Lectures on
Stochastic processes

Lasse Leskelä* and Eveliina Peltola†

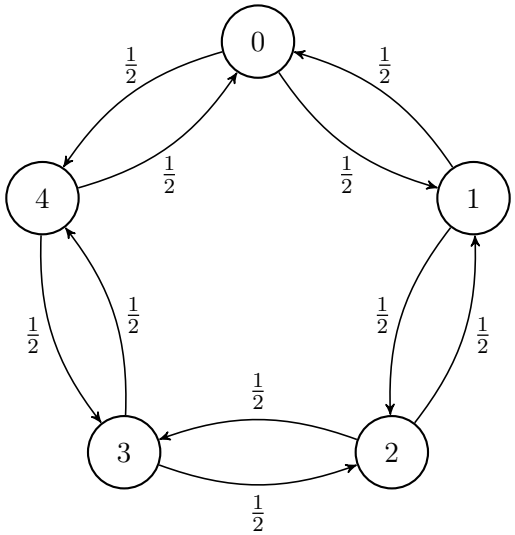
With contributions by Osama Abuzaid‡

November 12, 2025

Abstract

This course will get you introduced to stochastic processes, the theory of time-dependent random phenomena. You learn to mathematically model and analyze, for instance, particle and population flows using Markov processes and unpredictable time instants using Poisson processes. You are assumed to be familiar with basic notions of matrix algebra (e.g. [Str06]) and calculus (e.g. [AE21]), and basic concepts of probability, which are necessary for the treatment of stochastic processes, for instance as in the freely downloadable books [GS97, Dur12]. Other useful lecture notes related to the contents of the present course are, e.g., [LPW08, Les20] — see also the textbooks [Wil91, Kal21] some more advanced topics.

Most parts of these notes are directly based on the older notes [Les20] by the first author (LL). Any mistakes introduced in this new version are due to the second author (EP).



*Aalto University, Department of Mathematics and Systems Analysis. lasse.leskela@aalto.fi
†Aalto University, Department of Mathematics and Systems Analysis. eveliina.peltola@aalto.fi, University of Bonn, Institute for Applied Mathematics. eveliina.peltola@hcm.uni-bonn.de
‡Aalto University, Department of Mathematics and Systems Analysis. osama.abuzaid@aalto.fi

What is a stochastic process?

Roughly speaking, the word "stochastic" means "random" (unknown, modelled by a mathematical model), and the word "process" means something "evolving in time" (for example movement of particles, stock prices, evolution of some population, spread of a disease, data in the internet, a machine learning algorithm, etc.). The majority of this course will focus on a few prototypical mathematical models for random phenomena — which are, admittedly, oversimplifications, but still have proven to be extremely useful in modelling a plethora of phenomena.

The first of these is the simple model of a *Markov chain*; starting from Chapter 1, we will discuss basic theory, examples, and variants of Markov chains. Another very useful model especially for waiting times and queueing theory is the *Poisson process* (Chapters 8–9). *Branching processes* discussed in Chapter 7 are widely used, for example, as population models. In the end of the course, we will also learn basics of Markov chains in continuous time (Chapters 10–11) and some algorithmic sampling, "Monte Carlo," methods (Chapter 12).

Contents

1 Markov chains and stochastic models

A finite-state Markov chain is a stochastic (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 $S = \{x_1, x_2, \ldots, x_n\}$, and for now it is assumed to be a finite set with cardinality (size) $|S| = n \in \mathbb{N} = \{1, 2, \ldots\}$. The collection of transition probabilities[1]

$$P \;=\; \{P(x,y) : x, y \in S\}$$

is called the *transition matrix* (*siirtymämatriisi*). The transition matrix is a square matrix of size $|S| \times |S|$, with rows and columns indexed by the possible states $x, y \in S$. Being probabilities, the entries of the transition matrix must satisfy

$$0 \;\le\; P(x,y) \;\le\; 1, \qquad \text{for all } x, y \in S,$$

and because the Markov chain certainly moves to *some* state, each row-sum is equal to 1 by the law of total probability: for the x:th row, we have

$$\sum_{y \in S} P(x,y) \;=\; 1, \qquad \text{for all } x \in S. \tag{1.1}$$

Note that we can also allow the Markov chain to move back to the same state (or, "stay put") with probability $P(x,x)$. This means that the system does not change at that time step.

Reminder. In general, a *probability distribution* (*todennäköisyysjakauma*) on S is a function $\mu : S \to [0,1]$ such that the *law of total probability* holds:

$$\sum_{y \in S} \mu(y) \;=\; 1.$$

In the context of Markov chains, we will usually interpret probability distributions as row-vectors $\mu = (\mu(y) : y \in S)$ indexed by the states $y \in S$. For example, $(P(x,y) : y \in S)$ for fixed x. ♣

1.1 Markov property

Definition. An S-valued stochastic process (random sequence) $X = (X_0, X_1, X_2, \ldots)$ is a (time-homogeneous) *Markov chain* (*Markov-ketju*) with state space S and transition matrix P if X is "conditionally independent of the past," i.e.,

$$\mathbb{P}(X_{t+1} = y \mid X_t = x,\, H_{t-}) \;=\; P(x,y), \tag{1.2}$$

for all states $x, y \in S$, all times $t \ge 0$, and for all events $H_{t-} = \{X_0 = x_0, \ldots, X_{t-1} = x_{t-1}\}$ such that $\mathbb{P}(X_t = x,\, H_{t-}) > 0$.

In words, 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.

Reminder. For two events A and B such that $\mathbb{P}(B) > 0$, the *conditional probability* (*ehdollinen todennäköisyys*) of A given the event B is

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

where the symbol "$\cap$" means "and," i.e., $\mathbb{P}(A \cap B) = \mathbb{P}(A,\, B) = \mathbb{P}(A \text{ and } B)$. (Similarly, the symbol "$\cup$" means "or," i.e., $\mathbb{P}(A \cup B) = \mathbb{P}(A \text{ or } B)$.) ♣

[1] As is conventional and for better readability, we will use the notation $P(x,y)$ instead of the more usual notation $P_{x,y}$ for the matrix element of a matrix P at its x:th row and y:th column.

Equation (1.2) is called *Markov property* (*Markov-ominaisuus*) (after the Russian mathematician Andrey Markov (1856-1922)). The Markov property can be defined analogously also for stochastic processes with continuous time parameter and infinite state spaces (Chapter 5). The class of general Markov processes includes several important stochastic models such as Poisson processes, Random walks, and Brownian motions, some of which will be discussed later.

The following fundamental result tells that the past history H_{t-} may be ignored in (1.2). The idea is that the right-hand side of (1.2) only depends on x, y, and the transition matrix P.

Theorem 1.1 (*Markov property*). *For any (finite-state, time-homogeneous) Markov chain $X = (X_0, X_1, X_2, \ldots)$ with transition matrix P, we have*

$$\mathbb{P}(X_{t+1} = y \mid X_t = x) = P(x, y), \tag{1.3}$$

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

(The proof can be skipped at the first reading. It uses basic probability tools.)

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

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

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

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

and the Markov property (1.2) can be rephrased as

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

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

$$\sum_{x_0, \ldots, x_{t-1} \in S} f_{t+1}(x_0, \ldots, x_{t-1}, x, y) = P(x, y) \sum_{x_0, \ldots, x_{t-1} \in S} f_t(x_0, \ldots, x_{t-1}, x). \tag{1.4}$$

By the law of total probability, the left side of (1.4) equals $\mathbb{P}(X_t = x, X_{t+1} = y)$ and the right side equals $P(x, y) \cdot \mathbb{P}(X_t = x)$. Hence, we see that

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

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

1.2 Transition matrix and transition diagram

The structure of a Markov chain is usually best illustrated visually by a transition diagram. The *transition diagram* (*siirtymäkaavio*) of a transition matrix P and the corresponding Markov chain is a directed graph with node set being the state space S and link set comprising the 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. More examples will be discussed in the exercise sessions and later chapters.

Example 1.2 (*Weather model*)**.** The summer weather of day $t \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ in Espoo can be modeled using a stochastic process in state space $S = \{1, 2\}$, where

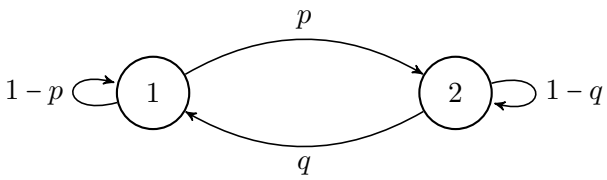
$$\text{state '1' = 'cloudy'} \qquad \text{and} \qquad \text{state '2' = 'sunny'}.$$

(Note that this model is not, of course, very realistic. However, for learning mathematical concepts it is much better to first look at very simple "toy" examples.)

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 states of the weather model can be represented as a Markov chain $X = (X_0, X_1, \ldots)$ 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 diagram



Let us assume that Monday (day $t = 0$) is cloudy. Then the weather model predicts Tuesday to be cloudy with probability $P(1, 1) = 1 - p$ and sunny with probability $P(1, 2) = p$, so that

$$\mathbb{P}(X_1 = 1 \mid X_0 = 1) = 1 - p \qquad \text{and} \qquad \mathbb{P}(X_1 = 2 \mid X_0 = 1) = p.$$

The probability that it is cloudy also on Wednesday is obtained by conditioning on the possible states of Tuesday's weather, in the following manner. (This is a very useful trick!)

Reminder. Using the law of total probability, for today's weather X_t at day $t \geq 1$, we have

$$\begin{aligned} \mathbb{P}(X_t = 1) &= \sum_{x \in \{1,2\}} \mathbb{P}(X_t = 1,\, X_{t-1} = x) \\ &= \sum_{x \in \{1,2\}} \mathbb{P}(X_t = 1 \mid X_{t-1} = x) \cdot \mathbb{P}(X_{t-1} = x) \\ &= \mathbb{P}(X_t = 1 \mid X_{t-1} = 1) \cdot \mathbb{P}(X_{t-1} = 1) + \mathbb{P}(X_t = 1 \mid X_{t-1} = 2) \cdot \mathbb{P}(X_{t-1} = 2). \end{aligned}$$

♣

Applying this to $t = 2$ with the initial condition $X_0 = 1$ gives

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

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

Does this seem like a practical model? Well, it is mathematically simple, and one can compute predictions via matrix multiplication. For long-time predictions it would be better to use a computer, as we will see soon. However, one can also use the powerful theory of Markov chains to simplify the computations significantly. The key is the Markov property (1.2) (conditional independence property from the past). See Theorem 1.5.

The following, more complicated example is typical in applications related to industrial engineering and management. More examples of similar kind are available 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 counts 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 Poi(λ) distributed with mean $\lambda = 3.5$. Customers finding an empty stock at an instant of purchase go to buy their laptops elsewhere. We develop a Markov chain to model the state of the inventory[2].

Let X_t be a random variable describing the number of laptops in stock on Monday 10:00 during week $t \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$. Denote by D_t a random variable modeling the demand of laptops during the corresponding week. Because D_t is Poi(λ)-distributed, we know that

$$\mathbb{P}(D_t = k) = \begin{cases} e^{-\lambda}\frac{\lambda^k}{k!}, & k \geq 0, \\ 0, & k < 0. \end{cases} \tag{1.5}$$

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, we have

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

Hence, the state space of the random process $X = (X_0, X_1, \ldots)$ 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 is a Markov chain.

Let us next determine the transition probabilities $P(i, j)$. Consider first the case where $i = 2$ 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 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 using the Poisson distribution 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 $H_{t-} = \{X_0 = x_0, \ldots, X_{t-1} = x_{t-1}\}$. 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

$$\begin{aligned} P(i,j) &= \mathbb{P}(X_{t+1} = j \mid X_t = i, X_{t-1}, \ldots, X_0) \\ &= \mathbb{P}(X_t - D_t = j \mid X_t = i, X_{t-1}, \ldots, X_0) \\ &= \mathbb{P}(i - D_t = j \mid X_t = i, X_{t-1}, \ldots, X_0) \\ &= \mathbb{P}(D_t = i - j), \qquad \text{for all } i \in \{2,3,4,5\} \text{ and } j \in \{2,3,4\}. \end{aligned}$$

Using the Poisson distribution, 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, we have

$$\begin{aligned} P(i,5) &= \mathbb{P}(X_{t+1} = 5 \mid X_t = i, X_{t-1}, \ldots, X_0) \\ &= \mathbb{P}(X_t - D_t \leq 1 \mid X_t = i, X_{t-1}, \ldots, X_0) \\ &= \mathbb{P}(i - D_t \leq 1 \mid X_t = i, X_{t-1}, \ldots, X_0) \\ &= \mathbb{P}(D_t \geq i - 1), \qquad \text{for all } i \in \{2,3,4\}. \end{aligned}$$

[2]You may skip this example in this section for the first reading. It will be elaborated further in Chapter 3.

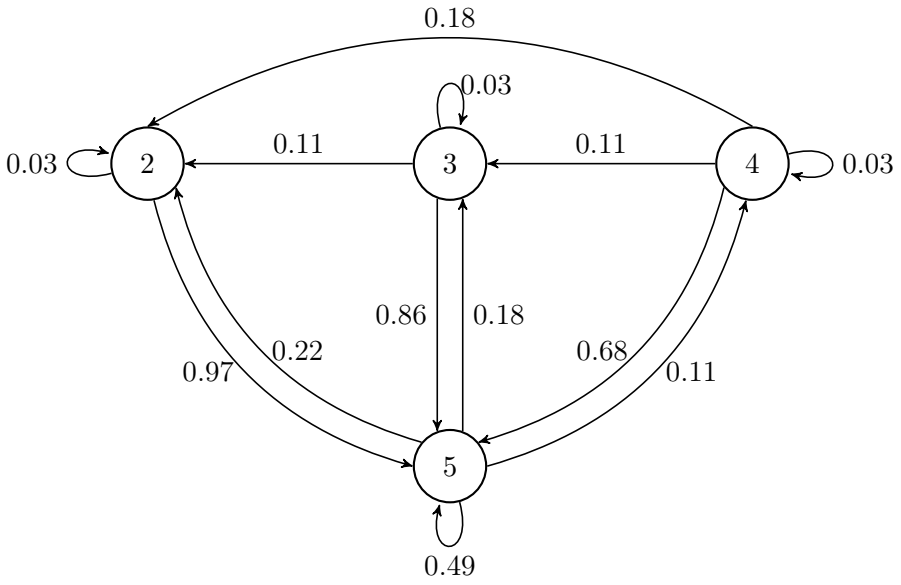
Finally, we need the value $P(5,5)$. A transition from state $i = 5$ to state $j = 5$ occurs in two cases: either there is no demand during week t, or the demand is 4 or more. Therefore,

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

By computing the probabilities of D_t from the Poisson distribution (1.5), we may write the transition 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



■

Markov chains encountered in applications in science and technology can have huge state spaces. The state space of the following example contains billions of nodes and grows all the time. (In Section 5 we will consider models for Markov chains with infinite state spaces.)

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 PageRank, which is defined as follows[3].

Consider a directed graph where the nodes consist of all web pages in the world, and links correspond to hyperlinks between the pages. Denote the set of nodes by S (state space), and define the *adjacency matrix* of the graph as a square matrix A with entries

$$A(x,y) = \begin{cases} 1, & \text{if there is a link from } x \text{ to } y, \\ 0, & \text{otherwise.} \end{cases} \tag{1.6}$$

[3]This example will be elaborated further in the exercise sessions.

Then, define a transition matrix on state space S by the formula[4]

$$P_c(x,y) \;=\; c\frac{1}{n} + (1-c)\,\frac{A(x,y)}{\sum_{z\in S} A(x,z)},$$

where $n = |S|$ 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 ($t \to \infty$). Whether or not this definition makes sense is not at all trivial. In Section 2, 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. ■

1.3 Time-dependent distributions

The *(time-dependent) distribution* (*(hetkittäinen) (tila)jakauma*) of a Markov chain describes its behavior in a finite time-horizon. It can be handily computed using Theorem 1.5.

Definition. The *distribution* (*jakauma*) of Markov chain $X = (X_0, X_1, X_2, \ldots)$ at time instant $t \geq 0$ is the probability distribution of the random variable X_t and is denoted by

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

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

The distribution μ_t at time t is a row-vector with elements indexed by the possible states $x \in S$. Therefore, we can multiply it from the right by the transition matrix P as in Equation (1.7).

▷ Note that the initial state X_0 where the Markov chain X starts can be random, or deterministic (known). The *initial distribution* gives the probabilities for X_0 to be in a given initial state $x \in S$, that is, the probabilities

$$\mu_0(x) \;=\; \mathbb{P}(X_0 = x), \qquad x \in S.$$

For example, μ_0 could be determined from some data collected about the phenomenon one wants to model, and the time-evolution would then be determined by only the initial distribution μ_0 and the transition matrix P, according to Theorem 1.5 below.

▷ In Example 1.2, the initial distribution corresponding to the deterministic initial state $X_0 = 1$ (today is cloudy) equals the *Dirac distribution* (*Dirac-jakauma*) at state '1':

$$\mu_0(1) \;=\; \mathbb{P}(X_0 = 1) = 1 \qquad \text{and} \qquad \mu_0(2) \;=\; \mathbb{P}(X_0 = 2) = 0.$$

This can be written as a row-vector $\mu_0 = [1, 0]$.

▷ Note that the law of total probability holds at all times:

$$\sum_{x\in S} \mu_t(x) \;=\; 1, \qquad t \geq 0.$$

[4]The formula is valid for graphs where the outdegree $\deg(x) = \sum_z A(x,z)$ 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.

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

$$\mathbb{P}(X_t = y) \;=\; \sum_{x \in S} \mathbb{P}(X_{t-1} = x) \cdot \mathbb{P}(X_t = y \mid X_{t-1} = x).$$

By applying (1.3) from Theorem 1.1, the above equation can be written as

$$\mu_t(y) \;=\; \sum_{x \in S} \mu_{t-1}(x) \cdot P(x, y), \qquad t \geq 1. \tag{1.7}$$

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

$$\mu_t \;=\; \mu_{t-1} \cdot P. \tag{1.8}$$

This observation leads to the following important result.

Theorem 1.5 (*Time-dependent distribution*)**.** *The distribution μ_t of any (time-homogeneous) Markov chain $X = (X_0, X_1, X_2, \ldots)$ at an arbitrary time instant $t \geq 0$ can be computed from the initial distribution μ_0 using the formula*

$$\mu_t \;=\; \mu_0 \cdot P^t, \tag{1.9}$$

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

Proof. We prove the claim (1.9) by mathematical induction. 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 (1.8) and the associativity of matrix multiplication, it follows that

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

and hence, the claim also holds for time instant $t+1$. Thus, according to the induction principle, the claim (1.9) holds for all $t \geq 0$. □

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 deterministic 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 (1.9), the weather distribution of Wednesday can be computed using the formula

$$\mu_2 = \mu_0 \cdot P^2,$$

so that

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

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

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

We will study long-term behavior of a similar model as $t \to \infty$ in Section 3 (Example 3.4). ■

1.4 Many-step transition probabilities

The entry $P(x, y)$ of the transition matrix gives the probability of moving from state x to state y during one time step. The following result shows that $P^t(x, y)$ gives the probability of moving from state x to state y during exactly t time steps.

Theorem 1.7. *The probability that (time-homogeneous) Markov chain $X = (X_0, X_1, \ldots)$ moves from state x to state y during t time steps can be computed using the formula*

$$\mathbb{P}(X_t = y \mid X_0 = x) = P^t(x, y), \qquad \text{for all } x, y \in S, \tag{1.10}$$

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. We prove the claim (1.10) by mathematical induction. The claim is true at time instant $t = 0$ because the identity matrix $P^0 = I$ satisfies

$$P^0(x, y) = I(x, y) = \delta_x(y) = \begin{cases} 1, & \text{if } x = y, \\ 0, & \text{if } x \neq y. \end{cases}$$

If the claim is true for some time instant $t \geq 0$, then by conditioning on all the possible states of X_t and applying the induction hypothesis $\mathbb{P}(X_t = z \mid X_0 = x) = P^t(x, z)$ and the Markov property (1.2) we find that

$$\begin{aligned} \mathbb{P}(X_{t+1} = y \mid X_0 = x) &= \sum_{z \in S} \mathbb{P}(X_{t+1} = y \mid X_t = z, X_0 = x) \cdot \mathbb{P}(X_t = z \mid X_0 = x) \\ &= \sum_{z \in S} P(z, y) \cdot P^t(x, z) \\ &= \sum_{z \in S} P^t(x, z) \cdot P(z, y) = P^{t+1}(x, y). \end{aligned}$$

Hence, the claim also holds for time instant $t + 1$. Thus, according to the induction principle, the claim (1.10) holds for all $t \geq 0$. □

Example 1.8 (*Holiday weather*)**.** 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 (1.10) (and a computer) from Theorem 1.7 that

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

The expected net cost of the holiday using the travel insurance is hence

$$300 \text{ EUR} + (1 - 0.571) \cdot 1900 \text{ EUR} = 1151 \text{ EUR},$$

so that travel insurance appears to be a good investment. ■

1.5 Path probabilities

The initial distribution μ_0 and the transition matrix P of a Markov chain X determine the probabilities all possible finite trajectories (that is, paths in S that the Markov chain can take). The following result tells how these can be computed for each given trajectory.

Theorem 1.9. *For any (time-homogeneous) Markov chain $X = (X_0, X_1, X_2, \ldots)$ with transition matrix P and for any time instant $t \geq 1$, we have*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = \mu_0(x_0) \cdot P(x_0, x_1) \cdots P(x_{t-1}, x_t), \tag{1.11}$$

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

(The proof is similar to the earlier ones, and can be skipped at the first reading.)

Proof. We prove the claim (1.11) by mathematical induction. The claim is true at time instant $t = 1$ by definition of conditional probability:

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

To proceed by induction, assume that (1.11) is true for some $t \geq 1$, and denote the event that the trajectory of the Markov chain up to time t equals a particular list of states $(x_0, \ldots, x_t)$ in S by $A_t = \{X_0 = x_0, \ldots, X_t = x_t\}$. Then, by noting that[5] $A_{t+1} = A_t \cap \{X_{t+1} = x_{t+1}\}$, we find that

$$\mathbb{P}(A_{t+1}) = \mathbb{P}(A_{t+1} \mid A_t) \cdot \mathbb{P}(A_t) = \mathbb{P}(X_{t+1} = x_{t+1} \mid A_t) \cdot \mathbb{P}(A_t).$$

Furthermore, the Markov property (1.2) implies that

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

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

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

which shows that the claim (1.11) also holds for time instant $t + 1$. Thus, according to the induction principle, the claim (1.11) holds for all $t \geq 1$. □

1.6 Occupancy of states

It is often interesting to estimate *how often* a given state occurs for the Markov chain in the long run. The expected value of these occurrences can be measured in terms of their *occupancy*, or *frequency*. These are collected in the *occupancy matrix* of the Markov chain.

To analyze frequencies of states, we employ the following common notation.

▷ The *indicator* (*indikaattori*) random variable of an event A is a binary random variable $\mathbb{1}(A)$ such that

$$\mathbb{1}(A) = \begin{cases} 1, & \text{if event } A \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases} \tag{1.12}$$

[5] Here, we use the common notation for intersections of events, so $A_{t+1} = \{X_0 = x_0, \ldots, X_t = x_t, X_{t+1} = x_{t+1}\} = \{X_0 = x_0\} \cap \cdots \cap \{X_t = x_t\} \cap \{X_{t+1} = x_{t+1}\} = A_t \cap \{X_{t+1} = x_{t+1}\}$. (So the symbol "∩" means "and.")

Reminder. The expectation of the indicator random variable of an arbitrary event A equals

$$\begin{aligned}\mathbb{E}(\mathbb{1}(A)) &= 0 \cdot \mathbb{P}(\mathbb{1}(A) = 0) + 1 \cdot \mathbb{P}(\mathbb{1}(A) = 1) \\ &= \mathbb{P}(\mathbb{1}(A) = 1) \\ &= \mathbb{P}(A). \end{aligned} \tag{1.13}$$

♣

▷ The number of times that a given state y occurs in a trajectory $(X_0, \ldots, X_{t-1})$ realized by the Markov chain at its first t steps is a random integer

$$N_t(y) = \sum_{s=0}^{t-1} \mathbb{1}(X_s = y). \tag{1.14}$$

We also call $N_t(y)$ the *number of visits* (*vierailujen lukumäärä*) to state y by Markov chain X during its first t time steps. From (1.12) we see that this is just counting with weight one the times $s \geq 0$ when $X_s = y$, and with weight zero the times $s \geq 0$ when X_s is in some other state than y.

▷ The *(relative) frequency* (*(suhteellinen) esiintyvyys*) of the state y is

$$\frac{N_t(y)}{t} = \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{1}(X_s = y). \tag{1.15}$$

▷ The *occupancy time* (*odotusarvoinen esiintyvyys*) of state y for initial state x is

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

In the literature, $G_t(x, y)$ is also called *Green's function*, due to its relation to potential theory [LPW08].

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

Theorem 1.10 (*Occupancy matrix*)**.** *The occupancy matrix of (time-homogeneous) Markov chain $X = (X_0, X_1, \ldots)$ can be computed from powers of the transition matrix P using the formula*

$$G_t = \sum_{s=0}^{t-1} P^s. \tag{1.16}$$

(The proof can be skipped at the first reading. It uses basic probability tools.)

Proof. Using (1.13) and linearity of the expectation, the number of visits (1.14) to y by the Markov chain X has expected value

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

With initial condition $X_0 = x$, because $\mathbb{P}(X_s = y \mid X_0 = x) = P^s(x, y)$ due to (1.10) in Theorem 1.7, we see that

$$G_t(x, y) = \mathbb{E}(N_t(y) \mid X_0 = x) = \sum_{s=0}^{t-1} \mathbb{P}(X_s = y \mid X_0 = x) = \sum_{s=0}^{t-1} P^s(x, y),$$

which is an entry-by-entry representation of the matrix equation (1.16). □

Example 1.11 (*Weather model*)**.** Let us 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 $G_7(2,1)$ of the occupancy matrix G_7 at time $t = 7$. Applying (1.16) from Theorem 1.10 (and a computer), we find that

$$G_7 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix} + \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}^2 + \cdots + \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}^6 = \begin{bmatrix} 5.408 & 1.592 \\ 3.980 & 3.020 \end{bmatrix}.$$

According to the prediction, the expected number of cloudy days is hence $G_7(2,1) = 3.980$. ■

2 Markov chains in the long run

In the previous lecture, we learned to compute the time-dependent distributions μ_t of a Markov chain with given initial distribution μ_0 using the formula $\mu_t = \mu_0 \cdot P^t$ (Theorem 1.5). When looking at a very long time-horizon, it is natural to ask the following questions:

1. Do the time-dependent distributions μ_t have a limit as $t \to \infty$ (in a sense detailed below)?
2. Does such a limit, if exists, depend on the initial distribution μ_0?
3. How can a limit be computed in practise?

We will answer these questions for Markov chains on finite state spaces in this chapter, and return to them in the case of countably infinite state spaces in Chapter 5.

2.1 Invariant (i.e., stationary) and limiting distributions

Consider a Markov chain $X = (X_0, X_1, X_2, \ldots)$ on finite state space S with time-dependent distribution

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

Definition. Starting from a given initial distribution μ_0, if the limit

$$\lim_{t\to\infty} \mu_t(x) = \mu_\infty(x), \qquad \text{for all } x \in S,$$

exists, then we say that μ_∞ is the *limiting distribution* (*rajajakauma*) of the Markov chain started at the initial distribution μ_0. (Note that μ_∞ is always a probability distribution for finite state spaces S, but not necessarily for infinite state spaces (Chapter 5).)

Let us emphasize that the *realizations* of the random sequence $(X_0, X_1, X_2, \ldots)$ do *not* in general converge to any *fixed* point in S. Instead, the limit describes a *statistical equilibrium* (*tilastollinen tasapaino*) where the Markov chain will settle in the long run[6].

In applications, one is usually interested in computing the limiting distribution. It turns out that any limiting distribution is invariant under the dynamics of the Markov chain (Theorem 2.2).

Definition. A probability distribution $\pi = (\pi(x) : x \in S)$ is called an *invariant distribution* (*tasapainojakauma*) of a transition matrix P and the corresponding Markov chain $X = (X_0, X_1, X_2, \ldots)$ if it satisfies the *balance equations* (*tasapainoyhtälöt*)

$$\sum_{x\in S} \pi(x) \cdot P(x, y) = \pi(y), \qquad \text{for all } y \in S, \tag{2.1}$$

or in matrix form (with π interpreted as a row-vector indexed by $y \in S$),

$$\pi \cdot P = \pi. \tag{2.2}$$

Such a distribution π is also called a *stationary distribution* (*stationaarinen jakauma*).

[6] For mathematically oriented readers, let us note that this is usually expressed mathematically in the form $X_t \xrightarrow{(\mathrm{d})} X_\infty$, which means that the random sequence $(X_0, X_1, X_2, \ldots)$ "converges in distribution" towards a random variable X_∞ which is distributed according to a probability distribution μ_∞.

▷ It is important to note that, since $\pi = (\pi(x) : x \in S)$ is probability distribution, the law of total probability holds for any invariant distribution:

$$\sum_{x \in S} \pi(x) \;=\; 1. \tag{2.3}$$

Thus, to find an invariant distribution, one has to solve the system of balance equations (2.1) *together with the normalization* (2.3).

▷ In fact, when the state space is *finite*, $|S| < \infty$, one can always find a solution to Equations (2.1, 2.3) (Theorem 2.1). However, the solution might not be *unique* (Theorem 2.8).

▷ Note also that the balance equation in matrix form (2.2) defines an invariant distribution as a *left eigenvector* of the transition matrix P with eigenvalue 1.

▷ The balance equation (2.1) can be viewed as a conservation-of-mass property in the sense that it is equivalent to[7]

$$\sum_{x \in S} \pi(x) \cdot P(x,y) \;=\; \pi(y) \cdot \sum_{z \in S} P(y,z) \;=\; \sum_{z \in S} \pi(y) \cdot P(y,z), \qquad \text{for all } y \in S. \tag{2.4}$$

Indeed, interpreting each $\pi(x)$ as the probability mass at state $x \in S$, the left-hand side of identity (2.4) can be interpreted as the flow of mass coming into state y, while the right-hand side of identity (2.4) can be interpreted as the flow of mass going out from state y.

Theorem 2.1. *Every finite-state Markov chain has an invariant distribution π.*

(The proof can be skipped at the first reading. It uses basic linear algebra.)

Proof. Let $\mathbf{1} = [1, 1, \ldots, 1]$ be the row-vector whose each entry is 1. Multiplying $\mathbf{1}$ by the transition matrix P from the left gives

$$(P \cdot \mathbf{1})(x) \;=\; \sum_{y \in S} P(x,y) \cdot \mathbf{1}(y) \;=\; \sum_{y \in S} P(x,y) \overset{(1.1)}{=} 1 \;=\; \mathbf{1}(x), \qquad x \in S.$$

This shows that $P \cdot \mathbf{1} = \mathbf{1}$, so 1 is an eigenvalue of P. In particular, there exists a non-zero left eigen(row)vector $\tilde{\pi} \neq 0$ of P with eigenvalue 1, so that $\tilde{\pi} \cdot P = \tilde{\pi}$, that is, $\tilde{\pi}$ satisfies the balance equations (2.2). Since S is a finite set, we have $\Pi := \sum_{x \in S} \tilde{\pi}(x) < \infty$, so

$$\pi(x) \;=\; \tfrac{1}{\Pi} \cdot \tilde{\pi}(x)$$

also satisfies the balance equations (2.2) and, moreover, the normalization (2.3). We conclude that π is an invariant distribution for the transition matrix P. □

It is, however, important to note that a *limiting distribution* might not exist, even if invariant distributions do exist. We will study this phenomenon in Example 2.4 and Theorem 2.14.

Theorem 2.2 (*Limiting distribution is also invariant*)**.** *If μ_∞ is a* limiting *distribution of a finite-state Markov chain, then it is also an* invariant *distribution.*

[7] You can see this by inserting the row-sum (1.1) (equaling 1) into the right-hand side of (2.1).

Proof. By the associativity of matrix multiplication, we see that

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

which can be written entry-by-entry in the form

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

If we assume that as the time grows, $t \to \infty$, we have $\mu_t(x) \to \mu_\infty(x)$ for every $x \in S$, then we see by taking limits on both sides of the above equation that

$$\mu_\infty(y) \;=\; \lim_{t\to\infty} \mu_{t+1}(y) \;=\; \sum_{x \in S} \Big(\lim_{t\to\infty} \mu_t(x) \Big) \cdot P(x, y) \;=\; \sum_{x \in S} \mu_\infty(x) \cdot P(x, y).$$

Hence, the balance equation (2.1) holds. Moreover, because μ_t is a probability distribution,

$$\sum_{x \in S} \mu_t(x) \;=\; 1, \qquad \text{for all } t \geq 0.$$

By taking limits on both sides of the above equation as $t \to \infty$, we now see that also the law of total probability (2.3) holds, so μ_∞ is indeed a probability distribution on S. □

Observe that, if Markov chain $X = (X_0, X_1, \ldots)$ is started from an invariant initial distribution $\mu_0 = \pi$, we find using Theorem 1.5 and the associativity of matrix multiplication that

$$\mu_t \;=\; \pi \cdot P^t \;=\; (\pi \cdot P) \cdot P^{t-1} \;=\; \pi \cdot P^{t-1} \;=\; \cdots \;=\; \pi \cdot P \;=\; \pi.$$

Hence, for any Markov chain with a random initial state X_0 distributed according to an invariant distribution, *the distribution of X_t remains invariant for all time instants* $t \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$.

2.2 Examples

Theorem 2.2 tells that a limiting distribution (if exists) can be determined as a solution of the linear system of equations (2.1, 2.3). Let us now look into some examples, which show how the limiting distribution can be computed in practise, how it may depend on the initial distribution μ_0, and how it can happen that a limiting distribution does not exist but invariant distributions do exist. (We will discuss uniqueness of invariant distributions a little bit later.)

Example 2.3 (*Brand loyalty*)**.** A smartphone market is dominated by three manufacturers:

$$\text{state '1' = 'Apple'}, \qquad \text{state '2' = 'Nokia'}, \qquad \text{and} \qquad \text{state '3' = 'Samsung'}.$$

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 (uniformly). Assume that

$$\beta_1 = 0.8, \qquad \beta_2 = 0.6, \qquad \text{and} \qquad \beta_3 = 0.4,$$

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 $X = (X_0, X_1, \ldots)$ 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 easily compute powers of P using a computer:

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

$$P^{10} = \begin{bmatrix} 0.5471287 & 0.2715017 & 0.1813696 \\ 0.5430034 & 0.2745217 & 01824748 \\ 0.5441087 & 0.2737123 & 0.1821790 \end{bmatrix}, \ldots,$$

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

The above computations indicate that after 20 phone purchases, an initial customer of manufacturer "i" is a customer of manufacturer "1" (Apple) with probability

$$P^{20}(i,1) \approx 0.545.$$

Because the rows of P^{20} are approximately equal, we can see that the effect of initial state $i \in S = \{1,2,3\}$ becomes negligible over time. Hence, it appears that the market shares indeed stabilize towards a (unique) *limiting distribution*

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

The balance equations (2.1), $\pi \cdot P = \pi$, and the normalization $\sum_{x=1}^{3} \pi(x) = 1$, can be written as

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

The unique solution of the above system of linear equations is

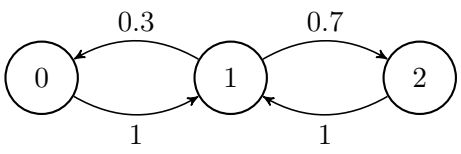
$$\pi = \left[\frac{6}{11}, \frac{3}{11}, \frac{2}{11} \right] \approx [0.5454545, 0.2727273, 0.1818182],$$

close to the numerically found limiting distribution, as it should according to Theorem 2.2. ■

The limiting distribution in Example 2.3 does not depend on the choice of the initial state of the Markov chain (or its initial distribution): there is just one distribution π which is the *unique invariant distribution* of the Markov chain, and also the unique limiting distribution *started from any initial state*. The next example shows that there might not exist a limiting distribution.

Example 2.4 (*Ehrenfest Markov chain*)**.** The Ehrenfest Markov chains, named after the physicist Paul Ehrenfest (1180–1933), are simple, discrete models for the exchange of gas molecules between containers. The easiest such model has two containers (left and right) and two particles (molecules or such). Denote by $X_t \in \{0,1,2\}$ the number of particles in the right container at time $t \in \mathbb{N}_0 = \{0,1,2,\ldots\}$. Consider $X = (X_0, X_1, X_2, \ldots)$ as a Markov chain on state space $S = \{0,1,2\}$ with initial state $X_0 = 0$, and transition matrix

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



By computing powers of P, we see that

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

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

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 Markov chain with the deterministic initial state $X_0 = 0$ (corresponding to initial distribution $\mu_0 = [1, 0, 0]$ where there are no particles in the right container) hence satisfies

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

Clearly such a Markov chain has no limiting distribution, as its time-dependent distribution jumps between the two possibilities $[0,\ 1,\ 0]$ and $[0.3,\ 0,\ 0.7]$. However, a direct computation still shows that average of these two distributions,

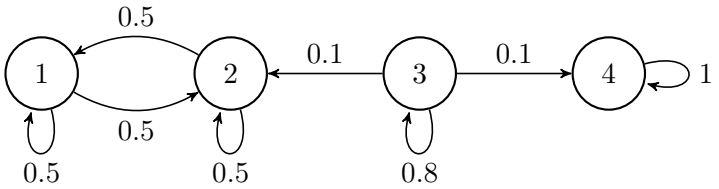
$$\pi = [0.15,\ 0.50,\ 0.35],$$

is an invariant distribution for the Ehrenfest Markov chain. ■

The next example shows that there might exist several different limiting distributions, depending on the initial state of the Markov chain. This also implies that there are several different invariant distributions (by Theorem 2.2).

Example 2.5 (*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

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

As a consequence (by Theorem 2.2), both[8]

$$\pi^{(12)} = [0.5,\ 0.5,\ 0,\ 0] \qquad \text{and} \qquad \pi^{(4)} = [0,\ 0,\ 0,\ 1]$$

are invariant distributions of P. In fact, by linearity, one can verify that every probability distribution of the form (that is, a *convex* combination of the two distributions $\pi^{(12)}$ and $\pi^{(4)}$)

$$\pi = \alpha \cdot \pi^{(12)} + (1 - \alpha) \cdot \pi^{(4)}, \qquad 0 \leq \alpha \leq 1,$$

is an invariant distribution of P. ■

[8]Here, we use the notation $\pi^{(12)}$ and $\pi^{(4)}$ respectively for the different invariant distributions related to the different absorbing components $\{1, 2\}$ and $\{4\}$ of the Markov chain.

2.3 Irreducibility and uniqueness of invariant distribution

One can verify directly that all Markov chains in Examples 1.2, 1.3, and 2.3 have a unique invariant distribution π. The common feature of these examples is that one can get from any state to any other state by steps of the Markov chain, that is, they are *irreducible*.

Given a transition matrix P, we denote $x \rightsquigarrow y$ if the corresponding transition diagram contains a (directed) path from x to y. Here we allow paths of length zero, so that $x \rightsquigarrow x$.

Definition. A transition matrix P and the corresponding Markov chain X is called[a] *irreducible* (*yhtenäinen*), if $x \rightsquigarrow y$ for all $x, y \in S$.

[a] In graph-theoretical terms, a Markov chain is irreducible if and only if its transition diagram is a strongly connected directed graph.

Otherwise, P and X are called *reducible* (*epäyhtenäinen*).

Example 2.6 (*Irreducible Markov chains*)**.** The following Markov chains are irreducible:

- ▷ Weather model (Example 1.2),
- ▷ Inventory model (Example 1.3),
- ▷ Brand loyalty (Example 2.3).

■

Example 2.7 (*PageRank*)**.** How about PageRank Markov chain in Example 1.4? Does its behavior depend on the underlying graph or the damping factor $c \in [0, 1]$? ■

In fact, only reducibility can prevent the uniqueness of the invariant distribution. (Moreover, for reducible Markov chains, we saw in Example 2.5 that convex combinations of invariant distributions for the different components of the Markov chain are also invariant.)

Theorem 2.8 (*Uniqueness of invariant distribution*)**.** *Consider a finite-state Markov chain X. If X is* irreducible*, then its invariant distribution π is* unique*.*

(The proof sketch can be skipped at the first reading. It uses basic linear algebra.)

Proof sketch. The fact that the invariant distribution is unique can be justified by first verifying that for an irreducible transition matrix P, all column-vectors solving $P \cdot v = v$ must have the form $v = [a, a, \ldots, a]^T$, for some constant $a \in \mathbb{R}$, so that the null space of $P - I$ is one-dimensional. Using basic facts of linear algebra, one can then conclude from this that also the linear space of (row-vector) solutions to $\mu \cdot (P - I) = 0$ has dimension one. In particular, this space contains at most one solution satisfying the normalization constraint (2.3): $\sum_x \mu(x) = 1$. Hence, an irreducible transition matrix P may have at most one invariant distribution. For mathematically oriented readers, a complete proof is available in [LPW08, Section 1.5]. □

How can one then verify the irreducibility? One useful criterion is the following result.

Theorem 2.9 (*Irreducibility*)**.** *A transition matrix P and the corresponding Markov chain X is irreducible* if and only if *for all $x, y \in S$ there exists an integer $t \geq 1$ such that*

$$P^t(x, y) > 0.$$

(The proof can be skipped at the first reading.) The key to the proof is Theorem 1.7:

$$P^t(x,y) \;=\; \mathbb{P}\,(X_t = y \mid X_0 = x), \qquad \text{for all } x,y \in S. \tag{2.5}$$

Proof. We first prove the implication "⇒". Assume that P (and X) is irreducible and pick some states $x \neq y$. Then, the transition diagram contains a path $x = x_0 \to x_1 \to \cdots \to x_t = y$, so that

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

As a consequence, we have

$$\begin{aligned}
P^t(x,y) \;&=\; \mathbb{P}\,(X_t = y \mid X_0 = x) && \text{[by (2.5)]}\\
&=\; \mathbb{P}\,(X_t = x_t \mid X_0 = x_0)\\
&\geq\; \mathbb{P}\,(X_t = x_t,\, X_{t-1} = x_{t-1},\, \ldots, X_1 = x_1 \mid X_0 = x_0) && \text{[by monotonicity of } \mathbb{P}\text{]}\\
&=\; P(x_0,x_1)\cdot P(x_1,x_2)\cdots P(x_{t-1},x_t) && \text{[by (1.11)]}\\
&>\; 0.
\end{aligned}$$

We then prove the implication "⇐". Pick some states $x \neq y$ and assume that $P^t(x,y) > 0$ for some integer $t \geq 1$. Then, $\mathbb{P}\,(X_t = y \mid X_0 = x) > 0$ by (2.5), so that the Markov chain starting at x can be located in state y after t time instants with positive probability. This is only possible if the transition diagram contains a path of length t from x to y, so that $x \rightsquigarrow y$. This shows that P is irreducible, since x and y were arbitrary states. □

The structure of Markov chains can be analyzed by defining a symmetric relation by denoting

$$x \leftrightsquigarrow y \qquad \text{if and only if} \qquad x \rightsquigarrow y \quad \text{and} \quad y \rightsquigarrow x.$$

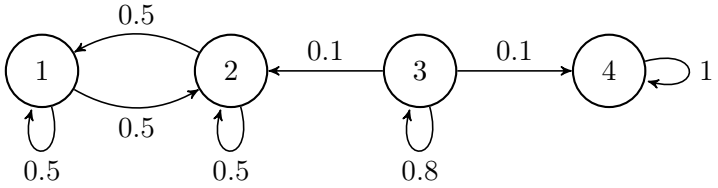
This equivalence relation partitions the state space into equivalence classes

$$C(x) = \{y \in S : y \leftrightsquigarrow x\}, \qquad x \in S,$$

called the *components* (*komponentit*) of X. An irreducible Markov chain has only one component, which contains all states of the state space.

- ▷ A component is called called *absorbing* (*absorboiva*) if the Markov chain cannot exit the component.
- ▷ Otherwise, the component is called *transient* (*väistyvä*).

Example 2.10 (*Reducible Markov chain*)**.** The Markov chain in Example 2.5 is not irreducible because it cannot move away from state 4. The transition diagram



of this Markov chain has 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. ■

2.4 Periodicity and aperiodicity

Recall that a Markov chain may not always admit limiting distributions: the distribution may jump forever between a number of possibilities (Example 2.4). We now investigate this precisely.

Definition. The *period* (*jakso*) of state $x \in S$ for Markov chain X with transition matrix P is the *greatest common divisor (gcd)* (*suurin yhteinen tekijä (syt)*) of all the time instants at which the Markov chain started at $X_0 = x$ may return to its initial state.

