Markov Processes on Discrete State Spaces

Theoretical Background and Applications.

Christof Schuette[1] & Wilhelm Huisinga[2]

[1] Fachbereich Mathematik und Informatik
Freie Universität Berlin & DFG Research Center Matheon,
Berlin, Germany
schuette@mi.fu-berlin.de

[2] Hamilton Institute
Dublin, Ireland
wilhelm.huisinga@nuim.ie

Berlin, June 8, 2011

Contents

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.

This manuscript is partially based on contributions of Eike Meerbach and Philipp Metzner (both FU Berlin). Without their work the present version would not exist. Thanks!

Wilhelm Huisinga and Christof Schuette

2 Setting the scene

2.1 Introductory example

We will start with an example that illustrate some features of Markov chains: Imagine a surfer on the world-wide-web (WWW). At an arbitrary instance in time he is looking at some webpage out of the WWW universe of N different webpages. Let us ignore the question of whether a specific webpage has to be counted as one of the individual webpages or may just be a subpage of another one. Let us simply assume that we have a listing $1, \ldots, N$ of all of them (like GOOGLE, for example, is said to have), and furthermore that we know which webpage includes links to which other webpage (GOOGLE is said to have this also). That is, for each webpage $w \in \{1, \ldots, N\}$ we do have a list $L_w \subset \{1, \ldots, N\}$ of webpages to which w includes links.

To make things simple enough for an introduction, let us assume that we are *not* interested in what an individual surfer will really do but imagine that we are interested in producing a webbrowser and that our problem is to find a good but not too complicated *model* of how often the *average surfer* will be looking at a certain webpage. We plan to use the probability p_w of a visit of the *average surfer* on webpage $w \in \{1, \ldots, N\}$ for ranking all the webpages: the higher the probability, the higher the ranking.

To make live even easier let us equip our average surfer with some simple properties:

(P1) He looks at the present website for exactly the same time, say $\Delta t = 1$, before surfing to the next one. So, we can write down the list of webpage he visits one after the other: $w_0, w_1, w_2, \ldots$, where $w_0 \in \{1, \ldots, N\}$ is his start page, and w_j is the page he is looking at at time j.

(P2) When moving from webpage w to the next webpage he just chooses from one of the webpages in L_w with equal probability. His choice is independent of the instance at which he visits a certain webpage.

Thus, the sequence of webpages our surfer is visiting can be completely described as a sequence of random variable w_j with joint state space $\{1, \ldots, N\}$, and *transition probabilities*

$$\mathbb{P}[w_{j+1} = w' | w_j = w] = \begin{cases} 1/|L_w| & \text{if } w' \in L_w, \text{and } L_w \neq \emptyset \\ 0 & \text{otherwise} \end{cases},$$

where $|L_w|$ denotes the number of elements in L_w. We immediately see that we get trapped whenever there is a webpage w so that $L_w = \emptyset$ which may well happen in the real WWW. Therefore, let us adapt property (P2) of our average surfer a little bit:

(P2a) When moving from webpage w to the next webpage the average surfer behaves as follows: With some small probability α he jumps to an arbitrary webpage in the WWW (random move). With probability $(1-\alpha)$ he just chooses from one of the webpages in L_w with equal probability; if $L_w = \emptyset$ he then stays at the present webpage. His choice is independent of the instance at which he visits a certain webpage.

Under (P1) and (P2a) our sequence of random variable w_j are connected through *transition probabilities*

$$\mathbb{P}[w_{j+1} = w' | w_j = w] = \begin{cases} \frac{\alpha}{N} + \frac{1-\alpha}{|L_w|} & \text{if } w' \in L_w, \text{ and } L_w \neq \emptyset \\ \frac{\alpha}{N} & \text{if } w' \notin L_w, \text{ and } L_w \neq \emptyset \\ \frac{\alpha}{N} + 1 - \alpha & \text{if } w' = w, \text{ and } L_w = \emptyset \\ \frac{\alpha}{N} & \text{if } w' \neq w, \text{ and } L_w = \emptyset \end{cases} . \quad (1)$$

Now our average surfer cannot get trapped and will move through the WWW till eternity.

So what now is the probability p_w of a visit of the average surfer on webpage $w \in \{1, \ldots, N\}$. The obvious mathematical answer is an asymptotic answer: p_w is the $T \to \infty$ limit of the probability of visits on webpage w in a long trajectory (length T) of the random surfer. This leaves many obvious questions open. For example: How can we compute p_w? And, if there is an explicit equation for this probability, can we compute it efficiently for very large N?

We will give answers to these questions (see Remarks on pages 9, 24), 36, and 57)! But before we are able to do this, we will have to digest some of the theory of Markov Chains: the sequence $w_0, w_1, w_2, \ldots$ of random variables described above form a (discrete-time) Markov chain. They have the characteristic property that is sometimes stated as "The future depends on the past only through the present": The next move of the average surfer depends just on the present webpage and on nothing else. This 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. But whatever we may consider: we always have to be aware that Markov chains, as in our introductory example, are often simplistic *mathematical models* of the real-world process they try to describe.

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 $X_0 = y$:

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

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***

$$\mathbb{P}[X_{k+1} = z | X_k = y, X_{k-1} = x_{k-1}, \ldots, X_0 = x_0] = \mathbb{P}[X_{k+1} = z | X_k = y] \quad (2)$$

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 (2) are defined[1] *and, moreover, the right hand side of (2) does not depend on k, hence*

$$\mathbb{P}[X_{k+1} = z | X_k = y] \ = \ \ldots \ = \ \mathbb{P}[X_1 = z | X_0 = y]. \quad (3)$$

[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.

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

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

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*

$$p_{xy} \geq 0, \ \textit{and} \ \sum_{y \in \mathbf{S}} p_{xy} = 1 \tag{4}$$

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

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

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

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

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

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 [12, Thm. 3.2.1] will now make this more precise.

Remark. In our introductory example (random surfer on the WWW), we can easily check that the matrix $P = (p_{w,w'})_{w,w'=1,\ldots,N}$ with

$$p_{w,w'} = \mathbb{P}[w_{j+1} = w' | w_j = w]$$

according to (1) is a stochastic matrix in which all entries are positive.

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. (4). 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. 1.*

Figure 1: 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. 2.

Figure 2: 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. 3. 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 (4). Then,* P *defines a Markov chain on the state space* $\mathbf{S}$.

Figure 3: 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*

$$X_{k+1} = f(X_k, \xi_k) \tag{5}$$

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 (5) 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

$$f(x,u) = z \quad \text{for} \quad \sum_{y=1}^{z-1} P(x,y) \leq u < \sum_{y=1}^{z} P(x,y). \tag{6}$$

Note that *every* homogeneous Markov chain has a representation (5) with the function f defined in (6).

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 (6).

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

$$||\mu||_{TV} = \sum_{x \in \mathbf{S}} |\mu(x)|. \tag{7}$$

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

Figure 4: 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.

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**

$$P^k(x, y) \quad = \quad \mathbb{P}[X_k = y | X_0 = x], \tag{8}$$

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

$$\mu_k \quad = \quad \mu_{k-1}\, P^1 \;=\; \mu_{k-2}\, P^2 \;=\; \ldots \;=\; \mu_0\, P^k. \tag{9}$$

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

$$P^{m+k}(x, z) \quad = \quad \sum_{y\in\mathbf{S}} P^m(x, y) P^k(y, z) \tag{10}$$

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 4. 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.

Definition 2.6 *A probability distribution π satisfying*

$$\mathbb{P}_\pi[X_1 = y] \quad = \quad \pi(y) \tag{11}$$

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

$$\pi = \pi P \tag{12}$$

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 4). Moreover, we analytically compute the stationary distribution of the two state Markov chain. Here, $\pi = (\pi(1), \pi(2))$ has to satisfy

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

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. (11) 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

$$\begin{aligned} \mathrm{flux}_\pi(A,B) &= \mathbb{P}_\pi[X_1 \in B, X_0 \in A] \\ &= \sum_{x\in A} \pi(x) P(x,B) \;=\; \sum_{x\in A} \sum_{y\in B} \pi(x) P(x,y). \end{aligned} \tag{13}$$

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*

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

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 (11) 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

$$\sum_{k=0}^{\infty} \left(\frac{p}{1-p}\right)^k = \begin{cases} \infty; \text{ if } p \geq 1/2 \\ (1-p)/(1-2p); \text{ if } p < 1/2. \end{cases} \tag{14}$$

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. (14) 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

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

Figure 5: Typical realization of a Markov chain in state space $\mathbf{S} = \{1, 2, 3, 4\}$ as considered in example 5.7 below (upper panel) and resulting running average $S_n(f)$ (lower panel).

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.

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

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 the communication relation is a so–called *partial equivalence relation*, hence it is

1. symmetric: $x \leftrightarrow y$ implies $y \leftrightarrow x$,

2. transitive: $x \leftrightarrow y$ and $y \leftrightarrow z$ imply $x \leftrightarrow z$.

In general, it is *not* reflexive ($x \leftrightarrow x$). Let us study the so–called equivalence classes defined as

$$[x] \ := \ \{y \in \mathbf{S} : y \leftrightarrow x\}.$$

It is easy to justify that we have

$$\begin{aligned} x, y \in \mathbf{S},\ x \not\leftrightarrow y \qquad &\Leftarrow \qquad [x] \cap [y] = \emptyset \\ x \not\leftrightarrow x \qquad &\Leftrightarrow \qquad [x] = \emptyset \end{aligned}$$

That is, if we collect all states that do not belong to communication classes,

$$D = \{x \in \mathbf{S} : \ [x] = \emptyset\}, \tag{15}$$

then the communication relation is a full equivalence relation on $D^c = \mathbf{S} \setminus D$. Recall that every equivalence relation induces a partition $D^c = C_0 \cup \cdots \cup C_{r-1}$ of the remaining state space D^c 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 $x_0, \ldots, x_{r-1} \in D^c$. 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 the excluded set D and possibly all remaining communication classes. Then

$$\mathbf{S} \quad = \quad \Big(C_0 \cup \ldots \cup C_{r-1} \Big) \cup D^*. \tag{16}$$

Example 3.3 *Consider the three-state Markov chain with transition matrix*

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

We find $D = \{1\}$ and that $D^c = \{2, 3\}$ is one closed communication class.

The following proposition states that we may restrict the Markov chain to its closed communication classes that hence can be analyzed separately [12, Prop. 4.1.2].

Proposition 3.4 *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 [12, p.84], for reducible Markov chains we can analyze at least the closed subsets in the decomposition (16) 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.

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

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" [12, 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:

Theorem 3.5 (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 3.5 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.

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

Definition 3.6 ([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].

Theorem 3.7 *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 (10) 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 $P^k(x, y)P^m(y, x) > 0$ and thus $P^{k+j+m}(x, x) > 0$. To handle $j > 0$ we first define $D_y = \{j \in \mathbb{N} : \ P^j(y, y) > 0\}$ such that $d(y) = \gcd D_y$. Choosing an arbitrary $j \in D_y$ yields $P^{k+j+m}(x, x) > 0$ and thus $d(x)|(k+j+m)$ in addition to $d(x)|(k+m)$. Therefore we have $d(x)|j$. Since this holds for all $j \in D_y$ we get $d(x)|d(y) = \gcd D_y$. 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 3.5). Combining Definition 3.6 with Theorem 3.7, we get the following useful criterion for aperiodicity:

Corollary 3.8 *An irreducible Markov chain $\{X_k\}_{k\in\mathbb{N}}$ is aperiodic, if there exists some state $x \in \mathbf{S}$ such that $\mathbb{P}[X_1 = x|X_0 = x] > 0$.*

Remark. In our introductory example (random surfer on the WWW), we can easily check that the matrix $P = (p_{w,w'})_{w,w'=1,\ldots,N}$ with

$$p_{w,w'} = \mathbb{P}[w_{j+1} = w'|w_j = w]$$

according to (1) is a stochastic matrix in which all entries are positive. Thus, the associated chain is irreducible and aperiodic.

3.1.1 Additional Material: Proof of Theorem 3.5

We now start to prove Theorem 3.5. The proof will be a simple consequence of the following three propositions.

Proposition 3.9 *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 (10), 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 [16, 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 3.9 can be used to define an equivalence relation on $\mathbf{S}$, which gives rise to the cyclic classes in Theorem 3.5: 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 3.5.

Proposition 3.10 *Let P denote the transition function of an irreducible Markov chain with period d and define $E_0, \dots, E_{d-1}$ 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 3.11 *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. 3.9 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 3.10, 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 3.12 *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 3.13 *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 3.14 *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 3.15 *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 3.16 (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 $X_0, \ldots, X_T$.*

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

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

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

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

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

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

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

Theorem 3.16 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

$$\mathbb{P}_x[N_y = m] = \mathbb{P}_x[T_y < \infty] \, \mathbb{P}_y[T_y < \infty]^{m-1} \, \mathbb{P}_y[T_y = \infty] \tag{17}$$

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

Theorem 3.17 *Consider some state $x \in \mathbf{S}$. Then*

$$x \text{ is recurrent} \Leftrightarrow \mathbb{P}_x[N_x = \infty] = 1 \Leftrightarrow \mathbb{E}_x[N_x] = \infty, \tag{18}$$

and

$$x \text{ is transient} \Leftrightarrow \mathbb{P}_x[N_x = \infty] = 0 \Leftrightarrow \mathbb{E}_x[N_x] < \infty. \tag{19}$$

The above equivalence in general fails to hold for the denumerable, more general (continuous) state space case—here, one has to introduce the notion of Harris recurrent [12, Chapt. 9].

Proof: Let us first show the first $\Rightarrow$ in (18): If x is recurrent then $\mathbb{P}_x[T_x < \infty] = 1$. Hence, due to eq. (17)

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

vanishes, since every summand is zero. Consequently, $\mathbb{P}_x[N_x = \infty] = 1$.

Next, let us address the first $\Rightarrow$ in (19): If x is transient, then $\mathbb{P}_x[T_x < \infty] < 1$ and hence

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

which immediately implies $\mathbb{P}_x[N_x = \infty] = 0$.

Furthermore, we can easily see the second $\Rightarrow$ in (19): If x is transient, then $\mathbb{P}_x[T_x < \infty] < 1$ and hence

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

The remaining assertions follow by negation. For example, whenever $\mathbb{E}_x[N_x] = \infty$ holds and we assume $\mathbb{P}_x[T_x < \infty] < 1$ in addition then this implies a contradiction since the last identity yields $\mathbb{E}_x[N_x] < \infty$ then. Thus, $\mathbb{E}_x[N_x] = \infty$ implies $\mathbb{P}_x[T_x < \infty] = 1$, that is the second $\Leftarrow$ in (18). □

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 3.18 *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. 3.17.

2. Proof left as an excercise (Hint: Use Proposition 3.18 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 3.19 *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. 3.17. □

As a consequence of Theorem 3.19, 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 3.19 and Prop. 3.18 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 3.20 *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*

$$\mu(y) = \mathbb{E}_x \left[\sum_{n=1}^{T_x} 1_{\{X_n = y\}} \right], \tag{20}$$

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. (20).*

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. (20) 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 (20) 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. (20) 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. (20) 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 $X_n=y$, $n\le T_x$ and $x\neq y$ implies $n+1\le T_x$. 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

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

On the other hand,

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

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

$$\nu(z) \geq \sum_{n=1}^{k} \mathbb{P}_x[X_n = z, n \leq T_x] \tag{21}$$

holds for every $z \in \mathbf{S}$. Note that the right hand side of eq. (21) 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. (21) 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. (21) 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 3.20 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

$$\mathbb{E}_x[T_x] = \sum_{n=1}^{\infty} n \, \mathbb{P}_x[T_x = n]. \tag{22}$$

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

Definition 3.21 *A recurrent state $x \in \mathbf{S}$ is called **positive recurrent**, if*

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

*and **null recurrent** otherwise.*

In view of eq. (22) 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 3.22 *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*

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

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

Proof: Theorem 3.20 states that an irreducible and recurrent Markov chain admits an invariant measure μ defined through (20) 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 3.20. On the other hand, if there exists a stationary distribution the Markov process must be recurrent because otherwise $\pi(x)$ would be zero for all $x \in \mathbf{S}$ according to Theorem 3.18. Positive recurrence follows from the uniqueness of π and the consideration above. □

Our considerations in the proof of Theorem 3.22 easily leads to a criteria to distinguish positive recurrence from null recurrence.

Corollary 3.23 *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 and important statement.

Theorem 3.24 (Irreducibility + Finite State Space ⇒ Recurrence) *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 3.25 *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. □

Remark. In our introductory example (random surfer on the WWW) we considered the transition matrix $P = (p_{w,w'})_{w,w'=1,\ldots,N}$ with

$$p_{w,w'} = \mathbb{P}[w_{j+1} = w' | w_j = w]$$

Figure 8: Different recurrent behavior of irreducible, aperiodic Markov chains.

according to (1). The associated chain is irreducible (one closed communication class), aperiodic and positive recurrent. Its stationary distribution π thus is unique, positive everywhere and given by either $\pi P = \pi$ or $\pi(w) = 1/\,\mathbb{E}_w[T_w]$. The value $\pi(w)$ is a good candidate for the probability p_w of a visit of the average surfer on webpage $w \in \{1, \dots, N\}$. That this is in fact true will be shown in the next chapter.

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. 3.18 □

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 3.24):

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., [9, 14]).

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 $A \subset \mathbf{S}$

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

because $\mathbb{P}[X \in A, X = Y] = \mathbb{P}[Y \in A, X = Y]$. Since

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

the statement directly follows. □

Remark 4.5 *Above (page 13) we used the equation $\|\mu\|_{TV} = \sum_{x \in \S} |\mu(x)|$ for the total variation norm of some function $\mu \in \mathcal{M} = \{\mu : \mathbf{S} \to \mathbb{C} : ||\mu||_{TV} < \infty\}$. Here we used the total variation distance between two probability distributions $\|\nu - \pi\|_{TV} = 2 \sup_{A \subset \mathbf{S}} |\nu(A) - \pi(A)|$. Both equations are special cases of the general definition of the total variation of some complex-valued measure $\mu : \mathcal{A} \to \mathbb{C}$, where $\mathcal{A}$ denotes the σ-algebra of all measurable subsets of $\mathbf{S}$: $\|\mu\|_{TV} = |\mu|(\mathbf{S}) = \sup_D \sum_{D_i} |\mu(D_i)|$, where the sum is taken over all possible finite partitions $D = (D_1, \ldots, D_m) \in \mathcal{A}^m$ of $\mathbf{S}$, $\mathbf{S} = \cup_{i=1}^m D_i$.*

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 4.6 *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 9: The construction of the coupled process X' as needed in the proof of Prop. 4.6. 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. 9) 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 3.16 (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 4.6 enables us to prove the convergence of ν_k to π by proving that $\mathbb{P}[k < T_c]$ converges to zero.

Theorem 4.7 *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 an integer $k^* > 0$ such that $P^{k^*}(v, w) > 0$ *and* $P^{k^*}(x, y) > 0$ holds, see Prop. 3.9. 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. 3.22. By Thm. 3.19, this implies $\mathbb{P}[T_c^* < \infty] = 1$, which completes the proof of the statement. □

Fig. 10 summarizes the long run behavior of irreducible and aperiodic Markov chains.

Figure 10: Long run behaviour of an irreducible aperiodic Markov chain.

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

$$\mathbb{P}_\pi[Y_0 = y, Y_1 = x] \quad = \quad \mathbb{P}_\pi[X_0 = x, X_1 = y] \tag{23}$$

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 (23) we obtain

$$\pi(y)Q(y,x) \quad = \quad \pi(x)P(x,y). \tag{24}$$

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. (24) we deduce by summing over all $x \in \mathbf{S}$ that π is some stationary distribution of the Markov chain.

Definition 4.8 *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*

$$Q(y,x) = \frac{\pi(x)P(x,y)}{\pi(y)} \tag{25}$$

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

Example 4.9 *Consider the two state Markov chain given by*

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

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

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

for $a, b, c \in [0,1]$. Denote by π the stationary distribution (which exists due to Theorem 3.24). Then, $\pi = \pi P$ is equivalent to $a\pi(1) = b\pi(2) = c\pi(3)$. A short calculation reveals

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

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 4.11 *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**

$$\pi(x)P(x,y) \;=\; \pi(y)P(y,x) \tag{26}$$

is satisfied for every $x, y \in \mathbf{S}$. Eq. (26) has a nice interpretation in terms of the probability flux defined in (13). Recall that the flux from x to y is defined by $\text{flux}_\pi(x, y) = \pi(x)P(x, y)$. Thus, eq. (26) 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 4.12 *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 4.12 yields that the three state Markov chain defined in Example 4.10 cannot be reversible.

Example 4.13 *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 4.12. 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.

Last but not least let us observe that the detailed balance condition can also be used *without* assuming π to be a stationary distribution:

Theorem 4.14 *Consider some Markov chain $X = \{X_k\}_{k\in\mathbb{N}}$ with transition function P and a probability distribution $\pi > 0$. Let the detailed balance condition be formally satisfied:*

$$\pi(x)P(x,y) \quad = \quad \pi(y)P(y,x).$$

Then, π is a stationary distribution of X, i.e., $\pi P = \pi$.

Proof: The statement is an easy consequence of the detailed balance condition and $\sum_y P(x,y) = 1$ for all $y \in \mathbf{S}$:

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

□

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

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

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., [21])). 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 4.15 *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 4.16 *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 3.18.

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. (24)), 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**

$$\nu P \quad = \quad \lambda \nu. \tag{27}$$

The function ν is called an (left) **eigenvector** corresponding to the eigenvalue λ. Note that not every function ν satisfying (27) 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., [21, 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

$$Tu(x) = \mathbb{E}_x[u(X_1)] = \sum_{y \in \mathbf{S}} P(x,y)u(y) \tag{28}$$

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., [21, 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 4.17 *Consider some transfer operator P acting on $\mathcal{M}_1(\pi)$. Then $\pi P = \pi$ (since $\mathbf{1}$ is in $l^1(\pi)$) 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 4.18 *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 4.18 is due to the following fact about reversible Markov chains that emphasizes their importance.

