Markov Chains for Everybody

An Introduction to the theory of discrete time Markov chains on countable state spaces.

Wilhelm Huisinga, &
Eike Meerbach

Fachbereich Mathematik und Informatik
Freien Universität Berlin &
DFG Research Center Matheon, Berlin
huisinga@mi.fu-berlin.de
meerbach@mi.fu-berlin.de

Berlin, January 26, 2005

Contents

1 A short introductional note **3**

2 Setting the scene **4**
2.1 Introductory example 4
2.2 Markov property, stochastic matrix, realization, density propagation 5
2.3 Realization of a Markov chain 10
2.4 The evolution of distributions under the Markov chain 12
2.5 Some key questions concerning Markov chains 17

3 Communication and recurrence **19**
3.1 Irreducibility and (A)periodicity 19
3.2 Recurrence and the existence of stationary distributions . . . 24

4 Asymptotic behavior **35**
4.1 k-step transition probabilities and distributions 35
4.2 Time reversal and reversibility 39
4.3 Some spectral theory 42
4.4 Evolution of transfer operators 47

5 Empirical averages **50**
5.1 The strong law of large numbers 50
5.2 Central limit theorem 53
5.3 Monte Carlo Methods 55
5.4 Simulated annealing 55

6 Identification of macroscopic properties **56**
6.1 Identification of communication classes 56
6.2 Identification of cyclic classes 61
6.3 Almost invariant communication classes 64
6.4 Metastability 64

1 A short introductional note

This script is a personal compilation of introductory topics about discrete time Markov chains on some countable state space. The choice of a countable state space is motivated by the fact that it is mathematically richer than the finite state space case, but still not as technically as general state space case. Furthermore, it allows for an easier generalization to the general state space Markov chains. Of course, this is only an introductory script that obviously lacks a lot of (important) topic— we explicitly encourage any interested student to study further, by referring to the literature provided at the end of this script. Furthermore we did our best to avoid any errors, but for sure there are still some typos out there, if you spot one, do not hesitate to contact us.

Some additional information may be found under URL:

http://biocomputing.mi.fu-berlin.de/Lehre/MarkovKetten_WS04/.

Wilhelm Huisinga & Eike Meerbach

2 Setting the scene

2.1 Introductory example

We will start with two examples that illustrate some features of Markov chains.

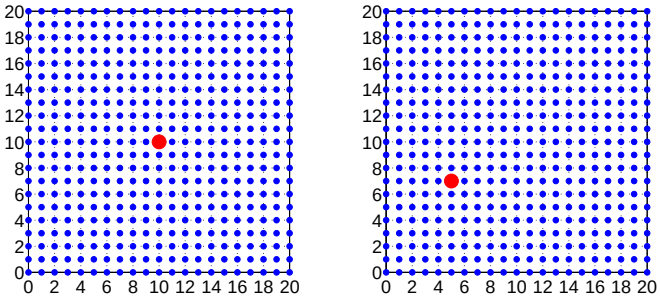


Figure 1: A particle (red) on a two-dimensional finite grid with $N = 20$. The smaller (blue) dots indicate all possible positions (the state space).

Some particle moving in a box. Imagine some particle in a test tube filled with gas or liquid. The particle will move "randomly" in the test tube driven by forces due to collisions with much smaller particles, i.e., the particles of the gas or the liquid. [You might know that this sort of movement is in general modelled by a stochastic process called Brownian motion, and in fact Brownian motion is an example of a Markov process in continuous time.] To keep things simple, we assume that our particle is moving on a two-dimensional finite grid such that each possible position can be labelled by coordinates $X = (a, b)$ with $a, b \in \{0, 1, 2, \ldots, N\}$. The set of possible positions $\mathbf{S} = \{(a, b) | a, b \in \{0, 1, 2, \ldots, N\}\}$ is called **state space**, an element of $\mathbf{S}$ is called a **state** (Fig. 1). The movement of a single particle at the discrete instances of time $t = 0, 1, 2, \ldots$ is then described by a sequence of states $X_0, X_1, X_2, \ldots$.

If we want to analyze or simulate the "randomly" movement of the particle, we have to specify the term "randomly", and our first attempt is based on the following

> Moving rule: Assume that the particle is at position $X_k = (a_k, b_k)$ at time k. Then, at time $k+1$ it moves to any other position $X_{k+1} = (a_{k+1}, b_{k+1})$ with equal probability $p = 1/(N+1)^2$.

At the beginning, we just place the particle randomly at one of the states with equal probability. Then, it moves around according to the specified

moving rule.

A possible observation/realization of the particle moving around could be:

$$X_0 = (11, 11),\ X_1 = (11, 20),\ X_3 = (1, 2),\ X_4 = (15, 7),\ X_5 = (3, 6) \dots$$

The dynamic of the particle we just specified is already a Markov chain on the state space $\mathbf{S}$. In fact, it is a very special Markov chain, since the position X_{k+1} of the particle at time $k+1$ does not depend on the previous position at time k; it is in fact an independent chain. As a model of some particle moving around, it is very "unnatural": It is rather unlikely that the particle moves from every position to every other position with equal probability. Rather positions closer to the current position of the particle should have a higher probability to be visited by the particle than states at more distance. We can incorporate this in our moving rule, by introducing some dependence of the future position on the current one.

This time, the particle moves around according to a different

> Moving rule: Assume that the particle is at position $X_k = (a_k, b_k)$ at time k. Then, the particle moves with equal probability to one of its neighboring positions, where two positions $(a, b) \in \mathbf{S}$ and $(c, d) \in \mathbf{S}$ are called neighboring, if $|a - c| \leq 1$ and $|b - d| \leq 1$.

This time, the particle is moving more "naturally", since the future position of the particle depends on the current one. This "memory effect", sometimes stated as "The future depends on the past only through the present", is known as the Markov property. Similar dependence on the history might be used to model the evolution of stock prices, the behavior of telephone customers, molecular networks etc.

2.2 Markov property, stochastic matrix, realization, density propagation

When dealing with randomness, some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is usually involved; Ω is called the **sample space**, $\mathcal{A}$ the set of all possible events (the σ–algebra) and $\mathbb{P}$ is some **probability measure** on Ω. Usually, not much is known about the probability space, rather the concept of random variables is used to deal with randomness. A function $X_0 : \Omega \to \mathbf{S}$ is called a (discrete) **random variable**, if for every $y \in \mathbf{S}$:

$$\{X_0 = y\} \quad := \quad \{\omega \in \Omega : X_0(\omega) = y\} \ \in \ \mathcal{A}.$$

In the above definition, the set $\mathbf{S}$ is called the **state space**, the set of all possible "outcomes" or "observations" of the random phenomena. Throughout this manuscript, the state space is assumed to be countable; hence it is

either finite, e.g., $\mathbf{S} = \{0, \ldots, N\}$ for some $N \in \mathbb{N}$ or countable infinite, e.g., $\mathbf{S} = \mathbb{N}$. Elements of the state space are denoted by $x, y, z, \ldots$ The definition of a random variable is motivated by the fact, that it is well-defined to assign a probability to the outcome or observation $X_0 = y$:

$$\mathbb{P}[X_0 = y] \;=\; \mathbb{P}[\{X_0 = y\}] \;=\; \mathbb{P}[\{\omega \in \Omega : X_0(\omega) = y\}].$$

The function $\mu_0 : \mathbf{S} \to \mathbb{R}$ with $\mu_0(y) = \mathbb{P}[X_0 = y]$ is called the **distribution** or law of the random variable X_0. Most of the time, a random variable is characterized by its distribution rather than as a function on the sample space Ω.

A sequence $X = \{X_k\}_{k\in\mathbb{N}}$ of random variables $X_k : \Omega \to \mathbf{S}$ is called a **discrete-time stochastic process** on the state space $\mathbf{S}$. The index k admits the convenient interpretation as time: if $X_k = y$, the process is said to be in state y at time k. For some given $\omega \in \Omega$, the $\mathbf{S}$–valued sequence

$$X(\omega) = \{X_0(\omega), X_1(\omega), X_2(\omega), \ldots\}$$

is called a **realization** (trajectory, sample path) of the stochastic process X associated with ω. In order to define the stochastic process properly, it is necessary to specify all distributions of the form

$$\mathbb{P}[X_m = x_m, X_{m-1=}x_{m-1}, \ldots, X_0 = x_0]$$

for $m \in \mathbb{N}$ and $x_0, \ldots, x_m \in \mathbf{S}$. This, of course, in general is a hard task. As we will see below, for Markov chains it can be done quite easily.

Definition 2.1 (Homogeneous Markov chain) *A discrete-time stochastic process $\{X_k\}_{k\in\mathbb{N}}$ on a countable state space $\mathbf{S}$ is called a **homogeneous Markov chain**, if the so–called **Markov property***

$$\mathbb{P}[X_{k+1} = z | X_k = y, X_{k-1} = x_{k-1}, \ldots, X_0 = x_0] = \mathbb{P}[X_{k+1} = z | X_k = y] \quad (1)$$

holds for every $k \in \mathbb{N}$, $x_0, \ldots, x_{k-1}, y, z \in \mathbf{S}$, implicitly assuming that both sides of equation (1) are defined[1] and, moreover, the right hand side of (1) does not depend on k, hence

$$\mathbb{P}[X_{k+1} = z | X_k = y] \;=\; \ldots \;=\; \mathbb{P}[X_1 = z | X_0 = y]. \quad (2)$$

For a given homogeneous Markov chain, the function $P : \mathbf{S} \times \mathbf{S} \to \mathbb{R}$ with

$$P(y, z) \;=\; \mathbb{P}[X_{k+1} = z | X_k = y]$$

[1]The conditional probability $\mathbb{P}[A|B]$ is only defined if $\mathbb{P}[B] \neq 0$. We will assume this throughout the manuscript whenever dealing with conditional probabilities.

is called the ***transition function***[2]*; its values* $P(y,z)$ *are called the (conditional)* ***transition probabilities*** *from* y *to* z*. The probability distribution* μ_0 *satisfying*

$$\mu_0(x) \quad = \quad \mathbb{P}[X_0 = x]$$

is called the ***initial distribution****. If there is a single* $x \in \mathbf{S}$ *such that* $\mu_0(x) = 1$*, then* x *is called the* ***initial state****.*

Often, one writes $\mathbb{P}_{\mu_0}$ or $\mathbb{P}_x$ to indicate that the initial distribution or the initial state is given by μ_0 or x, respectively. We also define the conditional transition probability

$$P(y, C) = \sum_{z \in C} P(y, z).$$

from some state $y \in \mathbf{S}$ to some subset $C \subset \mathbf{S}$.

There is a close relation between Markov chains, transition functions and stochastic matrices that we want to outline next. This will allow us to easily state a variety of examples of Markov chains. To do so, we need the following

Definition 2.2 *A matrix* $P = (p_{xy})_{x,y \in \mathbf{S}}$ *is called* ***stochastic****, if*

$$p_{xy} \geq 0, \ \textit{and} \ \sum_{y \in \mathbf{S}} p_{xy} = 1 \tag{3}$$

for all $x, y \in \mathbf{S}$*. Hence, all entries are non–negative and the row-sums are normalized to one.*

By Def. 2.1, every Markov chain defines via its transition function a stochastic matrix. The next theorem states that a stochastic matrix also allows to define a Markov chain, if additionally the initial distribution is specified. This can already be seen from the following short calculation: A stochastic process is defined in terms of the distributions

$$\mathbb{P}_\mu[X_m = x_m, X_{m-1=} x_{m-1}, \ldots, X_0 = x_0]$$

for every $m \in \mathbb{N}$ and $x_0, \ldots, x_m \in \mathbf{S}$. Exploiting the Markov property, we deduce

$$\begin{aligned}
&\mathbb{P}_{\mu_0}[X_m = x_m, X_{m-1=} x_{m-1}, \ldots, X_0 = x_0] \\
&\quad = \quad \mathbb{P}[X_m = x_m | X_{m-1} = x_{m-1}, \ldots, X_0 = x_0] \cdot \ldots \\
&\qquad \mathbb{P}[X_2 = x_2 | X_1 = x_1, X_0 = x_0] \cdot \mathbb{P}[X_1 = x_1 | X_0 = x_0] \cdot \mathbb{P}_{\mu_0}[X_0 = x_0] \\
&\quad = \quad \mathbb{P}[X_m = x_m | X_{m-1} = x_{m-1}] \cdot \ldots \cdot \mathbb{P}[X_2 = x_2 | X_1 = x_1] \\
&\qquad \mathbb{P}[X_1 = x_1 | X_0 = x_0] \cdot \mathbb{P}_{\mu_0}[X_0 = x_0] \\
&\quad = \quad P(x_{m-1}, x_m) \cdots P(x_1, x_2) \cdot P(x_0, x_1) \cdot \mu(x_0).
\end{aligned}$$

[2] Alternative notations are stochastic transition function, transition kernel, Markov kernel.

Hence, to calculate the probability of a specific sample path, we start with the initial probability of the first state and successively multiply by the conditional transition probabilities along the sample path. Theorem 2.3 [7, Thm. 3.2.1] will now make this more precise.

Remark. Above, we have exploited Bayes's rules. There are three of them [2]:
Bayes's rule of retrodiction. With $\mathbb{P}[A] > 0$, we have

$$\mathbb{P}[B|A] \quad = \quad \frac{\mathbb{P}[A|B] \cdot \mathbb{P}[B]}{\mathbb{P}[A]}.$$

Bayes's rule of exclusive and exhaustive causes. For a partition of the state space

$$\mathbf{S} \quad = \quad B_1 \cup B_2 \cup \ldots$$

and for every A we have

$$\mathbb{P}[A] \quad = \quad \sum_k \mathbb{P}[A|B_k] \cdot \mathbb{P}[B_k].$$

Bayes's sequential formula. For any sequence of events $A_1, \ldots, A_n$,

$$\mathbb{P}[A_1, \ldots, A_n] \quad = \quad \mathbb{P}[A_1] \cdot \mathbb{P}[A_2|A_1] \cdot \mathbb{P}[A_3|A_2, A_1] \cdot \ldots \cdot \mathbb{P}[A_n|A_{n-1}, \ldots, A_1].$$

Theorem 2.3 *For some given stochastic matrix $P = (p_{xy})_{x,y \in \mathbf{S}}$ and some initial distribution μ_0 on a countable state space $\mathbf{S}$, there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P}_{\mu_0})$ and a Markov chain $X = \{X_k\}_{k \in \mathbb{N}}$ satisfying*

$$\mathbb{P}_{\mu_0}[X_{k+1} = y | X_k = x, X_{k-1} = x_{k-1} \ldots, X_0 = x_0] = p_{xy}.$$

for all $x_0, \ldots, x_{k-1}, x, y \in \mathbf{S}$.

Often it is convenient to specify only the transition function of a Markov chain via some stochastic matrix, without further specifying its initial distribution. This would actually correspond to specifying a family of Markov chains, having the same transition function but possibly different initial distributions. For convenience, we will not distinguish between the Markov chain (with initial distribution) and the family of Markov chain (without specified initial distribution) in the sequel. No confusion should result from

this usage.

Exploiting Theorem 2.3, we now give some examples of Markov chains by specifying their transition function in terms of some stochastic matrix.