The set of possible return times can be written as

$$\begin{aligned} \mathcal{T}_x &= \{t \geq 1 : \mathbb{P}(X_t = x \mid X_0 = x) > 0\} \\ &= \{t \geq 1 : P^t(x,x) > 0\}, \qquad \text{[by (1.10)]} \end{aligned}$$

so that the period of x is given by 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, one can determine the period 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 this number d is the period of x.
- ▷ Note also that if $P(x,x) > 0$, then the period of state x is 1. Indeed, the set of possible return times in this case is $\mathcal{T}_x = \{1, 2, 3, \dots\}$, with the greatest common divisor 1.
- ▷ If P is irreducible, then the period of any state is the same. Indeed, for any two states $x, y \in S$, by irreducibility (Theorem 2.9), there are times $t, s \geq 0$ such that $P^t(x,y) > 0$ and $P^s(y,x) > 0$. Now, by drawing a picture, you can easily verify the following claims:
  - * The time $t + s \in \mathcal{T}_x \cap \mathcal{T}_y$.
  - * For any $u \in \mathcal{T}_x$, we have $u + t + s \in \mathcal{T}_y$.
  - * For any $u \in \mathcal{T}_y$, we have $u + t + s \in \mathcal{T}_x$.

  This shows that the periods of x and y must be the same.

Definition. A transition matrix P and the corresponding Markov chain X is called *aperiodic* (*jaksoton*) if every state has period 1; and otherwise, *periodic* (*jaksollinen*).

Example 2.11 (*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 2.3).

■

Example 2.12 (*Periodic Markov chain*)**.** The Ehrenfest Markov chain in Example 2.4 is periodic with every state having period 2. ■

Example 2.13 (*PageRank*)**.** How about PageRank Markov chain in Example 1.4? Does its behavior depend on the underlying graph or the damping factor $c \in [0, 1]$? ■

2.5 Markov Chain Convergence Theorem

In fact, any aperiodic Markov chain has a limiting distribution started at any initial distribution μ_0. The limiting distribution can, however, depend on the choice of the initial distribution μ_0. To guarantee that the limiting distribution is independent of this choice, one has to assume in addition that the Markov chain is irreducible (i.e., has only one component).

Theorem 2.14 (*Convergence Theorem*)**.** *Consider a finite-state Markov chain X.*

1. *If X is* aperiodic*, then it admits a* limiting distribution *starting from any given initial distribution μ_0, that is, the following limit exists:*

$$\lim_{t\to\infty} \mu_0 \cdot P^t = \lim_{t\to\infty} \mu_t = \mu_\infty.$$

   (*However, μ_∞ may depend on the initial distribution μ_0*).

2. *If X is* aperiodic *and* irreducible*, then the limiting distribution μ_∞ is independent of the initial distribution μ_0, and equals the unique* invariant distribution *of X. It can be determined as the unique solution $\pi = \mu_\infty$ to balance equations* (2.1):

$$\sum_{x\in S} \pi(x) \cdot P(x,y) = \pi(y), \qquad \text{for all } y \in S,$$

   and the normalization (2.3):

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

(The proof sketch can be skipped at the first reading. We return to this in Chapter 5.)

Proof sketch. Theorem 2.2 says that the limiting distribution is also an invariant distribution. The existence of the limit can be proved by (at least) two methods. The precise proof requires techniques beyond this course. Students majoring in mathematics are recommended to have a look at [LPW08, Sections 4-5], where both proof techniques in the irreducible case are explained in detail. There, [LPW08, Theorem 4.9] is the main statement, phrased in terms of so-called "total variation distance" for distributions. The first proof essentially uses methods of matrix analysis [LPW08, Section 4], while the second proof relies on a very useful general technique known as *stochastic coupling* [LPW08, Section 5], that we will also use later for Theorem 5.6. □

3 Markov additive processes and ergodicity

3.1 Ergodicity

Ergodicity refers to a phenomenon where time-averages and space-averages become the same in the long run. So far, we have learned that the distribution of an irreducible and aperiodic Markov chain converges to the unique invariant distribution π of the Markov chain (Theorem 2.14). This distribution π can be viewed as a *space-average*: for each state $y \in S$ in the state space S, the value $\pi(y)$ is the probability that y occurs with respect to the distribution π:

$$\text{if} \quad Y \sim \pi, \qquad \text{then} \quad \pi(y) = \mathbb{P}[Y = y].$$

For any function $\phi : S \to \mathbb{R}$, the *expected value* of $\phi(Y)$ with respect to π is, by definition,

$$\mathbb{E}[\phi(Y)] = \sum_{y \in S} \pi(y) \cdot \phi(y).$$

The following results provide alternative interpretations for the invariant distribution π:

- ▷ The long-term *time-average* of the random sequence $X_0, X_1, X_2, \ldots$ is close to the *expected value* of the invariant distribution, in the sense of Equation (3.1) in Theorem 3.1. This phenomenon is called *ergodicity* (*ergodisuus*) property. It is a sort of *law of large numbers*.
- ▷ The long-term relative *frequency* of X to visit any state y is close to the probability $\pi(y)$ of y in the invariant distribution, see Equation (3.3) in Theorem 3.3.

Because of the following result, irreducible Markov chains are also termed *ergodic Markov chains* in the literature. By Theorem 2.8, every irreducible finite-state Markov chain has a unique invariant distribution π. Even though there is no guarantee for the existence of a limiting distribution, we have a limit for the time-averages. (The proof will be skipped in this course.)

Theorem 3.1 (*Ergodic theorem*)**.** *For any* irreducible *Markov chain* $X = (X_0, X_1, X_2, \ldots)$ *with invariant distribution* π *on finite state space* S*, we have*

$$\frac{1}{t} \sum_{s=0}^{t-1} \phi(X_s) \longrightarrow \sum_{y \in S} \pi(y) \cdot \phi(y), \qquad \text{as } t \to \infty, \tag{3.1}$$

for any function $\phi : S \to \mathbb{R}$ *with probability one, regardless of the initial state of the Markov chain.*

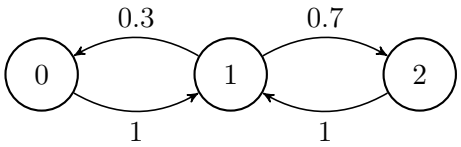
Note that periodicity is not an issue here, because the time-averages smoothen out any possible periodic effects present in the model. See Example 3.2.

(The proof sketch can be skipped at the first reading. It uses basic probability tools, and the details presented in [LPW08] are quite illuminating for mathematically oriented readers.)

Proof sketch. The ergodicity can be proved by fixing an initial state $X_0 = x$ and keeping track of successive visits of the Markov chain to x over time. The Markov property (1.2) implies that the paths between successive visits are stochastically independent, and the ergodicity property (3.1) can be proved by applying a strong law of large numbers — see [LPW08, Appendix C]. □

Example 3.2 (*Ehrenfest Markov chain*)**.** Consider the irreducible Markov chain $X = (X_0, X_1, \ldots)$ on state space $S = \{0, 1, 2\}$ of Example 2.4, where each state has period 2, with transition matrix

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



As X is periodic, it does not have a limiting distribution, but it does have a unique invariant distribution (by Theorem 2.8, π is unique — or you can check it directly)

$$\pi = [\pi(0), \pi(1), \pi(2)] = [0.15, 0.50, 0.35].$$

We can use the Ergodic Theorem 3.1 (with the identity function $\phi(x) = x$) to find that the long-term time-average of the number of particles in the right container is

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} X_s = 0 \cdot \pi(0) + 1 \cdot \pi(1) + 2 \cdot \pi(2) = 0.50 + 2 \cdot 0.35 = 1.2.$$

What would it be if the transition probabilities would be symmetric: $P(0,1) = 0.5 = P(1,2)$? ■

3.2 Long-term relative frequencies and occupation in the long run

As an important consequence, we obtain the following result regarding (empirical) relative frequencies. Recall that the *(relative) frequency* of state y is the time-average (1.15) of the number $N_t(y)$ of visits (1.14) of X to y during the first t time steps:

$$\frac{N_t(y)}{t} = \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{1}(X_s = y), \qquad y \in S. \tag{3.2}$$

Note that (3.2) is a random number depending on the state y of interest, determined by the realized trajectory of the Markov chain at its first t steps $(X_0, X_1, \ldots, X_{t-1})$. The following result confirms that the value of the invariant distribution $\pi(y)$ of X can be interpreted as[9] the *long-term average time that the Markov chain spends in state* y.

Theorem 3.3 (*Long-term frequencies of visits*)**.** *For any* irreducible *Markov chain* $X = (X_0, X_1, X_2, \ldots)$ *with invariant distribution* π *on finite state space* S*, the relative frequencies satisfy*

$$\lim_{t\to\infty} \frac{N_t(y)}{t} = \pi(y), \qquad \text{for all } y \in S, \tag{3.3}$$

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

Moreover, the occupancy matrix

$$G_t(x, y) = \mathbb{E}\left(N_t(y) \mid X_0 = x\right)$$

of the Markov chain also satisfies

$$\lim_{t\to\infty} \frac{G_t(x, y)}{t} = \pi(y) \qquad \text{for all } x, y \in S. \tag{3.4}$$

In particular, the limit (3.4) *is independent of the initial state* x*.*

[9] I.e., long-term time-average of the occupancy time, or long-term relative frequency.

Proof. The number $N_t(y)$ of visits (3.2) can be written as

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

where $\phi : S \to \mathbb{R}$ is the function

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

By applying Theorem 3.1, we conclude that

$$\lim_{t\to\infty} \frac{N_t(y)}{t} \;=\; \sum_{x\,\in\,S} \pi(x)\cdot\phi(x) \;=\; \sum_{x\,\in\,S} \pi(x)\cdot\mathbb{1}\{x = y\} \;=\; \pi(y),$$

with probability one, regardless of the initial state. This proves (3.3).

To prove (3.4), note that the relative frequency of state y is bounded by

$$0 \;\le\; \frac{N_t(y)}{t} \;\le\; 1, \qquad \text{with probability one for all } t \in \mathbb{N}_0.$$

By taking the limit $t \to \infty$ inside the expectation[10] and applying (3.3), it follows that

$$\lim_{t\to\infty} \frac{G_t(x,y)}{t} \;=\; \lim_{t\to\infty} \mathbb{E}\Big(\frac{N_t(y)}{t} \;\Big|\; X_0 = x\Big) \;=\; \mathbb{E}\Big(\lim_{t\to\infty} \frac{N_t(y)}{t} \;\Big|\; X_0 = x\Big) \;=\; \pi(y).$$

□

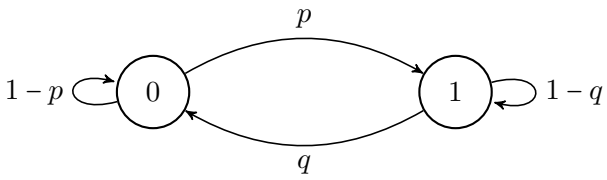
Example 3.4 (*November rain*)**.** The November weather of day $t \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ in Lahti can be modeled using a stochastic process in state space $\{0, 1\}$, where

$$\text{state '0' = 'dry'} \qquad \text{and} \qquad \text{state '1' = 'rain'}.$$

The weather is represented as a Markov chain $X = (X_0, X_1, X_2, \ldots)$ with transition matrix

$$P \;=\; \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

(note that the rows and columns are indexed by the states $x = 0, 1$), and transition diagram



where $p, q \in [0, 1]$. For concreteness, let us assume that $p = 0.75$ and $q = 0.5$.

Because the Markov chain is irreducible and aperiodic, by Theorem 2.14 it converges to a unique limiting distribution, which is also its invariant distribution $\pi = [\pi(0), \pi(1)] = [0.4, 0.6]$ — solved from the balance equations (2.1) and the normalization (2.3):

$$\begin{aligned} 0.25\cdot\pi(0) \;+\; 0.5\cdot\pi(1) \;&=\; \pi(0), \\ 0.75\cdot\pi(0) \;+\; 0.5\cdot\pi(1) \;&=\; \pi(1), \\ \pi(0) \;+\; \pi(1) \;&=\; 1. \end{aligned}$$

[10] Allowed for bounded random sequences by Lebesgue's dominated convergence theorem [Kyt20, Thm. VII.22].

We can use the Ergodic Theorem 3.1 (with the identity function $\phi(x) = x$) to find that the time-average of the "raininess" is

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} X_s \;=\; 0\cdot\pi(0) \,+\, 1\cdot\pi(1) \;=\; \pi(1) \;=\; 0.6,$$

and Theorem 3.3 to find that (with probability one) the relative "rain frequency" is given by (3.3):

$$\lim_{t\to\infty} \frac{N_t(1)}{t} \;=\; \pi(1) \;=\; 0.6,$$

where $N_t(1)$ is the number of rainy days among the first t days.

In finite time-horizon, similarly as in Example 1.11, we can predict the expected number of rainy days during November (30 days): according to Theorem 1.10, this is the entry $G_{30}(x,1)$ of the occupancy matrix G_{30} at time $t = 30$, where $x \in \{0,1\}$ is the state of day zero. Applying (1.16) from Theorem 1.10 (and a computer), we find that

$$G_{30} \;=\; \begin{bmatrix} G_{30}(0,0) & G_{30}(0,1) \\ G_{30}(1,0) & G_{30}(1,1) \end{bmatrix} \;=\; \sum_{s=0}^{29} \begin{bmatrix} 0.25 & 0.75 \\ 0.5 & 0.5 \end{bmatrix}^s \;=\; \begin{bmatrix} 12.5 & 17.5 \\ 11.7 & 18.3 \end{bmatrix}.$$

According to the prediction, the expected number of rainy days is hence

$$\begin{cases} G_{30}(0,1) \;=\; 17.5, & \text{if } x = 0, \\ G_{30}(1,1) \;=\; 18.3, & \text{if } x = 1. \end{cases}$$

The time-averages of these values are

$$\begin{cases} \dfrac{G_{30}(0,1)}{30} \;=\; 0.58, & \text{if } x = 0, \\ \dfrac{G_{30}(1,1)}{30} \;=\; 0.61, & \text{if } x = 1. \end{cases}$$

These are not very far from the equilibrium values, given by Equation (3.4) in Theorem 3.3:

$$\lim_{t\to\infty} \frac{G_t(x,1)}{t} \;=\; \pi(1) \;=\; 0.6 \qquad \text{independently of } x \in \{0,1\}.$$

■

Markov additive processes provide more sophisticated models based on Markov chains. To illustrate the idea, we modify the weather model by introducing temperature in Example 3.5.

3.3 Markov additive processes — cost/profit models

In many applications, we need to analyze sums of random numbers which depend on the realized trajectory of a Markov chain. Examples include cumulative rewards in reinforcement learning, cost/profit 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.

In general, Markov additive processes have an underlying *Markov component* and an *additive component* that represents cumulative values of some quantity (e.g. profit). To illustrate the idea, let us first consider an example.

Example 3.5 (*November snow*)**.** A simple model for snowy days in November in Lahti consists of a Markov chain $X = (X_0, X_1, X_2, \ldots)$ with state space $S = \{-25, -24, \ldots, -1, 0, +1, \ldots, +24, +25\}$ modeling the daily temperature[11], and a random variable U with two possible values:

$$U = 0 = \text{'dry'} \qquad \text{and} \qquad U = 1 = \text{'rain'}.$$

For concreteness, let us model the probability of rainy versus dry weather by the statistical equilibrium values of the weather model in Example 3.4: $U \sim \pi$, that is,

$$\mathbb{P}(U = 1) = 0.6 \qquad \text{and} \qquad \mathbb{P}(U = 0) = 0.4.$$

We will also assume that U (rain indicator) is independent of Markov chain X (daily temperatures). We then define the random variables

$$C(y) = \mathbb{1}\{y \leq -1\} \cdot U, \qquad y \in S,$$

which indicate whether the weather is snowy:

$$C(y) = 1 = \text{'it snows'} \qquad \text{and} \qquad C(y) = 0 = \text{'it does not snow'},$$

since it snows if and only if the temperature is below zero and it rains. The events when it does not snow include both rainy days with non-negative temperature, and dry days with any temperature (for simplicity, we are only interested in snow in this example).

Thus, the number V_t of *snowy days* among the first t days of the month can be expressed as

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

where for each time instant $s \geq 0$ and each state $y \in S$, the random variable $C_s(y)$ is just an independent copy of $C(y)$. Now, we ask: What is the expected number of snowy days during November (30 days), if the first of November had temperature zero ($X_0 = 0$)? To answer this question, we need to find the expected number of days when it rains and the temperature is below zero. The answer is handily provided to us by Theorem 3.6 proven below: it holds that

$$\begin{aligned}
\mathbb{E}(V_{30} \mid X_0 = 0) &= \sum_{y=-25}^{+25} \sum_{s=0}^{29} P^s(0, y) \cdot \mathbb{E}(C(y)) \\
&= \sum_{y=-25}^{+25} \sum_{s=0}^{29} P^s(0, y) \cdot \mathbb{1}\{y \leq -1\} \cdot \mathbb{P}(U = 1) \\
&= 0.6 \cdot \sum_{y=-25}^{-1} \sum_{s=0}^{29} P^s(0, y). \qquad (3.5)
\end{aligned}$$

Thus, given the transition matrix P for the temperature Markov chain X, and the probability distribution of the rain indicator U, we could use a computer to answer the above question.

The transition matrix of the Markov component X has size 51×51 in general. For the purposes of modeling snowfall, it is quite reasonable to model the temperature with a smaller Markov chain on state space $\{-2, -1, 0, +1, +2\}$, where

$$\begin{aligned}
&\text{state '}-2\text{'} = \text{'temperature} \leq -2\text{'}, \\
&\text{state '}-1\text{'} = \text{'temperature} = -1\text{'}, \\
&\text{state } \quad \text{'}0\text{'} = \text{'temperature} = 0\text{'}, \\
&\text{state '}+1\text{'} = \text{'temperature} = +1\text{'}, \\
&\text{state '}+2\text{'} = \text{'temperature} \geq +2\text{'},
\end{aligned}$$

[11] Do you think a simple Markov chain is good for modeling temperature? Remember that we try to study very simplified "toy models" in this course, and especially the weather models are not very realistic.

and with transition matrix

$$P = \begin{bmatrix} P(-2,-2) & P(-2,-1) & P(-2,0) & P(-2,+1) & P(-2,+2) \\ P(-1,-2) & P(-1,-1) & P(-1,0) & P(-1,+1) & P(-1,+2) \\ P(0,-2) & P(0,-1) & P(0,0) & P(0,+1) & P(0,+2) \\ P(+1,-2) & P(+1,-1) & P(+1,0) & P(+1,+1) & P(+1,+2) \\ P(+2,-2) & P(+2,-1) & P(+2,0) & P(+2,+1) & P(+2,+2) \end{bmatrix}$$

$$= \begin{bmatrix} 0.8 & 0.2 & 0 & 0 & 0 \\ 0.1 & 0.8 & 0.1 & 0 & 0 \\ 0 & 0.1 & 0.8 & 0.1 & 0 \\ 0 & 0 & 0.1 & 0.8 & 0.1 \\ 0 & 0 & 0 & 0.2 & 0.8 \end{bmatrix}. \tag{3.6}$$

Then, we can use a computer to easily compute (3.5):

$$\mathbb{E}\,(V_t \mid X_0 = 0) \;=\; 0.6 \cdot \sum_{y=-2}^{-1} \sum_{s=0}^{29} P^s(0,y) \;\approx\; 5.8.$$

■

General setup. To introduce the general setup, consider Markov chain $X = (X_0, X_1, X_2, \ldots)$ on finite state space S. Assume that at each time $s \geq 0$, a cost/profit depending on the state $X_s = y$ occurs. Let us model the cost/profit at state y by a *real-valued random variable* $C(y)$. To allow for a time-evolution of the cost/profit process as well, for each time instant $s \geq 0$, we shall take *independent copies* of $C(y)$: that is, for every state $y \in S$, we have a collection

$$\{C_s(y) : s \in \mathbb{N}_0\},$$

of *independent and identically distributed (iid)* random variables:

$$\mathbb{P}[C_s(y) = a] \;=\; \mathbb{P}[C(y) = a], \qquad \text{for all } y \in S \text{ and for all } a \in \mathbb{R}.$$

We will also assume that the profits/costs $C_s(y)$ are independent of $X_0, X_1, \ldots, X_s$, i.e., the Markov chain up to time s. This prevents the Markov chain to decide its transitions based on *future* profits/costs, but, crucially, allows for the transition from X_t to X_{t+1} be based on *current* cost/profit $C_t(X_t)$.

Reminder. A collection $\{U_j : j \in \mathbb{N}_0\}$ of random variables is said to be (iid): *independent and identically distributed* (*riippumattomat ja samoin jakautuneet*) if each random variable U_j has the same probability distribution as the others and all of them are mutually independent. ♣

Definition. A (time-homogeneous) *Markov additive process* (*Markov-additiivinen prosessi*) is a pair (X, V) of processes, where $X = (X_0, X_1, X_2, \ldots)$ is a Markov chain and[a]

$$V_t \;=\; \sum_{s=0}^{t-1} C_s(X_s).$$

▷ The random variables $\{C(y) : y \in S\}$ are called the *increments* (*lisäykset*) of (X, V).

▷ The process $X = (X_0, X_1, \ldots)$ is called the *Markov component* (*Markov-komponentti*).

▷ The process $V = (V_0, V_1, \ldots)$ is called the *additive component* (*summa-komponentti*).

[a] For time $t = 0$ we set $V_0 = 0$ by usual convention.

V_t is the *total cost/profit up to time t*. The *expected* total cost/profit up to time t is

$$\mathbb{E}(V_t) = \mathbb{E}\Big(\sum_{s=0}^{t-1} C_s(X_s)\Big) = \sum_{s=0}^{t-1} \mathbb{E}(C_s(X_s)).$$

To simplify some formulas, we will denote the expected increment at state y as

$$c(y) = \mathbb{E}(C(y)), \qquad y \in S.$$

The values $c = \{c(y) : y \in S\}$ form a column-vector with elements indexed by the possible states $y \in S$. Hence, we can multiply it from the left by the occupancy matrix G_t from Equation (1.16). The next result shows that we thus obtain the expected value of the total cost/profit.

Theorem 3.6 (*Expected total cost/profit in finite time-horizon*)**.** *For a Markov additive process (X, V) on a finite state space S, the expected total cost/profit up to time t can be computed using the occupancy matrix G_t as*

$$\mathbb{E}(V_t \mid X_0 = x) = \sum_{y \in S} \Big(\sum_{s=0}^{t-1} P^s(x,y)\Big) \cdot c(y) = \sum_{y \in S} G_t(x,y) \cdot c(y), \qquad \text{for all } x \in S, \quad (3.7)$$

or in matrix form (with c interpreted as a column-vector indexed by $y \in S$),

$$\mathbb{E}(V_t \mid X_0 = x) = (G_t \cdot c)(x), \qquad \text{for all } x \in S.$$

(The proof can be skipped at the first reading. It uses basic probability tools.)

Proof. By conditioning on all the possible states of X_s, we have

$$\begin{aligned}
\mathbb{E}(C_s(X_s) \mid X_0 = x) &= \sum_{y \in S} \underbrace{\mathbb{E}(C_s(X_s) \mid X_s = y,\ X_0 = x)}_{= \mathbb{E}(C(y))} \cdot \mathbb{P}(X_s = y \mid X_0 = x) \\
&= \sum_{y \in S} \mathbb{E}(C(y)) \cdot P^s(x,y), && \text{[by (1.10)]} \\
&= \sum_{y \in S} c(y) \cdot P^s(x,y), && [\mathbb{E}(C(y)) = c(y)]
\end{aligned}$$

where $\mathbb{E}(C_s(X_s) \mid X_s = y,\ X_0 = x) = \mathbb{E}(C(y))$ by the independence of the increment $C_s(y)$ and states X_s and X_0. We obtain

$$\begin{aligned}
\mathbb{E}(V_t \mid X_0 = x) = \mathbb{E}\Big(\sum_{s=0}^{t-1} C_s(X_s) \,\Big|\, X_0 = x\Big) &= \sum_{s=0}^{t-1} \mathbb{E}(C_s(X_s) \mid X_0 = x) \\
&= \sum_{s=0}^{t-1} \sum_{y \in S} c(y) \cdot P^s(x,y) \\
&= \sum_{y \in S} \Big(\sum_{s=0}^{t-1} P^s(x,y)\Big) \cdot c(y) \\
&= \sum_{y \in S} G_t(x,y) \cdot c(y) && \text{[by (1.16)]} \\
&= (G_t \cdot c)(x).
\end{aligned}$$

This is the asserted formula (3.7). □

Example 3.7 (*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 (so that the revenue from selling one laptop is (790-590) EUR=200 EUR), and that the storage expenses per week is 50 EUR for every laptop in stock at the beginning of a week. Determine the expected net profit from ten forthcoming weeks, when in the beginning of the first week there are five laptops in stock.

Denote by V_t the net profit (i.e., sales revenue minus storage expenses) 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 expenses (EUR) for the week equals $50\,X_t$, 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 D_t are mutually independent and identically distributed and D_t is also independent of $(X_0, \ldots, X_t)$, we can model the net profit as a Markov additive process.

Let $D \sim \mathrm{Poi}(\lambda)$ be a Poisson distributed random variable with mean $\lambda = 3.5$:

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

Then, since $X = (X_0, X_1, \ldots)$ is a Markov chain (Example 1.3) and the net profit can be written in the form

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

with increments $\{C(y) : y \in S\}$ distributed as

$$C(y) = (790 - 590)\min(y, D) - 50y, \qquad y \in S, \tag{3.8}$$

we see that (X, V) is a Markov additive process with Markov component X, additive component V, and increments (3.8). To compute the expectation of V_t using Theorem 3.6, we need to first compute the expectation

$$c(y) = \mathbb{E}(C(y)) = (790 - 590)\,\mathbb{E}(\min(y, D)) - 50y, \qquad y \in S.$$

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

$$\begin{aligned} \mathbb{E}(\min(y, D)) &= \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}\min(y, k) \\ &= \sum_{k=0}^{y} e^{-\lambda}\frac{\lambda^k}{k!}k + \Big(1 - \sum_{k=0}^{y} e^{-\lambda}\frac{\lambda^k}{k!}\Big)y \\ &= y - \sum_{k=0}^{y} e^{-\lambda}\frac{\lambda^k}{k!}(y - k), \end{aligned}$$

and hence, we obtain

$$c(y) = (790 - 590)\bigg(y - \sum_{k=0}^{y} e^{-\lambda}\frac{\lambda^k}{k!}(y - k)\bigg) - 50y.$$

The other ingredient in Theorem 3.6 is the occupancy matrix G_t, which can be computed from powers of the transition matrix (that we already found in Example 1.3):

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

Powers of P can be easily evaluated using a computer, and the multiplication of the column-vector $c = \{c(y) : y \in S\}$ from the left by the s:th matrix power P^s as well. We find that (recall that column vectors are indexed by the states $x = 2, 3, 4, 5$)

$$c = \begin{bmatrix} 266.78 \\ 352.61 \\ 395.29 \\ 400.20 \end{bmatrix} \qquad \text{and} \qquad \sum_{s=0}^{9} P^s \cdot c = \begin{bmatrix} 3627.24 \\ 3704.00 \\ 3735.81 \\ 3735.00 \end{bmatrix}.$$

We conclude that the expected net profit from next ten weeks is 3735 EUR.

Note that the expected net profit 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 profits

$$P^0 \cdot c = I \cdot c = c = \begin{bmatrix} 266.78 \\ 352.61 \\ 395.29 \\ 400.20 \end{bmatrix}$$

satisfying

$$395.29 = c(4) < c(5) = 400.20,$$

and indicates that actions which maximize one-week outcomes may not be optimal for longer time-horizons. We return to this model again in Example 3.11. ■

3.4 Behavior of time-averages in the long run

For a Markov additive process (X, V), the additive component $V = (V_0, V_1, V_2, \ldots)$ usually does not converge to a statistical equilibrium even if the underlying Markov chain X would. Rather, V_t might tend to infinity or minus infinity in the long run — note that V_t comprises sums of real-valued random variables. Therefore, it makes sense to analyze the long-term *growth rates* V_t/t. The following result tells that the expected growth rate has a limit as $t \to \infty$, which moreover does not depend on the initial state of the Markov component.

Theorem 3.8 (*Long-term time-average of expected total cost/profit*). *Assume that the Markov chain X is irreducible. Then, for Markov additive process (X, V) on finite state space S, the time-average of the expected total cost/profit satisfies*

$$\lim_{t\to\infty} \frac{1}{t} \mathbb{E}(V_t) = \sum_{y \in S} \pi(y) \cdot c(y) = \pi \cdot c, \tag{3.9}$$

regardless of the initial state of the Markov component.

Proof. By Theorem 3.6, we have

$$\frac{1}{t} \mathbb{E}(V_t \mid X_0 = x) = \frac{1}{t} (G_t \cdot c)(x), \qquad \text{for all } x \in S.$$

Taking the limit $t \to \infty$ of both sides and recalling that by (3.4) in Theorem 3.3,

$$\lim_{t\to\infty} \frac{G_t(x, y)}{t} = \pi(y) \qquad \text{for all } x, y \in S,$$

we obtain the asserted limit (3.9), regardless of the initial state $X_0 = x$ of the Markov chain. □

Example 3.9 (*November snow*)**.** Returning to Example 3.5, consider the simplified temperature Markov chain with transition matrix P given by Equation (3.6). Because P is irreducible, it has a unique invariant distribution

$$\pi \;=\; [0.125,\, 0.25,\, 0.25,\, 0.25,\, 0.125], \tag{3.10}$$

which can be solved from the balance equations (2.1) and the normalization (2.3). Theorem 3.8 gives the long-term time-average of the expected number of snowy days:

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t}\,\mathbb{E}\,(V_t) \;&=\; \sum_{y\,=\,-2}^{+2} \pi(y)\cdot c(y) && \text{[by (3.9)]}\\
&=\; 0.6\cdot \sum_{y\,=\,-2}^{+2} \pi(y)\cdot \mathbb{1}\{y \,\le\, -1\} && \text{[by (3.11)]}\\
&=\; 0.6\cdot \sum_{y\,=\,-2}^{-1} \pi(y)\\
&=\; 0.6\cdot (0.125 \,+\, 0.25) && \text{[by (3.10)]}\\
&=\; 0.225,
\end{aligned}$$

where we used the knowledge from Example 3.5 for $c(y) = \mathbb{E}\,(C(y))$:

$$c(y) \;=\; \mathbb{1}\{y \,\le\, -1\}\cdot \mathbb{P}\,(U = 1) \;=\; 0.6\cdot \mathbb{1}\{y \,\le\, -1\}. \tag{3.11}$$

In conclusion, the model predicts 22.5% of days to be snowy. ■

In fact, one can prove via similar arguments as for Ergodic Theorem 3.1 that the above result also holds for the process V_t/t itself.

Theorem 3.10 (*Long-term time-average of total cost/profit*)**.** *Assume that the Markov chain X is irreducible. Then, for Markov additive process (X, V) on finite state space S, the time-average of the total cost/profit satisfies*

$$\lim_{t\to\infty} \frac{1}{t}\, V_t \;=\; \sum_{y\,\in\, S} \pi(y)\cdot c(y) \;=\; \pi\cdot c$$

with probability one, regardless of the initial state of the Markov component.

Proof. This is a good exercise for students majoring in mathematics: check out the proof of Theorem 3.1 from [LPW08, Appendix C] and think how to obtain the asserted result. □

Example 3.11 (*Inventory model*)**.** Let us continue the analysis of Example 3.7. What is the long-term (expected) profit rate for the Katiskakauppa.com company?

Because the Markov chain X is irreducible, it has a unique invariant distribution π which can be solved from the balance equations (2.1) and the normalization (2.3). By applying Theorems 3.8 and 3.10, we conclude that the long-term (expected) profit rate equals

$$\lim_{t\to\infty} \frac{1}{t}\,\mathbb{E}\,(V_t) \;=\; \sum_{y\,\in\, S} \pi(y)\cdot c(y) \;=\; \lim_{t\to\infty} \frac{1}{t}\, V_t,$$

independently of the initial state X_0 of the inventory. After computing the numerical values, we find that the long-term profit rate equals 371.29 EUR per week. This corresponds to approximately 3713 EUR profit rate per a 10-week period, and is quite close to the expected cumulative profit computed in Example 3.7 — which depend on the initial state. ■

4 Passage times and hitting probabilities

4.1 Passage times

Consider Markov chain $X = (X_0, X_1, X_2, \ldots)$ on finite state space S.

Definition. The *passage time* (*kulkuaika*) of X into set $A \subset S$ on state space S is

$$T_A = \min\{t \geq 0 : X_t \in A\}, \tag{4.1}$$

with the notational convention that $T_A = \infty$ if the process X never visits A.

The passage time is a random variable which takes values in the extended set of integers

$$\mathbb{N}_0 \cup \{\infty\} = \{0, 1, 2, \ldots\} \cup \{\infty\}.$$

The *expected passage time* (*odotettu kulkuaika*) into set A for X starting at state $X_0 = x$ is

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

Theorem 4.1 (*Expected passage time*)**.** *For any (time-homogeneous) Markov chain with transition matrix P on finite state space S, the expected passage times into set $A \subset S$,*

$$\{k_A(x) : x \in S\},$$

satisfy the system of equations

$$k_A(x) = \begin{cases} 1 + \sum_{y \in S} P(x,y) \cdot k_A(y), & \text{if } x \notin A, \\ 0, & \text{if } x \in A. \end{cases} \tag{4.2}$$

▷ Note that since $k_A(y) = 0$ for all $y \in A$, we have

$$\sum_{y \in S} P(x,y) \cdot k_A(y) = \sum_{y \notin A} P(x,y) \cdot k_A(y).$$

▷ From the harmonic analysis point of view, the system of equations (4.2) corresponds to a discrete *Poisson equation* (*Poisson-yhtälö*) on the complement $A^c = S \setminus A$,

$$\Delta f(x) = -1, \qquad \text{for all } x \in A^c,$$

with boundary condition on $\partial A^c = \{y \in A : \exists\, x \in A^c \text{ s.t. } P(x,y) > 0\}$

$$f(x) = 0, \qquad \text{for all } x \in \partial A^c,$$

where $\Delta = P - I$ is the *discrete Laplace operator* (*diskreetti Laplace-operaattori*)[12] associated to transition matrix P. In some literature, the operator Δ is also termed the *drift matrix* (*virtausmatriisi*) of the Markov chain.

[12] Note that there are slightly different conventions in defining the discrete Laplace operator in the literature.

Example 4.2. Consider an undirected graph G with finite node set S such that the degree[13] $\deg(x)$ of every node x in G is at least 1. A random walk X on G is a Markov chain that proceeds by moving at each step to a neighboring node selected uniformly at random:

$$P(x,y) \;=\; \frac{\mathbb{1}(x\sim y)}{\deg(x)} \;=\; \begin{cases} \frac{1}{\deg(x)}, & \text{if } x \text{ and } y \text{ are neighbors,} \\ 0, & \text{otherwise,} \end{cases} \qquad x,y\in S,$$

where we write $x \sim y$ if x and y are neighbors, meaning that there is an (undirected) edge in the graph G between them.

Because $P(x,y) = 0$ if x and y are not neighbors, we see that the discrete Laplace operator $\Delta = P - I$ associated to this Markov chain is operating on functions $f : S \to \mathbb{R}$ by

$$\begin{aligned} (\Delta f)(x) \;&=\; \sum_{y\in S} (P(x,y) - \delta_x(y))\cdot f(y) \\ &=\; \sum_{y\in S} P(x,y)\cdot f(y) \;-\; f(x) \\ &=\; \frac{1}{\deg(x)} \sum_{\substack{y\in S \\ x\sim y}} (f(y) - f(x)), \qquad x\in S. \end{aligned}$$

This is the average difference of the values of f around x. ■

To prove Theorem 4.1, we will use a very useful technique for Markov chains, termed *first-step analysis*, that is, by conditioning on the possible states of the very first step, X_1. In words, by the Markov property (1.2), we may first investigate the very first transition $X_0 \to X_1$ of the Markov chain, and then the future after this will have similar dynamics, because it only depends on the state X_1 and not anymore on X_0. This idea will feature a lot in the sequel.

Proof of Theorem 4.1. [14] If the initial state $x \in A$, then we surely have $T_A = 0$, so that $k_A(x) = 0$. Thus, let us assume that $x \notin A$ and consider the first line in (4.2). Note that when $X_0 = x \notin A$,

$$\begin{aligned} T_A \;&=\; \min\{t \geq 1 : X_t \in A\} \;=\; 1 \;+\; \min\{s \geq 0 : X_{s+1} \in A\} && \text{[writing } s = t-1\text{]} \\ &=\; 1 \;+\; \min\{s \geq 0 : \hat{X}_s \in A\} && \text{[writing } \hat{X}_s = X_{s+1}\text{]} \end{aligned}$$

where we define $\hat{X}_s = X_{s+1}$ for $s = 0, 1, 2, \ldots$ the "future" of the Markov chain after the first step. Using the Markov property (1.2), we may regard the process $\hat{X} = (X_1, X_2, X_3, \ldots)$ as the same Markov chain but started at $X_1 = \hat{X}_0$, so that we can write

$$\mathbb{E}\,(T_A \mid X_1 = y,\, X_0 = x) \;=\; 1 \;+\; \mathbb{E}\,(T_A \mid \hat{X}_0 = y) \;=\; 1 + k_A(y).$$

By conditioning on the possible values of X_1, we now find that

$$\begin{aligned} k_A(x) \;&=\; \sum_{y\in S} \mathbb{E}\,(T_A \mid X_1 = y,\, X_0 = x)\cdot \mathbb{P}\,(X_1 = y \mid X_0 = x) \\ &=\; \sum_{y\in S} (1 + k_A(y))\cdot P(x,y). \end{aligned}$$

The claimed equations (4.2) follow from this after recalling that the row-sums of P equal one. □

[13] Here by *degree* we mean the outdegree $\deg(x) = \sum_z A(x,z)$ (each edge is oriented in two ways, and contributes to both the outdegree $\deg(x) = \sum_z A(x,z)$ and the indegree $\sum_z A(z,x)$, where A is the adjacency matrix of the graph defined in Equation (1.6).

[14] Theorem 4.1 can also be proven using a more general result, Theorem 4.6 in Section 4.3, which interested readeers can check out from Section 4.3 below.

In fact, k_A is the *smallest* nonnegative solution to the linear system (4.2). It can be found numerically as follows. First set $f_0(x) = 0$ for all $x \in S$, and then iteratively compute

$$f_{n+1}(x) = \begin{cases} 1 + \sum_{y \notin A} P(x,y) \cdot f_n(y), & x \notin A, \\ 0, & x \in A, \end{cases} \qquad n = 1, 2, 3, \ldots.$$

Then, it is possible to prove that $f_0, f_1, f_2, \ldots$ forms a nondecreasing sequence of functions with pointwise limit $f(x) = \lim_{n\to\infty} f_n(x)$. The limit f takes values in the extended number set $[0, \infty]$ and is the smallest nonnegative solution of (4.2) — so $f = k_A$. (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.)

Example 4.3 (*Human resource management*)**.** Kalvonvääntäjät Oyj is management consulting company which has 100 employees divided into three salary categories:

$$\text{'1'} = \text{'junior'}, \qquad \text{'2'} = \text{'senior'}, \qquad \text{and} \qquad \text{'3'} = \text{'partner'}.$$

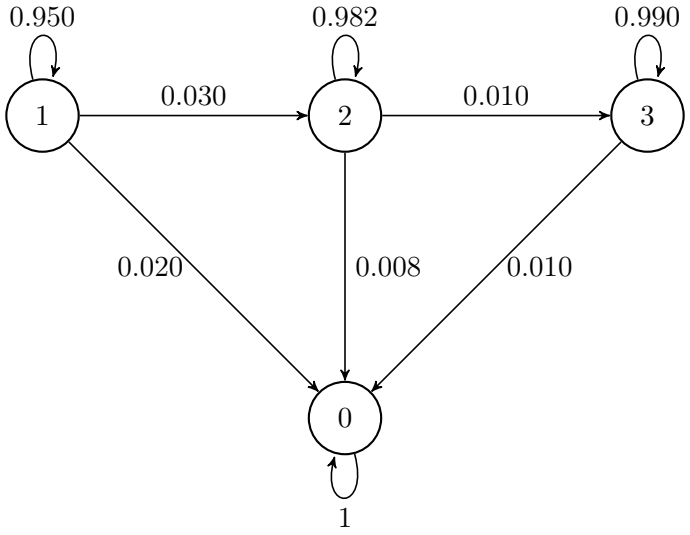
An employee holding a junior position in the beginning of a month gets promoted to senior with probability 0.030, leaves the company with probability 0.020, and otherwise continues in the same position in the beginning of next month. Similarly, a senior gets promoted to a partner with probability 0.010, 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 the absorbing state 0 means that the employee has left the company, and states $\{1, 2, 3\}$ are transient. The transition matrix and transition diagram are

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



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 $k_{\{0\}}(x) = k(x)$ satisfy the system of equations (4.2):

$$k(x) = 1 + \sum_{y=0}^{3} P(x,y) \cdot k(y) = 1 + \sum_{y=1}^{3} P(x,y) \cdot k(y), \qquad x \in \{1,2,3\},$$

with boundary condition $k(0) = 0$. Using the transition matrix in (4.3) we obtain the equations

$$\begin{cases} k(1) = 1 + 0.950 \cdot k(1) + 0.030 \cdot k(2), \\ k(2) = 1 + 0.982 \cdot k(2) + 0.010 \cdot k(3), \\ k(3) = 1 + 0.990 \cdot k(3). \end{cases}$$

These linear equations can be solved easily by first setting

$$\begin{aligned} & & k(3) &= \frac{1}{1-0.990} = 100, \\ \text{then} & & k(2) &= \frac{1+0.010 \cdot k(3)}{1-0.982} = 111.11, \\ \text{and finally} & & k(1) &= \frac{1+0.030 \cdot k(2)}{1-0.950} = 86.67. \end{aligned}$$

Hence we conclude that[15]

- ▷ 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.

■

4.2 Hitting probabilities

Consider Markov chain $X = (X_0, X_1, X_2, \ldots)$ on finite state space S. Select a non-empty set of states $A \subset S$. An irreducible Markov chain will surely visit every state, but a reducible chain might not. What is the probability that X starting at $X_0 = x$ eventually visits A?

Definition. The *hitting probability* (*osumatodennäköisyys*) by X of set A from initial state $X_0 = x$ is

$$h_A(x) = \mathbb{P}(X_t \in A \text{ for some } t \geq 0 \mid X_0 = x) = \mathbb{P}(T_A < \infty \mid X_0 = x). \tag{4.4}$$

Theorem 4.4 (*Hitting probabilities*). *For any (time-homogeneous) Markov chain with transition matrix P on finite state space S, the hitting probabilities into set $A \subset S$,*

$$\{h_A(x) : x \in S\},$$

satisfy the system of equations

$$h_A(x) = \begin{cases} \sum_{y \in S} P(x,y) \cdot h_A(y), & \text{if } x \notin A, \\ 1, & \text{if } x \in A. \end{cases} \tag{4.5}$$

[15] Similarly, the expected amount of time that a newly recruited employee has to wait before having been promoted to a partner equals $k_{\{3\}}(1)$. In this case, Theorem 4.1 gives $k_{\{3\}}(1) = \infty$. Does this make sense?

▷ The system of equations (4.5) can be interpreted in harmonic analytic terms as a Poisson equation, namely a discrete *Laplace equation* (*Laplace-yhtälö*) on the complement $A^c = S \setminus A$,

$$\Delta f(x) = 0, \qquad \text{for all } x \in A^c,$$

with boundary condition

$$f(x) = 1, \qquad \text{for all } x \in \partial A^c.$$

▷ In fact, h_A is the *smallest* nonnegative solution to the system (4.5).

The proof of Theorem 4.4 below uses, again, first-step analysis.

Proof of Theorem 4.4. [16] If the initial state $x \in A$, then X surely visits A, so that $h_A(x) = 1$. Thus, let us assume that $x \notin A$. By conditioning on the possible values of X_1, we now find that

$$\begin{aligned} h_A(x) &= \mathbb{P}(T_A < \infty \mid X_0 = x) \\ &= \sum_{y \in S} \mathbb{P}(T_A < \infty \mid X_1 = y,\, X_0 = x) \cdot \mathbb{P}(X_1 = y \mid X_0 = x) \\ &= \sum_{y \in S} \mathbb{P}(T_A < \infty \mid X_1 = y) \cdot P(x, y) \qquad \text{[by Markov property (1.2)]} \\ &= \sum_{y \in S} h_A(y) \cdot P(x, y), \end{aligned}$$

which is the asserted equation (4.5) for $x \notin A$. □

Example 4.5 (*Human resource management*)**.** Consider the Kalvonvääntäjät company described in Example 4.3. What is the probability that a freshly recruited new employee eventually becomes a partner in the company?

