' : X(\omega') = X(\omega)\}).$$

This concept becomes clearer when investigating the following concrete example.

Example 11.1 (Poker hands). Two players both receive two cards from a standard deck of 52 cards. Denote by X_i the number aces in hand $i = 1, 2$. Elementary combinatorial reasoning (a good exercise) shows that

$$\mathbb{E}(X_2 | X_1 = k) = 2 \cdot \frac{4 - k}{50}, \quad k = 0, 1, 2.$$

Then from an external observer's viewpoint, the value of X_2 expected by

- player 1 (who knows X_1) equals $\mathbb{E}(X_2 | X_1) = 2 \cdot \frac{4 - X_1}{50}$,
- player 2 (who knows X_2) equals $\mathbb{E}(X_2 | X_2) = X_2$,
- the dealer (who knows nothing) equals $\mathbb{E}(X_2) = 2 \cdot \frac{4}{52}$.

■

Conditional expectations can also be defined with respect to vector-valued information. Let X_1, \dots, X_n ja Y be some random numbers which take on finitely many values. Then $\mathbb{E}(Y | X_1, \dots, X_n)$ is the value of Y expected by an

insider who knows the random vector $X = (X_1, \dots, X_n)$, seen from an outsider's point of view. This is mathematically defined as

$$\mathbb{E}(Y | X_1, \dots, X_n) = h(X_1, \dots, X_n),$$

where the deterministic function h is defined by

$$h(x) = \mathbb{E}(Y | X = x)$$

for those $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ such that $\mathbb{P}(X = x) > 0$.

11.1.2 Rules

We get introduced to the most important rules for computing with conditional expectations. First we assume that all random variables take on only finitely many values. We will use the shorthand $Z \in \sigma(X_1, \dots, X_n)$ to indicate that $Z = h(X_1, \dots, X_n)$ for some deterministic function h .

Example 11.2. Let us throw two dice and denote by X_i be the outcome of die $i = 1, 2$. Denote $Z_1 = X_1 + X_2$ and $Z_2 = X_1 - X_2$. Then obviously $Z_1, Z_2 \in \sigma(X_1, X_2)$. Moreover, $f(Z_1, Z_2) \in \sigma(X_1, X_2)$ for every deterministic function $f : \mathbb{Z}^2 \rightarrow \mathbb{R}$. On the other hand, $X_2 \in \sigma(X_1, Z_1)$ but $X_2 \notin \sigma(X_1)$. ■

Theorem 11.3. *The following rules are valid for all random variables with finitely many possible values.*

(i) *Unbiasedness:*

$$\mathbb{E}(\mathbb{E}(Y | X_1, \dots, X_n)) = \mathbb{E}(Y). \quad (11.2)$$

(ii) *Pulling out known factors:*

$$\mathbb{E}(ZY | X_1, \dots, X_n) = Z \mathbb{E}(Y | X_1, \dots, X_n) \quad (11.3)$$

for all $Z \in \sigma(X_1, \dots, X_n)$.

(iii) *Removing independent information:*

$$\mathbb{E}(Y | X_1, \dots, X_n) = \mathbb{E}(Y) \quad (11.4)$$

whenever Y and (X_1, \dots, X_n) are independent.

(iv) *Removing redundant information:*

$$\mathbb{E}(Y | X_1, \dots, X_n, X'_1, \dots, X'_n) = \mathbb{E}(Y | X_1, \dots, X_n) \quad (11.5)$$

whenever $X'_1, \dots, X'_n \in \sigma(X_1, \dots, X_n)$.

Proof. Denote by $S = \{x \in \mathbb{R}^n : \mathbb{P}(X = x) > 0\}$ the set of possible values of $X = (X_1, \dots, X_n)$ and define

$$h(x) = \mathbb{E}(Y | X = x), \quad x \in S.$$

(i) For all $x \in S$,

$$h(x) = \sum_y y \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}.$$

Hence

$$\begin{aligned} \mathbb{E}(\mathbb{E}(Y | X)) &= \mathbb{E}h(X) = \sum_{x \in S} h(x) \mathbb{P}(X = x) \\ &= \sum_{x \in S} \sum_y y \mathbb{P}(X = x, Y = y) \\ &= \sum_y y \mathbb{P}(Y = y) = \mathbb{E}(Y). \end{aligned}$$

(ii) If $Z \in \sigma(X_1, \dots, X_n)$, then $Z = \phi(X)$ for some deterministic function $\phi(x)$. Then for all $x \in S$,

$$\tilde{h}(x) = \mathbb{E}(ZY | X = x) = \phi(x)\mathbb{E}(Y | X = x) = \phi(x)h(x),$$

so that

$$\mathbb{E}(ZY | X) = \tilde{h}(X) = \phi(X)h(X) = Z\mathbb{E}(Y | X).$$

(iii) If X and Y are independent, then for all $x \in S$,

$$h(x) = \sum_y y \mathbb{P}(Y = y | X = x) = \sum_y y \mathbb{P}(Y = y) = \mathbb{E}(Y).$$

Therefore $\mathbb{E}(Y | X) = h(X) = \mathbb{E}(Y)$.

(iv) Denote $X' = (X'_1, \dots, X'_n)$. Then $X' = \phi(X)$ for some deterministic function ϕ and the set of possible values of random vector (X, X') can be expressed as $\tilde{S} = \{(x, \phi(x)) : \mathbb{P}(X = x) > 0\}$. Moreover, for all $(x, x') \in \tilde{S}$

$$\tilde{h}(x, x') = \mathbb{E}(Y | X = x, X' = x') = \mathbb{E}(Y | X = x) = h(x).$$

Hence

$$\mathbb{E}(Y | X, X') = \tilde{h}(X, X') = h(X) = \mathbb{E}(Y | X).$$

□

11.1.3 General definition

Defining the conditional expectation with respect to a random variable taking values in an uncountable state space is not straightforward because the function $h(x) = \mathbb{E}(Y | X = x)$ cannot be defined using formula (11.1). To arrive at a general definition, let us first write down a generalization of (11.2).

Lemma 11.4 (Conditional unbiasedness). *For any random vector X and random number Y with finitely many possible values, the random number $\hat{Y} = \mathbb{E}(Y|X)$ satisfies*

$$\mathbb{E}(\hat{Y} | X \in A) = \mathbb{E}(Y | X \in A)$$

for all A such that $\mathbb{P}(X \in A) > 0$.

Proof. By applying (11.2) and (11.3), we see that the indicator random variable $Z = 1(X \in A) \in \sigma(X)$ satisfies

$$\mathbb{E}(\hat{Y}Z) = \mathbb{E}(\mathbb{E}(Y|X)Z) = \mathbb{E}(\mathbb{E}(YZ|X)) = \mathbb{E}(YZ),$$

so that

$$\mathbb{E}(\hat{Y} | X \in A) = \frac{\mathbb{E}(\hat{Y}1(X \in A))}{\mathbb{P}(X \in A)} = \frac{\mathbb{E}(\hat{Y}Z)}{\mathbb{P}(X \in A)} = \frac{\mathbb{E}(YZ)}{\mathbb{P}(X \in A)} = \mathbb{E}(Y | X \in A).$$

□

A Russian mathematician Andrey Kolmogorov (1903–1987) introduced in 1933 a general definition of conditional expectation based on the conditional unbiasedness property. This definition is valid for any \mathbb{R}^n -valued random vectors, and is based on the following theorem.