Example 2.4 *1.* ***Two state Markov chain.*** *Consider the state space* $\mathbf{S} = \{0, 1\}$. *For any given parameters* $p_0, p_1 \in [0, 1]$ *we define the transition function as*

$$P = \begin{pmatrix} 1 - p_0 & p_0 \\ p_1 & 1 - p_1 \end{pmatrix}.$$

Obviously, P *is a stochastic matrix—see cond. (3). The transition matrix is sometimes represented by its* ***transition graph*** $\mathcal{G}$*, whose vertices (nodes) are identified with the states of* $\mathbf{S}$*. The graph has an oriented edge from node* x *to node* y *with weight* p*, if the transition probability from* x *to* y *equals* p*, i.e.,* $P(x, y) = p$*. For the two state Markov chain, the transition graph is shown in Fig. 2.*

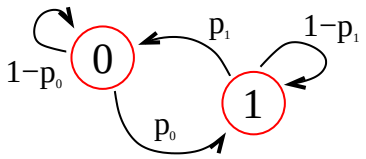


Figure 2: Transition graph of the two state Markov chain

2. ***Random walk on*** $\mathbb{N}$***.*** *Consider the state space* $\mathbf{S} = \{0, 1, 2, \ldots\}$ *and parameters* $p_k \in (0, 1)$ *for* $k \in \mathbb{N}$*. We define the transition function as*

$$P = \begin{pmatrix} 1 - p_0 & p_0 & & & \\ 1 - p_1 & 0 & p_1 & & \\ 0 & 1 - p_2 & 0 & p_2 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

Again, P *is a stochastic matrix. The transition graph corresponding to the random walk on* $\mathbb{N}$ *is shown in Fig. 3.*

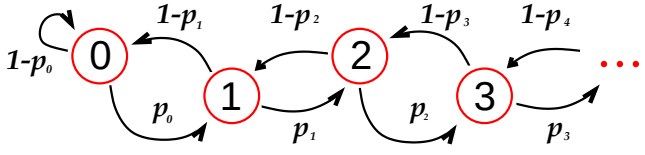


Figure 3: Transition graph of the random walk on $\mathbb{N}$.

3. ***A nine state Markov chain.*** *Consider the state space* $\mathbf{S} = \{1, \ldots, 9\}$ *and a transition graph specified in Fig. 4. If, as usually done, non–zero transition probabilities between states are indicated by an edge, while omitted edges are assumed to have zero weight, then the corresponding transition function has the form*

$$P = \begin{pmatrix} & p_{12} & & & & & & & \\ & & p_{23} & & & & & & \\ p_{31} & & & & p_{35} & & & & \\ & & p_{43} & p_{44} & & & & & \\ & p_{52} & & p_{54} & p_{55} & p_{56} & & & \\ & & & & p_{65} & p_{66} & & p_{68} & \\ & & & p_{74} & & p_{76} & p_{77} & & \\ & & & & & & p_{87} & p_{88} & p_{89} \\ & & & & & & & p_{98} & p_{99} \end{pmatrix}$$

Assume that the parameters p_{xy} *are such that* P *satisfies the two conditions (3). Then,* P *defines a Markov chain on the state space* $\mathbf{S}$.

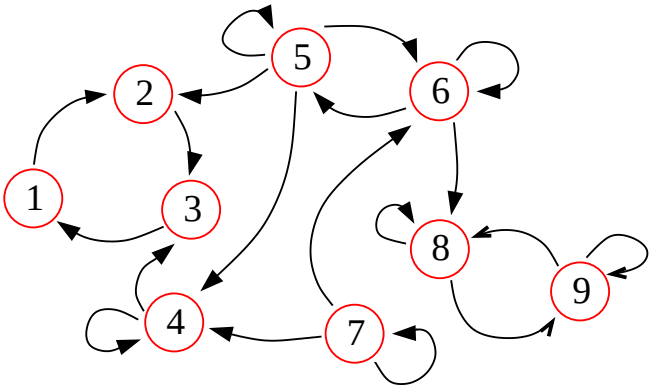


Figure 4: Transition graph of a nine state Markov chain.

2.3 Realization of a Markov chain

We now address the question of how to simulate a given Markov chain $X = \{X_k\}_{k\in\mathbb{N}}$,i.e., how to compute a realization $X_0(\omega), X_1(\omega), \ldots$ for some $\omega \in \Omega$. With this respect, the following theorem will be of great use.

Theorem 2.5 (Canonical representation) *[2, Sec. 2.1.1] Let* $\{\xi_k\}_{k\in\mathbb{N}}$ *denote some independent and identically distributed (i.i.d.) sequence of random variables with values in some space* $\mathbf{Y}$*, and denote by* X_0 *some random variable with values in* $\mathbf{S}$ *and independent of* $\{\xi_k\}_{k\in\mathbb{N}}$*. Consider some function* $f : \mathbf{S} \times \mathbf{Y} \to \mathbf{S}$*. Then the* ***stochastic dynamical system*** *defined by*

the recurrence equation

$$X_{k+1} \quad = \quad f(X_k, \xi_k) \tag{4}$$

defines a homogeneous Markov chain $X = \{X_k\}_{k\in\mathbb{N}}$ *on the state space* $\mathbf{S}$.

As a simple illustration of the canonical representation, let $(\xi_k)_{k\in\mathbb{N}}$ denote a sequence of i.i.d. random variables, independent of X_0, taking values in $\mathbf{Y} = \{-1, +1\}$ with probability

$$\mathbb{P}[\xi_k = 1] = q \quad \text{and} \quad \mathbb{P}[\xi_k = -1] = 1 - q$$

for some $q \in (0, 1)$. Then, the Markov chain $\{X_k\}_{k\in\mathbb{N}}$ on $\mathbf{S} = \mathbb{Z}$ defined by

$$X_{k+1} \quad = \quad X_k + \xi_k$$

corresponding to $f : \mathbb{Z} \times \mathbf{Y} \to \mathbb{Z}$ with $f(x, y) = x + y$ is a homogeneous Markov chain, called the random walk on $\mathbb{Z}$ (with parameter q).

Given the canonical representation, the transition function P of the Markov chain is defined by

$$P(x, y) \quad = \quad \mathbb{P}[f(x, \xi_0) = y].$$

The proof is left as an exercise. On the other hand, if some Markov chain $\{X_k\}_{k\in\mathbb{N}}$ is given in terms of its stochastic transition matrix P, we can define the canonical representation (4) for $\{X_k\}_{k\in\mathbb{N}}$ as follows: Let $\{\xi_k\}_{k\in\mathbb{N}}$ denote an i.i.d. sequence of random variables uniformly distributed on $[0, 1]$. Then, the recurrence relation $X_{k+1} = f(X_k, \xi_k)$ holds for $f : \mathbf{S} \times [0, 1] \to \mathbf{S}$ with

$$f(x, u) = z \quad \text{for} \quad \sum_{y=1}^{z-1} P(x, y) \le u < \sum_{y=1}^{z} P(x, y). \tag{5}$$

Note that *every* homogeneous Markov chain has a representation (4) with the function f defined in (5).

Two particular classes of functions f are of further interest: If f is a function of x alone and does not depend on u, then the thereby defined Markov chain is in fact deterministic and the recurrence equation is called a **deterministic dynamical system** with possibly random initial data. If, however, f is a function of u alone and does not depend on x, then the recurrence relation defines a sequence of independent random variables. This way, Markov chains are a mixture of deterministic dynamical systems and independent random variables.

Now, we come back to the task of computing a realization of a Markov chain. Here, the canonical representation proves extreme useful, since it directly implies an algorithmic realization: In order to simulate a Markov chain $\{X_k\}_{k\in\mathbb{N}}$, choose a random number x_0 according to the law of X_0 and choose a sequence of random numbers $w_0, w_1, \ldots$ according to the law of ξ_0 (recall that the ξ_k are i.i.d.). Then, the realization $x_0, x_1, \ldots$ of $\{X_k\}_{k\in\mathbb{N}}$ is recursively defined by $x_{k+1} = f(x_k, w_k)$. If the Markov chain is specified in terms of some transition function P and some initial distribution X_0, then the same holds with the sequence of ξ_k being i.i.d. uniform in $[0, 1)$ distributed random variables and f is defined in terms of P via relation (5).

2.4 The evolution of distributions under the Markov chain

One important task in the theory of Markov chains is to determine the distribution of the Markov chain while it evolves in time. Given some initial distribution μ_0, the distribution μ_k of the Markov chain at time k is given by

$$\mu_k(z) = \mathbb{P}_{\mu_0}[X_k = z]$$

for every $z \in \mathbf{S}$. A short calculation reveals

$$\begin{aligned} \mu_k(z) &= \mathbb{P}_{\mu_0}[X_k = z] \\ &= \sum_{y\in\mathbf{S}} \mathbb{P}_{\mu_0}[X_{k-1} = y]\ \mathbb{P}[X_k = z | X_{k-1} = y] \\ &= \sum_{y\in\mathbf{S}} \mu_{k-1}(y)\, P(y, z) \end{aligned}$$

To proceed we introduce the notion of transfer operators, which is closely related to transition functions and Markov chains.

Given some distribution $\mu : \mathbf{S} \to \mathbb{C}$, we define the **total variation norm** $||\cdot||_{TV}$ by

$$||\mu||_{TV} = \sum_{x\in\mathbf{S}} |\mu(x)|.$$

Based on the total variation norm, we define the function space

$$\mathcal{M} = \{\mu : \mathbf{S} \to \mathbb{C} : ||\mu||_{TV} < \infty\}.$$

Note that $\mathcal{M}$ equipped with the total variation norm is a Banach space. Given some Markov chain in terms of its transition function P, we define the **transfer operator** $P : \mathcal{M} \to \mathcal{M}$ acting on distributions by $\mu \mapsto \mu P$ with

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

We are aware of the fact that the term P has multiple meanings. It serves to denote (i) some transition function corresponding to a Markov chain, (ii) some stochastic matrix, and (iii) some transfer operator. However, no confusion should result form the multiple usage, since it should be clear from the context what meaning we are referring to. Moreover, actually, the three meanings are equivalent expressions of the same fact.

Given some transfer operator P, we define the kth power P^k of P recursively by $\mu P^k = (\mu P)P^{k-1}$ for $k > 0$ and $P^0 = \mathrm{Id}$, the identity operator. As can be shown, P^k is again a transfer operator associated with the k-step Markov chain $Y = (Y_n)_{n\in\mathbb{N}}$ with $Y_n = X_{kn}$. The corresponding transition function Q is identical to the so-called k–**step transition probability**

$$P^k(x, y) \;=\; \mathbb{P}[X_k = y | X_0 = x], \tag{6}$$

denoting the (conditional) transition probability from x to y in k steps of the Markov chain X. Thus, we have

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

In the notion of stochastic matrices, P^k is simply the kth power of the stochastic matrix P.

Exploiting the notion of transfer operators acting on distributions, the evolution of distributions under the Markov chain can be formulated quite easily. In terms of powers of P, we can rewrite μ_k as follows

$$\mu_k \;=\; \mu_{k-1}\, P^1 \;=\; \mu_{k-2}\, P^2 \;=\; \ldots \;=\; \mu_0\, P^k. \tag{7}$$

There is an important relation involving k–step transition probability, namely the **Chapman-Kolmogorov equation** stating that

$$P^{m+k}(x, z) \;=\; \sum_{y\in\mathbf{S}} P^m(x, y)P^k(y, z) \tag{8}$$

holds for every $m, k \in \mathbb{N}$ and $x, y, z \in \mathbf{S}$. In terms of transfer operators, the Chapman-Kolmogorov equation reads $P^{m+k} = P^m P^k$, which is somehow an obvious statement.

To illustrate the evolution of densities, consider our nine state Markov chain with suitable chosen parameters for the transition matrix. The initial distribution μ_0 and some iterates, namely, μ_1, μ_3, μ_{15}, μ_{50} are shown in Figure 5. We observe that μ_k changes while evolving in time. However, there also exist distributions that do not change in time; as we will see in the course of this manuscript, these are of special interest.

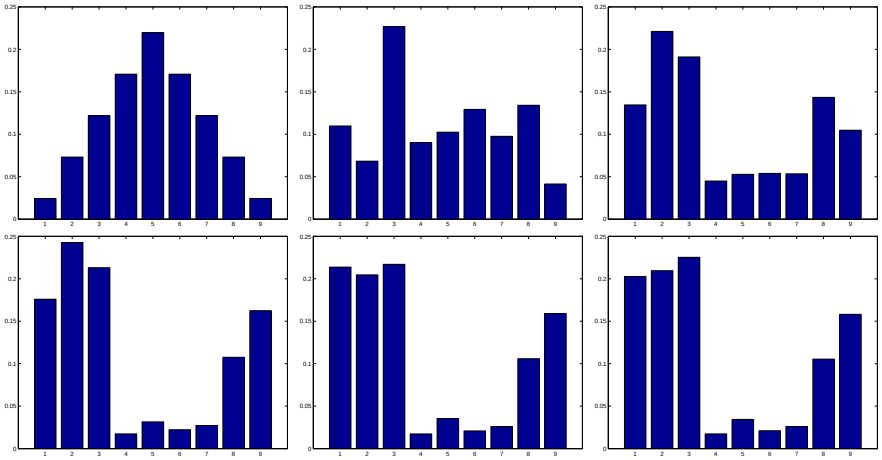


Figure 5: Evolution of some density μ_k in time. Top: at time $k = 0, 1, 3$ (left to right). Bottom: at time $k = 15, 50$. The stationary density is shown at the bottom, right.

Definition 2.6 *A probability distribution π satisfying*

$$\mathbb{P}_\pi[X_1 = y] \quad = \quad \pi(y) \tag{9}$$

is called a ***stationary distribution*** *or* ***invariant probability measure*** *of the Markov chain $\{X_k\}_{k\in\mathbb{N}}$. Equivalently, it is*

$$\pi = \pi P \tag{10}$$

in terms of its transfer operator P.

Note that $\pi = \pi P$ implies $\pi = \pi P^k$ to hold for every $k \in \mathbb{N}$. To illustrate the above definition, we have computed the stationary density for the nine state Markov chain (see Figure 5). Moreover, we analytically compute the stationary distribution of the two state Markov chain. Here, $\pi = (\pi(1), \pi(2))$ has to satisfy

$$\pi \quad = \quad \pi \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

resulting in the two equations

$$\begin{aligned} \pi(1) &= \pi(1)(1-a) + \pi(2)b \\ \pi(2) &= \pi(1)a + \pi(2)(1-b). \end{aligned}$$

This is a dependent system reducing to the single equation $\pi(1)a = \pi(2)b$, to which the additional constraint $\pi(1) + \pi(2) = 1$ must be added (why?). We obtain

$$\pi \;=\; \left(\frac{b}{a+b}, \frac{a}{a+b}\right).$$

Stationary distributions need neither to exist (as we will see below) nor they need to be unique! As an example for the latter statement, consider a Markov chain with the identity matrix as transition function. Then, every probability distribution is stationary.

When calculating stationary distributions, two strategies can be quite useful. The first one is based on the interpretation of eq. (9) as an eigenvalue problem, the second one is based on the notion of the probability flux. While we postpone the eigenvalue interpretation, we will now exploit the probability flux idea in order to calculate the stationary distribution of the random walk on $\mathbb{N}$.

Assume that the Markov chain exhibits a stationary distribution π and let $A, B \subset \mathbf{S}$ denote two subsets of the state space. Then, the **probability flux** from A to B is defined by

$$\begin{aligned} \mathrm{flux}_\pi(A,B) &= \mathbb{P}_\pi[X_1 \in B, X_0 \in A] \\ &= \sum_{x\in A} \pi(x) P(x,B) \;=\; \sum_{x\in A}\sum_{y\in B} \pi(x)P(x,y). \end{aligned} \tag{11}$$

For a Markov chain possessing a stationary distribution, the flux from some subset A to its complement A^c is somehow balanced:

Theorem 2.7 ([3]) *Let $\{X_k\}_{k\in\mathbb{N}}$ denote a Markov chain with stationary distribution π and $A \subset \mathbf{S}$ an arbitrary subset of the state space. Then*

$$\mathrm{flux}_\pi(A, A^c) \;=\; \mathrm{flux}_\pi(A^c, A),$$

hence the probability flux from A to its complement A^c is equal to the reverse flux from A^c to A.