The answer is the hitting probability $h_A(1)$ of set $A = \{3\}$ from initial state $X_0 = 1$. According to Theorem 4.4, the hitting probabilities $h_{\{3\}}(x) = h(x)$ satisfy the system of equations (4.5):

$$h(x) = \sum_{y=0}^{3} P(x, y) \cdot h(y), \qquad x = 0, 1, 2,$$

with boundary condition $h(3) = 1$. Using the transition matrix in (4.3) we obtain the equations

$$\begin{cases} h(0) = h(0), \\ h(1) = 0.020 \cdot h(0) + 0.950 \cdot h(1) + 0.030 \cdot h(2), \\ h(2) = 0.008 \cdot h(0) + 0.982 \cdot h(2) + 0.010 \cdot h(3), \\ h(3) = 1. \end{cases}$$

Because there is no access from state 0 to state 3, we know that $h(0) = 0$. In light of this, we may solve the other equations to obtain

$$h = [0,\, 0.333,\, 0.556,\, 1].$$

We conclude that the probability that a freshly recruited junior eventually becomes a partner equals $h(1) = h_A(1) = 0.333$. Note that the entries of h do not sum into one, even though they are probabilities. (Not all vectors of probabilities represent probability distributions.) ■

[16] Theorem 4.4 can also be proven using a more general result, Theorem 4.6 in Section 4.3, which interested readeers can check out from Section 4.3 below.

4.3 General Poisson type equation for accumulated cost at passage time

Recall that the transition diagram of X is a directed graph with node set being the state space S and link set comprising the ordered node pairs (x, y) such that $P(x, y) > 0$. Assume that at each state y and at each transition (x, y), a *deterministic* cost (or profit) occurs:

$$\{c(y) : y \in S\}, \qquad \{c(x, y) : x, y \in S,\ P(x, y) > 0\},$$

where by convention we set $c(x, y) = 0$ when $P(x, y) = 0$. The *total cost accumulated at the passage time* into A is

$$W_A = \sum_{s=0}^{T_A} c(X_s) + \sum_{s=0}^{T_A - 1} c(X_s, X_{s+1}). \tag{4.6}$$

For the Markov chain started at initial state $X_0 = x$, the expected total cost is $\mathbb{E}(W_A \mid X_0 = x)$.

Theorem 4.6 (*Expected total cost*). *Consider a (time-homogeneous) Markov chain $X = (X_0, X_1, X_2, \ldots)$ with transition matrix P on finite state space S. Select a non-empty set of states $A \subset S$. The expected total cost accumulated at the passage time into A,*

$$w_A(x) = \mathbb{E}(W_A \mid X_0 = x), \tag{4.7}$$

satisfies the system of equations

$$w_A(x) = \begin{cases} c(x) + \sum_{y \in S} P(x, y) \cdot \big(c(x, y) + w_A(y)\big), & \text{if } x \notin A, \\ c(x), & \text{if } x \in A. \end{cases} \tag{4.8}$$

Proof. If the initial state $x \in A$, we already have $w_A(x) = c(x)$, since $T_A = 0$. Assume next that $x \notin A$. As before, the proof proceeds by applying first-step analysis, that is, by conditioning on the possible states of the first step X_1. By the Markov property (1.2), we have

$$\begin{aligned} \mathbb{E}(W_A \mid X_1 = y, X_0 = x) &= c(x) + c(x, y) + \mathbb{E}\left(\sum_{s=1}^{T_A} c(X_s) + \sum_{s=1}^{T_A - 1} c(X_s, X_{s+1}) \,\middle|\, X_1 = y, X_0 = x \right) \\ &= c(x) + c(x, y) + \mathbb{E}(W_A \mid X_0 = y) \\ &= c(x) + c(x, y) + w_A(y), \qquad y \in S. \end{aligned}$$

Hence, we obtain

$$\begin{aligned} \mathbb{E}(W_A \mid X_0 = x) &= \sum_{y \in S} \mathbb{E}(W_A \mid X_1 = y, X_0 = x) \cdot \mathbb{P}(X_1 = y \mid X_0 = x) \\ &= \sum_{y \in S} c(x) \cdot P(x, y) + \sum_{y \in S} \big(c(x, y) + w_A(y)\big) \cdot P(x, y) \\ &= c(x) + \sum_{y \in S} P(x, y) \cdot \big(c(x, y) + w_A(y)\big), \end{aligned}$$

which is the asserted equation (4.8) for $x \notin A$. □

We can now easily derive the Poisson equation (4.2) from the above more general result.

Proof of Theorem 4.1 using Theorem 4.6. If in Theorem 4.6, we choose the costs to be

$$\begin{cases} c(x) = 1, & x \notin A, \\ c(x) = 0, & x \in A, \end{cases} \qquad \text{and} \qquad c(x,y) = 0, \qquad x, y \in S,$$

then system (4.2) coincides with system (4.8). Hence, it remains to show that $k_A(x) = w_A(x)$, where w_A is given by (4.6, 4.7) as in Theorem 4.6 in Section 4.3.

By definition of the passage time T_A, we have $X_{T_A} \in A$ and $X_s \notin A$ for all $0 \le s \le T_A - 1$, so

$$k_A(x) = \mathbb{E}\,(T_A \mid X_0 = x) = \mathbb{E}\left(\sum_{s=0}^{T_A - 1} 1 \mid X_0 = x \right) = \mathbb{E}\left(\sum_{s=0}^{T_A} c(X_s) \mid X_0 = x \right) = w_A(x),$$

using the notation (4.7) from Section 4.3. Theorem 4.6 then yields the asserted system (4.2). □

Similarly, we can easily derive the Laplace equation (4.5).

Proof of Theorem 4.4 using Theorem 4.6. If in Theorem 4.6, we choose the costs to be

$$\begin{cases} c(x) = 1, & x \in A, \\ c(x) = 0, & x \notin A, \end{cases} \qquad \text{and} \qquad c(x,y) = 0, \qquad x, y \in S,$$

then system (4.5) coincides with system (4.8). Hence, it remains to show that $h_A(x) = w_A(x)$, where w_A is given by (4.6, 4.7) as in Theorem 4.6 in Section 4.3.

By definition of the passage time T_A, we have $X_{T_A} \in A$ and $X_s \notin A$ for all $0 \le s \le T_A - 1$, so

$$W_A = \sum_{s=0}^{T_A} c(X_s) = \begin{cases} c(X_{T_A}), & T_A < \infty, \\ \sum_{s=0}^{\infty} 0, & T_A = \infty, \end{cases} = \begin{cases} 1, & T_A < \infty, \\ 0, & T_A = \infty, \end{cases}$$

$$= \mathbb{1}\{T_A < \infty\},$$

using the notation (4.7) from Section 4.3. Now, its expected value is

$$w_A(x) = \mathbb{E}\,(W_A \mid X_0 = x) = \mathbb{E}\,(\mathbb{1}\{T_A < \infty\} \mid X_0 = x) = \mathbb{P}\,(T_A < \infty \mid X_0 = x) = h_A(x).$$

Theorem 4.6 then yields the asserted system (4.5). □

4.4 Random walk on finite state space and gambler's ruin

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

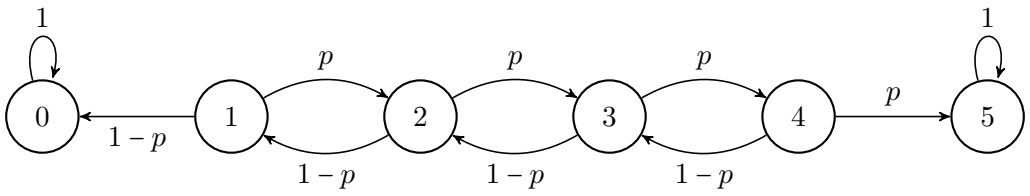


Figure 4.1: Transition diagram of a random walk with $M = 5$.

In gambling context, the associated Markov chain $X = (X_0, X_1, X_2, \ldots)$ represents the wealth X_t of a gambler after t rounds in a game where the gambler wins 1 EUR with probability p and loses 1 EUR with probability $1 - p$, and where all game rounds are independent of each other. The game ends[17] if the wealth X_t hits the value M (gambler's target) or the value 0 (gambler's money is all gone, and they are ruined).

A basic question here is to determine the probability of the gambler hitting the target, given that the initial wealth equals a given value x. That is, we wish to compute the probability

$$h(x) \;=\; \mathbb{P}\,(X_t \;=\; M \text{ for some } t \geq 0 \mid X_0 = x). \tag{4.9}$$

To avoid trivialities, we assume that $p \in (0,1)$. Then, the Markov chain surely eventually hits either 0 or M, and we see that the probability of the gambler's eventual ruin equals

$$1 - h(x).$$

The probability $h(x)$ in (4.9) equals the hitting probability $h_A(x)$ in (4.4) for the singleton set $A = \{M\}$. Hence, by Theorem 4.4 the function $h(x)$ solves the system of equations (4.5), which in the present case take the form

$$\begin{cases} h(0) \;=\; h(0), \\ h(x) \;=\; (1-p)\cdot h(x-1) \;+\; p\cdot h(x+1), \qquad 0 < x < M, \\ h(M) \;=\; 1. \end{cases}$$

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 solution to the equation

$$h(x) \;=\; (1-p)\cdot h(x-1) \;+\; p\cdot h(x+1), \qquad 0 < x < M, \tag{4.10}$$

with boundary conditions $h(0) = 0$ and $h(M) = 1$. There are various ways to solve these equations: one can use a computer (extra exercise), an ansatz (that we make below), or one can apply generating functions (see Section 6).

Reminder. Consider a homogeneous n:th order linear difference equation

$$f(x+n) \;=\; a_1\, f(x+n-1) \;+\; a_2\, f(x+n-2) \;+\; \cdots \;+\; a_{n-1}\, f(x+1) \;+\; a_n\, f(x). \tag{4.11}$$

The *characteristic equation* is the polynomial equation

$$p(z) \;=\; z^n \;-\; a_1\, z^{n-1} \;-\; a_2\, z^{n-2} \;-\; \cdots \;-\; a_{n-1}\, z \;-\; a_n \;=\; 0.$$

If the roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the characteristic polynomial $p(z)$ are all *distinct*, then all solutions of (4.11) have the form

$$f(x) \;=\; c_1\, \lambda_1^x \;+\; \cdots \;+\; c_n\, \lambda_n^x,$$

where the coefficients $c_1, \ldots, c_n$ can be determined from boundary conditions. ♣

Let us first solve $h(x)$ in the *asymmetric* case where $p \in (0,1)$ is such that $p \neq 1/2$. Formula (4.10) is a *second-order homogeneous linear difference equation*, for which it is useful to make the ansatz $h(x) = z^x$ for some real number $z > 0$. Substituting this into (4.10) leads to

$$z^x \;=\; (1-p)\cdot z^{x-1} \;+\; p\cdot z^{x+1},$$

[17] The expected hitting time will be computed in Example 4.9.

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

$$pz^2 - z + (1-p) = 0,$$

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

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

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

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

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

$$h(x) = \frac{\alpha^x - 1}{\alpha^M - 1}. \tag{4.12}$$

To obtain the solution of (4.10) in the symmetric case with $p = 1/2$, we may inspect how the solution of (4.12) behaves as a function of p as $p \to 1/2$. In this case, we have

$$\alpha = \frac{1-p}{p} \quad \longrightarrow \quad 1, \qquad \text{as } p \to 1/2.$$

and by l'Hôpital's rule, it follows that

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

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 4.7 (*Gambler's ruin*). *The probability that a random walk on $\{0, 1, \ldots, M\}$ described in Figure 4.1 with $p \in (0,1)$ started at x eventually hits M equals*

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

The main message of Theorem 4.7 is that when $p \leq 1/2$, the probability of ever reaching a state M from any initial state x tends to zero as $M \to \infty$ (a greedy player desiring an infinite profit). In other words, the probability of eventual ruin $1 - h(x)$ tends to one. See Figure 4.2.

Example 4.8 (*Gambling: doubling strategy*). 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 $p = 18/37$, and the probability of losing 1 EUR is $1 - p$. If a gambler targets to double their initial wealth x, then the probability $h(x)$ of successfully ending the game is obtained by applying Theorem 4.7 with $M = 2x$: ■

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

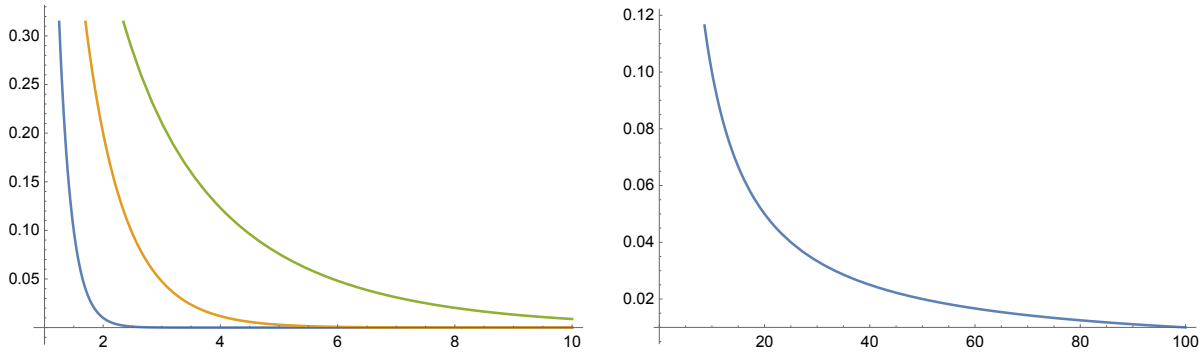


Figure 4.2: For random walk started at $x = 1$, plotting the "hitting the target" probability $h(1) = \left(\left(\frac{1-p}{p}\right) - 1\right)\left(\left(\frac{1-p}{p}\right)^M - 1\right)^{-1}$ of Theorem 4.7 as a function of M for some values of $p < 1/2$ (left) and $p = 1/2$ (right). As $M \to \infty$, we see that it tends to zero.

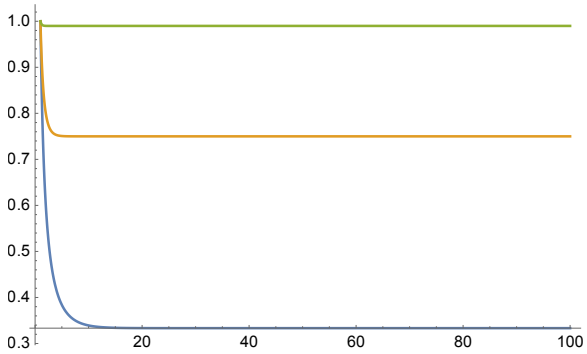


Figure 4.3: Plotting $h(1) = \left(\left(\frac{1-p}{p}\right) - 1\right)\left(\left(\frac{1-p}{p}\right)^M - 1\right)^{-1}$ of Theorem 4.7 as a function of M for some values of $p > 1/2$. As $M \to \infty$, we see that it tends to a non-zero value.

Example 4.9 (*Expected duration of the game*)**.** Consider the expected passage time

$$k(x) \;=\; \mathbb{E}\,(T_{\{0,M\}} \mid X_0 = x)$$

to $A = \{0, M\}$ by a random walk on $\{0, 1, \ldots, M\}$ described in Figure 4.1 with $p \in (0, 1)$ started at $X_0 = x$. Theorem 4.1 gives the system of equations (4.2) for $k(x)$, which take the form

$$\begin{cases} k(0) \;=\; 0, \\ k(x) \;=\; 1 \;+\; (1-p)\cdot k(x-1) \;+\; p\cdot k(x+1), & 0 < x < M, \\ k(M) \;=\; 0. \end{cases}$$

This is an *inhomogeneous second order linear difference equation.* E.g. using a computer, we can solve in the asymmetric case

$$k(x) \;=\; \left(\frac{x}{1-2p}\right) - \left(\frac{\left(\frac{1-p}{p}\right)^x - 1}{\left(\frac{1-p}{p}\right)^M - 1}\right)\left(\frac{M}{1-2p}\right), \qquad p \;\neq\; 1/2,$$

and in the symmetric case

$$k(x) \;=\; (M-x)\,x, \qquad p \;=\; 1/2.$$

We see that when $p \geq 1/2$, the expected passage time $k(x) \to \infty$ as $M \to \infty$, while for $p < 1/2$, the limit is finite: $k(x) \to \frac{x}{1-2p}$ as $M \to \infty$. We will get back to this in Section 5. ■

5 Markov chains and random walks in countably infinite spaces

5.1 Basic definitions: generalization of finite state spaces

We will now study stochastic discrete-time processes with values in a general countable (finite or countably infinite) state space S. The assumption that S is *countable* (*numeroituva*) means that its elements can be numbered using positive integers according to $S = \{x_1, x_2, \ldots\}$, 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 $\mathbb{Z}$ of integers, the set $\mathbb{Q}$ of rational numbers, and the set $\mathbb{N}$ of natural numbers.
- ▷ The set $\mathbb{Z}^d$ of vectors $(x_1, \ldots, x_d)$ with integer coordinates $x_j \in \mathbb{Z}$ for all j.
- ▷ The set of finite strings composed of letters from a finite alphabet.

The following sets can be shown to be uncountably infinite:

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

■

The *sum* (*summa*) of a nonnegative function $f : S \to [0, \infty)$ on a countably infinite space $S = \{x_1, x_2, \ldots\}$ is defined[18] by

$$\sum_{x \in S} f(x) = \sum_{j=1}^{\infty} f(x_j) = \lim_{n \to \infty} \sum_{j=1}^{n} f(x_j).$$

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

Definition.

- ▷ A *probability distribution* (*todennäköisyysjakauma*) on S is a function $\mu : S \to [0, 1]$ such that the law of total probability holds:

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

  In the context of Markov chains, we interpret probability distribution μ as a (possibly infinite) row-vector indexed by the states $x \in S$.

- ▷ A *transition matrix* (*siirtymämatriisi*) on S is a function $P : S \times S \to [0, 1]$ such that

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

  which means that each row sum of the (infinite) square matrix P equals 1.

[18] The sum always exists because $\sum_{j=1}^{n} f(x_j)$ is non-decreasing in n, but it may be infinite: $\sum_{j=1}^{\infty} f(x_j) \in [0, \infty]$.

Matrix multiplication with countably infinite matrices is defined similarly as in the finite case.

▷ If μ is a probability distribution on S, we define $\mu \cdot P$ by the formula

$$(\mu \cdot P)(y) \;=\; \sum_{x \in S} \mu(x) \cdot P(x,y), \qquad \text{for all } y \in S.$$

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

$$R(x,z) \;=\; \sum_{y \in S} P(x,y) \cdot Q(y,z), \qquad \text{for all } x, z \in S.$$

▷ Matrix powers are defined in the usual way as $P^0 = I$ and recursively $P^{t+1} = P^t \cdot P$ for $t \in \mathbb{N}_0$, where the identity matrix $I : S \times S \to [0,1]$ is given by

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

Lemma 5.2. *If μ is a probability distribution and P, Q are transition matrices on S, then*

▷ *$\mu \cdot P$ is a probability distribution on S,*

▷ *$R = P \cdot Q$ is a transition matrix on S,*

▷ *P^t is a transition matrix on S for all $t \in \mathbb{N}_0$.*

Proof. Clearly $(\mu \cdot P)(x) \geq 0$ for all $x \in S$. Moreover, by changing the order of summation[19], we see that the vector $\mu \cdot P = \{(\mu \cdot P)(x) : x \in S\}$ satisfies

$$\sum_{y \in S} (\mu \cdot P)(y) \;=\; \sum_{y \in S} \sum_{x \in S} \mu(x) \cdot P(x,y) \;=\; \sum_{x \in S} \mu(x) \Big(\sum_{y \in S} P(x,y) \Big) \;=\; 1.$$

Hence, $\mu \cdot P$ is a probability distribution on S.

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

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

Hence, $R = P \cdot Q$ is a transition matrix on S. Lastly, one can use this and mathematical induction to prove that P^t is a transition matrix on S for all $t \in \mathbb{N}_0$. □

Markov chains on countably infinite state spaces are defined precisely in the same way as in Section 1. The only difference is that the transition matrix P may have infinitely many rows and columns. We can view the infinite transition matrix as a function $P : S \times S \to [0,1]$ which maps a pair of states (x,y) into the probability

$$P(x,y) = \mathbb{P}\,(X_{t+1} = y \mid X_t = x), \qquad \text{for each } t \in \mathbb{N}_0.$$

(The proof of Theorem 1.1 can also be adjusted to the present setting.)

[19] Allowed when the terms are nonnegative by so-called Fubini-Tonelli theorem [Kyt20, Thm. IX.9(a)].

Definition. An S-valued stochastic process (random sequence) $X = (X_0, X_1, X_2, \ldots)$ is a (time-homogeneous) *Markov chain* (*Markov-ketju*) with state space S and transition matrix P if X is "conditionally independent of the past", i.e.,

$$\mathbb{P}(X_{t+1} = y \mid X_t = x,\, H_{t-}) \;=\; P(x, y), \tag{5.2}$$

for all states $x, y \in S$, all times $t \geq 0$, and for all events $H_{t-} = \{X_0 = x_0, \ldots, X_{t-1} = x_{t-1}\}$ such that $\mathbb{P}(X_t = x,\, H_{t-}) > 0$.

Theorem 5.3 (*Time-dependent distribution*)**.** *The distribution*

$$\mu_t(x) \;=\; \mathbb{P}(X_t = x), \qquad x \in S,$$

of any (time-homogeneous) Markov chain $X = (X_0, X_1, X_2, \ldots)$ at an arbitrary time instant $t \geq 0$ can be computed from the initial distribution μ_0 using the formula

$$\mu_t \;=\; \mu_0 \cdot P^t, \tag{5.3}$$

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

The probability that X moves from state x to state y during t time steps equals

$$\mathbb{P}(X_t = y \mid X_0 = x) \;=\; P^t(x, y), \qquad \text{for all } x, y \in S. \tag{5.4}$$

Proof. After inspecting them, we can see that the proofs of Theorems 1.5 and 1.7 work also for countably infinite state spaces. This implies (5.3) and (5.4). □

In Chapter 2, we investigated long-term behavior of finite-state Markov chains. Let us return to these questions in the case of countably infinite state spaces:

1. Do the time-dependent distributions μ_t have a limit as $t \to \infty$ (in a sense detailed below)? Does such a limit, if exists, depend on the initial distribution μ_0?
2. Does an invariant distribution π satisfying $\pi \cdot P = \pi$ exist, and is it unique?

As we shall see in this chapter, the answers to these questions are more delicate in the case of countably infinite state spaces.

Let us begin with the basic definitions, which are similar to the case of finite state spaces.

Definition. Starting from a given initial distribution μ_0, if the limit

$$\lim_{t \to \infty} \mu_t(x) \;=\; \mu_\infty(x), \qquad \text{for all } x \in S,$$

exists and defines a probability distribution, then we say that μ_∞ is the *limiting distribution* (*rajajakauma*) of the Markov chain started at the initial distribution μ_0.

Definition. A probability distribution $\pi = (\pi(x) : x \in S)$ is called an *invariant distribution* (*tasapainojakauma*) of a transition matrix P and the corresponding Markov chain $X = (X_0, X_1, X_2, \ldots)$ if it satisfies the *balance equations* (*tasapainoyhtälöt*)

$$\pi \cdot P = \pi.$$

Such a distribution π is also called a *stationary distribution* (*stationaarinen jakauma*).

▷ We learned in Theorem 2.2 that any limiting distribution μ_∞ is also an *invariant* distribution. (The same proof works also for countably infinite state spaces, remembering that we require μ_∞ to be a probability distribution.)

▷ In particular, starting from an invariant initial distribution $\mu_0 = \pi$, we find using Theorem 5.3 and the associativity of matrix multiplication that

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

Hence, for a Markov chain with a random initial state $X_0 \sim \pi$ distributed according to an invariant distribution, *the distribution of X_t remains invariant for all time instants t.*

5.2 Invariant distribution, recurrence, and irreducibility

The long-term analysis of Markov chains on infinite state spaces has one fundamental difference compared to Markov chains on finite spaces:

an invariant distribution might not exist,

and neither does irreducibility guarantee the existence of an invariant distribution. (Irreducibility only guarantees uniqueness, see Theorem 5.5.) The problem is that, while one can always solve for a left eigenvector of the transition matrix P with eigenvalue 1 (as in the proof of Theorem 2.1), one still might not be able to normalize it to give a *probability distribution*.

Every irreducible Markov chain on a finite state space visits all states infinitely often with probability one. In infinite spaces, this may or may not be the case. To study this, a key quantity is the *positive passage time* (*positiivinen kulkuaika*) of X into set $A \subset S$,

$$T_A^+ = \min\{t \geq 1 : X_t \in A\}, \tag{5.5}$$

with the notational convention that $T_A = \infty$ if the process X never visits A after starting[20]. If the set $A = \{y\}$ is a singleton, we simply write $T_{\{y\}}^+ = T_y^+$.

An invariant distribution can also be sought by noticing that

$$\tilde{\pi}(x) = \frac{1}{\mathbb{E}\,(T_x^+ \mid X_0 = x)} \tag{5.6}$$

satisfies the balance equations (2.1):

$$\tilde{\pi} \cdot P = \tilde{\pi}.$$

In finite state spaces one can normalize $\tilde{\pi}$ to find a probability distribution π satisfying the system of equations (2.1, 2.3). However, the normalization (2.3) might not be possible in general (see the random walk in Section 5.6 for an example): it is possible that $\sum_{x \in S} \tilde{\pi}(x) = \infty$.

[20] Note that T_A^+ differs from the passage time T_A in (4.1) in the sense that $T_A = 0$ if the Markov chain already starts from A, while T_A^+ is always nonzero.

Definition. For two states $x, y \in S$, the *visiting probability* (*vierailutodennäköisyys*) of X to y after started at x is denoted as

$$\rho(x,y) \;=\; \mathbb{P}\left(X_t = y \text{ for some } t \geq 1 \mid X_0 = x\right) \;=\; \mathbb{P}\left(T_y^+ < \infty \mid X_0 = x\right).$$

In particular, the *return probability* (*paluutodennäköisyys*) of X to x is

$$\rho(x,x) \;=\; \mathbb{P}\left(X_t = x \text{ for some } t \geq 1 \mid X_0 = x\right) \;=\; \mathbb{P}\left(T_x^+ < \infty \mid X_0 = x\right).$$

State x is called

- ▷ *recurrent* (*palautuva*) if it has return probability $\rho(x,x) = 1$, and
- ▷ *transient* (*väistyvä*) otherwise.

Recall that, given a transition matrix P, we denote $x \rightsquigarrow y$ if the corresponding transition diagram contains a (directed) path from x to y. We define the *components* (*komponentit*) of the corresponding Markov chain X to be the equivalence classes

$$C(x) = \{y \in S : y \leftrightsquigarrow x\}, \qquad x \in S,$$

where $x \leftrightsquigarrow y$ if and only if $x \rightsquigarrow y$ and $y \rightsquigarrow x$. As before, we say that transition matrix P and the corresponding Markov chain X are *irreducible* (*redusoitumaton*) if there is only one component.

We use the next result to prove the Markov Chain Convergence Theorem in the next section.

Theorem 5.4. *If an* irreducible *Markov chain on a countable state space S has an invariant distribution π, then*

1. *all states have positive mass:*

$$\pi(y) \;>\; 0, \qquad \text{for all } y \in S, \tag{5.7}$$

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

For mathematically oriented readers, we prove Theorem 5.4 in Section 5.7.

Theorem 5.5 (*Uniqueness of invariant distribution*)**.** *Every* irreducible *Markov chain on a countable state space admits* at most one *invariant distribution π.*

Proof. The Convergence Theorem 5.6 (discussed below) gives the claim if the Markov chain is irreducible and aperiodic. If the Markov chain is periodic, we can modify it to become aperiodic by considering the transition matrix $\tilde{P} = \frac{1}{2}(P+I)$. Note that $\pi \cdot P = \pi$ is equivalent to $\pi \cdot \tilde{P} = \pi$. □

5.3 Long-term behavior: Convergence Theorem

A key result in the theory of Markov chains is the Convergence Theorem 5.6[21].

[21] It is a generalization of Theorem 2.14.

Theorem 5.6 (*Convergence Theorem*). *Consider an* irreducible *and* aperiodic *Markov chain X on a countable state space S. Then, X admits* at most one *invariant distribution. Moreover, if the invariant distribution π exists, then it also equals the limiting distribution,*

$$\lim_{t\to\infty} \mathbb{P}(X_t = y \mid X_0 = x) \;=\; \pi(y), \qquad \text{for all } x, y \in S,$$

which is independent of the initial state of the Markov chain X.

The proof relies on a very useful general technique known as *stochastic coupling* [LPW08, Section 5]. Note that if $|S| < \infty$, then the invariant distribution π always exists (Theorem 2.1). One issue with infinite state spaces is that the Markov chain can "escape" to infinity and fail to have an invariant distribution. See the random walk example in Section 5.6 and Figure 5.1.

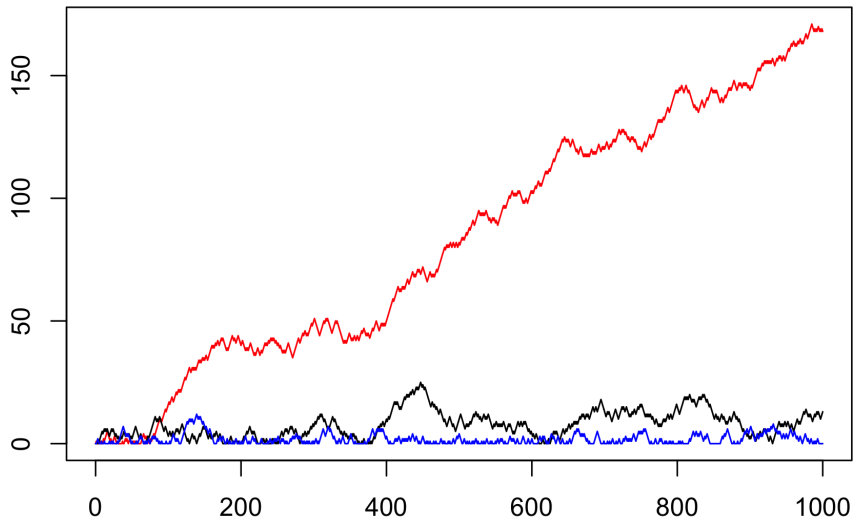


Figure 5.1: Simulated paths of the random walk on nonnegative integers $\mathbb{N}_0$ defined by transition matrix (5.16) for $p = 0.4$ (blue), $p = 0.5$ (black), $p = 0.6$ (red). Can you observe different behavior depending on whether $p < 1/2$ or $p > 1/2$? We discuss this in Section 5.6.

Proof sketch. Let X and Y be independent copies of our Markov chain, both having transition matrix P, and such that X has initial distribution μ_0 and Y has some initial distribution. Let

$$\tau \;=\; \min\{t \geq 0 : X_t = Y_t\} \tag{5.8}$$

be the first time instant (possibly ∞) at which the paths of the Markov chains meet each other. Observe next that, for any $s \leq t$, we have the symmetric relations

$$\begin{aligned}
\mathbb{P}(X_t = y, \tau = s) \;&=\; \sum_{x\in S} \mathbb{P}(\tau = s, X_s = x, X_t = y) \\
&=\; \sum_{x\in S} \mathbb{P}(X_t = y \mid \tau = s, X_s = x) \cdot \mathbb{P}(\tau = s, X_s = x) \\
&=\; \sum_{x\in S} \mathbb{P}(Y_t = y \mid \tau = s, Y_s = x) \cdot \mathbb{P}(\tau = s, Y_s = x) \qquad \text{[by (5.8)]} \\
&=\; \mathbb{P}(Y_t = y, \tau = s), \qquad y \sin S.
\end{aligned}$$

By summing this over $s \leq t$, it follows that

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

which implies that *we can bound the differences of the time-dependent probabilities of the chains X and Y by the tail probability of their meeting time τ*:

$$\begin{aligned}\sum_{y\in S}\big|\mathbb{P}(X_t=y)-\mathbb{P}(Y_t=y)\big| &= \sum_{y\in S}\big|\mathbb{P}(X_t=y,\tau>t)-\mathbb{P}(Y_t=y,\tau>t)\big|\\ &\le \sum_{y\in S}\mathbb{P}(X_t=y,\tau>t)+\sum_{y\in S}\mathbb{P}(Y_t=y,\tau>t)\\ &= 2\mathbb{P}(\tau>t). \end{aligned} \tag{5.9}$$

(Note that up to this point of the proof, we have not used any invariant distribution π yet.)

Next, when X is started at a deterministic point x and Y is started at a random initial state distributed according to an invariant distribution π, the upper bound (5.9) becomes

$$\sum_{y\in S}\big|\,\mathbb{P}(X_t=y\mid X_0=x)\;-\;\pi(y)\,\big| \;=\; \sum_{y\in S}\big|P^t(x,y)-\pi(y)\big| \;\le\; 2\mathbb{P}(\tau>t). \tag{5.10}$$

To finish the proof, it suffices to show that the upper bound $\mathbb{P}(\tau>t)\to 0$ as $t\to\infty$, which is equivalent to showing that $\mathbb{P}(\tau<\infty)=1$. For this purpose, we use the Markov chain $\{(X_t,Y_t): t\ge 0\}$ on the product space $S\times S$, with transition matrix $\tilde{P}$ defined by

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

It is straightforward to verify that $\tilde{\pi}(x,y)=\pi(x)\cdot\pi(y)$ 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 — think of what can happen in the periodic case!). In terms of the product Markov chain (X,Y), we see that τ is the first passage time T_D of the product chain into the diagonal set $D=\{(x,y)\in S\times S: x=y\}$, which is clearly bounded from above by $T_D\le T_{(x,x)}$ for any $x\in S$. By Theorem 5.4, we know that $T_{(x,x)}$ is finite with probability one, and hence so is $\tau=T_D$. □

Remark. The idea of the proof of Theorem 5.6 can in certain cases be used also to argue that the invariant (and limiting) distribution exists. Indeed, by bounding the so-called *total variation distance* (defined in Equation (12.12) in Chapter 12) of the two copies of the Markov chains started at two arbitrary initial distributions by the tail probability of their meeting time τ, one can derive the existence of the limit. This requires, however, strong enough control of the meeting time τ. Students majoring in mathematics are recommended to look into the lecture notes [LPW08] for more details about the coupling method.

5.4 Reversibility and detailed balance equations

Recall that in the case of finite state spaces, it was possible to find an invariant distribution π by solving the balance equations (2.1) under the normalization condition (2.3). For infinite state spaces, these equations (2.1, 2.3) become an infinite system of linear equations and do not guarantee that a solution exists. It is sometimes handy to consider another infinite set of linear equations — the *detailed balance equations* (5.11) given below.

Definition. A transition matrix P and the corresponding Markov chain X is called *reversible* (*kääntyvä*) with respect to a probability distribution π (π-reversible) if the following *detailed balance equations* (*pareittaiset tasapainoyhtälöt*) hold:

$$\pi(x)\cdot P(x,y) \;=\; \pi(y)\cdot P(y,x), \qquad \text{for all } x,y\in S. \tag{5.11}$$

Note that reversibility also includes the condition that π is a probability distribution, i.e.,

$$\sum_{x\in S}\pi(x) \;=\; 1. \tag{5.12}$$

The detailed balance equations (5.11) are a stronger condition than the balance equations (2.1). Indeed, the next result shows that the former implies the latter.

Theorem 5.7 (*Reversibility guarantees existence of invariant distribution*)**.** *If transition matrix P and the corresponding Markov chain X is* π-reversible *with respect to a probability distribution π, then π is an invariant distribution of P and X.*

Proof. If the detailed balance equations (5.11) hold, then, for all $y \in S$, we have

$$\begin{aligned}(\pi \cdot P)(y) &= \sum_{x \in S} \pi(x) \cdot P(x,y) \\ &= \sum_{x \in S} \pi(y) \cdot P(y,x) = \pi(y) \sum_{x \in S} P(y,x) = \pi(y), \qquad \text{for all } y \in S.\end{aligned}$$

Hence $\pi \cdot P = \pi$, which shows that π is an invariant distribution of P and X. □

Corollary 5.8 (*Summary of irreducible/reversible/aperiodic cases*)**.** *Suppose that Markov chain X with transition matrix P on countable state space S is* irreducible *and* π-reversible *for a probability distribution π. Then, π is the* unique *invariant distribution of X. It can be determined as the unique solution π to detailed balance equations* (5.11):

$$\pi(x) \cdot P(x,y) = \pi(y) \cdot P(y,x), \qquad \text{for all } x, y \in S,$$

and the normalization (5.12):

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

Moreover, in the aperiodic *case, π also equals the unique limiting distribution,*

$$\lim_{t \to \infty} \mathbb{P}(X_t = y \mid X_0 = x) = \pi(y), \qquad \text{for all } x, y \in S.$$

Proof. This is an immediate consequence of Theorems 5.6, 5.5, and 5.7. □

Reversibility can be interpreted as follows. Consider a Markov chain X with transition matrix P which is π-reversible. Pick the initial distribution μ_0 of X to be π:

$$X_0 \sim \pi, \qquad \text{that is,} \quad \mathbb{P}[X_0 = x] = \pi(x), \qquad \text{for all } x \in S.$$

Since by Theorem 5.7, π is an invariant distribution of P and X, we have

$$X_t \sim \pi, \qquad \text{that is,} \quad \mathbb{P}[X_t = x] = \pi(x), \qquad \text{for all } x \in S \text{ and for all } t \geq 0.$$

By applying the detailed balance equations (5.11), we find that

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

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

5.5 Birth–death chains

An important class of reversible Markov chains is discussed next.

Definition. A *birth–death chain* (*syntymiskuolemisketju*) is a Markov chain on a state space $S \subset \mathbb{N}_0$ with a transition matrix P such that

$$P(x, y) = 0 \qquad \text{for } |x - y| > 1.$$

Examples of birth–death chains include *random walk on finite state space* discussed in Section 4.4, and *random walk on* $\mathbb{N}_0$, discussed in Section 5.6. Importantly, a birth–death chain can only move to its nearby states, which implies that the flow of mass from each state can only go to two directions. As a consequence, birth–death chains are reversible whenever they admit an invariant distribution, as the following result, Theorem 5.9, shows.

Theorem 5.9 (*Invariant distribution for birth–death chain*)**.** *If a birth–death chain has an invariant distribution π, then the chain is π-reversible.*

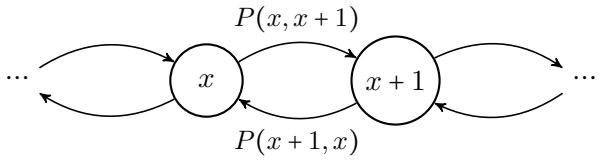
The key idea to prove Theorem 5.9 is the conservation-of-mass property encoded into the balance equations, and refined by the detailed balance equations (5.11) — see Lemma 5.10.

Proof. We have to show that the balance equations (5.11) hold for π. Fix a state $s \in S$.

- ▷ If $x = y$, then (5.11) is trivially true.
- ▷ When $|x - y| > 1$, both sides of (5.11) are zero, so the detailed balance is also trivially true.
- ▷ Hence, the only case that we need to investigate is the one where we assume that $x, y \in S$ are such that $y = x + 1$. (The case $y = x - 1$ is symmetric to this.) We want to prove that

$$\pi(x) \cdot P(x, x+1) = \pi(x+1) \cdot P(x+1, x). \tag{5.13}$$

The key observation is that for any birth–death chain, the sets $A_x = \{v \in S : v \leq x\}$ and $S \setminus A_x = \{u \in S : u \geq x+1\}$ are connected in the transition diagram only via transitions between the states x and $x+1$. By the conservation of probability mass, the flow of mass from x to $x+1$ equals the flow of mass from $x+1$ to x (see Lemma 5.10 for a general version of this property).



Recall that we assumed that π is an invariant distribution. The balance equation $\pi = \pi \cdot P$ reads

$$\pi(v) = \sum_{u \in S} \pi(u) \cdot P(u, v), \qquad v \in S,$$

and by summing over $v \in A_x$, we find that[22]

$$\begin{aligned}\sum_{v\in A_x}\pi(v) &= \sum_{u\in S}\pi(u)\sum_{v\in A_x}P(u,v)\\ &= \sum_{u\in S\setminus A_x}\pi(u)\sum_{v\in A_x}P(u,v) + \sum_{u\in A_x}\pi(u)\sum_{v\in A_x}P(u,v),\end{aligned}$$

which after rearranging and relabeling indices gives

$$\begin{aligned}&\sum_{z\in A_x}\pi(z)\underbrace{\Big(1-\sum_{w\in A_x}P(z,w)\Big)}_{=\sum_{u\in S\setminus A_x}P(z,u)} = \sum_{u\in S\setminus A_x}\pi(u)\sum_{z\in A_x}P(u,z)\\ \implies\quad &\sum_{z\in A_x}\sum_{u\in S\setminus A_x}\pi(z)\cdot P(z,u) = \sum_{u\in S\setminus A_x}\sum_{z\in A_x}\pi(u)\cdot P(u,z) \qquad (5.14)\end{aligned}$$

Now, recall that $P(a,b)$ can only be nonzero when $|b-a|\leq 1$. The only pair $(z,u)\in A_x\times S\setminus A_x$ in (5.14) satisfying $|z-u|\leq 1$ is $z=x$ and $u=x+1$, which gives the only possibly nonzero terms:

$$(5.14)\qquad\implies\qquad \pi(x)\cdot P(x,x+1) = \pi(x+1)\cdot P(x+1,x).$$

This shows the desired equation (5.13). □

The idea of the proof above can be cast into the following useful more general result, concerning the probability mass flow at a set A. In essence, in statistical equilibrium, the flow of probability mass away from A equals the flow of probability mass into A. As such, this generalizes Equation (2.4), which covers the case where A consists only of one state, $A=\{y\}$.

Lemma 5.10 (*Extended balance equations*)**.** *Let $X=(X_0,X_1,X_2,\ldots)$ be a Markov chain on a (finite or countably infinite) state space S with transition matrix P. If π is an invariant distribution for X, then for any subset $A\subset S$ of states, we have*

$$\sum_{x\in A}\sum_{y\in S\setminus A}\pi(x)\cdot P(x,y) = \sum_{y\in S\setminus A}\sum_{x\in A}\pi(y)\cdot P(y,x). \qquad (5.15)$$

Proof. By summing the balance equation $\pi=\pi\cdot P$ over $v\in A$, we find that

$$\begin{aligned}\sum_{v\in A}\pi(v) &= \sum_{u\in S}\pi(u)\sum_{v\in A}P(u,v)\\ &= \sum_{u\in S\setminus A}\pi(u)\sum_{v\in A}P(u,v) + \sum_{u\in A}\pi(u)\sum_{v\in A}P(u,v),\end{aligned}$$

which after rearranging and relabeling indices gives

$$\begin{aligned}&\sum_{x\in A}\pi(x)\underbrace{\Big(1-\sum_{w\in A}P(x,w)\Big)}_{=\sum_{y\in S\setminus A}P(x,y)} = \sum_{y\in S\setminus A}\pi(y)\sum_{x\in A}P(y,x)\\ \implies\quad &\sum_{x\in A}\sum_{y\in S\setminus A}\pi(x)\cdot P(x,y) = \sum_{y\in S\setminus A}\sum_{x\in A}\pi(y)\cdot P(y,x),\end{aligned}$$

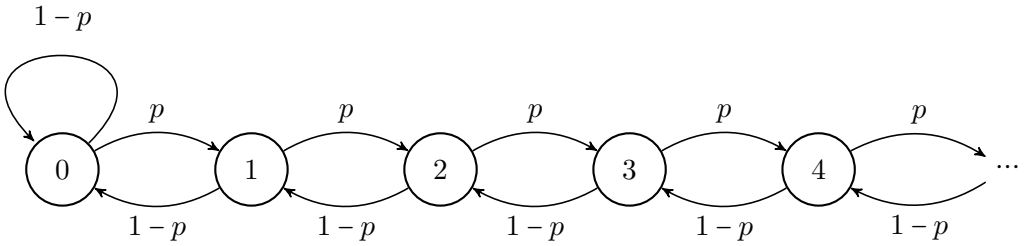
as desired. □

[22] Here, $\sum_{z\in A_x}P(u,z)$ is the flow of mass from state $u\in S\setminus A_x$ into the component A_x.

5.6 Random walk on the nonnegative integers

An irreducible Markov chain on a *finite* state space always has a unique invariant distribution π (by Theorem 2.8). If the Markov chain is also aperiodic, then the distribution μ_t of X_t converges to π as $t \to \infty$, regardless of the initial state (by Theorem 2.14). In the context of *infinite* state spaces, even the invariant distribution π might not exist in general, as we shall see shortly.