Theorem 11.5. *If $\mathbb{E}|Y| < \infty$, then there exists a unique (with probability one) random number $\hat{Y} \in \sigma(X_1, \dots, X_n)$ such that $\mathbb{E}|\hat{Y}| < \infty$ and*

$$\mathbb{E}(\hat{Y} | X \in A) = \mathbb{E}(Y | X \in A)$$

for all $A \subset \mathbb{R}^n$ such that $\mathbb{P}(X \in A) > 0$.

The proof of the above theorem [Wil91, Theorem 9.2] requires technical preliminaries which are treated in advanced courses of probability theory and analysis. The conditional expectation of a random number Y with respect to information (X_1, \dots, X_n) is then defined by

$$\mathbb{E}(Y | X_1, \dots, X_n) = \hat{Y}$$

where \hat{Y} is the random number appearing in Theorem 11.5. The theorem does not provide an explicit formula for \hat{Y} . However, this is usually not a problem because in practice the functional form can be determined from the context. From a theoretical point of view, it usually suffices to know rules of computation with conditional expectations. These are presented next [Wil91, Theorem 9.7]. A random number Y is called *integrable* (*integrõituv*) when $\mathbb{E}|Y| < \infty$.

Theorem 11.6. For integrable random numbers Y, Y_n, Z, YZ the rules of Theorem 11.3 are valid and moreover,

- *Conditional unbiasedness:* $\mathbb{E}(\mathbb{E}(Y|X_1, X_2)|X_1) = \mathbb{E}(Y|X_1)$.
- *Conditional pulling out known factors* $\mathbb{E}(Y|X_1, X_2) = \mathbb{E}(Y|X_1)$ for all $X_2 \perp\!\!\!\perp (X_1, Y)$
- *Linearity:* $\mathbb{E}(a_1Y_1 + a_2Y_2 | X) = a_1\mathbb{E}(Y_1|X) + a_2\mathbb{E}(Y_2|X)$.
- *Monotonicity:* $Y_1 \leq Y_2 \implies \mathbb{E}(Y_1|X) \leq \mathbb{E}(Y_2|X)$.
- *Monotone continuity:* Every nondecreasing random sequence $0 \leq Y_1 \leq Y_2 \leq Y_3 \leq \dots$ satisfies

$$Y_n \rightarrow Y \implies \mathbb{E}(Y_n|X) \rightarrow \mathbb{E}(Y|X).$$

- *Dominated continuity:* Every random sequence dominated by $|Y_n| \leq Z$ with $\mathbb{E}Z < \infty$ satisfies

$$Y_n \rightarrow Y \implies \mathbb{E}(Y_n|X) \rightarrow \mathbb{E}(Y|X).$$

11.2 Martingales

A random sequence (M_0, M_1, \dots) is a *martingale* (*martingaali*) with respect to random sequence (X_0, X_1, \dots) if

- (i) $\mathbb{E}|M_t| < \infty$,
- (ii) $M_t \in \sigma(X_0, \dots, X_t)$, and
- (iii) $\mathbb{E}(M_{t+1} | X_0, \dots, X_t) = M_t$.

A random sequence (M_t) satisfying (i) and (ii) is a *submartingale* (*alimartingaali*) if

- (iii)' $\mathbb{E}(M_{t+1} | X_0, \dots, X_t) \geq M_t$.

and a *supermartingale* (*ylimartingaali*) if

- (iii)'' $\mathbb{E}(M_{t+1} | X_0, \dots, X_t) \leq M_t$.

Property (i) is a technical condition which guarantees that the relevant expectations and conditional expectations are well defined. Property (ii) means that the state of a martingale at time t can be determined by the information (X_0, X_1, \dots, X_t) up to time t . Property (iii) is the essential martingale property, and says that the best predictor of M_{t+1} for an observer who knows information (X_0, X_1, \dots, X_t) equals M_t . In this sense the martingale property is natural for publicly traded assets in efficient markets: the expected tomorrow's value of an asset M_{t+1} based on available market data (X_0, \dots, X_t) up to time t is the present value M_t .

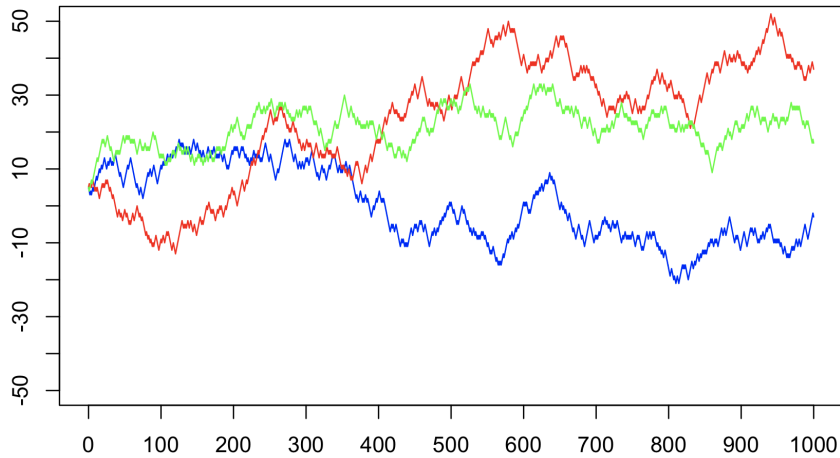


Figure 11.1: Three paths of a symmetric random walk.

Example 11.7 (Random walk). Let $S_t = S_0 + X_1 + \cdots + X_t$, where $\mathbb{E}|S_0| < \infty$ and X_1, X_2, \dots are identically distributed, and independent of each other and the initial state S_0 , with mean m . Then

$$\mathbb{E}|S_t| \leq \mathbb{E}|S_0| + t\mathbb{E}|X_1| < \infty$$

and $S_t \in \sigma(S_0, X_1, \dots, X_t)$ for all $t \geq 0$. Moreover,

$$\begin{aligned} \mathbb{E}(S_{t+1} | S_0, X_1, \dots, X_t) &= \mathbb{E}(S_t + X_{t+1} | S_0, X_1, \dots, X_t) \\ &= \mathbb{E}(S_t | S_0, X_1, \dots, X_t) + \mathbb{E}(X_{t+1} | S_0, X_1, \dots, X_t) \\ &= S_t + \mathbb{E}(X_{t+1}) \\ &= S_t + m. \end{aligned}$$

From this we see that the random walk (S_t) with respect to (S_0, X_1, X_2, \dots) is

$$\begin{cases} \text{supermartingale,} & \text{when } m < 0, \\ \text{martingale,} & \text{when } m = 0, \\ \text{submartingale,} & \text{when } m > 0. \end{cases}$$

In all cases, the centered random walk $t \mapsto S_t - mt$ a martingale (exercise). Three simulated paths of a symmetric random walk started at $S_0 = 5$ with X_t being uniformly distributed in $\{-1, +1\}$ are plotted in Figure 11.1 using the R code below.

```
# Simulate three paths of a symmetric random walk
s0 <- 5
T <- 1000
t <- 0:T
S <- matrix(0, 3, T+1)
for (omega in 1:3) {
  X <- sample(c(-1,+1), T, replace=TRUE)
  S[omega,] <- s0 + c(0,cumsum(X))
}
```

```

}

# Plot the paths
cols <- c("blue", "red", "green")
plot(NULL, xlim=c(0,T), ylim=c(-50,50), xaxt="n", yaxt="n", xlab="", ylab="")
axis(side=1,at=seq(0,T,by=100)); axis(side=2,at=seq(-50,50,by=10));
for (omega in 1:3) {
  lines(t, S[omega,], col=cols[omega])
}

```

■

Example 11.8 (Prediction martingale). If Z is an integrable random number and (X_0, X_1, \dots) some random sequence, then the best predictor of Z based on information up to time t equals

$$M_t = \mathbb{E}(Z | X_0, \dots, X_t).$$

The unbiasedness of conditional expectation implies $\mathbb{E}|M_t| \leq \mathbb{E}|Z| < \infty$. The definition of conditional expectation implies that $M_t \in \sigma(X_0, \dots, X_t)$. Conditional unbiasedness (Theorem 11.6) in turn implies that

$$\mathbb{E}(M_{t+1} | X_{0:t}) = \mathbb{E}(\mathbb{E}(Z | X_{0:t+1}) | X_{0:t}) = \mathbb{E}(Z | X_{0:t}) = M_t.$$

Therefore (M_0, M_1, \dots) is a martingale. This process is called the *prediction martingale* (*ennustusmartingaali*) of Z .

