Chapter 1

Markov Chains

This chapter introduces Markov chains[1], a special kind of random process which is said to have "no memory": the evolution of the process in the future depends only on the present state and not on where it has been in the past. In order to be able to study Markov chains, we first need to introduce the concept of a stochastic process.

1.1 Stochastic processes

Definition 1.1 (Stochastic process). *A* stochastic process X *is a family* $\{X_t : t \in T\}$ *of random variables* $X_t : \Omega \to S$. T *is hereby called the* index set *("time") and* S *is called the* state space.

We will soon focus on stochastic *processes in discrete time*, i.e. we assume that $T \subset \mathbb{N}$ or $T \subset \mathbb{Z}$. Other choices would be $T = [0, \infty)$ or $T = \mathbb{R}$ (*processes in continuous time*) or $T = \mathbb{R} \times \mathbb{R}$ (*spatial process*).

An example of a stochastic process in discrete time would be the sequence of temperatures recorded every morning at Braemar in the Scottish Highlands. Another example would be the price of a share recorded at the opening of the market every day. During the day we can trace the share price continuously, which would constitute a stochastic process in continuous time.

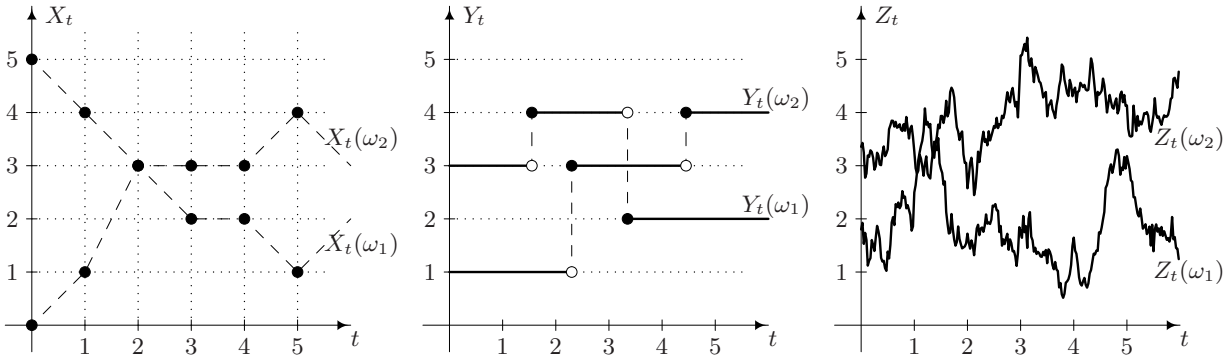
We can distinguish between processes not only based on their index set T, but also based on their state space S, which gives the "range" of possible values the process can take. An important special case arises if the state space S is a countable set. We shall then call X a *discrete process*. The reasons for treating discrete processes separately are the same as for treating discrete random variables separately: we can assume without loss of generality that the state space are the natural numbers. This special case will turn out to be much simpler than the case of a general state space.

Definition 1.2 (Sample Path). *For a given realisation* $\omega \in \Omega$ *the collection* $\{X_t(\omega) : t \in T\}$ *is called the* sample path *of* X *at* ω.

If $T = \mathbb{N}_0$ (discrete time) the sample path is a sequence; if $T = \mathbb{R}$ (continuous time) the sample path is a function from $\mathbb{R}$ to S.

Figure 1.1 shows sample paths both of a stochastic process in discrete time (panel (a)), and of two stochastic processes in continuous time (panels (b) and (c)). The process in panel (b) has a discrete state space, whereas the process in panel (c) has the real numbers as its state space ("continuous state space"). Note that whilst certain stochastic processes have sample paths that are (almost surely) continuous or differentiable, this does not need to be the case.

[1] named after the Andrey Andreyevich Markov (1856–1922), a Russian mathematician.



(a) Two sample paths of a discrete process in discrete time.

(b) Two sample paths of a discrete process in continuous time.

(c) Two sample paths of a continuous process in continuous time.

Figure 1.1. Examples of sample paths of stochastic processes.