Proof: The proof is left as an exercise. □

Now, we want to exploit the above theorem to calculate the stationary distribution of the random walk on $\mathbb{N}$. For sake of illustration, we take $a_k = p \in (0,1)$ for $k \in \mathbb{N}$. Hence, with probability p the Markov chain moves to the right, while with probability $1-p$ it moves to the left (with exception of the origin). Then, the equation of stationarity (9) reads

$$\begin{aligned} \pi(0) &= \pi(0)(1-p) + \pi(1)(1-p) \text{ and} \\ \pi(k) &= \pi(k-1)p + \pi(k+1)(1-p) \end{aligned}$$

for $k > 0$. The first equation can be rewritten as $\pi(1) = \pi(0)p/(1-p)$. Instead of exploiting the second equation (try it), we use Theorem 2.7 to proceed. For some $k \in \mathbb{N}$ consider $A = \{0, \ldots, k\}$ implying $A^c = \{k+1, k+2, \ldots\}$; then

$$\begin{aligned} \text{flux}_\pi(A, A^c) &= \sum_{x \in A} \pi(x) P(x, A^c) = \pi(k)p \\ \text{flux}_\pi(A^c, A) &= \sum_{x \in A^c} \pi(x) P(x, A) = \pi(k+1)(1-p) \end{aligned}$$

It follows from Theorem 2.7, that

$$\pi(k)p = \pi(k+1)(1-p)$$

and therefore

$$\pi(k+1) = \pi(k)\left(\frac{p}{1-p}\right) = \ldots = \pi(0)\left(\frac{p}{1-p}\right)^{k+1}.$$

The value of $\pi(0)$ is determined by demanding that π is a probability distribution:

$$1 = \sum_{k=0}^{\infty} \pi(k) = \pi(0) \sum_{k=0}^{\infty} \left(\frac{p}{1-p}\right)^k.$$

Depending on the parameter p, we have

$$\sum_{k=0}^{\infty} \left(\frac{p}{1-p}\right)^k = \begin{cases} \infty; \text{ if } p \geq 1/2 \\ (1-p)/(1-2p); \text{ if } p < 1/2. \end{cases} \tag{12}$$

Thus, we obtain for the random walk on $\mathbb{N}$ the following dependence on the parameter p:

- for $p < 1/2$, the stationary distribution is given by

$$\pi(0) = \frac{1-2p}{1-p} \quad \text{and} \quad \pi(k) = \pi(0)\left(\frac{p}{1-p}\right)^k$$

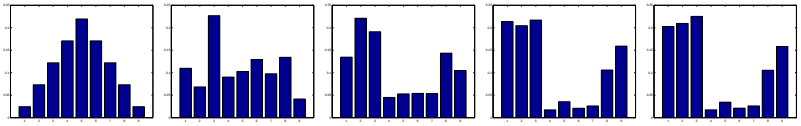
- for $p \geq 1/2$ there does not exist a stationary distribution π, since the normalisation in eq. (12) fails.

A density or measure π satisfying $\pi = \pi P$, without the requirement $\sum \pi(x) = 1$, is called **invariant**. Trivially, every stationary distribution is invariant, but the reverse statement is not true. Hence, for $p \geq 1/2$, the family of mesures π with $\pi(0) \in \mathbb{R}^+$ and $\pi(k) = \pi(0)p/(1-p)$ are invariant measures of the random walk on $\mathbb{N}$ (with parameter p) .

2.5 Some key questions concerning Markov chains

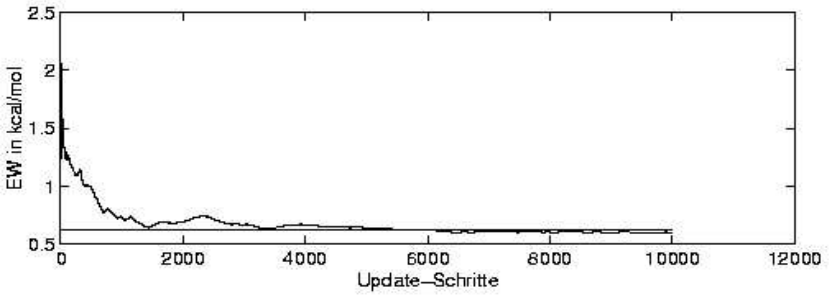
1. Existence of unique invariant measure and corresponding convergence rates

$$\mu_n \longrightarrow \pi \quad \text{or} \quad P^n = 1\pi^t + \mathcal{O}(n^{m_2}|\lambda_2|^n).$$

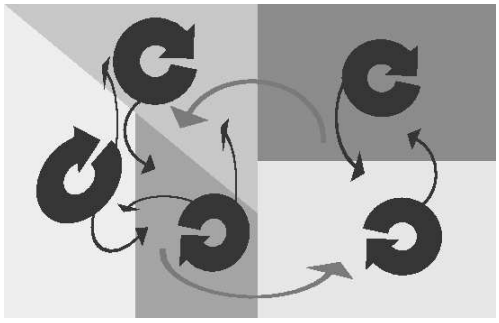


2. Evaluation of expectation values and corresponding convergence rates, including sampling of the stationary distribution

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \longrightarrow \sum_{x\in\mathbf{S}} f(x)\pi(x)$$

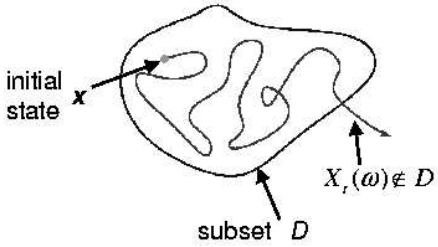


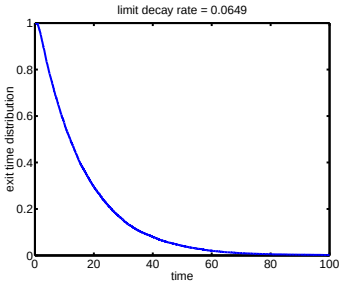
3. Identification of macroscopic properties like, e.g., cyclic or metastable behaviour, coarse graining of the state space.



4. Calculation of return and stopping times, exit probabilities and probabilities of absorption.

$$\sigma_D(x) \quad = \quad \inf\{t > 0 : X_t \notin D, X_0 = x\}$$


initial
state
x
$X_t(\omega) \notin D$
subset D


limit decay rate = 0.0649
exit time distribution
time
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100

3 Communication and recurrence

3.1 Irreducibility and (A)periodicity

This section is about the topology of Markov chains. We start with some

Definition 3.1 *Let $\{X_k\}_{k\in\mathbb{N}}$ denote a Markov chain with transition function P, and let $x, y \in \mathbf{S}$ denote some arbitrary pair of states.*

1. *The state x **has access to** the state y, written $x \to y$, if*

$$\mathbb{P}[X_m = y | X_0 = x] \quad > \quad 0$$

   for some $m \in \mathbb{N}$ that possibly depends on x and y. In other words, it is possible to move (in m steps) from x to y with positive probability.

2. *The states x and y **communicate**, if x has access to y and y access to x, denoted by $x \leftrightarrow y$.*

3. *The Markov chain (equivalently its transition function) is said to be **irreducible**, if all pairs of states communicate.*

The communication relation $\leftrightarrow$ can be exploited to analyze the Markov chain in more detail. It is easy to prove that communication relation is a so–called equivalence relation, hence it is

1. reflexive: $x \leftrightarrow x$
2. symmetric: $x \leftrightarrow y$ implies $y \leftrightarrow x$,
3. transitive: $x \leftrightarrow y$ and $y \leftrightarrow z$ imply $x \leftrightarrow z$.

Recall that every equivalence relation induces a partition $\mathbf{S} = C_0 \cup \cdots \cup C_{r-1}$ of the state space $\mathbf{S}$ into so–called equivalence classes defined as

$$C_k \quad = \quad [x_k] \; := \; \{y \in \mathbf{S} : y \leftrightarrow x_k\}$$

for $k = 0, \ldots, r-1$ and suitable states $x_0, \ldots, x_{r-1} \in \mathbf{S}$. In the theory of Markov chains, the elements $C_0, \ldots, C_{r-1}$ of the induced partition are called **communication classes**.

Why are we interested in communication classes? The partition into communication classes allows to break down the Markov chain into easier to handle and separately analyzable subunits. This might be interpreted as finding some normal form for the Markov chain. If there is only one communication class, hence all states communicate, then nothing can be further partitioned, and the Markov chain is already in its normal form. There are some additional properties of communication classes:

Definition 3.2 *A communication class C is called **closed** (invariant or absorbing) if none of the states in C has access to the complement $C^c = \mathbf{S}\backslash C$ of C, i.e, for every $x \in C$ and every $y \in C^c$ we have $x \nrightarrow y$. In terms of transition probabilities, this is equivalent to*

$$\mathbb{P}[X_m \in C^c | X_0 = x] = 0$$

for every $x \in C$ and every $m \geq 0$.

Now assume that the Markov chain is not irreducible. Let $C_0, \ldots, C_{r-1}$ denote the closed communication classes and D the collection of all remaining communication classes. Then

$$\mathbf{S} \quad = \quad \Big(C_0 \cup \ldots \cup C_{r-1}\Big) \cup D. \tag{13}$$

The following proposition states that we may restrict the Markov chain to its closed communication classes that hence can be analyzed separately [7, Prop. 4.1.2].

Proposition 3.3 *Suppose that C is some closed communication class. Let P_C denote the transition function $P = (P(x,y))_{x,y\in\mathbf{S}}$ restricted to C, i.e.,*

$$P_C \quad = \quad \big(P(x,y)\big)_{x,y\in C}.$$

Then there exists an irreducible Markov chain $\{Y_n\}_{n\in\mathbb{N}}$ whose state space is C and whose transition function is given by P_C.

Proof: We only have to check that P_C is a stochastic matrix. Then the Proposition follows from Theorem 2.3. □

According to [7, p.84], for reducible Markov chains we can analyze at least the closed subsets in the decomposition (13) as separate chains. The power of this decomposition lies largely in the fact that any Markov chain on a countable state space can be studied assuming irreducibility. The irreducible parts can then be put together to deduce most of the properties of the original (possible reducible) Markov chain. Only the behavior of the remaining part D has to be studied separately, and in analyzing stability properties the part of the state space corresponding to D may often be ignored.

For the states $x \in D$ only two things can happen: either they reach one of the closed communication classes C_i, in which case they get absorbed, or the only other alternative, the Markov chain leaves every finite subset of D and "heads to infinity" [7, p.84].

Another important property is periodicity, somehow a leftover of the deterministic realm within the stochastic world of Markov chains. It is best illustrated by the following theorem, which we prove at the end of this section:

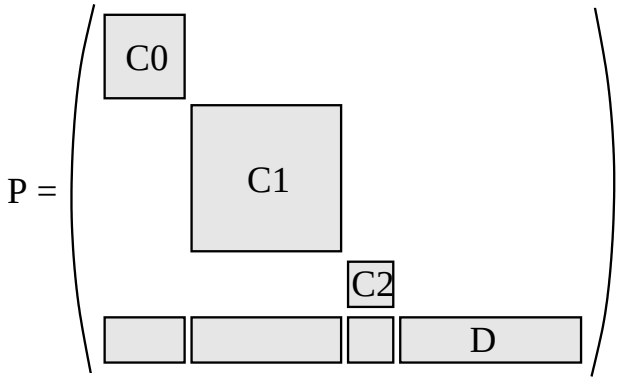


Figure 6: Normal form of the transition function for $r = 3$.

Theorem 3.4 (cyclic structure [2]) *For any* irreducible *Markov chain* $\{X_k\}_{k\in\mathbb{N}}$*, there exists a unique partition of the state space* $\mathbf{S}$ *into* d *so–called* ***cyclic classes*** $E_0, \ldots, E_{d-1}$ *such that*

$$\mathbb{P}[X_1 \in E_{k+1} | X_0 = x] = 1$$

for every $x \in E_k$ *and* $k = 0, \ldots, d-1$ *(by convention* $E_d = E_0$*). Moreover,* d *is maximal in the sense that there exists no partition into more than* d *classes with the same property.*

Hence the Markov chain moves cyclically in each time step from one class to the next. The number d in Theorem 3.4 is called the **period** of the Markov chain (respectively the transition function). If $d = 1$, then the Markov chain is called **aperiodic**. Later on, we will see, how to identify (a)periodic behavior and, for $d > 1$ the cyclic classes.

The transition matrix of a periodic irreducible Markov chain has a special structure. After renumbering of the states of $\mathbf{S}$ (if necessary), the transition function has a block structure as illustrated in Fig. 7. There is a more arithmetic but much less intuitive definition of the period that in addition does not rely on irreducibility of the Markov chain.

Definition 3.5 ([2]) *The period* $d(x)$ *of some state* $x \in \mathbf{S}$ *is defined as*

$$d(x) \quad = \quad \gcd\{k \geq 1 : \mathbb{P}[X_k = x | X_0 = x] > 0\},$$

with the convention $d(x) = \infty$*, if* $\mathbb{P}[X_k = x | X_0 = x] = 0$ *for all* $k \geq 1$*. If* $d(x) = 1$*, then the state* x *is called aperiodic.*

Hence, different states may have different periods. As the following theorem states, this is only possible for reducible Markov chains [2].

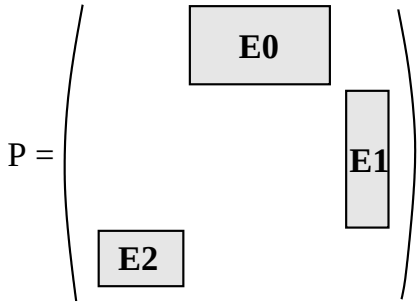


Figure 7: Block structure of a periodic, irreducible Markov chain with period $d = 3$.

Theorem 3.6 *The period is a class property, i.e., all states of a communication class have the same period.*

Proof: Let C be a communication class and $x, y \in C$. As a consequence, there exist $k, m \in \mathbb{N}$ with $P^k(x, y) > 0$ and $P^m(y, x) > 0$. Moreover, $d(x) < \infty$ and $d(y) < \infty$. From the Chapman-Kolmogorov Equation (8) we get

$$P^{k+j+m}(x, x) \geq P^k(x, y)P^j(y, y)P^m(y, x)$$

for all $j \in \mathbb{N}$. Now, for $j = 0$ we infer that $d(x)$ divides $k + m$, in short $d(x)|(k + m)$, since $P^k(x, y)P^m(y, x) > 0$. Whereas choosing j such that $P^j(y, y) > 0$ yields $d(x)|(k+j+m)$. Therefore we have $d(x)|j$, which means that $d(x)|d(y)$. By symmetry of the argument, we obtain $d(y)|d(x)$, which implies $d(x) = d(y)$. □

In particular, if the Markov chain is irreducible, all states have the same period d, and we may call d the period of the Markov chain (cf. Theorem 3.4). Combining Definition 3.5 with Theorem 3.6, we get the following useful criterion for aperiodicity:

Corollary 3.7 *An irreducible Markov chain $\{X_k\}_{k\in\mathbb{N}}$ is aperiodic, if there exists some state $x \in \mathbf{S}$ such that $\mathbb{P}[X_1 = x|X_0 = x] > 0$.*

We now start to prove Theorem 3.4. The proof will be a simple consequence of the following three propositions.

Proposition 3.8 *Let $\{X_k\}_{k\in\mathbb{N}}$ be an irreducible Markov chain with transition function P and period d. Then, for any states $x, y \in \mathbf{S}$, there is an $k_0 \in \mathbb{N}$ and $m \in \{0, \ldots, d-1\}$, possibly depending on x and y, such that*

$$P^{kd+m}(x, y) > 0$$

for every $k \geq k_0$.

Proof: For now asumme that $x = y$, then, by the Chapman-Kolmogorov equation (8), the set $G_x = \{k \in \mathbb{N} : P^k(x,x) > 0\}$ is closed under addition, since $k, k' \in G_x$ implies

$$P^{k+k'}(x,x) \geq P^k(x,x)P^{k'}(x,x) > 0,$$

and therefore $k + k'$ is an element of G_x. This enables us to use a number theoretic result [9, Appendix A]: *A subset of the natural numbers which is closed under addition, contains all, except a finite number, multiples of its greatest common divisor.* By definition, the gcd of G_x is the period d, so there is a $k_0 \in \mathbb{N}$ with $P^{kd}(x,x) > 0$ for $k \geq k_0$. Now, if $x \neq y$ then irreducibility of the Markov chain ensures that there is an $m \in \mathbb{N}$ with $P^m(x,y) > 0$ and therefore

$$P^{kd+m}(x,y) \geq P^{kd}(x,x)P^m(x,y) > 0$$

for $k \geq k_0$. Of course k_0 can be chosen in such a way that $m < d$. □

Proposition 3.8 can be used to define an equivalence relation on $\mathbf{S}$, which gives rise to the cyclic classes in Theorem 3.4: Fix an arbitrary state $z \in \mathbf{S}$ and define x and y to be equivalent, denoted by $x \leftrightarrow_z y$, if there is an $m \in \{0, \ldots, d-1\}$ and an $k_0 \in \mathbb{N}$ such that

$$P^{kd+m}(z,x) > 0 \text{ and } P^{kd+m}(z,y) > 0$$

for every $k \geq k_0$. The relation $x \leftrightarrow_z y$ is indeed an equivalent relation (the proof is left as an exercise) and therefore defines a disjoint partition of the state space $S = E_0 \cup E_1 \cup E_2 \cup \ldots E_{d-1}$ with

$$E_m = \{x \in \mathbf{S} : P^{kd+m}(z,x) > 0 \text{ for } k \geq k_0\}$$

for $m = 0, \ldots, d-1$. The next proposition confirms that these are the cyclic classes used in Theorem 3.4.