As a concrete example, consider the prediction martingale of a random number $Z =_{\text{st}} \text{Unif}(0, 1)$ with respect to information sequence $X_t = \lfloor 2^t Z \rfloor$, so that

$$X_1 = \begin{cases} 0, & 0 < Z < \frac{1}{2}, \\ 1, & \frac{1}{2} < Z < 1, \end{cases} \quad X_2 = \begin{cases} 0, & 0 < Z < \frac{1}{4}, \\ 1, & \frac{1}{4} < Z < \frac{1}{2}, \\ 2, & \frac{1}{2} < Z < \frac{3}{4}, \\ 3, & \frac{3}{4} < Z < 1, \end{cases}$$

and so on. Then by applying the fact that a uniformly distribution conditioned on some interval (a, b) is uniform on (a, b) , one can verify that the prediction martingale at time t equals

$$M_t = \frac{1}{2} \left(\frac{X_t}{2^t} + \frac{X_t + 1}{2^t} \right).$$

Three simulated trajectories of (M_t) are plotted in the figure below, using the R code below.

```

# Simulate three paths of a prediction martingale
t <- 0:10
M <- matrix(0, 3, 11)
for (omega in 1:3) {
  Z <- runif(1)
  X <- floor(2^t*Z)
  M[omega,] <- (X+1/2)/2^t
}

```

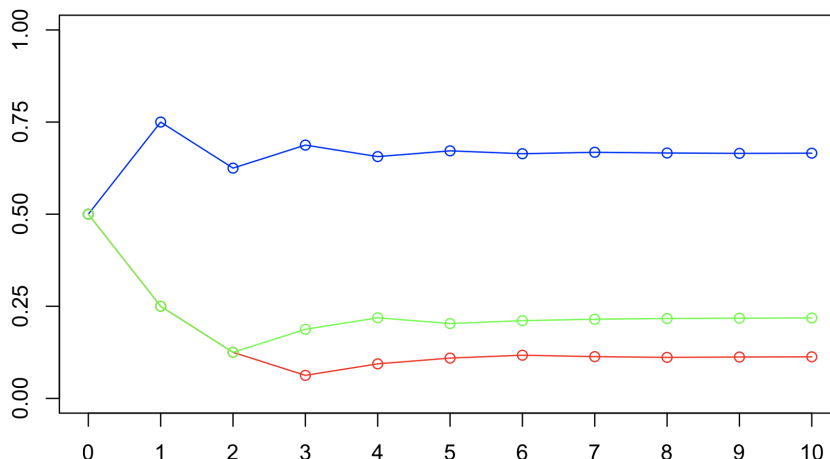


Figure 11.2: Three paths of a prediction martingale.

```
# Plot the paths
cols <- c("blue", "red", "green")
plot(NULL, xlim=c(0,10), ylim=c(0,1), xaxt="n", yaxt="n", xlab="", ylab="")
axis(side=1,at=t); axis(side=2,at=seq(0,1,by=1/4));
for (omega in 1:3) {
  lines(t, M[omega,], col=cols[omega])
  points(t,M[omega,], col=cols[omega])
}
```



11.3 Properties of martingales

It is common to call (M_t) a martingale without mentioning the information process in the background. In this case it is meant that (M_t) is a martingale with respect to itself. The following result describes the role of the information process in the definition of a martingale.

Theorem 11.9. *If (M_0, M_1, \dots) is a martingale with respect to (X_0, X_1, \dots) , then it is a martingale also with respect to itself.*

Proof. Under the assumption of the theorem, it is clear that $\mathbb{E}|M_t| < \infty$ for all t . Moreover, trivially $M_t \in \sigma(M_0, \dots, M_t)$. Now by denoting $M_{0:t} = (M_0, \dots, M_t)$ and $X_{0:t} = (X_0, \dots, X_t)$, we see by applying conditional unbiasedness (Theorem 11.6) that

$$\mathbb{E}(M_{t+1} | M_{0:t}) = \mathbb{E}(\mathbb{E}(M_{t+1} | X_{0:t}, M_{0:t}) | M_{0:t}).$$

Because $M_{0:t} \in \sigma(X_{0:t})$, it follows by removing redundant information (Theorem 11.6) and applying the martingale property that

$$\mathbb{E}(M_{t+1} | X_{0:t}, M_{0:t}) = \mathbb{E}(M_{t+1} | X_{0:t}) = M_t.$$

By combining the above two formulas we find that

$$\mathbb{E}(M_{t+1} | M_{0:t}) = \mathbb{E}(M_t | M_{0:t}) = M_t.$$

□

The following result characterizes how the expectation of a martingales evolve over time.

Theorem 11.10. *The map $t \mapsto \mathbb{E}(M_t)$ is¹*

$$\begin{cases} \text{increasing,} & \text{when } (M_t) \text{ is a submartingale,} \\ \text{constant,} & \text{when } (M_t) \text{ is a martingale,} \\ \text{decreasing,} & \text{when } (M_t) \text{ is a supermartingale.} \end{cases}$$

Proof. If (M_t) is a submartingale with respect to (X_t) , then

$$\mathbb{E}(M_{t+1} | X_0, \dots, X_t) \geq M_t.$$

By the unbiasedness and monotonicity of conditional expectations we see that

$$\mathbb{E}(M_{t+1}) = \mathbb{E}(\mathbb{E}(M_{t+1} | X_0, \dots, X_t)) \geq \mathbb{E}(M_t).$$

The cases for martingales and supermartingales are obtained analogously. □

Although martingales remain constant in expectation by Theorem 11.10, the statistical behavior of a martingale may nevertheless change significantly.

Example 11.11 (Random walk). In Example 11.7, the random walk

$$S_t = S_0 + X_1 + \dots + X_t$$

is a martingale when $m = \mathbb{E}(X_1) = 0$. The variance of this random walk equals

$$\text{Var}(S_t) = \text{Var}(S_0) + \sigma^2 t,$$

where $\sigma^2 = \text{Var}(X_1)$. When $\sigma^2 > 0$, it follows that the random variability of S_t grows to infinity as $t \rightarrow \infty$. This increasing variability is also visible in the simulated paths in Figure 11.1. ■

11.4 Long-term behavior of martingales

The long-term behavior of martingales is summarized by the following two important results. Their proofs [Wil91, 11.7,14.1] require deeper probability theoretic background, and are here omitted.

¹In these lecture notes a function is called *increasing*, when $s \leq t \implies f(s) \leq f(t)$ and *strictly increasing*, when $s < t \implies f(s) < f(t)$.