Let us consider the following generalization of the random walk in Section 4.4. A particle moves in the infinite set $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ 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 $1 - p$, independently of the past steps. With the boundary condition $P(0,0) = 1 - p$, we get the transition diagram



and the infinite transition matrix

$$P = \begin{bmatrix} 1-p & p & 0 & \cdots & & & & \\ 1-p & 0 & p & 0 & \cdots & & & \\ 0 & 1-p & 0 & p & 0 & \cdots & & \\ 0 & 0 & 1-p & 0 & p & 0 & \cdots & \\ 0 & 0 & 0 & 1-p & 0 & p & 0 & \cdots \\ \vdots & \vdots & & & \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{5.16}$$

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

5.6.1 Invariant distribution

Let us first study whether or not this random walk has an invariant distribution. It is an instance of a birth–death chain, so by Theorem 5.9, any possible invariant distribution π of P must satisfy the detailed balance equations (5.11):

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

or equivalently,

$$p \cdot \pi(x) = (1-p) \cdot \pi(x+1), \qquad x \geq 0,$$

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

$$\pi(x) = \Big(\frac{p}{1-p}\Big)^x \cdot \pi(0), \qquad x \geq 0.$$

In order for π to be a probability distribution, it must satisfy the law of total probability

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

If $p < q$, or equivalently $p < 1/2$, this normalisation is possible by choosing $\pi(0) = 1 - \frac{p}{1-p}$. However if $p \geq 1/2$, this is not possible. We conclude that:

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

Table 1: Properties of the random walk on nonnegative integers $\mathbb{N}_0$ defined by (5.16).

▷ If $p < 1/2$, the unique invariant distribution of the random walk is the geometric distribution

$$\pi(x) = \Big(1 - \frac{p}{1-p}\Big)\Big(\frac{p}{1-p}\Big)^x, \qquad \text{on } \mathbb{N}_0 = \{0, 1, 2, \dots\}.$$

▷ If $p \geq 1/2$, the random walk does *not* have an invariant distribution.

5.6.2 Recurrence and transience

Let us then investigate how the random walk behaves when $p \geq 1/2$. We study the question whether or not it ever returns to state 0 after leaving it: so we investigate the *return probability*

$$\rho(0,0) \;=\; \mathbb{P}\,(X_t = 0 \text{ for some } t \geq 1 \mid X_0 = 0) \;=\; \mathbb{P}\,(T_0^+ < \infty \mid X_0 = 0).$$

Recall that state 0 is *recurrent* if and only if $\rho(0,0) = 1$ (and *transient* otherwise).

Now, it is useful to notice that by the Markov property, it follows by a similar first-step analysis as in Chapter 4 (like[23] in the proof of Theorem 4.1, but instead of taking just one time-step, waiting for the first time that X is at state 1 after leaving from $X_0 = 0$) that *the probability that the random walk ever returns to* 0 *can also be written as*

$$\rho(0,0) \;=\; \mathbb{P}\,(T_0 < \infty \mid X_0 = 1) \;=\; \lim_{M \to \infty} \mathbb{P}\,(T_0 < T_M \mid X_0 = 1),$$

where T_x is the first passage time (4.1) into state x. Observe that $\mathbb{P}\,(T_0 < T_M \mid X_0 = 1)$ also equals a gambler's ruin probability with initial wealth 1 and target wealth M (like[24] in Section 4.4), so by Theorem 4.7, we can conclude that

$$\mathbb{P}\,(T_0 < T_M \mid X_0 = 1) \;=\; 1 - h(1) \;=\; \begin{cases} \dfrac{\left(\frac{1-p}{p}\right) - 1}{\left(\frac{1-p}{p}\right)^M - 1}, & p \neq 1/2, \\[2ex] \dfrac{1}{M}, & p = 1/2. \end{cases}$$

Hence, the probability that the random walk returns to state 0 after leaving it equals

$$\mathbb{P}\,(T_0 < \infty \mid X_0 = 1) \;=\; \begin{cases} 1, & p \leq 1/2, \\ \frac{1-p}{p}, & p > 1/2. \end{cases}$$

This means that the states of the random walk are

[23] Here, we may omit the hat notation used in the proof of Theorem 4.1, because the probabilities associated to the future states of the random walk are anyway the same as for the initial random walk.

[24] Since we stop the random walk when it returns to state 0, the difference of the transition probabilities at state 0 in this Chapter versus in Chapter 4 is insignificant for our computation here.

▷ *recurrent* for $p \le 1/2$, and

▷ *transient* for $p > 1/2$.

The case $p = 1/2$ is special in that although the random walk eventually returns to every state, one can show that the expected return time is infinite (see Example 4.9). Table 1 summarizes key properties of the random walk. Figure 5.1 describes paths of the random walk.

5.7 Additional material on recurrence

This section contains useful results for return properties of Markov chains and the proof of Theorem 5.4. It can be skipped when focusing on the core contents of this course. The material here does not need any further mathematical background — and it is advised that students majoring in mathematics and oriented to probability theory have a look at this section.

Lemma 5.11 (*Recurrence*)**.** *State $x \in S$ is recurrent* if and only if *starting from any initial state in $C(x)$, state x is visited infinitely often with probability one.*

Proof. Consider the event A_x that the Markov chain X visits state x *only finitely many times*. By dividing according to the number of visits to x, we can write event A_x as a disjoint union

$$A_x = \bigcup_{T=0}^{\infty} \{X_T = x,\, X_s \neq x \text{ for all } s > T\},$$

where T is the last time instant when X visits x. By the Markov property (5.2), it follows that

$$\begin{aligned} \mathbb{P}(X_T = x,\, X_s \neq x \text{ for all } s > T) &= \mathbb{P}(X_T = x) \cdot \mathbb{P}(X_s \neq x \text{ for all } s > T \mid X_T = x) \\ &= \mathbb{P}(X_T = x) \cdot \mathbb{P}(T_x^+ = \infty \mid X_0 = x) \qquad \text{[by (5.2)]} \\ &= \mathbb{P}(X_T = x) \cdot (1 - \rho(x,x)). \end{aligned}$$

Therefore, we obtain

$$\mathbb{P}(A_x) = \sum_{T=0}^{\infty} \mathbb{P}(X_T = x,\, X_s \neq x \text{ for all } s > T) = (1 - \rho(x,x)) \cdot \sum_{T=0}^{\infty} \mathbb{P}(X_T = x). \tag{5.17}$$

Applying[25] Theorem 2.9 in the component $C(x)$, we know that for all $z \in C(x)$ there exists $T \in \mathbb{N}$ such that $P^T(z,x) > 0$. Therefore, the above sum has at least one nonzero term. Thus,

$$\mathbb{P}(A_x) = 0 \qquad \Longleftrightarrow \qquad \rho(x,x) = 1. \tag{5.18}$$

It remains to observe that $\mathbb{P}(A_x) = 0$ means that its complementary event, that is, that state x is visited infinitely often, has probability one. Thus, the equivalence (5.18) proves the claim. □

Lemma 5.12. *If $x \in S$ is recurrent, then for all states $y \in S$ which are reachable from x, we have $\rho(y,x) = 1$.*

(The proof of this lemma is technical, and can be skipped for the first reading.)

[25] Note that the proof of Theorem 2.9 works also for countably infinite state spaces.

Proof. Let $t \geq 0$ be the length of the shortest path from x to y in the transition diagram of the Markov chain. Then, the transition diagram contains a t-hop path $x = x_0 \to x_1 \to \cdots \to x_t = y$ which is such that x does not belong to $\{x_1, \ldots, x_t\}$ and

$$P(x_0, x_1) \cdot P(x_1, x_2) \cdots P(x_{t-1}, x_t) \; > \; 0. \tag{5.19}$$

By the Markov property (5.2), the probability that after starting at x, the Markov chain never returns to x can be bounded as

$$\begin{aligned} 1 - \rho(x,x) \;=\; \mathbb{P}\,(T_x^+ = \infty \mid X_0 = x) \;&=\; \mathbb{P}\,(X_t \neq x \text{ for all } t \geq 1 \mid X_0 = x) \\ &\geq\; P(x_0, x_1) \cdot P(x_1, x_2) \cdots P(x_{t-1}, x_t) \cdot \mathbb{P}\,(T_x^+ = \infty \mid X_0 = y) \\ &=\; \underbrace{P(x_0, x_1) \cdot P(x_1, x_2) \cdots P(x_{t-1}, x_t)}_{> 0} \cdot (1 - \rho(y,x)). \end{aligned}$$

Because state x is recurrent by assumption, we have $\rho(x,x) = 1$. Hence, we see from (5.19) that it must be the case that $1 - \rho(y,x) = 0$, that is, $\rho(y,x) = 1$. □

We summarize the main properties in the next theorem.

Theorem 5.13 (*Recurrence in components*)**.** *If state $x \in S$ is* recurrent*, then for all states $y, z \in C(x)$ in the same component, the following hold.*

- ▷ $\rho(y,z) = 1$.
- ▷ *y (and z) is also recurrent.*
- ▷ *Starting from z, state y is visited infinitely often with probability one.*

Note that the statements hold in particular for state x itself.

(The proof of this result is technical, and can be skipped for the first reading.)

Proof. We first prove the claims for $z = x$. Thanks to Lemma 5.11, it remains to show that any state $y \in C(x)$ is recurrent, that is, $\rho(y,y) = 1$. For this, we will study the expected number of visits to state y overall, using the occupancy times from Section 1.6.

Recall that the occupancy time of state y for initial state x is

$$G_t(x,y) \;=\; \mathbb{E}\,(N_t(y) \mid X_0 = x),$$

where $N_t(y)$ is the number of visits to state y during the first t time steps (1.14). Because for each state y the map $t \mapsto N_t(y)$ is non-decreasing for times $t = 0, 1, 2, \ldots$, the limit

$$\begin{aligned} \lim_{t\to\infty} G_t(x,y) \;=\; \lim_{t\to\infty} \mathbb{E}\,(N_t(y) \mid X_0 = x) \;&=\; \mathbb{E}\Big(\lim_{t\to\infty} N_t(y) \mid X_0 = x\Big) \\ &=\; \mathbb{E}\Big(\sum_{s=0}^{\infty} \mathbb{1}(X_s = y) \,\Big|\, X_0 = x\Big) \\ &=\; \sum_{s=0}^{\infty} \mathbb{P}\,(X_s = y \mid X_0 = x) \;\in\; [0, \infty] \end{aligned}$$

exists and takes values in the extended number set $[0, \infty]$. We denote this limit as

$$\begin{aligned} G(x,y) \;&=\; \sum_{s=0}^{\infty} \mathbb{P}\,(X_s = y \mid X_0 = x) \\ &=\; \sum_{s=0}^{\infty} P^s(x,y) \;\in\; [0, \infty]. \qquad \text{[by (5.4) in Theorem 5.3]} \end{aligned}$$

It represents the expected number of visits to state y starting from x for all time. In the literature, $G(x,y)$ is also called *Green's function*, due to its relation to potential theory [LPW08].

Recall that we are going to prove that y is a recurrent state, that is, $\rho(y,y) = 1$. Our computation in Equation (5.17) shows that the probability of the event A_y that the Markov chain X visits state y only finitely many times is

$$(1-\rho(y,y))\cdot G(x,y) \;=\; (1-\rho(y,y))\cdot \sum_{T=0}^{\infty} \mathbb{P}\,(X_T = y \mid X_0 = x) \;=\; \mathbb{P}\,(A_y \mid X_0 = x) \;\in\; [0,1].$$

Hence, we see that if $G(x,y) = \infty$, then $1-\rho(y,y) = 0$. Thus, to finish the proof we aim at showing that $G(x,y) = \infty$.

By irreducibility, Theorem 2.9 shows that there exists a time instant $T \in \mathbb{N}$ such that $P^T(x,y) > 0$. Hence, we may conclude using (5.4) and the Markov property (5.2) that

$$P^s(x,x)\cdot P^T(x,y) \;\le\; P^{s+T}(x,y), \qquad s \ge 0.$$

Therefore, we obtain

$$G(x,y) \;\ge\; \sum_{s=0}^{\infty} P^{t+s}(x,y) \;\ge\; P^t(x,y)\cdot \sum_{s=0}^{\infty} P^s(x,x) \;=\; P^t(x,y)\cdot G(x,x) \;=\; \infty,$$

since $G(x,x) = \infty$, as x is recurrent by assumption, and it is surely visited infinitely many times by Lemma 5.11. This shows that $G(x,y) = \infty$ as well.

This proves the claims for $z = x$. To address the case of arbitrary $y, z \in C(x)$, note that Lemma 5.12 shows that $\rho(y,x) = 1$, and from the first part of the proof, we know that z also recurrent and $\rho(x,z) = 1$. Hence, we obtain

$$\rho(y,z) \;\ge\; \rho(y,x)\cdot\rho(x,z) \;=\; 1.$$

The other claims for general $z \in C(x)$ follow from the first part of the proof. □

Proof of Theorem 5.4. Let us first verify the positivity (5.7). Because $\sum_x \pi(x) = 1$, we can choose a state x_0 such that $\pi(x_0) > 0$. By irreducibility, the transition diagram contains a path from x_0 to y, so that $P^t(x_0,y) > 0$, where $t \in \mathbb{N}_0$ is the length of the path. Because by the balance equation $\pi\cdot P = \pi$, we also have $\pi\cdot P^t = \pi$, and we obtain the positivity (5.7):

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

Equation (5.17) holds for any state $y \in S$ and for any initial distribution of the chain. Especially, if we denote by $\mathbb{P}_\pi$ the distribution of the Markov chain corresponding to the initial distribution $\mu_0 = \pi$, then because $\mathbb{P}_\pi(X_T = y) = \pi(y)$, it follows that

$$\mathbb{P}_\pi(A_y) \;=\; (1-\rho(y,y))\cdot \sum_{T=0}^{\infty} \mathbb{P}_\pi(X_T = y) \;=\; (1-\rho(y,y))\cdot \sum_{T=0}^{\infty} \pi(y).$$

Because the terms of the sum do not depend on T, we must have $\pi(y)(1-\rho(y,y)) = 0$. Furthermore, by (5.7), we know that $\pi(y) > 0$, so we conclude that $\rho(y,y) = 1$ for any $y \in S$. Thus, all states are recurrent. Lemma 5.11 then implies that they are also visited infinitely often. □

6 Generating functions

6.1 Why generating functions?

For any sequence $a_0, a_1, a_2, \ldots$, the function

$$F(z) = \sum_{k=0}^{\infty} a_k z^k$$

is called the *generating function* (*generoiva funktio*) of the sequence. Note that the series defining $F(z)$ might not necessarily converge anywhere.

Using generating functions one can solve easily many recursion problems, even if the solution would be very complicated. To illustrate the method, let us look at a simple example. You can find an extensive collection of examples and applications in the online book [Wil94].

Example 6.1 (*Fibonacci sequence*)**.** The Fibonacci recurrence is

$$f_0 = 0, \qquad f_1 = 1, \qquad f_{k+1} = f_k + f_{k-1}, \qquad k = 1, 2, 3, \ldots. \tag{6.1}$$

We form the generating function

$$F(z) = \sum_{k=0}^{\infty} f_k z^k.$$

Plugging into $F(z)$ the recurrence (6.1), we obtain

$$\begin{aligned}
F(z) &= f_0 + f_1 z + \sum_{k=2}^{\infty} f_k z^k \\
&= f_0 + f_1 z + \sum_{\ell=1}^{\infty} f_{\ell+1} z^{\ell+1} \\
&= z + \sum_{\ell=1}^{\infty} (f_\ell + f_{\ell-1}) z^{\ell+1} && \text{[by (6.1)]} \\
&= z + z \cdot \sum_{k=0}^{\infty} f_k z^k + z^2 \cdot \sum_{k=0}^{\infty} f_k z^k \\
&= z + z \cdot F(z) + z^2 \cdot F(z).
\end{aligned}$$

Rearranging this, we obtain

$$F(z) = \frac{z}{1 - z - z^2}.$$

So the generating function has quite a simple formula. Now, in order to find the coefficients $\{f_k : k = 2, 3, 4, \ldots\}$, we just have to expand $F(z)$ as a power series. This can be done using the *partial fraction decomposition*. Note that the denominator in $F(z)$ is a polynomial

$$1 - z - z^2 = -(z - \alpha_+)(z - \alpha_-)$$

with roots

$$\alpha_\pm = \frac{-1 \pm \sqrt{5}}{2}.$$

Note that $\alpha_+ - \alpha_- = \sqrt{5}$. We write

$$F(z) = \frac{z}{1 - z - z^2} = \frac{-z}{(z - \alpha_+)(z - \alpha_-)} = \frac{A(z)}{z - \alpha_+} + \frac{B(z)}{z - \alpha_-}$$

for suitable A and B:

$$\frac{A}{z-\alpha_+} + \frac{B}{z-\alpha_-} = \frac{A(z-\alpha_-) + B(z-\alpha_+)}{(z-\alpha_+)(z-\alpha_-)}.$$

Equating this with $F(z)$ gives

$$A(z-\alpha_-) + B(z-\alpha_+) = -z.$$

Evaluating at $z = \alpha_+$ gives

$$A = -\frac{\alpha_+}{\alpha_+ - \alpha_-} = \frac{\sqrt{5}-5}{10},$$

and evaluating at $z = \alpha_-$ gives

$$B = \frac{\alpha_-}{\alpha_+ - \alpha_-} = \frac{-\sqrt{5}-5}{10}.$$

Hence, we have

$$\begin{aligned} F(z) &= \frac{A(z)}{z-\alpha_+} + \frac{B(z)}{z-\alpha_-} \\ &= -\frac{\left(\frac{\alpha_+}{\alpha_+-\alpha_-}\right)}{z-\alpha_+} + \frac{\left(\frac{\alpha_-}{\alpha_+-\alpha_-}\right)}{z-\alpha_-} = \frac{1}{\alpha_+-\alpha_-}\left(\frac{\alpha_-}{z-\alpha_-} - \frac{\alpha_+}{z-\alpha_+}\right). \end{aligned}$$

Notice that we can write this in terms of the *geometric series* (*geometrinen sarja*):

$$\frac{\alpha}{z-\alpha} = \frac{\alpha}{-\alpha(1-\frac{z}{\alpha})} = -\frac{1}{1-\frac{z}{\alpha}} = -\sum_{n=0}^{\infty}\left(\frac{z}{\alpha}\right)^n.$$

Therefore, we conclude that the generating function reads

$$\begin{aligned} F(z) &= \frac{1}{\alpha_+-\alpha_-}\left(\frac{\alpha_-}{z-\alpha_-} - \frac{\alpha_+}{z-\alpha_+}\right) \\ &= \frac{1}{\alpha_+-\alpha_-}\left(-\sum_{n=0}^{\infty}\left(\frac{z}{\alpha_-}\right)^n + \sum_{n=0}^{\infty}\left(\frac{z}{\alpha_+}\right)^n\right) = \frac{1}{\alpha_+-\alpha_-}\sum_{n=0}^{\infty}\left(\alpha_+^{-n} - \alpha_-^{-n}\right)z^n, \end{aligned}$$

and we can directly read the coefficients

$$f_n = \frac{1}{\alpha_+-\alpha_-}\left(\alpha_+^{-n} - \alpha_-^{-n}\right) = \frac{1}{\sqrt{5}}\left(\left(\frac{2}{-1+\sqrt{5}}\right)^n - \left(\frac{2}{-1-\sqrt{5}}\right)^n\right),$$

which we can double-check to indeed satisfy $f_0 = 0$, $f_1 = 1$, $f_2 = 1$, $f_3 = 2$, $f_4 = 3$, etc. ■

Reminder. Recall the useful formula for the *geometric series*:

$$\sum_{n=0}^{\infty} z^n = 1 + z + z^2 + z^3 + \cdots = \frac{1}{1-z}, \qquad |z| < 1.$$

This is used very frequently when dealing with generating functions. Another useful and familiar series is the *exponential series* (*exponenttifunktion sarja*)

$$e^z = \sum_{n=0}^{\infty}\frac{z^n}{n!}, \qquad z \in \mathbb{R}, \tag{6.2}$$

where $n! = 1\cdot 2\cdot 3\cdots(n-1)\cdot n$. ♣

6.2 Probability generating functions

We will next gather some of the key aspects of generating functions describing a probability distribution, which will be repeatedly applied in this course. For more background and applications of generating functions in probability theory, see [GS97, Chapter 10].

Definition. The *probability generating function* (*todennäköisyydet generoiva funktio*) of a random integer Y in $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ distributed according to $\mathbb{P}(Y = k) = p_k$ is

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

for those values of $s \in \mathbb{R}$ (or $\mathbb{C}$) for which the sum on the right side converges.

▷ The probability generating function ϕ_Y is always defined for $s \in [-1, 1]$: since $\{p_k : k \in \mathbb{N}_0\}$ forms a probability distribution, we have

$$\sum_{k=0}^{\infty} |p_k| \cdot |s|^k \leq \sum_{k=0}^{\infty} p_k = \phi_Y(1) = 1, \qquad s \in [-1, 1]. \tag{6.4}$$

▷ ϕ_Y is also defined for other values of s if the probabilities p_k vanish quickly enough for large values of k.

▷ ϕ_Y can also be defined for real-valued random variables Y via the formula $\phi_Y(s) = \mathbb{E}(s^Y)$.

Theorem 6.2 (*Probability generating function determines distribution*)**.** *The values of*

$$\phi_Y(s) = \sum_{k=0}^{\infty} \mathbb{P}(Y = k) \cdot s^k, \qquad s \in [-1, 1],$$

determine the probability distribution of Y uniquely. Moreover, we have[a]

$$\mathbb{P}(Y = k) = \frac{\phi_Y^{(k)}(0)}{k!}, \qquad k = 0, 1, 2, \ldots, \tag{6.5}$$

and if $\phi_Y'(1)$ exists, then it gives the expected value of Y,

$$\phi_Y'(1) = \mathbb{E}(Y). \tag{6.6}$$

[a] Here, $\phi^{(k)}(s) = (\frac{\mathrm{d}}{\mathrm{d}s})^k \phi(s)$ denotes the k:th derivative of the function ϕ.

▷ Note that in particular, $\mathbb{P}(Y = 0) = \phi_Y(0) \in [0, 1]$ as a special case of (6.5).

▷ In fact, a slightly stronger property than (6.6) also holds: in general, we have

$$\phi_Y'(1-) = \lim_{s \to 1-} \phi_Y'(s) = \mathbb{E}(Y). \tag{6.7}$$

Proof. We will see in Lemma 6.7 that the series (6.3) defining ϕ_Y can be differentiated infinitely many times. The probabilities (6.5) determine the probability distribution of Y uniquely. The formula (6.6) is proved in Lemma 6.3 below. □

Let us collect some very useful properties of probability generating functions. The proofs of some of them are beyond this course, but we will give a sketch proof for interested readers.

Lemma 6.3. *We have*

- ▷ $\phi_Y(0) = \mathbb{P}(Y=0) \in [0,1]$ *and*
- ▷ $\phi_Y(1) = 1$.

Suppose that the derivatives $\phi'_Y(s)$ and $\phi''_Y(s)$ exist at $s=1$. Then, we have

- ▷ $\phi'_Y(1) = \mathbb{E}(Y)$,
- ▷ $\phi''_Y(1) = \mathbb{E}(Y^2) - \mathbb{E}(Y)$,
- ▷ $\operatorname{var}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \phi''_Y(1) + \phi'_Y(1) - (\phi'_Y(1))^2$.

We also have the scaling property

$$\phi_Y(s^n) = \phi_{nY}(s), \qquad \text{for all } n \in \mathbb{N}. \tag{6.8}$$

Proof. All claims are straightforward to compute from definition (6.3). First, we have

$$\phi_Y(s) = \sum_{k=0}^{\infty} p_k s^k = \begin{cases} p_0 \cdot 0^0 + 0 + 0 + \cdots = p_0, & s = 0, \\ \sum_{k=0}^{\infty} p_k = 1, & s = 1, \end{cases}$$

and

$$\phi'_Y(s) = \sum_{k=1}^{\infty} k\, p_k\, s^{k-1} \qquad \text{and} \qquad \phi''_Y(s) = \sum_{k=2}^{\infty} k(k-1)\, p_k\, s^{k-2}. \tag{6.9}$$

These evaluate at $s=1$ to

$$\phi'_Y(1) = \sum_{k=1}^{\infty} k\, p_k = \sum_{k=1}^{\infty} k \cdot \mathbb{P}(Y=k) = \mathbb{E}(Y),$$

and (since $k^2 = k$ when $k=1$)

$$\phi''_Y(1) = \sum_{k=2}^{\infty} k^2 p_k - \sum_{k=2}^{\infty} k\, p_k = \sum_{k=1}^{\infty} k^2 p_k - \sum_{k=1}^{\infty} k\, p_k = \mathbb{E}(Y^2) - \mathbb{E}(Y).$$

The variance $\operatorname{var}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \phi''_Y(1) + \phi'_Y(1) - (\phi'_Y(1))^2$ follows from these. Lastly, the scaling relation (6.8) also follows using definition (6.3):

$$\phi_Y(s^n) = \mathbb{E}((s^n)^Y) = \mathbb{E}(s^{nY}) = \phi_{nY}(s).$$

□

Example 6.4 (*Bernoulli random variable*)**.** For a Bernoulli random variable $Y \sim \operatorname{Ber}(p)$ with values $\{0,1\}$, that is,

$$\mathbb{P}(Y=1) = p \qquad \text{and} \qquad \mathbb{P}(Y=0) = 1-p,$$

the probability generating function is the polynomial $\phi_Y(s) = p\,s + (1-p)$.

One can also consider a Bernoulli random variable with different values, such as $\{0, n\}$, that is,

$$\mathbb{P}(Y = n) = q \qquad \text{and} \qquad \mathbb{P}(Y = 0) = 1 - q.$$

Its probability generating function is $\phi_Y(s) = q\,s^n + (1-q)$. (The probability generating function depends not only on the probabilities but also on the values of the random variable.) ■

Example 6.5 (*Geometric random variable*)**.** For a geometric random variable Y with

$$\mathbb{P}(Y = k) = (1-p)^k\,p,$$

the probability generating function is

$$\phi_Y(s) = \sum_{k=0}^{\infty} (1-p)^k\,p\,s^k = p \sum_{k=0}^{\infty} ((1-p)\,s)^k = \frac{p}{1-(1-p)\,s},$$

where we used the geometric series to evaluate the sum. From ϕ_Y and Lemma (6.3), we easily obtain the expected value

$$\mathbb{E}(Y) = \phi_Y'(1) = \left(\frac{p(1-p)}{(1-(1-p)\,s)^2}\right)\Big|_{s=1} = \frac{1-p}{p},$$

and the variance

$$\begin{aligned}\operatorname{var}(Y) &= \phi_Y''(1) + \phi_Y'(1) - (\phi_Y'(1))^2 \\ &= \left(\frac{2p(1-p)^2}{(1-(1-p)\,s)^3} + \frac{p(1-p)}{(1-(1-p)\,s)^2} - \frac{p^2(1-p)^2}{(1-(1-p)\,s)^4}\right)\Big|_{s=1} = \frac{1-p}{p^2}.\end{aligned}$$

■

Example 6.6 (*Poisson random variable*)**.** For a Poisson distributed random variable $Y \sim \mathrm{Poi}(\lambda)$ with

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

the probability generating function is

$$\phi_Y(s) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\,s^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{-\lambda}\,e^{\lambda s} = e^{\lambda(s-1)},$$

where we used the series expansion (6.2) of the exponential function to evaluate the sum. From ϕ_Y and Lemma (6.3), we easily obtain the expected value

$$\mathbb{E}(Y) = \phi_Y'(1) = \lambda\,e^{\lambda(s-1)}\Big|_{s=1} = \lambda,$$

and the variance

$$\operatorname{var}(Y) = \phi_Y''(1) + \phi_Y'(1) - (\phi_Y'(1))^2 = \left(\lambda^2\,e^{\lambda(s-1)} + \lambda\,e^{\lambda(s-1)} - \lambda^2\,e^{2\lambda(s-1)}\right)\Big|_{s=1} = \lambda.$$

■

Lemma 6.7. *The probability generating function ϕ_Y satisfies the following properties.*

- ▷ *It is continuous as a map $s \mapsto \phi_Y(s)$.*
- ▷ *It is convex (konveksi): for all $s, t \in [0,1]$, we have*

$$\phi_Y\big(\alpha\, t + (1-\alpha)\, s\big) \;\le\; \alpha\, \phi_Y(t) \;+\; (1-\alpha)\, \phi_Y(s), \qquad \text{for all } \alpha \in [0,1]. \tag{6.10}$$

- ▷ *It is non-decreasing on $[0,1]$ and satisfies $\phi_Y(s) \in [0,1]$ for all $s \in [0,1]$.*
- ▷ *It is infinitely many times differentiable on $(-1,1)$, and*

$$\mathbb{P}(Y = k) \;=\; p_k \;=\; \frac{\phi_Y^{(k)}(0)}{k!}, \qquad k = 0, 1, 2, \ldots. \tag{6.11}$$

Convexity means that the straight line between any pair of points on the curve of ϕ_Y is above or just meets the graph of ϕ_Y. Indeed, varying $\alpha \in [0,1]$, we see that the right-hand side in (6.10) is the straight line between the points $(s, \phi_Y(s))$ and $(t, \phi_Y(t))$, while the left-hand side is the graph of ϕ_Y between the points t and s. See Figure 6.1.

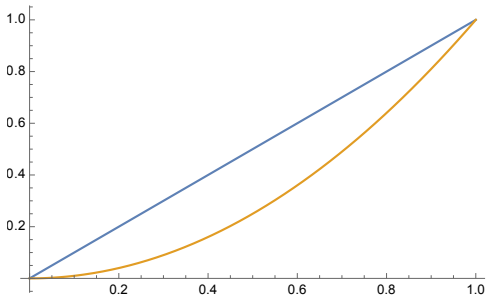


Figure 6.1: The linear function $s \mapsto s$ (blue) and an example of a convex function $s \mapsto s^2$ (orange) on the unit interval $[0,1]$.

Proof. By (6.4) the convergence radius[26] of the power series on the right side of (6.3) is always at least 1, and therefore the series defining ϕ_Y is continuous and can be differentiated infinitely many times[27] term be term at every point in $(-1,1)$. In particular, by differentiating it k times at zero, we find (6.11).

Next, by investigating the derivatives (6.9), we see that both $\phi_Y'(s)$ and $\phi_Y''(s)$ are non-negative for $s \in [0,1]$.

- ▷ Hence, since $\phi_Y'(s) \ge 0$ for all $s \in [0,1]$, the function $s \mapsto \phi_Y(s)$ is non-decreasing on $[0,1]$. Because $\phi_Y(1) = 1$ and $\phi_Y(0) = p_0 \ge 0$, we see that $\phi_Y(s) \in [0,1]$ for all $s \in [0,1]$.
- ▷ It is a good exercise for a mathematically oriented reader to prove using the Mean Value Theorem (väliarvolause)[28] that for any function $f : [0,1] \to \mathbb{R}$, it holds that if $f''(s) \ge 0$ for all $s \in [0,1]$, then f is convex, i.e., (6.10) holds.

[26] The convergence radius of a power series $\sum_k p_k\, x^k$ is the largest number $R \ge 0$ such that $\sum_k |p_k|\, |x|^k < \infty$ whenever $|x| < R$. This means that the series converges absolutely.

[27] This is discussed in the course Differentiaali- ja integraalilaskenta 1 (MS-A010X).

[28] This is also discussed in the course Differentiaali- ja integraalilaskenta 1 (MS-A010X).

▷ It therefore follows from the property $\phi_Y''(s) \geq 0$ for all $s \in [0,1]$ that ϕ_Y is convex.

□

6.3 Multiplicativity properties of probability generating functions

The key usefulness of probability generating functions is that they behave well for sums of independent random variables.

Theorem 6.8 (*Multiplicativity*). *Consider* independent *random integers* $Y_1, Y_2, \ldots, Y_n$ *in* $\mathbb{N}_0$. *Then, their sum* $Y = Y_1 + Y_2 + \cdots + Y_n$ *is also a random integer in* $\mathbb{N}_0$ *and its probability generating function is*

$$\phi_Y(s) = \phi_{Y_1}(s) \cdot \phi_{Y_2}(s) \cdots \phi_{Y_n}(s). \tag{6.12}$$

A special case of Theorem 6.8 is when $Y_1, Y_2, \ldots, Y_n$ are iid random numbers in $\mathbb{N}_0$:

$$\phi_{Y_1+Y_2+\cdots+Y_n}(s) = (\phi_{Y_1}(s))^n, \tag{6.13}$$

since for iid random numbers, $\phi_{Y_1}(s) = \phi_{Y_2}(s) = \cdots = \phi_{Y_n}(s)$ by Theorem 6.2.

Proof. The left-hand side of (6.12) is

$$\phi_Y(s) = \mathbb{E}(s^{Y_1+Y_2+\cdots+Y_n}) = \mathbb{E}(s^{Y_1} \cdot s^{Y_2} \cdots s^{Y_n}),$$

and because $Y_1, Y_2, \ldots, Y_n$ are independent, it factorizes to

$$\mathbb{E}(s^{Y_1} \cdot s^{Y_2} \cdots s^{Y_n}) = \mathbb{E}(s^{Y_1}) \cdot \mathbb{E}(s^{Y_2}) \cdots \mathbb{E}(s^{Y_n}) = \phi_{Y_1}(s) \cdot \phi_{Y_2}(s) \cdots \phi_{Y_n}(s),$$

which is the right-hand side of (6.12). □

The following result generalizes Theorem 6.8 to the case where the number of summands is a random variable. (An empty sum $\sum_{j=1}^{0} Y_j$ is defined as zero in the formula below.)

Theorem 6.9 (*Composition*). *Consider* independent *random integers* N *and* $Y_1, Y_2, \ldots$ *in* $\mathbb{N}_0$. *If* $Y_1, Y_2, \ldots$ *are* identically distributed (iid), *then the probability generating function of the random sum*

$$Y = \sum_{j=1}^{N} Y_j$$

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

Proof. By conditioning on the possible values of N and applying independence and identity (6.12) from Theorem 6.8 we find that

$$\begin{aligned}
\phi_Y(s) &= \sum_{n=0}^{\infty} \mathbb{P}(N = n) \cdot \mathbb{E}(s^{Y_1+Y_2+\cdots+Y_n} \mid N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N = n) \cdot \mathbb{E}(s^{Y_1+Y_2+\cdots+Y_n}) && [N \text{ and } Y_1, Y_2, \ldots \text{ independent}] \\
&= \sum_{n=0}^{\infty} \mathbb{P}(N = n) \cdot (\phi_{Y_1}(s))^n && [\text{by (6.13), since } Y_1, Y_2, \ldots \text{ iid}] \\
&= \phi_N(\phi_{Y_1}(s)). && [\text{by (6.3)}]
\end{aligned}$$

□

Theorem 6.10 (*Wald's identity*). *Consider* independent *random integers* N *and* $Y_1, Y_2, \ldots$ *in* $\mathbb{N}_0$. *Suppose furthermore that* $Y_1, Y_2, \ldots$ *are* identically distributed (iid), *and that* $\mathbb{E}(N) < \infty$ *and* $\mathbb{E}(Y_1) < \infty$. *Then, we have*

$$\mathbb{E}\Big(\sum_{j=1}^{N} Y_j\Big) = \mathbb{E}(N) \cdot \mathbb{E}(Y_1). \tag{6.14}$$

Proof. The assumptions $\mathbb{E}(N) < \infty$ and $\mathbb{E}(Y_1) < \infty$ imply that we may differentiate both sides of the identity from Theorem 6.9,

$$\phi_Y(s) = \phi_N(\phi_{Y_1}(s)),$$

at $s = 1$ and use the chain rule to obtain

$$\phi_Y'(s) = \phi_N'(\phi_{Y_1}(s)) \cdot \phi_{Y_1}'(s) \quad \xrightarrow{s \to 1} \quad \phi_N'(\phi_{Y_1}(1)) \cdot \phi_{Y_1}'(1) = \phi_N'(1) \cdot \phi_{Y_1}'(1) = \mathbb{E}(N) \cdot \mathbb{E}(Y_1),$$

using the values $\phi_{Y_1}(1) = 1$, and $\phi_N'(1) = \mathbb{E}(N)$, and $\phi_{Y_1}'(1) = \mathbb{E}(Y_1)$ from Lemma 6.3. □

Example 6.11. Otaniemi Eulers take part in the championship of the robot football league every spring. The probability that Eulers wins a game is $p \in (0,1)$, and each match is considered independent. Alas, Eulers do not make it to the championship finals every year: the number of years between two finals that Eulers get to play is random and obeys the Poisson distribution $\mathrm{Poi}(\lambda)$ with mean $\lambda > 0$. What is the expected number Y of years between the consecutive wins that Eulers gain in the finals?

We model the numbers of years between the instances when Eulers are in the championship final by iid random variables $Y_1, Y_2, \ldots$ following the Poisson distribution

$$\mathbb{P}(Y_1 = k) = \begin{cases} e^{-\lambda}\frac{\lambda^k}{k!}, & k \geq 0, \\ 0, & k < 0. \end{cases}$$

We model the number of attempts needed before a win by a geometric random variable N,

$$\mathbb{P}(N = k) = (1-p)^k p.$$

Then, every time Eulers lose the final, another Y_j years have to pass before the next chance, and our count stops when they eventually win. Thus, we have

$$Y = \sum_{j=0}^{N} Y_j.$$

By Theorem 6.10, the expected value of Y equals

$$\mathbb{E}(Y) = \mathbb{E}(N) \cdot \mathbb{E}(Y_1) = \frac{1-p}{p}\lambda.$$

■

6.4 Moment generating functions

Definition. The *moment generating function* (*momentit generoiva funktio*) of a random integer Y in $\mathbb{N}_0$ distributed according $\mathbb{P}(Y = k) = p_k$ is

$$M_Y(s) = \mathbb{E}(e^{sY}) = \sum_{k=0}^{\infty} p_k e^{sk}, \tag{6.15}$$

for those values of s for which the sum on the right side converges.

▷ The moment generating function M_Y is always defined for $s \in (-\infty, 0]$: since $\{p_k : k \in \mathbb{N}_0\}$ forms a probability distribution, we have

$$\sum_{k=0}^{\infty} |p_k| \cdot e^{sk} \leq \sum_{k=0}^{\infty} p_k = 1, \qquad s \in (-\infty, 0].$$

▷ Note that $\phi_Y(s) = M_Y(\log s)$, for all $s \in (0, 1)$

▷ Note that $M_Y(0) = 1$ and $\lim_{s \to -\infty} M_Y(s) = 0$.

▷ The moment generating function M_Y can also be defined for real-valued random variables Y via the formula $M_Y(s) = \mathbb{E}(e^{sY})$.

Example 6.12.

▷ In Example 6.5, the moment generating function of a geometric random variable Y is

$$M_Y(s) = \phi_Y(e^s) = \frac{p}{1 - (1-p)e^s}.$$

▷ In Example 6.6, the moment generating function of a Poisson random variable Y is

$$M_Y(s) = \phi_Y(e^s) = e^{\lambda(e^s - 1)}.$$

■

Theorem 6.13 (*Moments*). *Suppose that the derivatives $M_Y^{(k)}(s)$ exist at $s = 0$. Then, we have*

$$\mathbb{E}(Y^k) = M_Y^{(k)}(0), \qquad k = 0, 1, 2, \ldots.$$

Theorem 6.14 (*Multiplicativity*). *Consider* independent *random integers $Y_1, Y_2, \ldots, Y_n$ in $\mathbb{N}_0$. The moment generating function of $Y = Y_1 + Y_2 + \cdots + Y_n$ is*

$$M_Y(s) = M_{Y_1}(s) \cdot M_{Y_2}(s) \cdots M_{Y_n}(s).$$

Proof. We leave the proofs of Theorems 6.13 and 6.14 as an exercise. □

Lemma 6.15. *The moment generating function M_Y satisfies the following properties.*

- ▷ *It is continuous as a map $s \mapsto M_Y(s)$.*
- ▷ *It is convex: for all $s, t \in (-\infty, 0]$, we have*

$$M_Y\big(\alpha\, t + (1-\alpha)\, s\big) \;\leq\; \alpha\, M_Y(t) \;+\; (1-\alpha)\, M_Y(s), \qquad \text{for all } \alpha \in [0,1].$$

Proof. This can be proven similarly as Lemma 6.7. □

6.5 Using generating functions to solve difference equations

In the exercises, you will see how generating functions can be used also to find solutions to linear difference equations.

7 Branching processes

7.1 Branching processes as a Markov chain and its transition matrix

For more background of branching processes, see, e.g. [GS97, Chapter 10].

Definition. A *branching process* (*haarautumisprosessi*) is a countable-state Markov chain $X = (X_0, X_1, X_2, \dots)$ on state space $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ 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 parametrized by an *offspring distribution* (*lisääntymisjakauma*)

$$p = \{p(k) : k \in \mathbb{N}_0\} = \{p(k) : k = 0, 1, 2, \dots\} \tag{7.1}$$

which is a probability distribution on $\mathbb{N}_0 = \{0, 1, 2, \dots\}$, where the entry $p(k)$ equals the probability that an individual produces k children.

X_t represents the number of individuals in t:th generation and is defined inductively by

$$X_{t+1} = \sum_{j=1}^{X_t} Y_j, \tag{7.2}$$

where $Y_1, Y_2, \dots$ are independent and identically distributed (iid) p-distributed random integers representing the offspring of each individual in the t:th generation.

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* (*Galton–Watson prosessi*). Branching processes are applied to several types 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.
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 branching process into state zero? (Recall (4.4) with $A = \{0\}$.) We will answer this question in Theorem 7.6.

▷ If there are $X_0 = n$ individuals in the zero:th generation, then the size of generation 1 is

$$X_1 = Y_1 + \cdots + Y_n,$$

where $Y_1, Y_2, \dots$ are iid p-distributed random integers representing the offspring of each individual in the initial population. Note in particular that if $X_0 = 1$ (interpreted as tracking the evolution line of one individual), then $X_1 = Y_1 \sim p$ is just p-distributed with (7.1).