A stochastic process is not only characterised by the marginal distributions of X_t, but also by the dependency structure of the process. This dependency structure can be expressed by the *finite-dimensional distributions* of the process:

$$\mathbb{P}(X_{t_1} \in A_1, \ldots, X_{t_k} \in A_k)$$

where $t_1, \ldots, t_k \in T$, $k \in \mathbb{N}$, and $A_1, \ldots, A_k$ are measurable subsets of S. In the case of $S \subset \mathbb{R}$ the finite-dimensional distributions can be represented using their joint distribution functions

$$F_{(t_1,\ldots,t_k)}(x_1, \ldots, x_k) = \mathbb{P}(X_{t_1} \in (-\infty, x_1], \ldots, X_{t_k} \in (-\infty, x_k]).$$

This raises the question whether a stochastic process X is fully described by its finite dimensional distributions. The answer to this is given by Kolmogorov's existence theorem. However, in order to be able to formulate the theorem, we need to introduce the concept of a consistent family of finite-dimensional distributions. To keep things simple, we will formulate this condition using distributions functions. We shall call a family of finite dimensional distribution functions *consistent* if for any collection of times $t_1, \ldots t_k$, for all $j \in \{1, \ldots, k\}$

$$F_{(t_1,\ldots,t_{j-1},t_j,t_{j+1},\ldots,t_k)}(x_1, \ldots, x_{j-1}, +\infty, x_{j+1}, \ldots, x_k) = F_{(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_k)}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k) \tag{1.1}$$

This consistency condition says nothing else than that lower-dimensional members of the family have to be the marginal distributions of the higher-dimensional members of the family. For a discrete state space, (1.1) corresponds to

$$\sum_{x_j} p_{(t_1,\ldots,t_{j-1},t_j,t_{j+1},\ldots,t_k)}(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_k) = p_{(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_k)}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k),$$

where $p_{(\ldots)}(\cdot)$ are the joint probability mass functions (p.m.f.). For a continuous state space, (1.1) corresponds to

$$\int f_{(t_1,\ldots,t_{j-1},t_j,t_{j+1},\ldots,t_k)}(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_k)\, dx_j = f_{(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_k)}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$$

where $f_{(\ldots)}(\cdot)$ are the joint probability density functions (p.d.f.).

Without the consistency condition we could obtain different results when computing the same probability using different members of the family.

Theorem 1.3 (Kolmogorov). *Let $F_{(t_1,\ldots,t_k)}$ be a family of consistent finite-dimensional distribution functions. Then there exists a probability space and a stochastic process X, such that*

$$F_{(t_1,\ldots,t_k)}(x_1, \ldots, x_k) = \mathbb{P}(X_{t_1} \in (-\infty, x_1], \ldots, X_{t_k} \in (-\infty, x_k]).$$

Proof. The proof of this theorem can be for example found in (Gihman and Skohorod, 1974). □

Thus we can specify the distribution of a stochastic process by writing down its finite-dimensional distributions. Note that the stochastic process X is not necessarily uniquely determined by its finite-dimensional distributions. However, the finite dimensional distributions uniquely determine all probabilities relating to events involving an at most countable collection of random variables. This is however, at least as far as this course is concerned, all that we are interested in.

In what follows we will only consider the case of a stochastic process in discrete time i.e. $T = \mathbb{N}_0$ (or $\mathbb{Z}$). Initially, we will also assume that the state space is discrete.

1.2 Discrete Markov chains

1.2.1 Introduction

In this section we will define Markov chains, however we will focus on the special case that the state space S is (at most) countable. Thus we can assume without loss of generality that the state space S is the set of natural numbers $\mathbb{N}$ (or a subset of it): there exists a bijection that uniquely maps each element to S to a natural number, thus we can relabel the states $1, 2, 3, \ldots$.

Definition 1.4 (Discrete Markov chain). *Let X be a stochastic process in discrete time with countable ("discrete") state space. X is called a* Markov chain (with discrete state space) *if X satisfies the* Markov property

$$\mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t, \ldots, X_0 = x_0) = \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t)$$

This definition formalises the idea of the process depending on the past only through the present. If we know the current state X_t, then the next state X_{t+1} is independent of the past states $X_0, \ldots X_{t-1}$. Figure 1.2 illustrates this idea.[2]