Proposition 3.9 *Let P denote the transition function of an irreducible Markov chain with period d and define $E_0, \ldots, E_{d-1}$ as above.*

If $P^r(x,y) > 0$ for some $r > 0$ and $x \in E_m$ then $y \in E_{m+r}$, where the indices are taken modulo d. In particular, if $P(x,y) > 0$ and $x \in E_m$ then $y \in E_{m+1}$ with the convention $E_d = E_0$.

Proof: Let $P^r(x,y) > 0$ and $x \in E_m$, then there is a k_0, such that $P^{kd+m}(z,x) > 0$ for all $k \geq k_0$, and hence

$$P^{kd+m+r}(z,y) \geq P^{kd+m}(z,x)P^r(x,y) > 0,$$

for every $k > k_0$, therefore $y \in E_{m+r}$. □

There is one thing left to do: We have to prove that the partition of $\mathbf{S}$ into cyclic classes is unique, i.e., it does not depend on the $z \in \mathbf{S}$ chosen to define $\leftrightarrow_z$.

Proposition 3.10 *For two given states $z, z' \in \mathbf{S}$, the partitions of the state space induced by $\leftrightarrow_z$ and $\leftrightarrow_{z'}$ are equal.*

Proof: Let E_m and $E'_{m'}$ denote two arbitrary subsets from the partitions induced by $\leftrightarrow_z$ and $\leftrightarrow_{z'}$, respectively. We prove that the two subsets are either equal or disjoint. Assume that E_m and $E'_{m'}$ are not disjoint and consider some $x \in E_m \cap E'_{m'}$. Consider some $y \in E_m$. Then, due to Props. 3.8 there exist $k_0 \in \mathbb{N}$ and $s < d$ such that $P^{kd+s}(x, y) > 0$ for $k \geq k_0$. Due to 3.9, we infer $y \in E_{(kd+s)+m}$, hence s is a multiple of d. Consequently, $P^{kd}(x, y) > 0$ for $k \geq k''_0$. By definition of $E'_{m'}$, there is an $k'_0 \in \mathbb{N}$, such that $P^{kd+m'}(z', x) > 0$ for $k \geq k'_0$, and therefore

$$P^{(k+k''_0)d+m'}(z', y) \geq P^{kd+m'}(z', x)P^{k''_0 d}(x, y) > 0$$

for $k \geq k'_0$. Equivalently, $P^{k'd+m'}(z', y) > 0$ for $k' \geq k'_0 + k''_0$, so that $y \in E'_{m'}$. □

3.2 Recurrence and the existence of stationary distributions

In Section 3.1 we have investigated the topology of a Markov chain. Recurrence and transience is somehow the next detailed level of investigation. It is in particular suitable to answer the question, whether a Markov chain admits a unique stationary distribution.

Consider an irreducible Markov chain on the state space $\mathbf{S} = \mathbb{N}$. By definition we know that each two states communicate. Hence, given $x, y \in \mathbf{S}$ there is always a positive probability to move from x to y and vice versa. Consequently, there is also a positive probability to start in x and return to x via visiting y. However, there might also exist the possibility that the Markov chain never returns to x within finite time. This is often an undesirable feature; in a sense the Markov chain is unstable.

A better notion of stability is that of recurrence, when the Markov chain returns to any state infinitely often. The strongest results are obtained, when in addition the average return time to any state is finite. We start by introducing the necessary notions.

Definition 3.11 *A random variable $T : \Omega \to \mathbb{N} \cup \{\infty\}$ is called a **stopping time** w.r.t. the Markov chain $\{X_k\}_{k\in\mathbb{N}}$, if for every integer $k \in \mathbb{N}$ the event $\{T = k\}$ can be expressed in terms of $X_0, X_1, \ldots, X_k$.*

We give two prominent examples.

Example 3.12 *For every $c \in \mathbb{N}$, the random variable $T = c$ is a stopping time.*

The so–called first return time plays a crucial role in the analysis of recurrence and transience.

Definition 3.13 *The stopping time* $T_x : \Omega \to \mathbb{N} \cup \{\infty\}$ *defined by*

$$T_x \quad = \quad \min\{k \geq 1 : X_k = x\},$$

with the convention $T_x = \infty$*, if* $X_k \neq x$ *for all* $k \geq 1$*, is called the* ***first return time*** *to state* x*.*

Note that T_x is a random variable. Hence, for a given realization ω with $X_0(\omega) = y$ for some initial state $y \in \mathbf{S}$, the term

$$T_x(\omega) = \min\{k \geq 1 : X_k(\omega) = x\}$$

is an integer, or infinite. Using the first return time, we can specify how often and how likely the Markov chain returns to some state $x \in \mathbf{S}$. The following considerations will be of use:

- The probability of starting initially in $x \in \mathbf{S}$ and returning to x in exactly n steps: $\mathbb{P}_x[T_x = n]$.
- The probability of starting initially in $x \in \mathbf{S}$ and returning to x in a finite number of steps: $\mathbb{P}_x[T_x < \infty]$.
- The probability of starting initially in $x \in \mathbf{S}$ and not returning to x in a finite number of steps: $\mathbb{P}_x[T_x = \infty]$.

Of course, the relation among the three above introduced probabilities is

$$\mathbb{P}_x[T_x < \infty] = \sum_{n=1}^{\infty} \mathbb{P}_x[T_x = n] \quad \text{and} \quad \mathbb{P}_x[T_x < \infty] + \mathbb{P}_x[T_x = \infty] = 1.$$

We now introduce the important concept of recurrence. We begin by defining a recurrent state, and then show that recurrence is actually a class property, i.e., the states of some communication class are either all recurrent or none of them is.

Definition 3.14 *Some state* $x \in \mathbf{S}$ *is called* ***recurrent*** *if*

$$\mathbb{P}_x[T_x < \infty] \quad = \quad 1,$$

and ***transient*** *otherwise.*

The properties of recurrence and transience are intimately related to the number of visits to a given state. To do so, we need a generalization of the Markov property, the so-called **strong Markov property**. It states that the Markov property, i.e. the independence of past and future given the present state, holds even if the present state is determined by a stopping time.

Theorem 3.15 (Strong Markov property) *Let $\{X_k\}_{k\in\mathbb{N}}$ be a homogenous Markov chain on a countable state space $\mathbf{S}$ with transition matrix P and initial distribution μ_0. Let T denote a stopping time w.r.t. the Markov chain. Then, conditional on $T<\infty$ and $X_T = z \in \mathbf{S}$, the sequence $(X_{T+n})_{n\in\mathbb{N}}$ is a Markov chain with transition matrix P and initial state z that is independent of $X_0,\dots,X_T$.*

Proof: Let $H \subset \Omega$ denote some event determined by $X_0,\dots,X_T$, e.g., $H = \{X_0 = y_0,\dots,X_T = y_T\}$ for $y_0,\dots,y_T \in \mathbf{S}$. Then, the event $H \cap \{T = m\}$ is determined by $X_0,\dots,X_m$. By the Markov property at time $t = m$ we get

$$\begin{aligned}
&\mathbb{P}_{\mu_0}[X_T = x_0,\dots,X_{T+n} = x_n, H, X_T = z, T = m] \\
&\quad = \mathbb{P}_{\mu_0}[X_T = x_0,\dots,X_{T+n} = x_n | H, X_m = z] \\
&\qquad \mathbb{P}_{\mu_0}[H, X_T = z, T = m] \\
&\quad = \mathbb{P}_z[X_0 = x_0,\dots,X_n = x_n]\,\mathbb{P}_{\mu_0}[H, X_T = z, T = m].
\end{aligned}$$

Hence, summation over $m = 0, 1, \dots$ yields

$$\begin{aligned}
&\mathbb{P}_{\mu_0}[X_T = x_0,\dots,X_{T+n} = x_n, H, X_T = z, T < \infty] \\
&\quad = \mathbb{P}_z[X_0 = x_0,\dots,X_n = x_n]\,\mathbb{P}_{\mu_0}[H, X_T = z, T < \infty],
\end{aligned}$$

and dividing by $\mathbb{P}_{\mu_0}[X_T = z, T < \infty]$, we finally obtain

$$\begin{aligned}
&\mathbb{P}_{\mu_0}[X_T = x_0,\dots,X_{T+n} = x_n, H | X_T = z, T < \infty] \\
&\quad = \mathbb{P}_z[X_0 = x_0,\dots,X_n = x_n]\,\mathbb{P}_{\mu_0}[H | X_T = z, T < \infty].
\end{aligned}$$

This is exactly the statement of the strong Markov property. □

Theorem 3.15 states that if a Markov chain is stopped by any "stopping time rule" at, say $X_T = x$, and the realization after T is observed, it can not be distinguished from the Markov chain started at x (with the same transition function, of course). Now, we are ready to state the relation between recurrence and the number of visits $N_y : \Omega \to \mathbb{N} \cup \{\infty\}$ to some state $y \in \mathbf{S}$ defined by

$$N_y = \sum_{k=1}^{\infty} \mathbf{1}_{\{X_k = y\}}.$$

Exploiting the strong Markov property and by induction [2, Thm. 7.2], it can be shown that

$$\mathbb{P}_x[N_y = m] = \mathbb{P}_x[T_y < \infty]\,\mathbb{P}_y[T_y < \infty]^{m-1}\,\mathbb{P}_y[T_y = \infty] \tag{14}$$

for $m > 0$, and $\mathbb{P}_x[N_y = 0] = \mathbb{P}_x[T_y = \infty]$.

Theorem 3.16 *Consider some state $x \in \mathbf{S}$. Then*

$$x \text{ is recurrent} \Leftrightarrow \mathbb{P}_x[N_x = \infty] = 1 \Leftrightarrow \mathbb{E}_x[N_x] = \infty,$$

and

$$x \text{ is transient} \Leftrightarrow \mathbb{P}_x[N_x = \infty] = 0 \Leftrightarrow \mathbb{E}_x[N_x] < \infty.$$

The above equivalence in general fails to hold for the denumerable, more general state space case—here, one has to introduce the notion of Harris recurrent [7, Chapt. 9].

Proof: Now, if x is recurrent then $\mathbb{P}_x[T_x < \infty] = 1$. Hence, due to eq. (14)

$$\mathbb{P}_x[N_x < \infty] = \sum_{m=0}^{\infty} \mathbb{P}_x[N_x = m] = \sum_{m=0}^{\infty} \mathbb{P}_x[T_x < \infty]^m \, \mathbb{P}_x[T_x = \infty],$$

vanishes, since every summand is zero. Consequently, $\mathbb{P}_x[N_x = \infty] = 1$. Now, if x is transient, then $\mathbb{P}_x[T_x < \infty] < 1$ and hence

$$\mathbb{P}_x[N_x < \infty] = \mathbb{P}_x[T_x = \infty] \sum_{m=0}^{\infty} \mathbb{P}_x[T_x < \infty]^m = \frac{\mathbb{P}_x[T_x = \infty]}{1 - \mathbb{P}_x[T_x < \infty]} = 1.$$

Furthermore

$$\begin{aligned}
\mathbb{E}_x[N_x] &= \sum_{m=1}^{\infty} m \, \mathbb{P}_x[N_x = m] = \sum_{m=1}^{\infty} m \, \mathbb{P}_x[T_x < \infty]^m \, \mathbb{P}_x[T_x = \infty] \\
&= \mathbb{P}_x[T_x < \infty] \, \mathbb{P}_x[T_x = \infty] \, \frac{\mathrm{d}}{\mathrm{d} \, \mathbb{P}_x[T_x < \infty]} \sum_{m=1}^{\infty} \mathbb{P}_x[T_x < \infty]^m \\
&= \mathbb{P}_x[T_x < \infty] \, \mathbb{P}_x[T_x = \infty] \, \frac{\mathrm{d}}{\mathrm{d} \, \mathbb{P}_x[T_x < \infty]} \frac{1}{1 - \mathbb{P}_x[T_x < \infty]} \\
&= \frac{\mathbb{P}_x[T_x < \infty] \, \mathbb{P}_x[T_x = \infty]}{(1 - \mathbb{P}_x[T_x < \infty])^2} = \frac{\mathbb{P}_x[T_x < \infty]}{1 - \mathbb{P}_x[T_x < \infty]}.
\end{aligned}$$

Hence, $\mathbb{E}_x[N_x] < \infty$ implies $\mathbb{P}_x[T_x < \infty] < 1$, and vice versa. The remaining implications follow by negation. □

A Markov chain may possess both, recurrent and transient states as, e.g., the two state Markov chain given by

$$P = \begin{pmatrix} 1 - a & a \\ 0 & 1 \end{pmatrix}.$$

for some $a \in (0, 1)$. This example is actually a nice illustration of the next proposition.

Proposition 3.17 *Consider a Markov chain $\{X_k\}_{k\in\mathbb{N}}$ on a state space $\mathbf{S}$.*

1. *If $\{X_k\}_{k\in\mathbb{N}}$ admits some stationary distribution π and $y \in \mathbf{S}$ is some transient state then $\pi(y) = 0$.*

2. *If the state space $\mathbf{S}$ is finite, then there exists at least some recurrent state $x \in \mathbf{S}$.*

Proof: 1. Assume we had proven that $\mathbb{E}_x[N_y] < \infty$ for arbitrary $x \in \mathbf{S}$ and transient $y \in \mathbf{S}$, which implies $P^k(x, y) \to 0$ for $k \to \infty$. Then

$$\pi(y) = \sum_{x\in\mathbf{S}} \pi(x)P^k(x, y)$$

for every $k \in \mathbb{N}$, and finally

$$\pi(y) = \lim_{k\to\infty} \sum_{x\in\mathbf{S}} \pi(x)P^k(x, y) = \sum_{x\in\mathbf{S}} \pi(x) \lim_{k\to\infty} P^k(x, y) = 0.$$

where exchanging summation and limit is justified by the theorem of dominated convergence (e.g., [2, Appendix]), which proves the statement. Hence, it remains to prove $\mathbb{E}_x[N_y] < \infty$.

If $\mathbb{P}_y[T_y < \infty] = 0$, then $\mathbb{E}_x[N_y] = 0 < \infty$. Now assume that $\mathbb{P}_y[T_y < \infty] > 0$. Then, we obtain

$$\begin{aligned} \mathbb{E}_x[N_y] &= \sum_{m=1}^{\infty} m\, \mathbb{P}_x[T_y < \infty]\, \mathbb{P}_y[T_y < \infty]^{m-1}\, \mathbb{P}_y[T_y = \infty] \\ &= \frac{\mathbb{P}_x[T_y < \infty]}{\mathbb{P}_y[T_y < \infty]}\, \mathbb{E}_y[N_y] < \infty \end{aligned}$$

where the last inequality is due to transience of y and Thm. 3.16.

2. Proof left as an excercise (Hint: Use Proposition 3.17 and think on properties of stochastic matrices). □

The following theorem gives some additional insight into the relation between different states. It states that recurrence and transience are class properties.