▷ If there are no individuals in generation t, then no children are born and hence also the next generation is empty. State 0 is hence *absorbing* for the branching process Markov chain. The interpretation is that when X enters 0, the population becomes *extinct*. Note that

$$X_s = 0 \quad \Longrightarrow \quad X_t = 0, \qquad \text{for all } t \geq s.$$

Theorem 7.1 (*Transition matrix*). *For branching process $X = (X_0, X_1, X_2, \ldots)$, the transition probability from state $x \geq 1$ to state $y \geq 0$ equals*

$$P(x,y) \;=\; \mathbb{P}\,(Y_1 + \cdots + Y_x = y), \tag{7.3}$$

and the transition probability from state $x = 0$ to state $y \geq 0$ equals

$$P(0,y) \;=\; \begin{cases} 1, & y \;=\; 0, \\ 0, & \text{otherwise.} \end{cases} \tag{7.4}$$

Proof. Formula (7.3) follows from the definition of branching process, specifically from (7.2):

$$X_{t+1} \;=\; \sum_{k\,=\,1}^{X_t} Y_k. \tag{7.5}$$

To obtain formula (7.4), just note that if $X_t = 0$, then the above sum is empty, so $X_{t+1} = 0$. □

After the offspring distribution p has been given, formulas (7.3)–(7.4) uniquely determine the entries of a infinite transition matrix P with rows and columns indexed by $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. The only problem is that computing numerical values of the entries of P can be difficult from (7.3). For example, to determine the entry $P(3,9)$ requires computing the complicated sum

$$P(3,9) \;=\; \sum_{y_1\,=\,0}^{\infty} \sum_{y_2\,=\,0}^{\infty} \sum_{y_3\,=\,0}^{\infty} \mathbb{1}\{y_1 + y_2 + y_3 = 9\} \cdot p(y_1) \cdot p(y_2) \cdot p(y_3).$$

Generating functions (Section 6) provide a powerful tool for treating such formulas. Indeed, if the branching process starts with n individuals, $X_0 = n$, then we see from Theorem 6.8 and (6.13)

$$\phi_{X_1}(s) \;=\; \phi_{Y_1+Y_2+\cdots+Y_n}(s) \;=\; (\phi_{Y_1}(s))^n. \tag{7.6}$$

Hence, the entry $P(3,9)$ of the transition matrix could by computed by writing the probability generating function $(\phi_{Y_1}(s))^3$ defined by (6.3) as a power series, and finding out the term corresponding to s^9:

$$(\phi_{Y_1}(s))^3 \;=\; \left(\sum_{k\,=\,0}^{\infty} p(k)\, s^k \right)^3 \;=\; \sum_{k_1\,=\,0}^{\infty} \sum_{k_2\,=\,0}^{\infty} \sum_{k_3\,=\,0}^{\infty} p(k_1) \cdot p(k_2) \cdot p(k_3) \cdot s^{k_1+k_2+k_3}.$$

Much more conveniently, recall that by formula (6.5) in Theorem 6.2,

$$\mathbb{P}\,(X_1 = k) \;=\; \frac{\phi_{X_1}^{(k)}(0)}{k!}, \qquad k = 0, 1, 2, \ldots,$$

so taking again $X_0 = 3$, the entry $P(3,9)$ can the obtained by differentiating $(\phi_{Y_1}(s))^3$ nine times at zero and dividing the outcome by the factorial of 9:

$$\mathbb{P}\,(Y_1 + Y_2 + Y_3 = 9 \mid X_0 = 3) \;=\; \frac{1}{9!} \Big(\frac{\mathrm{d}}{\mathrm{d}s} \Big)^9 (\phi_{Y_1}(s))^3 \Big|_{s=0}.$$

Example 7.2. If the offspring distribution is the Bernoulli distribution with values $\{0, 3\}$,

$$\mathbb{P}\,(Y_1 = 3) \;=\; q \qquad \text{and} \qquad \mathbb{P}\,(Y_1 = 0) \;=\; 1 - q,$$

then the probability generating function is $\phi_{Y_1}(s) = q\, s^3 + (1-q)$ (cf. Example 6.4), and we obtain

$$\mathbb{P}\,(Y_1 + Y_2 + Y_3 = 9 \mid X_0 = 3) \;=\; \frac{1}{9!} \Big(\frac{\mathrm{d}}{\mathrm{d}s} \Big)^9 (\phi_{Y_1}(s))^3 \Big|_{s=0} \;=\; \frac{1}{9!} \Big(\frac{\mathrm{d}}{\mathrm{d}s} \Big)^9 \big(q\, s^3 + (1-q) \big)^3 \Big|_{s=0} \;=\; q^3.$$

Could you find it just from the definition of the model? How about more complicated cases? ■

Theorem 7.3 (*Probability generating function*). *For branching process $X = (X_0, X_1, \ldots)$ starting with one individual $(X_0 = 1)$, the probability generating function of the t:th generation X_t is*

$$\phi_{X_t}(s) = \phi_{X_{t-1}}(\phi_{Y_1}(s)) = \underbrace{(\phi_{Y_1} \circ \phi_{Y_1} \circ \cdots \circ \phi_{Y_1})}_{t}(s). \tag{7.7}$$

Proof. Since $X_0 = 1$, we have $X_1 = Y_1$, so $\phi_{X_1} = \phi_{Y_1}$. More generally, by Theorem 6.9 the probability generating function (6.3) of the t:th generation (7.2) is

$$\phi_{X_t}(s) = \phi_{X_{t-1}}(\phi_{Y_1}(s)).$$

We prove the second equality in claim (7.7) by mathematical induction. The base case $t = 1$ is clear from $\phi_{X_1} = \phi_{Y_1}$. Also, if for some time instant $t \geq 1$, the claim is true:

$$\phi_{X_t}(s) = \underbrace{(\phi_{Y_1} \circ \phi_{Y_1} \circ \cdots \circ \phi_{Y_1})}_{t}(s),$$

then we know by Theorem 6.9 that

$$\begin{aligned}
\phi_{X_{t+1}}(s) &= \phi_{X_t}(\phi_{Y_1}(s)) && \text{[by Theorem 6.9]} \\
&= \underbrace{(\phi_{Y_1} \circ \phi_{Y_1} \circ \cdots \circ \phi_{Y_1})}_{t}(\phi_{Y_1}(s)) && \text{[by induction hypothesis]} \\
&= \underbrace{(\phi_{Y_1} \circ \phi_{Y_1} \circ \cdots \circ \phi_{Y_1})}_{t+1}(s),
\end{aligned}$$

so the second equality in claim (7.7) also holds for time instant $t + 1$. Thus, according to the induction principle, claim (7.7) holds for all $t \geq 0$. □

7.2 Expected population size

The following result helps to compute the expected population size as a function of time for branching process $X = (X_0, X_1, X_2, \ldots)$ with offspring distribution $p = \{p(k) : k = 0, 1, 2, \ldots\}$, where

$$\mathrm{m} = \mathbb{E}(Y_1) = \sum_{k=0}^{\infty} k \cdot p(k)$$

is the expected number of children produced by an individual distributed as $Y_1 \sim p$. As a consequence, we see that the population size tends to zero when $\mathrm{m} < 1$ and grows exponentially fast to infinity when $\mathrm{m} > 1$. We will discuss the case of $\mathrm{m} = 1$ in Section 7.4 (see Theorem 7.8.)

Theorem 7.4 (*Expected generation size*). *For branching process $X = (X_0, X_1, \ldots)$, the expected size of generation t is*

$$\mathbb{E}(X_t) = \mathbb{E}(X_0) \cdot \mathrm{m}^t, \qquad t = 0, 1, 2, \ldots, \tag{7.8}$$

where $\mathrm{m} = \mathbb{E}(Y_1)$. In particular, for a branching process started with $X_0 = n$ individuals, we have

$$\mathbb{E}(X_t) = n \cdot \mathrm{m}^t, \qquad t = 0, 1, 2, \ldots.$$

We can prove this easily using Wald's identity (6.14) from Theorem 6.10.

Proof. We prove claim (7.8) by mathematical induction. It is obviously true for $t = 0$:

$$\mathbb{E}(X_0) \;=\; \mathbb{E}(X_0) \cdot \mathrm{m}^0 \;=\; \mathbb{E}(X_0).$$

If the claim is true for some time instant $t \geq 0$, then using Theorem 6.10 with $N = X_t$, we obtain

$$\begin{aligned}\mathbb{E}(X_{t+1}) \;=\; \mathbb{E}\Big(\sum_{j=1}^{X_t} Y_j\Big) \;&=\; \mathbb{E}(X_t) \cdot \mathbb{E}(Y_1) \;=\; \mathbb{E}(X_t) \cdot \mathrm{m} \qquad \text{[by (6.14)]} \\ &=\; \mathbb{E}(X_0) \cdot \mathrm{m}^t \cdot \mathrm{m} \;=\; \mathbb{E}(X_0) \cdot \mathrm{m}^{t+1}, \quad \text{[by ind. hypo. } \mathbb{E}(X_t) = \mathbb{E}(X_0) \cdot \mathrm{m}^t]\end{aligned}$$

and hence, the claim also holds for time instant $t+1$. Thus, according to the induction principle, claim (7.8) holds for all $t \geq 0$. Lastly, if $X_0 = n$ is nonrandom, then of course $\mathbb{E}(X_0) = n$. □

7.3 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 $X_0 = 1$, and that the branches of the initial individuals are *mutually independent*. Therefore, if the initial generation contains $n \geq 1$ individuals, then *the probability of eventual extinction is the probability of all individual family lines becoming extinct.* This probability equals (as argued more precisely in the proof of Theorem 7.7)

$$\mathbb{P}(\text{extinction} \mid X_0 = n) \;=\; \big(\mathbb{P}(\text{extinction} \mid X_0 = 1)\big)^n \;=\; \eta^n, \tag{7.9}$$

$$\text{denoting} \qquad \eta \;=\; \mathbb{P}(\text{extinction} \mid X_0 = 1) \;=\; \mathbb{P}(T_0 < \infty \mid X_0 = 1), \tag{7.10}$$

the extinction probability of a branching process starting with one individual, $X_0 = 1$. In practise, the extinction probability η can be found as a *fixed point* of the probability generating function ϕ_{Y_1} of the offspring distribution, as Theorem 7.6 confirms.

Example 7.5 (*Binary tree population model*)**.** During its lifetime, each individual produces two children with probability q and no children otherwise. What is the probability that the family line of a particular individual eventually becomes extinct?

The offspring distribution is that of a Bernoulli random variable $Y \sim \mathrm{Ber}(q)$ with $\{0, 2\}$,

$$\mathbb{P}(Y = 2) \;=\; q \qquad \text{and} \qquad \mathbb{P}(Y = 0) \;=\; 1 - q.$$

Assume that $X_0 = 1$ so that we track the family line of a particular individual. Let us consider the two possible scenarios of the value of X_1:

- ▷ $X_1 = 0$, which happens if the number of children of the initial individual is zero (this has probability $(1 - q)$). Then, the family line becomes immediately extinct.
- ▷ $X_1 = 2$, which happens if the number of children of the initial individual is two (this has probability q). In this case, we need to investigate the family lines of these two children. Note that the whole process becomes eventually extinct if both family lines from these two individuals become eventually extinct. This has probability η^2. Thus, we find that

$$\eta \;=\; (1 - q) + q\,\eta^2.$$

We can recognize this as the probability generating function $\phi_Y(s) = q\,s^2 + (1-q)$ of the offspring distribution (Example 6.4). The fixed points of ϕ_Y are the solutions of

$$\phi_Y(s) \;=\; s \qquad \Leftrightarrow \qquad qs^2 \;-\; s \;+\; (1 - q) \;=\; 0$$

which we can solve easily:

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

Theorem 7.6 below says that the extinction probability is the *smaller* of these solutions:

$$\eta = \begin{cases} 1, & q \le 1/2, \\ \frac{1-q}{q}, & q > 1/2. \end{cases}$$

■

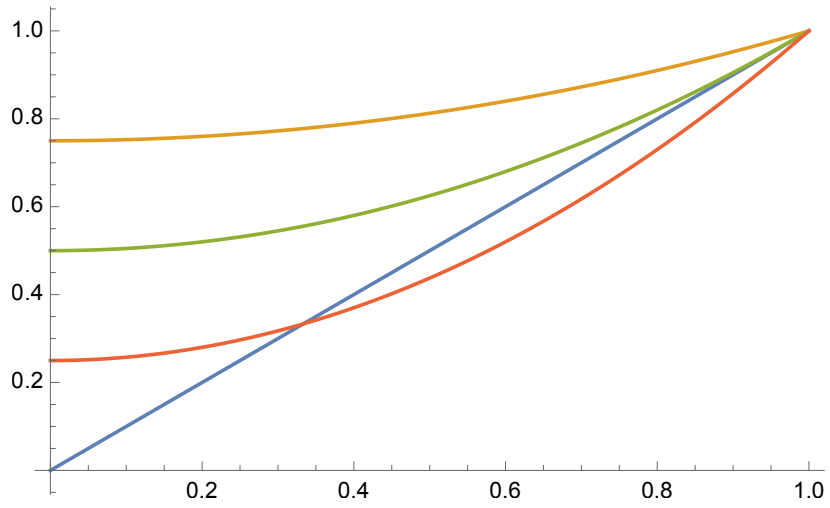


Figure 7.1: The linear function $s \mapsto s$ (blue) and examples of convex functions with various fixed points: $s \mapsto \frac{1}{4}s^2 + \frac{3}{4}$ (orange), $s \mapsto \frac{1}{2}s^2 + \frac{1}{2}$ (green), and $s \mapsto \frac{3}{4}s^2 + \frac{1}{4}$ (red), corresponding to the probability generating functions in Example 7.5 with $q = 1/4$, $q = 1/2$, and $q = 3/4$, respectively. Here, the fixed points of the functions are the intersections of them with the linear (blue) function. We see that on the interval $[0,1]$, the point 1 is always a fixed point, and the orange and green functions have no other fixed points (since $q \le 1/2$), while the red function has also the fixed point at $1/3$ (since $q > 1/2$, so $\frac{1-q}{q}$ is a fixed point). See also Theorem 7.8.

Theorem 7.6 (*Extinction probability*)**.** *For branching process $X = (X_0, X_1, \ldots)$ starting with one individual $(X_0 = 1)$, the extinction probability η is the smallest solution of*

$$\phi_{Y_1}(s) = s, \qquad s \in [0,1],$$

that is, we have $\phi_{Y_1}(\eta) = \eta$, and if $\phi_{Y_1}(a) = a$ for some $a \in [0,1]$, then $\eta \le a$.

In fact, the same proof also shows that η is the smallest *nonnegative* solution of $\phi_{Y_1}(s) = s$.

Proof. **Step 1.** We first verify that η is indeed a fixed point: $\phi_{Y_1}(\eta) = \eta$. We can write the event $\{T_0 < \infty\}$ defining η in (7.10) as

$$\{T_0 < \infty\} \;=\; \bigcup_{s=1}^{\infty} \{X_s = 0\},$$

since the branching process hits zero overall if and only if it hits zero at some time $s \geq 1$. Hence,

$$\eta \;=\; \mathbb{P}\Big(\bigcup_{s=1}^{\infty} \{X_s = 0\} \Big) \;=\; \lim_{t\to\infty} \mathbb{P}\Big(\bigcup_{s=1}^{t} \{X_s = 0\} \Big)$$

by continuity of probability measures[29]. Next, we observe that since 0 is an absorbing state,

$$\begin{aligned} X_s \;=\; 0 \qquad &\Longrightarrow \qquad X_t \;=\; 0, \qquad \text{for all } t \;\geq\; s \\ &\Longrightarrow \qquad \bigcup_{s=1}^{t} \{X_s = 0\} \;=\; \{X_t = 0\}, \end{aligned}$$

and we may write

$$\eta \;=\; \lim_{t\to\infty} \mathbb{P}\Big(\bigcup_{s=1}^{t} \{X_s = 0\} \Big) \;=\; \lim_{t\to\infty} \mathbb{P}\,(X_t = 0) \;=\; \lim_{t\to\infty} \eta(t),$$

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

Recall now from Theorem 6.2 that $\mathbb{P}\,(X_t = 0) = \phi_{X_t}(0)$, where $\phi_{X_t}(s)$ is the probability generating function from Theorem 7.3 of the t:th generation (7.2),

$$\phi_{X_t}(s) \;=\; \underbrace{(\phi_{Y_1} \circ \phi_{Y_1} \circ \cdots \circ \phi_{Y_1})}_{t}(s) \;=\; \phi_{Y_1}(\phi_{X_{t-1}}(s)).$$

Hence, we have

$$\eta(t) \;=\; \mathbb{P}\,(X_t = 0) \;=\; \phi_{X_t}(0) \;=\; \phi_{Y_1}(\eta(t-1)), \qquad t \geq 1. \tag{7.11}$$

Now, we can take the limit as $t \to \infty$ to obtain

$$\eta \;=\; \lim_{t\to\infty} \eta(t) \;=\; \lim_{t\to\infty} \phi_{Y_1}(\eta(t-1)) \;=\; \phi_{Y_1}\Big(\lim_{t\to\infty} \eta(t-1) \Big) \;=\; \phi_{Y_1}(\eta),$$

since $s \mapsto \phi_{Y_1}(s)$ is continuous by Lemma 6.7. Hence, $\phi_{Y_1}(\eta) = \eta$, as desired.

Step 2. We then prove that η is the smallest fixed point of ϕ_{Y_1} on $[0,1]$: if $\phi_{Y_1}(a) = a$ for some $a \in [0,1]$, then $\eta \leq a$. Since $\phi_{Y_1} : [0,1] \to [0,\infty)$ is non-decreasing by Lemma 6.7, we see that

$$\eta(1) \;=\; \mathbb{P}\,(X_1 = 0) \;=\; \phi_{X_1}(0) \;=\; \phi_{Y_1}(0) \;\leq\; \phi_{Y_1}(a) \;=\; a,$$

since $X_1 = Y_1$. Similarly, using (7.11) we have

$$\eta(2) \;=\; \phi_{Y_1}(\eta(1)) \;\leq\; \phi_{Y_1}(a) \;=\; a,$$

and we may inductively conclude that $\eta(t) \leq a$ for all $t \geq 1$. Hence, we obtain

$$\eta \;=\; \lim_{t\to\infty} \eta(t) \;\leq\; a,$$

which is what we sought to prove. □

[29] This is discussed in the course Probability theory (MS-E1600), see [Kyt20].

Theorem 7.7 (*Extinction probability with general initial condition*). *For branching process $X = (X_0, X_1, \ldots)$ with initial distribution $X_0 \sim \mu_0$, the extinction probability is*

$$\mathbb{P}(\text{extinction} \mid X_0 \sim \mu_0) = \phi_{X_0}(\eta),$$

where η is the extinction probability when starting with one individual $(X_0 = 1)$.

In particular, for a branching process started with $X_0 = n$ individuals, the extinction probability equals η^n, as we claimed in Equation (7.9).

Proof. By the law of total probability, we have

$$\mathbb{P}(\text{extinction}) = \sum_{n=0}^{\infty} \mathbb{P}(\text{extinction} \mid X_0 = n) \cdot \mathbb{P}(X_0 = n). \tag{7.12}$$

Each individual in the initial population starts an independent copy of the branching process. Hence, given $X_0 = n$, extinction of X happens exactly when each of the n independent branching processes with one initial individual go extinct, each having probability η. The extinction probability conditioned on $X_0 = n$ is thus given by

$$\mathbb{P}(\text{extinction} \mid X_0 = n) = \eta^n.$$

Plugging this into (7.12) gives

$$\mathbb{P}(\text{extinction}) = \sum_{n=0}^{\infty} \eta^n \cdot \mathbb{P}(X_0 = n) = \mathbb{E}(\eta^{X_0}) = \phi_{X_0}(\eta),$$

which is what we sought to prove. □

7.4 Sure extinction

Let us finally derive the following fundamental result: *a branching process can never reach a statistical equilibrium with a sustainable nonzero population size.* As before, we write

$$\mathrm{m} = \mathbb{E}(Y_1) = \sum_{k=0}^{\infty} k \cdot p(k), \qquad Y_1 \sim p,$$

for the expected number of children for an individual. Then, the only case where the population does *not* become eventually extinct is the one with expected number of children $\mathrm{m} > 1$, in which case the population even *grows to infinity* exponentially fast, according to Theorem 7.4. This is sometimes called a *Malthusian property*, after an English scholar Thomas Malthus (1766–1834).

Theorem 7.8 (*Growth of population*). *Assume that $p(0) = \mathbb{P}(Y_1 = 0) > 0$. Then, for branching process $X = (X_0, X_1, \ldots)$, the extinction probability η when starting with one individual $(X_0 = 1)$ satisfies*

- ▷ *$\eta = 1$, for $\mathrm{m} \le 1$, while*
- ▷ *$\eta \in (0, 1)$, for $\mathrm{m} > 1$.*

The case where $\mathrm{m} = 1$ is often called *critical*. On the one hand, we see from Theorem 7.4 that when $\mathrm{m} = 1$,

$$\mathbb{E}(X_t) = \mathrm{m}^t = 1, \qquad t = 0, 1, 2, \ldots.$$

On the other hand, Theorem 7.8 shows that the branching process surely becomes extinct also in the case of $\mathrm{m} = 1$. See also Figure 7.1 which shows some fixed points.

Proof sketch of Theorem 7.8. We know from Theorem 7.6 that η is the smallest fixed point of ϕ_{Y_1} on $[0,1]$. Recall the following properties of the probability generating function ϕ_{Y_1}:

- $\phi_{Y_1}(1) = 1$ (cf. Lemma 6.3),
- the map $s \mapsto \phi_Y(s)$ is convex (6.10) on $[0,1]$ (cf. Lemma 6.7),
- by (6.7), the expected number of children for an individual equals

$$m = \mathbb{E}(Y_1) = \phi'_Y(1-).$$

Hence, we can make the following observations.

- If $m \leq 1$, then by sketching a plot of ϕ_{Y_1} on the interval $[0,1]$ we see that ϕ_{Y_1} does not have any fixed points $[0,1)$ — so the smallest fixed point of ϕ_{Y_1} on $[0,1]$ is $\eta = 1$.
- If $m > 1$, then again by plotting ϕ_{Y_1} on the interval $[0,1]$ we see that ϕ_{Y_1} has precisely one fixed point on $(0,1)$. This fixed point is thus 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 $\phi_{Y_1}(s)$ around zero and around one. We leave this analysis for a mathematically oriented reader. □

8 Point processes, counting processes, and the Poisson process

We now begin to discuss stochastic processes in continuous time. *Point processes* in one dimension are just random particles in $\mathbb{R}$; and point processes in $[0,\infty) \subset \mathbb{R}$ can be thought of as collections of random time instants. The associated *counting processes*, which tell how many random time instants happened during a given time interval, are examples of stochastic processes *in continuous time*. One of the most important examples and a central topic in this course is the *Poisson process*, which is Markovian in a certain sense (cf. Theorems 8.7 and 8.13).

For more background and examples on Poisson processes, see, e.g., [Dur12, Chapter 2]. We refer the mathematically oriented readers to [SW08, Chapter 3] for more details, and [Kal21] for a general abstract mathematical framework.

8.1 Point processes and counting processes

Point processes are used in many applications: economics, epidemiology, materials science, neuroscience, spatial data analysis, telecommunications, as well as in abstract probability theory. In this course, we will consider the following types of random point collections on intervals $I \subset \mathbb{R}$ of the real line $\mathbb{R}$. As usual, the notation (s,t) refers to the *open interval* $\{x \in (0,1) : s < x < t\}$, and $(s,t]$ refers to the *half-open interval* $\{x \in (0,\infty) : s < x \le t\}$.

Definition. A *point process* (*satunnainen pistekuvio*) on an interval $I \subset \mathbb{R}$ is a locally finite[a] random subset of I.

[a] A subset X of an interval I is *locally finite* (*lokaalisti äärellinen*) if $X \cap K$ is finite whenever $K \subset I$ is closed and bounded (in particular, when $K = [a,b]$ is a closed interval).

In particular, *random time instants* related to a random phenomenon under study can be modeled as point processes on $(0,\infty)$. Note that the elements in a point process X are usually not independent. Their distribution is given by their joint probability distribution.

It is usually of interest to find out properties such as the average density of the point process in $\mathbb{R}$ or $(0,1]$. The *counting measure* (*laskurimitta*) of a point process X on $I \subset \mathbb{R}$ is a random function which returns the point count of X restricted to subsets of I:

$$N(B) = |X \cap B|, \qquad \text{for all } B \subset I.$$

For point processes on $(0,\infty)$, the point count on each interval $(0,t]$ provides the most important object of study in the scope of point processes in this course. We therefore give it a special name.

Definition. The *counting process* (*laskuriprosessi*) of point process X on $(0,\infty)$ is

$$N(t) = N((0,t]), \qquad t \in (0,\infty).$$

The definition implies that the point count of X in interval $(s,t]$ can be expressed in terms of the *increments* of the counting process N:

$$|X \cap (s,t]| = N((s,t]) = N(t) - N(s), \qquad \text{for all } 0 \le s < t.$$

The random function $t \mapsto N(t)$ is a *continuous-time stochastic process* with countable state space $\mathbb{N}_0 = \{0,1,2,\ldots\}$. We will study soon an analogue of the Markov property for such processes (see Section 8.5 and later Section 10). Let us first focus on some important examples.

Example 8.1 (*Uniform point process*)**.** Let $U_1, \ldots, U_n$ be independent and uniformly distributed random numbers on the interval $(0,1)$. Then, $X = \{U_1, \ldots, U_n\}$ is a point process on $I = (0,1)$, consisting of n random elements of $(0,1)$. Its counting process can be written as

$$N(t) = N((0,t]) = \sum_{j=1}^{n} \mathbb{1}\{U_j \le t\}, \qquad t \in (0,1).$$

■

Example 8.2 (*Poisson dominated points*)**.** Let $Z \sim \mathrm{Poi}(\lambda)$ be a random integer which follows a Poisson distribution with mean $\lambda > 0$, so that

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

Then, $X = \{k \in \mathbb{N}_0 : k \le Z\}$ is a point process on $I = (0,\infty)$. Its counting process can be written as

$$N(t) = N((0,t]) = |\{k \in \mathbb{N}_0 : k \le \min\{Z,t\}\}|, \qquad t \in (0,\infty).$$

■

The following example is very important in the theory of point processes, as we will see later: according to Theorem 8.6, the point process in Example 8.3 is in fact a *Poisson point process*, which is the most important continuous-time stochastic process discussed in this course.

Example 8.3 (*Exponential arrivals*)**.** Define random numbers $T_1, T_2, \ldots$ by the formula

$$T_n = \tau_1 + \tau_2 + \cdots + \tau_n, \qquad n \ge 1,$$

where $\tau_1, \tau_2, \ldots$ are iid exponentially distributed random numbers, $\tau_1 \sim \mathrm{Exp}(\lambda)$, so that

$$\mathbb{P}(\tau_1 \le t) = 1 - e^{-\lambda t}, \qquad t \in [0,\infty).$$

The parameter λ is usually called the *arrival rate* (*saapumistahti*). The mean (average) of the exponential random variable $\tau_1 \sim \mathrm{Exp}(\lambda)$ is inverse of the rate,

$$\mathbb{E}(\tau_1) = \frac{1}{\lambda}.$$

The numbers $X = \{T_1, T_2, \ldots\}$ form a point process on $I = (0,\infty)$, and λ is called the *intensity* (*intensiteetti*) of X. Its counting process can be written as

$$N(t) = N((0,t]) = \sum_{j=1}^{\infty} \mathbb{1}\{T_j \le t\}, \qquad t \in (0,\infty).$$

This process plays a crucial role in the next Section 8.2. ■

8.2 Poisson process and exponential waiting times

We now introduce the most important continuous-time stochastic process discussed in this course: the ubiquitous *Poisson process*. It is the counting process of a point process on $(0,\infty)$ with two special properties: homogeneity and independence (discussed precisely in Section 8.5). In fact, by Theorem 8.13 the Poisson process is the *only possible* process arising in this way.

Definition. Random function $N : (0,\infty) \to \mathbb{N}_0 = \{0,1,2,\ldots\}$ is a (homogeneous) *Poisson process* (*Poisson-prosessi*) with *intensity* (*intensiteetti*) $\lambda \in [0,\infty)$ if

1. $N(t) - N(s) \sim \mathrm{Poi}(\lambda(t-s))$ for all $(s,t] \subset (0,\infty)$, and
2. N has *independent increments* (*riippumattomat lisäykset*), that is,

$$N(t_1) - N(s_1),\ \ldots,\ N(t_k) - N(s_k)$$

are independent whenever $(s_1,t_1],\ \ldots,\ (s_k,t_k] \subset (0,\infty)$ are disjoint.

In particular, we have $\mathbb{E}(N(1)) = \lambda$ and

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

Note in particular that the intensity of a Poisson process is the expected number of points on the unit interval $(0,1]$ in the associated point process (called *Poisson point process*):

$$\mathbb{E}(N(1)) = \lambda.$$

As can be seen in Figure 8.1, the paths of a Poisson process are piecewise constant, and grow with unit jumps at random time instants $T_1, T_2, \ldots$. 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 T_n of Poisson process N can be written as

$$T_n = \min\{t \geq 0 : N(t) = n\}, \qquad n = 1,2,\ldots.$$

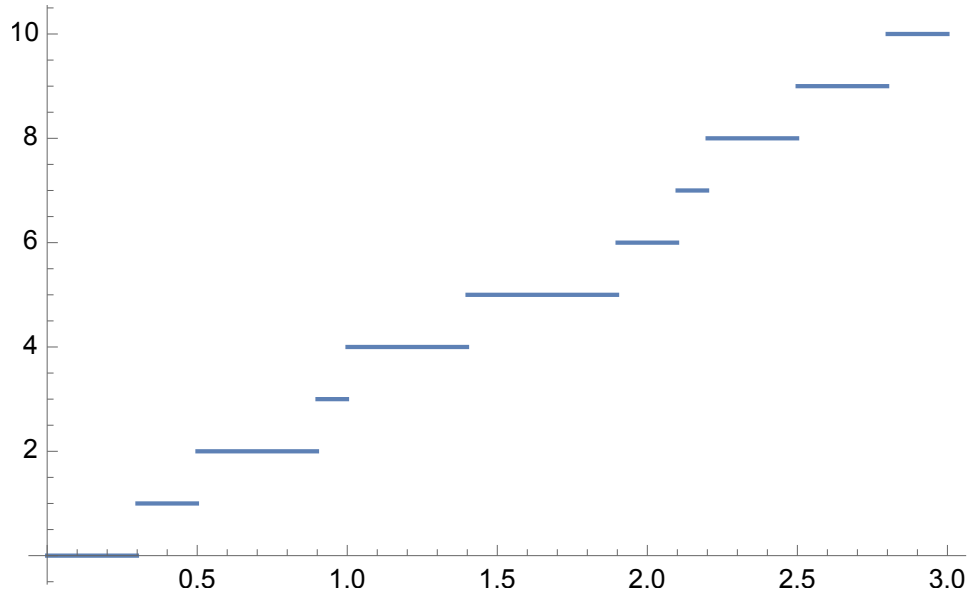


Figure 8.1: A typical sample path of a Poisson process.

Definition. Let N be a Poisson process with intensity λ. The collection $\{T_1, T_2, \ldots\}$ of jump instants forms a point process on $(0, \infty)$ with counting process

$$N(t) = \sum_{j=1}^{\infty} \mathbb{1}\{T_j \le t\}, \qquad t \in (0, \infty).$$

$\{T_1, T_2, \ldots\}$ is called a *Poisson point process* (*Poisson pistekuvio*) with intensity λ.

▷ The random variables $T_1, T_2, \ldots$ are often viewed as the *events* of the Poisson process, and then, the difference $N(t) - N(s)$ tells *the number of events* during 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[30].

▷ Conversely, the point process in Example 8.3 with exponentially distributed arrivals is in fact a Poisson point process, see Theorem 8.6.

Theorem 8.4 (*Jump times*). *For Poisson process N with intensity $\lambda \in (0, \infty)$,*

▷ *the* first jump instant

$$T_1 = \min\{t \ge 0 : N(t) = 1\}$$

is exponentially distributed with rate λ, that is, $T_1 \sim \mathrm{Exp}(\lambda)$, so that

$$\mathbb{P}(T_1 \le t) = 1 - e^{-\lambda t}, \qquad t \in (0, \infty),$$

▷ *the n:*th jump instant

$$T_n = \min\{t \ge 0 : N(t) = n\}$$

is Gamma distributed, that is, $T_n \sim \mathrm{Gamma}(n, \lambda)$, so that

$$\mathbb{P}(T_n \le t) = \int_0^t \frac{\lambda^n}{(n-1)!} u^{n-1} e^{-\lambda u} \, \mathrm{d}u, \qquad t \in (0, \infty). \tag{8.1}$$

Proof. We leave the proof as an exercise. □

Theorem 8.5 (*Waiting/arrival times*). *For Poisson process N with intensity $\lambda \in (0, \infty)$, the distances between jumps are iid and exponentially distributed with rate λ:*

$$T_n - T_{n-1} \sim \mathrm{Exp}(\lambda), \qquad \text{for all } n = 1, 2, 3, \ldots,$$

where we take $T_0 = 0$ by convention.

[30] This follows from the fact that distribution of T_n is continuous — in fact, T_n follows a Gamma distribution (Theorem 8.4). The distances $T_n - T_{n-1}$ between jump instants follow the exponential distribution (Theorem 8.5).

Proof. We leave it as an exercise to check that $T_n - T_{n-1} \sim \mathrm{Exp}(\lambda)$. We will see later in Theorem 11.2 that (conditioned on the event $\{N_0 = 0, N_{T_1} = 1, N_{T_2} = 2, \ldots, N_{T_{n-1}} = n-1\}$ which has probability one), the arrival times $T_n - T_{n-1}$ are independent for different $n \in \mathbb{N}_0$ (this follows since N is a continuous-time Markov process, as discussed in more detail in Chapters 10–11). □

Crucially, the converse of Theorem 8.5 also holds: the point process where the arrival times are independent and exponentially distributed (as in Example 8.3) is necessarily a Poisson point process. This is perhaps the most practical way to construct a Poisson point process.

Theorem 8.6 (*Exponential arrivals are counted by Poisson process*)**.** *Consider the point process $X = \{T_1, T_2, \ldots\}$ on $I = (0, \infty)$, where*

$$T_n = \tau_1 + \tau_2 + \cdots + \tau_n, \qquad n \geq 1,$$

and where $\tau_1, \tau_2, \ldots$ are iid exponentially distributed random numbers with rate λ, so that $\tau_1 \sim \mathrm{Exp}(\lambda)$. Then, the counting process of X,

$$N(t) = \sum_{j=1}^{\infty} \mathbb{1}\{T_j \leq t\}, \qquad t \in (0, \infty), \tag{8.2}$$

is a Poisson process with intensity λ.

Proof. We already know from Theorem 8.5 that, for any Poisson process N with intensity $\lambda \in (0, \infty)$, for the corresponding Poisson point process $\{T_1, T_2, \ldots\}$ the differences $\tau_n = T_n - T_{n-1}$, for $n = 1, 2, \ldots$, are iid and exponentially distributed, with $\tau_1 \sim \mathrm{Exp}(\lambda)$. Therefore, they coincide stochastically with those in Example 8.3, and hence the counting process (8.2) is N. □

8.3 Memoryless property of exponential distribution

An analogue of *Markov property* in the context of continuous-time processes is provided by the *memoryless property* of the times between jumps. We will get back to this in Sections 10–11.

Definition. We say that random variable T on $[0, \infty)$ satisfies the *memoryless property* (*muistittomuusominaisuus*) if

$$\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t), \qquad \text{for all } s, t \in [0, \infty). \tag{8.3}$$

In fact, the memoryless property is a special property that completely characterizes the exponential distribution (and thus gives a kind of Markovian nature for it):

Theorem 8.7 (*Memoryless property of exponential distribution*)**.** *Random variable T on $[0, \infty)$ is exponentially distributed with some rate parameter $\lambda \in [0, \infty)$, i.e., $T \sim \mathrm{Exp}(\lambda)$,* if and only if *$T$ satisfies memoryless property* (8.3).

Thus, *the times $T_n - T_{n-1} \sim \mathrm{Exp}(\lambda)$ between jumps for a Poisson process are uniquely characterized by the memoryless property.* However, note that the jump instants T_n themselves, other than T_1, are not memoryless — the Gamma distribution (8.1) does not enjoy the property (8.3). Indeed, knowing that $T_n > s$, it is more likely that $T_n > s + t$. Hence, we have

$$\mathbb{P}(T_n > s + t \mid T_n > s) \geq \mathbb{P}(T_n > s + t), \qquad s, t \in [0, \infty).$$

Proof of Theorem 8.7. We leave it as an exercise to verify that $T \sim \mathrm{Exp}(\lambda)$ satisfies (8.3).

Conversely, suppose that random variable T satisfies memoryless property (8.3). Then, the tail distribution function $H(t) = \mathbb{P}(T > t)$ satisfies the multiplicativity

$$H(t+s) \;=\; H(t) \cdot H(s), \qquad \text{for all } s,t \in [0,\infty). \tag{8.4}$$

Because H is nonincreasing, it follows by the theory of Cauchy's functional equations that H must have the form

$$H(t) \;=\; e^{-\lambda t}, \qquad \text{for some } \lambda \in [0,\infty).$$

▷ When $\lambda > 0$, this shows that the random variable T is $\mathrm{Exp}(\lambda)$-distributed.

▷ When $\lambda = 0$, it follows that

$$\mathbb{P}(T = \infty) \;=\; \lim_{n\to\infty} \mathbb{P}(T > n) \;=\; 1,$$

which corresponds to an exponential distribution with rate parameter zero.

□

Example 8.8 (*Triathlon race*)**.** Consider two competitors Y_1, Y_2 in a triathlon race, whose arrival times to the finish line are independent and exponentially distributed:

$$Y_1 \;\sim\; \mathrm{Exp}(\lambda_1) \qquad \text{and} \qquad Y_2 \;\sim\; \mathrm{Exp}(\lambda_2).$$

▷ What is the probability that Y_1 wins the race?

▷ What is the winning time $T = \min\{Y_1, Y_2\}$?

Recall that the exponential distribution has density $f(t) = \lambda e^{-\lambda t}$. Using this, because Y_1, Y_2 are independent, we can compute

$$\begin{aligned}
\mathbb{P}(Y_1 < Y_2) \;&=\; \int_0^\infty \Big(\int_0^t \lambda_1 e^{-\lambda_1 s}\, \mathrm{d}s \Big) \cdot \lambda_2 e^{-\lambda_2 t}\, \mathrm{d}t \\
&=\; \int_0^\infty \big(1 - e^{-\lambda_1 t}\big) \cdot \lambda_2 e^{-\lambda_2 t}\, \mathrm{d}t \\
&=\; \frac{\lambda_1}{\lambda_1 + \lambda_2}.
\end{aligned}$$

Because Y_1, Y_2 are independent, we can also compute

$$\mathbb{P}(T > t) \;=\; \mathbb{P}(Y_1 > t,\, Y_2 > t) \;=\; \mathbb{P}(Y_1 > t) \cdot \mathbb{P}(Y_2 > t) \;=\; e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} \;=\; e^{-(\lambda_1+\lambda_2)t}.$$

This shows that $T \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$. This property similarly holds for more than two exponential random variables (see Theorem 9.10 in Section 9.4). ■

8.4 Binomial approximation of Poisson distribution

The next lemma says that Poisson random variables can be approximated by binomial random variables. Its proof is meant for additional information for students majoring in mathematics, and can also be skipped at the first reading. (We use it in the proof of Theorem 8.13.)

Lemma 8.9 (*Law of small numbers*)**.** *For each $n \in \mathbb{N}$, let Z_n be a $\mathrm{Bin}(n, q_n)$-distributed random integer, and assume that $n\,q_n \longrightarrow \alpha \in (0,\infty)$ as $n \to \infty$. Then, the following convergence holds*:

$$\lim_{n\to\infty} \mathbb{P}(Z_n = k) \;=\; e^{-\alpha}\, \frac{\alpha^k}{k!}, \qquad \text{for all } k = 0, 1, 2, \ldots.$$

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

$$\begin{aligned} \mathbb{P}(Z_n = k) &= \frac{n!}{k!(n-k)!}(1-q_n)^{n-k}q_n^k \\ &= \frac{n!}{n^k(n-k)!}\frac{1}{(1-q_n)^k}\frac{(nq_n)^k}{k!}\Big(1-\frac{nq_n}{n}\Big)^n. \end{aligned} \tag{8.5}$$

We will analyze the right-hand side of (8.5) as $n \to \infty$. The first term satisfies

$$\frac{n!}{n^k(n-k)!} = \frac{1}{n^k}\prod_{j=0}^{k-1}(n-j) = \prod_{j=0}^{k-1}(1-j/n) \quad \overset{n\to\infty}{\longrightarrow} \quad 1.$$

Because $q_n \to 0$, also the second term on the right side of (8.5) satisfies

$$\frac{1}{(1-q_n)^k} \quad \overset{n\to\infty}{\longrightarrow} \quad 1.$$

Furthermore, the assumption $n\,q_n \to \alpha$ implies that the third term on the right of (8.5) scales as

$$\frac{(nq_n)^k}{k!} \quad \overset{n\to\infty}{\longrightarrow} \quad \frac{\alpha^k}{k!}.$$

Hence, the claim follows after verifying that

$$\lim_{n\to\infty}\left(1-\frac{nq_n}{n}\right)^n = e^{-\alpha},$$

which we leave as an exercise for an interested reader to check. □

8.5 Homogeneous and independent scattering

A useful notion of independence for point processes is that information about the points of X within a set A is irrelevant when predicting how the point process behaves outside A. Such *independent scattering* is a very restrictive assumption, which only few point processes satisfy.

Definition. Point process X is *independently scattered* (*riippumattomasti sironnut*) if random variables $N(A_1), \ldots N(A_k)$ are independent whenever sets $A_1, \ldots, A_k$ are disjoint.

Example 8.10. Is the point process $X = \{U_1, \ldots, U_n\}$ of Example 8.1 independently scattered? The answer is *no*, because the number of points in a subinterval of $(0, 1)$ depends on the number of points in its complement. Indeed, by dividing the open unit interval into intervals $A_1 = (0, 1/2]$ and $A_2 = (1/2, 1)$, we see that

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

while the corresponding conditional probability given $\{N(A_2) = n\}$ equals

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

because by definition, the equation $N(A_1) + N(A_2) = n$ surely holds. ■

Example 8.11. Is the point process of Example 8.2 independently scattered? ■

Example 8.12. The point process of Example 8.3 is independently scattered. ■

The following result characterizes how independent scattering, an intrinsically algebraic property, automatically yields a quantitative description of the distribution of point counts of the point process. The result also underlines the central role of the *Poisson distribution as a universal distribution* describing point counts of independently scattered point processes.

Definition. Point process X on $(0,\infty)$ is *homogeneous* (*tasakoosteinen*) if its counting process satisfies

$$N(B+t) \sim N(B)$$

for all sets $B \subset (0,\infty)$ and for all $t \in [0,\infty)$, where $B+t = \{x+t : x \in B\}$.

The *intensity* (*intensiteetti*) of homogeneous point process X is the expected point count $\mathbb{E}(N((0,1])$ on the unit interval $(0,1]$.

For a homogeneous and independently scattered point process X on $(0,\infty)$, any interval $I \subset (0,\infty)$ can have arbitrarily many points and the points in X can have an arbitrarily long distance. There is essentially only one type of such processes, as the following result verifies.

Theorem 8.13 (*Homogeneous independently scattered must be Poisson point process*)**.** *The counting process $N(t) = N((0,t])$ of a homogeneous independently scattered point process X is a* Poisson process *with intensity $\lambda = \mathbb{E}(N((0,1]))$.*

The proof can be skipped for the first reading.

Proof sketch. **Step 1.** It follows immediately from the assumption that X is a homogeneous independently scattered point process that its counting process N has independent increments.

Step 2. We first show that the probability $t \mapsto H(t) = \mathbb{P}(N(t) = 0)$ that there are no points of X in the interval $(0,t]$ satisfies the multiplicativity property (8.4). Indeed, because

$$N(0,s+t] = 0 \quad \Longleftrightarrow \quad N(0,s] = 0 \quad \text{and} \quad N(s,s+t] = 0,$$

we see that

$$\begin{aligned} H(s+t) &= \mathbb{P}(N(0,s+t] = 0) \\ &= \mathbb{P}(N(0,s] = 0,\ N(s,s+t] = 0) \\ &= \mathbb{P}(N(0,s] = 0) \cdot \mathbb{P}(N(s,s+t] = 0) && \text{[by independence]} \\ &= \mathbb{P}(N(0,s] = 0) \cdot \mathbb{P}(N(0,t] = 0) && \text{[by homogeneity]} \\ &= H(s) \cdot H(t). \end{aligned}$$

Because H is nonincreasing, it follows by the theory of Cauchy's functional equations that H must have the form

$$H(t) = e^{-\alpha t}, \qquad \text{for some } \alpha \in [0,\infty). \tag{8.6}$$

Step 3. Write $q_n = 1 - H(t/n)$. We next prove that

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

The idea is that

$$\mathbb{P}(N(t) = k) \approx \mathbb{P}(Z_n = k),$$

where $n \geq k$ is a large number and Z_n is a $\mathrm{Bin}(n, q_n)$-distributed random integer, as in Lemma 8.9.

To see this, we divide the interval $(0, t]$ into equally sized subintervals $I_{n,j} = (\frac{j-1}{n}t, \frac{j}{n}t]$ of length t/n, with $j = 1, \ldots, n$. Consider the indicator random variables

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

Then, $Z_n = \theta_1 + \cdots + \theta_n$ is the number of subintervals which contains points of X. Due to independent scattering, $\theta_1, \ldots, \theta_n$ are iid, and they have the Bernoulli distribution

$$\mathbb{P}(\theta_1 = 1) = q_n \qquad \text{and} \qquad \mathbb{P}(\theta_1 = 0) = 1 - q_n,$$

with parameter q_n. Hence, by Theorem 6.8 and identity (6.13), the probability generating function of $Z_n = \theta_1 + \cdots + \theta_n$ is (recalling Example 6.4)

$$\phi_{Z_n}(s) = \big(q_n s + (1 - q_n)\big)^n = \sum_{k=0}^{n} \binom{n}{k} q_n^k (1 - q_n)^{n-k} s^k,$$

which implies that $Z_n \sim \mathrm{Bin}(n, q_n)$ by Theorem 6.2. Note that by (6.5), we have

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

Denote by E_n the event that each subinterval contains at most one point. Now, on the event E_n, we have $N(t) = Z_n$, which implies that

$$\mathbb{P}(N(t) = k) = \mathbb{P}(Z_n = k) + \varepsilon_n, \tag{8.7}$$

where $\varepsilon_n = \mathbb{P}(N(t) = k,\, E_n^c) - \mathbb{P}(Z_n = k,\, E_n^c)$. To conclude, we gather the following facts.

▷ The right-hand side of (8.7) converges as

$$\mathbb{P}(Z_n = k) \quad \overset{n\to\infty}{\longrightarrow} \quad e^{-\alpha}\frac{\alpha^k}{k!}. \tag{8.8}$$

This follows from Lemma 8.9 with the limit

$$n\,q_n = n\,(1 - e^{-\alpha t/n}) = \frac{1 - e^{-\alpha t/n}}{1/n} \quad \overset{n\to\infty}{\longrightarrow} \quad \alpha\, t$$

(to see this, use Equation (8.6) and l'Hôpital's rule).

▷ We have $\varepsilon_n \longrightarrow 0$ as $n \to \infty$ (this is Lemma 8.14).

From these facts, we see that that

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

We may thus conclude that $N(t)$ is Poisson distributed with mean $\alpha\, t = \lambda\, t$. □

The next lemma is a technical tool for proving Theorem 8.13. It is meant for additional information for students majoring in mathematics, and can also be skipped at the first reading.