Theorem 4.19 *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 $\mathcal{M}_1(\pi)$ [7]; 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 4.20 *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 [8, 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 [7]. 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, 16]). 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 4.21 (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., [16]. □

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

Theorem 4.22 *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 [7], $\lambda = 1$ 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 [7], $\lambda = 1$ 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 [7], 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 4.22 (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 4.23 *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 $x \in \mathbf{S}$*

$$||P^k(x, \cdot) - \pi||_{TV} \leq Cr^k \tag{29}$$

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

Theorem 4.24 *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. 4.22.*

Proof: Apply Thm. 4.24 of [7] 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, 18]) *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*

$$\sum_{x\in\mathbf{S}} |f(x)|\pi(x) \quad < \quad \infty. \tag{30}$$

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

$$\frac{1}{n+1}\sum_{k=0}^{n} f(X_k) \longrightarrow \mathbb{E}_\pi[f] \tag{31}$$

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 3.20). For the random variable $U_0 = \sum_{k=0}^{T_x} f(X_k)$ the expectation is given by

$$\begin{aligned}\mathbb{E}[U_0] = \mathbb{E}_x\left[\sum_{k=0}^{T_x} f(X_k)\right] &= \mathbb{E}_x\left[\sum_{k=0}^{T_x}\sum_{y\in\mathbf{S}} f(y)\mathbf{1}_{\{X_k=y\}}\right] \\ &= \sum_{y\in\mathbf{S}} f(y)\,\mathbb{E}_x\left[\sum_{k=0}^{T_x}\mathbf{1}_{\{X_k=y\}}\right] = \sum_{y\in\mathbf{S}} f(y)\nu(y)\end{aligned} \tag{32}$$

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 3.16) that $U_0, U_1, U_2, \ldots$ are i.i.d. random variables. Since from (30) and (32) 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

$$\frac{1}{N_x(n)}\sum_{k=0}^{\tau_{N_x(n)}} f(X_k) \leq \frac{1}{N_x(n)}\sum_{k=0}^{n} f(X_k) \leq \frac{1}{N_x(n)}\sum_{k=0}^{\tau_{N_x(n)+1}} f(X_k). \tag{33}$$

Since the Markov chain is recurrent $\lim_{n\to\infty} N_x(n) = \infty$, so that the extremal terms in (33) 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 (30), 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 (31)) is equal to the space average (right hand side of (31)). 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. (5)) . 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.

Remark. In our introductory example (random surfer on the WWW) we considered the transition matrix $P = (p_{w,w'})_{w,w'=1,\dots,N}$ with

$$p_{w,w'} = \mathbb{P}[w_{j+1} = w' | w_j = w]$$

according to (1). The associated chain is irreducible, and aperiodic on finite state space. Its stationary distribution π thus is unique, positive everywhere and given by either $\pi P = \pi$. The probability p_w of a visit of the average surfer on webpage $w \in \{1, \dots, N\}$ is the relative frequency of visits in w in some infinitely long realization of the chain. It is given by its stationary distribution

$$p_w = \lim_{n\to\infty} \frac{1}{n+1} \sum_{j=0}^{n} \mathbf{1}\{w_j = w\} = \pi(w).$$

Thus the ranking of webpage can be computed by solving the eigenvalue problem $\pi P = \pi$, which is very large since the number N of webpages in the WWW is extemely large (N is in the range of 10^{10}). Thus direct numerical solution of the eigenvalue problem is prohibitive, and only iterative solvers can lead to success. Google's famous page rank algorithm is an iterative scheme for computing the solution π iteratively via the application of the power method to $\pi P = \pi$.

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 $n\pi(u)$.

Remark 5.4 *By putting together the two above forms of the strong law of large numbers we get for $n \to \infty$:*

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

since $\sum_{k=0}^{n}\mathbf{1}\{X_k = u\} \to \pi(u)$. This result tells us that for estimating the transition matrix of a Markov chain from an observation of its realization we just need to count:

$$\frac{1}{n+1}\sum_{k=0}^{n}\frac{\mathbf{1}\{X_k = u, X_{k+1} = v\}}{\mathbf{1}\{X_k = u\}} = \frac{\textit{no. of transitions } u \to v}{\textit{no. of visits to } u}.$$

Remark 5.5 (Periodic versus aperiodic chains) *Now we are in a position to compare our previous insights into period and aperiodic irreducible Markov chains. For the sake of simplicity we will use a finite state space $\mathbf{S} = \{1,2\}$:*

	irreducible + periodic	***irreducible + aperiodic***
trans. matrix	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$ $a, b \in (0,1)$
stationary distribution	$\pi = (1/2, 1/2)$	$\pi = (\frac{b}{a+b}, \frac{b}{b+a})$
expected first return	$\mathbb{E}_x(T_x) = 2 = 1/\pi(x)$ $T_x = 2$	$\mathbb{E}_x(T_x) = 1/\pi(x)$
evolution of distributions	*no convergence*	$\|P^k(x,\cdot) - \pi\|_{TV} \to 0$ $k \to \infty$
law of large numbers	$\frac{1}{n+1}\sum_{k=0}^{n} f(X_k) \to \mathbb{E}_\pi[f]$ $N \to \infty$	$\frac{1}{n+1}\sum_{k=0}^{n} f(X_k) \to \mathbb{E}_\pi[f]$ $N \to \infty$

5.2 Central limit theorem

We are now ready to state the important central limit theorem:

Theorem 5.6 (central limit theorem) *Let $\{X_k\}_{k\in\mathbb{N}}$ denote an irreducible reversible Markov chain with stationary distribution π. Let moreover $f : \mathbf{S} \to \mathbb{R}$ denote some observable satisfying*

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

Then, for every initial state $X_0 = x \in \mathbf{S}$:

(i) The variance of the empirical averages satisfies

$$n \operatorname{Var}\left[\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)\right] \quad \longrightarrow \quad \sigma^2(f)$$

*as $n \to \infty$, where $\sigma^2(f)$ is called the **asymptotic variance** w.r.t. f.*

(ii) If $\sigma(f) < \infty$, then the distribution of the empirical averages satisfies

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \mathbb{E}_\pi[f]\right) \longrightarrow \mathcal{N}\left(0, \sigma^2(f)\right) \tag{34}$$

for $n \to \infty$, where the convergence is understood in distribution, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left[\left\{\omega : \sqrt{n}\left(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k(\omega)) - \mathbb{E}_\pi[f]\right) \le z\right\}\right] \quad = \quad \Phi_{\sigma(f)}(z)$$

for every $z \in \mathbb{R}$. Here $\Phi : \mathbb{R} \to \mathbb{R}$ given by

$$\Phi_{\sigma(f)}(z) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2(f)}} \int_{-\infty}^{z} \exp\left(-\frac{y^2}{2\sigma^2(f)}\right) \mathrm{d}y$$

denotes the distribution function of the normal distribution with mean zero and variance $\sigma^2(f)$.

Remarks. For P acting on $\mathcal{M}_2(\pi)$ define the quantity $\Lambda_{\max}(P) = \sup\{\lambda \in \sigma(P) : \lambda \neq 1\} \le 1$.

1. The convergence process related to the asymptotic variance can be characterized more precisely by

$$\operatorname{Var}\left[\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)\right] = \frac{\sigma^2(f)}{n} + \mathcal{O}\left(\frac{1}{n^2}\right).$$

2. Whenever the observable f decays quickly enough, i.e., for all $f \in l^{2+\epsilon}(\pi)$ with some $\epsilon > 0$, we have an explicit formula for the asymptotic variance:
$$\sigma^2(f) = \langle f', f' \rangle_\pi + 2 \sum_{k=1}^{\infty} \langle f', T^k f' \rangle_\pi, \tag{35}$$
where $f' = f - \mathbb{E}_\pi(f)$ and T is the transfer operator on $l^2(\pi)$, the adjoint of P, as introduced in (28).

3. As a consequence of the last statement we have the upper bound on the asymptotic variance in terms of $\Lambda_{\max}(P)$. If $\Lambda_{\max}(P) < 1$, then
$$\sigma^2(f) \quad \leq \quad \frac{1 + \Lambda_{\max}(P)}{1 - \Lambda_{\max}(P)} \langle f, f \rangle_\pi < \infty.$$

4. For finite state space case we always have $\Lambda_{\max}(P) < 1$.

5. The asymptotic variance in the central limit theorem is related to the convergence rates of the empirical averages $S_n(f)$. The smaller the asymptotic variance $\sigma^2(f)$, the better the convergence of $S_n(f)$ to its limit values $\mathbb{E}_\pi[f]$.

6. Equation (34) is often interpreted in the following way to quantify convergence for a given observable f. For large n we approximately have
$$\left(\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - \mathbb{E}_\pi[f] \right) \quad \approx \quad \frac{1}{\sqrt{n}} \, \mathcal{N}\big(0, \sigma^2(f)\big)$$
in a distributional sense. Hence if we quadruple the length of the Markov chain realization, we gain only the double precision.

7. Note that the asymptotic variance depends on the observable f. Hence, for one and the same Markov chain the asymptotic variance may be small for one observable, but large for another.

Example 5.7 *Let us consider the state space* $\mathbf{S} = \{1, 2, 3, 4\}$ *and the transition matrix*
$$P \quad = \quad \begin{pmatrix} 0.7 & 0.3 & 0 & 0 \\ 0.5 & 0.3 & 0.2 & 0 \\ 0 & 0.1 & 0.85 & 0.05 \\ 0 & 0 & 0.1 & 0.9 \end{pmatrix}.$$
The associated Markov chain $\{X_k\}_{k \in \mathbb{N}}$ *is irreducible, reversible, and aperiodic and has stationary distribution*
$$\pi \quad = \quad \begin{pmatrix} 0.2941 & 0.1765 & 0.3529 & 0.1765 \end{pmatrix}.$$

Figure 11: Typical realization of the chain considered in example 5.7 (upper panel) and resulting running average $S_n(f)$ (lower panel).

Figure 12: Variance $\mathrm{Var}(S_n(f))$ of the running averages of $L = 5000$ independent samples of the chain considered in example 5.7 (upper panel) and resulting $n\mathrm{Var}(S_n(f))$ (lower panel).

The eigenvalues of P are

$$\lambda \;=\; \begin{pmatrix} 0.0457 & 0.7885 & 0.9158 & 1 \end{pmatrix},$$

such that we have $\Lambda_{\max}(P) = 0.9158$. Let us consider the observable $f(x) = x$ such that $E_\pi(f) = 2.4118$. A typical realization of $\{X_k\}_{k\in\mathbb{N}}$ is shown in Fig. 11 (upper panel). The lower panel of Fig. 11 shows the resulting dependence of $S_n(f) = \sum_{k=0}^{n} f(X_k)/(n+1)$ on n. Next, we consider $L = 5000$ independent realization of $\{X_k\}_{k\in\mathbb{N}}$ such that we we get L independent samples of the random variable $S_n(f)$. The variance $\mathrm{Var}(S_n(f))$ computed from this samples is shown in Fig. 12 (left panel), while $n\mathrm{Var}(S_n(f))$ is shown in the right hand panel. We observe that $n\mathrm{Var}(S_n(f))$ shows some kind of convergence to a constant value of about 25; the exact value computed according to (35) is $\sigma^2(f) = 25.117$. This is what the central limit theorem is stating. Last, Fig. 13 shows the histograms of the ensemble of $S_n(f)$ for $n = 40$ and $n = 2000$. We observe that the histogram for $n = 40$ is far from a normal distribution while for $n = 2000$ the histogram is close to a normal distribution.

5.3 Markov Chain Monte Carlo (MCMC)

We will now give a very short summary of the idea behind Markov Chain Monte Carlo (MCMC). There are many different (and very general) approaches to the construction of MCMC methods; however, we will concentrate completely on *Metropolis MCMC*.

The background. We consider the following problem: Let π be a probability distribution on state space $\mathbf{S}$. We are interested in computing the expectation value of some observable $f : \mathbf{S} \to \mathbb{R}$,

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

Figure 13: Histograms of the ensemble of $S_n(f)$ for $n = 40$ (left panel) and $n = 2000$ (right panel) for the chain considered in example 5.7.

For the sake of simplification we consider π and f to be everywhere positive both. If the state space is gigantic the actual computation of the expectation value may be a very hard or even seemingly infeasible task. Furthermore, in many such applications the value of $\pi(x)$ for some arbitrary $x \in \mathbf{S}$ cannot be computed *explicitly* but has the form

$$\pi(x) = \frac{1}{Z}\mu(x),$$

with some μ for which $\mu(x)$ is explicitly given and easily computable but the normalization constant

$$Z = \mathbb{E}_\pi[\mathbf{1}] = \sum_{x \in \mathbf{S}} \pi(x),$$

is not!

The idea. Given π we want to construct an irreducible Markov Chain $\{X_k\}_{k\in\mathbb{N}}$ such that

(C1) π is a stationary distribution of $\{X_k\}_{k\in\mathbb{N}}$, and

(C2) realizations of $\{X_k\}_{k\in\mathbb{N}}$ can be computed by evaluations of μ only (without having to compute Z).

Then we can exploit the law of large numbers and approximate the desired expectation value by mean values $S_n(f)$ of finite realization of $\{X_k\}_{k\in\mathbb{N}}$:

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

Theory. We start with the assumption that we have a transition function $Q : \mathbf{S} \times \mathbf{S} \to \mathbb{R}$ that belongs to an irreducible Markov Chain. We can take *any* irreducible transition function; it does not need to have any relation to π but should have the additional property of being efficiently computable. Surprisingly this is enough to construct an irreducible Markov chain with above properties (C1) and (C2):

Theorem 5.8 *Let $\pi > 0$ be the given probability distribution and $Q : \mathbf{S} \times \mathbf{S} \to \mathbb{R}$ a transition function of an arbitrary irreducible Markov Chain which satisfies*

$$Q(x, y) \neq 0 \quad \Leftrightarrow \quad Q(y, x) \neq 0,$$

which we will call the proposal function. *Furthermore let $\Psi : (0, \infty) \to (0, 1]$ be a function satisfying*

$$\frac{\Psi(x)}{\Psi(1/x)} = x, \qquad \textit{for all } x \in (0, \infty).$$

Then, the acceptance function $A : \mathbf{S} \times \mathbf{S} \to (0,1]$ *is defined by*

$$A(x,y) = \begin{cases} \Psi\left(\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right) \in (0,1] & \text{if } Q(x,y) \neq 0 \\ 0 & \text{otherwise} \end{cases},$$

and the transition function $P : \mathbf{S} \times \mathbf{S} \to \mathbb{R}$ *by*

$$P(x,y) = \begin{cases} Q(x,y)A(x,y) & \text{if } x \neq y \\ 1 - \sum\limits_{z \in \mathbf{S}, z \neq x} Q(x,z)A(x,z) & \text{if } x = y \end{cases}. \tag{36}$$

Then the Markov Chain $\{X_k\}_{k\in\mathbb{N}}$ *associated with* P *is irreducible, and has* π *as stationary distribution, i.e.,* $\pi P = \pi$.

Proof: First, we observe that P indeed is a transition function:

$$\sum_{y \in \mathbf{S}} P(x,y) = \sum_{z \in \mathbf{S}, z \neq x} Q(x,z)A(x,z) + 1 - \sum_{z \in \mathbf{S}, z \neq x} Q(x,z)A(x,z) = 1.$$

Furthermore, since $\pi > 0$ and $Q \geq 0$ per definition, we have $A \geq 0$ and $P \geq 0$.

In particular, if $Q(x,y) > 0$ for some $x \not= y$ then $Q(y,x) > 0$ also. But then $A(x,y) > 0$, $A(y,x) > 0$ and $P(x,y) > 0$, $P(y,x) > 0$. Thus P inherits the irreducibility from Q.

Next, assume that $x, y \in \mathbf{S}$ are chosen such that $Q(x,y) \neq 0$ and therefore $Q(y,x) \neq 0$; then also $P(x,y) > 0$ and therefore $P(y,x) > 0$, and the same for A. Hence, we compute by exploitation of the property of Ψ:

$$\frac{A(x,y)}{A(y,x)} = \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)},$$

and therefore finally

$$\pi(x)P(x,y) = \pi(y)P(y,x). \tag{37}$$

The same equation is obviously satisfied if $x = y$. If $x \neq y$ and such that $Q(x,y) = 0$ then also $P(x,y) = 0$ such that (37) is satisfied again. Thus, the detailed balance condition (37) is valid for all $x, y \in \mathbf{S}$. Then, by Theorem 4.14, π is a stationary distribution of $\{X_k\}_{k\in\mathbb{N}}$ and $\pi P = \pi$. □

The usual choice for the function Ψ is the so-called Metropolis function

$$\Psi(x) = \min\{1, x\}.$$

In contrast to this, the transition function Q needs to be chosen individually for each application under consideration.

Algorithmic realization. Computing a realization $(x_k)_{k=0,1,2,\ldots}$ of the Markov Chain $\{X_k\}_{k\in\mathbb{N}}$ associated with the transition function P is easier than expected:

1. Start in some state x_0 with iteration index $k = 0$.
2. Let x_k be the present state.
3. Draw y from the proposal distribution $Q(x_k, \cdot)$. This can be done as described in Sec. 2.3.
4. Compute the acceptance probability
$$a = A(x_k, y) = \Psi\Big(\frac{\pi(y)Q(y, x_k)}{\pi(x_k)Q(x_k, y)}\Big) = \Psi\Big(\frac{\mu(y)Q(y, x_k)}{\mu(x_k)Q(x_k, y)}\Big).$$
5. Draw $r \in [0, 1)$ randomly from a uniform distribution in $[0, 1)$.
6. Set
$$x_{k+1} = \begin{cases} y & \text{if } r \le a \\ x_k & \text{if } r > a \end{cases}.$$
7. Set $k := k + 1$ and return to step 2.

Thus, $a = A(x_k, y)$ really is an acceptance probability since the proposed state y is taken as the next state x_{k+1} with probability a, while we remain in the present state ($x_k = x_{k+1}$) with probability $1 - a$.

If in addition Q is symmetric, i.e., $Q(x, y) = Q(y, x)$ for all pairs $x, y \in \mathbf{S}$, then the acceptance probability takes the particularly simple form

$$A(x_k, y) = \Psi\Big(\frac{\mu(y)}{\mu(x_k)}\Big).$$

General remarks. We observe that we do not need the normalization constant Z from $\pi = \mu/Z$; instead we only need fractions of the form $\mu(x)/\mu(y)$. Thus, we achieved everything we wanted, and we can compute the desired expectation values from the mean values $S_n(f)$ of realization of $\{X_k\}_{k\in\mathbb{N}}$. The speed of convergence $S_n(f) \to \mathbb{E}_\pi[f]$ is given by the central limit theorem; we can apply Theorem 5.6 since the chain $\{X_k\}_{k\in\mathbb{N}}$ is reversible by construction. From our remarks to the central limit theorem on page 59 we learn that the variance of the random variables $S_n(f)$ will decay with $\mathcal{O}(n^{-1}\sigma^2(f))$ with an asymptotic variance $\sigma^2(f)$ that essentially depends on largest eigenvalues of P (not counting $\lambda = 1$). It is an art to choose Q in a way that minimizes these largest eigenvalues of P and thus minimizes the asymptotic variance.

Figure 14: Illustration of a grid of $N = 36$ spins. Locations marked with 'o' have spin -1, while locations with '+'-marks have spin $+1$.

Figure 15: Illustration of the neighborhood $\mathcal{N}_i$ of spin location i on the spin grid: the eight spin locations in $\mathcal{N}_i$ are enlarged.

5.4 Example for MCMC: the Ising Model

The so-called *Ising model* is one of the simplest atomistic models for studying magnetization of crystals. It is named after the physicists E. Ising and is defined via a discrete collection of variables called spins, which can take on the value 1 or -1. These spins are located on a regular grid or a graph or network, respectively; for a two-dimensional illustration see Fig. 14. Let us enumerate the grid location by $i = 1, \ldots, N$, that is, we consider the state space

$$\mathbf{S} = \{s = (s_1, \ldots, s_N) : \ s_i = \pm 1, \ i = 1, \ldots, N\},$$

which (for large N) is a quite large space: $|\mathbf{S}| = 2^N$. For each spin location i we define a neighborhood $\mathcal{N}_i \subset \mathbf{S}$; see illustration in Fig. 15. Typically the number of spins in the neighborhood of a typical spin i is a constant of the model, $|\mathcal{N}_i| = M$ independent of i, with possible exceptions at the boundaries of the grid.

Only neighboring spins interact physically. The energy of the spin state $s \in \mathbf{S}$ is given by

$$E(s) = -\frac{1}{2} \sum_{\substack{i=1,\ldots,N \\ j \in \mathcal{N}_i}} J_{ij} s_i s_j \ - \ M \sum_{i=1}^{N} s_i,$$

where M is the external magnetic field. Notice that the product of spins is either $+1$ if the two spins are the same (aligned), or -1 (anti-aligned). The coupling factor J_{ij} is half the difference in energy between the two possibilities; it normally does not depend on the pair (i, j). Magnetic interactions seek to align spins relative to one another. Typically, if $J_{ij} > 0$ the interaction is called ferromagnetic, for $J_{ij} < 0$ antiferromagnetic. A ferromagnetic interaction tends to align spins, and an antiferromagnetic tends to anti-align them. Spins become randomized when thermal energy is greater than the strength of the interaction, that is, with increasing temperature random perturbations of the optimal alignment of the spins increases. The stationary distribution of physical relevance then is

$$\pi(s) = \frac{1}{Z}\mu(s), \quad \mu(s) = \exp(-\beta E(s)), \quad Z = \sum_{s \in \mathbf{S}} \mu(s),$$

where $\beta > 0$ is the inverse temperature. This is an obvious case where Z cannot be computed directly for large N, as are the expectation values $\mathbb{E}_\pi(f)$ for some given observable $f : \mathbf{S} \to \mathbb{R}$.