Theorem 11.12. *Every nonnegative martingale (M_t) converges according to*

$$\lim_{t \rightarrow \infty} M_t = M_\infty$$

with probability one, where the limit M_∞ is a finite nonnegative random number.

Martingales which may take on positive and negative values may not converge in the long run. The random walk in Example 11.11 cannot converge because its variance grows to infinity. Bounded martingales will nevertheless converge. A random number X is called *bounded* if there exists a constant c such that $\mathbb{P}(|X| \leq c) = 1$. A stochastic process (X_t) is called *bounded* if there exists a constant c such that $\mathbb{P}(|X_t| \leq c \text{ for all } t) = 1$.

Theorem 11.13. *Every bounded martingale (M_t) converges according to*

$$\lim_{t \rightarrow \infty} M_t = M_\infty$$

with probability one, where the limit M_∞ is a bounded random number.

The nature of convergence in the above theorems is stronger than the distributional convergence which have seen in the context of Markov chains. Here every path of the martingale converges to a limit (as in Figure 11.2), whereas the paths of irreducible Markov chains keep on visiting all states infinitely often and thus never converge. The pathwise convergence with probability one implies that $M_t \rightarrow M_\infty$ in distribution, but not vice versa in general.

Theorem 11.13 can also be generalized to martingales which are bounded in a weaker sense, namely to martingales which are uniformly integrable according to $\sup_t \mathbb{E}(|M_t| 1_{(|M_t| > K)}) \rightarrow 0$ as $K \rightarrow \infty$. It can also be shown that every uniformly integrable martingale can be represented as a prediction martingale of the limiting random variable M_∞ (recall Example 11.8).

11.4.1 Martingales and Markov chains

Let (X_0, X_1, \dots) be a discrete-time Markov chain with transition matrix P on a countable state space S . Let $f : S \rightarrow \mathbb{R}$ be some function, modeling our observation of the Markov chain. Then it is natural to ask when $M_t = f(X_t)$ is a martingale.

The above question is natural to set in a slightly general context where the evaluation function $f_t : S \rightarrow \mathbb{R}$ is allowed to also depend on the time parameter. Hence we will study a random process $M_t = f_t(X_t)$. When we condition on the event $X_t = x$, the expected value of $M_{t+1} = f_{t+1}(X_{t+1})$ is obtained from the formula

$$\mathbb{E}(M_{t+1} | X_t = x) = \sum_y \mathbb{P}(X_{t+1} = y | X_t = x) f_{t+1}(y) = \sum_y P(x, y) f_{t+1}(y).$$

When the map $x \mapsto \sum_y P(x, y)f_{t+1}(y)$ is interpreted as a matrix product of a square matrix P and column vector Pf_{t+1} , the above result can be written as

$$\mathbb{E}(M_{t+1} | X_t = x) = Pf_{t+1}(x).$$

By the Markov property, the above expectation does not depend on the past states of the chain, so that

$$\mathbb{E}(M_{t+1} | X_0, \dots, X_t) = \mathbb{E}(M_{t+1} | X_t) = Pf_{t+1}(X_t).$$

The computations lead us to the following theorem. Note that the equalities and inequalities in the theorem concerning column vectors are considered entrywise, so that for example $Pf_{t+1} = f_t$ means that

$$Pf_{t+1}(x) = f_t(x) \quad \text{for all } x \in S.$$

Theorem 11.14. *Assume that $\sum_y P(x, y)|f_t(y)| < \infty$ for all $x \in S$ and all t . Then the random process $M_t = f_t(X_t)$ with respect to (X_t) is a*

$$\begin{cases} \text{supermartingale,} & \text{if } Pf_{t+1} \leq f_t, \\ \text{martingale,} & \text{if } Pf_{t+1} = f_t, \\ \text{submartingale,} & \text{if } Pf_{t+1} \geq f_t. \end{cases}$$

Example 11.15 (Gambling with unit bets). A casino offers a game where every round produces one euro win ($X_t = +1$) with probability p , and one euro loss ($X_t = -1$) with probability $q = 1 - p$, independently of other rounds. The wealth of a gambler after t rounds equals

$$S_t = S_0 + X_1 + \dots + X_t,$$

where $S_0 = 100$ is the gambler's initial wealth. A discounted value of the gambler's wealth can be expressed as $M_t = r^{S_t}$ for some discount factor $r > 0$. Is (M_t) a martingale?

The wealth process (S_0, S_1, \dots) is a discrete-time Markov chain on \mathbb{Z} , with transition matrix P such that $P(x, x+1) = p$ and $P(x, x-1) = q$ for all $x \in \mathbb{Z}$. For the function $f(x) = r^x$ considered as a column vector, we have

$$Pf(x) = \sum_{y=-\infty}^{\infty} P(x, y)r^y = pr^{x+1} + qr^{x-1}.$$

When $r = q/p$, it hence follows that $Pf(x) = f(x)$ for all x , and Theorem 11.14 implies that $(q/p)^{S_t}$ is a martingale. This is called de Moivre's martingale². (Note: Also the centered random walk $S_t - (p - q)t$ is a martingale, by Example 11.7). ■

²Named after French mathematician Abraham de Moivre (1667–1754) who applied this process to solve a gambler's ruin problem in his classic book *The Doctrine of Chances* (1718).

Example 11.16 (Normalised branching process). Let (X_0, X_1, \dots) be a branching process with offspring distribution $p = (p(0), p(1), \dots)$ having mean $m = \sum_{x=0}^{\infty} xp(x) < \infty$. Is the normalised process $M_t = r^{-t}X_t$ a martingale for some constant $r > 0$?

The normalised process can be represented as $M_t = f_t(X_t)$, where $f_t(x) = r^{-t}x$. According to Theorem 6.2, the conditional expectation satisfies

$$\mathbb{E}_x(X_1) = \mathbb{E}(X_1 | X_0 = x) = mx,$$

so that

$$Pf_{t+1}(x) = \mathbb{E}_x(f_{t+1}(X_1)) = \mathbb{E}_x(r^{-t-1}X_1) = r^{-t-1}mx = (m/r)f_t(x).$$

By choosing $r = m$ we therefore have $Pf_{t+1}(x) = f_t(x)$ for all $x \in S$ and $t \geq 0$, so that by Theorem 11.14 the process $M_t = m^{-t}X_t$ is a martingale with respect to (X_0, X_1, \dots) . Because (M_t) is a nonnegative martingale, Theorem 11.12 implies that for some finite random number M_∞ ,

$$M_\infty = \lim_{t \rightarrow \infty} m^{-t}X_t.$$

Hence we may express the population size at generation t approximately as

$$X_t \approx M_\infty m^t.$$

This represents exponential growth with a random constant factor M_∞ . The event that $M_\infty = 0$ corresponds to the case where the population becomes extinct, whereas on the event $M_\infty > 0$ (which occurs with positive probability when $m > 1$) the population approaches infinity at an exponential rate. ■

Chapter 12

Stopped martingales and optional times

In the context of gambling, a *martingale* (*martingaali*) is a betting strategy where the bet is doubled after every losing round. In this section we learn to analyze various betting strategies using martingales and random optional times.