Theorem 3.18 *Consider two states $x, y \in \mathbf{S}$ that communicate. Then*

1. *If x is recurrent then y is recurrent;*

2. *If x is transient then y is transient.*

Proof: Since x and y communicate, there exist integers $m, n \in \mathbb{N}$ such that $P^m(x, y) > 0$ and $P^n(y, x) > 0$. Introducing $q = P^m(x, y)P^n(y, x) > 0$, and exploiting the Chapman-Kolmogorov equation, we get $P^{n+k+m}(x, x) \geq$

$P^m(x,y)P^k(y,y)P^n(y,x) = qP^k(y,y)$ and $P^{n+k+m}(y,y) \geq qP^k(x,x)$, for $k \in \mathbb{N}$. Consequently,

$$\begin{aligned}\mathbb{E}_y[N_y] &= \mathbb{E}_y\left[\sum_{k=1}^{\infty} 1_{\{X_k=y\}}\right] = \sum_{k=1}^{\infty} \mathbb{E}_y[1_{\{X_k=y\}}] = \sum_{k=1}^{\infty} \mathbb{P}_y[X_k = y] \\ &= \sum_{k=1}^{\infty} P^k(y,y) \leq \frac{1}{q} \sum_{k=m+n}^{\infty} P^k(x,x) \leq \frac{1}{q}\,\mathbb{E}_x[N_x].\end{aligned}$$

Analogously, we get $\mathbb{E}_x[N_x] \leq \mathbb{E}_y[N_y]/q$. Now, the two statements directly follow by Thm. 3.16. □

As a consequence of Theorem 3.18, all states of an irreducible Markov chain are of the same nature: We therefore call an irreducible Markov chain recurrent or transient, if one of its states (and hence all) is recurrent, respectively, transient. Let us summarize the stability properties introduced so far. Combining Theorem 3.18 and Prop. 3.17 we conclude:

- Given some finite state space Markov chain
  - (i) that is not irreducible: there exists at least one recurrent communication class that moreover is closed.
  - (ii) that is irreducible: all states are recurrent, hence so is the Markov chain.
- Given some countable infinite state space Markov chain
  - (i) that is not irreducible: there may exist recurrent as well as transient communication classes.
  - (ii) that is irreducible: all states are either recurrent or transient.

We now address the important question of existence and uniqueness of invariant measures and stationary distributions. The following theorem states that for irreducible and recurrent Markov chains there always exists a unique invariant measure (up to a multiplicative factor).

Theorem 3.19 *Consider an irreducible and recurrent Markov chain. For an arbitrary state $x \in \mathbf{S}$ define $\mu = \big(\mu(y)\big)_{y\in\mathbf{S}}$ with*

$$\mu(y) = \mathbb{E}_x\left[\sum_{n=1}^{T_x} 1_{\{X_n=y\}}\right], \tag{15}$$

the expected value for the number visits in y before returning to x. Then

1. $0 < \mu(y) < \infty$ *for all* $y \in \mathbf{S}$. *Moreover,* $\mu(x) = 1$ *for the state* $x \in \mathbf{S}$ *chosen in the eq. (15).*
2. $\mu = \mu P$.
3. *If* $\nu = \nu P$ *for some measure* ν, *then* $\nu = \alpha\mu$ *for some* $\alpha \in \mathbb{R}$.

The interpretation of eq. (15) is this: for some fixed $x \in \mathbf{S}$ the invariant measure $\mu(y)$ is proportional to the number of visits to y before returning to x. Note that the invariant measure μ defined in (15) in general depends on the state $x \in \mathbf{S}$ chosen, since $\mu(x) = 1$ per construction. This reflects the fact that μ is only determined up to some multiplicative factor (stated in (iii)). We further remark that eq. (15) defines for every $x \in \mathbf{S}$ some invariant distribution, however for some arbitrarily given invariant measure μ, in general there does not exist an $x \in \mathbf{S}$ such that eq. (15) holds.

Proof: 1. Note that due to recurrence of x and definition of μ we have

$$\begin{aligned}\mu(x) &= \mathbb{E}_x\left[\sum_{n=1}^{T_x} 1_{\{X_n=x\}}\right] = \sum_{n=1}^{\infty} \mathbb{E}_x[1_{\{X_n=x\}} 1_{\{n\le T_x\}}] \\ &= \sum_{n=1}^{\infty} \mathbb{P}_x[X_n = x, n \le T_x] = \sum_{n=1}^{\infty} \mathbb{P}_x[T_x = n] = \mathbb{P}_x[T_x < \infty] = 1,\end{aligned}$$

which proves $\mu(x) = 1$. We postpone the second part of the first statement and prove

2. Observe that for $n \in \mathbb{N}$, the event $\{T_x \ge n\}$ depends only on the random variables $X_0, X_1, \ldots, X_{n-1}$. Thus

$$\mathbb{P}_x[X_n = z, X_{n-1} = y, T_x \ge n] = \mathbb{P}_x[X_{n-1} = y, T_x \ge n]\, P(y, z).$$

Now, we have for arbitrary $z \in \mathbf{S}$

$$\begin{aligned}\sum_{y\in\mathbf{S}} \mu(y)P(y,z) &= \mu(x)P(x,z) + \sum_{y\neq x} \mu(y)P(y,z) \\ &= P(x,z) + \sum_{y\neq x}\sum_{n=1}^{\infty} \mathbb{P}_x[X_n = y, n \le T_x]\, P(y,z) \\ &= P(x,z) + \sum_{n=1}^{\infty}\sum_{y\neq x} \mathbb{P}_x[X_{n+1} = z, X_n = y, n \le T_x] \\ &= \mathbb{P}_x[X_1 = z] + \sum_{n=1}^{\infty} \mathbb{P}_x[X_{n+1} = z, n+1 \le T_x] \\ &= \mathbb{P}_x[X_1 = z, 1 \le T_x] + \sum_{n=2}^{\infty} \mathbb{P}_x[X_n = z, n \le T_x] \\ &= \sum_{n=1}^{\infty} \mathbb{P}_x[X_n = z, n \le T_x] = \mu(z),\end{aligned}$$

where for the second equality we used $\mu(x) = 1$ and for the fourth equality we used that $X_n = y$, $n \leq T_x$ and $x \neq y$ implies $n+1 \leq T_x$. Thus we proved $\mu P = \mu$.

1. (continued) Since P is irreducible, there exist integers $k, j \in \mathbb{N}$ such that $P^k(x,y) > 0$ and $P^j(y,x) > 0$ for every $y \in \mathbf{S}$. Therefore, for every $k \in \mathbb{N}$ and exploiting statement 2.), we have

$$0 < \mu(x)P^k(x,y) \leq \sum_{z \in \mathbf{S}} \mu(z)P^k(z,y) = \mu(y).$$

On the other hand,

$$\mu(y) = \frac{\mu(y)P^j(y,x)}{P^j(y,x)} \leq \frac{\sum_{z \in \mathbf{S}} \mu(z)P^j(z,x)}{P^j(y,x)} = \frac{\mu(x)}{P^j(y,x)} < \infty.$$

Hence, the first statement has been proven.

3. The first step to prove the uniqueness of μ is to show that μ is minimal, which means that $\nu \geq \mu$ holds for any other invariant measure ν satisfying $\nu(x) = \mu(x) = 1$. We prove by induction that

$$\nu(z) \geq \sum_{n=1}^{k} \mathbb{P}_x[X_n = z, n \leq T_x] \tag{16}$$

holds for every $z \in \mathbf{S}$. Note that the right hand side of eq. (16) converges to $\mu(z)$ as $k \to \infty$ (cmp. proof of 1.). For $k = 1$ it is

$$\nu(z) = \sum_{y \in \mathbf{S}} \nu(y)P(y,z) \geq P(x,z) = \mathbb{P}_x[X_1 = z, 1 \leq T_x].$$

Now, assume that eq. (16) holds for some $k \in \mathbb{N}$. Then

$$\begin{aligned}
\nu(z) &\geq \nu(x)P(x,z) + \sum_{y \neq x} \nu(y)P(y,z) \\
&\geq P(x,z) + \sum_{y \neq x} \sum_{n=1}^{k} \mathbb{P}_x[X_n = y, n \leq T_x]P(y,z) \\
&= \mathbb{P}_x[X_1 = z, 1 \leq T_x] + \sum_{n=1}^{k} \mathbb{P}_x[X_{n+1} = z, n+1 \leq T_x] \\
&= \sum_{n=1}^{k+1} \mathbb{P}_x[X_n = z, n \leq T_x].
\end{aligned}$$

Therefore, eq. (16) holds for every $k \in \mathbb{N}$, and in the limit we get $\nu \geq \mu$. Define $\lambda = \nu - \mu$; since P is irreducible, for every $z \in \mathbf{S}$ there exists some integer $k \in \mathbb{N}$ such that $P^k(z,x) > 0$. Thus

$$0 = \lambda(x) = \sum_{y \in \mathbf{S}} \lambda(y)P^k(y,x) \geq \lambda(z)P^k(z,x),$$

implying $\lambda(z) = 0$ and finally $\nu = \mu$. Now, if we relax the condition $\nu(x) = 1$, then statement 3. follows with $c = \nu(x)$. □

We already know that the converse of Theorem 3.19 is false, since there are transient irreducible Markov chains that possess invariant measures. For example, the random walk on $\mathbb{N}$ is transient for $p > 1/2$, but admits an invariant measure. At the level of invariant measures, nothing more can be said. However, if we require that the invariant measure is a probability measure, then it is possible to give necessary and sufficient conditions. These involve the expected return times

$$\mathbb{E}_x[T_x] \quad = \quad \sum_{n=1}^{\infty} n \, \mathbb{P}_x[T_x = n]. \tag{17}$$

Depending on the behaviour of $\mathbb{E}_x[T_x]$, we further distinguish two types of states:

Definition 3.20 *A recurrent state $x \in \mathbf{S}$ is called **positive recurrent**, if*

$$\mathbb{E}_x[T_x] \quad < \quad \infty$$

*and **null recurrent** otherwise.*

In view of eq. (17) the difference between positive and null recurrence is manifested in the decay rate of $\mathbb{P}_x[T_x = n]$ for $n \to \infty$. If $\mathbb{P}_x[T_x = n]$ decays too slowly as $n \to \infty$, then $\mathbb{E}_x[T_x]$ is infinite and the state is null recurrent. On the other hand, if $\mathbb{P}_x[T_x = n]$ decays rapidly in the limit $n \to \infty$, then $\mathbb{E}_x[T_x]$ will be finite and the state is positive recurrent.

As for recurrence, positive and null recurrence are class properties [2]. Hence, we call a Markov chain positive or null recurrent, if one of its states (and therefore all) is positive, respectively, null recurrent. The next theorem illustrates the usefulness of positive recurrence and gives an additional useful interpretation of the stationary distribution.

Theorem 3.21 *Consider an irreducible Markov chain. Then the Markov chain is positive recurrent, if and only if there exists a stationary distribution. Under these conditions, the stationary distribution is unique and positive everywhere, with*

$$\pi(x) \quad = \quad \frac{1}{\mathbb{E}_x[T_x]}.$$

Hence $\pi(x)$ can be interpreted as the inverse of the expected first return time to state $x \in \mathbf{S}$.

Proof: Theorem 3.19 states that an irreducible and recurrent Markov chain admits an invariant measure μ defined through (15) for an arbitrary $x \in \mathbf{S}$. Thus

$$\begin{aligned}\sum_{y\in\mathbf{S}} \mu(y) &= \sum_{y\in\mathbf{S}} \mathbb{E}_x\left[\sum_{n=1}^{T_x} 1_{\{X_n=y\}}\right] = \mathbb{E}_x\left[\sum_{n=1}^{\infty}\sum_{y\in\mathbf{S}} 1_{\{X_n=y\}} 1_{\{n\le T_x\}}\right] \\ &= \mathbb{E}_x\left[\sum_{n=1}^{\infty} 1_{\{n\le T_x\}}\right] = \sum_{n=1}^{\infty} \mathbb{P}_x[T_x \ge n] \\ &= \sum_{n=1}^{\infty}\sum_{k=n}^{\infty} \mathbb{P}_x[T_x = k] = \sum_{k=1}^{\infty} k\,\mathbb{P}_x[T_x = k] = \mathbb{E}_x[T_x],\end{aligned}$$

which is by definition finite in the case of positive recurrence. Therefore the stationary distribution can be obtained by normalization of μ with $\mathbb{E}_x(T_x)$ yielding

$$\pi(x) = \frac{\mu(x)}{\mathbb{E}_x(T_x)} = \frac{1}{\mathbb{E}_x(T_x)}.$$

Since the state x was chosen arbitrary this is true for all $x \in \mathbf{S}$. Uniqueness and positivity of π follows from Theorem 3.19. On the other hand, if there exists a stationary distribution the Markov chain must be recurrent because otherwise $\pi(x)$ would be zero for all $x \in \mathbf{S}$ according to Theorem 3.17. Positive recurrence follows from the uniqueness of π and the consideration above. □

Our considerations in the proof of Theorem 3.21 easily leads to a criteria to distinguish positive recurrence from null recurrence.

Corollary 3.22 *Consider an irreducible recurrent Markov chain $\{X_k\}_{k\in\mathbb{N}}$ with invariant measure $\mu = (\mu(x))_{x\in\mathbf{S}}$. Then*

1. *$\{X_k\}_{k\in\mathbb{N}}$ positive recurrent $\Leftrightarrow \sum_{x\in\mathbf{S}} \mu(x) < \infty$,*
2. *$\{X_k\}_{k\in\mathbb{N}}$ null recurrent $\Leftrightarrow \sum_{x\in\mathbf{S}} \mu(x) = \infty$.*

Proof: The proof is left as an exercise. □

For the finite state space case, we have the following powerful statement.

Theorem 3.23 *Every irreducible Markov chain on a finite state space is positive recurrent and therefore admits a unique stationary distribution that is positive everywhere.*

Proof: The proof is left as an exercise. □

For the general possibly non-irreducible case, the results of this section are summarized in the next

Proposition 3.24 *Let $C \subset \mathbf{S}$ denote a communication class corresponding to some Markov chain $\{X_k\}_{k\in\mathbb{N}}$ on the state space $\mathbf{S}$.*

1. *If C is not closed, then all states in C are transient.*
2. *If C is closed and finite, then all states in C are positive recurrent.*
3. *If all state in C are null recurrent, then C is necessarily infinite.*

Proof: The proof is left as an exercise. □

4 Asymptotic behavior

The asymptotic behavior of distributions and transfer operators is closely related to so-called ergodic properties of the Markov chain. The term ergodicity is not consistently used in literature. In ergodic theory, it roughly refers to the fact that space and time averages coincide (as, e.g., stated in the strong law of large numbers by Thm. 5.1). In the theory of Markov chain, however, the meaning is slightly different. Here, ergodicity is related to the convergence of probability distributions ν_0 in time, i.e., $\nu_k \to \pi$ as $k \to \infty$, and assumes aperiodicity as a necessary condition.

4.1 k-step transition probabilities and distributions

We prove statements for the convergence of k-step probabilities involving transient, null recurrent and finally positive recurrent states.

Proposition 4.1 *Let $y \in \mathbf{S}$ denote a transient state of some Markov chain with transition function P. Then, for any initial state $x \in \mathbf{S}$*

$$P^k(x, y) \to 0$$

as $k \to \infty$. Hence, the y-th column of P^k tends to zero as $k \to \infty$.

Proof: This has already been proved in the proof of Prop. 3.17 □

The situation is similar for an irreducible Markov chain that is null recurrent (and thus defined on a infinite countable state space due to Theorem 3.23):

Theorem 4.2 (Orey's Theorem) *Let $\{X_k\}_{k\in\mathbb{N}}$ be an irreducible null recurrent Markov chain on $\mathbf{S}$. Then, for all pairs of states $x, y \in \mathbf{S}$*

$$P^k(x, y) \to 0$$

as $k \to \infty$.

Proof: See, e.g., [2], p.131. □

In order to derive a result for the evolution of k-step transition probabilities for positive recurrent Markov chains, we will exploit a powerful tool from probability theory, the **coupling method** (see, e.g., [6, 8]).