In order to apply MCMC to this case, we first have to define the proposal transition function. To this end, we define for each $s \in \mathbf{S}$:

$$\text{switch}(s) = \{s' \in \mathbf{S}; \quad s'_j = s_j, \text{ except for exactly one } i \in \{1, \ldots, N\} \text{ with } s'_i = -s_i\},$$

such that $|\text{switch}(s)| = N$ for all $s \in \mathbf{S}$. For the proposal step we randomly draw a position i uniformly from $\{1, \ldots, N\}$, and then propose to move from s to s' by switching the spin at location i:

$$Q(s, s') = \begin{cases} 1/N & \text{if } s' \in \text{switch}(s) \\ 0 & \text{otherwise} \end{cases}.$$

Obviously, the Markov chain defined by Q is irreducible, since we can move from each spin state s to another s' by at most N spin flips. Furthermore, we find that

$$s' \in \text{switch}(s) \Leftrightarrow s \in \text{switch}(s'),$$

and therefore

$$Q(s, s') = Q(s', s).$$

As a consequence we have that the acceptance probability gets the form

$$A(s, s') = \begin{cases} \Psi\left(\frac{\mu(s')}{\mu(s)}\right) & \text{if } s' \in \text{switch}(s) \\ 0 & \text{otherwise} \end{cases},$$

where

$$\frac{\mu(s')}{\mu(s)} = \exp\left(-\beta \Delta E_{s,s'}\right), \quad \text{with } \Delta E_{s,s'} = E(s') - E(s) = \sum_{j \in \mathcal{N}_i} J_{ij} s_i s_j + 2M,$$

where i denotes the location of the switch. If we moreover choose $\Psi(x) = \min\{1, x\}$, then

$$A(s, s') = \begin{cases} 1 & \text{if } s' \in \text{switch}(s) \text{ and } \Delta E_{s,s'} \leq 0 \\ \exp\left(-\beta \Delta E_{s,s'}\right) & \text{if } s' \in \text{switch}(s) \text{ and } \Delta E_{s,s'} > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (38)$$

which can be computed very efficiently since evaluation of $\exp\left(-\beta \Delta E_{s,s'}\right)$ just involves a sum over the neighborhood of the switch location i.

As a consequence of our above considerations MCMC for the Ising model takes the following simple form:

1. Start in some spin state $s^{(0)}$ with iteration index $k = 0$.

2. Let $s^{(k)}$ be the present spin state.

3. Determine the proposal $s' \in \text{switch}(s^{(k)})$ by switching the spin at location i in $s^{(k)}$, where i is drawn randomly from $\{1, \ldots, N\}$.

4. Compute $\Delta E_{s^{(k)},s'} = E(s') - E(s^{(k)}) = \sum_{j \in \mathcal{N}_i} J_{ij} s_i^{(k)} s_j^{(k)} + 2M$.

5. Compute the acceptance probability $a = A(s^{(k)}, s')$ according to equation (38).

6. Draw $r \in [0, 1)$ randomly from a uniform distribution in $[0, 1)$.

7. Set
$$s^{(k+1)} = \begin{cases} s' & \text{if } r \leq a \\ s^{(k)} & \text{if } r > a \end{cases}.$$

8. Set $k := k + 1$ and return to step 2.

Thus, in contrast to the huge state space with 2^N states, each iteration of the MCMC scheme is computationally very cheap, that is, produces computational effort of just $|\mathcal{N}_i|$ additions and multiplications and the computation of two random variables. Thus, the Law of Large Numbers can be used to approximate expectation values via running averages of the MCMC chain:

$$\frac{1}{n+1} \sum_{k=0}^{n} f(s^{(k)}) \to \mathbb{E}_\pi(f) = \sum_{s \in \mathbf{S}} \pi(x) f(x) \quad \text{for } n \to \infty.$$

Generalization. Whenever the stationary distribution of interest has the form $\pi(x) = \exp(-\beta E(x))/Z$ on some huge state space the acceptance step of the resulting MCMC algorithm will have essentially the form we derived above. Such cases are very frequent in the physical, chemical and biological literature and range from applications to materials and molecular sciences to network evolution in biology. In all these diverse applications one fundamental problem appears over and over again: Whenever the state space contains disjoint subsets that are separated by high energy barriers our Monte Carlo Markov Chain typically exhibits $\Lambda_{\max}(P) \approx 1$ such that its convergence as given by the central limit theorem slows down drastically (compare remarks on page 59). More precisely, let there be subsets $A, B \subset \mathbf{S}$ such that the transition probability from A to B, $P(A, B) = \sum_{x \in A} \pi(x) P(x, B)/\pi(A)$, is small, $P(A, B) = \epsilon$, then our Monte Carlo Markov Chain exhibits $1 - \Lambda_{\max}(P) < C\epsilon$ (see Sec. 6.4 below for more details). If there is an energy barrier ΔE between A and B then one often finds that ϵ scales like $\exp(-\beta \Delta E)$, that is, $\Lambda_{\max}(P)$ goes to 1 exponentially fast with decreasing temperature, i.e., with growing inverse temperature β. This problem is often referred to as the *trapping problem*.

Figure 16: Typical realization of the MCMC sampler described in the text for $\beta = 3$ and $\beta = 5$.

Figure 17: Convergence of the running means $S_n(f)$ (for $f(x) = x$) of the MCMC sampler described in the text for $\beta = 3$ and $\beta = 5$.

Example 5.9 *Let us consider the state space* $\mathbf{S} = \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ *and the stationary distribution* $\pi_\beta(x) = \exp(-\beta E(x))/Z$ *induced by*

$$E(x) = ((x/10)^2 - 1)^2.$$

The energy landscape exhibits an energy barrier of size $\Delta E = 1$ *between the two energy minima at* $x = \pm 10$. *As proposal transition function we choose* $Q(x, x+1) = Q(x, x-1) = 1/2$ *and* $Q(x, y) = 0$ *for all* $y \neq x - 1, x + 1$. *Fig. 16 shows typical realizations of the resulting MCMC sampler for* $\beta = 3$ *and* $\beta = 5$. *We observe that for* $\beta = 5$ *the energy barrier is overcome much less frequent as for* $\beta = 3$ *indicating a trapping problem. In fact, we find* $\Lambda_{\max}(P) = 0.999754$ *for* $\beta = 5$ *and* $\Lambda_{\max}(P) = 0.998888$ *for* $\beta = 3$. *Fig. 17 shows the resulting convergence of the running means of MCMC sampler for* $f(x) = x$ *with* $\mathbb{E}_{\pi_\beta}[f] = 0$ *independent of* β. *Again, we observe that for* $\beta = 3$ *we can see convergence of the running means* $S_n(f)$ *to* $\mathbb{E}_{\pi_\beta}[f] = 0$ *after* 1.000.000 *steps while* $S_n(f)$ *still deviates significantly from* 0 *for* $\beta = 5$. *Accordingly, the asymptotic variance* $\sigma^2(f)$ *differs between* 152.520 *(for* $\beta = 3$*) and* 720.260 *(for* $\beta = 5$*). This means that for* $\beta = 5$ *we still have a standard variation of about* 0.83 *for* $S_n(f)$ *after* 1.000.000 *steps which fits to the final deviation between* $S_n(f)$ *and* $\mathbb{E}_{\pi_\beta}[f] = 0$ *in the left panel of Fig. 17.*

5.5 Replica Exchange Monte Carlo

Replica exchange Monte-Carlo (REMC) has been designed to circumvent the trapping problem. The idea behind REMC is simple: That is, let P_{β_1}, and P_{β_2} be Markov chain Monte Carlo transition functions of the form discussed in the last section for inverse temperatures β_1 and $\beta_2 < \beta_1$ and associated stationary distributions $\pi_i(x) = \exp(-\beta_i E(x))/Z_i$, $i = 1, 2$. Let us assume that the form of the energy landscape E induces the trapping problem. Then, the chain $X^{(\beta_1)}$ given by P_{β_1} will converge significantly slower than the chain $X^{(\beta_2)}$ with transition function P_{β_2}, because $X^{(\beta_2)}$ will overcome energy barriers more often than $X^{(\beta_1)}$. Consequently, if we could sometimes *exchange* states between the two chains then $X^{(\beta_1)}$ would sometimes be pushed over energy barriers by inheriting a state on the opposite side of the barrier from $X^{(\beta_2)}$. In order to understand how such an exchange can be

done *without* changing the stationary distribution of the two chains let us first introduce the chain

$$Z = (X^{(\beta_1)}, X^{(\beta_2)}), \quad \text{i.e., } Z_k = (X_k^{(\beta_1)}, X_k^{(\beta_2)}),\ k = 0, 1, 2, \ldots,$$

with transition functions $P_Z((x,y),(u,v)) = P_{\beta_1}(x,u)P_{\beta_2}(y,v)$ and stationary distribution $\pi_Z(x,y) = \pi_1(x)\pi_2(y)$. This Markov chain satisfies the detailed balance condition and can be realized by *parallel* execution of two MCMC algorithms of the form given on page 68. Now, we will change the transition function P_Z to the replica exchange transition function P_{RE}. To this end, let $S(u,v)$ denote the probability of the exchange $(u,v) \to (v,u)$. Then, the probability of first a step from state (x,y) to (u,v) and then an exchange resulting in (v,u) is

$$P_{RE}((x,y),(v,u)) = P_Z((x,y),(u,v)) \cdot S(u,v),$$

and P_{RE} satisfies the detailed balance condition (DBC) if

$$\pi_1(u)\pi_2(v)S(u,v) = \pi_1(v)\pi_2(u)S(v,u).$$

If we find $S : \mathbf{S} \times \mathbf{S} \to (0,1]$ such that the above DBC is valid, we in fact have achieved what we wanted: π_Z is a stationary distribution of P_{RE}. Fortunately, we already know how to find such a function S: If

$$\begin{aligned} S(u,v) &= \Psi(g(u,v)), \\ g(u,v) &= \frac{\pi_1(v)\pi_2(u)}{\pi_1(u)\pi_2(v)} = \exp\Big(-(\beta_1 - \beta_2)(E(v) - E(u))\Big), \end{aligned}$$

with some function Ψ satisfying $\Psi(g)/\Psi(1/g) = g$, then

$$\frac{S(u,v)}{S(v,u)} = \frac{\Psi(g(u,v))}{\Psi(g(v,u))} = g(u,v) = \frac{\pi_1(v)\pi_2(u)}{\pi_1(u)\pi_2(v)}.$$

Sumarizing, we thus derived the following statement

Theorem 5.10 *Let $\Psi : (0,\infty) \to (0,1]$ be a function satisfying $\Psi(x)/\Psi(1/x) = x$, for all $x \in (0,\infty)$. Furthermore let P_{β_i}, $\pi_i = \exp(-\beta_i E(x))/Z_i$, $i = 1,2$, and $P_Z((x,y),(u,v)) = P_{\beta_1}(x,u)P_{\beta_2}(y,v)$ be as introduced above. Define the* exchange function $S : \mathbf{S} \times \mathbf{S} \to (0,1]$ *by*

$$S(u,v) = \Psi\Big[\exp\Big(-(\beta_1 - \beta_2)(E(v) - E(u))\Big)\Big] \in (0,1],$$

and the transition function $P_{RE} : \mathbf{S}^2 \times \mathbf{S}^2 \to \mathbb{R}$ by

$$P_{RE}((x,y),(u',v')) = P_Z((x,y),(u,v)) \cdot \begin{cases} S(u,v) & \text{if } (u',v') = (v,u) \\ 1 - S(u,v) & \text{if } (u',v') = (u,v) \end{cases}.$$

Figure 18: REMC sampler described in the text for $\beta_2 = 3$ and $\beta_1 = 5$. Left panel: Typical realization of the $\beta_1 = 5$-component. Right panel: Convergence of the running mean for the first component ($\beta_1 = 5$).

Then the Markov Chain $Z^{(RE)} = (X_k^{(1)}, X_k^{(2)})_{k=0,1,2,\ldots}$ *associated with* P_{RE} *is irreducible and reversible with* π_Z *as stationary distribution. Therefore it allows to compute expectation values for two different temperatures:*

$$\frac{1}{n+1}\sum_{k=0}^{n} f(X_k^{(i)}) \to \mathbb{E}_{\pi_i}[f], \quad i = 1, 2,$$

for $n \to \infty$.

As a consequence, algorithmic realization of REMC (for $\Psi(x) = \min(1, x)$) takes the following simple form:

1. Start in some joint state $z^{(0)} = (x_1^{(0)}, x_2^{(0)})$ with iteration index $k = 0$.
2. Let $z^{(k)} = (x_1^{(k)}, x_2^{(k)})$ be the present joint state.
3. Draw the next regular joint state $z^* = (x_1^*, x_2^*)$ from $P_Z(z^{(k)}, \cdot)$, i.e., determine x_i^* via realization of one step of $P_{\beta_i}(x_i^{(k)}, \cdot)$ for $i = 1, 2$ like in regular MCMC.
4. Compute the exchange probability
$$s = S(x_1^*, x_2^*) = \min\Big[1, \exp\Big(-(\beta_1 - \beta_2)(E(x_2^*) - E(x_1^*))\Big)\Big].$$
5. Draw $r \in [0, 1)$ randomly from a uniform distribution in $[0, 1)$.
6. Accept the exchange if $s \leq r$, that is, set
$$z^{(k+1)} = (x_1^{(k+1)}, x_2^{(k+1)}) = \begin{cases} (x_2^*, x_1^*) & \text{if } r \leq s \\ (x_1^*, x_2^*) & \text{if } r > s \end{cases}.$$
7. Set $k := k + 1$ and return to step 2.

Example 5.11 *Let us revisit the situation of Example 5.9 from above. Now, we use the REMC sampler for the stationary distribution* $\pi_\beta(x) = \exp(-\beta E(x))/Z$ *for* $\beta_2 = 3$ *and* $\beta_1 = 5$. *The REMC sampler introduces exchanges between the two standard MCMC samplers described in Example 5.9. Fig. 18 shows a typical realization the resulting convergence of the running means of the* first *component of the REMC sampler, i.e., of the component that is distributed according to* $\pi_{\beta_1=5}$, *for* $f(x) = x$ *with* $\mathbb{E}_{\pi_{\beta_1}}[f] = 0$. *In comparison to the*

respective left panels of Figs. 16 and Figs. 17 we observe that trapping is much reduced for $\beta = 5$ *and that convergence of the running means* $S_n(f)$ *to* $\mathbb{E}_{\pi_\beta}[f] = 0$ *is significantly faster. When we compute the asymptotic variance* $\sigma^2(f)$ *of the running means* $S_n()$ *of the first component (*$\beta = 5$*) approximately by studying the limit* $n\mathrm{Var}(S_n(f))$ *large enough step numbers* n *we find a value of about* 145.000 *which shows that our simple REMC converges a rough factor of 5 faster than the pure MCMC at* $\beta = 5$.

6 Identification of macroscopic properties

6.1 Identification of communication classes

Before we state the main result of this section, we need the following two notions. Given a probability distribution ν, the **support** of ν is defined as the set of all states $x \in \mathbf{S}$, restricted to which ν is positive, i.e.,

$$\text{supp}(\nu) \quad = \quad \{x \in \mathbf{S} : \nu(x) > 0\}.$$

We say that the probability densities $\nu_1, \ldots, \nu_d$ have mutually disjoint supports, if

$$\text{supp}(\nu_k) \cap \text{supp}(\nu_m) \quad = \quad \emptyset$$

for all $x \in \mathbf{S}$ and $k, m = 1, \ldots, d$ with $k \neq m$. In other words, $\nu_1, \ldots, \nu_d$ have mutually disjoint supports, if

$$\nu_k(x)\nu_m(x) \quad = \quad 0$$

for any $x \in \mathbf{S}$ and every $k, m = 1, \ldots, d$ with $k \neq m$.

Theorem 6.1 *Let P denote a stochastic transition matrix with maximal stationary distribution π and satisfying Assumption R. Then the following two statements are equivalent:*

1. *the eigenvalue $\lambda = 1$ is of multiplicity d, and there exist exactly d stationary distributions $\nu_1, \ldots, \nu_d$ with mutually disjoint supports. (Note that in this case $\pi = \frac{1}{d}(\nu_1 + \cdots + \nu_d)$ is a maximal stationary distribution.)*

2. *there exists a decomposition of the state space*

$$\mathbf{S} \quad = \quad C_1 \cup \ldots \cup C_d \cup D$$

into d mutually disjoint invariant positive recurrent communication classes $C_1, \ldots, C_d$ satisfying $\pi(C_k) > 0$ for $k = 1, \ldots, d$, and a (possibly empty) collection of null–recurrent and/or transient communication classes D with $\pi(D) = 0$.

In broad terms, Theorem 6.1 states that we can identify the number of invariant positive recurrent communication classes simply by solving the (left) eigenvalue problem $\nu P = \nu$ and counting the multiplicity of the eigenvalue $\lambda = 1$, hence determining the number of linear independent eigenvectors corresponding to $\lambda = 1$.

Example 6.2 *Consider the nine state Markov chain defined by*

$$P = \begin{pmatrix} & 1.0 & & & & & & & \\ & & 1.0 & & & & & & \\ 1.0 & & & & & & & & \\ & & 0.9 & 0.1 & & & & & \\ & 0.2 & & 0.1 & 0.4 & 0.3 & & & \\ & & & & 0.4 & 0.1 & & 0.5 & \\ & & & 0.4 & & 0.4 & 0.2 & & \\ & & & & & & & 0.7 & 0.3 \\ & & & & & & & 0.2 & 0.8 \end{pmatrix}$$

on the state space $S = \{1, \ldots, 9, \}$. *We already know that there exist two invariant positive recurrent communication classes* $C_1 = \{1, 2, 3\}$ *and* $C_2 = \{8, 9\}$, *hence* $d = 2$, *while* $D = \{4, 5, 6, 7\}$ *is the collection of transient communication classes. Solving the (left) eigenvalue problem* $\nu P = \lambda\nu$ *yields*

$$\sigma(P) = \{1.0, 1.0, 0.63, 0.5, 0.2, 0.1, -0.13, -0.5 \pm 0.87i\},$$

where we have counted the eigenvalues according to their multiplicity. Hence, the eigenvalue $\lambda = 1$ *is two–fold, the corresponding stationary densities with disjoint support are*

$$\begin{aligned} \nu_1 &= \Big(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0, 0, 0, 0\Big), \\ \nu_2 &= \Big(0, 0, 0, 0, 0, 0, 0, \frac{2}{5}, \frac{3}{5}\Big). \end{aligned}$$

Moreover, $\pi = 0.5(\nu_1 + \nu_2)$ *is a maximal stationary distribution. In accordance with Theorem 6.1 we have* $\pi(C_1) > 0$ *and* $\pi(C_2) > 0$, *while* $\pi(D) = 0$.

Note that we can also identify the decomposition of the state space given by $\mathbf{S} = C_1 \cup C_2 \cup D$ *from the knowledge of* ν_1, ν_2 *and* π *by defining the sets according to*

$$\begin{aligned} C_1 &= \{x \in \mathbf{S} : \nu_1(x) > 0\}, \\ C_2 &= \{x \in \mathbf{S} : \nu_2(x) > 0\}, \\ D &= \{x \in \mathbf{S} : \pi(x) = 0\}. \end{aligned}$$

Alternatively, we can set $D = \{x \in \mathbf{S} : \nu_1(x) = \nu_2(x) = 0\}$.

The next proposition states that, as demonstrated in the above example, we can always exploit the stationary distributions corresponding to the eigenvalue $\lambda = 1$ in order to identify the communication classes of the corresponding to the Markov chain.

Proposition 6.3 *Let P denote a stochastic transition matrix with maximal stationary distribution π and satisfying Assumption R. Moreover, assume that the eigenvalue $\lambda = 1$ is d–fold with corresponding stationary distributions $\nu_1, \ldots, \nu_d$ that have mutually disjoint support. Then*

$$C_k \quad = \quad \{x \in \mathbf{S} : \nu_k(x) > 0\}$$

for $k = 1, \ldots, d$ defines exactly the invariant positive recurrent communication classes of the corresponding Markov chain, while

$$D \quad = \quad \{x \in \mathbf{S} : \nu_1(x) = \ldots = \nu_d(x) = 0\}$$

defines the collection of all null–recurrent and/or transient communication classes.

In order to exploit Proposition 6.3 numerically, we face the following problem: Assume we have solved the (left) eigenvalue problem $\nu P = \lambda\nu$ for $\lambda = 1$ by applying some eigenvalue solver (like *eig* to P^T in MATLAB). The result is a certain number of linear independent (left) eigenvectors $\eta_1, \ldots, \eta_d$ solving $\eta_k P = \eta_k$. Almost always, the eigenvectors η_k will not be stationary distributions with disjoint support that we are looking for in view of Proposition 6.3. Even worse, most of the time the η_k will be not even non–negative. The next statement, however, will be useful to overcome this problem and to compute the desired stationary distributions from the output of the eigenvalue solver.

Proposition 6.4 *Given a stochastic transition matrix P and some eigenvector η satisfying $\eta P = \eta$. Decompose the eigenvector into its positive part η^+ given by*

$$\eta^+(x) = \begin{cases} \eta(x); & \text{if } \eta(x) > 0 \\ 0; & \text{otherwise .} \end{cases}$$

and its negative part (defined analogously) such that $\eta = \eta^+ - \eta^-$. Then, we have

$$\eta^+ P = \eta^+ \quad \text{and} \quad \eta^- P = \eta^-,$$

hence the positive as well as the negative part of η are also eigenvectors of P corresponding to $\lambda = 1$.

As an example, consider the stochastic transition matrix P from Example 4.16. Then $\eta P = \eta$ for, e.g., $\eta = (-3, 2, -1, 0)$. By Proposition 6.4, we have that also $\eta^+ P = \eta^+$ with $\eta^+ = (0, 2, 0, 0)$, and $\eta^- P = \eta^-$ with $\eta^- = (3, 0, 1, 0)$ are eigenvectors corresponding to $\lambda = 1$.

In order to continue the computational identification of communication classes, consider the 9×9 stochastic transition matrix

$$P = \begin{pmatrix} 0.1 & & & & & & & 0.9 & \\ & 0.1 & & 0.5 & 0.4 & & & & \\ & 0.2 & 0.3 & & & 0.2 & 0.3 & & \\ & 0.5 & & 0.2 & 0.3 & & & & \\ & 0.2 & & 0.1 & 0.7 & & & & \\ & & & & & 0.5 & & & 0.5 \\ & & 0.1 & 0.3 & 0.1 & 0.2 & 0.2 & & 0.1 \\ 0.4 & & & & & & & 0.6 & \\ & & & & & 0.3 & & & 0.7 \end{pmatrix}. \tag{39}$$

Solving the (left) eigenvalue problem yields

$$\sigma(P) = \{1.00, 1.00, 1.00, 0.43, 0.36, 0.20, 0.07, -0.30, -0.36\};$$

thus, the eigenvalue $\lambda = 1$ is threefold ($d = 3$) with corresponding eigenvectors

$$\begin{aligned} \eta_1 &= (0.32, -0.03, 0.00, -0.02, -0.06, 0.33, 0.00, 0.72, 0.55) \\ \eta_2 &= (0.09, +0.08, 0.00, +0.07, +0.18, 0.35, 0.00, 0.20, 0.58) \\ \eta_3 &= (0.39, +0.07, 0.00, +0.07, +0.17, 0.41, 0.00, 0.88, 0.68). \end{aligned}$$

From this, we can compute a maximal stationary distribution exploiting Proposition 6.4

$$\pi = \frac{1}{Z} \sum_{k=1}^{d} \left(\eta_k^+ + \eta_k^-\right) = (0.14, 0.05, 0.00, 0.04, 0.10, 0.12, 0.00, 0.34, 0.21).$$

By Proposition 6.3 we find that

$$D = \{x \in \mathbf{S} : \pi(x) = 0\} = \{3, 7\}$$

is the collection of transient communication classes. Hence, it remains to identify the $d = 3$ invariant positive recurrent communication classes.