12.1 Gambling with unit bets

The cumulative net profit from t rounds can be written as

$$M_t = \sum_{s=1}^t X_s \quad (M_0 = 0),$$

where X_s is the profit from round s with mean $m = \mathbb{E}(X_s)$. When the profits per rounds are independent and identically distributed random integers, it follows that (M_t) is a random walk on \mathbb{Z} . According to Example 11.7, the process (M_t) is with respect to (M_0, X_1, X_2, \dots)

$$\begin{cases} \text{supermartingale,} & \text{when } m \leq 0, \\ \text{martingale,} & \text{when } m = 0, \\ \text{submartingale,} & \text{when } m \geq 0. \end{cases}$$

Example 12.1. In a typical casino the game of roulette with a unit bet on red produces

$$X_s = \begin{cases} +1 & \text{with probability } 18/37, \\ -1 & \text{with probability } 19/37, \end{cases} \quad m = -1/37,$$

and a unit bet on a selected number produces

$$X_s = \begin{cases} +31 & \text{with probability } 1/37, \\ -1 & \text{with probability } 36/37, \end{cases} \quad m = -5/37.$$

In both games the expected return per unit bet is negative, so the corresponding unit return processes are supermartingales. ■

According to Theorem 11.10, the expected net profit $t \mapsto M_t$ in an unfavorable game (supermartingale) is decreasing, and in a fair game (martingale) constant. Hence an unfavorable game does not produce profits with unit bets. The it is natural to ask whether a positive profit can be made by a suitable adaptive betting strategy.

12.2 Doubling strategy

In a doubling strategy the bet is doubled after every losing round. This is continued until either the player hits a selected target value or the player runs out of money. Consider a game where you win or lose one euro at every round. Table 12.1 describes the evolution of net profit for a player using the doubling strategy in a simulated scenario where the the first four rounds are losing rounds and the fifth round is a winning round. The initial bet is one euro.

t	1	2	3	4	5
Bet in round t	1	2	4	8	16
Outcome of round t	Loss	Loss	Loss	Loss	Win
Profit of round t	-1	-2	-4	-8	+16
Net profit from t rounds	-1	-3	-7	-15	+1

Table 12.1: Evolution of net profit in a doubling strategy.

In the above scenario, the net profit becomes +1 after the first winning round. This observation holds indeed in general. Namely, in a scenario with t losing rounds before a winning round, the cumulative losses from the first t rounds are $1 + 2 + \dots + 2^{t-1}$ euros, and the amount won round $t + 1$ equals 2^t euros. Hence the wealth of a player, starting with W_0 euros, after $t + 1$ rounds equals

$$W_{t+1} = W_0 - (1 + 2 + \dots + 2^{t-1}) + 2^t = W_0 - \frac{2^t - 1}{2 - 1} + 2^t = W_0 + 1.$$

Hence a player following the doubling strategy surely ends up with net profit of one euro after the first winning round. In this analysis no assumptions were made about probabilities of winning. The only essential requirements is that a winning round will eventually happen. According to Theorem 12.2 below, it is sufficient to assume that the outcomes of the rounds are independent, and the probability of winning is bounded away from zero.

Theorem 12.2. *Let X_0, X_1, \dots be independent $\{0, 1\}$ -valued random variables such that $\mathbb{P}(X_t = 1) \geq \epsilon > 0$ for all $t \geq 0$. Then*

$$\mathbb{P}(X_t = 1 \text{ for some } t \geq 0) = 1.$$

Proof. Let $T = \min\{t \geq 0 : X_t = 1\} \in [0, \infty]$ first hitting time of (X_t) into state 1. Then

$$\begin{aligned}\mathbb{P}(T > t) &= \mathbb{P}(X_0 = 0, \dots, X_t = 0) \\ &= \mathbb{P}(X_0 = 0) \cdots \mathbb{P}(X_t = 0) \\ &\leq (1 - \epsilon)^{t+1}.\end{aligned}$$

When $\epsilon > 0$, it follows by the monotone continuity of probability measures that

$$\mathbb{P}(T = \infty) = \mathbb{P}(\cap_{t=0}^{\infty} \{T > t\}) = \lim_{t \rightarrow \infty} \mathbb{P}(T > t) \leq \lim_{t \rightarrow \infty} (1 - \epsilon)^{t+1} = 0.$$

Hence

$$\mathbb{P}(X_t = 1 \text{ for some } t \geq 0) = \mathbb{P}(T < \infty) = 1.$$

□

If we denote the time index of the first winning round by a random variable T , then under the above assumptions, $T < \infty$ with probability one. Moreover, on this random instant we have $W_T = W_0 + 1$ with probability one. This is what happens regardless of how small the probability of winning is. Hence it appears the doubling strategy provides a sure way to make profit in an arbitrary game. Is this really the case? We will consider this in more detailed in what follows.

12.3 Adaptive betting

If we bet H_s euros on round s and the profit per unit bet is X_s euros, the the wealth of a player after round t equals

$$W_t = W_0 + \sum_{s=1}^t H_s X_s.$$

In analyzing general betting strategies we need to keep in mind that the when choosing the bet amount for round t , the player only knows the realizations of random variables W_0, X_1, \dots, X_{t-1} . Hence the player chooses the bet amount for round t as a deterministic function of $(W_0, X_1, \dots, X_{t-1})$, and it follows

$$H_t \in \sigma(W_0, X_1, \dots, X_{t-1}), \quad t \geq 1.$$

In this case the sequence (H_1, H_2, \dots) is called *previsible* (*ennakoitava*) with respect to the information sequence (W_0, X_1, X_2, \dots) . The *unit yield process* (*yksikkötuottoprosessi*) of the game is defined by

$$M_t = \sum_{s=1}^t X_s, \quad t = 0, 1, \dots$$

Because $X_s = M_s - M_{s-1}$, we can represent the wealth process corresponding to a general betting strategy (H_1, H_2, \dots) as

$$W_t = W_0 + (H \cdot M)_t,$$

where

$$(H \cdot M)_t = \sum_{s=1}^t H_s(M_s - M_{s-1}), \quad t = 0, 1, 2, \dots, \quad (12.1)$$

is the *integral process* (*integraaliprosessi*) of the sequence (H_1, H_2, \dots) against the unit yield process (M_0, M_1, \dots) . A stock market interpretation of the above formula is obtained by considering M_t as the price of a stock in the end of trading day t , and H_t as the amount of stock in the portfolio during day t .¹

Theorem 12.3. *Let (H_1, H_2, \dots) be a previsible sequence of integrable random numbers with respect to (X_0, X_1, \dots) such that $(H \cdot M)_t$ is integrable for all t .*

- (i) *If (M_t) is a martingale, then $(H \cdot M)_t$ is a martingale.*
- (ii) *If (M_t) is a submartingale and $H_t \geq 0$ for all t , then $(H \cdot M)_t$ is a submartingale.*
- (iii) *If (M_t) is a supermartingale and $H_t \geq 0$ for all t , then $(H \cdot M)_t$ is a supermartingale.*

Before proving Theorem 12.3 we note that for the integrability condition $(H \cdot M)_t$ it suffices to assume that the random numbers of either (M_0, M_1, \dots) or (H_1, H_2, \dots) are *bounded*². If for example $|H_s| \leq c_s$ for s , then by applying the triangle inequality and (12.1) we see that

$$\mathbb{E}|(H \cdot M)_t| \leq \sum_{s=1}^t c_s(\mathbb{E}|M_s| + \mathbb{E}|M_{s-1}|) < \infty,$$

because the random numbers of every martingale are integrable by definition. A corresponding reasoning also guarantees the integrability $\mathbb{E}|(H \cdot M)_t| < \infty$ in the case where the random numbers M_0, M_1, \dots are bounded and H_1, H_2, \dots integrable.