Definition 4.3 *A **coupling** of two random variables $X, Y : \Omega \to \mathbf{S}$ is a random variable $Z : \Omega \to \mathbf{S} \times \mathbf{S}$ such that*

$$\sum_{y\in\mathbf{S}} \mathbb{P}[Z = (x, y)] = \mathbb{P}[X = x], \text{ and } \sum_{x\in\mathbf{S}} \mathbb{P}[Z = (x, y)] = \mathbb{P}[Y = y]$$

for every $x \in \mathbf{S}$, and for every $y \in \mathbf{S}$, respectively. Hence, the coupling Z has X and Y as its marginals.

Note that, except for artificial cases, there exists infinitely many couplings of two random variables. The coupling method exploits the fact that the total variation distance between the two distributions $\mathbb{P}[X \in A]$ and $\mathbb{P}[Y \in \mathcal{A}]$ can be bounded in terms of the coupling Z.

Proposition 4.4 (Basic coupling inequality) *Consider two independent random variables $X, Y : \Omega \to \mathbf{S}$ with distributions ν and π, respectively, defined via $\nu(x) = \mathbb{P}[X = x]$ and $\pi(y) = \mathbb{P}[Y = y]$ for $x, y \in \mathbf{S}$. Then*

$$\|\nu - \pi\|_{TV} \leq 2\,\mathbb{P}[X \neq Y],$$

with $[X \neq Y] = \{\omega \in \Omega : X(\omega) \neq Y(\omega)\}$.

Proof: We have for some subset $A \subset \mathbf{S}$

$$\begin{aligned} |\nu(A) - \pi(A)| &= |\,\mathbb{P}[X \in A] - \mathbb{P}[Y \in A]| \\ &= |\,\mathbb{P}[X \in A, X = Y] + \mathbb{P}[X \in A, X \neq Y] \\ &\quad - \mathbb{P}[Y \in A, X = Y] - \mathbb{P}[Y \in A, X \neq Y]| \\ &= |\,\mathbb{P}[X \in A, X \neq Y] - \mathbb{P}[Y \in A, X \neq Y]| \\ &\leq \mathbb{P}[X \neq Y]. \end{aligned}$$

Since

$$\|\nu - \pi\|_{TV} = 2 \sup_{A \subset \mathbf{S}} |\nu(A) - \pi(A)|$$

the statement directly follows. □

Note that the term $\mathbb{P}[X \neq Y]$ in the basic coupling inequality can be stated in terms of the coupling Z:

$$\mathbb{P}[X \neq Y] = \sum_{x,y \in \mathbf{S}, x \neq y} \mathbb{P}[Z = (x, y)] = 1 - \sum_{x \in \mathbf{S}} \mathbb{P}[Z = (x, x)].$$

Since there are many couplings the aim is two construct a coupling Z such that $\sum_{x \neq y} \mathbb{P}[Z = (x, y)]$ is as small, or $\sum_x \mathbb{P}[Z = (x, x)]$ is as large as possible. To prove convergence results for the evolution of the distribution of some Markov chain, we exploit a specific (and impressive) example of the coupling method.

Consider an irreducible, aperiodic, positive recurrent Markov chain $X = \{X_k\}_{k \in \mathbb{N}}$ with stationary distribution π and some initial distribution ν_0. Moreover, define another independent Markov chain $Y = \{Y_k\}_{k \in \mathbb{N}}$ that has the same transition function as X, but the stationary distribution π as

initial distribution. Observe that Y is a stationary process, i.e., the induced distribution of Y_k equals π for all $k \in \mathbb{N}$. Then, we can make use of the coupling method by interpreting the Markov chains as random variables $X, Y : \Omega \to \mathbf{S}^{\mathbb{N}}$ and consider some coupling $Z : \Omega \to \mathbf{S}^{\mathbb{N}} \times \mathbf{S}^{\mathbb{N}}$. Define the **coupling time** $T_c : \Omega \to \mathbb{N}$ by

$$T_c = \min\{k \geq 1 : X_k = Y_k\};$$

T_c is the first time at which the Markov chains X and Y met; moreover, it is stopping time for Z. The next proposition bounds the distance between the distributions ν_k and π at time k in terms of the coupling time T_c.

Proposition 4.5 *Consider some irreducible, aperiodic, positive recurrent Markov chain with initial distribution ν_0 and stationary distribution π. Then, the distribution ν_k at time k satisfies*

$$\|\nu_k - \pi\|_{TV} \leq 2\,\mathbb{P}[k < T_c]$$

for every $k \in \mathbb{N}$, where T_c denote the coupling time defined above.

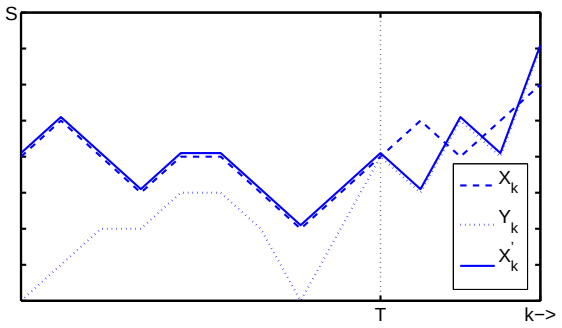


Figure 8: The construction of the coupled process X' as needed in the proof of Prop. 4.5. Here, T denotes the value of the coupling time T_c for this realization.

Proof: We start by defining a new stochastic process $X' = \{X'_k\}_{k\in\mathbb{N}}$ with $X'_k : \Omega \to \mathbf{S}$ (see Fig. 8) according to

$$X'_k = \begin{cases} X_k; & \text{if } k < T_c, \\ Y_k; & \text{if } k \geq T_c. \end{cases}$$

Due to the strong Markov property 3.15 (applied to the coupled Markov chain $(X_k, Y_k)_{k\in\mathbb{N}}$), X' is a Markov chain with the same transition probabilities as X and Y. As a consequence of the definition of X' we have

$\mathbb{P}_{\nu_0}[X_k \in A] = \mathbb{P}_{\nu_0}[X'_k \in A]$ for $k \in \mathbb{N}$ and every $A \subset \mathbf{S}$, hence the distributions of X_k and X'_k are the same. Hence, from the basic coupling inequality, we get

$$|\nu_k(A) - \pi(A)| = |\,\mathbb{P}[X'_k \in A] - \mathbb{P}[Y_k \in A]| \leq 2\,\mathbb{P}[X'_k \neq Y_k].$$

Since $\{X'_k \neq Y_k\} \subset \{k < T_c\}$, we finally obtain

$$\mathbb{P}[X'_k \neq Y_k] \leq \mathbb{P}[k < T_c],$$

which implies the statement. □

Proposition 4.5 enables us to prove the convergence of ν_k to π by proving that $\mathbb{P}[k < T_c]$ converges to zero.

Theorem 4.6 *Consider some irreducible, aperiodic, positive recurrent Markov chain with stationary distribution π. Then, for any initial probability distribution ν_0, the distribution of the Markov chain at time k satisfies*

$$\|\nu_k - \pi\|_{TV} \to 0$$

as $k \to \infty$. In particular, choosing the initial distribution to be a delta distribution at $x \in \mathbf{S}$, we obtain

$$\|P^k(x, \cdot) - \pi\|_{TV} \to 0$$

as $k \to \infty$.

Proof: It suffices to prove $\mathbb{P}[T_c < \infty] = 1$. Moreover, if we fixe some state $x^* \in \mathbf{S}$ and consider the stopping time

$$T_c^* = \inf\{k \geq 1; X_k = x^* = Y_k\},$$

then $\mathbb{P}[T_c < \infty] = 1$ follows from $\mathbb{P}[T_c^* < \infty] = 1$. To prove the latter statement, consider the coupling $Z = (Z_k)_{k\in\mathbb{N}}$ with $Z_k = (X_k, Y_k) \in \mathbf{S} \times \mathbf{S}$ with $X = \{X_k\}_{k\in\mathbb{N}}$ and $Y = \{Y_k\}_{k\in\mathbb{N}}$ defined as above. Because X and Y are independent, the transition matrix P_Z of Z is given by

$$P_Z\big((v, w), (x, y)\big) = P(v, w)P(x, y)$$

for all $v, w, x, y \in \mathbf{S}$. Obviously, Z has a stationary distribution given by

$$\pi_Z(x, y) = \pi(x)\pi(y).$$

Furthermore the coupled Markov chain is irreducible: consider $(v, w), (x, y) \in \mathbf{S} \times \mathbf{S}$ arbitrary. Since X and Y are irreducible *and* aperiodic we can choose

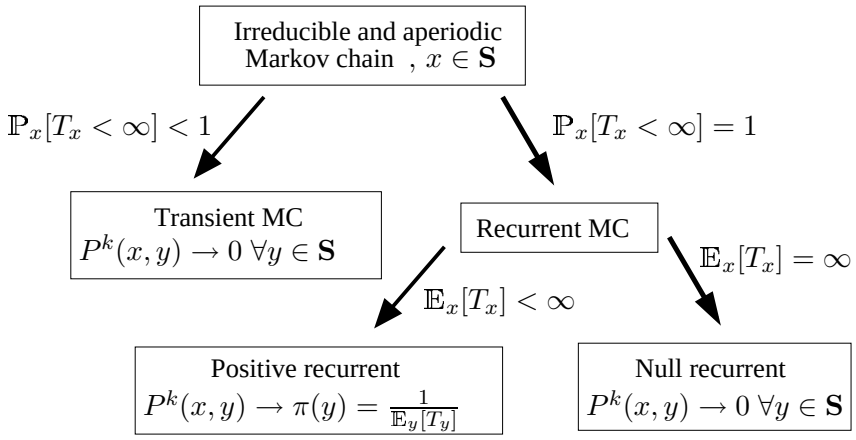


Figure 9: Long run behaviour of an irreducible aperiodic Markov chain.

an integer $k^* > 0$ such that $P^{k^*}(v,w) > 0$ *and* $P^{k^*}(x,y) > 0$ holds, see Prop. 3.8. Therefore

$$P_Z^{k^*}\big((v,w),(x,y)\big) = P^{k^*}(v,w)P^{k^*}(x,y) > 0.$$

Hence Z is irreducible. Finally observe that T_c^* is the first return time of the coupled Markov chain to the state (x^*,x^*). Since Z is irreducible and has a stationary distribution, it is positive recurrent according to Thm. 3.21. By Thm. 3.18, this implies $\mathbb{P}[T_c^* < \infty] = 1$, which completes the proof of the statement. □

Fig. 9 summarizes the long run behavior of irreducible and aperiodic Markov chains.

4.2 Time reversal and reversibility

The notions of time reversal and time reversibility are very productive, in particular w.r.t. the spectral theory, the central limit theory and theory of Monte Carlo methods, as we will see.

Chang [3] has a nice motivation of time reversibility: Let $X_0, X_1, \ldots$ denote a Markov chain with transition function P. Imagine that I recorded a movie of the sequence of states $(X_0, \ldots, X_n)$, and I am showing you the movie on my fancy machine that can play the tape forward or backward equally well. Can you tell by watching the sequence of transitions on the movie whether I am showing it forward or backward?

To answer this question, we determine the transition probabilities of the Markov chain $\{Y_k\}_{k\in\mathbb{N}}$ obtained by reversing time for the original Markov

chain $\{X_k\}_{k\in\mathbb{N}}$. Given some probability distribution $\pi > 0$, we require that

$$\mathbb{P}_\pi[Y_0 = x_m, \ldots, Y_m = x_0] \quad = \quad \mathbb{P}_\pi[X_0 = x_0, \ldots, X_m = x_m]$$

holds for every $m \in \mathbb{N}$ and every $x_0, \ldots, x_m \in \mathbf{S}$ in the case of reversibility. For the special case $m = 1$ we have

$$\mathbb{P}_\pi[Y_0 = y, Y_1 = x] \quad = \quad \mathbb{P}_\pi[X_0 = x, X_1 = y] \tag{18}$$

for $x, y \in \mathbf{S}$. Denote by Q and P the transition functions of the Markov chains $\{Y_k\}$ and $\{X_k\}$, respectively. Then, by equation (18) we obtain

$$\pi(y)Q(y,x) \quad = \quad \pi(x)P(x,y). \tag{19}$$

Note that the diagonals of P and Q are always equal, hence $Q(x,x) = P(x,x)$ for every $x \in \mathbf{S}$. Moreover, from eq. (20) we deduce by summing over all $x \in \mathbf{S}$ that π is some stationary distribution of the Markov chain.

Definition 4.7 *Consider some Markov chain $X = \{X_k\}_{k\in\mathbb{N}}$ with transition function P and stationary distribution $\pi > 0$. Then, the Markov chain $\{Y_k\}_{k\in\mathbb{N}}$ with transition function Q defined by*

$$Q(y,x) = \frac{\pi(x)P(x,y)}{\pi(y)} \tag{20}$$

*is called the **time-reversed Markov chain** (assoziated with X).*

Example 4.8 *Consider the two state Markov chain given by*

$$P \quad = \quad \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

for $a, b \in [0,1]$. The two state Markov chain is an exceptionally simple example, since we know on the one hand that the diagonal entries of Q and P are identical, and on the other hand that Q is a stochastic matrix. Consequently $Q = P$.

Example 4.9 *Consider a Markov chain on the state space $\mathbf{S} = \{1,2,3\}$ given by*

$$P \quad = \quad \begin{pmatrix} 1-a & a & 0 \\ 0 & 1-b & b \\ c & 0 & 1-c \end{pmatrix}.$$

for $a, b, c \in [0,1]$. Denote by π the stationary distribution (which exists due to Theorem 3.23). Then, $\pi = \pi P$ is equivalent to $a\pi(1) = b\pi(2) = c\pi(3)$. A short calculation reveals

$$\pi \quad = \quad \frac{1}{ab + ac + bc}\big(bc, ac, ab\big).$$

Once again, we have only to compute the off–diagonal entries of Q. We get

$$Q = \begin{pmatrix} 1-a & 0 & a \\ b & 1-b & 0 \\ 0 & c & 1-c \end{pmatrix}.$$

For illustration, consider the case $a = b = c = 1$. Then P is periodic with period $d = 3$; it moves deterministically: $1 \to 2 \to 3 \to 1 \ldots$. By construction, the matrix Q corresponds to the time reversed Markov chain that moves like: $3 \to 2 \to 1 \to 3 \ldots$, but this is exactly the dynamics defined by Q.

Definition 4.10 *Consider some Markov chain $X = \{X_k\}_{k\in\mathbb{N}}$ with transition function P and stationary distribution $\pi > 0$, and its associated time-reversed Markov chain with transition function Q. Then X is called* ***reversible*** *w.r.t. π, if*

$$P(x, y) = Q(x, y)$$

for all $x, y \in \mathbf{S}$.

The above definition can be reformulated: a Markov chain is reversible w.r.t. π, if and only if the **detailed balance condition**

$$\pi(x)P(x, y) = \pi(y)P(y, x) \tag{21}$$

is satisfied for every $x, y \in \mathbf{S}$. Eq. (21) has a nice interpretation in terms of the probability flux defined in (11). Recall that the flux from x to y is defined by $\text{flux}_\pi(x, y) = \pi(x)P(x, y)$. Thus, eq. (21) states that the flux from x to y is the same as the flux from y to x—it is *locally balanced* between each pair of states: $\text{flux}_\pi(x, y) = \text{flux}_\pi(y, x)$ for $x, y \in \mathbf{S}$. This is a much stronger condition than the *global balance* condition that characterizes stationarity. The global balance condition that can be rewritten as $\sum_x \pi(y)P(y, x) = \pi(y) = \sum_x \pi(x)P(x, y)$ states that the total flux leaving state x is the same as the total flux into state x: $\text{flux}_\pi(x, \mathbf{S} \setminus \{x\}) = \text{flux}_\pi(\mathbf{S} \setminus \{x\}, x)$.

Corollary 4.11 *Given some Markov chain with transition function P and stationary distribution π. If there exist a pair of states $x, y \in \mathbf{S}$ with $\pi(x) > 0$ such that*

$$P(x, y) > 0, \quad \textit{while} \quad P(y, x) = 0$$

then the detailed balance condition cannot hold for P, hence the Markov chain is not reversible. This is in particular the case, if the Markov chain is periodic with period $d > 2$.

Application of Corollary 4.11 yields that the three state Markov chain defined in Example 4.9 cannot be reversible.