Lemma 8.14. *Let X be a point process on interval $J \subset \mathbb{R}$ with counting measure N. Divide the real axis $\mathbb{R}$ into intervals $I_{n,j} = (\frac{j-1}{n}, \frac{j}{n}]$ of length $1/n$, indexed by integers $j \in \mathbb{Z}$. Then, for any interval $J \subset I$ such that $\mathbb{E}(N(J)) < \infty$, we have*

$$\lim_{n\to\infty} \mathbb{P}\big(N(J \cap I_{n,j}) \le 1 \text{ for all } j \in \mathbb{Z}\big) = 1.$$

Proof. Consider the random number

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

which is the smallest distance of two points in the point process 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, on the event $\{D > 1/n\}$, we have

$$Z_n = \sup_j N(A \cap I_{n,j}) = \sup_j |X \cap A \cap I_{n,j}| \le 1. \tag{8.9}$$

The assumption $\mathbb{E}(N(A)) < \infty$ implies that the set $X \cap A$ is finite with probability one. Hence, we have $D > 0$ with probability one, and the above inequality (8.9) shows that

$$\lim_{n\to\infty} \mathbb{1}\{Z_n \le 1\} = 1$$

with probability one. Now by applying Lebesgue's dominated convergence theorem [31] to justify interchanging the limit and the expectation below, it follows that

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le 1) = \lim_{n\to\infty} \mathbb{E}\big(\mathbb{1}(Z_n \le 1)\big) = \mathbb{E}\Big(\lim_{n\to\infty} \mathbb{1}(Z_n \le 1)\Big) = 1,$$

which is what we sought to prove. □

[31] See [Kyt20, Thm. VII.22].

9 Variants of Poisson processes

9.1 Superposed Poisson processes

The following theorem confirms the intuitively natural fact that by superposing several mutually independent Poisson processes we obtain again a Poisson process. This can be handily proven using generating functions from Section 6.

Example 9.1. Consider independent Poisson distributed random variables $Y_1, Y_2, \dots, Y_n$ with parameters $\lambda_1, \lambda_2, \dots, \lambda_n$. What is the distribution of the sum

$$Y = \sum_{j=1}^{n} Y_j \quad ? \tag{9.1}$$

From Example 6.6, we know that the probability generating functions of the summands are $\phi_{Y_j}(s) = e^{\lambda_j(s-1)}$, for $j = 1, 2, \dots, n$. Using Theorem 6.8 we easily compute the probability generating function of the sum (9.1):

$$\phi_Y(s) = \phi_{Y_1}(s) \cdot \phi_{Y_2}(s) \cdots \phi_{Y_n}(s) = e^{\lambda_1(s-1)} \cdot e^{\lambda_2(s-1)} \cdots e^{\lambda_n(s-1)} = e^{(\lambda_1+\lambda_2+\cdots+\lambda_n)(s-1)}.$$

This shows that Y follows the Poisson distribution with intensity $\lambda = \sum_{j=1}^{n} \lambda_j$. ■

Theorem 9.2 (*Superposed Poisson processes*)**.** *If $N_1, N_2, \dots, N_n$ are independent Poisson processes with intensities $\lambda_1, \lambda_2, \dots, \lambda_n$, then*

$$N(t) = \sum_{j=1}^{n} N_j(t), \qquad t \in (0, \infty), \tag{9.2}$$

is a Poisson process with intensity $\lambda = \sum_{j=1}^{n} \lambda_j$.

In the sum above, the index set could also be countably infinite. In that case, the same proof works, but we need to assume that $\sum_{j=1}^{\infty} \lambda_j < \infty$.

Proof. Let us verify the three conditions in the definition from Section 8.2 for the process (9.2).

▷ First, we prove that $N(t) - N(s) \sim \mathrm{Poi}(\lambda(t-s))$ for all $(s,t] \subset (0,\infty)$. Indeed, recall from Example 9.1 that the sum of Poisson random variables is also Poisson distributed with parameter being the sum of the parameters. Hence, we have

$$N(t) - N(s) = \sum_{j=1}^{n} \underbrace{(N_j(t) - N_j(s))}_{\sim \mathrm{Poi}(\lambda_j(t-s))} \sim \mathrm{Poi}(\underbrace{(\lambda_1 + \lambda_2 + \cdots + \lambda_n)}_{=\lambda}(t-s)).$$

▷ Second, we prove that N has *independent increments*, that is,

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

are independent whenever $(s_1, t_1], \ \dots, \ (s_n, t_n] \subset (0,\infty)$ are disjoint. To see this, we use the following two facts.

⋆ Each part $N_j(t)$ in (9.2) is a Poisson process, so its increments

$$N_j(t_1) - N_j(s_1), \ \dots, \ N_j(t_k) - N_j(s_k)$$

are independent.

* Each $N_j(t)$ is independent of the other processes $N_i(t)$ for $i \neq j$, so their increments are mutually independent.

Now, the increments of the process (9.2) have the form

$$N(t_i) - N(s_i) = \sum_{j=1}^{n} (N_j(t_i) - N_j(s_i)), \qquad i = 1, 2, \ldots, k,$$

so from the above two facts, we see that they are independent for all $i = 1, 2, \ldots, k$.

This proves that (9.2) is a Poisson process with intensity $\lambda = \sum_{j=1}^{n} \lambda_j$. □

9.2 Compound Poisson processes

Poisson process $N(t)$ models the number of homogeneous and independently scattered random time instants during times $(0, t]$. If the time instants are generated as a superposition of many sparse event sequences, then the net counting process can also be quite accurately modeled as a Poisson process (called *compound* 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), etc., are not too big.

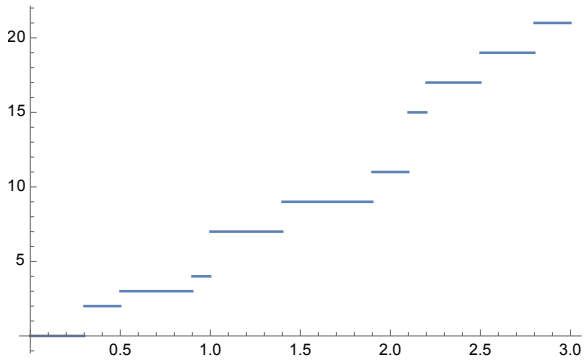


Figure 9.1: A sample path of a compound Poisson process.

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 9.3 (*Traffic flow*)**.** The average flow of cars crossing the Helsinki–Espoo border on Länsiväylä during weekdays equals $\lambda = 40$ cars per minute, and the average number of people per car is m = 1.9 with an estimated standard deviation of $\sigma = 1.2$ (variance σ^2). We 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 people crossing the border per hour.

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

$$R(t) = \sum_{j=1}^{\infty} Z_j \cdot \mathbb{1}\{T_j \leq t\} = \sum_{j=1}^{N(t)} Z_j, \qquad t \in (0, \infty),$$

where N is the counting process of the time instants $\{T_1, T_2, \ldots\}$, and $\{Z_1, Z_2, \ldots\}$ are random jump sizes at these times – see Figure 9.1. In this example, we assume that the times $\{T_1, T_2, \ldots\}$

form a Poisson point process of intensity $\lambda = 40$, and we require that the random jump sizes Z_j take values in set $S = \{1, 2, 3, 4, 5\}$, and have expected value $\mathbb{E}(Z_j) = \mathrm{m}$ and variance $\mathrm{var}(Z_j) = \sigma^2$, where $\mathrm{m} = 1.9$ and $\sigma = 1.2$. We will analyze this model further in Example 9.5. ■

In general, we can add randomness to point process $X = \{T_1, T_2, \ldots\}$ on $(0, \infty)$ by defining a new point process on $(0, \infty) \times S$,

$$\tilde{X} \;=\; \big\{(T_1, Z_1), (T_2, Z_2), \ldots\big\},$$

where $Z_1, Z_2, \ldots$ are random variables with values in some state space S. When the random variables Z_j are real-valued, we may view them as a cost (or reward) at time instant T_j. Then, the net reward up to time t can be written as

$$R(t) \;=\; \sum_{j=1}^{\infty} Z_j \cdot \mathbb{1}\{T_j \le t\} \;=\; \sum_{j=1}^{N(t)} Z_j, \qquad t \in (0, \infty),$$

where $N(t)$ is the counting process of the time instants $\{T_1, T_2, \ldots\}$,

$$N(t) \;=\; \sum_{j=1}^{\infty} \mathbb{1}\{T_j \le t\}, \qquad t \in (0, \infty).$$

Definition. When N is a Poisson process with intensity λ and the costs $Z_1, Z_2, \ldots$ are iid and also independent of N, then the stochastic process

$$R(t) \;=\; \sum_{j=1}^{N(t)} Z_j, \qquad t \in (0, \infty), \tag{9.3}$$

is called a *compound Poisson process* (*yhdistetty Poisson-prosessi*). The random variables $\{Z_1, Z_2, \ldots\}$ are sometimes called *weights* (*painot*), or *jump sizes*, of R.

Theorem 9.4 (*Compound Poisson process*)**.** *A compound Poisson process R has independent increments, and the mean and variance at time $t \in (0, \infty)$ can be computed as*

$$\begin{aligned} \mathbb{E}(R(t)) \;&=\; \lambda\,\mathrm{m}\,t, \\ \mathrm{var}(R(t)) \;&=\; \lambda\,(\mathrm{m}^2 + \sigma^2)\,t, \end{aligned} \tag{9.4}$$

where $\mathrm{m} = \mathbb{E}(Z_1)$ and $\sigma^2 = \mathrm{var}(Z_1)$, and $\lambda \in [0, \infty)$ is the intensity of R.

Proof sketch. The independence of increments is intuitively clear, since the weights are independent[32]. The values $N(t)$ the counting process are Poisson distributed with mean λt, so

$$\mathbb{E}(N(t)) \;=\; \mathrm{var}(N(t)) \;=\; \lambda t,$$

(e.g., by Example 6.6). Hence, the first claim in (9.4) follows from Theorem 6.10:

$$\mathbb{E}(R(t)) \;=\; \mathbb{E}\Big(\sum_{j=1}^{N(t)} Z_j\Big) \;=\; \mathbb{E}(N(t)) \cdot \mathbb{E}(Z_1) \;=\; \lambda\,\mathrm{m}\,t.$$

[32] Proving this rigorously can be done by conditioning on events of the form $\{N(s_k) = m_k,\ N(t_k) = m_k + r_k\}$.

The second claim in (9.4) can be proven similarly: we first differentiate both sides of the identity for the probability generating functions from Theorem 6.9,

$$\phi_{R(t)}(s) \;=\; \phi_{N(t)}(\phi_{Z_1}(s)),$$

at $s = 1$, and use the chain rule and the value $\phi_{Z_1}(1) = 1$ to obtain

$$\phi''_{R(t)}(1) \;=\; \phi''_{N(t)}(1) \cdot (\phi'_{Z_1}(1))^2 \;+\; \phi'_{N(t)}(1) \cdot \phi''_{Z_1}(1).$$

Recalling also the formula $\phi'_Y(1) = \mathbb{E}(Y)$ from Lemma 6.3, we see that

$$\phi''_{R(t)}(1) \;=\; \mathrm{m}^2 \cdot \phi''_{N(t)}(1) \;+\; \lambda t \cdot \phi''_{Z_1}(1).$$

By Lemma 6.3, the variance of $R(t)$ is

$$\begin{aligned} \operatorname{var}(R(t)) \;&=\; \phi''_{R(t)}(1) \;+\; \phi'_{R(t)}(1) \;-\; (\phi'_{R(t)}(1))^2 \\ &=\; \phi''_{R(t)}(1) \;+\; \mathbb{E}(R(t)) \;-\; (\mathbb{E}(R(t)))^2 \\ &=\; \phi''_{R(t)}(1) \;+\; \lambda \mathrm{m} t \;-\; (\lambda \mathrm{m} t)^2, \end{aligned}$$

the variance of $N(t)$ is

$$\begin{aligned} \lambda t \;=\; \operatorname{var}(N(t)) \;&=\; \phi''_{N(t)}(1) \;+\; \phi'_{N(t)}(1) \;-\; (\phi'_{N(t)}(1))^2 \\ &=\; \phi''_{N(t)}(1) \;+\; \mathbb{E}(N(t)) \;-\; (\mathbb{E}(N(t)))^2 \\ &=\; \phi''_{N(t)}(1) \;+\; \lambda t \;-\; (\lambda t)^2, \end{aligned}$$

which shows in particular that $\phi''_{N(t)}(1) = (\lambda t)^2$, and the variance of Z_1 is

$$\begin{aligned} \sigma^2 \;=\; \operatorname{var}(Z_1) \;&=\; \phi''_{Z_1}(1) \;+\; \mathbb{E}(Z_1) \;-\; (\mathbb{E}(Z_1))^2 \\ &=\; \phi''_{Z_1}(1) \;+\; \mathrm{m} \;-\; \mathrm{m}^2, \end{aligned}$$

which shows in particular that $\phi''_{Z_1}(1) = \sigma^2 - \mathrm{m} + \mathrm{m}^2$. Putting these together yields the claimed

$$\begin{aligned} \operatorname{var}(R(t)) \;&=\; \phi''_{R(t)}(1) \;+\; \lambda \mathrm{m} t \;-\; (\lambda \mathrm{m} t)^2 \\ &=\; \big(\mathrm{m}^2 \cdot \phi''_{N(t)}(1) \;+\; \lambda t \cdot \phi''_{Z_1}(1)\big) \;+\; \lambda \mathrm{m} t \;-\; (\lambda \mathrm{m} t)^2 \\ &=\; \big(\mathrm{m}^2 \cdot (\lambda t)^2 \;+\; \lambda t \cdot (\sigma^2 - \mathrm{m} + \mathrm{m}^2)\big) \;+\; \lambda \mathrm{m} t \;-\; (\lambda \mathrm{m} t)^2 \\ &=\; \lambda (\mathrm{m}^2 + \sigma^2)\, t. \end{aligned}$$

□

Example 9.5 (*Traffic flow*)**.** Let us come back to Example 9.3. By Theorem 9.4, at time instant $t = 60$, we have

$$\mathbb{E}(R(t)) \;=\; \lambda \mathrm{m} t \;=\; 40 \cdot 1.9 \cdot 60 \;=\; 4560$$

and

$$\operatorname{var}(R(t)) \;=\; \lambda(\mathrm{m}^2 + \sigma^2)\, t \;=\; 40 \cdot (1.9^2 + 1.2^2) \cdot 60 \;=\; 12120.$$

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

9.3 Thinned Poisson processes

In Section 9.1 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. This is also a special case of a compound Poisson process, with costs taking values in $\{0, 1\}$, which corresponds to omitting some of the events. This is commonly referred to as *thinning*.

Definition. A *thinned Poisson process* (*harvennettu Poisson-prosessi*) is a compound Poisson process

$$N_1(t) = \sum_{j=1}^{N(t)} \theta_j, \qquad t \in (0, \infty), \tag{9.5}$$

where the weights $\theta_1, \theta_2, \ldots$ are iid Bernoulli random variables with values in $\{0, 1\}$, independent of the original Poisson process N.

Example 9.6 (*Thinned traffic flow*)**.** The average flow of cars crossing the Helsinki–Espoo border on Länsiväylä highway during weekdays equals $\lambda = 40$ cars per minute. Of these cars, $p = 30\%$ contain only one person. What is the probability that during a particular minute, at most 20 cars contain only one person, given that at least 30 cars contain more than one person?

We build a statistical model for the traffic using Poisson processes, taking into account the number of people in each car. For each time interval $[0, t]$, we denote

▷ the total number of cars by

$$N(t) = \sum_{j=1}^{\infty} \mathbb{1}\{T_j \le t\},$$

▷ the number of cars containing only one person by

$$N_1(t) = \sum_{j=1}^{\infty} \theta_j \cdot \mathbb{1}\{T_j \le t\} = \sum_{j=1}^{N(t)} \theta_j, \tag{9.6}$$

▷ and the number of cars containing more than one person by

$$N_2(t) = \sum_{j=1}^{\infty} (1 - \theta_j) \cdot \mathbb{1}\{T_j \le t\} = \sum_{j=1}^{N(t)} (1 - \theta_j), \tag{9.7}$$

where $\theta_j \in \{0, 1\}$ is the indicator random variable for the event that the j:th car contains only one person.

If we assume that $\theta_1, \theta_2, \ldots$ are independent, the counting process $N_1(t)$ thus obtained is a thinned Poisson process, which is obtained by removing 70% of the events of the original Poisson process $N(t)$ by independent sampling (see Theorem 9.7). Analogously, also $N_2(t)$ is a thinned Poisson process. We will return to this model in Example 9.8. ■

The following result confirms that independently thinned Poisson processes are Poisson processes — and more strikingly, the thinned processes are *mutually independent*.

Theorem 9.7 (*Thinned Poisson processes*). *Let N be a Poisson process with intensity λ, and let $\theta_1, \theta_2, \ldots$ be a collection of iid Bernoulli random variables with values in $\{0,1\}$, independent of N, and satisfying $\theta_1 \sim \mathrm{Ber}(p)$. Then, the thinnings (9.6, 9.7) are mutually independent Poisson processes, with intensities $p\,\lambda$ and $(1-p)\,\lambda$, respectively.*

Proof. **Step 1.** We first show that N_1 is a Poisson process. The same argument also shows that N_2 is a Poisson process. We will verify the three conditions in the definition from Section 8.2.

▷ First, we prove that $N_1(t) - N_1(s) \sim \mathrm{Poi}(p\,\lambda)$ for all $(s,t] \subset (0,\infty)$. Recall that

  * the probability generating function of $N(t)$ is $\phi_{N(t)}(s) = e^{\lambda t(s-1)}$ (Example 6.6),
  * the probability generating function of θ_1 is $\phi_{\theta_1}(s) = p\,s + (1-p)$ (Example 6.4).

  Hence, we may apply Theorem 6.9 to conclude that

$$\phi_{N_1(t)}(s) \;=\; \phi_{N(t)}(\phi_{\theta_1}(s)) \;=\; \exp\big(\lambda t\,(p\,s + (1-p) - 1)\big) \;=\; \exp\big(\lambda t p\,(s-1)\big),$$

  which implies that $N_1(t) \sim \mathrm{Poi}(p\lambda t)$. In precisely the same way, we can verify that its increments satisfy $N_1(t) - N_1(s) \sim \mathrm{Poi}(p\lambda(t-s))$.

▷ Second, note that because N_1 is a compound Poisson process, it follows from Theorem 9.4 that N_1 has independent increments. Hence, N_1 is a Poisson process with intensity $p\,\lambda$.

Step 2. It remains to verify that N_1 and N_2 are independent. Note that 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 that

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

Hence, we see that the random variables $N_1(t)$ and $N_2(t)$ are independent for every time t. This argument can be generalized in a straightforward manner to show that the random vectors $(N_1(t_1), \ldots, N_1(t_k))$ and $(N_2(t_1), \ldots, N_2(t_k))$ are independent for arbitrary distinct times $t_1, \ldots, t_k$, which corresponds to the independence of the processes N_1 and N_2. □

Example 9.8 (*Thinned traffic flow*). For the model of Example 9.6, it follows by Theorem 9.7 that the traffic flows corresponding cars containing one person and containing more people are mutually independent. Therefore, the probability that during a particular minute, at most 20 cars contain only one person, given that at least 30 cars contain more than one person equals

$$\mathbb{P}\,(N_2(1) \le 20 \mid N_1(1) \ge 30) \;=\; \mathbb{P}\,(N_2(1) \le 20).$$

Thus, information about cars containing more than one person has no relevance in predicting how many cars contain only one person. ■

The above independence is rather counterintuitive, because by definition, we must have

$$N_1(t) + N_2(t) = N(t)$$

with probability one. The independence property is one of the magical properties of Poisson processes — which is not valid in general for other counting processes. The result of Theorem 9.7 can be generalized to thinnings with more general random variables compared to coin flips.

Theorem 9.9 (*General thinned Poisson processes*)**.** *Let N be a Poisson process with intensity λ, and let $Z_1, Z_2, \ldots$ be a collection of iid random variables with values in S, independent of N. Then, the thinnings*

$$N_x(t) = \sum_{j=1}^{N(t)} \mathbb{1}\{Z_j = x\}, \qquad x \in S,$$

are mutually independent Poisson processes, with intensities $\lambda_x = \lambda \cdot \mathbb{P}(Z_j = x)$.

Proof. This is a good exercise for students majoring in mathematics: check out the proof of Theorem 9.7 and think how to obtain the asserted result. □

9.4 Memoryless races

Recall from Example 8.8 that the minimum of two exponentially distributed random variables is also exponentially distributed:

$$\begin{cases} Y_1 \sim \text{Exp}(\lambda_1), \\ Y_2 \sim \text{Exp}(\lambda_2) \end{cases} \qquad \Longrightarrow \qquad \min\{Y_1, Y_2\} \sim \text{Exp}(\lambda_1 + \lambda_2).$$

We now prove that this property similarly holds for more than two exponential random variables.

Consider a set of competitors labeled by $j \in I$ participating in a race. Assume that the final time of competitor j equals $Y_j \sim \text{Exp}(\lambda_j)$, and that the times of the competitors are independent. Then, the winning time of the race equals

$$Y_{\min} = \min_{j \in I} Y_j$$

and the label of the winner is

$$J_{\min} = \arg\min_{j \in I} Y_j.$$

Being independent random numbers with a continuous distribution, the times Y_j 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 besides the exponential.

Theorem 9.10. *If $\lambda = \sum_{j \in I} \lambda_j < \infty$ (e.g., when index set I is finite), then the winning time $Y_{\min}$ is $\text{Exp}(\lambda)$-distributed with rate parameter λ, and $J_{\min}$ is distributed as*

$$\mathbb{P}(J_{\min} = j) = \frac{\lambda_j}{\lambda}, \qquad j \in I. \tag{9.8}$$

Moreover, $Y_{\min}$ and $J_{\min}$ are independent.

Proof. **Step 1.** We first determine the distribution of the winning time. Because

$$\mathbb{P}(Y_{\min} > t) \;=\; \mathbb{P}(Y_j > t \text{ for all } j \in I) \;=\; \prod_{j\in I} e^{-\lambda_j t} \;=\; e^{-\lambda t}, \tag{9.9}$$

we may conclude that $Y_{\min} \sim \mathrm{Exp}(\lambda)$.

Step 2. Note that competitor j wins the race precisely when $Y_j < Z_j = \min_{k\neq j} Y_k$, where the random number Z_j is the best time among the rivals of j. By Step 1, we see that $Z_j \sim \mathrm{Exp}(\kappa_j)$, where $\kappa_j = \sum_{k\neq j} \lambda_k$. Because Y_j and Z_j are independent from each other, we see that

$$\begin{aligned}
\mathbb{P}(Y_{\min} > t,\, J_{\min} = j) \;&=\; \mathbb{P}(t < Y_j < Z_j) \\
&=\; \int_0^\infty \int_0^\infty \mathbb{1}\{t < t_j < s\} \cdot \lambda_j e^{-\lambda_j t_j} \cdot \kappa_j e^{-\kappa_j s}\, \mathrm{d}t_j\, \mathrm{d}s \\
&=\; \lambda_j \int_t^\infty e^{-\lambda_j t_j} \cdot \Big(\kappa_j \int_{t_j}^\infty e^{-\kappa_j s}\, \mathrm{d}s\Big)\, \mathrm{d}t_j \\
&=\; \lambda_j \int_t^\infty e^{-\lambda_j t_j} \cdot e^{-\kappa_j t_j}\, \mathrm{d}t_j \\
&=\; \lambda_j \int_t^\infty e^{-\lambda t_j}\, \mathrm{d}t_j \qquad\qquad [\text{since } \lambda_j + \kappa_j = \lambda] \\
&=\; \frac{\lambda_j}{\lambda} e^{-\lambda t}.
\end{aligned}$$

From this and (9.9), we may conclude that

$$\mathbb{P}(Y_{\min} > t,\, J_{\min} = j) \;=\; \frac{\lambda_j}{\lambda} \cdot \mathbb{P}(Y_{\min} > t). \tag{9.10}$$

In particular, by substituting $t = 0$ into (9.10), we see that (9.8) holds.

Step 3. Lastly, substituting (9.8) into (9.10), we obtain

$$\mathbb{P}(Y_{\min} > t,\, J_{\min} = j) \;=\; \mathbb{P}(J_{\min} = j) \cdot \mathbb{P}(Y_{\min} > t),$$

which shows that $Y_{\min}$ and $J_{\min}$ are independent. □

10 Continuous-time Markov chains

We will now study more general continuous-time stochastic processes $X = (X_t)_{t\in[0,\infty)}$ with values in a countable (finite or countably infinite) state space S.

10.1 Poisson process as a continuous-time Markov chain

Poisson process $N = (N_t)_{t\in[0,\infty)}$ is a natural example of a continuous-time Markov chain. Recall that it has jumps at random time instants $T_1, T_2, \ldots$, whose differences (arrival times) $\tau_1, \tau_2, \ldots$ are exponentially distributed (Theorem 8.5). The jump instants can be written in the form

$$T_n = \tau_1 + \cdots + \tau_n = \min\{t \geq 0 : N(t) = n\}, \qquad n = 1, 2, \ldots,$$

and Poisson process is the counting process (see also Figure 8.1)

$$N_t = N(t) = \sum_{j=1}^{\infty} \mathbb{1}\{T_j \leq t\}, \qquad t \in (0, \infty).$$

Example 10.1 (*Poisson process*)**.** Consider Poisson process $N = (N_t)_{t\in[0,\infty)}$ with intensity λ. As the associated Poisson point process $\{T_1, T_2, \ldots\}$ is homogeneous and independently scattered (Theorem 8.13), we find that for any event H_{s-} determined by the values of the Poisson process up to time $[0, s]$, the future values of the Poisson process only depend on N_s:

$$\begin{aligned}
\mathbb{P}(N_{s+t} = k \mid N_s = j,\, H_{s-}) &= \mathbb{P}(N_{s+t} - N_s = k - j \mid N_s = j,\, H_{s-}) \\
&= \mathbb{P}(N_{s+t} - N_s = k - j \mid N_s = j) \\
&= \mathbb{P}(N_{s+t} - N_s = k - j) \\
&= \mathbb{P}(N_t - N_0 = k - j) \\
&= \mathbb{P}(N_t = k - j), \qquad t \geq 0.
\end{aligned}$$

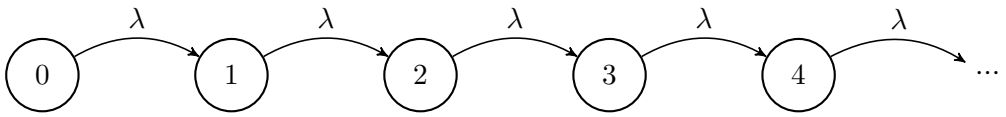
Because the random variable $N_t \sim \mathrm{Poi}(\lambda t)$, it follows that the time evolution of $N = (N_t)_{t\in[0,\infty)}$ can be encoded into a family of *transition matrices*

$$P_t(j,k) = \mathbb{P}(N_{s+t} = k \mid N_s = j) = \begin{cases} e^{-\lambda t}\dfrac{(\lambda t)^{k-j}}{(k-j)!}, & k \geq j, \\ 0, & \text{otherwise.} \end{cases} \tag{10.1}$$

Hence, $N = (N_t)_{t\in[0,\infty)}$ is a continuous-time Markov process on the countably infinite state space $\mathbb{N}_0$: it satisfies the continuum analogue (10.2) of the Markov property (as defined below).

The continuous-time process N makes jumps *at random time instants* T_j in $(0, \infty)$, that are exponentially distributed with mean $1/\lambda$. The locations of the jumps follow an underlying discrete-time Markov chain, which is a simple birth–death chain with transition diagram

In the setting of continuous-time Markov chains, it is conventional to encode the jump rates (1/mean) into the transition diagram:



■

The Poisson process can be thought of as a continuous-time Markov chain following the above transition diagram and evolving in time as follows. Starting at state $N_0 = 0$, the process

1. spends a random Exp(λ)-distributed time in state 0,
2. thereafter, the process jumps from state 0 to state 1,
3. then, it spends a random Exp(λ)-distributed time in state 1,
4. thereafter, the process jumps from state 1 to state 2,

and so on. In this simple example, the Poisson process N always increases, as can be seen directly from the transition diagram. We will see more complicated examples shortly.

10.2 Continuous-time Markov chains

Let S be a countable (finite or countably infinite) state space. We consider stochastic processes $X = (X_t)_{t\in[0,\infty)}$ running in continuous time, with values in S. The transition matrix in continuous time is rather a collection of transition matrices for each time instant $t \in [0,\infty)$: we write it as $P = (P_t)_{t\in[0,\infty)}$, where each $P_t : S \times S \to [0,1]$ satisfies

$$\sum_{y\in S} P_t(x,y) \;=\; 1, \qquad \text{for all } x \in S.$$

The definition of a continuous-time Markov chain is very similar to the discrete-time case. The main difference is of mathematical nature: one has to be a bit careful with what is meant by the history of the process up to a give time $s \geq 0$. This leads to the mathematical notion of *measurability* — which is beyond the scope of the present course, but will be discussed, for example, in the course Probability theory (MS-E1600), see [Kyt20].

Definition. An S-valued stochastic process $X = (X_t)_{t\in[0,\infty)}$ is a (time-homogeneous) *continuous-time Markov chain* (*jatkuva-aikainen Markov-ketju*) with state space S and transition matrices $P = (P_t)_{t\in[0,\infty)}$ if X is "conditionally independent of the past", i.e.,

$$\mathbb{P}\,(X_{s+t} = y \mid X_s = x,\, H_{s-}) \;=\; P_t(x,y), \tag{10.2}$$

for all states $x, y \in S$, all times $t, s \geq 0$, and for all events[a] $H_{s-} = \{(X_u)_{u\leq s} \in A\}$ such that $\mathbb{P}\,(X_s = x,\, H_{s-}) > 0$.

[a] Here, $A \subset \{f : [0,t] \to S\}$ is any suitable set of functions from time interval $[0,s]$ to state space S.

▷ As in Theorem 1.1, the Markov property (10.2) can be written in the form

$$\mathbb{P}\,(X_{s+t} = y \mid X_s = x) \;=\; \mathbb{P}\,(X_t = y \mid X_0 = x) \;=\; P_t(x,y), \tag{10.3}$$

for all times $t, s \geq 0$ and for all states $x, y \in S$ such that $\mathbb{P}\,(X_s = x) > 0$.

▷ The Markov property (10.2) can also be extended to concern several future states and times of the process X: for any suitable set $B \subset \{f : [t,\infty) \to S\}$ of functions from time interval $[t,\infty)$ to state space S,

$$\begin{aligned}\mathbb{P}\,((X_t)_{t\geq s} \in B \mid X_s = x,\, H_{s-}) \;&=\; \mathbb{P}\,((X_t)_{t\geq s} \in B \mid X_s = x) \\ &=\; \mathbb{P}\,((X_t)_{t\geq 0} \in B \mid X_0 = x),\end{aligned} \tag{10.4}$$

for all states $x, y \in S$, all times $s \geq 0$, and for all events $H_{s-} = \{(X_u)_{u\leq s} \in A\}$ such that $\mathbb{P}\,(X_s = x,\, H_{s-}) > 0$. This property is sometimes referred to as the *extended Markov property* (*laajennettu Markov-ominaisuus*), see [Kal21, Lemma 11.1].

Example 10.2 (*Satellite*)**.** A satellite that has been launched in space has a random operational time, which is assumed to be exponentially distributed: $T \sim \mathrm{Exp}(\kappa)$, with mean $1/\kappa = 10$ years. When the satellite breaks, it will not be repaired. The state of the satellite can be described as a simple stochastic process

$$X_t = \begin{cases} 1, & \text{if satellite is operational at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

Given that the event $\{X_s = 1\}$ occurs at time s, we know that the satellite is still operational at time s, and nothing has so far happened to the system. Therefore, by applying the memoryless property of exponential distributions from Theorem 8.7, we see that for any event H_{s-} determined by the past values $(X_u)_{u \le s}$ of the process, we have

$$\begin{aligned} \mathbb{P}(X_{s+t} = 1 \mid X_s = 1,\, H_{s-}) &= \mathbb{P}(X_{s+t} = 1 \mid X_s = 1) && \text{[by (10.2)]} \\ &= \mathbb{P}(T > s + t \mid T > s) \\ &= \mathbb{P}(T > t) = e^{-\kappa t}. \end{aligned}$$

Thus, by the law of total probability, the probability of the complementary event equals

$$\mathbb{P}(X_{s+t} = 0 \mid X_s = 1,\, H_{s-}) = 1 - e^{-\kappa t}.$$

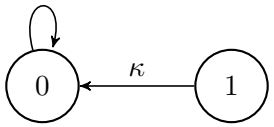
Because a broken satellite remains broken forever, we see that

$$\mathbb{P}(X_{s+t} = 0 \mid X_s = 0,\, H_{s-}) = 1 \qquad \text{and} \qquad \mathbb{P}(X_{s+t} = 1 \mid X_s = 0,\, H_{s-}) = 0.$$

In conclusion, we see that $X = (X_t)_{t \in [0,\infty)}$ is a continuous-time Markov chain on state space $\{0, 1\}$, and its time-t transition matrix is

$$P_t = \begin{bmatrix} P_t(0,0) & P_t(0,1) \\ P_t(1,0) & P_t(1,1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 - e^{-\kappa t} & e^{-\kappa t} \end{bmatrix}, \qquad t \ge 0.$$

Its underlying discrete-time Markov chain has transition diagram (including the jump rate κ)

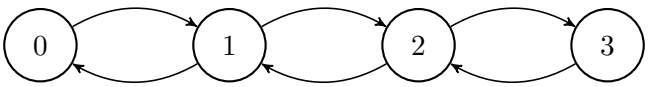


■

In Example 10.2, we were able to compute the transition matrices P_t. However, more often than not, computing transition matrices directly from a description of a continuous time Markov chain is difficult — even if the Markov chain is relatively simple as in the following example.

Example 10.3 (*Taxis near Christmastime*)**.** During December, the town Pussinperä has three taxis operating around town. Each taxi ride (including the return time to the taxi stand) lasts on average m = 20 minutes, and customers arrive at the taxi stand with rate $\lambda = 2$ customers per hour. If upon a customers' arrival all taxis are busy, then the customer will go elsewhere. What is the probability that all taxis are busy when a customer arrives?

We model the taxis by a continuous-time Markov chain $X = (X_t)_{t \in [0,\infty)}$, where X_t is the number of busy taxis at time t. The state space is $\{0, 1, 2, 3\}$, and the underlying discrete-time model follows transition diagram



The probability that all taxis are busy when a customer arrives is $\mathbb{P}(X_t = 3)$. Let us make some further assumptions that will enable us to mathematically analyze this model.

- ▷ Assume that the arrivals of the customers occur independently and exponentially with rate $\lambda = 2$ customers per hour, as in Example 8.3. In other words, we assume that the iid waiting times $\tau_1, \tau_2, \ldots$ between customers' arrivals follow the distribution $\tau_1 \sim \mathrm{Exp}(\lambda)$.
- ▷ Assume that the duration $\sigma_1, \sigma_2, \ldots$ of each taxi ride is also exponentially distributed and independent of the durations of the other taxi rides, as well as of the arrivals of the customers: $\sigma_1 \sim \mathrm{Exp}(\kappa)$. Since each taxi ride lasts on average $\mathrm{m} = 20$ minutes, we see that

$$\mathbb{E}[\sigma_1] = \frac{1}{\kappa} = 20\,\text{minutes} = \frac{1}{3}\,\text{hours},$$

  so that the rate is $\kappa = 3$ per hour.

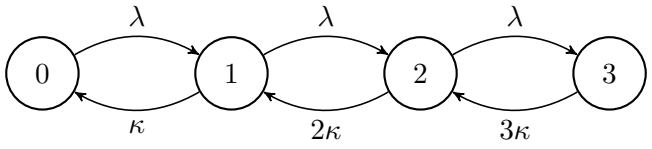
The behavior of this continuous-time Markov chain over time can be characterized as follows. Starting at state $x \in \{0, 1, 2, 3\}$, the Markov chain $X = (X_t)_{t \in [0,\infty)}$

- ▷ spends a random exponentially distributed time in state x,
- ▷ thereafter, it jumps from state x to some state y with some jump probability,
- ▷ then, it spends a random exponentially distributed time in state y,
- ▷ thereafter, it jumps from state y to some state z with some jump probability,

and so on. To characterize the model, we thus have to find the exponential jump rates and the transition probabilities. We begin by deducing the rates of the exponential waiting times.

- ▷ Because customers arrive according to $\mathrm{Exp}(\lambda)$-distributed waiting times, the jump rates for the Markov chain from state $x \in \{0, 1, 2\}$ to state $x + 1 \in \{1, 2, 3\}$ are all equal to λ.
- ▷ If at time instant s, one taxi is busy ($X_s = 1$), then in order for all of the taxis to be available again, we have to wait for time duration $\sigma_1 \sim \mathrm{Exp}(\kappa)$. Thus, the jump rate for the Markov chain from state 1 to state 0 is κ.
- ▷ If at time instant s, two of the taxis are busy ($X_s = 2$), then in order for *one more taxi* to be available again, we have to wait for time duration $\min\{\sigma_1, \sigma_2\}$, which is the time when *one of the two taxis* becomes available. Now, recall from Example 8.8 that $\min\{\sigma_1, \sigma_2\} \sim \mathrm{Exp}(2\kappa)$. Thus, the jump rate for the Markov chain from state 2 to state 1 is 2κ.
- ▷ Analogously, if at time instant s, all taxis are busy ($X_s = 3$), then in order for one more taxi to be available again, we have to wait for time duration $\min\{\sigma_1, \sigma_2, \sigma_3\}$. Similarly as in Example 8.8 and Theorem 9.10, one can prove that $\min\{\sigma_1, \sigma_2, \sigma_3\} \sim \mathrm{Exp}(3\kappa)$, so that the jump rate for the Markov chain from state 3 to state 2 is 3κ.

Hence, including the jump rates, our model follows transition diagram



To find the probability $\mathbb{P}(X_t = 3)$ that all taxis are busy when a customer arrives, we will have to analyze the transition matrices P_t of the continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$. But here, we run into an issue: since the Markov chain is in continuous time, in any time interval $t \in [0, T]$ the Markov chain can make *arbitrarily many* jumps. Hence, to compute the transition probability $P_t(x, y)$ we should sum over *all possible paths* between x and y, which there are infinitely many of. This is computationally unfeasible. However, not all hope is lost! Indeed, in Theorem 11.7 we will show a *systematic way* to compute P_t. Furthermore, the long-term behaviour of the Markov chain *can* be easily analysed using transition diagrams. We will come back to this in Examples 10.9 and 10.11. ■

10.3 Transition matrices and semigroup property

Recall that for discrete-time Markov chains, the time evolution of the process is encoded in powers of the transition matrix (Theorem 5.3). For continuous-time Markov chains, one would expect a similar property, but because non-integer powers of the transition matrices do not make sense, one has to instead use the time-dependent matrices $P_t : S \times S \to [0, 1]$.

Theorem 10.4 (*Time-dependent distribution*)**.** *The distribution*

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

of any continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ at an arbitrary time instant $t \geq 0$ can be computed from the initial distribution μ_0 using the formula

$$\mu_t = \mu_0 \cdot P_t, \tag{10.5}$$

where P_t is the time-t transition matrix of the Markov chain.

Proof. By conditioning on the possible values of the initial state X_0, we find that

$$\begin{aligned}\mathbb{P}(X_t = y) &= \sum_{x\in S} \mathbb{P}(X_t = y \mid X_0 = x) \cdot \mathbb{P}(X_0 = x) \\ &= \sum_{x\in S} \mathbb{P}(X_0 = x) \cdot \mathbb{P}(X_t = y \mid X_0 = x) = \sum_{x\in S} \mu_0(x) \cdot P_t(x, y),\end{aligned}$$

which is the asserted equation in matrix form. □

The notion of associativity of the matrix powers is provided by the so-called *semigroup property*: if the Markov chain first evolves for s amount of time and then t amount of time, the total evolution should be stochastically the same as for $s + t$ amount of time. In mathematical terms, this means that the collection $(P_t)_{t\geq 0}$ forms a *transition semigroup* (*siirtymäpuoliryhmä*).

Theorem 10.5 (*Transition semigroup*)**.** *The transition matrices of a continuous-time Markov chain form a transition semigroup $P = (P_t)_{t\in[0,\infty)}$, that is,*

$$P_s \cdot P_t = P_{s+t}, \qquad \text{for all } s, t \geq 0. \tag{10.6}$$

Equation (10.6) is also called the *Chapman-Kolmogorov equation* (*Chapman-Kolmogorov-yhtälöt*).

Proof. By summing over all possible states $X_s = y$ and using the Markov property, we compute

$$\begin{aligned} P_{s+t}(x,z) &= \mathbb{P}(X_{s+t} = z \mid X_0 = x) \\ &= \sum_{y \in S} \mathbb{P}(X_{s+t} = z,\, X_s = y \mid X_0 = x) \\ &= \sum_{y \in S} \mathbb{P}(X_{s+t} = z \mid X_s = y,\, X_0 = x) \cdot \mathbb{P}(X_s = y \mid X_0 = x) \\ &= \sum_{y \in S} P_t(y,z) \cdot P_s(x,y) \qquad \text{[by (10.2)]} \\ &= \sum_{y \in S} P_s(x,y) \cdot P_t(y,z), \end{aligned}$$

which is the asserted equation in matrix form. □

10.4 Generator matrix

From the semigroup property, one could heuristically anticipate that P_t could be determined from small time instants as

$$P_t = P_{n \cdot \frac{t}{n}} = P_{\frac{t}{n} + \cdots + \frac{t}{n}} = P_{t/n}^n.$$

However, taking the limit $n \to \infty$ of $P_{t/n}$ would only give the identity matrix, which gives no information about the behavior of the Markov chain. The issue is that, contrast to discrete-time Markov chains, there is no "next time step" in continuous time. Instead, we would like to retain information about the transition probabilities only in "infinitesimal future", or in "infinitesimal neighbourhood" of time $t = 0$. Derivatives capture only this kind of infinitesimal information, which motivates the following definition.

Definition. For continuous-time Markov chain $X = (X_t)_{t \in [0,\infty)}$ with transition matrices $P = (P_t)_{t \in [0,\infty)}$, the *generator matrix* (*generaattorimatriisi*), if exists, is

$$Q = \lim_{t \to 0} \frac{1}{t}(P_t - P_0) = \frac{\mathrm{d}}{\mathrm{d}t} P_t \Big|_{t=0}. \tag{10.7}$$

▷ We will see in Theorem 11.4 in Section 11 that the limit (10.7) exists when the jump rates of the Markov chain are all uniformly bounded — for instance, when state space S is finite.

▷ Equation (10.7) gives for the matrix entries

$$Q(x,y) = \begin{cases} \lim_{t \to 0} \frac{1}{t} P_t(x,y), & x \neq y \\ \lim_{t \to 0} \frac{1}{t}(P_t(x,y) - 1), & x = y, \end{cases} = \frac{\mathrm{d}}{\mathrm{d}t} P_t(x,y) \Big|_{t=0}.$$

For $x \neq y$, this describes the *jump rate of the Markov chain from state x to state y.*

To demonstrate some general features of generator matrices, let us return to Example 10.2.

Example 10.6 (*Satellite*)**.** Consider the transition matrices of Example 10.2

$$P_t = \begin{bmatrix} 1 & 0 \\ 1 - e^{-\kappa t} & e^{-\kappa t} \end{bmatrix}, \qquad t \geq 0$$

on state space $\{0,1\}$. As $t \to 0$, we see that

$$P_t \xrightarrow{t\to 0} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

while as $t \to \infty$, we see that

$$P_t \xrightarrow{t\to \infty} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

In particular, we would expect that the Markov chain has invariant (and limiting) distribution

$$\pi = [\pi(0), \pi(1)] = [1, 0].$$

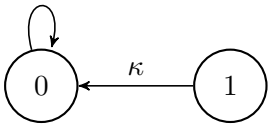
(This means that the satellite is eventually broken.)

We can compute the generator matrix using l'Hôpital's rule:

$$Q = \frac{1}{t}\big(P_t - P_0\big) \xrightarrow{t\to 0} \lim_{t\to 0} \begin{bmatrix} 0 & 0 \\ \frac{1-e^{-\kappa t}}{t} & \frac{e^{-\kappa t}-1}{t} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \kappa & -\kappa \end{bmatrix}$$

The entry $Q(1,0) = \kappa$ is the *jump rate* of the Markov chain from state 1 to state 0, and the entry $Q(0,1) = 0$ is the jump rate of the Markov chain from state 0 to state 1. The diagonal entries $Q(0,0) = 0$ and $Q(1,1) = -\kappa$ do not have such an obvious interpretation.

▷ Recall that the jump rate κ appears also in the transition diagram



▷ We can also observe that $\pi \cdot Q = 0$, which is actually a general fact (see Theorem 10.10).

■

Let us next observe that the generator matrix Q governs how the transition matrices P_t change over time via so-called Kolmogorov's backward differential equation.