Proof of Theorem 12.3. (i) Denote the integral process by $W_t = (H \cdot M)_t$ and let us also use the shorthand notation $X_{s:t} = (X_s, \dots, X_t)$. By the integrability assumption $\mathbb{E}|W_t| < \infty$ for all $t \geq 0$. Because $H_{1:t}$ is determined by $X_{0:(t-1)}$ and $M_{0:t}$ is determined by $X_{0:t}$, we may conclude using (12.1) that

$$W_t \in \sigma(X_0, \dots, X_t). \quad (12.2)$$

¹In continuous-time investment models the integral process is defined as a stochastic Itô-integral $(H \cdot M)_t = \int_0^t H_s dM_s$. This setting is discussed for example on the course MS-E1601 Brownian motion and stochastic analysis.

²A random number Z is *bounded* (*rajoitettu*) if $\mathbb{P}(|Z| \leq c) = 1$ for some constant c .

Because (M_0, M_1, \dots) is a martingale,

$$\mathbb{E}(M_{t+1} | X_{0:t}) = M_t = \mathbb{E}(M_t | X_{0:t}),$$

so that the by the linearity of conditional expectations,

$$\mathbb{E}(M_{t+1} - M_t | X_{0:t}) = 0. \quad (12.3)$$

The definition of the integral process in turn implies that

$$W_{t+1} - W_t = H_{t+1}(M_{t+1} - M_t).$$

Previsibility implies that $H_{t+1} \in \sigma(X_{0:t})$, so by pulling out a known factor (Theorem 11.3) and applying (12.3),

$$\begin{aligned} \mathbb{E}(W_{t+1} - W_t | X_{0:t}) &= \mathbb{E}(H_{t+1}(M_{t+1} - M_t) | X_{0:t}) \\ &= H_{t+1} \mathbb{E}(M_{t+1} - M_t | X_{0:t}) \\ &= 0. \end{aligned} \quad (12.4)$$

By linearity of conditional expectations and (12.2), this implies that

$$\begin{aligned} \mathbb{E}(W_{t+1} | X_{0:t}) &= \mathbb{E}(W_t | X_{0:t}) + \mathbb{E}(W_{t+1} - W_t | X_{0:t}) \\ &= \mathbb{E}(W_t | X_{0:t}) \\ &= W_t. \end{aligned}$$

Hence (W_0, W_1, \dots) is a martingale with respect to (X_0, X_1, \dots) .

(ii) When (M_0, M_1, \dots) is a submartingale, we may verify in a similar way that $\mathbb{E}|W_t| < \infty$ and $W_t \in \sigma(X_0, \dots, X_t)$. Analogously we may also prove that $\mathbb{E}(W_{t+1} | X_{0:t}) \leq W_t$. In this case '=' gets replaced with ' \geq ' in (12.3) and in the last equality of (12.4). Note that in justifying

$$\mathbb{E}(M_{t+1} - M_t | X_{0:t}) \geq 0 \implies H_t \mathbb{E}(M_{t+1} - M_t | X_{0:t}) \geq 0$$

we need to apply the extra assumption $H_t \geq 0$.

(iii) When (M_0, M_1, \dots) is a supermartingale, the proof is analogous to the proof (ii). \square

Consider a game where the unit yield process (M_t) is a supermartingale with respect to the available information (X_t) . Assume also that the terms of the unit yield process are bounded random variables. The by Theorem 12.3 we see that the gambler's wealth process

$$W_t = W_0 + (H \cdot M)_t, \quad t \geq 0,$$

is a supermartingale for every previsible betting strategy with $H_t \geq 0$ and $\mathbb{E}(H_t) < \infty$. Therefore the gambler's wealth $t \mapsto W_t$ is nonincreasing by expectation (Theorem 11.10). We may hence conclude that in such unfavorable games there is no way to make profit using previsible betting strategies. In this context the condition $H_t \geq 0$ forbids the gambler to act as a casino.

Example 12.4 (Doubling strategy). Let us continue analyzing the doubling strategy in Example 12.2. Denote the unit yield of round t by $X_t \in \{-1, +1\}$, and denote the index of the first winning round by

$$T = \min\{t \geq 1 : X_t = +1\}.$$

Assume that X_1, X_2, \dots are mutually independent with $\mathbb{P}(X_t = +1) = p$ and $\mathbb{P}(X_t = -1) = q$, where $0 < p < q$. Then the unit yield process $M_t = \sum_{s=1}^t X_s$ is a supermartingale with respect to information sequence $(0, X_1, X_2, \dots)$ and saw earlier (Theorem 12.2) that T is finite with probability one.

The betting strategy (H_1, H_2, \dots) corresponding to the doubling strategy can be recursively expressed as $H_1 = 1$ and $H_{t+1} = 2H_t 1(t < T)$ for all $t \geq 0$. This implies that

$$H_t = 2^{t-1} 1(t \leq T), \quad t = 0, 1, \dots$$

Because T is random, also the numbers H_t are random variables. By writing the bet of round t as

$$H_t = \begin{cases} 2^{t-1}, & \text{if } X_s = -1 \text{ for all } s = 0, \dots, t-1, \\ 0, & \text{else,} \end{cases}$$

we see that $H_t \in \sigma(0, X_1, \dots, X_{t-1})$ for all $t \geq 1$, so that the betting process is previsible with respect to information sequence $(0, X_1, X_2, \dots)$. Further, because H_t is bounded by $0 \leq H_t \leq 2^{t-1}$, we may conclude using Theorem 12.3 that the gambler's wealth process

$$W_t = W_0 + (H \cdot M)_t, \quad t \geq 0,$$

is a supermartingale, and hence decreasing by expectation (Theorem 11.10). Hence expected net profit is negative by expectation at any deterministic time instant $t \geq 1$. ■

The result of Example 12.4 is in apparent conflict with the analysis in Section 12.2, where we saw that using doubling strategy can be used to make sure profit. The conflict can be explained by noting that the expected net profit for gambling with the doubling strategy is negative at every *deterministic* time instant t . The huge losses made in those games where the first winning round occurs after t cause the expected net profit at time t to be negative.

12.4 Optional times

In stock markets a natural investment strategy is to buy a stock and sell it at some random time instant T when certain conditions are met, for example when the stock price reaches a certain target level. A betting process (H_1, H_2, \dots) corresponding to this strategy is given by

$$H_t = \begin{cases} 1, & t \leq T, \\ 0, & \text{else.} \end{cases}$$

Assuming that the investor does have crystal ball helping to see the future, the decision whether or not to sell the stock at time t should be based on the observed values of available information X_0, \dots, X_t up to time t . Mathematically this requirement can be formulated using optional times. A random time instant $T \in \mathbb{Z}_+ \cup \{\infty\}$ is called an *optional time* (*valintahetki*) with respect to information sequence (X_0, X_1, \dots) if

$$1(T = t) \in \sigma(X_0, \dots, X_t) \quad \text{for all } t \geq 0.$$

This means that we can decide whether or not $T = t$ based on some deterministic function of (X_0, \dots, X_t) . The following result underlines the connection between optional times and previsible betting strategies.

Theorem 12.5. *A random process $H_t = 1(t \leq T)$ is previsible if and only if T is an optional time.*

Proof. Assume first that the sequence (H_1, H_2, \dots) defined by $H_t = 1(t \leq T)$ is previsible with respect to (X_0, X_1, \dots) . Then both H_t and H_{t+1} are determined by (X_0, \dots, X_t) , so that by the equation

$$1(T = t) = 1(t \leq T) - 1(t \leq T - 1) = H_t - H_{t+1}$$

it follows that also $1(T = t)$ is determined by (X_0, \dots, X_t) . Hence T is an optional time with respect to (X_0, X_1, \dots) .