Example 4.12 *Consider the random walk on* $\mathbb{N}$ *with fixed parameter* $p \in (0, \frac{1}{2})$*. The Markov chain is given by*

$$P(x, x+1) = p \quad \textit{and} \quad P(x+1, x) = 1 - p$$

for $x \in \mathbf{S}$ *and* $P(0,0) = 1 - p$*, while all other transition probabilities are zero. It is irreducible and admits a unique stationary distribution given by*

$$\pi(0) \;=\; \frac{1-2p}{1-p} \quad \textit{and} \quad \pi(k) = \pi(0) \left(\frac{p}{1-p} \right)^k$$

for $k > 0$*. Obviously, we expect no trouble due to Corollary 4.11. Moreover, we have*

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

for arbitrary $x \in \mathbf{S}$*; hence, the detailed balance condition holds for* P *and the random walk on* $\mathbb{N}$ *is reversible.*

4.3 Some spectral theory

We now introduce the necessary notions from spectral theory in order to analyze the asymptotic behavior of transfer operators. Throughout this section, we assume that π is some stationary distribution of a Markov chain with transition function P. Note that π is neither assumed to be unique nor positive everywhere.

We start by introducing the Banach spaces (of equivalence classes)

$$l^r(\pi) = \{v : \mathbf{S} \to \mathbb{C} : \sum_{x \in \mathbf{S}} |v(x)|^r \pi(x) < \infty\},$$

for $1 \leq r < \infty$ with corresponding norms

$$\|v\|_r = \left(\sum_{x \in \mathbf{S}} |v(x)|^r \pi(x) \right)^{1/r}$$

and

$$l^\infty(\pi) = \{v : \mathbf{S} \to \mathbb{C} : \pi\text{-}\sup_{x \in \mathbf{S}} |v(x)| < \infty\},$$

with supremums norm defined by

$$\|v\|_\infty = \pi\text{-}\sup_{x \in \mathbf{S}} |v(x)| = \sup_{x \in \mathbf{S}, \pi(x) > 0} |v(x)|.$$

Given two functions $u, v \in l^2(\pi)$, the π–weighted **scalar product** $\langle \cdot, \cdot \rangle_\pi : \mathbf{S} \times \mathbf{S} \to \mathbb{C}$ is defined by

$$\langle u, v \rangle_\pi \;=\; \sum_{x \in \mathbf{S}} u(x) \bar{v}(x) \pi(x),$$

where $\bar{v}$ denotes the conjugate complex of v. Note that $l^2(\pi)$, equipped with the scalar product $\langle \cdot, \cdot \rangle_\pi$, is a **Hilbert space**.

Remark. In general, the elements of the above introduced function spaces are equivalence classes of functions $[f] = \{g : \mathbf{S} \to \mathbb{C} : g(x) = f(x), \text{ if } \pi(x) > 0\}$ rather than single functions $f : \mathbf{S} \to \mathbb{C}$ (this is equivalent to the approach of introducing equivalence classes of Lebesgue-integrable functions (see, e.g., [11])). Hence, functions that differ on a set of points with π-measure zero are considered to be equivalent. However, if the probability distribution π is positive everywhere, we regain the interpretation of functions as elements.

Before proceeding, we need the following two definitions.

Definition 4.13 *Given some Markov chain with stationary distribution π.*

1. *Some measuere $\nu \in \mathcal{M}$ is said to be **absolutely continuous** w.r.t. π, in short $\nu \ll \pi$, if*

$$\pi(x) = 0 \quad \Rightarrow \quad \nu(x) = 0$$

   *for every $x \in \mathbf{S}$. In this case, there exists some function $f : \mathbf{S} \to \mathbb{C}$ such that $\nu = f\pi$. The function f is called the **Radon-Nikodym derivative** of ν w.r.t. π and sometimes denoted by $\mathrm{d}\nu/\mathrm{d}\pi$.*

2. *The stationary distribution π is called **maximal**, if every other stationary distribution ν is absolutely continuous w.r.t. π.*

In broad terms, a stationary distribution is maximal, if it possesses as many non–zero elements as possible. Note that a maximal stationary distribution need not be unique.

Example 4.14 *Consider the state space $\mathbf{S} = \{1, 2, 3, 4\}$ and a Markov chain with transition function*

$$P \;=\; \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Then $\nu_1 = (1, 0, 0, 0), \nu_2 = (0, 1, 0, 0), \nu_3 = (0, 0, 1, 0)$ are stationary distributions of P, but none of them is obviously maximal. In contrast to that, both $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$ and $\sigma = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0)$ are maximal stationary distributions. Note that since state $x = 4$ is transient, every stationary distribution ν satisfies $\nu(4) = 0$ due to Proposition 3.17.

To this end, we consider some Markov chain with maximal stationary distribution π. We now restrict the transfer operator P from the space of complex finite measures $\mathcal{M}$ to the space of complex finite measures that are absolutely continuous w.r.t. π. We define for $1 \leq r \leq \infty$

$$\mathcal{M}_r(\pi) = \{\nu \in \mathcal{M} : \nu \ll \pi \text{ and } \mathrm{d}\nu/\mathrm{d}\pi \in l^r(\pi)\}$$

with corresponding norm $\|\nu\|_{\mathcal{M}_r(\pi)} = \|\mathrm{d}\nu/\mathrm{d}\pi\|_r$. Note that $\|\nu\|_{\mathcal{M}_1(\pi)} = \|\nu\|_{TV}$ and $\mathcal{M}_1(\pi) \supseteq \mathcal{M}_2(\pi) \supseteq \ldots$. We now define the transfer operator $P : \mathcal{M}_1(\pi) \to \mathcal{M}_1(\pi)$ by

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

It can be shown by exploiting Hölders inequality that P is well-defined on any $\mathcal{M}_r(\pi)$ for $1 \leq r \leq \infty$.

It is interesting to note that the transfer operator P on $\mathcal{M}_1(\pi)$ induces some transfer operator $\mathcal{P}$ on $l^1(\pi)$: Given some $\nu \in \mathcal{M}_1(\pi)$ with derivative $v = \mathrm{d}\nu/\mathrm{d}\pi$, it follows that $\nu P \ll \pi$ (if π is some stationary measure with $\pi P = \pi$ then $\pi(y)$ implies $p(x, y)$ for every $x \in \mathbf{S}$ with $\pi(x) > 0$. Now, the statement directly follows). Hence, we define $\mathcal{P}$ by $(v\pi)P = (v\mathcal{P})\pi$. More precisely, it is $\mathcal{P} : l^1(\pi) \to l^1(\pi)$ given by

$$v\mathcal{P}(y) = \sum_{x \in \mathbf{S}} Q(y, x) v(x)$$

for $v \in l^1(\pi)$. Above, Q with $Q(y, x) = \pi(x)P(x, y)/\pi(y)$ is the transition function of the time-reversed Markov chain (see eq. (20)), which is an interesting relation between the original Markov chain and the time-reversed one. Actually, we could formulate all following results also in terms of the transfer operator $\mathcal{P}$, which is usually done for the general state space case. Here, however, we prefer to state the results related to the function space $\mathcal{M}_1(\pi)$, since then there is a direct relation to the action of the transfer operator and the (stochastic) matrix-vector multiplication from the left. In terms of $l^1(\pi)$, this important relation would only hold after some suitable reweighting (of the stochastic matrix). From a functional analytical point of view, however, the two function spaces $(\mathcal{M}_1(\pi), \|\cdot\|_{TV})$ and $(l^1(\pi), \|\cdot\|_1)$ are equivalent.

Central for our purpose will be notion of eigenvalues and eigenvectors of some transfer operator $P : \mathcal{M}_1(\pi) \to \mathcal{M}_1(\pi)$. Some number $\lambda \in \mathbb{C}$ is called an **eigenvalue** of P, if there exists some $\nu \in \mathcal{M}_1(\pi)$ with $\nu \neq 0$ satisfying the **eigenvalue equation**

$$\nu P \quad = \quad \lambda \nu. \tag{22}$$

The function ν is called an (left) **eigenvector** corresponding to the eigenvalue λ. Note that not every function ν satisfying (22) is an eigenvector, since ν has to fulfill the integrability condition $||\nu||_{TV} < \infty$ by definition (which, of course, is always satisfied in the finite state space case). The subspace of all eigenvectors corresponding to some eigenvalue λ is called the eigenspace corresponding to λ. By $\sigma(P)$ we denote the **spectrum** of P, which contains all eigenvalues of P. In the finite state space case, we have $\sigma(P) = \{\lambda \in \mathbb{C} : \lambda \text{ is eigenvalue of } P\}$, while for the infinite state space case, it may well contain elements that are not eigenvalues (see, e.g., [11, Kap. VI]).

The transfer operators considered above is closely related to a transfer operator acting on bounded (measurable) functions. Define $T : l^\infty(\pi) \to l^\infty(\pi)$ by

$$Tu(x) = \mathbb{E}_x[u(X_1)] = \sum_{y \in \mathbf{S}} P(x,y)u(y)$$

for $u \in l^\infty(\pi)$. We remark that for the important class of reversible Markov chains, T is simply given by $Tv(x) = \sum_y P(x,y)v(y)$ (which corresponds to the matrix vector multiplication from the right). For some function $\nu \in \mathcal{M}_1(\pi)$ and $u \in l^\infty(\pi)$, define the duality bracket $\langle \cdot, \cdot \rangle : \mathcal{M}_1(\pi) \times l^\infty(\pi)$ by

$$\langle \nu, u \rangle = \sum_{x \in \mathbf{S}} \nu(x)u(x).$$

Then, we have

$$\langle \nu P, u \rangle = \sum_{x \in \mathbf{S}} \sum_{y \in \mathbf{S}} \nu(y)P(y,x)u(x) = \sum_{y \in \mathbf{S}} \nu(y) \sum_{x \in \mathbf{S}} P(y,x)u(x) = \langle \nu, Tu \rangle,$$

hence T is the adjoint operator of P, or $P^* = T$. This fact can be widely exploited when dealing with spectral properties of P, since the spectrum of some operator is equal to the spectrum of its adjoint operator (see, e.g., [11, Satz VI.1.2]). Hence, if $\lambda \in \sigma(P)$, then there exists some non-vanishing function $u \in l^\infty(\pi)$ with $Tu = \lambda u$ (and analogously for the reversed implication).

Example 4.15 *Consider some transfer operator P acting on $\mathcal{M}_1(\pi)$. Then $\pi P = \pi$ (since $\mathbf{1}$ is in $\mathcal{M}_1(\pi)$) and consequently the $\lambda = 1$ is an eigenvalue of P.*

The next proposition collects some useful facts about the spectrum of the transfer operator.

Proposition 4.16 *Consider a transition function P on a countable state space with stationary distribution π. Then, for the associated transfer operator $P : \mathcal{M}_1(\pi) \to \mathcal{M}_1(\pi)$ holds:*

(a) *The spectrum of P is contained in the unit disc, i.e. $\lambda \in \sigma(P)$ implies $|\lambda| \leq 1$.*

(b) *$\lambda = 1$ is an eigenvalue of P, i.e., $1 \in \sigma(P)$.*

(c) *If $\lambda = a + ib$ is some eigenvalue of P, so is $\eta = a - ib$. Hence, the spectrum $\sigma(P)$ is symmetric w.r.t. the real axis.*

(d) *If the transition function is reversible, then the spectrum of P acting on $\mathcal{M}_2(\pi)$ is real–valued, i.e., $\sigma(P) \subset [-1, +1]$.*

Item (d) of Proposition 4.16 is due to the following fact about reversible Markov chains that emphasizes their importance.

Theorem 4.17 *Let $T : l^2(\pi) \to l^2(\pi))$ denote some transfer operator corresponding to some reversible Markov chain with stationary distribution π. Then T is self–adjoint w.r.t. to $\langle \cdot, \cdot \rangle_\pi$, i.e.,*

$$\langle Tu, v \rangle_\pi \quad = \quad \langle u, Tv \rangle_\pi$$

for arbitrary $u, v \in l^2(\pi)$. Since $P^* = T$, the same result holds for P on $\mathcal{M}_2(\pi)$.

Below, we will give a much more detailed analysis of the spectrum of P such that it is possible to infer structural properties of the corresponding Markov chain.

In the sequel, we often will assume that the following assumption on the spectrum of P as an operator action on $\mathcal{M}_1(\pi)$ holds.

> **Assumption R**. There exists some constant $R < 1$ such that there are only finitely many $\lambda \in \sigma(P)$ with $|\lambda| > R$, each being an eigenvalue of finite multiplicity[3].

Assumption R is, e.g., a condition on the so–called essential spectral radius of P in $\mathcal{M}_1(\pi)$ [4]; it is also closely related to the so-called *Doeblincondition*. Assumption R is necessary only for the infinite countable state space case, since for the finite state space case, it is trivially fulfilled.

Proposition 4.18 *Given some Markov chain $\{X_k\}_{k \in \mathbb{N}}$ on $\mathbf{S}$ with maximal stationary distribution $\pi > 0$. Let $P : \mathcal{M}_1(\pi) \to \mathcal{M}_1(\pi)$ denote the associated transfer operator. Then, condition R is satisfied, if*

1. *the state space is finite; in this case it is $R = 0$.*

[3]For the general definition of multiplicity see [5, Chap. III.6]. If P is in addition reversible, then the eigenvalue $\lambda = 1$ is of finite multiplicity, if there exist only finitely many mutually linear independent corresponding eigenvectors.

2. *the transition function P fulfills the* ***Doeblin condition****, i.e., there exist $\epsilon, \delta > 0$ and some $m \in \mathbb{N}$ such that for every $y \in \mathbf{S}$*

$$\pi(y) \leq \epsilon \implies P^m(x, y) \leq 1 - \delta.$$

for all $x \in \mathbf{S}$. In this case, it is $R = (1 - \delta)^{1/m}$.

3. *the transfer operator is* ***constrictive****, i.e., there exist $\epsilon, \delta > 0$ and some $m_0 \in \mathbb{N}$ such that for every $\nu \in \mathcal{M}_1(\pi)$*

$$\pi(y) \leq \epsilon \implies \nu P^m(y) \leq 1 - \delta.$$

for all $m \geq m_0$. In this case, it is $R = (1 - \delta)^{1/m_0}$.

Proof: The statements 2. and 3. follow from Thm. 4.13 in [4]. 1. follows from 2. or 3. by choosing $\epsilon < \min_{y \in \mathbf{S}} \pi(y)$, which is positive due to the finiteness of the state space. Now, choose $\delta = 1$ and $m = m_0 = 1$. □

4.4 Evolution of transfer operators

We start by stating the famous Frobenius–Perron theorem for transfer operators related to Markov chains on some *finite* state space (see, e.g., [1, 2, 9]). We then state the result for the infinite state space case. To do so, we define, based on stationary distribution π, the transition function $\Pi = (\Pi(x, y))_{x,y \in \mathbf{S}}$ by

$$\Pi(x, y) \quad = \quad \pi(y)$$

Hence, each row of Π is identical to π, and the Markov chain associated with Π is actually a sequence of i.i.d. random variables, each distributed according to π. We will see that Π is related to the asymptotic behaviour of the powers of the transition matrix P. In matrix notation, it is $\Pi = \mathbf{1}\pi^t$, where $\mathbf{1}$ is the function constant 1.

Theorem 4.19 (Frobenius–Perron theorem) *Let P denote an $n \times n$ transition matrix that is irreducible and aperiodic. Then*

1. *The eigenvalue $\lambda_1 = 1$ is simple and the corresponding left and right eigenvectors can be chosen positive. More precisely, $\pi P = \pi$ for $\pi > 0$ and $P\mathbf{1} = \mathbf{1}$ for $\mathbf{1} = (1, \ldots, 1)$.*

2. *Any other eigenvalue μ of P is strictly smaller (in modulus) than $\lambda_1 = 1$, i.e., $|\mu| < 1$ for any $\mu \in \sigma(P)$ with $\mu \neq 1$.*

3. *Let $\lambda_1, \lambda_2, \ldots, \lambda_r$ with some $r \leq n$* [4] *denote the eigenvalues of P ordered in such a way that*

$$\lambda_1 > |\lambda_2| \geq |\lambda_3 \geq \ldots \geq |\lambda_r|.$$

Let moreover m denote the algebraic multiplicity[5] *of λ_2. Then*

$$P^n \quad = \quad 1\pi^t + \mathcal{O}(n^{m-1}|\lambda_2|^n).$$

Proof: See, e.g., [9]. □

We now state an extended result for the infinite state space case.