Corollary 10.7 (*Kolmogorov's backward equation*)**.** *For continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ with generator matrix Q, the transition semigroup $P = (P_t)_{t\in[0,\infty)}$ satisfies* Kolmogorov's backward differential equation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} P_t &= Q \cdot P_t, \qquad t \in [0,\infty) \\ P_0 &= I. \end{aligned} \tag{10.8}$$

For matrix entries, Equation (10.8) gives the initial value problem

$$\begin{aligned} \Big(\frac{\mathrm{d}}{\mathrm{d}t} P_t\Big)(x,y) &= (Q \cdot P_t)(x,y), \qquad t \in [0,\infty), \\ P_0(x,y) &= \mathbb{1}\{x = y\}, \end{aligned}$$

for all states $x, y \in S$. This gives means for solving P_t if the generator matrix Q is known.

Proof. By definition (10.7) and the semigroup property (10.6) from Theorem 10.5, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}P_t \;=\; \lim_{s\to 0}\frac{P_{t+s}-P_t}{s} \;=\; \lim_{s\to 0}\Big(\frac{P_s-P_0}{s}\Big)\cdot P_t \;=\; Q\cdot P_t, \qquad t\in[0,\infty),$$

(where the matrix P_t can be taken outside of the limit since its row-sums equal one). □

The matrix initial value problem (10.8) is analogous to that for the exponential function:

$$f(t) \;=\; e^{at} \;=\; \sum_{n=0}^{\infty}\frac{(at)^n}{n!}$$

satisfies the ordinary differential equation

$$\begin{aligned} f'(t) \;&=\; a\,f(t), \qquad t\in[0,\infty),\\ f(0) \;&=\; 1. \end{aligned}$$

In fact, we will show in Theorem 11.7 in Section 11 that[33] *matrix exponentials* e^{tQ} of the generator matrix Q give rise to the transition matrices

$$P_t \;=\; e^{tQ}, \qquad t\in[0,\infty),$$

In light of Corollary 10.7, this does not come as a surprise.

Lemma 10.8. *Generator matrix Q of a continuous-time Markov chain has*

▷ *zero row-sums*:

$$\sum_{y\in S}Q(x,y) \;=\; 0, \qquad \text{for all } x\in S, \tag{10.9}$$

▷ *nonnegative off-diagonal entries*:

$$Q(x,y) \;\geq\; 0, \qquad \text{for all } x,y\in S \text{ such that } x\neq y,$$

▷ *and nonpositive diagonal entries, given by*

$$Q(x,x) \;=\; -\sum_{y\neq x}Q(x,y), \qquad \text{for all } x\in S. \tag{10.10}$$

Proof. It follows from the definition

$$Q(x,y) \;=\; \lim_{t\to 0}\frac{P_t(x,y)-I(x,y)}{t}, \qquad x,y\in S, \tag{10.11}$$

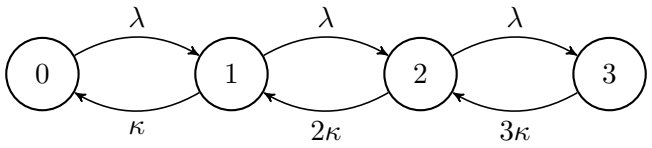
that the row-sums of Q must be zero, because P_t and I have unit row-sums. The above formula (10.11) also implies that the off-diagonal entries of Q satisfy

$$Q(x,y) \;=\; \lim_{t\to 0}\frac{P_t(x,y)}{t} \;\geq\; 0, \qquad x\neq y.$$

Because the row-sums of Q are zero, we see that its diagonal entries must satisfy (10.10). □

[33] When the jump rates of the Markov chain are all uniformly bounded.

Example 10.9 (*Taxis near Christmastime*)**.** Let us come back to Example 10.3. While it is not very practical to try to compute the transition matrices for this model directly, one can analyze its generator matrix. From the jump rates in the transition diagram



one can read the generator matrix

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 \\ \kappa & -\lambda-\kappa & \lambda & 0 \\ 0 & 2\kappa & -\lambda-2\kappa & \lambda \\ 0 & 0 & 3\kappa & -3\kappa \end{bmatrix}. \tag{10.12}$$

Its off-diagonal entries $Q(x,y)$ with $x \neq y$ are just the labels (jump rates) of the transition diagram, and its diagonal entries $Q(x,x)$ are chosen so that the row-sums (10.9) are zero. ■

10.5 Invariant distributions

Invariant distributions for a continuous-time Markov chain are defined similarly as in the discrete-time case. The main result of practical importance in this section is Theorem 10.10, which gives balance equations from which one can compute the invariant distribution.

Definition. $\pi = \{\pi(x) : x \in S\}$ is an *invariant distribution* (*tasapainojakauma*) of transition matrices $P = (P_t)_{t\in[0,\infty)}$ and the corresponding Markov chain $X = (X_t)_{t\in[0,\infty)}$ if it satisfies the *balance equations* $\pi \cdot P_t = \pi$ for all times $t \geq 0$, that is,

$$\sum_{x\in S} \pi(x) \cdot P_t(x,y) = \pi(y), \qquad \text{for all } y \in S \text{ and } t \in [0,\infty),$$

and the *law of total probability*

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

Theorem 10.10 (*Invariant distribution*)**.** *The following are equivalent for a continuous-time* finite-state *Markov chain and for any probability distribution* π:

1. π *is an invariant distribution of the Markov chain.*
2. $\pi \cdot Q = 0$*, where* Q *is the generator matrix of the Markov chain.*

Because the row-sums of Q *are zero, equation* $\pi \cdot Q = 0$ *can be written in the form*

$$\sum_{x\neq y} \pi(x) \cdot Q(x,y) = \pi(y) \cdot \sum_{z\neq y} Q(y,z). \tag{10.13}$$

Thanks to the equivalence in Theorem 10.10, equations $\pi \cdot Q = 0$ are also termed *balance equations* (*tasapainoyhtälöt*) for generator matrix Q and the corresponding Markov chain X.

▷ In (10.13), the left side describes the long-term average rate of jumps *into state* y.

▷ The right side of (10.13) describes the corresponding rate of *out from state y*.

Theorem 10.10 can be also generalized to countably infinite state spaces, provided that the matrix entries of Q are all uniformly bounded (see Section 11).

Proof of Theorem 10.10. We first show that any invariant distribution π of the Markov chain satisfies the equation $\pi \cdot Q = 0$. To this end, by differentiating the formula

$$(\pi \cdot P_t)(y) \;=\; \sum_{x \in S} \pi(x) \cdot P_t(x, y)$$

term by term with respect to t (which is possible since we assumed that S is finite), we see by Kolmogorov's backward differential equation (10.8) from Corollary 10.7 that

$$\frac{\mathrm{d}}{\mathrm{d}t}(\pi P_t) \;=\; \pi \cdot \frac{\mathrm{d}}{\mathrm{d}t} P_t \;=\; \pi \cdot (Q \cdot P_t) \;=\; (\pi \cdot Q) \cdot P_t. \tag{10.14}$$

If π is invariant, this implies that $0 = (\pi \cdot Q) \cdot P_t$. By substituting $t = 0$, we see that indeed,

$$0 \;=\; \pi \cdot Q \cdot P_0 \;=\; \pi \cdot Q \cdot I \;=\; \pi \cdot Q.$$

Conversely, suppose that $\pi \cdot Q = 0$. Then, formula (10.14) shows that $\frac{\mathrm{d}}{\mathrm{d}t}(\pi \cdot P_t) = 0$, so the mapping $t \mapsto \pi \cdot P_t$ is constant over time. Therefore, we see that

$$\pi \cdot P_t = \pi \cdot P_0 = \pi, \qquad \text{for all } t \in [0, \infty),$$

that is, π is invariant for P_t. □

Example 10.11 (*Taxis near Christmastime*)**.** Let us again come back to Example 10.3. While it is not very practical to try to compute the transition matrices for this model directly, one can analyze its generator matrix (10.12)

$$Q \;=\; \begin{bmatrix} -\lambda & \lambda & 0 & 0 \\ \kappa & -\lambda - \kappa & \lambda & 0 \\ 0 & 2\kappa & -\lambda - 2\kappa & \lambda \\ 0 & 0 & 3\kappa & -3\kappa \end{bmatrix}$$

By Theorem 10.10, the invariant distribution π can be handily computed using the generator matrix Q from the balance equations

$$\pi \cdot Q = 0.$$

Solving these equations for the Markov chain in this example yields

$$\pi(1) \;=\; \pi(0) \cdot \frac{\lambda}{\kappa}, \qquad \pi(2) \;=\; \pi(1) \cdot \frac{\lambda}{2\kappa}, \qquad \pi(3) \;=\; \pi(2) \cdot \frac{\lambda}{3\kappa},$$

and requiring that π is a probability distribution, we have

$$\pi(0) \;+\; \pi(1) \;+\; \pi(2) \;+\; \pi(3) \;=\; 1,$$

from which we can solve for

$$\pi(0) \;=\; \frac{1}{1 + \frac{\lambda}{\kappa} + \frac{1}{2}\left(\frac{\lambda}{\kappa}\right)^2 + \frac{1}{6}\left(\frac{\lambda}{\kappa}\right)^3}.$$

Here, λ/κ is the ratio of the arrival rate λ and the service rate κ. Plugging in $\lambda = 2$ and $\kappa = 3$, we find that

$$\mathbb{P}\left(X_t = 3 \mid X_0 \sim \pi\right) \;=\; \pi(3) \;=\; 0.025.$$

This is the probability at the statistical equilibrium that all taxis are busy when a customer arrives, which we were after in Example 10.3. ■

10.6 Irreducibility, reversibility, and convergence theorems

Irreducibility for continuous-time Markov chains is defined in the same way[34] as in discrete time. Note that the transition diagram of a continuous-time Markov chain is a directed graph with nodes being the states, and links being the pairs (x, y) for which $Q(x, y) > 0$. Thus, a generator matrix Q and the 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.

Theorem 10.12 (*Uniqueness of invariant distribution*)**.** *Any* irreducible *continuous-time Markov chain has* at most one *invariant distribution.*

If the invariant distribution π exists, then it also equals the unique limiting distribution,

$$\lim_{t\to\infty} P_t(x, y) = \pi(y), \qquad \text{for all } x \in S. \tag{10.15}$$

- ▷ The balance equations (10.13) provide a practical way to find the invariant distribution, if it exists. (Computing the limit (10.15) is usually not practical if the state space is large.)
- ▷ Any *irreducible* continuous-time Markov chain on a *finite* state space has a unique invariant distribution. This is not always true for infinite state spaces.

Proof. This is an analogue of Theorem 5.6. See [Dur12, Theorem 4.4] for a detailed proof. □

Reversibility of a continuous-time Markov chain is defined similarly as in the discrete-time case: a generator matrix Q and the corresponding Markov chain X is called *reversible* with respect to a probability distribution π (π-reversible) if the *detailed balance equations* hold:

$$\pi(x) \cdot Q(x, y) = \pi(y) \cdot Q(y, x), \qquad \text{for all } x, y \in S \text{ such that } x \neq y. \tag{10.16}$$

Note that reversibility also includes the condition (10.17) that π is a probability distribution.

Theorem 10.13 (*Existence and uniqueness of invariant distribution from reversibility*)**.** *Every* irreducible *and* reversible *Markov chain on a countable state space admits a* unique *invariant distribution π, satisfying the detailed balance equations* (10.16):

$$\pi(x) \cdot Q(x, y) = \pi(y) \cdot Q(y, x), \qquad \text{for all } x, y \in S \text{ such that } x \neq y,$$

and the normalization

$$\sum_{x\in S} \pi(x) = 1. \tag{10.17}$$

Moreover, π also equals the unique limiting distribution,

$$\lim_{t\to\infty} P_t(x, y) = \pi(y), \qquad \text{for all } x \in S.$$

Proof. This is an analogue of Theorem 5.7. See [Dur12, Chapter 4] for a detailed proof. □

[34] For continuous-time Markov chains, we never need to worry about periodicity issues, because all continuous-time chains are automatically aperiodic.

Theorem 10.14 (*Ergodic theorem*)**.** *For any* irreducible *continuous-time Markov chain* $X = (X_t)_{t\in[0,\infty)}$ *with invariant distribution* π*, we have*

$$\frac{1}{t}\int_0^t \phi(X_s)\,\mathrm{d}s \longrightarrow \sum_{y\in S}\pi(y)\cdot\phi(y), \qquad \text{as } t\to\infty,$$

for any function $\phi : S \to \mathbb{R}$ *with probability one, regardless of the initial state of the Markov chain.*

Proof. This is an analogue of Theorem 3.1. Its proof will be skipped in this course. □

11 Analysis of continuous-time Markov chains

11.1 Jump rates and jump probabilities

Consider continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ on countable state space S. It is of great interest to understand the time instant when the Markov chain first exits its initial state (i.e., when it *jumps* to a different state), which is an extended random number in $[0,\infty]$.

Indeed, we will see that the behavior of any Markov chain is completely determined by its random jump instants together with its jump probabilities that tell the distribution of the possible positions of X after its jumps. It turns out that the Markov property is very restrictive:

- ▷ the jump instants must be *memoryless* (thus exponentially distributed) (see Theorems 11.1 & 11.2),
- ▷ while the jump probabilities form a transition matrix called *jump probability matrix* P_* (see Section 11.2).

Definition. For continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ started at $X_0 = x$, the *first jump instant* (*ensimmäinen hyppyhetki*) of X is

$$\tau(x) = \min\{t \geq 0 : X_t \neq X_0 = x\}, \tag{11.1}$$

with the notational convention that $\tau(x) = \infty$ if X never leaves its initial state.

Definition. The *total jump rate* (*hyppyvauhti*) of X away from state $x \in S$ is

$$\lambda(x) = \frac{1}{\mathbb{E}(\tau(x))}, \tag{11.2}$$

with the notational convention that $\lambda(x) = 0$ when the denominator is infinite.

We say that jump rates $\{\lambda(x) : x \in S\}$ are *uniformly bounded* (*tasaisesti rajoitettu*) if there exists a constant $\Lambda \in (0,\infty)$ such that

$$\lambda(x) < \Lambda, \qquad \text{for all } x \in S. \tag{11.3}$$

In this course, we will always assume that the jump rates are uniformly bounded. This is the case, for example, when state space S is finite (which holds for most applications in our course).

The following result confirms a fact already observed in the examples: *any continuous-time Markov chain spends an exponentially distributed random time*[35] *in every state that it visits.*

Theorem 11.1 (*First jump instant*)**.** *The first jump instant of continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ started at $X_0 = x$ is* exponentially distributed *with rate $\lambda(x)$:*

$$\tau(x) \sim \text{Exp}(\lambda(x)).$$

[35] Here, we interpret an exponential distribution Exp(0) with rate zero as the distribution of a random variable which is infinite with probability one. This corresponds to staying in an absorbing state.

Proof. By applying the extended Markov property (10.4), we can verify that

$$\begin{aligned}
\mathbb{P}(\tau(x) > s+t \mid \tau(x) > s) &= \mathbb{P}(X_u = x \text{ for all } u \in [s, s+t] \mid X_r = x \text{ for all } r \in [0,s]) && \text{[by (11.1)]}\\
&= \mathbb{P}(X_u = x \text{ for all } u \in [s, s+t] \mid X_s = x) && \text{[by (10.4)]}\\
&= \mathbb{P}(X_u = x \text{ for all } u \in [0, t] \mid X_0 = x) && \text{[by (10.4)]}\\
&= \mathbb{P}(\tau(x) > t). && \text{[by (11.1)]}
\end{aligned}$$

This means that the distribution of $\tau(x)$ is memoryless (cf. Equation (8.3)), so it follows from Theorem 8.7 that $\tau(x) \sim \mathrm{Exp}(\lambda(x))$. □

▷ The *jump instants* (*hyppyhetket*) of Markov chain X can be defined recursively as $T_0 = 0$,

$$\begin{aligned}
T_1 &= \min\{t \geq 0 : X_t \neq X_0\},\\
T_n &= \min\{t \geq T_{n-1} : X_t \neq X_{T_{n-1}}\}, \qquad n = 2, 3, \ldots.
\end{aligned}$$

▷ Then, the *arrival times* (*saapumisajat*), or *waiting times* (*odotusajat*) of the jumps are

$$\tau_n = T_n - T_{n-1}, \qquad n = 1, 2, 3, \ldots.$$

(In the literature, they are also called *interpoint distances* of the point process $\{T_1, T_2, \ldots\}$.)

Theorem 11.2 (*Waiting/arrival times*). *For continuous Markov chain $X = (X_t)_{t\in[0,\infty)}$, conditioned on the event $\{X_0 = x_0, X_{T_1} = x_1, X_{T_2} = x_2, \ldots, X_{T_{n-1}} = x_{n-1}\}$, arrival times $\tau_1, \tau_2, \ldots, \tau_n$ are independent and exponentially distributed, with rates $\tau_k \sim \mathrm{Exp}(\lambda(x_{k-1}))$.*

Moreover, conditioned on the event $\{X_{T_{n-1}} = x\}$, we have $\tau_n \sim \mathrm{Exp}(\lambda(x))$.

Proof. This very intuitive fact can be proven using the Markov property, but applied to random time instant T_{n-1}. The mathematically precise proof is a bit involved — see [Kal21, Theorem 13.1 and Lemma 13.2] (see also [Kal21, Theorem 13.6] for Theorem 8.5 about Poisson process). □

To fully describe the behavior of a continuous-time Markov chain, we need to know how it jumps to the next states. The Markov property guarantees that each new state is selected independently of the past trajectory, only depending on the present state.

Theorem 11.3 (*Jumps are independent of arrival times*). *Fix a state $x \in S$. Then, conditioned on the event $\{X_{T_{n-1}} = x\}$, the random variable X_{T_n} is independent of τ_n, and its distribution only depends on state x.*

Proof. Let us prove the claim in the case $n = 1$, where we have $T_1 = \tau_1$. Fix $y \neq x$. Then,

$$\mathbb{P}(X_{\tau_1} = y \mid \tau_1 \geq t, X_0 = x) = \mathbb{P}(X_{\sigma_t} = y \mid X_s = x \text{ for all } s \in [0,t]), \qquad t \in [0, \infty),$$

where $\sigma_t = \min\{u \geq t : X_u \neq X_t\}$. By the extended Markov property (10.4), X_{σ_t} only depends on X_t and we obtain

$$\begin{aligned}
\mathbb{P}(X_{\tau_1} = y \mid \tau_1 \geq t, X_0 = x) &= \mathbb{P}(X_{\sigma_t} = y \mid X_t = x) = \mathbb{P}(X_{\sigma_0} = y \mid X_0 = x) && \text{[by (10.4)]}\\
&= \mathbb{P}(X_{\tau_1} = y \mid X_0 = x), && \text{[since } \sigma_0 = \tau_1\text{]}
\end{aligned}$$

As this holds for every $y \neq x$ and $t \geq 0$, we may conclude that X_{τ_1} and τ_1 are independent. The general case can be proved using the "strong[36] Markov property" [Kal21, Theorem 13.1]. □

[36] One can apply the Markov property to suitable natural random time instants (called *stopping*, or *optional* times). This is called *strong* Markov property, see [Kal21, Theorem 13.1].

11.2 Embedded discrete-time Markov chain

By Theorems 11.2 and 11.3, the behavior of *any* continuous-time Markov chain over time can be characterized as follows. Starting at state $x \in S$, the Markov chain $X = (X_t)_{t\in[0,\infty)}$

- ▷ spends a random exponentially distributed time $\tau_1(x) \sim \mathrm{Exp}(\lambda(x))$ in state x,
- ▷ thereafter, it jumps from state x to some state y with some jump probability $P_*(x,y)$,
- ▷ then, it spends a random exponentially distributed time $\tau_2(y) \sim \mathrm{Exp}(\lambda(y))$ in state y,
- ▷ thereafter, it jumps from state y to some state z with some jump probability $P_*(y,z)$,
- ▷ and so on.

By the Markov property (10.3), the new state is selected *independently of the past trajectory* of the Markov chain. The Markov chain evolves as above as long as it visits states with a nonzero jump rate. If the Markov chain hits a state with jump rate zero, it remains stuck there (this is an absorbing state). The discrete-time Markov chain with transition probabilities $P_*(x,y)$ is called the *embedded Markov chain* (*upotettu Markov-ketju*).

Definition. The *jump probability matrix* (*hyppytodennäköisyysmatriisi*) of continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ is the (possibly infinite) matrix P_* with rows and columns indexed by states $x, y \in S$ and entries given by

$$P_*(x,y) = \mathbb{P}(X_{\tau(x)} = y \mid X_0 = x).$$

It is a transition matrix on S (related to the underlying discrete-time Markov chain).

- ▷ Note that the diagonal entries $P_*(x,x) = 0$ with $\lambda(x) > 0$ are zero, because the Markov chain changes its state on each jump instant.
- ▷ For states with jump rate $\lambda(x) = 0$, it is usual to define the jump rates as $P_*(x,y) = \mathbb{1}\{x = y\}$, although these entries have no effect on the behavior of the Markov chain. Note that $\lambda(x) = 0$ means that the Markov chain will never jump away from state x, so $P_*(x,x) = 1$.

Theorem 11.4 (*Relationship of generator matrix Q and jump probability matrix P_**)**.** *For continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ with* uniformly bounded jump rates (11.3) *and jump probability matrix P_*, the generator matrix* (10.7, 10.7)

$$Q = \lim_{t\to 0} \frac{1}{t}(P_t - P_0) = \frac{\mathrm{d}}{\mathrm{d}t} P_t(x,y)\Big|_{t=0} \tag{11.4}$$

exists and its entries are given by

$$Q(x,y) = \begin{cases} \lambda(x)\cdot P_*(x,y), & x \neq y, \\ -\lambda(x), & x = y. \end{cases} \tag{11.5}$$

In particular, if the total jump rate $\lambda(x) \neq 0$ is nonvanishing for state x, then we have

$$\lambda(x) = -Q(x,x) \qquad \text{and} \qquad P_*(x,y) = \frac{Q(x,y)}{\lambda(x)}, \qquad y \neq x.$$

Proof sketch. The off-diagonal entries of Q can be obtained by the following heuristics — we leave it as an exercise to mathematically oriented students to fill in the details to make the proof precise. First, for $X_t \neq X_0$ to be possible, the Markov chain has to have made *at least one jump* before time t, so that $\tau_1 \leq t$. Moreover, the event $\{T_2 \leq t\}$ that the Markov chain has made *more than one jump* before time t to hold would also require $\tau_2 \leq t$.

Since by Theorem 11.2, the arrival times $\tau_1 \sim \mathrm{Exp}(\lambda(X_0))$ and $\tau_2 \sim \mathrm{Exp}(\lambda(X_{\tau_1}))$ are independent when conditioned on the event $\{X_0 = x_0, X_{T_1} = x_1\}$, we obtain[37]

$$\begin{aligned}
&\mathbb{P}(T_2 \leq t \mid X_0 = x_0, X_{T_1} = x_1) \\
\leq\ &\mathbb{P}(\tau_1 \leq t, \tau_2 \leq t \mid X_0 = x_0, X_{T_1} = x_1) \\
=\ &\mathbb{P}(\tau_1 \leq t \mid X_0 = x_0, X_{T_1} = x_1) \cdot \mathbb{P}(\tau_2 \leq t \mid X_0 = x_0, X_{T_1} = x_1) && \text{[by Theorem 11.2]} \\
\leq\ &(1 - e^{-\Lambda t})^2 \ =\ (\Lambda t)^2 - (\Lambda t)^3 + \cdots, && \text{[by (11.3)]}
\end{aligned}$$

Summing over the possible states $X_0 = x_0$ and $X_{T_1} = x_1$ then yields

$$\begin{aligned}
\mathbb{P}(T_2 \leq t) \ &= \sum_{x_0 \in S} \sum_{x_1 \in S} \mathbb{P}(T_2 \leq t \mid X_0 = x_0, X_{T_1} = x_1) \cdot \mathbb{P}(X_0 = x_0, X_{T_1} = x_1) \\
&\leq\ (1 - e^{-\Lambda t})^2 \underbrace{\sum_{x_0 \in S} \sum_{x_1 \in S} \mathbb{P}(X_0 = x_0, X_{T_1} = x_1)}_{=1} \\
&=\ (1 - e^{-\Lambda t})^2 \ =\ (\Lambda t)^2 - (\Lambda t)^3 + \cdots.
\end{aligned}$$

This shows that, up to linear order in t, the probability $\mathbb{P}(T_2 \leq t)$ equals zero (so it is quite small). Note that the above estimate works regardless of the initial distribution $X_0 \sim \mu_0$.

Next, we take two distinct states $x \neq y$ and start Markov chain X at $X_0 = x$. Note that if $T_2 > t$, then the event $\{X_t = y\}$ is equivalent with the event $\{\tau_1 \leq t \text{ and } X_{\tau_1} = y\}$. Moreover, by Theorem 11.3 τ_1 and X_{τ_1} are independent when conditioned to the event $\{X_0 = x\}$. Since up to linear order in t, the event $\{T_2 > t\}$ happens with probability 1, we have the approximation

$$\begin{aligned}
P_t(x,y) \ &\approx\ \mathbb{P}(\tau_1 \leq t,\ X_{\tau_1} = y \mid X_0 = x) \\
&=\ \mathbb{P}(\tau_1 \leq t \mid X_0 = x) \cdot \underbrace{\mathbb{P}(X_{\tau_1} = y \mid X_0 = x)}_{P_*(x,y)} && \text{[by Theorem 11.3]} \\
&=\ (1 - e^{-\lambda(x)t}) \cdot P_*(x,y), && \text{[as } \tau_1 \sim \mathrm{Exp}(\lambda(x)) \text{ by Theorem 11.1]}
\end{aligned}$$

up to linear order in t. Now, since the *time-derivative at $t = 0$ only recovers the linear order in time*, we can compute the matrix element $Q(x,y)$ by differentiating the right hand side of the above approximation as in definition (10.7):

$$Q(x,y) \ =\ \frac{\mathrm{d}}{\mathrm{d}t} P_t(x,y)\Big|_{t=0} \ =\ \frac{\mathrm{d}}{\mathrm{d}t}(1 - e^{-\lambda(x)t})\Big|_{t=0} \cdot P_*(x,y) \ =\ \lambda(x) \cdot P_*(x,y). \tag{11.6}$$

This gives the off-diagonal entries in the asserted equality (11.5).

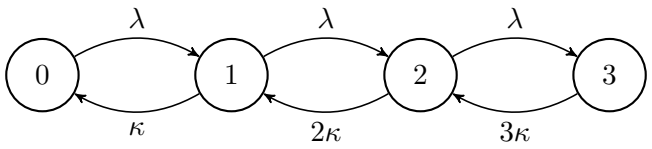
To compute the diagonal entries $Q(x,x)$ assuming $\lambda(x) > 0$, we can use the facts that the row sums of Q equal zero by Lemma 10.8, the row sums of P_* equal one since P_* is a transition matrix, and $P_*(x,x) = 0$ if $\lambda(x) > 0$, to obtain

$$Q(x,x) \ =\ -\sum_{y \neq x} Q(x,y) \ =\ -\sum_{y \neq x} \lambda(x) P_*(x,y) \ =\ -\lambda(x) \sum_{y \in S} P_*(x,y) \ =\ -\lambda(x).$$

On the other hand, if $\lambda(x) = 0$, then $Q(x,y) = 0$ for every $y \neq x$, and since the row sums of Q equal zero, we also must have $Q(x,x) = 0 = -\lambda(x)$, as claimed. □

[37] The jump rates $\lambda(X_0), \lambda(X_{\tau_1})$ are random, but all assumed to be uniformly bounded by Λ as in (11.3).

Example 11.5 (*Taxis near Christmastime*)**.** We again return to Examples 10.3, 10.9, and 10.11. Including the jump rates, Markov chain $X = (X_t)_{t\in[0,\infty)}$ follows transition diagram



from which we can read the generator matrix encoding the jump rates:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 \\ \kappa & -\lambda-\kappa & \lambda & 0 \\ 0 & 2\kappa & -\lambda-2\kappa & \lambda \\ 0 & 0 & 3\kappa & -3\kappa \end{bmatrix}$$

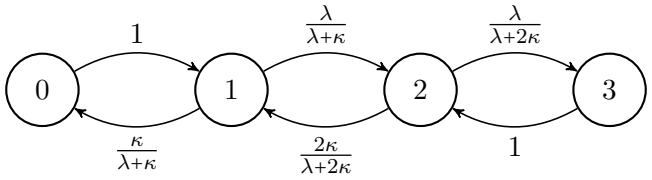
The total jump rate $\lambda(x)$ *away* from state $x \in \{0,1,2,3\}$ is

$$\lambda(0) = \lambda, \qquad \lambda(1) = \lambda+\kappa, \qquad \lambda(2) = \lambda+2\kappa, \qquad \lambda(3) = 3\kappa.$$

From Theorem 11.4, we see that the jump probability is

$$P_* = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{\kappa}{\lambda+\kappa} & 0 & \frac{\lambda}{\lambda+\kappa} & 0 \\ 0 & \frac{2\kappa}{\lambda+2\kappa} & 0 & \frac{\lambda}{\lambda+2\kappa} \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{11.7}$$

In conclusion, the underlying *embedded discrete-time* Markov chain has transition diagram



where the labels now are the *discrete-time* transition probabilities $P_*(x,y)$. ■

> Note that as a directed graph, the discrete-time transition diagram agrees with that of the continuous-time transition diagram of $X = (X_t)_{t\in[0,\infty)}$, while *the labels are different*:
>
> ▷ the continuous-time transition diagram has the jump rates $Q(x,y)$ as labels, whereas
>
> ▷ the discrete-time transition diagram has the transition probabilities $P_*(x,y)$ as labels.

Example 11.6 (*Taxis near Christmastime, continued*)**.** To demonstrate the usefulness of Theorem 11.4, let us find the jump probabilities $P_*(x,y)$, for $x,y \in \{0,1,2,3\}$, in Example 11.5 by hand. We see that the computation can be rather tedious in general.

Some of the matrix entries of P_* are clear from the model:

$$P_*(0,1) = 1, \qquad P_*(0,x) = 0, \quad x \in \{0,2,3\},$$
$$P_*(3,2) = 1, \qquad P_*(3,x) = 0, \quad x \in \{0,1,3\},$$

and since the jump probabilities satisfy the law of total probability,

$$\sum_{y \in S} P_*(x,y) \;=\; 1, \qquad \text{for all } x \in S,$$

we see that

$$P_*(1,0) \;+\; P_*(1,2) \;=\; 1 \qquad \text{and} \qquad P_*(2,1) \;+\; P_*(2,3) \;=\; 1.$$

Hence, we only have to solve for the two probabilities $P_*(1,0)$ and $P_*(2,1)$, say.

▷ We find using Theorem 9.10 (or Example 8.8) that

$$\begin{aligned} P_*(1,0) \;&=\; \mathbb{P}\,(X_t = 0 \mid X_0 = 1) \\ &=\; \mathbb{P}\,(\text{the taxi ride ends before a new customer arrives}) \\ &=\; \mathbb{P}\,(\sigma_1 < \tau_1) \;=\; \frac{\kappa}{\lambda + \kappa}, \end{aligned}$$

where $\tau_1 \sim \mathrm{Exp}(\lambda)$ and $\sigma_1 \sim \mathrm{Exp}(\kappa)$. Hence, we also have $P_*(1,2) \;=\; 1 - P_*(1,0) \;=\; \frac{\lambda}{\lambda+\kappa}$.

▷ Similarly, we find that

$$\begin{aligned} P_*(2,1) \;&=\; \mathbb{P}\,(X_t = 1 \mid X_0 = 2) \\ &=\; \mathbb{P}\,(\text{one of the two taxi rides ends before a new customer arrives}) \\ &=\; \mathbb{P}\,(\min\{\sigma_1, \sigma_2\} < \tau_1) \;=\; \frac{2\kappa}{\lambda + 2\kappa}, \end{aligned}$$

where $\tau_1 \sim \mathrm{Exp}(\lambda)$ and $\min\{\sigma_1, \sigma_2\} \sim \mathrm{Exp}(2\kappa)$, and thus, $P_*(2,3) \;=\; 1 - P_*(2,1) \;=\; \frac{\lambda}{\lambda+2\kappa}$.

This gives the same matrix (11.7) as above. ■

11.3 Transition semigroup for general continuous-time Markov chains

Recall from Corollary 10.7 that the transition semigroup $P = (P_t)_{t\in[0,\infty)}$ of continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ with generator matrix Q satisfies Kolmogorov's backward differential equation (10.8):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} P_t \;&=\; Q \cdot P_t, \qquad t \in [0,\infty), \\ P_0 \;&=\; I \end{aligned} \tag{11.8}$$

By comparison to the analogous ordinary differential equation $f'(t) = af(t)$, we may infer that the transition semigroup is given by

$$P_t \;=\; e^{tQ} \;=\; \sum_{n=0}^{\infty} \frac{(tQ)^n}{n!}, \qquad t \in [0,\infty).$$

This is indeed true, as we shall show below.

Definition. The *matrix exponential* (*exponenttimatriisi*) of a (possibly infinite) square-matrix A is defined as a square-matrix

$$e^A \;=\; \sum_{n=0}^{\infty} \frac{A^n}{n!}$$

so that the (x,y):th entry of e^A equals

$$e^A(x,y) \;=\; \lim_{N\to\infty} \sum_{n=0}^{N} \frac{1}{n!} A^n(x,y) \;=\; \sum_{n=0}^{\infty} \frac{1}{n!} A^n(x,y). \tag{11.9}$$

It can be shown that the limit in (11.9) exists and gives a well-defined square-matrix whenever

$$\|A\| \;=\; \max_{x\in S} \sum_{y\,\in\, S} |A(x,y)| \;<\; \infty.$$

The quantity $\|A\|$ defines a norm on those matrices for which it is finite. This holds in particular for the scaled generator matrix $t\,Q$ of any continuous-time Markov chain with uniformly bounded jump rates (11.3). Indeed, by Lemma 10.8 and Theorem 11.4, we have

$$\begin{aligned}\sum_{y\,\in\, S} |t\,Q(x,y)| \;&=\; t\,|Q(x,x)| \;+\; t \sum_{y\neq x} Q(x,y) \;=\; 2\,t|Q(x,x)| \\ &=\; 2\,t\lambda(x) \;\leq\; 2\,t\Lambda, \qquad \text{for all } x \in S,\end{aligned}$$

which shows that

$$\|t\,Q\| \;=\; \max_{x\in S} \sum_{y\,\in\, S} |t\,Q(x,y)| \;\leq\; 2\,t\Lambda \;<\; \infty. \tag{11.10}$$

Theorem 11.7 (*Transition semigroup as matrix exponentials*)**.** *Transition matrices* $P = (P_t)_{t\in[0,\infty)}$ *of continuous-time Markov chain* $X = (X_t)_{t\in[0,\infty)}$ *with* uniformly bounded jump rates (11.3) *are given in terms of the generator matrix* Q *by the* matrix exponential

$$P_t \;=\; e^{t\,Q} \;=\; \sum_{n\,=\,0}^{\infty} \frac{t^n Q^n}{n!}. \tag{11.11}$$

Proof sketch. By Theorem 11.4, the generator matrix $Q = \frac{\mathrm{d}}{\mathrm{d}t} P_t|_{t=0}$ exists, and $\|t\,Q\| < \infty$ by (11.10). Hence, the matrix exponential $e^{t\,Q}$ is well-defined. One can also show that Kolmogorov's backward differential equation (11.8) in Corollary 10.7 with initial condition $P_0 = I$ has a unique solution. It thus remains to show that the matrix exponential

$$\tilde{P}_t \;=\; e^{t\,Q} \;=\; \sum_{n\,=\,0}^{\infty} \frac{(tQ)^n}{n!} \tag{11.12}$$

satisfies the same initial value problem (11.8). Separating the $n = 0$ term from the sum (11.12) gives

$$\tilde{P}_t \;=\; \frac{(tQ)^0}{0!} + \sum_{n\,=\,1}^{\infty} \frac{(tQ)^n}{n!} \;=\; I + \sum_{n\,=\,1}^{\infty} \frac{t^n Q^n}{n!}.$$

Differentiating $\tilde{P}_t$ term-by-term (which can be justified since $\|tQ\| < \infty$) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{P}_t \;=\; \frac{\mathrm{d}}{\mathrm{d}t} I + \sum_{n\,=\,1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}t} \frac{t^n Q^n}{n!} \;=\; 0 + \sum_{n\,=\,1}^{\infty} \frac{n t^{n-1} Q^n}{n!} \;=\; Q \sum_{n\,=\,1}^{\infty} \frac{t^{n-1} Q^{n-1}}{(n-1)!} \;=\; Q\tilde{P}_t,$$

where in the last equality we recognize that the sum is the same as in (11.12) (the exchange of infinite sum and multiplication by Q in the second-to-last equality can be justified since $\|Q\| < \infty$). Hence $\tilde{P}_t$ satisfies Kolmogorov's backward differential equation (11.8), and it also has the correct initial value:

$$\tilde{P}_0 \;=\; I + \sum_{n\,=\,1}^{\infty} \frac{(0Q)^n}{n!} \;=\; I.$$

By uniqueness of the solution to (11.8) we conclude that $P_t = \tilde{P}_t = e^{tQ}$ for every $t \in [0,\infty)$. □

Let us summarize the properties of continuous-time Markov chains. The following hold for any continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ with uniformly bounded jump rates (11.3).

▷ (Theorem 11.4): The generator matrix exists as the limit

$$Q = \lim_{t\to 0} \frac{1}{t}(P_t - P_0) = \frac{\mathrm{d}}{\mathrm{d}t} P_t(x,y)\Big|_{t=0}. \tag{11.13}$$

▷ (Theorem 11.7): The transition matrices $P = (P_t)_{t\in[0,\infty)}$ are given in terms of Q as

$$P_t = e^{tQ} = \sum_{n=0}^{\infty} \frac{t^n Q^n}{n!}.$$

▷ (Theorem 10.5): The transition matrices $P = (P_t)_{t\in[0,\infty)}$ form a transition semigroup:

$$P_s \cdot P_t = P_{s+t}, \qquad \text{for all } s,t \geq 0.$$

▷ (Corollary 10.7): The transition matrices solve Kolmogorov's backward differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t = Q \cdot P_t, \qquad t \in [0,\infty),$$

and Kolmogorov's forward differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t = P_t \cdot Q, \qquad t \in [0,\infty).$$

(The latter can be proven completely similarly to Corollary 10.7.)

▷ (Theorem 10.10): Probability distribution π is an *invariant distribution* of X if and only if the *balance equation* $\pi \cdot Q = 0$ holds (cf. (10.13)):

$$\sum_{x\neq y} \pi(x)\cdot Q(x,y) = \pi(y) \cdot \sum_{z\neq y} Q(y,z).$$

▷ (Theorems 11.2 & 11.3): The arrival times $\tau_1, \tau_2, \ldots$ of jumps for $X = (X_t)_{t\in[0,\infty)}$ are independent and identically distributed,

$$\tau_n \sim \mathrm{Exp}(\lambda(x)) \qquad \text{on the event } \{X_{T_{n-1}} = x\},$$

and the random variable X_{T_n} is independent of τ_n, conditioned on the event $\{X_{T_{n-1}} = x\}$.

11.4 Poisson modulated Markov chains

So-called *Poisson modulated chains* provide a rich and versatile class of continuous-time Markov chains. In fact, *all* continuous-time Markov chains with uniformly bounded jump rates (11.3) can be represented as Poisson modulated chains, see Section 11.5. The idea is to take an underlying discrete-time Markov chain and add randomness to *when* the jumps happen.

Definition. A *Poisson modulated chain* (*Poisson-moduloitu Markov-ketju*) on countable state space S is a Markov chain $X = (X_t)_{t\in[0,\infty)}$ of the form

$$X_t = Y_{N(t)}, \qquad t \in [0,\infty),$$

where

▷ $Y = (Y_n)_{n\in\mathbb{N}_0}$ is a discrete-time Markov chain on state space S, and

▷ $N = (N(t))_{t\in[0,\infty)}$ is a Poisson process with intensity λ which is independent of Y.

▷ Because the waiting times of the jumps of the Poisson process $N = (N(t))_{t\in[0,\infty)}$ are independent and $\mathrm{Exp}(\lambda)$-distributed (by Theorem 8.5), it possible to show using the memoryless property (Theorem 8.7) that $X = (Y_{N(t)})_{t\in[0,\infty)}$ is indeed a continuous-time Markov chain on state space S. We leave this as an exercise for an mathematically oriented reader.

▷ Any Poisson process can be seen as a special instance of a Poisson modulated chain, where $Y_n = n$ is a Markov chain on $\mathbb{N}_0$ which deterministically moves one step up at every discrete time step. (Recall Example 10.1.)

Theorem 11.8 (*Poisson modulated chain*)**.** *For any Poisson modulated Markov chain $X = (Y_{N(t)})_{t\in[0,\infty)}$, the generator matrix is given by*

$$Q = \lambda\cdot(P-I), \tag{11.14}$$

where P is the transition matrix of the discrete-time chain Y.

Furthermore, X has the same invariant distributions as Y.

Proof. Let $\{T_1, T_2, \ldots\}$ be the Poisson point process of jump instants of the Poisson process N. Then, by definition of $X = (X_t)_{t\in[0,\infty)} = (Y_{N(t)})_{t\in[0,\infty)}$, we have (see Figure 11.1)

$$X_t = \begin{cases} Y_0, & 0 \le t < T_1, \\ Y_1, & T_1 \le t < T_2, \\ Y_2, & T_2 \le t < T_3, \\ \vdots \end{cases}$$

We compute the transition matrices of X using powers of the underlying discrete-time transition matrix P. Conditioning on the Poisson-distributed number $N(t) \sim \mathrm{Poi}(\lambda t)$ gives

$$\begin{aligned} \mathbb{P}(X_t = y \mid X_0 = x) &= \sum_{n=0}^{\infty} \mathbb{P}(Y_n = y \mid Y_0 = x)\cdot\mathbb{P}(N(t) = n) \\ &= \sum_{n=0}^{\infty} P^n(x,y)\cdot e^{-\lambda t}\frac{(\lambda t)^n}{n!} = e^{-\lambda t}\sum_{n=0}^{\infty}\frac{(\lambda t)^n}{n!}\cdot P^n(x,y), \end{aligned}$$

which shows that the time-t transition matrix P_t of the Poisson modulated chain X is given by

$$P_t = e^{-\lambda t}\sum_{n=0}^{\infty}\frac{(\lambda t P)^n}{n!} = e^{-\lambda t I}\cdot e^{\lambda t P}, \tag{11.15}$$

where $e^{-\lambda t I}$ is the matrix exponential of $-\lambda t$ times the identity matrix I, and $e^{\lambda t P}$ is the matrix exponential of λt times the discrete-time transition matrix P. By applying the formula $e^A\cdot e^B = e^{A+B}$ (valid whenever the matrices A and B commute, i.e., $AB = BA$), we find that

$$P_t = e^{-\lambda t I}\cdot e^{\lambda t P} = e^{\lambda t(P-I)} = e^{tQ},$$

where $Q = \lambda\cdot(P-I)$ is the generator matrix for X by Theorem 11.7. This proves Equation (11.14).

To prove the second claim, we multiply (11.14) by a distribution π from both sides:

$$\pi Q = \lambda\cdot(\pi P - \pi).$$

From this, we see that $\pi Q = 0$ if and only if $\pi P = \pi$. Now, by Theorem 10.10 and Equation (2.2), these are precisely the conditions for π to be invariant for X and Y, respectively. □

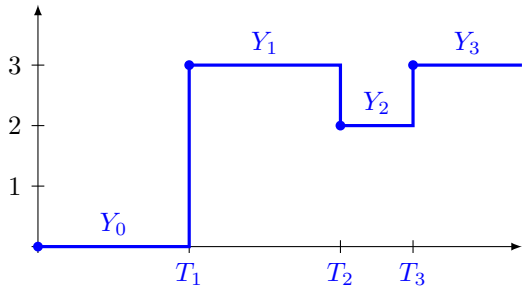


Figure 11.1: Path of a Poisson modulated chain on state space $\mathbb{N}_0$.

11.5 Uniformization of continuous-time Markov chains (overclocking)

We prove next that all continuous-time Markov chains with uniformly bounded jump rates (11.3) can be represented as Poisson modulated chains. This is a useful tool for *simulating* continuous-time Markov chains, as simulating a Poisson process is relatively easy. Furthermore, if one is just interested in the long-term behavior of a continuous Markov chain, then by Theorem 11.8 it suffices to simulate the discrete-time Markov chain under the Poisson-modulated chain.