To do so, we exploit the following two facts: First, assume that we know the d stationary distributions $\nu_1, \ldots, \nu_d$ with mutually disjoint support as given in Theorem 6.1. Since both the η_k as well as the ν_k are eigenvectors corresponding to $\lambda = 1$, we may write each η_m as a linear combination of the $\nu_1, \ldots, \nu_d$:

$$\eta_m = \sum_{k=1}^{d} c_{mk} \nu_k$$

for suitable coefficients $c_{m1}, \ldots, c_{md} \in \mathbb{R}$. Since every ν_k is different from zero only on C_k, we have $\eta_m(x) = c_{mk}\nu_k(x)$ for all $x \in C_k$, or equivalently

$$\frac{\eta_m}{\nu_k} \;=\; c_{mk} \;=\; \text{const on } C_k$$

for every $k, m = 1, \ldots, d$. As a second fact, we exploit that there exist positive coefficients $\alpha_1, \ldots, \alpha_d$ with $\alpha_1 + \ldots + \alpha_d = 1$ such that

$$\pi \;=\; \alpha_1\nu_1 + \ldots \alpha_d\nu_d.$$

Combining the two facts yields

Theorem 6.5 *Let P denote a stochastic transition matrix with maximal stationary distribution and satisfying Assumption R. Moreover, assume that the eigenvalue $\lambda = 1$ is d–fold with corresponding eigenvectors $\eta_1, \ldots, \eta_d$. Define the maximal stationary distribution π according to*

$$\pi \;=\; \frac{1}{Z}\sum_{k=1}^{d}\left(\eta_k^+ + \eta_k^-\right),$$

and assume that $\pi > 0$ on $\mathbf{S}$, hence $D = \emptyset$.[6]

1. *Define the reweighted vectors $u_1, \ldots, u_d$ by*

$$u_m \;=\; \frac{\eta_m}{\pi}.$$

Then, each vector u_k is constant on each of the invariant positive recurrent communication classes $C_1, \ldots, C_d$.

2. *Based on the reweighted vectors $u_1, \ldots, u_d$ define the transformation $V : \mathbf{S} \to \mathbb{R}^d$ given by*

$$x \;\mapsto\; V(x) \;=\; (u_1(x), \ldots, u_d(x)). \tag{40}$$

Then, V assumes exactly d different values $\gamma_1, \ldots, \gamma_d$, and the d invariant positive recurrent communication classes are given by

$$C_k \;=\; \{x \in \mathbf{S} : V(x) = \gamma_k\}$$

for $k = 1, \ldots, d$.

Theorem 6.5 can directly be exploited for computational purposes. To demonstrate this, let us continue the example based on the transition matrix (39). Computing the reweighted vectors $u_1, \ldots, u_d$ with $d = 3$ yields

$$\begin{aligned} u_1 &= (2.0, -1.3, *, -1.3, -1.3, 2.0, *, 2.0, 2.0) \\ u_2 &= (2.5, +2.7, *, +2.7, +2.7, 2.6, *, 2.5, 2.6) \\ u_3 &= (2.3, +2.9, *, +2.9, +2.9, 2.2, *, 2.3, 2.2), \end{aligned}$$

[6] Otherwise restrict to the subset of $\mathbf{S}$, on which π is positive.

where we marked the two transient states $x = 3, 7$ by $*$. From these, we compute according to (40)

$$V(1) = (2.0, 2.5, 2.3), \quad V(2) = (-1.3, 2.7, 2.9), \quad V(6) = (2.0, 2.6, 2.2).$$

As can be seen, these are the only different values that V assumes, and the invariant positive recurrent communication classes are given by

$$\begin{aligned} C_1 &= \{x \in \mathbf{S} : V(x) = (2.0, 2.5, 2.3)\} = \{1, 8\} \\ C_2 &= \{x \in \mathbf{S} : V(x) = (-1.3, 2.7, 2.9)\} = \{2, 4, 5\} \\ C_3 &= \{x \in \mathbf{S} : V(x) = (2.0, 2.6, 2.2)\} = \{6, 9\}. \end{aligned}$$

Permutation of the original transition matrix P given in Eq. (39) according to $(1, 8, 2, 4, 5, 6, 9, 3, 7)$ confirms the correct identification, since the permuted transition matrix $P_{permute}$ has the required normal form

$$P_{permute} = \begin{pmatrix} 0.1 & 0.9 & & & & & & & \\ 0.4 & 0.6 & & & & & & & \\ & & 0.1 & 0.5 & 0.4 & & & & \\ & & 0.5 & 0.2 & 0.3 & & & & \\ & & 0.2 & 0.1 & 0.7 & & & & \\ & & & & & 0.5 & 0.5 & & \\ & & & & & 0.3 & 0.7 & & \\ & & 0.2 & & & 0.2 & & 0.3 & 0.3 \\ & & & 0.3 & 0.1 & 0.2 & 0.1 & 0.1 & 0.2 \end{pmatrix}$$

as shown in Figure 6 on page 22.

6.2 Identification of cyclic classes

Assume that we have already identified the recurrent and transient communication classes. In the sequel, we therefore restrict our considerations to one single communication class. The following theorem states, how we can detect periodicity; similar to the previous section, it is based on spectral properties of the transition matrix corresponding to the Markov chain under investigation.

Theorem 6.6 *Let P denote an* irreducible *stochastic transition matrix with stationary distribution*[7] *π and satisfying Assumption R. Then the following two statements are equivalent:*

1. *there exist exactly d eigenvalue $\lambda_1, \ldots, \lambda_d$ of unit modulus, i.e., satisfying $|\lambda_k| = 1$ with $k = 1, \ldots, d$. These eigenvalues are located on the unit circle such that they are invariant under a rotation of the complex plane by $\theta = 2\pi/d$. In other words, after a suitable sorting, we have*

$$\lambda_1 = e^{0\theta i}, \lambda_2 = e^{1\theta i}, \ldots, \lambda_d = e^{(d-1)\theta i}$$

[7] Note that under these conditions π is unique, maximal and positive everywhere.

2. *the Markov chain is periodic with period d, and there exists a decomposition of the state space*

$$\mathbf{S} \quad = \quad E_1 \cup \ldots \cup E_d$$

into its cyclic classes $E_1, \ldots, E_d$.

Hence, as for the communication classes, we can detect periodicity by looking at spectral properties of the transition matrix P. More precisely, the period of P equals the number of eigenvalues *on* the unit circle.

As an example, consider the 9×9 stochastic transition matrix

$$P \quad = \quad \begin{pmatrix} & & & & & 0.5 & & 0.5 & \\ & & 0.4 & 0.6 & & & & & \\ 0.5 & & & & & 0.1 & & 0.4 & \\ 0.3 & & & & & 0.4 & & 0.3 & \\ & & 0.8 & 0.2 & & & & & \\ & & & & & & 0.3 & & 0.7 \\ & 0.6 & & & 0.4 & & & & \\ & & & & & & 0.6 & & 0.4 \\ & 0.4 & & & 0.6 & & & & \end{pmatrix}. \qquad (41)$$

Solving the (left) eigenvalue problems yield

$$\sigma(P) = \{1.0, i, -i, -1.0, 0.19 \pm 0.19i, 0, -0.19 \pm 0.19i\};$$

thus, there are $d = 4$ eigenvalues $\lambda_1 = 1.0, \lambda_2 = i, \lambda_3 = -i, \lambda_4 = -1.0$ of unit modulus. Note that these four eigenvalues are invariant under rotation of the complex plane by $\theta = 2\pi/4 = 90°$. As usual, the stationary distribution

$$\pi \quad = \quad (0.07, 0.08, 0.20, 0.13, 0.08, 0.04, 0.16, 0.06, 0.17)$$

is obtained by normalising the (left) eigenvector corresponding to $\lambda = 1$.

For the identification of the cyclic classes we can apply two different strategies.The first strategy is based on exploitation of the dth power P^d of P, hence considering a d–step version of the original Markov chain. Assume the decomposition $\mathbf{S} = E_1 \cup \ldots \cup E_d$; then $p(x_1, E_2) = 1$ for every $x_1 \in E_1$, and $p(x_2, E_3) = 1$ for every $x_2 \in E_2$ and so on. Consequently,

$$p^2(x_1, E_3) = 1, \ \ldots \ , \ p^{d-1}(x_1, E_d) = 1, \ p^d(x_1, E_1) = 1.$$

More general, after d steps of the Markov chain every state $z \in E_k$ returned to E_k. Thus the d–step transition matrix P^d has exactly d invariant positive recurrent communication classes that are identical to $E_1, \ldots, E_d$! Summarising, we can identify the cyclic classes of the one–step transition matrix P

by identifying the invariant positive recurrent communication classes of the d–step transition matrix P^d, hence exploiting the strategy presented in the previous section.

The second strategy is based on the fact that, as in the previous section, the eigenvectors corresponding to the eigenvalues of unit modulus have a special structure that can be directly exploited for identification purposes:

Theorem 6.7 *Let P denote an* irreducible *stochastic transition matrix with stationary distribution π and satisfying Assumption R. Moreover, assume that there are exactly d eigenvalue $\lambda_1, \ldots, \lambda_d$ of unit modulus, i.e., $|\lambda_k| = 1$ for $k = 1, \ldots, d$, with corresponding eigenvectors $\eta_1, \ldots, \eta_d$.*

1. *Define the reweighted vectors $u_1, \ldots, u_d$ by*

$$u_m \;=\; \frac{\eta_m}{\pi}.$$

   Then, each vector u_k is constant on each of the cyclic classes $E_1, \ldots, E_d$.

2. *Based on the reweighted vectors $u_1, \ldots, u_d$ define the transformation $V : \mathbf{S} \to \mathbb{C}^d$ given by*

$$x \;\mapsto\; V(x) \;=\; (u_1(x), \ldots, u_d(x)). \tag{42}$$

   Then, V assumes exactly d different values $\gamma_1, \ldots, \gamma_d$, and the d cyclic classes are given by

$$E_k \;=\; \{x \in \mathbf{S} : V(x) = \gamma_k\}$$

   for $k = 1, \ldots, d$.

In contrast to the Theorem 6.5, where the transformation V maps to $\mathbb{R}^d$, in Theorem 6.7 the corresponding transformation maps to $\mathbb{C}^d$, since in the former case the eigenvectors can chosen to be real–valued, while in the latter case all of them (except for the eigenvectors corresponding to $\lambda = \pm 1$) are complex–valued.

Theorem 6.7 can directly be exploited for computational purposes. To demonstrate this, let us continue the example based on the transition matrix (41). Computing the reweighted vectors $u_1, \ldots, u_d$ with $d = 4$ yields

$$\begin{aligned}
u_1 &= (+4.4, +4.4, -2.2, -2.2, +4.4, +4.4, -2.2, +4.4, -2.2) \\
u_2 &= (-4.4i, +4.4i, +2.2, +2.2, +4.4i, -4.4i, -2.2, -4.4i, -2.2) \\
u_3 &= (+4.4i, -4.4i, +2.2, +2.2, -4.4i, +4.4i, -2.2, +4.4i, -2.2) \\
u_4 &= (+4.4, +4.4, +2.2, +2.2, +4.4, +4.4, +2.2, +4.4, +2.2).
\end{aligned}$$

From these, we see that the transformation V attains only four different values and the cyclic classes are given by

$$\begin{aligned} E_1 &= \{x \in \mathbf{S} : V(x) = (4.4, 4.4i, -4.4i, 4.4)\} = \{2, 5\} \\ E_2 &= \{x \in \mathbf{S} : V(x) = (-2.2, 2.2, 2.2, 2.2)\} = \{3, 4\} \\ E_3 &= \{x \in \mathbf{S} : V(x) = (4.4, -4.4i, 4.4i, 4.4)\} = \{1, 6, 8\} \\ E_4 &= \{x \in \mathbf{S} : V(x) = (-2.2, -2.2, -2.2, 2.2)\} = \{7, 9\}. \end{aligned}$$

Permutation of the original transition matrix P given in Eq. (41) according to $(2, 5, 3, 4, 1, 6, 8, 7, 9)$ confirms the correct identification, since the permuted transition matrix $P_{permute}$ has the required normal form

$$P_{permute} = \begin{pmatrix} & & 0.4 & 0.6 & & & & & \\ & & 0.8 & 0.2 & & & & & \\ & & & & 0.5 & 0.1 & 0.4 & & \\ & & & & 0.3 & 0.4 & 0.3 & & \\ & & & & & & & 0.5 & 0.5 \\ & & & & & & & 0.3 & 0.7 \\ & & & & & & & 0.6 & 0.4 \\ 0.6 & 0.4 & & & & & & & \\ 0.4 & 0.6 & & & & & & & \end{pmatrix}$$

as shown in Figure 7 on page 23.

6.3 Almost invariant communication classes

6.4 Metastability

Let us assume that the Markov chain under investigation is reversible, has stationary distribution π and a transfer operator that satisfies Assumption R of page 50. In terms of the self-adjoint transfer operator $T : l^2(\pi) \to l^2(\pi))$ as defined in (28) this assumption states that:

T has $m + 1$ real eigenvalues $\lambda_0, ..., \lambda_m \in \mathbb{R}$

$$\lambda_0 = 1 > \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m, \tag{43}$$

with an orthonormal system of eigenvectors $(u_j)_{j=1,...,m}$, i.e.

$$Tu_j = \lambda_j u_j, \qquad \langle u_i, u_j \rangle_\pi = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{44}$$

and $u_0 = \mathbf{1}$, and that the remainder of the spectrum of T lies within a ball $B_R(0) \subset \mathbb{C}$ with radius $R < \lambda_m$. In addition we introduce the associated quantity

$$\eta = r/\lambda_1 < 1, \tag{45}$$

which we will call the *spectral gap* later on.

Projection onto dominant eigenspaces. Based on the above assumptions we can write

$$Tv = T\Pi v + T\Pi^{\perp} v = \sum_{j=0}^{m} \lambda_j \langle v, u_j \rangle_\pi u_j + T\Pi^{\perp} v, \tag{46}$$

where Π is the orthogonal projection onto $U = \mathrm{span}\{u_1, ..., u_m\}$

$$\Pi v = \sum_{j=0}^{m} \langle v, u_j \rangle_\pi u_j \tag{47}$$

and $\Pi^{\perp} = \mathrm{Id} - \Pi$ is the projection error with

$$\|T\Pi^{\perp}\| \leq r < \lambda_m, \qquad \mathrm{spec}(T) \setminus \{1, \lambda_1, ..., \lambda_m\} \subset B_r(0) \subset \mathbb{C}. \tag{48}$$

Furthermore, since T is assumed to be self-adjoint, the subspace U and its orthogonal complement do not mix under the action of T:

$$\Pi T \Pi^{\perp} = \Pi^{\perp} T \Pi = 0 \tag{49}$$

and therefore the dynamics can be studied by considering the dynamics of both subspaces separately

$$T^k = (T\Pi)^k + (T\Pi^{\perp})^k \ \forall k \geq 0, \tag{50}$$

where the operator $T\Pi$ is self-adjoint because of (44). In addition we also define the orthogonal projection Π_0 as

$$\Pi_0 v := \langle v, u_0 \rangle_\pi u_0 = \langle v, \mathbf{1} \rangle_\pi \mathbf{1}.$$

According to the above we have the asymptotic convergence rate $\|T^k - \Pi_0\| = \lambda_1^k$ for all $k \in \mathbb{N}$.

Projection onto stepfunctions Let us now consider some finite subdivision $A_1, ...A_n$ of $\mathbf{S}$, such that

$$\bigcup_{j=1}^{n} A_j = \mathbf{S} \qquad A_i \cap A_j = \emptyset \ \forall i \neq j, \tag{51}$$

with measurable sets A_j. Suppose $\mathbf{1}_A$ denotes the characteristic function of set A. We define the orthogonal projection Q onto the n dimensional space of step functions $D_n = \mathrm{span}\{\mathbf{1}_{A_1}, ..., \mathbf{1}_{A_n}\}$, i.e.

$$Qv = \sum_{i=1}^{n} \frac{\langle v, \mathbf{1}_{A_i} \rangle_\pi}{\pi(A_i)} \mathbf{1}_{A_i} = \sum_{i=1}^{n} \langle v, \phi_i \rangle_\pi \phi_i \tag{52}$$

with orthonormal basis $(\phi_i)_{i=1,\dots,n}$ of D_n

$$\phi_i = \frac{\mathbf{1}_{A_i}}{\sqrt{\pi(A_i)}}.$$

In short the orthogonal projection Q keeps the measure on the sets $A_1, .., A_n$, but on each of the sets the ensemble will be redistributed according to the invariant measure and the detailed information about the distribution inside of a set A_i is lost.

Since the sets $A_1, \dots, A_n$ form a full partition of $\mathbf{S}$ we have

$$Q\mathbf{1} = \mathbf{1}, \tag{53}$$

which implies

$$Q\Pi_0 = \Pi_0 Q = \Pi_0. \tag{54}$$

In order to measure the distance between the dominant eigenspace U and the stepfunction space D_n we define the *projection error* of the finite subdivision $A_1, ...A_n$ under consideration

$$\|Q^\perp u_j\| =: \delta_j \leq 1 \ \forall j, \qquad \delta(A) := \max_{j=1,\dots,m} \delta_j \tag{55}$$

where $Q^\perp = \mathrm{Id} - Q$ denotes the projection onto the orthogonal complement of D_n in $l^2(\pi)$.

Metastability. If we aggregate the original process $\{X_k\}_{k\in\mathbb{N}}$ with transfer operator T with respect to the finite subdivision $A_0, \dots, A_m$ of $\mathbf{S}$ with $m+1$ sets we get the aggregated transition matrix P with entries

$$P_{ij} = \mathbb{P}_\pi[X_1 \in A_j | X_0 \in A_i] = \frac{1}{\pi(A_i)} \sum_{x\in A_i} \pi(x)\mathbb{P}[X_1 \in A_j | X_0 = x]. \tag{56}$$

Let us define the metastability index of such a finite subdivision by the probability of remaining with the set A_i, i.e.,

$$\mathrm{Meta}(A_0, \dots, A_m) = \frac{1}{m+1} \sum_{i=0}^{m} P_{ii} = \frac{1}{m} \sum_{i=0}^{m} \mathbb{P}_\pi[X_1 \in A_i | X_0 \in A_i].$$

Now [15] gives the following theorem in which smallness of the projection error δ is related to the metastability index of a subdivision $A_0, \dots, A_m$ of state space into $m+1$ sets:

Theorem 6.8 *Let T be a self-adjoint transfer operator with properties as described above, in particular (43), (44), (48), and (49). The metastability*

of an arbitrary decomposition $A_0, \ldots, A_m$ of the state space is bounded from below and above by

$$1+(1-\delta_1^2)\lambda_1+\ldots+(1-\delta_m^2)\lambda_m+c \leq \sum_{j=0}^{m} \mathbb{P}_\pi[X_1 \in A_i | X_0 \in A_i] \leq 1+\lambda_1+\ldots+\lambda_m\,,$$

where $c = -r\,\left(\delta_1^2 + \ldots + \delta_m^2\right)$.

This result tells us that the minimization of δ with $m+1$ sets (for $m+1$ eigenvalues) corresponds to identifying the subdivision with maximum joint metastability.

6.4.1 MSM transfer operator

Now consider the projection of our transfer operator T onto D_n:

$$P = QTQ : l^2(\pi) \to D_n \subset l^2(\pi). \tag{57}$$

In order to understand the nature of P let us consider it as an operator on a finite dimensional space, $P : D_n \to D_n$. Here, as a linear operator on a finite-dimensional space, it has a matrix representation with respect to some basis. Let us take the basis $(\psi_i)_{i=1,\ldots,n}$ of probability densities given by

$$\psi_i = \frac{\mathbf{1}_{A_i}}{\pi(A_i)}. \tag{58}$$

By using the definition of T we get

$$\begin{aligned} P\psi_i &= QTQ\psi_i = QT\psi_i = \sum_{j=1}^{n} \frac{\langle T\psi_i, \mathbf{1}_{A_j}\rangle_\pi}{\pi(A_j)}\,\mathbf{1}_{A_j} = \sum_{j=1}^{n} \frac{\langle T\mathbf{1}_{A_i}, \mathbf{1}_{A_j}\rangle_\pi}{\pi(A_i)}\,\psi_j \\ &= \sum_{j=1}^{n} \Big(\int_{A_i} \mathbb{P}[X_1 \in A_j | X_0 = x]\,\pi(dx)\Big) \cdot \psi_j. \end{aligned}$$

such that

$$P\psi_i = \sum_{j=1}^{n} P_\pi[X_1 \in A_j | X_0 \in A_i] \cdot \psi_j. \tag{59}$$

So the transition matrix of the MSM Markov chain defined in Eq. (56) is identical with the matrix representation for the projected transfer operator P; therefore P is the MSM transfer operator. P inherits the ergodicity properties and the invariant measure of T as the following lemma shows:

Lemma 6.9 *For every $k \in \mathbb{N}$ we have*

$$\|P^k - \Pi_0\| \leq \|(TQ)^k - \Pi_0\| \leq \lambda_1^k.$$

Proof: Because of $\Pi_0 Q = Q\Pi_0 = \Pi_0$ and $\|T - \Pi_0\| = \lambda_1$ we have for $k = 1$:

$$\|TQ - \Pi_0\| = \|(T - \Pi_0)Q\| \leq \lambda_1. \tag{60}$$

Since furthermore $T\Pi_0 = \Pi_0$, and $T\Pi$ is self-adjoint we find for arbitrary $v \in l^2(\pi)$:

$$\begin{aligned} \Pi_0 T v &= \langle Tv, \mathbf{1}\rangle_\pi \mathbf{1} = \langle T\Pi v, \mathbf{1}\rangle_\pi \ \mathbf{1} + \langle T\Pi^\perp v, \Pi\mathbf{1}\rangle_\pi \, \mathbf{1} \\ &= \langle v, T\Pi\, \mathbf{1}\rangle_\pi \mathbf{1} + \langle \Pi T \Pi^\perp v, \mathbf{1}\rangle_\pi \mathbf{1} = \langle v, \mathbf{1}\rangle_\pi \mathbf{1} = \Pi_0 v, \end{aligned}$$

where the identity before the last follows from (49). Therefore

$$\Pi_0 T = T\Pi_0 = \Pi_0. \tag{61}$$

From this and $Q\Pi_0 = \Pi_0 Q = \Pi_0$ it follows that $(TQ - \Pi_0)^k = (TQ)^k - \Pi_0$ and thus with (60)

$$\|P^k - \Pi_0\| = \|Q(TQ)^k - Q\Pi_0\| \leq \|(TQ)^k - \Pi_0\| = \|(TQ - \Pi_0)^k\| \leq \|TQ - \Pi_0\|^k \leq \lambda_1^k,$$

which was the assertion. □

6.4.2 Approximation Error E

The approximation quality of the MSM Markov chain can be characterized by comparing the operators P^k and T^k for $k \in \mathbb{N}$ restricted to D_n:

$$E(k) = \|QT^kQ - P^k\| = \|QT^kQ - Q(TQ)^k\|. \tag{62}$$

Lemma 6.9 immediately implies that this error decays exponentially,

$$E(k) = \|QT^kQ - P^k\| \leq \|QT^kQ - \Pi_0\| + \|P^k - \Pi_0\| \leq \|Q(T^k - \Pi_0)Q\| + \|P^k - \Pi_0\| \leq 2\lambda_1^k, \tag{63}$$

independent of the choice of the sets $A_1, \ldots, A_n$. Since we want to understand how the choice of the sets and other parameters like the lag time τ influence the approximation quality we have to analyse the pre-factor in much more detail.