Assume next that T is an optional time with respect to (X_0, X_1, \dots) . Then the event $\{T = s\}$ is determined by (X_0, \dots, X_s) for all $s \geq 0$, and the equality

$$H_t = 1 - 1(T \leq t - 1) = 1 - \sum_{s=0}^{t-1} 1(T = s)$$

shows that the value of H_t can be determined as a deterministic function of X_0, \dots, X_{t-1} . Hence (H_1, H_2, \dots) is previsible with respect to (X_0, X_1, \dots) . \square

Example 12.6 (Optional and nonoptional times). Optional times with respect to sequence (X_0, X_1, \dots) are for example

- the time when (X_t) first hits a set A ,
- the time when (X_t) hits A for the fourth time,

with the usual convention that the above random times are infinite if the event under consideration never occurs. The following random times are in general not optional:

- the time when (X_t) reaches its all-time maximum value,
- the time when (X_t) visits A for the last time.

■

12.5 Stopped martingales

A random process $(M_t)_{t \geq 0}$ stopped at time instant $T \in [0, \infty]$ is a random process $(M_{t \wedge T})_{t \geq 0}$, where we use the shorthand notation $t \wedge T = \min\{t, T\}$. Figure 12.1 displays a path of a stochastic process (blue) and corresponding stopped process (red).

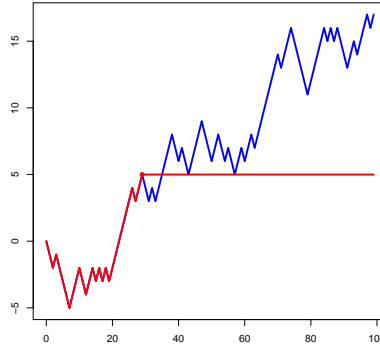


Figure 12.1: A simulated trajectory of a stochastic process stopped (red) at a random time instant $T = \min\{t \geq 0 : M_t = 5\}$, and the original stochastic process (blue).

Theorem 12.7. *Any (sub/super)martingale stopped at an optional time is a (sub/super)martingale.*

Proof. Let (M_t) be a submartingale and T an optional time with respect to (X_t) . The stopped process $M_{t \wedge T}$ can be written as

$$\begin{aligned} M_{t \wedge T} &= M_0 + \sum_{s=1}^{t \wedge T} (M_s - M_{s-1}) \\ &= M_0 + \sum_{s=1}^t H_s (M_s - M_{s-1}) = M_0 + (H \cdot M)_t, \end{aligned}$$

where $H_t = 1(t \leq T)$. According to Theorem 12.5 the sequence (H_1, H_2, \dots) is previsible. Moreover, because $0 \leq H_t \leq 1$ it follows that $(H \cdot M)_t$ is an integrable random number for all t . Theorem 12.3 then implies that $(H \cdot M)_t$ is a submartingale. Clearly, the constant process $t \mapsto M_0$ is also a submartingale. Then the linearity of conditional expectations implies that the sum of any submartingales is a submartingale. Hence the process $t \mapsto M_0 + M_{t \wedge T}$ is a submartingale.

Precisely the same argument works for proving the claims for martingales and supermartingales. \square

12.6 Optional stopping theorem

We will next analyze the value of a martingale M_T at a random time instant T . To be able to speak of this value, we must assume that T is finite with probability one. The following result is commonly known as (*Doob's optional stopping theorem*) after American mathematician Joseph Doob (1910–2004).

Theorem 12.8. *Let T be a finite optional time and let (M_0, M_1, \dots) be a martingale. Assume further that there exists an integrable random number Z such that*

$$|M_{t \wedge T}| \leq Z \quad \text{for all } t \geq 0. \quad (12.5)$$

Then $\mathbb{E}(M_T) = \mathbb{E}(M_0)$.

The technical integrability condition of Theorem (12.5) is valid for example in the case where the optional time T or the stopped process $t \mapsto |M_{t \wedge T}|$ is bounded from above with a deterministic constant.

Proof. When (M_t) is a martingale, then by Theorem 12.7 we know that $M_{t \wedge T}$ is a martingale, so that by Theorem 11.10 it follows that $t \mapsto M_{t \wedge T}$ is constant by expectation. Hence

$$\mathbb{E}(M_{t \wedge T}) = \mathbb{E}(M_{0 \wedge T}) = \mathbb{E}(M_0)$$

for all $t \geq 0$. On the other hand, the fact that T is finite with probability one guarantees that $\lim_{t \rightarrow \infty} M_{t \wedge T} = M_T$ with probability one. By the dominated continuity of expectations it follows that

$$\mathbb{E}(M_T) = \mathbb{E}\left(\lim_{t \rightarrow \infty} M_{t \wedge T}\right) = \lim_{t \rightarrow \infty} \mathbb{E}(M_{t \wedge T}) = \mathbb{E}(M_0).$$

□

The following example shows how Doob's optional stopping theorem can be applied to analyze hitting probabilities of random processes.

Example 12.9 (Random walk). Let (S_0, S_1, \dots) be a symmetric random walk on \mathbb{Z} , which moves one step left and one step right with equal probabilities $1/2$. Assume that the process is started at x such that $a < x < b$ for some integers a, b . What is the probability that the random walk hits b before a ?

According to Example 11.7 we know that (S_t) is a martingale. Let

$$T = \min\{t \geq 0 : S_t \in \{a, b\}\}$$

be the passage time of the process into $\{a, b\}$. Then T is an optional time with respect to (S_0, S_1, \dots) and it is known that T is finite with probability one. Because $a \leq S_{t \wedge T} \leq b$, it follows by Theorem 12.8 that

$$\mathbb{E}(S_T) = \mathbb{E}(S_0) = x.$$

On the other hand, at the random time instant T the random walk surely takes on value a or b , so that

$$\mathbb{E}(S_T) = a(1 - \mathbb{P}(S_T = b)) + b\mathbb{P}(S_T = b).$$

By combining these observations we conclude that

$$x = a(1 - \mathbb{P}(S_T = b)) + b\mathbb{P}(S_T = b),$$

from which we can solve

$$\mathbb{P}(S_T = b) = \frac{x - a}{b - a}.$$

We have derived this formula earlier using analysis of Markov chains. The analysis presented here is valid also for stochastic processes with more complicated dependence structures, as long as the martingale property holds. ■

Example 12.10 (Doubling strategy). Let us continue the analysis in Example 12.4. Here the time index of the first winning round

$$T = \min\{t \geq 1 : X_t = +1\}$$

is an optional time with respect to $(0, X_1, X_2, \dots)$. We also saw that the net profit process $W_t - W_0 = (H \cdot M)_t$ corresponding to the doubling strategy $H_t = 2^{t-1}1(t \leq T)$ is a supermartingale whenever the probability of winning is at most $1/2$, and hence decreasing in expectation. Nevertheless we have seen that $W_T - W_0 = 1$ with probability one. In this setting the statement of Doob's optional stopping theorem does not hold because $\mathbb{E}(W_T) = \mathbb{E}(W_0) + 1 > \mathbb{E}(W_0)$. The reason why the theorem is not applicable is that although T is finite, it is not bounded by any deterministic constant, and more seriously, the stopped process $|W_{t \wedge T}|$ is not bounded by any integrable random number.