Consider continuous-time Markov chain $X = (X_t)_{t\in[0,\infty)}$ on countable state space S, with jump probability matrix P_*. We will assume throughout that jump rates $\{\lambda(x) : x \in S\}$ are uniformly bounded by a constant $\Lambda \in (0,\infty)$ as in (11.3). This holds, for example, whenever state space S is finite. We aim to prove that X can be represented as *Poisson modulated chain* built from the data

$$\{\lambda(x) : x \in S\} \qquad \text{and} \qquad \{P_*(x,y) : x,y \in S\}.$$

To see this, we define a Poisson modulated chain

$$\hat{X}_t = \hat{Y}_{N(t)}, \qquad t \in [0,\infty), \tag{11.16}$$

where the jumps are governed by a Poisson process $(N(t))_{t\in[0,\infty)}$ with constant intensity Λ, and $\hat{Y} = (\hat{Y}_0, \hat{Y}_1, \hat{Y}_2, \ldots)$ is an independent discrete-time Markov chain with transition matrix

$$\hat{P}(x,y) = \frac{\lambda(x)}{\Lambda} \cdot P_*(x,y) + \Big(1 - \frac{\lambda(x)}{\Lambda}\Big) \cdot I(x,y) \tag{11.17}$$

$$= \frac{\lambda(x)}{\Lambda} \cdot P_*(x,y) + \Big(1 - \frac{\lambda(x)}{\Lambda}\Big) \cdot \mathbb{1}\{x = y\}, \qquad x,y \in S. \tag{11.18}$$

This matrix $\hat{P}$ represents a discrete-time Markov chain $\hat{Y}$ where at every time step we flip a coin, and

- ▷ with probability $\frac{\lambda(x)}{\Lambda}$ we move from x according to transition matrix $P_*(x,y)$, while
- ▷ with probability $1 - \frac{\lambda(x)}{\Lambda}$ we move from x according to transition matrix $I(x,y)$ (that is, we don't move anywhere).

We prove that the Poisson modulated Markov chain (11.16) thus defined is, in fact, stochastically the same as our original Markov chain X:

Theorem 11.9 (*Uniformization*). *Let $X = (X_t)_{t\in[0,\infty)}$ be a continuous-time Markov chain with bounded jump rates* (11.3)*, transition matrix Q, and jump probability matrix P_*. Let $(N(t))_{t\in[0,\infty)}$ be a Poisson process with intensity Λ, and $\hat{Y} = (\hat{Y}_0, \hat{Y}_1, \hat{Y}_2, \dots)$ a discrete-time Markov chain with transition matrix $\hat{P}$ with entries*

$$\hat{P}(x,y) \;=\; \frac{\lambda(x)}{\Lambda}\cdot P_*(x,y) \;+\; \Big(1-\frac{\lambda(x)}{\Lambda}\Big)\cdot I(x,y), \qquad x,y\in S.$$

Let $\hat{X} = (\hat{X}_t)_{t\in[0,\infty)}$ be the Poisson modulated chain

$$\hat{X}_t \;=\; \hat{Y}_{N(t)}.$$

Then, the chains $\hat{X}$ and X are stochastically the same.

Proof. Since by Theorems 11.2 and 11.3, continuous-time Markov chains are uniquely determined by the jump-rates $\lambda(x)$ and jump probabilities P_*, both of which are recoverable from the generator matrix Q by Theorem 11.4, it is sufficient to verify that the generator matrices of $\hat{X}$ and X agree. On the one hand, by Theorem 11.8 the generator matrix of $\hat{X}$ is given by

$$\begin{aligned}\hat{Q}(x,y) \;&=\; \Lambda\cdot(\hat{P}(x,y)-I(x,y)) \;=\; \lambda(x)\cdot P_*(x,y)-\lambda(x)\cdot\mathbb{1}\{x=y\} \qquad \text{[by (11.17)]}\\ &=\; \begin{cases}\lambda(x)\cdot P_*(x,y), & x\neq y\\ -\lambda(x), & x=y.\end{cases}\end{aligned} \tag{11.19}$$

On the other hand, by Theorem 11.4 the generator matrix of X is given by (11.5), which coincides with (11.19). □

12 Markov chain Monte Carlo methods

In this section, we again consider discrete-time Markov chains on countable state spaces.

Historically, "Monte Carlo methods" were initiated by physicists at Los Alamos Laboratory during World War II (in investigations related to nuclear bombs). It is a general term for various algorithmic methods that rely on random sampling (arguably, Monte Carlo is a famous casino). The original Monte Carlo algorithm is based on the law of large numbers and can already be quite effective for obtaining numerical results. Markov Chain Monte Carlo (MCMC) algorithms (such as the famous Metropolis algorithm) are improvements involving an exploration on the system of interest in the form of a random walk, or a more general Markov chain. These methods have gained magnificent success in various areas: optimization, numerical integration, various engineering subjects, probability theory, mathematical physics, risk management, business models, etc. Though, quoting Alan Sokal (one of the pioneers in the subject) [Sok89] — "*Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse*" — often there are no good alternatives!

The central idea in MCMC is to design a judicious Markov chain model with prescribed invariant distribution, and a sampling algorithm for it. The basis of MCMC is the Ergodic theorem (this is a generalization of Theorem 3.1 to countable state spaces):

Theorem 12.1 (*Ergodic theorem*)**.** *For any* irreducible *(discrete-time) Markov chain* $X = (X_0, X_1, X_2, \ldots)$ *with invariant distribution* π *on countable state space* S*, we have*

$$\frac{1}{t}\sum_{s=0}^{t-1} \phi(X_s) \longrightarrow \sum_{y\in S} \pi(y)\cdot\phi(y), \qquad \text{as } t\to\infty,$$

for any function $\phi : S \to \mathbb{R}$ *with probability one, regardless of the initial state of the Markov chain.*

Indeed, the Ergodic Theorem guarantees that the invariant distribution is well approximated by the empirical measures of the random states of the MCMC sampling algorithm.

12.1 Applications to numerical integration

The Monte Carlo method for numerical integration was developed by Stanislaw Ulam in the 1940s at Los Alamos Laboratory, where he was investigating neutron diffusion in the core of a nuclear weapon. Essentially, the problem is to estimate with good enough precision integrals of complicated functions, which are frequent in applications.

Example 12.2 (*Numerical integration*)**.** As the first example, consider function $f : [0,1] \to \mathbb{R}$. Sample iid points $X_1, X_2, \ldots, X_n$ in $[0,1]$ uniformly at random. Then if n is large, we have

$$\int_{[0,1]} f(x)\,\mathrm{d}x \approx \frac{1}{n}\sum_{k=0}^{n} f(X_k),$$

by the law of large numbers (or the Ergodic Theorem), since $X_1, X_2, \ldots, X_n$ are independent. The error made in this approximation of course depends on n,

$$\varepsilon(n) = \frac{1}{n}\sum_{k=0}^{n}\Big(f(X_k) - \int_{[0,1]} f(x)\,\mathrm{d}x\Big),$$

and we can estimate it quite easily. Since each X_k is sampled uniformly from $[0,1]$, the random numbers appearing in the sum,

$$Z_k \;=\; f(X_k) - \int_{[0,1]} f(x)\,\mathrm{d}x$$

have zero expected value (in mathematical terms, one often says that they are *centered*):

$$\mathbb{E}\,(Z_k) \;=\; 0.$$

Hence, we obtain

$$\begin{aligned}
\mathbb{E}\,(\varepsilon(n)^2) \;&=\; \mathbb{E}\left(\frac{1}{n^2}\sum_{k,\ell\,=\,0}^{n} Z_k \cdot Z_\ell\right)\\
&=\; \frac{1}{n^2}\sum_{k,\ell\,=\,0}^{n} \mathbb{E}\,(Z_k \cdot Z_\ell)\\
&=\; \frac{1}{n^2}\sum_{k\,=\,0}^{n} \mathbb{E}\,(Z_k^2)\\
&=\; \frac{1}{n^2}\sum_{k\,=\,0}^{n} \mathbb{E}\,(Z_1^2)\\
&=\; \frac{n}{n^2}\mathbb{E}\,(Z_1^2)\\
&=\; \frac{1}{n}\,\mathbb{E}\,(Z_1^2).
\end{aligned}$$

This goes to zero as $n \to \infty$ quite fast. ■

In the above method, when applying it numerically, it would be more effective to sample the test points X_k from intervals where the absolute value of the function f is large, which give the main contribution to the integral

$$\int_{[0,1]} f(x)\,\mathrm{d}x.$$

Therefore, one would want to focus on neighborhoods of the *local maxima* of the function $|f|$.

For functions $f : \mathbb{R}^k \to \mathbb{R}$ whose domain has high dimension k, the above method has clear drawbacks: sampling points uniformly from $\mathbb{R}^k$ is impossible, and finding the local maxima of the function $|f|$ is computationally extremely costly as k grows. See also Example 12.5.

In 1953 Metropolis, Rosenbluth, Teller, and Teller developed a method based on a refinement of the above ideas: instead of iid random test points, consider a *Markov chain* on the domain of interest (Example 12.4). In this method, it is presumed that there is a probability distribution π for the large values of $|f|$ and the goal is to design a Markov chain $X = (X_0, X_1, X_2, \ldots)$ with π as its invariant distribution.

12.2 Metropolis-Hastings algorithm

Suppose now that we have some complicated probability distribution π, and we should generate samples from π. The idea in MCMC methods is that one designs a suitable Markov chain $X = (X_0, X_1, X_2, \ldots)$ whose invariant distribution is π, and lets it evolve for a long enough time (how long is usually a difficult problem, while in practical applications the precision is often deemed good enough by inspection — see also Theorem 12.14 for a matehamtical result).

However, it is not so easy in general to find a transition matrix P with given invariant distribution π. Also, there might be many such transition matrices, and one of the difficulties is also to determine ones that are

- ▷ computationally efficient to simulate, and
- ▷ for which the convergence to the *statistical equilibrium* π is fast (which can be measured in terms of so-called mixing time, see Equation (12.13) and Theorem 12.14).

Since it is difficult to find directly a suitable transition matrix P, it turns out to be practical to start with a suitable *proposal* $\{p(x,y) : x,y \in S\}$ for a transition matrix, which is chosen to be as easy to manipulate as possible.

Definition. For each state $x \in S$, transition matrix p determines the *proposal distribution* (*ehdotusjakauma*) via

$$p(x,y) \;=\; \mathbb{P}\,(Y_1 = y \mid Y_0 = x), \qquad y \in S,$$

where $Y = (Y_0, Y_1, Y_2, \ldots)$ is the Markov chain with transition matrix p.

Recall that the detailed balance equations (5.11) (reversibility),

$$\pi(x) \cdot P(x,y) \;=\; \pi(y) \cdot P(y,x), \qquad \text{for all } x,y \in S, \tag{12.1}$$

guarantee the existence of an invariant distribution. Assume that our transition matrix p fails to satisfy (12.1): if we have

$$\pi(x) \cdot p(x,y) \;>\; \pi(y) \cdot p(y,x), \qquad \text{for some } x,y \in S, \tag{12.2}$$

then the proposal distribution causes the Markov chain to move from x to y too often. Therefore, we want to *tone down transitions from x to y*: we modify the transition probabilities in p by accepting transitions from x to y with probability $\alpha(x,y)$, called *acceptance probability* (*hyväksymistodennäköisyys*). Once we do this, we can make an ansatz

$$P(x,y) \;=\; \begin{cases} p(x,y) \cdot \alpha(x,y), & x \neq y, \\ 1 - \sum\limits_{z \neq x} p(x,z) \cdot \alpha(x,z), & x = y, \end{cases}$$

where it remains to determine the acceptance probabilities α. Since the Markov chain moves from y to x too rarely, we set $\alpha(y,x) = 1$. Then, from the detailed balance equations (12.1), we see that

$$\pi(x) \cdot p(x,y) \cdot \alpha(x,y) \;=\; \pi(y) \cdot p(y,x) \cdot \underbrace{\alpha(y,x)}_{=\,1} \;=\; \pi(y) \cdot p(y,x),$$

from which we can solve

$$\alpha(x,y) \;=\; \frac{\pi(y)}{\pi(x)} \cdot \frac{p(y,x)}{p(x,y)}.$$

Theorem 12.3 (*Metropolis-Hastings algorithm*). *Let π be a probability distribution on countable state space S. Let P and p be two transition matrices on S such that*

$$P(x,y) = \begin{cases} p(x,y)\cdot\alpha(x,y), & x \neq y, \\ 1 - \sum_{z\neq x} p(x,z)\cdot\alpha(x,z), & x = y, \end{cases} \tag{12.3}$$

where[a]

$$\alpha(x,y) = \min\left\{\frac{\pi(y)\cdot p(y,x)}{\pi(x)\cdot p(x,y)},\ 1\right\}. \tag{12.4}$$

Then, P and π satisfy the detailed balance equations (12.1)*, that is, P is* π-reversible.

[a] If $\pi(x)\cdot p(x,y) = 0$, Equation (12.4) is understood as $\alpha(x,y) = 1$.

Proof. One can verify this by a direct computation, which we leave as an exercise. □

Note that the Theorem 12.3 does not guarantee *uniqueness* of the invariant distribution. Indeed, if some states have zero probability, then the Markov chain with transition matrix P may contain several components, in which case the Metropolis-Hastings algorithm only explores one component and might not give a representative sample. However, if the proposal transition matrix p is irreducible and $\pi(x) > 0$ for all $x \in S$, then the transition diagram of the Markov chain coincides with the one given by p. In this case, the Markov chain is irreducible and π is its unique invariant distribution.

Importantly, to compute acceptance probabilities (12.4), we only have to compute the *ratios*

$$\frac{\pi(y)}{\pi(x)},$$

so that we do not in fact need to know the whole distribution π but only the *relative* proportions. Indeed, in many applications (especially in statistical physics) the distribution has the form

$$\pi(x) = \frac{\omega(x)}{\sum_{y\in S}\omega(y)},$$

where only the probability amplitudes $\omega(x)$ are known, but the normalizing factor

$$Z = \sum_{y\in S}\omega(y)$$

is hard to compute. Thus, it may be very difficult to analyze the distribution π directly, but for the Metropolis-Hastings algorithm, *one does not need to compute Z*. We will return to this in Examples 12.6 and 12.8.

Example 12.4. *Metropolis algorithm* (*Metropolis-algoritmi*) is a special case of Metropolis-Hastings algorithm[38]. In the Metropolis algorithm, the proposal distribution is assumed to be *symmetric*:

$$p(x,y) = p(y,x), \qquad \text{for all } x,y \in S.$$

[38] Historically, it was developed in 1953 by Metropolis, Rosenbluth, Teller, and Teller, for a specific stationary distribution. Hastings later in 1970 generalized it to the form above.

From the detailed balance (12.1) equations, we then see that

$$\alpha(x,y) \;=\; \frac{\pi(y)}{\pi(x)} \cdot \frac{p(y,x)}{p(x,y)} \;=\; \frac{\pi(y)}{\pi(x)}.$$

■

Example 12.5 (*Optimization*)**.** Consider function $f : V \to \mathbb{R}$, where $V = \{x_1, x_2, \ldots, x_n\}$ is the node set of a large undirected graph $G = (V, E)$ with edge set E. We call nodes x and y *neighbors* if $(x, y) \in E$ is an edge in G.

It is often of interest to find those areas in V where f is large — for example, when the graph approximates some high-dimensional domain of a real-valued function (e.g. in an analogue of Example 12.2 for a function from $\mathbb{R}^k$ to $\mathbb{R}$). However, for $n = |V|$ very large, searching the relevant areas in V may be computationally very expensive. Designing a suitable MCMC method can provide an algorithm in order $\log n$, as Theorem 12.15 discussed in Section 12.5 shows.

The *hill-climb* ("steepest ascent") algorithm is a Markov chain $X = (X_0, X_1, X_2, \ldots)$ which moves in V as follows:

- ▷ it starts at some node $x \in V$, so that $X_0 = x$,
- ▷ it checks whether at any neighbor of x, the value of f is (strictly) larger:

  $$f(y) \;>\; f(x) \text{ for some neighbor } y \in V \text{ of } x \qquad \Longrightarrow \qquad X_1 \;=\; \arg \max_{y\,:\,(x,y)\in E} f(y),$$

  that is, it moves to the neighbor where the value of f is the largest,
- ▷ and if there is no such neighbor, it stays put:

  $$f(y) \;\le\; f(x) \text{ for all neighbors } y \in V \text{ of } x \qquad \Longrightarrow \qquad X_1 \;=\; x.$$

This however is only a useful algorithm if one wants to find a *local maximum* of f. If f has many maxima, depending on where the Markov chain X starts from, it can get stuck in a different local maximum, which may not be the global maximum. One can improve the algorithm by adding randomness: allow X also to move to nodes with smaller values of f with some (small) probability. To demonstrate one possibility, define a distribution

$$\pi_a(x) \;=\; \frac{a^{f(x)}}{\sum\limits_{y\in S} a^{f(y)}} \;=\; \frac{a^{f(x)}}{Z_{\pi_a}},$$

where $a > 1$ is a constant, and the normalizing factor is

$$Z_{\pi_a} \;=\; \sum_{y\in S} a^{f(y)}.$$

Then, π_a is a probability distribution on V, which favors nodes where f is large (since $a > 1$).

As a convenient proposal distribution, consider random walk $Y = (Y_0, Y_1, Y_2, \ldots)$ on an undirected graph G (recall Example 4.2), that is, a Markov chain that proceeds by moving at each step to a neighboring node selected uniformly at random:

$$p(x,y) \;=\; \begin{cases} \frac{1}{\deg(x)}, & \text{if } x \text{ and } y \text{ are neighbors, i.e.,, } (x,y) \in E \\ 0, & \text{otherwise.} \end{cases}$$

The Metropolis-Hastings algorithm (Theorem 12.3) then gives a Markov chain on V whose invariant distribution is π_a. The acceptance probabilities (12.4) are given by

$$\alpha(x,y) \;=\; \min\left\{ a^{f(y)-f(x)} \cdot \frac{\deg(x)}{\deg(y)},\; 1 \right\}. \tag{12.5}$$

Indeed, when $a \to \infty$, we see that the chain becomes the deterministic hill-climb (steepest ascent) method. More precisely, writing

$$f_{\max} = \max_{y \in V} f(y) \qquad \text{and} \qquad V_{\max} = \{y \in V : f(y) = f_{\max}\},$$

we obtain

$$\begin{aligned}
\pi_a(x) = \frac{a^{f(x)}}{\sum_{y \in S} a^{f(y)}} &= \frac{a^{f(x)}}{\sum_{y \in V_{\max}} a^{f(y)} + \sum_{y \notin V_{\max}} a^{f(y)}} \\
&= a^{f(x)-f_{\max}} \cdot \frac{1}{\sum_{y \in V_{\max}} a^{f(y)-f_{\max}} + \sum_{y \notin V_{\max}} a^{f(y)-f_{\max}}} \\
&= a^{f(x)-f_{\max}} \cdot \frac{1}{\sum_{y \in V_{\max}} 1 + \sum_{y \notin V_{\max}} a^{f(y)-f_{\max}}} \\
&= a^{f(x)-f_{\max}} \cdot \frac{1}{|V_{\max}| + \sum_{y \notin V_{\max}} a^{f(y)-f_{\max}}} \\
&\xrightarrow{a \to \infty} \frac{\mathbb{1}\{x \in V_{\max}\}}{|V_{\max}|},
\end{aligned}$$

since $f(y) - f_{\max} < 0$ for $y \notin V_{\max}$ and $a > 1$, so $a^{f(y)-f_{\max}} \longrightarrow 0$ as $a \to \infty$. We conclude that

$$\lim_{a \to \infty} \pi_a(x) = \frac{\mathbb{1}\{x \in V_{\max}\}}{|V_{\max}|},$$

so π_a converges as $a \to \infty$ to the uniform distribution on the global maxima $V_{\max}$ of f. ■

12.3 Sampling a random function — local updates

In many applications, one would like to sample a function randomly. Generally, however, the set of all functions is very large, so direct sampling may be very difficult even for relatively simple distributions. Perhaps MCMC algorithms could help here? Below, we will consider random functions between two finite sets, but ideas presented here generalize also to more general settings.

For two sets A and V, we denote by $S = \mathrm{A}^V$ the set of all functions[39] from V to A:

$$S = \mathrm{A}^V = \{f : V \to \mathrm{A}\}.$$

Example 12.6 (*Boltzmann distributions*)**.** In statistical mechanics, one studies a large number of interacting particles with properties such as magnetization (Example 12.9), position, or velocity. One is often interested in the effect of temperature to the system. To numerically study this, one needs to be able to effectively simulate a statistical system for a wide range of temperatures.

A *state* of a statistical system is an assignment of a property to each particle. If we denote by V the (finite) set of particles of the system and by A the (finite) set of possible values that the property of each particle can take, then we can encode a state as a function $f : V \to \mathrm{A}$. The set of all possible states (that is, the state space) is thus naturally $S = \mathrm{A}^V$. The energy of

[39] The notation A^V can be understood as the set of A-valued vectors indexed by V, which in turn can be thought as a function from V to A. This justifies the notation.

the system depends on the state, so it is a function $H : S \to \mathbb{R}$ called the energy functional, or *Hamiltonian*. H commonly has the form

$$H(f) = \sum_{x \in V} U(f(x)) + \frac{1}{2} \sum_{\substack{x,y \in V \\ x \neq y}} I(f(x), f(y)), \tag{12.6}$$

where $U(f(x))$ represents the *potential energy* of particle x and $I(f(x), f(y))$ denotes the *interaction* between particles x and y. The interaction is assumed to be symmetric:

$$I(f(x), f(y)) = I(f(y), f(x)).$$

Now, if we denote by $\beta > 0$ the inverse temperature of the system, then (with some physically reasonable assumptions) the entropy of the system is maximized by the *Boltzmann distribution* (*Boltzmann-jakauma*) π on S given by

$$\pi(f) = \frac{1}{Z_\beta} e^{-\beta H(f)},$$

where the normalizing factor

$$Z_\beta = \sum_{f \in S} e^{-\beta H(f)}$$

is called the *partition function* in the physics context. The ratio of probabilities of states $f, g \in S$ equals

$$\frac{\pi(g)}{\pi(f)} = \frac{\frac{1}{Z_\beta} e^{-\beta H(g)}}{\frac{1}{Z_\beta} e^{-\beta H(f)}} = e^{\beta(H(f) - H(g))}. \tag{12.7}$$

Hence, to effectively apply Metropolis-Hastings algorithm to a Boltzmann distribution, one needs to be able to efficiently compute *differences* $H(f) - H(g)$ of energies between two states. With a clever choice of proposal distribution, plugging (12.6) to the difference $H(f) - H(g)$ results in cancellation of most terms in the sums. Thus, the difference of energies can be computed much faster than energy of a single state. We will return to this idea in Example 12.8. ■

Suppose now that we have a distribution π on the set $S = \mathrm{A}^V$ of functions that we want to sample using Metropolis-Hasting algorithm (Theorem 12.3). We shall build the proposal distribution $\{p(f, g) : f, g \in S\}$ via a procedure referred to as *local update* (*lokaali muutos*). It is performed by randomly changing the value of the function f in one place as follows.

Theorem 12.7 (*Local update*)**.** *Let V and A be finite sets, and let π be a distribution on the set $\mathrm{A}^V = \{f : V \to \mathrm{A}\}$ of functions. Let $X = (X_0, X_1, X_2, \ldots)$ be a Markov chain on state space $S = \mathrm{A}^V$ defined by* iterated local update: *given $X_t = f$, set X_{t+1} by the following algorithm:*

1. *Choose an element $v \in V$ uniformly at random.*
2. *Choose an element $a \in \mathrm{A} \setminus \{f(v)\}$ uniformly at random.*
3. *Define a function $\tilde{f}_v^a : V \to A$ by changing the value of $f(v)$ to a:*

$$\tilde{f}_v^a(x) = \begin{cases} f(x), & x \neq v, \\ a, & x = v. \end{cases} \tag{12.8}$$

4. *With probability $\alpha(f, \tilde{f}_v^a) = \min\{\frac{\pi(\tilde{f}_v^a)}{\pi(f)}, 1\}$ set $X_{t+1} = \tilde{f}_v^a$, and otherwise, keep $X_{t+1} = f$.*

Then, the Markov chain X is π-reversible, and π is an invariant distribution of X.

Note that by using (12.3), we can find the entries of the transition matrix P of X.

Proof. Let us write $g \leftrightarrow f$ for functions $f, g \in \mathrm{A}^V$ satisfying (12.8) with some v and a. As there are $|V|$ possible choices for $v \in V$ and $|\mathrm{A}| - 1$ choices for $a \in \mathrm{A} \setminus \{f(v)\}$, and we make these choices independently and uniformly at random, we see that the relevant proposal distribution is given by

$$p(f,g) = \begin{cases} \dfrac{1}{|V| \cdot (|A| - 1)}, & g \leftrightarrow f, \\ 0, & \text{otherwise.} \end{cases}$$

Note that, after obtaining $g = \tilde{f}_v^a$, we can get back to f by changing the value of $\tilde{f}_v^a(v)$ back to a. Hence, the proposal distribution is symmetric ($p(f,g) = p(g,f)$), so this is the special case of Metropolis algorithm (Example 12.4). The acceptance probabilities are thus simply given by

$$\alpha(f,g) = \min\left\{\frac{\pi(g)}{\pi(f)},\ 1\right\}.$$

Theorem 12.3 implies that X is π-reversible, and π is an invariant distribution of X. □

Example 12.8 (*MCMC on Boltzmann distributions*)**.** Returning to Example 12.6, let us compute the acceptance probabilities $\alpha(f, \tilde{f}_v^a) = \min\{\frac{\pi(\tilde{f}_v^a)}{\pi(f)}, 1\}$ from Theorem 12.7 for $f, \tilde{f}_v^a \in \mathrm{A}^V$ satisfying (12.8). From Equations (12.7, 12.6), we obtain

$$\begin{aligned} \frac{\pi(\tilde{f}_v^a)}{\pi(f)} &= \exp\big(\beta(H(f) - H(\tilde{f}_v^a))\big) \\ &= \exp\Big(\beta \sum_{x \in V} \big(U(f(x)) - U(\tilde{f}_v^a(x))\big) + \frac{\beta}{2} \sum_{\substack{x,y \in V \\ x \neq y}} \big(I(f(x), f(y)) - I(\tilde{f}_v^a(x), \tilde{f}_v^a(y))\big)\Big). \end{aligned}$$

Since $\tilde{f}_v^a(x) = f(x)$ for all $x \neq v$, in the first sum only the term $x = v$ survives, while in the second sum the terms with $x = v$ or $y = v$ survive. Since I is assumed to be symmetric, we can rewrite the above formula as

$$\frac{\pi(\tilde{f}_v^a)}{\pi(f)} = \exp\Big(\beta\Big(U(f(v)) - U(\tilde{f}_v^a(v))\Big) + \beta \sum_{\substack{x \in V \\ x \neq v}} \big(I(f(v), f(x)) - I(\tilde{f}_v^a(v), \tilde{f}_v^a(x))\big)\Big),$$

Hence, instead of computing energies between every pair of particles (the number of which is $|V|^2$), one needs to only compute energies between each particle with particle v, which we have updated (which only requires $|V|$ computations). If the long-range interaction between particles is small, it may even be enough to compute the interaction only for those $x \in V$ which are "near enough" to v for the interaction energy to contribute to the total energy. See Example 12.9. ■

Let us apply Example 12.8 to a system where particles are constrained to lie on a graph. Let us also assume that particles living on the nodes only interact with their neighboring particles.

Example 12.9 (*Ising model*)**.** Consider an undirected graph $G = (V, E)$ with finite node set V and edge set E. We call nodes x and y *neighboring* if $(x, y) \in E$ is an edge in G. A *spin configuration* is an assignment

$$\sigma \in \Sigma = \{-1, +1\}^V = \{f : V \to \{-1, +1\}\}$$

where -1 and $+1$ represent *spins* at the nodes. *Ising model* is a Boltzmann distribution on the set Σ of all spin configurations, modelling magnets each (node $x \in V$) having one of the two

possible orientations represented by -1 and $+1$. A spin configuration represents the random orientations of these magnets. The Hamiltonian (12.6) is[40]

$$H(\sigma) \;=\; -\sum_{\substack{x,y\in V\\(x,y)\in E}} \sigma(x)\cdot\sigma(y)$$

(so the potential energy equals zero, and the interaction is trough nearest neighbors).

Local updates according to Theorem 12.7 flip possibly one spin at some node $v \in V$ at a time. Recall that the associated Markov chain $X = (X_0, X_1, X_2, \ldots)$ on state space $S = \Sigma$ is given by the following transitions: if $X_t = \sigma$, then

1. choose an element $v \in V$ uniformly at random;
2. flip the spin $\sigma(v)$ at v (since there are just two possible spins, there is no choice in this step), so it becomes $-\sigma(v)$;
3. according to (12.8), define

$$\sigma_v(x) \;=\; \begin{cases}\sigma(x), & x \neq v,\\ -\sigma(v), & x = v;\end{cases} \tag{12.9}$$

4. and set $X_{t+1} = \sigma_v$ with acceptance probability $\alpha(\sigma,\sigma_v) = \min(\frac{\pi(\sigma_v)}{\pi(\sigma)}, 1)$, where by (12.7),

$$\frac{\pi(\sigma_v)}{\pi(\sigma)} \;=\; e^{\beta(H(\sigma)-H(\sigma_v))} \;=\; \exp\Big(-2\beta\sigma(v)\cdot\sum_{\substack{x\in V\\(x,v)\in E}}\sigma(x)\Big),$$

and otherwise, keep $X_{t+1} = \sigma$.

(This is discussed in more detail in the exercises.) ■

12.4 Convergence to statistical equilibrium

This section is primarily aimed for mathematically oriented readers.

A key result in the theory of Markov chains is the Convergence Theorem 5.6, which gives a uniqueness (but unfortunately not in general existence) criterion for π.

Theorem 12.10 (*Uniqueness of invariant distribution*)**.** *Every* irreducible *and* aperiodic *Markov chain on a countable state space admits* at most one *invariant distribution.*

If the invariant distribution π exists, then it also equals the unique limiting distribution,

$$\lim_{t\to\infty} \mathbb{P}\,(X_t = y \mid X_0 = x) \;=\; \pi(y), \qquad \text{for all } x, y \in S. \tag{12.10}$$

If the invariant distribution π exists (e.g., if the Markov chain can be verified to be π-reversible, recall Theorem 5.7) a basic but important problem of Markov chain theory concerns the *rate of convergence* in (12.10):

How long must the Markov chain be run to be "suitably close" to π? .

[40] Note that as $\beta \to 0$ (infinite temperature), the model just becomes the uniform distribution on Σ and the interaction plays no role. In contrast for large β the effects from the interaction dominate.

It is customary to measure distances between two probabilities by the so-called *total variation distance*. In essence, the total variation distance looks at the maximal difference of probabilities of events A measured by two probability distributions.

Definition. For two probability distributions μ and ν on state space S, the *total variation distance* (*kokonaisvaihteluetäisyys*) is

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subset S} |\mu(A) - \nu(A)|. \tag{12.11}$$

We can estimate the distance of the distribution μ_t^x of Markov chain X started at $X_0 = x$,

$$\mu_t^x(y) = \mathbb{P}(X_t = y \mid X_0 = x), \qquad y \in S,$$

(recall Theorem 5.3) to its statistical equilibrium π in terms of the total variation distance:

$$\|\mu_t^x - \pi\|_{\mathrm{TV}} = \max_{A \subset S} |\mu_t^x(A) - \pi(A)|. \tag{12.12}$$

Since in order to evaluate it (12.12), we would have to compute probabilities of all events, it is not very practical as such. However, there is an easier formula, given by the next result.

Theorem 12.11 (*Total variation distance*)**.** *For two probability distributions μ and ν on state space S, we have*

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_{y \in S} |\mu(y) - \nu(y)|.$$

Proof. Consider the set $B = \{y \in S : \mu(y) \geq \nu(y)\}$. Then, we have

$$\begin{aligned} \mu(A) - \nu(A) &\leq \mu(A \cap B) - \nu(A \cap B) \qquad [\text{as } \mu(y) - \nu(y) < 0 \text{ for all } y \in A \cap B^c] \\ &\leq \mu(B) - \nu(B), \end{aligned}$$

and similarly,

$$\nu(A) - \mu(A) \leq \nu(A \cap B^c) - \mu(A \cap B^c) \leq \nu(B^c) - \mu(B^c).$$

Now, observe that in fact, since μ and ν are probability distributions, the law of total probability gives

$$\mu(B) + \mu(B^c) = 1 = \nu(B) + \nu(B^c) \qquad \Longrightarrow \qquad \mu(B) - \nu(B) = \nu(B^c) - \mu(B^c).$$

Therefore, we obtain

$$|\mu(A) - \nu(A)| \leq \mu(B) - \nu(B), \qquad A \subset S,$$

and in particular,

$$|\mu(B) - \nu(B)| = \mu(B) - \nu(B) = \nu(B^c) - \mu(B^c),$$

so that the total variation distance (12.12) equals

$$\begin{aligned}\|\mu-\nu\|_{\mathrm{TV}} \;=\; \max_{A\subset S}|\mu(A)-\nu(A)| \;=\; |\mu(B)-\nu(B)| \;&=\; \frac{1}{2}\Big((\mu(B)-\nu(B))+(\nu(B^c)-\mu(B^c))\Big)\\ &=\; \frac{1}{2}\sum_{y\in S}|\mu(y)-\nu(y)|,\end{aligned}$$

as claimed. □

While Theorem 5.6 gives the convergence to the statistical equilibrium π (if it exists), it is in general a very hard problem to estimate the rate of convergence in

$$\|\mu_t^x-\pi\|_{\mathrm{TV}} \;=\; \max_{A\subset S}|\mu_t^x(A)-\pi(A)| \;=\; \frac{1}{2}\sum_{y\in S}|\mu_t^x(y)-\pi(y)| \qquad \stackrel{t\to\infty}{\longrightarrow} \qquad 0.$$

One can measure the required time that the Markov chain has to run before it is close to the statistical equilibrium by the so-called *mixing time*.

Definition. The *mixing time* (*sekoittumisaika*) for Markov chain X and parameter $\varepsilon>0$ is

$$t_{\mathrm{mix}}(\varepsilon) \;=\; \min\Big\{t\geq 0 : \max_{x\in S}\|\mu_t^x-\pi\|_{\mathrm{TV}} \;\leq\; \varepsilon\Big\}. \tag{12.13}$$

Note that it is, by definition, independent of the initial value $X_0=x$.

Rigorous upper bounds on mixing times provide us confidence that simulation studies or randomized algorithms indeed perform as advertised. See [LPW08] for a thorough discussion.

Example 12.12. Recall that in the proof of the Convergence Theorem 5.6, we found the upper bound (5.10),

$$\sum_{y\in S}\Big|\,\mu_t^x(y)\;-\;\pi(y)\,\Big| \;=\; \sum_{y\in S}\Big|\,\mathbb{P}\,(X_t=y\mid X_0=x)\;-\;\pi(y)\,\Big| \;=\; 2\mathbb{P}\,(\tau>t),$$

in terms of the first time (5.8)

$$\tau \;=\; \min\{t\geq 0 :\, X_t=Y_t\} \tag{12.14}$$

when two independent copies X and Y of our Markov chain (both having transition matrix P) such that X has initial state $X_0=x$ and Y has initial distribution π, meet. We see using Theorem 12.11 that

$$\|\mu_t^x-\pi\|_{\mathrm{TV}} \;=\; \frac{1}{2}\sum_{y\in S}|\mu_t^x(y)-\pi(y)| \;=\; \mathbb{P}\,(\tau>t).$$

Thus, the coupling method gives means to bound the total variation distance: estimates on the "tail distribution"[41] of the random meeting time τ, encoded in the probabilities $\{\mathbb{P}\,(\tau>t) : t\in[0,\infty]\}$, give estimates for the convergence rate of the Markov chain. ■

[41] This may a priori depend on the initial state $X_0=x$.

12.5 Convergence estimates for reversible chains

This section is primarily aimed for mathematically oriented readers.

In the Metropolis-Hastings algorithm, we constructed the transition matrix P to be π-reversible for simplicity. It turns out that π-reversibility is convenient also for studying convergence rates of the underlying Markov chain. Let us consider the case of a finite state space.

Suppose that $S = \{x_1, x_2, \ldots, x_n\}$ is finite with cardinality (size) $|S| = n$. Suppose that transition matrix P is π-reversible for some probability distribution π:

$$\begin{aligned} \pi(x) \cdot P(x,y) &= \pi(y) \cdot P(y,x), \qquad \text{for all } x, y \in S \\ \sum_{x \in S} \pi(x) &= 1. \end{aligned}$$

Then, there exists a diagonal matrix $\Lambda = \mathrm{diag}[\sqrt{\pi(1)}, \sqrt{\pi(2)}, \ldots, \sqrt{\pi(n)}]$ such that

$$P_{\mathrm{sym}} = \Lambda \cdot P \cdot \Lambda^{-1}$$

is a symmetric matrix. Indeed, we have

$$\begin{aligned} P_{\mathrm{sym}}(x,y) = (\Lambda \cdot P \cdot \Lambda^{-1})(x,y) &= \frac{\sqrt{\pi(x)}}{\sqrt{\pi(y)}} \cdot P(x,y) \\ &= \frac{\sqrt{\pi(y)}}{\sqrt{\pi(x)}} \cdot P(y,x) = P_{\mathrm{sym}}(y,x). \quad \text{[by detailed balance (12.1)]} \end{aligned}$$

(We know that $\pi(z) > 0$ by Theorem 5.4.) Hence, it is easy to study the spectral theory of P_{sym}:

- ▷ the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of P_{sym} are real,
- ▷ P_{sym} is diagonalizable and it has orthonormal eigenvectors $\psi_1, \psi_2, \ldots, \psi_n$,

$$P_{\mathrm{sym}} \cdot \psi_k = \lambda_k \cdot \psi_k, \qquad k = 1, 2, \ldots, n.$$

Since P is related to P_{sym} by conjugation by a diagonal matrix, also P has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{R}$, and corresponding right and left eigenvectors

$$\phi_k = \Lambda^{-1} \cdot \psi_k \qquad \text{and} \qquad \varphi_k = (\Lambda \cdot \psi_k)^T,$$

so that

$$P \cdot \phi_k = \lambda_k \cdot \phi_k \qquad \text{and} \qquad \varphi_k \cdot P = \varphi_k \cdot \lambda_k, \qquad k = 1, 2, \ldots, n.$$

We can write a spectral decomposition for the t:th power of P as

$$\begin{aligned} P^t &= \sum_{k=1}^{n} \varphi_k \cdot \lambda_k^t \cdot \phi_k \\ &= \sum_{k=1}^{n} (\Lambda \cdot \psi_k)^T \cdot \lambda_k^t \cdot \phi_k \\ &= \sum_{k=1}^{n} (\Lambda^2 \cdot \Lambda^{-1} \cdot \psi_k)^T \cdot \lambda_k^t \cdot \phi_k = \Lambda^2 \cdot \sum_{k=1}^{n} \phi_k^T \cdot \lambda_k^t \cdot \phi_k. \end{aligned}$$

Therefore, since $\Lambda^2 = \mathrm{diag}[\pi(1), \pi(2), \ldots, \pi(n)]$, we see that

$$P^t(x,y) = \pi(y) \cdot \sum_{k=1}^{n} \lambda_k^t \cdot \phi_k^T(x) \cdot \phi_k(y), \qquad \Longrightarrow \qquad \frac{P^t(x,y)}{\pi(y)} = \sum_{k=1}^{n} \lambda_k^t \cdot \phi_k(x)^T \cdot \phi_k(y). \tag{12.15}$$

(See [LPW08, Chapter 12] for more details.)

Lemma 12.13. *The following hold for any $n \times n$ transition matrix P.*

- ▷ *Its eigenvalues satisfy $|\lambda_k| \leq 1$ for all $k = 1, 2, \ldots, n$.*
- ▷ *1 is always an eigenvalue of P. The row-vector $[1, 1, \ldots, 1]^T$ is a right eigenvector of P with eigenvalue 1.*
- ▷ *If P is irreducible and aperiodic, then -1 is not an eigenvalue of P.*

Proof. Recall that transition matrix is a function $P : S \times S \to [0, 1]$ such that

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

The claims can be checked via relatively simple linear algebra using these properties[42]. □

By Lemma 12.13, we may order the eigenvalues of any π-reversible transition matrix P as

$$\lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 = 1.$$

Note also that $\lambda_n \geq -1$. Hence, with $\lambda_1 = 1$, the spectral decomposition from (12.15) yields

$$\frac{P^t(x, y)}{\pi(y)} = 1 + \sum_{k=2}^{n} \lambda_k^t \cdot \phi_k(x)^T \cdot \phi_k(y),$$

and if P is irreducible and aperiodic (so that -1 is not an eigenvalue), we see that the factors in the sum over $k \geq 2$ become small in absolute value as time grows:

$$|\lambda_k|^t \overset{t \to \infty}{\longrightarrow} 0.$$

Definition.

- ▷ The *spectral gap* (*spektrin rako*) of P is defined as
$$\gamma = 1 - \lambda_2.$$
- ▷ The *absolute spectral gap* (*spektrin itseisarvon rako*) of P is defined as
$$\gamma_* = 1 - \lambda_*,$$
where $\lambda_* = \max\{|\lambda_k| : \lambda_k \neq 1\}$ is the second largest eigenvalue of P in absolute value.

Note that, if P is irreducible and aperiodic, then it has a non-zero absolute spectral gap $\gamma_* > 0$ by Lemma 12.13 (since -1 is not an eigenvalue of P). The absolute spectral gap γ_* can be used to estimate the mixing time, as the next results shows.

[42] See [LPW08, Lemmas 12.1 and 12.2] for hints.

Theorem 12.14 (*Mixing time*). *Let P be irreducible, aperiodic, and π-reversible. Then, the mixing time* (12.13) *is bounded as*

$$\Big(\frac{1}{\gamma_*} - 1\Big) \log\Big(\frac{1}{2\varepsilon}\Big) \;\leq\; t_{\mathrm{mix}}(\varepsilon) \;\leq\; \frac{1}{\gamma_*} \log\Big(\frac{1}{\varepsilon\, \pi_{\min}}\Big)$$

where $\gamma_* = 1 - \lambda_*$ *and*

$$\pi_{\min} = \min_{y \in S} \pi(y) > 0.$$

Proof. See [LPW08, Theorems 12.4 and 12.5]. □

From Theorem 12.14, one can conclude that the convergence in Theorem 5.6 occurs at an exponential rate, that is governed by the second largest eigenvalue λ_* of P in absolute value.

Theorem 12.15 (*Exponential convergence*). *Let P be irreducible, aperiodic, and π-reversible. Then, the maximal total variation distance* (12.12) *of the time-t distribution of the corresponding Markov chain X to its statistical equilibrium π is bounded as*

$$\lim_{t \to \infty} \max_{x \in S} \| \mu_t^x - \pi \|_{\mathrm{TV}}^{1/t} = \lambda_*.$$

Proof. See [LPW08, Corollary 12.7]. □

We warmly recommend the lecture notes [LPW08, Corollary 12.7] for interested readers about mixing times. For more discussion and references especially related to MCMC methods, see [LPW08, Chapter 3] and [Sok89] (a bit old, but classic), as well as [RC10] highlighting applications.

References

[AE21] Robert A. Adams, Christopher Essex. *Calculus: A complete course.* Pearson Education Canada Inc., tenth edition, 2021.

[RC10] Christian Robert and George Casella. *Introducing Monte Carlo Methods with R.*. Springer, 2010.

[Dur12] Richard Durrett. *Essentials of Stochastic Processes.* Springer, second edition, 2012.

[GS97] Charles M. Grinstead, J. Laurie Snell. *Introduction to Probability.* American Mathematical Society, 1997.
`https://math.dartmouth.edu/~prob/prob/prob.pdf`.

[LPW08] David A. Levin, Yuval Peres, Elizabeth L. Wilmer. *Markov Chains and Mixing Times.* American Mathematical Society, 2008.
`https://pages.uoregon.edu/dlevin/MARKOV/mcmt2e.pdf`.

[Kal21] Olav Kallenberg. *Foundations of Modern Probability.* Springer, third edition, 2021.

[Kul16] Vidyadhar G. Kulkarni. *Modeling and Analysis of Stochastic Systems.* Chapman and Hall/CRC, third edition, 2016.

[Kyt20] Kalle Kytölä. *Probability theory.* Lecture notes at Aalto University, 2020.
`http://math.aalto.fi/~kkytola/files_KK/ProbaTh2019/ProbaTh-2019.pdf`

[Les20] Lasse Leskelä. Stochastic processes, 2020.
`https://math.aalto.fi/~lleskela/LectureNotes005.html`.

[SW08] Rolf Schneider and Wolfgang Weil. *Stochastic and Integral Geometry.* Springer, Berlin, 2008.

[Sok89] Alan Sokal. *Monte Carlo methods in statistical mechanics: foundations and new algorithms.* In Cours de Troisieme Cyle de la Physique en Suisse Romande, Lausanne, 1989.

[Str06] Gilbert Strang. *Linear Algebra and Its Applications.* Cengage Learning, fourth edition, 2006.

[Wil91] David Williams. *Probability with Martingales.* Cambridge University Press, 1991.

[Wil94] Herbert S. Wilf. Generatingfunctionology. Academic Press, Inc., 1994. `https://www2.math.upenn.edu/~wilf/gfology2.pdf`