Theorem 4.20 *Consider some Markov chain $X = \{X_k\}_{k\in\mathbb{N}}$ with maximal stationary distribution $\pi > 0$ and let $P : \mathcal{M}_1(\pi) \to \mathcal{M}_1(\pi)$ denote the associated transfer operator satisfying Assumption R. Then the following holds:*

1. *The Markov chain X is irreducible, if and only if the eigenvalue $\lambda = 1$ of P is simple, i.e., the multiplicity is equal to 1.*

2. *Assume that the Markov chain is irreducible. Then X is aperiodic, if and only if the eigenvalue $\lambda = 1$ of P is dominant, i.e., for any $\eta \in \sigma(P)$ with $\eta \neq 1$ implies $|\eta| < 1$.*

3. *If the Markov chain is irreducible and aperiodic, then $P^n \to \Pi$ as $n \to \infty$. More precisely, there exists constants $M > 0$ and $r < 1$ such that*

$$||P^n - \Pi||_{TV} \quad \leq \quad Mr^n$$

for $n \geq 1$. Defining $\Lambda_{abs}(P) = \sup\{|\lambda| : \lambda \in \sigma(P), |\lambda| < 1\}$, it is $r \leq \Lambda_{\text{abs}} + \epsilon$ for any $\epsilon > 0$ and $r = \Lambda_{\text{abs}}$ for reversible Markov chains.

Proof: 1.) By Thm. 4.14 of [4], $\lambda = 1$ simple is equivalent to a decomposition of the state space $\mathbf{S} = E \cup F$ with E being invariant ($\pi_E P = \pi_E$ with $\pi_E = \mathbf{1}_E \pi$) and F being of π–measure zero. Since $\pi > 0$ by assumption, F is empty and thus $E = \mathbf{S}$. By contradiction it follows that the Markov chain is irreducible.

2.) By Cor. 4.18 (ii) of [4], $\lambda = 1$ simple and dominant is equivalent to P being aperiodic (which in our case is equivalent to the Markov chain being aperiodic).

[4] If P is reversible than $r = n$ and there exists a complete basis of (orthogonal) eigenvectors.

[5] The algebraic multiplicity of λ_2 is defined as .... If P is reversible than m is equal to the dimension of the eigenspace corresponding to λ_2.

3.) By Cor. 4.22 of [4], the inequality $||P^n - \Pi||_{TV} \leq Mr^n$ is equivalent to P being ergodic and aperiodic (which in our case is equivalent to the Markov chain being irreducible and aperiodic—following from 1.) and 2.)). □

Theorem 4.20 (3.) states that for large n, the Markov chain X_n at time n is approximately distributed like π, and moreover it is approximately independent of its history, in particular of X_{n-1} and X_0. Thus the distribution of X_n for $n \gg 0$ is almost the same, namely π, regardless of whether the Markov chain started at $X_0 = x$ or $X_0 = y$ for some initial states $x, y \in \mathbf{S}$.

We end by relating a certain type of ergodicity condition to the above theorem.

Definition 4.21 *Let $X = \{X_k\}_{k\in\mathbb{N}}$ denote an irreducible Markov chain with transition function P and stationary distribution π. Then, X is called* ***uniformly ergodic****, if for every $x \in \mathbf{S}$*

$$||P^k(x, \cdot) - \pi||_{TV} \leq Cr^k \tag{23}$$

with positive constants $C \in \mathbb{R}$ and $r < 1$.

Theorem 4.22 *Let $\{X_k\}_{k\in\mathbb{N}}$ denote some uniformly ergordic Markov chain. Then, the Markov chain is irreducible, aperiodic and Assumption R is satisfied. Hence, $P^n \to \Pi$ for $n \to \infty$ as in Them. 4.20.*

Proof: Apply Thm. 4.24 of [4] and note that we required the properties to hold for every $x \in \mathbf{S}$ rather than for π almost every $x \in \mathbf{S}$. □

5 Empirical averages

5.1 The strong law of large numbers

Assume that we observe some realization $X_0(\omega), X_1(\omega), \ldots$ of a Markov chain. Is it possible to "reconstruct" the Markov chain by determining its transition probabilities just from the observed data?

In general the answer is 'no'; for example, if the Markov chain is reducible, we would expect to be able to approximate only the transition probabilities corresponding to one communication class. If the Markov chain is transient, the reconstruction attempt will also fail. However, under some reasonable conditions, the answer to our initial question is 'yes'.

In the context of Markov chain theory, a function $f : \mathbf{S} \to \mathbb{R}$ defined on the state space of the chain is called an **observable**. Observables allow to perform "measurements" on the system that is modelled by the Markov chain. Given some Markov chain $\{X_k\}_{k\in\mathbb{N}}$ we define the so–called **empirical average** $S_n(f)$ of the observable f by

$$S_n(f) \quad = \quad \frac{1}{n+1}\sum_{k=0}^{n} f(X_k).$$

Note that the empirical average is a random variable, hence $S_n(f) : \Omega \to \mathbb{R} \cup \{\pm\infty\}$. Under suitable conditions the empirical average converges to a probabilistic average, i.e., the expectation value

$$\mathbb{E}_\pi[f] \quad = \quad \sum_{x\in\mathbf{S}} f(x)\pi(x).$$

Theorem 5.1 (Strong law of large numbers [2, 10]) *Let $\{X_k\}_{k\in\mathbb{N}}$ denote an irreducible Markov chain with stationary distribution π, and let $f : \mathbf{S} \to \mathbb{R}$ be some observable such that*

$$\sum_{x\in\mathbf{S}} |f(x)|\pi(x) \quad < \quad \infty. \tag{24}$$

Then for any initial state $x \in \mathbf{S}$, i.e., $X_0 = x$

$$\frac{1}{n+1}\sum_{k=0}^{n} f(X_k) \longrightarrow \mathbb{E}_\pi[f] \tag{25}$$

as $n \to \infty$ and $\mathbb{P}_x$–almost surely, i.e,

$$\mathbb{P}_x\left[\left\{\omega : \lim_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} f\big(X_k(\omega)\big) = \mathbb{E}_\pi[f]\right\}\right] \quad = \quad 1.$$

Proof: Due to the assumption $\{X_k\}_{k\in\mathbb{N}}$ is irreducible and positive rekurrent, therefore $\nu(y) = \mathbb{E}_x[\sum_{n=0}^{T_x} \mathbf{1}_{\{X_n=y\}}]$ defines an invariant measure, while the stationary distribution is given by $\pi(y) = \frac{\nu(y)}{Z}$, with the normalization constant $Z = \sum_{y\in\mathbf{S}} \nu(y)$ (cp. Theorem 3.19). For the random variable $U_0 = \sum_{k=0}^{T_x} f(X_k)$ the expectation is given by

$$\begin{aligned}\mathbb{E}[U_0] = \mathbb{E}_x\left[\sum_{k=0}^{T_x} f(X_k)\right] &= \mathbb{E}_x\left[\sum_{k=0}^{T_x}\sum_{y\in\mathbf{S}} f(y)\mathbf{1}_{\{X_k=y\}}\right] \\ &= \sum_{y\in\mathbf{S}} f(y)\,\mathbb{E}_x\left[\sum_{k=0}^{T_x}\mathbf{1}_{\{X_k=y\}}\right] = \sum_{y\in\mathbf{S}} f(y)\nu(y)\end{aligned} \tag{26}$$

Now consider $U_p = \sum_{k=\tau_p+1}^{\tau_{p+1}} f(X_k)$, with $p \geq 1$ and $T_x = \tau_0, \tau_1, \tau_2, \ldots$ the successive return times to x. It follows from the strong Markov property (Theorem 3.15) that $U_0, U_1, U_2, \ldots$ are i.i.d. random variables. Since from (24) and (26) we have $\mathbb{E}[|U_0|] < \infty$, therefore the famous Strong Law of Large Numbers for i.i.d. random variables can be applied and yields with probability one, i.e. almost surely,

$$\lim_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} U_k = \sum_{y\in\mathbf{S}} f(y)\nu(y) \Leftrightarrow \lim_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{\tau_{n+1}} f(X_k) = \sum_{y\in\mathbf{S}} f(y)\nu(y).$$

For the moment assume that $f \geq 0$ and define $N_x(n) := \sum_{k=0}^{n} \mathbf{1}_{\{X_k=x\}}$, the number of visits in x within the first n steps. Due to

$$\tau_{N_x(n)} \leq n < \tau_{N_x(n)+1}$$

and $f \geq 0$ it follows that

$$\frac{1}{N_x(n)}\sum_{k=0}^{\tau_{N_x(n)}} f(X_k) \leq \frac{1}{N_x(n)}\sum_{k=0}^{n} f(X_k) \leq \frac{1}{N_x(n)}\sum_{k=0}^{\tau_{N_x(n)+1}} f(X_k). \tag{27}$$

Since the Markov chain is recurrent $\lim_{n\to\infty} N_x(n) = \infty$, so that the extremal terms in (27) converge to $\sum_{y\in\mathbf{S}} f(y)\nu(y)$ and therefore

$$\lim_{n\to\infty} \frac{1}{N_x(n)}\sum_{k=0}^{n} f(X_k) = \sum_{y\in\mathbf{S}} f(y)\nu(y) = Z\sum_{y\in\mathbf{S}} f(y)\pi(y).$$

Now consider the observable $g \equiv 1$, which is positive and fulfills condition (24), since $\{X_k\}_{k\in\mathbb{N}}$ is recurrent. By the equation above we have

$$\lim_{n\to\infty} \frac{1}{N_x(n)}\sum_{k=0}^{n} g(X_k) = \lim_{n\to\infty} \frac{n+1}{N_x(n)} = Z \Rightarrow \lim_{n\to\infty} \frac{N_x(n)}{n+1} = \frac{1}{Z},$$

and finally

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n} f(X_k) &= \lim_{n\to\infty}\frac{1}{N_x(n)}\frac{N_x(n)}{n+1}\sum_{k=0}^{n} f(X_k)\\ &= \frac{1}{Z}\sum_{y\in\mathbf{S}} f(y)\nu(y) = \sum_{y\in\mathbf{S}} f(y)\pi(y).\end{aligned}$$

For arbitrary f, consider $f^+ = \max(0, f)$ and $f^- = \max(0, -f)$ and take the difference between the obtained limits. □

Theorem 5.1 is often referred to as ergodic theorem. It states that the time average (left hand side of (25)) is equal to the space average (right hand side of (25)). The practical relevance of the strong law of large numbers is the following. Assume we want to calculate the expectation $\mathbb{E}_\pi[f]$ of some observable f w.r.t. the stationary distribution of the Markov chain $\{X_k\}_{k\in\mathbb{N}}$. Instead of first computing π and then $\mathbb{E}_\pi[f]$, we can alternatively compute some realization $X_0(\omega), X_1(\omega), \ldots$ and then determine the corresponding empirical average $S_n(f)$. By Theorem 5.1, $S_n(f)$ will be a good approximation to $\mathbb{E}_\pi[f]$ for "large enough" n and almost every realization $\omega \in \Omega$.

Why should we do so? There are many applications, for which the transition matrix of the Markov chain is not given explicitly. Instead, the Markov chain is specified by an algorithm of how to compute a realization of it (this is, e.g., the case, if the Markov chain is specified as a stochastic dynamics system like in eq. (4)) . In such situations, the strong law of large numbers can be extremely useful. Of course, we have to further investigate the approximation quality of the expectation by empirical averages, in particular try to specify how large "large enough" is.

Example 5.2 *Consider as observable $f : \mathbf{S} \to \mathbb{R}$ the indicator function of some subset $A \subset \mathbf{S}$, i.e.,*

$$f(x) = \mathbf{1}\{x \in A\} = \begin{cases} 1; & \text{if } x \in A \\ 0; & \text{otherwise .} \end{cases}$$

Then under the conditions of Theorem 5.1

$$\frac{1}{n+1}\sum_{k=0}^{n} \mathbf{1}\{X_k \in A\} = \frac{1}{n+1}\sum_{k=0}^{n} \mathbf{1}_A(X_k) \longrightarrow \pi(A)$$

as $n \to \infty$. Hence, $\pi(A)$ can be interpreted as the long time average number of visits to the subset A. Consequently for large enough n, $\pi(A)$ approximately denotes the probability of encountering the Markov chain after n steps in subset A.

To answer the initial question whether we can reconstruct the transition probabilities from a realization, we state the following

Corollary 5.3 (Strong law of large numbers II [2]) *Let $\{X_k\}_{k\in\mathbb{N}}$ denote an irreducible Markov chain with transition matrix $P = \big(P(x,y)\big)_{x,y\in\mathbf{S}}$ and stationary distribution π, and let $g : \mathbf{S}\times\mathbf{S} \to \mathbb{R}$ be some function such that*

$$\sum_{x,y\in\mathbf{S}} |g(x,y)|\pi(x)P(x,y) \quad < \quad \infty.$$

Then for any initial state $x \in \mathbf{S}$, i.e., $X_0 = x$ we have

$$\frac{1}{n+1}\sum_{k=0}^{n} g(X_k, X_{k+1}) \longrightarrow \mathbb{E}_{\pi,P}[g] \;=\; \sum_{x,y\in\mathbf{S}} g(x,y)\pi(x)P(x,y)$$

as $n \to \infty$ and $\mathbb{P}_x$–almost surely.

Proof: We leave this as an excercise. Prove that $\pi(x)P(x,y)$ is a stationary distribution of the bivariate Markov chain $Y_k = (X_k, X_{k+1})$. □

Corollary 5.3 is quite useful for our purpose. Consider the function $g : \mathbf{S}\times\mathbf{S} \to \mathbb{R}$ with

$$g(x,y) \quad = \quad \mathbf{1}_{(u,v)}(x,y) \;=\; \begin{cases} 1; \text{ if } x = u, y = v \\ 0; \text{ otherwise .} \end{cases}$$

Under the condition of Corollary 5.3

$$\frac{1}{n+1}\sum_{k=0}^{n} \mathbf{1}\{X_k = u, X_{k+1} = v\} \longrightarrow \pi(u)P(u,v)$$

as $n \to \infty$. Hence, if we first compute $\pi(u)$ as outlined in Example 5.2 with $A = \{u\}$, we can then approximate the transition probability $P(u,v)$ by computing the average number of "transitions $X_k = u, X_{k+1} = v$" with $0 \le k < n$ and divide it by $n\pi(u)$.

Acknowledgement Supported by the DFG Research Center MATHEON "Mathematics for key technologies".

References

[1] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979. Reprinted by SIAM, Philadelphia, 1994.

[2] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer, New York, 1999.

[3] J. Chang. Stochastic processes. Online available material for the course "Stochastic Processes", http://www.soe.ucsc.edu/classes/engr203/Spring99/, 1999.

[4] W. Huisinga. *Metastability of Markovian systems: A transfer operator approach in application to molecular dynamics*. PhD thesis, Free University Berlin, 2001.

[5] T. Kato. *Perturbation Theory for Linear Operators*. Springer, Berlin, 1995. Reprint of the 1980 edition.

[6] T. Lindvall. *Lectures on the Coupling Method*. Wiley-Interscience, 1992.

[7] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer, Berlin, 1993.

[8] J. W. Pittman. On coupling of Markov chains. *Z. Warsch. verw. Gebiete*, 35:315–322, 1976.

[9] E. Seneta. *Non–negative Matrices and Markov Chains*. Series in Statistics. Springer, second edition, 1981.

[10] L. Tierney. Introduction to general state-space Markov chain theory. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov chain Monte-Carlo in Practice*, pages 59–74. Chapman and Hall, London, 1997.

[11] D. Werner. *Funktionalanalysis*. Springer, Berlin, 2nd edition, 1997.