6.4.3 Main Result: An Upper Bound on E

The following theorem contains the main result of this article:

Theorem 6.10 *Let $T = T_\tau$ be a transfer operator for lag time $\tau > 0$ with properties as described above, in particular (43), (44), (48), and (49). Let the disjoint sets $A_1, ..., A_n$ form a full partition and define*

$$\|Q^\perp u_j\| =: \delta_j \leq 1 \ \forall j, \qquad \delta(A) := \max_{j=1,\ldots,m} \delta_j \tag{64}$$

where $Q^{\perp} = \mathrm{Id} - Q$ denotes the projection onto the orthogonal complement of D_n in L^2_π. Furthermore set

$$\eta(\tau) := \frac{r}{\lambda_1} = \exp(-\tau\Delta) < 1, \quad \text{with } \Delta > 0.$$

Then the error (62) is bounded from above by

$$E(k) \leq \min\Big[2\,;\, C(\delta(A), \eta(\tau), k)\Big] \cdot \lambda_1^k, \tag{65}$$

with a leading constant of following form

$$\begin{aligned} C(\delta, \eta, k) &= (m\delta + \eta)\Big[C_{\mathrm{sets}}(\delta, k) + C_{\mathrm{spec}}(\eta, k)\Big] && (66)\\ C_{\mathrm{sets}}(\delta, k) &= m^{1/2}(k-1)\,\delta && (67)\\ C_{\mathrm{spec}}(\eta, k) &= \frac{\eta}{1-\eta}(1-\eta^{k-1}). && (68) \end{aligned}$$

The proof can be found in Section **??**.

6.5 Interpretation and Observations

The theorem shows that the overall error can be made arbitrarily small by making the factor $[C_{\mathrm{sets}}(\delta, k) + C_{\mathrm{spec}}(\eta, k)]$ small. In order to understand the role of these two terms, consider for now the time of interest to be fixed at some $k \geq 2$. It can then be observed that:

1. The prefactor C_{sets} depends on the choice of the sets $A_1, \ldots, A_n$ only. It can be made smaller than any tolerance by choosing the sets appropriately and the number of sets, n large enough.

2. The prefactor C_{spec} is independent of the set definition and depends on the spectral gap Δ and the lag time τ only. While the spectral gap is given by the problem, the lag time may be chosen and thus C_{spec} can also be made smaller than any tolerance by choosing τ large enough. However, the factor C_{spec} will grow unboundedly for $\tau \to 0$ and $k \to \infty$, suggesting that using a large enough lagtime is essential to obtain an MSM with good approximation quality, even if the sets are well approximated.

If we are interested in a certain time scale, say T, then we have to set $k - 1 = T/\tau$. Then the pre-factor gets the form

$$C(\delta, \eta, T/\tau) = (m\delta + \eta)\Big[m^{1/2}\frac{T}{\tau}\,\delta + \frac{\eta}{1-\eta}F\Big], \quad F = 1 - \exp(-T\Delta),$$

in which the numbers m, and T have been chosen by the user, the spectral gap Δ is determined by the $(m+1)$st eigenvalue and thus, for given m,

a constant of the system. The error thus depends on the choice of the sets (*via* δ) and the lag time τ (and with it $\eta = \exp(-\tau\Delta)$). In the case where one is interested in the slowest process in the system, the time of interest may be directly related to the second eigenvalue *via* $T \propto 1/\Lambda_1$. For example, one may choose the half-life time of the decay of the slowest process: $T = \log 2/\Lambda_1$.

7 Markov jump processes

In the following two sections we will consider Markov processes on discrete state space but *continuous* in time. Many of the basic notions and definitions for discrete-time Markov chains will be repeated. We do this without too much reference to the former sections in order to make the following sections readable on their own.

7.1 Setting the scene

Consider some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where 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. A family $X = \{X(t) : t \geq 0\}$ of random variables $X(t) : \Omega \to \mathbf{S}$ is called a **continuous-time stochastic process** on the state space $\mathbf{S}$. The index t admits the convenient interpretation as time: if $X(t) = y$, the process is said to be in state y at time t. For some given $\omega \in \Omega$, the $\mathbf{S}$–valued set $\{X(t, \omega) : t \geq 0\}$ is called a **realization** (trajectory, sample path) of the stochastic process X associated with ω.

Definition 7.1 (Markov process) *A continuous-time stochastic process $\{X(t) : t \geq 0\}$ on a countable state space $\mathbf{S}$ is called a **Markov process**, if for any $t_{k+1} > t_k > \ldots > t_0$ and $B \subset \mathbf{S}$ the **Markov property***

$$\mathbb{P}[X(t_{k+1}) \in B | X(t_k), \ldots, X(t_0)] = \mathbb{P}[X(t_{k+1}) \in B | X(t_k)] \tag{69}$$

*holds. If, moreover, the right hand side of (69) does only depend on the time increment $t_{k+1} - t_k$, but not on t_k, then the Markov process is called **homogeneous**. Given a homogeneous Markov process, the function $p : \mathbb{R}^+ \times \mathbf{S} \times \mathbf{S} \to \mathbb{R}^+$ defined by*

$$p(t, x, y) = \mathbb{P}[X(t) = y | X(0) = x]$$

*is called the **stochastic transition function**; its values $p(t, y, z)$ are the (conditional) transition probabilities to move from x to y within time t. 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**.*

In the following, we will focus on homogeneous Markov process, and thus the term Markov process will always refer to a homogeneous Markov process, unless otherwise stated.

There are some subtleties in the realm of continuous time processes that are not present in the discrete-time case. These steam from the fact that

Figure 19: Illustration of a realization of a Markov jump process.

the uncountable union of measurable sets need not be measurable anymore. For example, the mapping $X(t,\cdot):\Omega\to\mathbf{S}$ is measurable for every $t\in\mathbb{R}^+$, i.e., $\{\omega\in\Omega: X(t,\omega)\in A\}\in\mathcal{A}$ for each given time t for every measurable subset $A\subset\mathbf{S}$. However,

$$\{\omega\in\Omega: X(t,\omega)\in A,\forall t\in\mathbb{R}^+\}=\bigcap_{t\in\mathbb{R}^+}\{\omega\in\Omega: X(t,\omega)\in A\}$$

need not be in $\mathcal{A}$ in general. This is related to functions like $\inf_{t\in\mathbb{R}^+}X(t)$ or $\sup_{t\in\mathbb{R}^+}X(t)$, since, e.g.,

$$\left\{\sup_{t\in\mathbb{R}^+}X(t)\le x\right\}=\bigcap_{t\in\mathbb{R}^+}\{\omega\in\Omega: X(t,\omega)\le x\}.$$

We will therefore impose some (quite natural) **regularity conditions** on the Markov process in order to exclude pathological cases (or too technical details). Throughout this chapter, we assume that

$$p(0,x,y)=\delta_{xy},\tag{70}$$

where $\delta_{xy}=1$, if $x=y$ and zero otherwise. This guarantees that no transition can take place at zero time. Moreover, we assume that the transition probabilities are continuous at $t=0$:

$$\lim_{t\to 0+}p(t,x,y)=\delta_{xy}\tag{71}$$

for every $x,y\in\mathbf{S}$. This guarantees (up to stochastic equivalence) that the realizations of $\{X(t): t\ge 0\}$ are right continuous functions (More precisely, it implies that the Markov process is stochastically continuous, separable and measurable on compact intervals. Moreover, there exists a separable version, being stochastically equivalent to $\{X(t): t\ge 0\}$ and all of whose realizations are continuous from the right; for details see reference [19, Chapt. 8.5]). Due to the fact that the state space is discrete, continuity from the right of the sampling functions implies that they are step functions, see Figure19 for an illustration, that is, for almost all $\omega\in\Omega$ and all $t\ge 0$ there exists $\Delta t(t,\omega)>0$ such that

$$X(t+\tau,\omega)=X(t,\omega);\quad \tau\in[0,\Delta t(t,\omega)).$$

This fact motivates the name **Markov jump process**.

For our further study, we recall to important random variables. A continuous random variable $\tau:\Omega\to\mathbb{R}^+$ satisfying

$$\mathbb{P}[\tau>s]=\exp(-\lambda s)$$

for every $s \geq 0$ is called an **exponential random variable** with parameter $\lambda \geq 0$. Its probability density $f : \mathbb{R}^+ \to \mathbb{R}^+$ is given by

$$f(s) = \lambda \exp(-\lambda s)$$

for $s \geq 0$ (and zero otherwise). Moreover, the expectation is given by

$$\mathbb{E}[\tau] = \frac{1}{\lambda}.$$

One of the most striking features of exponential random variables is their memoryless property, expressed as

$$\mathbb{P}[\tau > t + s | \tau > t] = \mathbb{P}[\tau > s]$$

for all $s, t \geq 0$. This is easily proven by noticing that the left hand side is per definition equal to $\exp(-\lambda(t+s))/\exp(-\lambda t) = \exp(-\lambda s)$, being equal to the right hand side.

A discrete random variable $N : \Omega \to \mathbb{N}$ with probability distribution

$$\mathbb{P}[N = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

for $k \in \mathbb{N}$ is called a **Poisson random variable** with parameter $\lambda \in \mathbb{R}^+$. Its expectation is given by

$$\mathbb{E}[N] = \lambda.$$

We now consider two examples of Markov jump processes that are of prototype nature.

Example 7.2 *Consider an iid. sequence $\{\tau_k\}_{k\in\mathbb{N}}$ of exponential random variable with parameter $\lambda > 0$ and define recursively the sequence of random variables $\{T_k\}_{k\in\mathbb{N}}$ by*

$$T_{k+1} = T_k + \tau_k$$

for $k \geq 1$ and $T_0 = 0$. Here, T_k is called the kth event time and τ_k the inter-event time. Then, the sequence of random variables $\{N(t) : t \geq 0\}$ are defined by

$$N(t) = \sum_{k=0}^{\infty} \mathbf{1}_{\{T_k \leq t\}} = \max\{k \geq 0 : T_k \leq t\}.$$

for $t \geq 0$ and with $N(0) = 0$. Its (discrete) distribution is given by the Poisson distribution with parameter λt:

$$\mathbb{P}[N(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t};$$

*for $k \geq 1$. That is why $\{N(t)\}$ is called a homogeneous **Poisson process** with intensity λ. Per construction, $N(t)$ is counting the number of events up to time t. Therefore, it is also sometime called the **counting process** associated with $\{T_k\}$.*

Remark: One could also determine the distribution of the event times T_k, the time at which the kth event happens to occur. Since per definition T_k is the sum of k iid. exponential random variables with parameter λ, the probability density is known as

$$f_{T_k}(t) = \frac{(\lambda t)^k}{k!}\lambda e^{-\lambda t}$$

for $t \geq 0$ and zero otherwise. This is the so-called Erlang distribution (a special type of Gamma distribution) with parameter $k+1$ and λ.

Example 7.3 *Consider some discrete-time Markov chain $\{E_k\}_{k\in\mathbb{N}}$ on a countable state space $\mathbf{S}$ with stochastic transition matrix $K = (k(x,y))_{xy\in\mathbf{S}}$, and furthermore, consider some homogeneous Poisson process $\{T_k\}_{k\in\mathbb{N}}$ on $\mathbb{R}^+$ with intensity $\lambda > 0$ and associated counting process $\{N(t) : t \geq 0\}$. Then, assuming independence of $\{E_k\}$ and $\{N(t)\}$, the process $\{X(t) : t \geq 0\}$ with*

$$X(t) = E_{N(t)}$$

*is called the **uniform Markov jump process** with clock $\{N(t)\}$ and subordinated chain $\{E_k\}$. The thus defined process is indeed a Markov jump process (exercise). Note that the jumps of $X(t)$ are events of the clock process $N(t)$, however, not every event of $N(t)$ corresponds to a jump (unless $E(x,x) = 0$ for all $x \in \mathbf{S}$). In order to compute its transition probabilities, note that*

$$\begin{aligned}
\mathbb{P}[X(t) = y | X(0) = x] &= \mathbb{P}[E_{N(t)} = y | E_0 = x] \\
&= \sum_{n=0}^{\infty} \mathbb{P}[E_n = y, N(t) = n | E_0 = x] \\
&= \sum_{n=0}^{\infty} \mathbb{P}[E_n = y | E_0 = x]\ \mathbb{P}[N(t) = n],
\end{aligned}$$

where the last equality is due to the assumed independence of E_n and $N(t)$. Hence, its transition probabilities are given by

$$p(t,x,y) = \sum_{n=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!}k^n(x,y), \tag{72}$$

for $t \geq 0$, $x, y \in \mathbf{S}$, where $k^n(x,y)$ is the corresponding entry of K^n, the n-step transition matrix of the subordinated Markov chain.

Solely based on the Markov property, we will now deduce some properties of Markov jump processes that illuminate the differences, but also the close relations to the realm of discrete-time Markov chains. We will see that the uniform Markov chain is in some sense *the* prototype of a Markov jump process. To do so, define for $t \in \mathbb{R}^+$ the **residual life time** $\tau(t) : \mathbf{S} \to [0, \infty]$ in state $X(t)$ by

$$\tau(t) = \inf\{s > 0 : X(t+s) \neq X(t)\}. \tag{73}$$

Obviously, $\tau(t)$ is a stopping time, i.e., it can be expressed in terms of $\{X(s) : 0 \leq s \leq t\}$ (since the sample paths are right-continuous). Hence, conditioned on $X(t)$ and $\tau(t) < \infty$, the next jump (state change) will occur at time $t + \tau(t)$. Otherwise the Markov process will not leave the state $X(t)$ anymore.

Proposition 7.4 *Consider some Markov process $\{X(t) : t \geq 0\}$ being in state $x \in \mathbf{S}$ at time $t \in \mathbb{R}^+$. Then, there exists $\lambda(x) \geq 0$, independent of the time t, such that*

$$\mathbb{P}[\tau(t) > s | X(t) = x] = \exp(-\lambda(x)s) \tag{74}$$

for every $s > 0$.

Therefore, $\lambda(x)$ is called the **jump rate** associated with the state $x \in \mathbf{S}$. Prop. 7.4 states that the residual life time decays exponentially in s.

Proof: Note that $\mathbb{P}[\tau(t) > s | X(t) = x] = \mathbb{P}[\tau(0) > s | X(0) = x]$, since the Markov jump process is homogeneous. Define $g(s) = \mathbb{P}[\tau(0) > s | X(0) = x]$ and compute

$$\begin{aligned} g(t+s) &= \mathbb{P}[\tau(0) > t+s | X(0) = x] = \mathbb{P}[\tau(0) > t, \tau(t) > s | X(0) = x] \\ &= \mathbb{P}[\tau(0) > t | X(0) = x] \, \mathbb{P}[\tau(t) > s | \tau(0) > t, X(0) = x] \\ &= g(t) \, \mathbb{P}[\tau(t) > s | \tau(0) > t, X(0) = x, X(t) = x] \\ &= g(t) \, \mathbb{P}[\tau(t) > s | X(t) = x] = g(t)g(s). \end{aligned}$$

In addition, $g(s)$ is continuous at $s = 0$, since the transition probabilities were assumed to be continuous at zero. Moreover, $0 \leq g(s) \leq 1$, which finally implies that the only solution must be

$$g(s) = \exp(-\lambda(x)s)$$

with $\lambda(x) \in [0, \infty]$ given by $\lambda(x) = -\ln(\mathbb{P}[\tau(0) > 1 | X(0) = x])$. □

To further illuminate the characteristics of Markov processes on countable state spaces, denote by $T_0 = 0 < T_1 < T_2 < \ldots$ the random **jump**

times or event times, at which the Markov process changes its state. Based on the jump times, define the sequence of random **life times** $(\tau_k)_{k\in\mathbb{N}}$ via the relation

$$\tau_k = T_{k+1} - T_k$$

for $k \in \mathbb{N}$. Due to Prop. 7.4 we know that

$$\mathbb{P}[\tau_k > s | X(T_k) = x] = \exp(-\lambda(x)s)$$

for $s \geq 0$. Moreover, the average life time of a state is

$$\mathbb{E}[\tau(t)|X(t) = x] = \frac{1}{\lambda(x)}.$$

In terms of the jump times, we have

$$X(t) = X(T_k); \quad t \in [T_k, T_{k+1}),$$

hence the Markov process is constant, except for the jumps.

Definition 7.5 *Consider a state $x \in \mathbf{S}$ with associated jump rate $\lambda(x)$. Then, x is called*

1. ***permanent**, if $\lambda(x) = 0$*
2. ***stable**, if $0 < \lambda(x) < \infty$,*
3. ***instantaneous**, if $\lambda(x) = \infty$ (not present for Markov processes with right continuous sample paths).*

Assume that $X(t) = x$ at time t; if x is

1. permanent, then $\mathbb{P}[X(s) = x|X(t) = x] = 1$ for every $s > t$, hence the Markov process stays in x forever,
2. stable, then $\mathbb{P}[0 < \tau(t) < \infty|X(t) = x] = 1$,
3. instantaneous, then $\mathbb{P}[\tau(t) = 0|X(t) = x] = 1$, hence the Markov process exists the state as soon as it enters it.

Due to our general assumption, we know that the Markov process has right continuous sample paths. As a consequence, the state space $\mathbf{S}$ does not contain instantaneous states.

Definition 7.6 *Consider some Markov process $\{X(t) : t \geq 0\}$ with right continuous sample paths. Then, the Markov process is called **regular** or **non-explosive**, if*

$$T_\infty := \sup_{k\in\mathbb{N}} T_k = \infty \quad (a.s.),$$

*where $T_0 < T_1 < \ldots$ denote the jump times of $\{X(t) : t \geq 0\}$. The random variable T_∞ is called the **explosion time**.*

If the Markov jump process is explosive, then $\mathbb{P}[T_\infty < \infty] > 0$. Hence there is a "substantial" set of realizations, for which the Markov process "blows up" in finite time. In such a situation, we assume that the Markov jump process is only defined for times smaller than the explosion time. The following proposition provides a quite general condition for a Markov process to be regular.

Proposition 7.7 *A Markov process $\{X(t) : t \geq 0\}$ on a countable state space $\mathbf{S}$ is regular, if and only if*

$$\sum_{k=1}^{\infty} \frac{1}{\lambda(X(T_k))} = \infty \quad (a.s.).$$

This is particularly the case, if (1) $\mathbf{S}$ is finite, or (2) if $\sup_{x \in \mathbf{S}} \lambda(x) < \infty$.

Proof: See Prop. 8.7.2 in [19]. □

Compare also Prop. 7.16. Based on the sequence of event times $(T_k)_{k \in \mathbb{N}}$, we define the following discrete-time $\mathbf{S}$-valued stochastic process $\{E_k\}$ by $E_k = X(T_k)$. As is motivated from the definition and the following results, $\{E_k\}_{k \in \mathbb{N}}$ is called the embedded Markov chain. However, it still remains to prove that $\{E_k\}_{k \in \mathbb{N}}$ is really well-defined and Markov. To do so, we need the following

Definition 7.8 *A Markov process $\{X(t) : t \geq 0\}$ on a state space $\mathbf{S}$ fulfills the **strong Markov property** if, for any stopping time τ, being finite a.s.,*

$$\mathbb{P}[X(s+\tau) \in A | X(\tau) = x, X(t), t < \tau] = \mathbb{P}_x[X(s) \in A]$$

for every $A \in \mathcal{A}$, whenever both sides are well-defined. Hence, the process $\{X(s+\tau) : s \geq 0\}$ is Markov and independent of $\{X(t), t < \tau\}$, given $X(\tau) = x$.

In contrast to the discrete-time case, not every continuous-time Markov process on a countable state space obeys the strong Markov property. However, under some suitable regularity conditions, this is true.

Theorem 7.9 *A regular Markov process on a countable state space fulfills the strong Markov property.*

Proof: See Thm. 4.1 in Chapter 8 of [2]. □

The next proposition states that the time, at which the Markov process jumps next, and the state, it jumps into, are independent.