To understand what this nonintegrability means in practice, let us investigate the expected net loss just before the winning round. The wealth just before winning equals

$$W_{T-1} = W_0 - \sum_{s=1}^{T-1} 2^{s-1} = W_0 + 1 - 2^{T-1},$$

and the probability of $\{T = t\}$ for $t \geq 1$ equals

$$\begin{aligned} \mathbb{P}(T = t) &= \mathbb{P}(X_1 = -1, \dots, X_{t-1} = -1, X_t = +1) \\ &= \mathbb{P}(X_1 = -1) \cdots \mathbb{P}(X_{t-1} = -1)\mathbb{P}(X_t = +1) \\ &= (1 - p)^{t-1}p. \end{aligned}$$

In an unfavorable ($0 < p \leq 1/2$) game we hence see that

$$\begin{aligned} \mathbb{E}(W_{T-1}) &= W_0 + 1 - \mathbb{E}(2^{T-1}) = W_0 + 1 - \sum_{t=1}^{\infty} 2^{t-1}(1-p)^{t-1}p \\ &= W_0 + 1 - p \sum_{t=0}^{\infty} (2(1-p))^t = -\infty. \end{aligned}$$

In an unfavorable game (supermartingale), a player following the doubling strategy is hence expected to make infinitely large loss before the winning round. ■

Appendix A

Suomi–English dictionary

suomi	englanti
alimartingaali	submartingale
alkio	element
alkujakauma	initial distribution
Bernoulli-jakauma	Bernoulli distribution
binomijakauma	binomial distribution
binomikerroin	binomial coefficient
diskreettiaikainen	discrete-time
diskreetti jakauma	discrete distribution
diskreetti satunnaismuuttuja	discrete random variable
ehdollinen jakauma	conditional distribution
ehdollinen odotusarvo	conditional expectation
ehdollinen tiheysfunktio	conditional density function
ehdollinen todennäköisyys	conditional probability
eksponenttijakauma	exponential distribution
elinaika	lifetime
ergodinen	ergodic
ergodisuus	ergodicity
esiintyvyys	occupancy, frequency
esiintyvyydsmatriisi	occupancy matrix
haarautumisprosessi	branching process
harha	bias
harhaton	unbiased
harvennettu	thinned
hetkittäinen jakauma	transient/time-dependent distribution
hyppytodennäköisyys	jump probability
hyppyvauhti	jump rate
indikaattori	indicator
indikaattorifunktio	indicator function
jakauma	distribution
jakso	period
jaksollinen	periodic
jaksoton	aperiodic
jatkuva-aikainen	continuous-time
jatkuva jakauma	continuous distribution
joukko	set, space
järjestetty lista	ordered list
järjestystunnusluku	order statistic
järjestämätön joukko	unordered set
kertoma	factorial
kertymäfunktio	cumulative distribution function
keskeinen raja-arvolause	central limit theorem
keskiarvo	average, mean
keskihajonta	standard deviation
keskineliövirhe	mean squared error
keskivirtaus	mean drift

kokonaisvaihteluetaisyys	total variation distance
kokovinoutettu	size-biased
komplementti	complement
konvoluutio	convolution
korrelaatio	correlation
korreloimaton	uncorrelated
korreloitu	correlated
kovarianssi	covariance
kulkuaika	passage time, first passage time, transition time
kustannuskertymä	cumulative cost
kustannusvauhti	cost rate
kvartiili	quartile
kääntyvä	reversible
laskurimitta	counting measure
laskuri-prosessi	counting process
leikkaus	intersection
lineaarinen riippuvuus	linear dependence
lisääntymisjakauma	offspring distribution
Markov-ketju	Markov chain
Markov-ominaisuus	Markov property
Markov-prosessi	Markov process
martingaali	martingale
mediaani	median
mitallinen	measurable
momentti	moment
moniulotteinen	multidimensional, multivariate
muuttuja	variable
normaaliapproksimaatio	normal approximation
normaalijakauma	normal distribution, Gaussian distribution
normitettu	normalized
normitettu normaalijakauma	standard normal distribution
odotettu	expected
odotusarvo	expectation, mean
osajoukko	subset
ositus	partition
osumatodennäköisyys	hitting probability
palautuva	recurrent
polku (satunnaisprosessin)	path (of a random process)
perusjoukko	sample space
raja-jakauma	limiting distribution
reuna-jakauma	marginal distribution
reunatiheysfunktio	marginal density function
riippumattomuus	independence
riippumattomasti sironnut	independently scattered
riippuvuus	dependence
satunnainen	random
satunnainen pistekuvio	random point pattern
satunnaisilmiö	random phenomenon
satunnaisjono	random sequence
satunnaiskenttä	random field
satunnaiskulku	random walk
satunnaisluku	random number
satunnaismatriisi	random matrix
satunnaismuuttuja	random variable
satunnaismuuttujan muunnos	transformation of a random variable
satunnaisvektori	random vector
satunnaisverkko	random graph
siirtymäkaavio	transition diagram
siirtymämatriisi	transition matrix
stokastiikka	stochastics
stokastinen	stochastic
stokastinen esitys	stochastic representation
stokastinen prosessi	stochastic process
stokastinen riippuvuus	stochastic dependence
suhteellinen esiintyvyys	relative frequency, relative occupancy
suhteellinen osuus	relative proportion
sukupuutto	extinction
supeta jakaumalta	converge in distribution

supeta stokastisesti	converge in probability
suurten lukujen laki	law of large numbers
syntymiskuolemisketju	birth–death chain
tapahtuma	event
tasaintegroituva	uniformly integrable
tasajakauma	uniform distribution
tasakoosteinen	homogeneous
tasapainojakauma	invariant/equilibrium/stationary distribution
tasapainoyhtälö	balance equation
tiheysfunktio (diskreetin jakauman)	density function, probability mass function
tiheysfunktio (jatkuvan jakauman)	density function, probability density function
tila (prosessin)	state (of a process)
tilajoukko (prosessin)	state space (of a process)
todennäköinen	probable, likely
todennäköisyys	probability
todennäköisyydet generoiva funktio	probability generating function
todennäköisyysjakauma	probability distribution
todennäköisyysmitta	probability measure
todennäköisyysteoria	probability theory
toteuma	realization, outcome
tulojoukko	product set, product space
uusiutumisosessi	renewal process
valinnaisen pysäyttämisen lause	optional stopping theorem
valintahetki	optional time, stopping time
varianssi	variance
vauhti	rate
väistyvä	transient
väliaika (uusiutumisosessin)	interevent time
yhdiste	union
yhteisjakauma	joint distribution
yhtenäinen ketju	irreducible chain
yhteysluokka	communicating class
yksiulotteinen	one-dimensional, univariate
ylimartingaali	supermartingale

Bibliography

- [Asm03] Søren Asmussen. *Applied Probability and Queues*. Springer, second edition, 2003.
- [BP98] Sergey Brin and Larry Page. The anatomy of a large-scale hyper-textual web search engine. In *7th International World-Wide Web Conference (WWW 1998)*, 1998.
- [Dur12] Richard Durrett. *Essentials of Stochastic Processes*. Springer, second edition, 2012.
- [Kal02] Olav Kallenberg. *Foundations of Modern Probability*. Springer, second edition, 2002.
- [Kul16] Vidyadhar G. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman and Hall/CRC, third edition, 2016.
- [LPW08] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, <http://pages.uoregon.edu/dlevin/MARKOV/>, 2008.
- [SW08] Rolf Schneider and Wolfgang Weil. *Stochastic and Integral Geometry*. Springer, Berlin, 2008.
- [Wil91] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.