Proposition 7.10 *Consider a regular Markov jump process on* $\mathbf{S}$*, and assume that* $T_{k+1} < \infty$ *a.s. Then, conditioned on* $X(T_k) = x$*, the random variables* τ_{k+1} *and* $X(T_{k+1})$ *are independent, i.e.,*

$$\begin{aligned} &\mathbb{P}[\tau_{k+1} > t, X(T_{k+1}) = y | X(T_k) = x] \\ &\quad = \mathbb{P}[\tau_{k+1} > t | X(T_k) = x] \cdot \mathbb{P}[X(T_{k+1}) = y | X(T_k) = x] \end{aligned} \tag{75}$$

Proof: Starting with (75) we get, by applying Bayes rule,

$$\begin{aligned} &\mathbb{P}[\tau_{k+1} > t, X(T_{k+1}) = y | X(T_k) = x] \\ &\quad = \mathbb{P}[\tau_{k+1} > t | X(T_k) = x] \cdot \mathbb{P}[X(T_{k+1}) = y | X(T_k) = x, \tau_{k+1} > t]. \end{aligned}$$

Using the Markov property we can rewrite he last factor

$$\begin{aligned} &\mathbb{P}[X(T_{k+1}) = y | X(T_k) = x, \tau_{k+1} > t] \\ &\quad = \mathbb{P}[X(T_{k+1}) = y, X(s) = x, T_k \le s < T_{k+1} | X(T_k + t) = x] \\ &\quad = \mathbb{P}[X(T_k + \tau(T_k + t)) = y, X(s) = x, T_k \le s < T_k + \tau(T_k + t) | X(T_k) = x] \\ &\quad = \mathbb{P}[X(T_{k+1}) = y | X(T_k) = x], \end{aligned}$$

where we used the homogeneity of the Markov process to proceed from the second to the third line. □

We are now ready to define the **embedded Markov chain** of the Markov jump process.

Definition 7.11 *Define the homogeneous Markov chain* $\{E_k\}_{k \in \mathbb{N}}$ *on the state space* $\mathbf{S}$ *in terms of the following transition function* $P = \big(p(x,y)\big)_{xy \in \mathbf{S}}$*. If* x *is permanent, set* $p(x,x) = 1$*. Otherwise, if* x *is stable, set*

$$p(x,y) = \mathbb{P}[X(T_1) = y | X(0) = x] \tag{76}$$

and consequently $p(x,x) = 0$*.*

Summarizing our results we obtain the following theorem.

Theorem 7.12 *Consider a regular Markov jump process and assume that the state space consists only of stable states. Then,* $\{X(T_k)\}_{k \in \mathbb{N}}$ *is a homogeneous Markov chain with transition function defined in (76). In other words, it is*

$$E_k = X(T_k)$$

for every $k \in \mathbb{N}$ *(in distribution).*

So, we obtain the following characterization of a Markov jump processes $\{X(t)\}$ on the state space $\mathbf{S}$. Assume that the process is at state x at time t, i.e., $X(t) = x$. If the state is permanent, then the Markov process will

stay in x forever, i.e., $X(s) = x$ for all $s > t$. If the state is stable, then the Markov process will leave the state x at a (random) time being exponentially distributed with parameter $0 < \lambda(x) < \infty$. It then jumps into some other state $y \neq x \in X$ with probability $p(x, y)$, hence according to the law of the embedded Markov chain $\{E_k\}_{k\in\mathbb{N}}$. Therefore knowing the rates $(\lambda(x))$ and the embedded Markov chain in terms of its transition function $P = (p(x, y))$ completely characterizes the Markov jump process.

The above characterization is also the basis for a numerical simulation of the Markov jump process. To do so, one might exploit the following important and well-known relation between an exponential random variable τ with parameter λ and some uniform random variable U on $[0, 1]$, given by

$$\tau = -\frac{1}{\lambda}\ln(U).$$

Hence, a numerical simulation of a Markov jump process can be based on randomly drawing two uniform random numbers for each jump event (one for the time, another one for the state change).

7.2 Communication and recurrence

This section is about the topology of *regular* Markov jump processes (unless stated otherwise). As in the case of Markov chains, we start with some

Definition 7.13 *Let $\{X(t) : t \geq 0\}$ denote a Markov process with transition function $P(t)$, 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*

$$\mathbb{P}[X(t) = y | X(0) = x] > 0$$

for some $t > 0$.

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 is said to be **irreducible**, if all pairs of states communicate.*

As for Markov chains, it can be proven that the communication relation $\leftrightarrow$ is an equivalence relation on the state space. We remark that periodicity plays no role for continuous-time Markov jump processes, since they are always **aperiodic**. We proceed by introducing the first return time.

Definition 7.14 1. *The stopping time $E_x : \Omega \to \mathbb{R}^+ \cup \{\infty\}$ defined by*

$$E_x = \inf\{t \geq 0 : X(t) \neq x, X(0) = x\}.$$

*is called the **first escape time** from state $x \in \mathbf{S}$.*

2. The stopping time $R_x : \Omega \to \mathbb{R}^+ \cup \{\infty\}$ *defined by*

$$R_x \quad = \quad \inf\{t > T_1 : X(t) = x\},$$

with $\inf\{\} = \infty$*, is called the* ***first return time*** *to state* x*.*

Note that $\mathbb{P}_x[E_x] = \mathbb{P}_x[\tau(0)]$ (see eq. (73)).

Analogous to the Markov chain theory, based on the first return time to a state, we may define recurrence and transience of a state.

Definition 7.15 *A state* $x \in \mathbf{S}$ *is called* ***recurrent****, if it is permanent or*

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

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

Again, recurrence and transience are class properties, i.e., the states of some communication class are either all recurrent or all transient. Interestingly and maybe not surprisingly, some (but not all, as we will see later) properties of states can be determined in terms of the embedded Markov chain.

Proposition 7.16 *Consider a regular and stable Markov jump process* $\{X(t) : t \geq 0\}$ *and the associated embedded Markov chain* $\{E_k\}_{k\in\mathbb{N}}$*, then the following holds true.*

a) The Markov jump process is irreducible, if and only if its embedded Markov chain is irreducible.

b) A state $x \in \mathbf{S}$ *is recurrent (transient) for the embedded Markov chain, if and only if is recurrent (transient) for the Markov jump process.*

c) A state $x \in \mathbf{S}$ *is recurrent for the Markov jump process, if and only if*

$$\int_0^\infty p(t, x, x)dt = \infty.$$

d) Recurrence and transience of the Markov process inherits to any discretization, i.e. if $h > 0$ *and* $Z_k := X(kh)$ *then recurrence of* $x \in \mathbf{S}$ *for the Markov process is equivalent to recurrence of* $x \in \mathbf{S}$ *for the discretization* $\{Z_k\}_{k\in\mathbb{N}}$*.*

Proof: We leave the proof of the first two statements as an exercise to the reader.

c) Remember the analogous formulation in the time-discrete case: if for some Markov chain, e.g. $\{E_k\}_{k\in\mathbb{N}}$, and some state, e.g. $x \in \mathbf{S}$, the random

variable N_x counts the number of visits in x, then x is recurrent if and only if

$$\begin{aligned}\mathbb{E}_x[N_x] &= \mathbb{E}_x\left[\sum_{k=0}^{\infty} \mathbb{1}_{E_k=x}\right] = \sum_{k=0}^{\infty} \mathbb{E}_x[\mathbb{1}_{E_k=x}] \\ &= \sum_{k=0}^{\infty} p^{(k)}(x,x) = \infty,\end{aligned}$$

where, as usual, $p^{(k)}(x,x)$ denotes the $k-$step transition probability $\mathbb{P}_x[E_k = x]$. Therefore we can prove the statement by showing that

$$\int_0^{\infty} p(t,x,x)dt = \frac{1}{\lambda(x)} \sum_{k=0}^{\infty} p^{(k)}(x,x).$$

This is done in the following, where we use Fubini's theorem to exchange integral and expectation and Beppo Levi's theorem to exchange summation and expectation:

$$\begin{aligned}\int_0^{\infty} p(t,x,x)dt &= \int_0^{\infty} \mathbb{E}_x[\mathbb{1}_{X(t)=x}]dt = \mathbb{E}_x\left[\int_0^{\infty} \mathbb{1}_{X(t)=x}dt\right] \\ &= \mathbb{E}_x\left[\sum_{k=0}^{\infty} \tau_{k+1}\mathbb{1}_{E_k=x}\right] = \sum_{k=0}^{\infty} \mathbb{E}_x[\tau_{k+1}|E_k = x]\,\mathbb{P}_x[E_k = x] \\ &= \sum_{k=0}^{\infty} \frac{1}{\lambda(x)} p^{(k)}(x,x)\end{aligned}$$

Be aware that the conditions to use Fubini's theorem are only met because X is a jump process.
d) That transience inherits to any discretization is obvious, so consider x recurrent. If t is constrained by $kh \le t < (k+1)h$, then

$$\begin{aligned}p((k+1)h,x,x) &\ge p((k+1)h - t,x,x)p(t,x,x) \\ &\ge \exp(-\lambda(x)((k+1)h - t)p(t,x,x) \\ &\ge \exp(-\lambda(x)h)p(t,x,x).\end{aligned}$$

Multiplication with $\exp(\lambda(x)h)$ yields

$$\exp(\lambda(x)h)p((k+1)h,x,x) \ge p(t,x,x) \text{ , for } kh \le t < (k+1)h.$$

This enables us to give an upperbound to the integral

$$\begin{aligned}\int_0^{\infty} p(t,x,x)dt &\le h\sum_{k=0}^{\infty} \exp(-\lambda(x)h)p((k+1)h,x,x) \\ &= h\exp(-\lambda(x)h)\sum_{k=1}^{\infty} p(kh,x,x).\end{aligned}$$

It follows from d) that $\sum_{k=1}^{\infty} p(kh, x, x) = \infty$, which is the sum over the transition probabilities of the discretized Markov chain.
□

Irreducible and recurrent Markov jump processes are regular, as the next theorem states.

Theorem 7.17 *An irreducible and recurrent Markov jump process is regular.*

Proof: Regularity means that the sequence of event times heads to infinity

$$\lim_{k\to\infty} T_k = \infty \Leftrightarrow \sum_{k=1}^{\infty} \tau_k = \infty.$$

Let $x \in \mathbf{S}$ be an arbitrary start position. Since the Markov process is irreducible and recurrent, we know that the embedded Markov chain $\{E_k\}_{k\in\mathbb{N}}$ visits x infinitely often. Denote by $\{N_k(x)\}_{k\in\mathbb{N}}$ the sequence of visits in x. Observe that if τ is $\lambda-$exponential distributed, then $\lambda\tau$ is $1-$exponential distributed (we pose that as an easy exercise), therefore we have

$$\begin{aligned} \infty &= \sum_{k=0}^{\infty} \lambda(E_{N_k(x)})\tau_{N_k+1} = \lambda(x) \sum_{k=0}^{\infty} \tau_{N_k+1} \\ &\leq \lambda(x) \sum_{k=0}^{\infty} \tau_{k+1}. \end{aligned}$$

□

As we will see below, it is also possible to characterize invariant measures in terms of the embedded Markov chain. However, the distinction between positive and null-recurrence and the existence of stationary distributions can not be examined in terms of the embedded Markov chain. Here, the rates $(\lambda(x))$ also have to come into play. That is why we postpone the corresponding analysis and first introduce the concept of infinitesimal generators, which is the more adequate object to study.

7.3 Infinitesimal generators and the master equation

We now come to the characterization of Markov jump processes that is not present for the discrete-time case. It is in terms of infinitesimal changes of the transition probabilities and based on the notion of generators. As in the preceding section, we assume throughout that the Markov jump process satisfies the two regularity conditions (70) and (71).

To start with, we introduce the **transition semigroup** $\{P(t) : t \geq 0\}$ with

$$P(t) = \big(p(t,x,y)\big)_{xy\in\mathbf{S}}.$$

Due to (70), it is $P(0) = \mathrm{Id}$ and due to (71), we have

$$\lim_{t\to 0+} P(t) = \mathrm{Id}.$$

In terms of the transition semigroup, we can also easily express the **Chapman-Kolmogorov equation** as

$$P(s+t) = P(s)P(t)$$

for $t, s \geq 0$ (which justifies the notion of a semigroup). In semigroup theory, one aims at characterizing $P(t)$ in terms of its infinitesimal generator Q. In broad terms, the goal is to prove and justify the notion $P(t) = \exp(tQ)$. In the following, we will proceed towards this goal.

Proposition 7.18 *Consider the semigroup $P(t)$ of a Markov jump process. Then, the limit*

$$A = \lim_{t\to 0+} \frac{P(t) - \mathrm{Id}}{t}$$

*exists (entrywise) and defines the **infinitesimal generator** $A = (a(x,y))_{xy\in\mathbf{S}}$ with $-\infty \leq a(x,x) \leq 0 \leq a(x,y) < \infty$.*

Note that we do not claim uniform convergence for all pairs of states $x, y \in \mathbf{S}$.

Proof: We first prove the result for the diagonal entries. Consider some state $x \in \mathbf{S}$ and define $h(t,x) = -\ln(p(t,x,x))$. Then, from the Chapman-Kolmogorov equation we deduce $p(t+s,x,x) \geq p(t,x,x)p(s,x,x)$. In terms of h, this implies $h(t+s,x) \leq h(t,x)h(s,x)$. Due to the general regularity condition (70), it is $h(0,x) = 0$, implying $h(t,x) \geq 0$ for all $t \geq 0$. Now, define

$$\sup_{0\leq t\leq\infty} \frac{h(t,x)}{t} =: c \in [0,\infty].$$

We now proof that c is in fact equal to the limit $\lim h(t,x)/t$ for $t \to +$, being equivalent to the statement that for every $b < c$ it is

$$b \leq \liminf_{t\to 0+} \frac{h(t,x)}{t} \leq \limsup_{t\to 0+} \frac{h(t,x)}{t} = c. \tag{77}$$

So, choose $b < c$ arbitrarily. Acc. to the definition of c, there exists $s > 0$ such that $b < h(s,x,x)/s$. Rewriting $s = nt + \Delta t$ with $t > 0$ and $0 \leq \Delta t < t$, we obtain

$$b < \frac{h(s,x)}{s} \leq \frac{nt}{s}\frac{h(t,x)}{t} + \frac{h(\Delta t,x)}{s}$$

Taking the joint limit $t \to 0+$, $\Delta t \to 0$ and $n \to \infty$ such that $nt/s \to 1$ proves the first inequality in (77) and thus completes the proof for the diagonal entries.

To prove the statement for the off-diagonal entries, assume that we are able to prove for every $\epsilon \in (1/2, 1)$ there exist $\delta > 0$ such that

$$p(ns, x, y) \geq (2\epsilon - 1)np(s, x, y) \tag{78}$$

for every $n \in \mathbb{N}$ and $s \geq 0$ such that $0 \leq ns < \delta$. Denote by $[x]$ the integer part of x. Then,

$$\frac{p(s, x, y)}{s} \leq \frac{p([t/s]s, x, y)}{[t/s]s(2\epsilon^2 - \epsilon)}$$

with $t, s < \delta$. Considering $s \to 0+$, we obtain for all $t > 0$

$$\limsup_{s \to 0+} \frac{p(s, x, y)}{s} \leq \frac{p(t, x, y)}{t(2\epsilon^2 - \epsilon)} < \infty$$

since $lim_{s \to 0+}[t/s]s = t$. Therefore,

$$\limsup_{s \to 0+} \frac{p(s, x, y)}{s} \leq \frac{1}{(2\epsilon^2 - \epsilon)} \liminf_{t \to 0+} \frac{p(t, x, y)}{t} < \infty.$$

Since ϵ can be chosen arbitrarily close to 1, we finally obtain the desired result. However, statement (78) still needs to be proven ... □

Sometimes, even for discrete-time Markov chains "generators" are defined; here $A = P - \mathrm{Id}$ mimics the properties of a infinitesimal generator (which it of course not is). In Graph Theory, such a matrix is known as **Laplace matrix**.

Example 7.19 *Rewriting eq. (72) in matrix form, we obtain for the transition semigroup of the uniform Markov process with intensity $\lambda > 0$ and subordinated Markov chain $K = (k(x, y))$*

$$P(t) = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} K^n = e^{t\lambda\left(K - \mathrm{Id}\right)}, \tag{79}$$

for $t \geq 0$. The infinitesimal generator is thus given by

$$A = \lambda\big(K - \mathrm{Id}\big), \tag{80}$$

which, entry-wise, corresponds to $a(x, x) = \lambda(1 - k(x, x))$, and $a(x, y) = \lambda k(x, y)$ for $x \neq y$. The result directly follows from eq. (79).

By now we have two different descriptions of a Markov jump process, one in of form sojourn times and the embedded Markov chain, the other

by the generator. We saw in the preceding sections that a Markov jump process is fully determined by its sojourn times and the embedded Markov chain. Why did we introduce the notion of a generator then? The answer is that more general Markov processes, i.e. in continuous state space, can not described by an embedded Markov chain anymore, while it is still possible to use the generator concept. But of course, in the case of a Markov jump process, it is possible to convey both description types into each other, like we did in the case of a uniform Markov jump process. This is what we do in the next paragraph. Therefore we will construct for a *given* generator a suitable Markov jump process and use this construction afterwards to get insight about the entries of a generator of a *given* Markov jump process.
As preparation we need

Proposition 7.20 *Consider a series of independent exponential distributed random variables* $\{N_k\}_{k\in\mathbb{N}}$ *with parameters* $\{\lambda_k\}_{k\in\mathbb{N}}$. *Assume that* $\sum_{k=0}^{\infty}\lambda_k = \lambda < \infty$ *and let*

$$T = \min\{N_0, N_1, N_2, \dots,\}$$

be the minimum value of the random variables series and J *such that* $N_J = T$. *Then* J *and* T *are independent random variables with*

$$\mathbb{P}[J = i, T \geq t] = \mathbb{P}[J = i]\,\mathbb{P}[T \geq t] = \frac{\lambda_i}{\lambda}\exp(-\lambda t).$$

Proof: Left as an exercise (use results about the distribution of a minimum of exponential distributed random variables, show the proposition for a finite number of random variables and then generalize to the case of an infinite number of random variables). □

Given a stable and conservative generator, i.e. a matrix A with

$$-\infty < -a(x,x) \leq 0,$$
$$0 \leq a(x,y) < \infty \text{ for } x \neq y$$

and $\sum_{y\neq x} a(x,y) = -a(x,x)$. To construct a jump process based on this matrix set $T_0 = 0$, $X(T_0) = X(0) = x_0$ and define recursively the following process

1. Assume $X(T_k) = x$.
2. If $a(x,x) = 0$ end the construction by setting $\tau_k = \infty$ and $X(t) = x$ for all $t \geq T_k$.
3. Otherwise $\tau_k = \min\{N_{x,0}, N_{x,1}, \dots, N_{x,x-1}, N_{x,x+1}, N_{x,x+2}, \dots\}$, where $N(x,y)$ is exponential distributed to the parameter $a(x,y)$.
4. Set $T_{k+1} = T_k + \tau_k$ and $X_{k+1} = X(T_{k+1}) = y$, where y is the state such that $\tau_k = N(x,y)$.

Theorem 7.21 *The previous constructed process is a homogeneous Markov jump process with generator A*

Proof: We leave the fact that the constructed process is a homogeneous Markov jump process to the careful reasoning of the reader. It remains to show that $a(x,y) = \lim_{t\to 0} \frac{P(t,x,y)-P(0,x,y)}{t}$, i.e. it is necessary to analyze the transition function P. The statement is trivial if $a(x,x) = 0$ (why?), therefore it is assumed in the following that $a(x,x) \neq 0$.
The first case to be considered is $x \neq y$, then

$$\begin{aligned} P(t,x,y) &= \mathbb{P}_x[T_2 \leq t, X(t) = y] + \mathbb{P}_x[T_2 > t, X(t) = y] \\ &= \mathbb{P}_x[T_2 \leq t, X(t) = y] + \mathbb{P}_x[T_2 > t, T_1 \leq t, X_1 = y] \\ &= \mathbb{P}_x[T_2 \leq t, X(t) = y] + \mathbb{P}_x[T_1 \leq t, X_1 = y] - \mathbb{P}_x[T_1 \leq t, X_1 = y, T_2 \leq t] \end{aligned}$$

The three terms on the right-hand side are now to be analyzed separately. For the second term we have, by 7.20,

$$\mathbb{P}_x[T_1 \leq t, X_1 = y] = (1 - \exp(-a(x,x)t)\frac{a(x,y)}{a(x,x)} =: f(t).$$

Since $f'(t) = a(x,y)\exp(a(x,x)t)$ we have $f'(0) = a(x,y)$ and

$$\lim_{t\to 0} \frac{f(t) - f(0)}{t} = \lim_{t\to 0} \frac{f(t)}{t} = a(x,y).$$

The first and the third term are both upper-bounded by $\mathbb{P}_x[T_2 \leq t]$, which can be bounded further by

$$\begin{aligned} \mathbb{P}_x[T_2 \leq t] &\leq \mathbb{P}_x[T_1 \leq t, \tau_1 \leq t] \\ &= \sum_{x\neq y} \mathbb{P}_x[T_1 \leq t, X_1 = y, \tau_1 \leq t] \\ &= \sum_{x\neq y} (1 - \exp(a(x,x)t)) \frac{a(x,x)}{-a(x,y)} (1 - \exp(a(y,y)t)) = f(t)g(t), \end{aligned}$$

where $f(t) := \exp(a(y,y)t) - 1$ and $g(t) := \sum_{x\neq y}(1 - \exp(a(x,x)t)\frac{a(x,x)}{a(x,y)}$. Now observe

$$\lim_{t\to 0} \frac{1}{t} f(t) = \lim_{t\to 0} \frac{\exp(a(y,y)t) - \exp(a(y,y)0)}{t - 0} = (\exp(a(y,y)t))'|_{t=0} = a(y,y)$$

and

$$\lim_{t\to 0} g(t) = \lim_{t\to 0} \sum_{x\neq y} (1 - \exp(a(x,x)t) \frac{a(x,x)}{a(x,y)} = 0,$$

(the exchange of limes and summation is allowed in this case, because all summands in the sum are positive, bounded and with an existing limes). This yields $\lim_{t\to 0} \frac{\mathbb{P}_x[T_2 \leq t]}{t} = \lim_{t\to 0} \frac{f(t)g(t)}{t} = 0$ and, putting it all together,

$$\lim_{t\to 0} \frac{P(t,x,y)}{t} = a(x,y).$$

It remains to show that $\lim_{t\to 0} \frac{P(t,x,x)-1}{t} = -a(x,x)$. This is very similar to the case $x \neq y$, in that $P(t,x,x)$ is decomposed by

$$\begin{aligned} P(t,x,x) &= \mathbb{P}_x[T_2 \leq t, X(t) = x] + \mathbb{P}_x[T_2 > t, X(t) = x] \\ &= \mathbb{P}_x[T_2 \leq t, X(t) = x] + \mathbb{P}_x[T_1 > t], \end{aligned}$$

which can be treated similar to the first case to show the assertion. □

As it is clear by now how to construct a Markov jump process to a given generator we proceed to the reverse direction. Assume a given Markov jump process $\{X(t) : t \geq 0\}$ with jump rates $\{\lambda(x)\}_{x\in\mathbf{S}}$ and conditional transition probabilities $\{k(x,y)\}_{x,y\in\mathbf{S}}$, furthermore $k(x,x) = 0$ and $\lambda(x) < \infty$ for all $x \in \mathbf{S}$. Let P be the transition function of this process. A matrix $\tilde{A}$ defined by

$$\tilde{a}(x,y) = \begin{cases} -\lambda(x) \text{ for } x = y \\ \lambda(x)k(x,y) \text{ for } x \neq y \end{cases}$$

fulfills obviously the conditions we posed on a generator, namely conservative and stable, to construct a Markov Process by the previous described procedure. Doing this we obtain another Markov Process with transition function $\tilde{P}$. By construction it is clear that $P = \tilde{P}$ (you should be able to figure that out!) and therefore the derivatives at 0 are equal, that is the generator of $\{X(t) : t \geq 0\}$ fulfills $A = \tilde{A}$. We state this important result in

Theorem 7.22 *Consider a homogeneous and regular Markov jump process on state space* $\mathbf{S}$ *with jump rates* $\{\lambda(x)\}_{x\in\mathbf{S}}$ *and conditional transition probabilities* $\{k(x,y)\}_{x,y\in\mathbf{S}}$*, where* $k(x,x) = 0$ *for all* $x \in \mathbf{S}$*. Then the generator* A *is given by*

$$\tilde{a}(x,y) = \begin{cases} -\lambda(x) \text{ for } x = y \\ \lambda(x)k(x,y) \text{ for } x \neq y \end{cases},$$

i.e. $A = \Lambda(K - Id)$*, where the* ***jump rate matrix*** Λ *is given by*

$$\Lambda = \begin{pmatrix} \lambda(0) & & & & \\ & \lambda(1) & & & \\ & & \lambda(2) & & \\ & & & \ddots & \end{pmatrix}. \tag{81}$$

Hence, the negative diagonal entry of the generator corresponds to the life time rate of the corresponding state, while the off-diagonal entries are proportional to the transition probabilities of the embedded Markov chain. We further remark that many infinitesimal generators can be represent in multiple ways, if represented by an intensity $\lambda \in \mathbb{R}^+$ and some subordinated Markov chain with transition matrix S. Assume $\sup_x \lambda(x) < \infty$. Then, for any choice of $\lambda \geq \sup_x \lambda(x)$, define

$$S = \lambda^{-1}\Lambda(K - \mathrm{Id}) + \mathrm{Id},$$

which indeed is a stochastic matrix (exercise). As a result, eq. (80) transforms into the **uniform representation**

$$A = \lambda(S - \mathrm{Id}), \tag{82}$$

which is the representation of the infinitesimal generator of an uniform Markov jump process with intensity λ and subordinated Markov chain represented by S. In the representation (82), every event is associated with a state change (since $p(x,x) = 0$). In contrast, events in the representation (82) need not necessarily correspond to a state change (since $s(x,x) \geq 0$, which might be positive).

Example 7.23 *A **birth and death process** with birth rates $(\alpha_x)_{x\in\mathbf{S}}$ and death rates $(\gamma_x)_{x\in\mathbf{S}}$ on the state space $\mathbf{S} = \mathbb{N}$ is a continuous-time Markov jump process with infinitesimal generator $A = (a(x,y))_{xy\in\mathbf{S}}$ defined by*

$$A = \begin{pmatrix} -\alpha_0 & \alpha_0 & 0 & 0 & 0 & \cdots \\ \gamma_1 & -(\gamma_1+\alpha_1) & \alpha_1 & 0 & 0 & \cdots \\ 0 & \gamma_2 & -(\gamma_2+\alpha_2) & \alpha_2 & 0 & \cdots \\ 0 & 0 & \gamma_3 & -(\gamma_3+\alpha_3) & \alpha_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}.$$

We assume that $\alpha(x), \gamma(x) \in (0,\infty)$ for $x \in \mathbf{S}$. The birth and death process is regular, if and only if the corresponding Reuters's criterion

$$\sum_{k=1}^{\infty} \left(\frac{1}{\alpha_k} + \frac{\gamma_k}{\alpha_k \alpha_{k-1}} + \cdots + \frac{\gamma_k \cdots \gamma_1}{\alpha_k \cdots \alpha_0} \right) = \infty$$

is satisfied [2, Chapt. 8, Thm. 4.5].

The next proposition shows how the generator can be used to evolve the semigroup of conditional transitions in time.

Proposition 7.24 *Consider a Markov jump process with transition semigroup $P(t)$ and infinitesimal generator $A = (a(x,y))$, satisfying $-a(x,x) < \infty$ for all $x \in \mathbf{S}$. Then, $P(t)$ is differentiable for all $t \geq 0$ and satisfies the **Kolmogorov backward equation***

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = AP(t). \tag{83}$$

If furthermore

$$\sum_y p(t,x,y)\lambda(y) < \infty \tag{84}$$

*is satisfied for all $t \geq 0$ and $x \in \mathbf{S}$, then also the **Kolmogorov forward equation***

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = P(t)A \tag{85}$$

holds.

Remark 7.25 *Condition (84) is always satisfied if the state space* $\mathbf{S}$ *is finite or* $\sup_x \lambda(x) < \infty$.

Proof: There is simple proof in the case of a finite state space. In the general case the proof is considerably harder, we refere to [19], Prop. 8.3.4, p.210.
By definition of a semigroup we have $P(t+s) = P(t)P(s) = P(s+t)$, therefore, under the assumption that $\mathbf{S}$ is finite,

$$\begin{aligned}\lim_{h\to 0}\frac{P(t+h)-P(t)}{h} &= \lim_{h\to 0} P(t)\frac{P(h)-Id}{h} = P(t)A \\ &= \lim_{h\to 0}\frac{P(h)-Id}{h}P(t) = AP(t)\end{aligned}$$

This does not work in the infinite case, because it is not sure if we can exchange the limes with the sum in the matrix-matrix multiplication. □

Remark 7.26 *Component wise the backward, resp. forward, equation reads*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P(t,x,y) &= -\lambda(x)P(t,x,y) + \sum_{z\neq x} a(x,z)P(t,z,y), \text{ resp.} \\ \frac{\mathrm{d}}{\mathrm{d}t}P(t,x,y) &= -\lambda(y)P(t,x,y) + \sum_{z\neq x} P(t,x,z)a(z,y)\end{aligned}$$

If the state space is finite the solution to eq. (83) is given by $P(t) = \exp(tA)$, where the matrix exponential function is defined via the series

$$\exp(tA) = \sum_{n\in\mathbb{N}} \frac{(tA)^n}{n!}$$

which is known to converge. The situation is quite easy if A is diagonalizable, i.e. we have $A = V^{-1}DV$ for a unitary matrix V and a diagonal matrix D. Then $A^n = V^{-1}D^nV$ and

$$\begin{aligned}\exp(tA) &= \sum_{n\in\mathbb{N}} \frac{(tV^{-1}DV)^n}{n!} = V^{-1}\sum_{n\in\mathbb{N}}\frac{(tD)^n}{n!}V \\ &= V^{-1}\mathrm{diag}(\exp(td_1),\exp(td_2),\ldots,\exp(td_r))V.\end{aligned}$$

In the non-diagonalizable case the exponential function can still be used for a reasonable approximation by computing only part of the sum.
¿From the Kolmogorov forward equation we can easily deduce the evolution equation for an arbitrary initial distribution $\mu_0 = (\mu_0(x))_{x\in\mathbf{S}}$ of the Markov jump process. As in the discrete-time case, we have

$$\mu(t) = \mu_0 P(t),$$

with $\mu(t) = (\mu(t,x))_{x\in\mathbf{S}}$ and $\mu(t,x) = \mathbb{P}_{\mu_0}[X(t) = x]$. Now, multiplying the Kolmogorov forward equation with μ_0 from the left, we get the so-called **Master equation**

$$\frac{\mathrm{d}\mu(t)}{\mathrm{d}t} = \mu(t)A$$

with initial condition $\mu(0) = \mu_0$. It describes on an infinitesimal scale the evolution of densities w.r.t. the Markov process. An alternative formulation can be given in terms of the jump rates $\{\lambda(x)\}$ and the embedded Markov chain $K = (k(x,y))$

$$\begin{aligned}\frac{\mathrm{d}\mu(t,z)}{\mathrm{d}t} &= \sum_{y\in\mathbf{S}} \mu(t,y)a(y,z) \\ &= \lambda(z)\sum_{y\in\mathbf{S}} \big(\mu(t,y) - \mu(t,z)\big)k(y,z)\end{aligned}$$

for every $z \in \mathbf{S}$.

7.4 Invariant measures and stationary distributions

This section studies the existence of invariant measures and stationary (probability) distributions. We will see that the embedded Markov chain is still a very useful object with this regard, but we will also see that not every property of the continuous-time Markov process can be specified in terms of the embedded discrete-time Markov chain. This is particularly true for the property of positive recurrence.

Definition 7.27 *A measure $\mu = (\mu(x))_{x\in\mathbf{S}}$ satisfying*

$$\mu = \mu P(t)$$

*for all $t \geq 0$ is called an **invariant** measure of the Markov jump process. If, moreover, μ is a probability measure satisfying $\mu(\mathbf{S}) = 1$, it is called a **stationary distribution**.*

We are now able to state for a quite large class of Markov jump processes the existence of invariant measures, and also to specify them.

Theorem 7.28 *Consider an irreducible and recurrent* regular *Markov jump process on $\mathbf{S}$ with transition semigroup $P(t)$. For an arbitrary state $x \in \mathbf{S}$ define $\mu = (\mu(y))_{y\in\mathbf{S}}$ via*

$$\mu(y) = \mathbb{E}_x\left[\int_0^{R_x} 1_{\{X(s)=y\}}\mathrm{d}s\right], \tag{86}$$

the expected time, the process visits y before returning to x. Then

1. $0 < \mu(y) < \infty$ *for all* $y \in \mathbf{S}$. *Moreover,* $\mu(x) = 1/\lambda(x)$ *for the state* $x \in \mathbf{S}$ *chosen in the eq. (86).*

2. $\mu = \mu P(t)$ *for all* $t \geq 0$.

3. *If* $\nu = \nu P(t)$ *for some measure* ν, *then* $\nu = \alpha\mu$ *for some* $\alpha \in \mathbb{R}$.

Proof: See Bremaud Thm. 5.1, p 357. We just prove here $\mu(x) = 1/\lambda(x)$. We have

$$\begin{aligned} \mu(x) &= \mathbb{E}_x \left[\int_0^{R_x} 1_{\{X(s)=x\}} \mathrm{d}s \right] && (87) \\ &= \mathbb{E}_x \left[\int_0^{E_x} 1_{\{X(s)=x\}} \mathrm{d}s \right] + \mathbb{E}_x \left[\int_{E_x}^{R_x} 1_{\{X(s)=x\}} \mathrm{d}s \right] && (88) \\ &= \mathbb{E}_x \left[E_x \right] + 0 \; = \; \frac{1}{\lambda(x)}. && (89) \end{aligned}$$

□

As one would expect, there is a close relation between the invariant measure of the transition semigroup, the infinitesimal generator and the embedded Markov chain.

Proposition 7.29 *Consider an irreducible and recurrent* regular *Markov jump process on* $\mathbf{S}$ *with transition semigroup* $P(t)$, *infinitesimal generator* A *and embedded Markov chain with transition matrix* K. *Then the following statements are equivalent:*

1. *There exists a measure* $\mu = (\mu(x))_{x \in \mathbf{S}}$ *such that*

$$\mu = \mu P(t)$$

   for all $t \geq 0$.

2. *There exists a measure* $\mu = (\mu(x))_{x \in \mathbf{S}}$ *such that*

$$0 = \mu A.$$

3. *There exists a measure* $\nu = (\nu(x))_{x \in \mathbf{S}}$ *such that*

$$\nu = \nu K$$

The relation between μ *and* ν *is given by* $\mu = \nu\Lambda^{-1}$, *which element-wise corresponds to*

$$\mu(x) = \frac{\nu(x)}{\lambda(x)}$$

for every $x \in \mathbf{S}$.

Proof: Exercise. □

Consider the expected return times from some state $x \in \mathbf{S}$ defined by

$$\mathbb{E}_x[R_x] = \mathbb{E}_x\left[\int_0^\infty \mathbf{1}_{\{s \le R_x\}} \mathrm{d}s\right]. \tag{90}$$

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

Definition 7.30 *A recurrent state $x \in \mathbf{S}$ is called **positive recurrent**, if*

$$\mathbb{E}_x[R_x] \;<\; \infty$$

*and **null recurrent** otherwise.*

As in the discrete-time case, we have the following result.

Theorem 7.31 *An irreducible* regular *Markov jump process with infinitesimal generator A is positive recurrent, if and only if there exists a probability distribution π on $\mathbf{S}$ such that*

$$0 = \pi A$$

holds. Under these conditions, the stationary distribution π is unique and positive everywhere, with

$$\pi(x) \;=\; \frac{1}{\lambda(x)\,\mathbb{E}_x[R_x]}.$$

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

Proof : Theorem 7.28 states that an irreducible and recurrent regular Markov jump process admits an invariant measure μ defined through (86) for an arbitrary $x \in \mathbf{S}$. Thus

$$\begin{aligned}
\sum_{y\in\mathbf{S}} \mu(y) &= \sum_{y\in\mathbf{S}} \mathbb{E}_x\left[\int_0^{R_x} 1_{\{X(s)=y\}} \mathrm{d}s\right] \\
&= \mathbb{E}_x\left[\int_0^\infty \sum_{y\in\mathbf{S}} 1_{\{X(s)=y\}} 1_{\{s\le R_x\}} \mathrm{d}s\right] \\
&= \mathbb{E}_x\left[\int_0^\infty 1_{\{s\le R_x\}} \mathrm{d}s\right] \;=\; \mathbb{E}_x[R_x],
\end{aligned}$$

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

$$\pi(x) = \frac{\mu(x)}{\mathbb{E}_x[R_x]} = \frac{1}{\lambda(x)\,\mathbb{E}_x[R_x]}.$$

Since the state x was chosen arbitrary this is true for all $x \in \mathbf{S}$. Uniqueness and positivity of π follows from Theorem 7.28. On the other hand, if there exists a stationary distribution of the Markov process, it satisfies $\pi = \pi P(t)$ for all $t \geq 0$ due to Prop. 7.29. Moreover if the Markov process were transient, then

$$\lim_{t\to\infty} \mathbf{1}_{\{X(t)=y\}} = 0 \quad \text{implying} \quad \lim_{t\to\infty} p(t,x,y) = 0$$

for $x, y \in \mathbf{S}$ by dominated convergence. In particular, $\pi P(t)$ would tend to zero for $t \to \infty$ component-wise, which would be in contradiction to $\pi = \pi P(t)$. Hence, the Markov process is recurrent. Positive recurrence follows from the uniqueness of π and the consideration above. □

Our considerations in the proof of Theorem 7.31 easily leads to a criteria to distinguish positive recurrence from null recurrence.

Corollary 7.32 *Consider an irreducible* regular *Markov jump process with invariant measure μ. Then*

1. $\{X(t) : t \geq 0\}$ *positive recurrent* $\Leftrightarrow \sum_{x\in\mathbf{S}} \mu(x) < \infty$,
2. $\{X(t) : t \geq 0\}$ *null recurrent* $\Leftrightarrow \sum_{x\in\mathbf{S}} \mu(x) = \infty$.

Proof: The proof is left as an exercise. □

It is important to notice that positive recurrence can not be characterized on the basis of the embedded Markov chain. This is due to the fact that given an irreducible regular Markov jump process with $0 = \mu A$, $\lambda(x)$, and $\nu = \nu K$, we know by by Prop. 7.29, that

$$\sum_{x=0}^{\infty} \mu(x) = \sum_{x=0}^{\infty} \frac{\nu(x)}{\lambda(x)}$$

So, whether the left hand side converges or not, depends on both, the asymptotic behavior of $(\nu(x))$ and of $(\lambda(x))$.

Example 7.33 *Consider the birth and death Markov jump process with embedded Markov chain given by*

$$K = \begin{pmatrix} 0 & 1 & & & \\ 1-p & 0 & p & & \\ & 1-p & 0 & p & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{91}$$

and jump rates $(\lambda(x))$, still to be specified. The embedded Markov chain is irreducible and recurrent for $0 < p \leq 1/2$. Hence, so is the thereby defined

Markov jump process, which in addition is regular due to Prop. 7.16. The invariant measure of the embedded Markov chain is given by

$$\nu(x) = \frac{1}{p}\left(\frac{p}{1-p}\right)^x \nu(0) \tag{92}$$

for $x \geq 1$ and $\nu(0) \in \mathbb{R}$. Computing the norm results in

$$\sum_{x=0}^{\infty} \nu(x) = \frac{2-2p}{1-2p}\,\nu(0).$$

Hence, the embedded Markov chain is null-recurrent for $p = 1/2$ and positive recurrent for $p < 1/2$. We now exemplify four possible setting:

1. *Set $\lambda(x) = x$ for $x = 1, 2, \ldots$ while $\lambda(0) = 2$, and $p = 1/2$. Then, we know that the embedded Markov chain is null-recurrent with invariant measure $\nu = (1/2, 1, 1, \ldots)$, and*

$$\sum_{x=0}^{\infty} \mu(x) = \sum_{x=0}^{\infty} \frac{1}{x} = \infty.$$

   Hence, the Markov jump process is null-recurrent, too.

2. *Set $\lambda(x) = x^2$ for $x = 1, 2, \ldots$, while $\lambda(0) = 2$, and $p = 1/2$. Again, the embedded Markov chain is null-recurrent, but now*

$$\sum_{x=0}^{\infty} \mu(x) = \sum_{x=0}^{\infty} \frac{1}{x^2} < \infty.$$

   Hence, now the Markov jump process is positive recurrent.

3. *Set $\lambda(x) = (1/3)^x$ for $x = 1, 2, \ldots$, while $\lambda(0) = 1/4$, and $p = 1/4$. Now, the embedded Markov chain is positive recurrent with stationary distribution $\nu(x) = 4(1/3)^{x+1}$ for $x \geq 1$ and $\nu(0) = 1/3$.*

$$\sum_{x=0}^{\infty} \mu(x) = \sum_{x=0}^{\infty} \frac{4}{3} = \infty.$$

   Hence, the Markov jump process is null-recurrent.

4. *Set $\lambda(x) = 4/3$ for $x = 1, 2, \ldots$, while $\lambda(0) = 1/3$, and $p = 1/4$. Again, the embedded Markov chain is positive recurrent. Finally, we have*

$$\sum_{x=0}^{\infty} \mu(x) = \sum_{x=0}^{\infty} \left(\frac{1}{3}\right)^x < \infty.$$

   Hence, the Markov jump process is positive recurrent.

In the same spirit, one can show that the existence of a stationary distribution of some irreducible Markov jump process (being not necessarily regular) does not guarantee positive recurrence. In other words, Theorem 7.31 is wrong, if one drops the assumption that the Markov jump process is regular.

Example 7.34 *We consider the embedded Markov chain with transition matrix given by eq. (91). If $p > 1/2$, then the Markov chain is transient. However, it does posses an invariant measure μ defined in eq. (92), where we choose $\nu(0) = p$. Define the jump rates by*

$$\lambda(x) = \left(\frac{1-p}{p}\right)^{2x}$$

for $x \geq 1$ and $\lambda(0) = p$. Then, we get that μ defined by $\mu(x) = \nu(x)/\lambda(x)$ is an invariant measure with

$$\sum_{x=0}^{\infty} \mu(x) = \sum_{x=0}^{\infty} \left(\frac{1-p}{p}\right)^{x} = \frac{p}{2p-1} < \infty.$$

since $p > 1/2$ and thus $(1-p)/p < 1$. Concluding, the Markov jump process is irreducible and possesses an stationary distribution. Due to Prop. 7.16, it moreover is transient (since the embedded Markov chain is). This can only be in accordance with Thm 7.31, if the Markov jump process is non-regular, hence explosive.

7.5 Reversibility and the law of large numbers

The concept of time reversibility for continuous-time Markov processes is basically the same as for discrete-time Markov chains. Consider some positive recurrent, irreducible regular Markov jump process with infinitesimal generator A and stationary distribution π. Then, define the time-reversed Markov jump process $\{Y(t) : t \geq 0\}$ in terms of its transition semigroup $\{Q(t); t \geq 0\}$ according to

$$q(t, y, x) = \frac{\pi(x)p(t, x, y)}{\pi(y)} \tag{93}$$

for $t \geq 0$ and $x, y \in \mathbf{S}$. As can easily be shown, $Q(t)$ fulfills all requirements for a transition semigroup. Defining the diagonal matrix $D_\pi = (\pi(x))_{x,y\in\mathbf{S}}$ with π on its diagonal, we may rewrite the above eq. (93) as

$$Q(t) = D_\pi^{-1} P(t)^T D_\pi.$$

Now, let us determine the infinitesimal generator $B = (b(x, y))_{x,y\in\mathbf{S}}$ of the semigroup $Q(t)$. It is

$$B \;=\; \lim_{t\to 0} \frac{Q(t) - \mathrm{Id}}{t} \;=\; \lim_{t\to 0} D_\pi^{-1} \frac{P(t)^T - \mathrm{Id}}{t} D_\pi \;=\; D_\pi^{-1} A^T D_\pi, \tag{94}$$

hence the inf. generator transforms in the same way as the semigroup does. From the inf. generator, we easily conclude that the jump rates $\lambda(x)$ are for both processes equal, since

$$\lambda_B(x) = -b(x,x) = -a(x,x) = \lambda_A(x).$$

Now, denote by K and L the transition matrix of the embedded Markov chains of the Markov jump processes $X(t)$ and $Y(t)$, respectively. Assume that K is positive recurrent with stationary distribution ν. Then, we get

$$\begin{aligned}
\Lambda(L - \mathrm{Id}) &= B = D_\pi^{-1} A^T D_\pi && (95)\\
&= D_\pi^{-1}(K^T - \mathrm{Id})\Lambda D_\pi && (96)\\
&= \Lambda\Lambda^{-1} D_\pi^{-1}(K^T - \mathrm{Id})\Lambda D_\pi && (97)\\
&= \Lambda(\Lambda^{-1} D_\pi^{-1} K^T \Lambda D_\pi - \mathrm{Id}) && (98)\\
&= \Lambda(D_\nu^{-1} K^T D_\nu - \mathrm{Id}) && (99)
\end{aligned}$$

since $\lambda(x)\pi(x) = a\nu(x)$ for all $x \in \mathbf{S}$ and some normalization constant $a > 0$, which implies $\Lambda D_\pi = a D_\nu$. Hence, we get the relation

$$L = D_\nu^{-1} K^T D_\nu$$

and thus that the embedded Markov chain L of the time-reversed Markov jump process $Y(t)$ equals the time-reversed embedded Markov chain K of the original Markov jump process $X(t)$.

As in the discrete time case, we have

Definition 7.35 *Consider an irreducible regular Markov jump process* $\{X(t) : t \geq 0\}$ *with infinitesimal generator* A *and stationary distribution* $\pi > 0$*, and its associated time-reversed Markov jump process with infinitesimal generator* B*. Then* $X(t)$ *is called* ***reversible*** *w.r.t.* π*, if*

$$A = B$$

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

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

$$\pi(x)a(x,y) = \pi(y)a(y,x) \qquad (100)$$

is satisfied for every $x, y \in \mathbf{S}$.

A measurable function $f : \mathbf{S} \to \mathbb{R}$ defined on the state space is called an **observable**. Observables allow to perform "measurements" on the system that is modelled by the Markov process. The expectation of f is defined as

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

Theorem 7.36 (Strong law of large numbers) *Let $\{X(t) : t \geq 0\}$ denote an irreducible regular Markov process with stationary distribution π, and let $f : \mathbf{S} \to \mathbb{R}$ be some observable such that*

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

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

$$\frac{1}{t}\int_0^t f(X(s))\mathrm{d}s \longrightarrow \mathbb{E}_\pi[f], \quad (a.s.)$$

as $t \to \infty$.

Proof: The proof is analogous to the proof for discrete-time Markov chains. □

7.6 Biochemical reaction kinetics

Consider a volume V containing molecules of N **chemically active species** $S_0, \ldots, S_{N-1}$ and possibly molecules of inert species. For $k = 0, \ldots, N-1$, denote by $X_k(t) \in \mathbb{N}$ the number of molecules of the chemical species S_k in V at time $t \in \mathbb{R}^+$, and set $X(t) = (X_0(t), \ldots, X_{N-1}(t)) \in \mathbb{N}^N$. Furthermore, consider M **chemical reactions** $R_0, \ldots, R_{M-1}$, characterized by a **reaction constant** c_k. The fundamental hypothesis of chemical reaction kinetics is that the rate of each reaction R_k can be specified in terms of a so-called **propensity function** $\alpha_k = \alpha_j(X(t))$, depending in general on the current state $X(t)$ and possible on time t. For the most common reaction type, it is (c = some generic reaction constant, r.p. = reaction products):

1. "spontaneous creation" $* \to$ r.p. , $\alpha(X(t), t) = c$,
2. mono-molecular reaction $S_j \to$ r.p. , $\alpha(X(t)) = c\, X_j(t)$,
3. bi-molecular reaction $S_j + S_k \to$ r.p. , $\alpha(X(t)) = c\, X_j(t) X_k(t)$,
4. bi-molecular reaction $S_j + S_j \to$ r.p. , $\alpha(X(t)) = c\, X_j(t)(X_j(t) - 1)/2$,

The change in numbers of molecules of described by the state change vectors $\nu_0, \ldots, \nu_{M-1} \in \mathbb{Z}^N$, such that $X(t) \to X(t) + \nu_k$, if reaction R_k occurs. The state change vectors are part of the stoichiometric matrix.

Example 7.37 *We consider here a model by Srivastava et al. [17] describing the intracellular growth of a T7 phage. The model comprises three chemical species, the viral nucleic acids classified into genomic (S_{gen}) and template (S_{tem}) and viral structural proteins (S_{struc}). The interaction network between the bacteriophage and the host is modelled by six reactions.*

No.	*reaction*	*propensity*	*state change*
R_0	$S_{gen} \xrightarrow{c_0} S_{tem}$	$\alpha_0 = c_0 \cdot X_{gen}$	η_0=(1,-1,0)
R_1	$S_{tem} \xrightarrow{c_1} \emptyset$	$\alpha_1 = c_1 \cdot X_{tem}$	η_1=(-1,0,0)
R_2	$S_{tem} \xrightarrow{c_2} S_{tem} + S_{gen}$	$\alpha_2 = c_2 \cdot X_{tem}$	η_2= (0,1,0)
R_3	$S_{gen} + S_{struc} \xrightarrow{c_3}$ "virus"	$\alpha_3 = c_3 \cdot X_{gen} \cdot X_{struc}$	η_3=(0,-1,-1)
R_4	$S_{tem} \xrightarrow{c_4} S_{tem} + S_{struc}$	$\alpha_4 = c_4 \cdot X_{tem}$	η_4=(0,0,1)
R_5	$S_{struc} \xrightarrow{c_5} \emptyset$	$\alpha_5 = c_5 \cdot X_{struc}$	η_5=(0,0,-1)

The reaction constants are given by $c_0 = 0.025$, $c_1 = 0.25$, $c_2 = 1.0$, $c_3 = 7.5 \cdot 10^{-6}$, $c_4 = 1000$, *and* $c_5 = 1.99$ *(day*$^{-1}$*). In the model, the volume of the cell is* $V = 1$. *The interesting scenario is the low infection level corresponding to the initial numbers of molecules* $X_{tem} = 1$, $X_{gen} = X_{struc} = 0$.

We now specify the dynamics of $\{X(t) : t \geq 0\}$, assuming that $X(t)$ is a regular Markov jump process on the state space $\mathbf{S} = \mathbb{N}^N$ that satisfies the regularity conditions (70) and (71), which seem to be a reasonable assumption for biochemical reaction systems. In terms of $X(t)$, the **fundamental hypothesis** of chemical reaction kinetics is that

$$\mathbb{P}[X(t+h) = x + \eta_k | X(t) = x] = \alpha_k(x)\, h + o(h)$$

as $h \to 0$ holds for $k = 0, \ldots, M-1$. This allows us to determine the infinitesimal generator $A = (a(x,y))_{xy \in \mathbf{S}}$. In view of

$$p(h, x, x + \eta_k) = a(x, x + \eta_k)h + o(h),$$

for $h \to 0$, we conclude that

$$a(x, x + \eta_k) = \alpha_k(x)$$

for $k = 0, \ldots, M-1$. As a consequence, the jump rates are given by the propensity functions $\{\alpha(x)\}_{x \in \mathbf{S}}$. Defining $\alpha(x) = \alpha_0(x) + \ldots + \alpha_{M-1}(x)$, the embedded Markov chain with transition matrix $K = (k(x,y))_{xy \in \mathbf{S}}$ is given by

$$k(x, x + \eta_k) = \frac{\alpha_k(x)}{\alpha(x)}$$

for $k = 0, \ldots, M-1$, and zero otherwise. The **algorithmic realization** of the chemical reaction kinetics is as follows:

1. Set initial time $t = t_0$ and initial numbers of molecules $X(t_0)$;

2. Generate independently two uniformly distributed random numbers $u_0, u_1 \sim U[0,1]$. Set $x = X(t)$.

3. Compute the next reaction time increment

$$\tau = -\frac{\ln(u_0)}{\alpha(x)};$$

4. Compute the next reaction R_k according to the discrete probability distribution

$$\left(\frac{\alpha_0(x)}{\alpha(x)}, \ldots, \frac{\alpha_{M-1}(x)}{\alpha(x)}\right);$$

5. Update molecular numbers $X(t+\tau) \leftarrow X(t)+\eta_k$, and time: $t \leftarrow t+\tau$. Go to Step 2.

This algorithmic scheme is known as the direct method [5, 6].

Consider some initial distribution u_0 and set $u(t,x) = \mathbb{P}[X(t) = x | X(0) \sim u_0]$. The evolution equation for u is given by the master equation, which in this context is called the **chemical master equation**. It takes the form

$$\begin{aligned}
\frac{\mathrm{d}u(t,x)}{\mathrm{d}t} &= \left(\sum_{y \neq x \in \mathbf{S}} u(t,y)a(y,x)\right) + u(t,x)a(x,x) \\
&= \sum_{k=0}^{M-1} \big(u(t,x-\eta_k)a(x-\eta_k,x) + u(t,x)a(x,x+\eta_k)\big) \\
&= \sum_{k=0}^{M-1} \big(\alpha_k(x-\eta_k)u(t,x-\eta_k) - \alpha_k(x)u(t,x)\big).
\end{aligned}$$

8 Transition path theory for Markov jump processes

Transition Path Theory (TPT) is concerned with transitions in Markov processes. The basic idea is to single out two disjoint subset in the state-space of the chain and ask what is the typical mechanism by which the dynamics transits from one of these states to the other. We may also ask at which rate these transitions occur.

The first object which comes to mind to characterize these transitions is the path of maximum likelihood by which they occur. However, this path can again be not very informative if the two states one has singled out are not metastable states. The main objective herein is to show that we can give a precise meaning to the question of finding typical mechanisms and rate of transition in discrete state spaces for continuous time processes which are neither metastable nor time-reversible. In a nutshell, given two subsets in state-space, TPT analyzes the statistical properties of the associated reactive trajectories, i.e., the trajectories by which transition occur between these sets. TPT provides information such as the probability distribution of these trajectories, their probability current and flux, and their rate of occurrence.

The framework of transition path theory (TPT) has first been developed in [4, 20, 10] in the context of diffusions. However, we will follow [11] and focus on continuous-time Markov chains, but we note that the results to be outlined can be straightforwardly extended to the case of discrete-time Markov chains.

8.1 Notation.

We consider a Markov jump process on the countable state space S with infinitesimal generator (or rate matrix) $L = (l_{ij})_{i,j\in S}$,

$$\begin{cases} l_{ij} \geq 0 & \text{for all } i,j \in S, i \neq j \\ \sum_{j\in S} l_{ij} = 0 & \text{for all } i \in S. \end{cases} \tag{101}$$

We assume that the process is irreducible and ergodic with respect to the unique, strictly positive invariant distribution $\pi = (\pi_i)_{i\in S}$ satisfying

$$0 = \pi^T L. \tag{102}$$

We will denote by $\{X_t\}$ a (right-continuous with left limits) trajectory of the Markov jump process. Finally, recall that if the infinitesimal generator satisfies the detailed balance equation,

$$\pi_i l_{ij} = \pi_j l_{ji} \quad \forall i,j \in S, \tag{103}$$

the process is reversible, i.e. the direct and the time-reversed process are statistically indistinguishable. In the following we assume that the process is irreducible and reversible.

Figure 20: Schematic representation of the first hitting time scenario. A trajectory hits the boundary ∂D at time τ the first time.

8.2 Hitting times and potential theory

In this paragraph we state a theorem which provides an elegant way to derive equations for the main objects in TPT. Suppose the state space S is decomposed into two disjoint sets D and $\partial D = S \setminus D$ where ∂D is called the *boundary* of D. Let $i \in D$ be an arbitrary state. Conditional on starting in i, the time $\tau(i)$ at which the process hits the boundary ∂D the first time is called the (first) *hitting time*. Formally, it is defined as

$$\tau(i) = \inf\{t > 0 : X_t \in \partial D, X_0 = i\}. \tag{104}$$

Notice that τ is a random variable and, particularly, τ is a stopping time. For a schematic presentation of the first hitting scenario see Fig. 20.

Next, suppose that two real-valued discrete functions $(c_i)_i, i \in S$ and $(f_i)_i, i \in S$ are given. The object of interest is the *potential* $\phi = (\phi_i), i \in S$ associated with the two functions and element-wise defined by

$$\phi_i = \mathbb{E}\left[\int_0^\tau c(X_t)dt + f(X_\tau)1_{\tau<\infty}|X_0 = i\right], \tag{105}$$

where τ denotes the *hitting time* of ∂D. In the following we assume that the hitting time τ is finite which is guaranteed by the irreducibility of the process.

Regarding the $(c_i)_i, i \in S$ and $(f_i)_i, i \in S$ as *costs*, the potential can be interpreted as an expected total cost: the process wanders around in D until it his the boundary ∂D where the cost of wandering around in D *per unit time* is given by $(c_i)_i, i \in S$. When the process hits the boundary, say in state j, a final cost or fee f_j is incurred. The next theorem states that the potential satisfies a discrete Dirichlet problem involving the generator $L = (l_{ij}), i, j \in S$ of the process.

Theorem 8.1 ([13], Sect. 4.2) *Under the assumption that the functions $(c_i)_i, i \in S$ and $(f_i)_i, i \in S$ are nonnegative, the potential ϕ is the nonnegative solution to*

$$\begin{cases} \sum_{j\in S} l_{ij}\phi_j = -c_i & \forall i \in D, \\ \phi_i = f_i & \forall i \in \partial D. \end{cases} \tag{106}$$

Consequently, the potential can simply numerically be computed by solving a system of linear equations with Dirichlet boundary conditions.

Figure 21: Schematic representation of a piece of an ergodic trajectory. The sub-piece connecting A to B (shown in thick black) is a reactive trajectory, and the collection of reactive trajectories is the ensemble of reactive trajectories.

8.3 Reactive trajectories.

Let A and B two nonempty, disjoint subsets of the state space S. By ergodicity, any equilibrium path $\{X_t\}$ oscillates infinitely many times between set A and set B. If we view A as a reactant state and B as a product state, each oscillation from A to B is a reaction event. To properly define and characterize the reaction events, we proceed by cutting a long ergodic trajectory $\{X_t\}$ into pieces that each connect A and B. To be more precise, a *reactive trajectory* is a piece of the ergodic trajectory which lives in $S \setminus (A \cup B)$ after leaving the set A and before hitting the set B. We shall then try to describe various statistical properties of the ensemble of reactive trajectories consisting of all these pieces. See Fig. 21 for a schematic illustration of a reaction trajectory. For details on the pruning procedure, see [11].

8.4 Committor functions

The fundamental objects of TPT are the committor functions. The *discrete forward committor* $q^+ = (q_i^+)_{i\in S}$ is defined as the probability that the process starting in $i \in S$ will reach first B rather than A. Analogously, we define the *discrete backward committor* $q^- = (q_i^-)_{i\in S}$ as the probability that the process arriving in state i has been started in A rather than B. It has been proven in [11] that the forward and backward committor satisfy a discrete Dirichlet problem that is the exact finite-dimensional analogue of the respective continuous problem [4], namely,

$$\begin{cases} \sum_{j\in S} l_{ij} q_j^+ = 0, & \forall i \in S \setminus (A \cup B) \\ q_i^+ = 0, & \forall i \in A \\ q_i^+ = 1, & \forall i \in B. \end{cases} \tag{107}$$

The forward committor equation (107) readily follows from Theorem 8.1. To see that set

$$c_i = 0 \quad \forall i \in S \setminus (A \cup B), \tag{108}$$

$$f_i = \begin{cases} 0, & \forall i \in A \\ 1, & \forall i \in B \end{cases} \tag{109}$$

and notice that the associated potential reduces to ($D = S \setminus (A \cup B)$):

$$\begin{aligned} \phi_i &= \mathbb{E}\left[\int_0^\tau c(X_t)dt + f(X_\tau)|X_0 = i\right] && (110)\\ &= \mathbb{E}\left[1_B(X_\tau)|X_0 = i\right] && (111)\\ &= q_i^+. && (112) \end{aligned}$$

Finally, we state without proof that the backward committor q_i^-, i.e., the probability that the process arriving in state i has been started in A rather than in B, is simply related to the forward committor function by

$$q_i^- = 1 - q_i^+. \tag{113}$$

That relation is a consequence of reversibility.

Another interesting quantity follows from setting

$$\begin{aligned} c_i &= 1 \quad \forall i \in S \setminus B, && (114)\\ f_i &= 0 \quad \forall i \in B. && (115) \end{aligned}$$

The associated potential is simply the expected or *mean first hitting time* of the process with respect to the set B,

$$\begin{aligned} \phi_i &= \mathbb{E}\left[\int_0^\tau c(X_t)dt + f(X_\tau)|X_0 = i\right] && (116)\\ &= \mathbb{E}\left[\tau|X_0 = i\right]. && (117) \end{aligned}$$

and satisfies the discrete Dirichlet problem

$$\begin{cases} \sum_{j\in S} l_{ij} m_j = -1, & \forall i \in S \setminus B \\ m_i = 0, & \forall i \in B. \end{cases} \tag{118}$$

8.5 Probability distribution of reactive trajectories.

The first relevant object for quantifying the statistical properties of the reactive trajectories is the *distribution of reactive trajectories* $m^R = (m_i^R)_{i\in S}$. The distribution m^R gives the equilibrium probability to observe a reactive trajectory at state i and at any time t.

How can we find an expression for m^R? Suppose we encounter the process X_t in a state $i \in S$. What is the probability that X_t be reactive? Intuitively, this is the probability that the process came rather from A than from B times the probability that the process will reach B rather than A in the future. Indeed, it was proven in [11] that the probability distribution of reactive trajectories is given by

$$m_i^R = q_i^- \pi_i q_i^+ = (1 - q^+)_i \pi_i q_i^+, \quad i \in S. \tag{119}$$

Probability current of reactive trajectories. Any equilibrated trajectory $\{X_t\}$ induces an equilibrium probability current f_{ij} between any pair (i, j) of states. In other words, f_{ij} is the average number of jumps from i to j per time unit observed in an infinitesimal time interval dt. In formula, that current is given by

$$f_{ij} = \pi_i l_{ij},$$

where l_{ij} is the jump rate from i to j and π_i is the probability to observe the equilibrated process in state i.

It should be clear that probability current induced by the ensemble of reactive trajectories differs from equilibrium probability current f_{ij}. How can be derive an expression for the average number of reactive trajectories flowing from state i to state j per unit of time? Again, intuitively, that current is given by the equilibrium probability current f_{ij} weighted with the probability that the process came rather from A than from B before jumping from i to j and weighted by the probability that the process will continue to B rather than to A after jumping from i to j.

Formally, the *probability current of reactive trajectories* $f^{AB} = (f_{ij}^{AB})_{i,j\in S}$ is given by [11]

$$f_{ij}^{AB} = \begin{cases} (1-q^+)_i \pi_i l_{ij} q_j^+, & \text{if } i \neq j \\ 0, & \text{otherwise.} \end{cases} \tag{120}$$

8.6 Transition rate and effective current.

Further we may ask for the average number of transitions from A to B per time unit or, equivalently, the average number of reactive trajectories observed per unit of time (transition rate). Let N_T be the number of reactive trajectories in the interval $[-T, T]$ in time. The *transition rate* k_{AB} is defined as

$$k_{AB} = \lim_{T\to\infty} \frac{N_T}{2T}. \tag{121}$$

Due to [11] the transition rate is given by

$$k_{AB} = \sum_{i\in A, j\in S} f_{ij}^{AB} = \sum_{j\in S, k\in B} f_{jk}^{AB}. \tag{122}$$

Notice that the rate equals

$$k_{AB} = \sum_{i\in A, j\in S} f_{ij}^{+}, \tag{123}$$

where the *effective current* is defined as

$$f_{ij}^{+} = \max(f_{ij}^{AB} - f_{ji}^{AB}, 0). \tag{124}$$

8.7 Reaction Pathways.

A *reaction pathway* $w = (i_0, i_1, \ldots, i_n),\ i_j \in S, j = 0, \ldots, n$ from A to B is a simple pathway with the property

$$i_0 \in A, i_n \in B, i_j \in (A \cup B)^c \quad j = 1, \ldots, n-1.$$

The crucial observation which leads to a characterization of bottlenecks of reaction pathways is that the amount of reactive trajectories which can be conducted by a reaction pathway per time unit is confined by the minimal effective current of a transition involved along the reaction pathway: the *min-current* of w is

$$c(w) = \min_{e=(i,j)\in w} \{f_{ij}^+\}. \tag{125}$$

Accordingly we shall characterize the "best" reaction pathway as the one with the *maximal min-current*, and, eventually, we can rank all reaction pathways according to the respective weight $c(w)$. Efficient graph algorithms for computing the hierarchy of transition pathways can be found in [11].

Acknowledgement Supported by the DFG Research Center "Mathematics for key technologies" (FZT 86) in Berlin.

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. E and E. Vanden-Eijnden. Towards a theory of transition paths. *J. Statist. Phys.*, 123(3):503–523, 2006.

[5] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys*, 22:403–434, 1976.

[6] D. T. Gillespie. Exact stochastic simulating of coupled chemical reactions. *J Phys Chem*, 81:2340–2361, 1977.

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

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

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

[10] P. Metzner, C. Schütte, and E. Vanden-Eijnden. Illustration of transition path theory on a collection of simple examples. *J. Chem. Phys.*, 125(8):084110, 2006.

[11] P. Metzner, C. Schütte, and E. Vanden-Eijnden. Transition path theory for Markov jump processes. *Mult. Mod. Sim.*, 7(3):1192–1219, 2009.

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

[13] J. R. Norris. *Markov chains*. Cambridge University Press, 1998.

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

[15] C. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In P. G. Ciarlet, editor, *Handbook of Numerical Analysis*, volume Special Volume Computational Chemistry, pages 699–744. North–Holland, 2003.

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

[17] R. Srivastava, L. You, J. Summer, and J. Yin. Stochastic vs. deterministic modeling of intracellular viral kinetics. *J theor Biol*, 218:309–321, 2002.

[18] 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.

[19] P. Todorovic. *An Introduction to Stochastic Processes and Their Applications*. Springer Series in Statitics. Springer, 1992.

[20] E. Vanden-Eijnden. Transition path theory. In M. Ferrario, G. Ciccotti, and K. Binder, editors, *Computer Simulations in Condensed Matter: from Materials to Chemical Biology*, volume 2 of *703*. Springer Verlag, 2006.

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