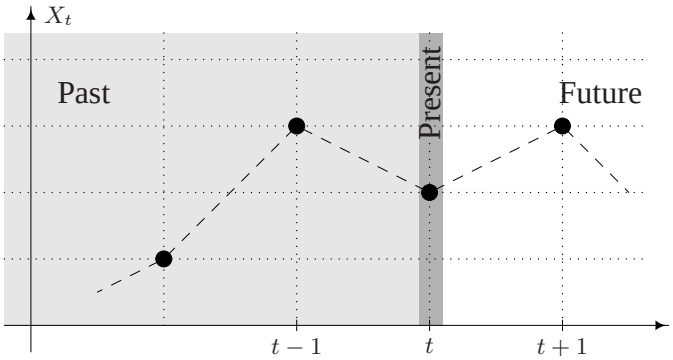


Figure 1.2. Past, present, and future of a Markov chain at t.

Proposition 1.5. *The Markov property is equivalent to assuming that for all $k \in \mathbb{N}$ and all $t_1 < \ldots < t_k \leq t$*

$$\mathbb{P}(X_{t+1} = x_{t+1} | X_{t_k} = x_{t_k}, \ldots, X_{t_1} = x_{t_1}) = \mathbb{P}(X_{t+1} = x_{t+1} | X_{t_k} = x_{t_k}).$$

Proof. (homework) □

Example 1.1 (Phone line). Consider the simple example of a phone line. It can either be busy (we shall call this state 1) or free (which we shall call 0). If we record its state every minute we obtain a stochastic process $\{X_t : t \in \mathbb{N}_0\}$.

[2] A similar concept (*Markov processes*) exists for processes in continuous time. See e.g. `http://en.wikipedia.org/wiki/Markov_process`.

If we assume that $\{X_t : t \in \mathbb{N}_0\}$ is a Markov chain, we assume that probability of a new phone call being ended is independent of how long the phone call has already lasted. Similarly the Markov assumption implies that the probability of a new phone call being made is independent of how long the phone has been out of use before.
The Markov assumption is compatible with assuming that the usage pattern changes over time. We can assume that the phone is more likely to be used during the day and more likely to be free during the night. ◁

Example 1.2 (Random walk on $\mathbb{Z}$). Consider a so-called *random walk* on $\mathbb{Z}$ starting at $X_0 = 0$. At every time, we can either stay in the state or move to the next smaller or next larger number. Suppose that independently of the current state, the probability of staying in the current state is $1-\alpha-\beta$, the probability of moving to the next smaller number is α and that the probability of moving to the next larger number is β, where $\alpha, \beta \geq 0$ with $\alpha + \beta \leq 1$. Figure 1.3 illustrates this idea. To analyse this process in more detail we write X_{t+1} as

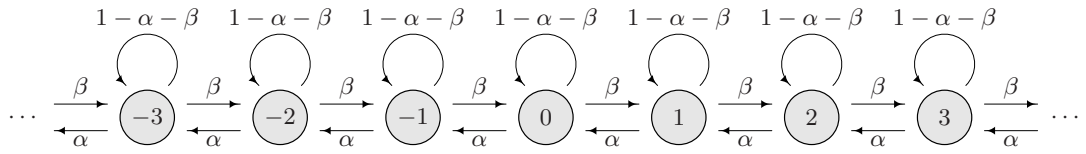


Figure 1.3. Illustration ("Markov graph") of the random walk on $\mathbb{Z}$.

$$X_{t+1} = X_t + E_t,$$

with the E_t being independent and for all t

$$\mathbb{P}(E_t = -1) = \alpha \qquad \mathbb{P}(E_t = 0) = 1-\alpha-\beta \qquad \mathbb{P}(E_t = 1) = \beta.$$

It is easy to see that

$$\mathbb{P}(X_{t+1} = x_t - 1 | X_t = x_t) = \alpha \quad \mathbb{P}(X_{t+1} = x_t | X_t = x_t) = 1-\alpha-\beta \quad \mathbb{P}(X_{t+1} = x_t + 1 | X_t = x_t) = \beta$$

Most importantly, these probabilities do not change when we condition additionally on the past $\{X_{t-1} = x_{t-1}, \ldots, X_0 = x_0\}$:

$$\begin{aligned} & \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t, X_{t-1} = x_{t-1} \ldots, X_0 = x_0) \\ = \quad & \mathbb{P}(E_t = x_{t+1} - x_t | E_{t-1} = x_t - x_{t-1}, \ldots, E_0 = x_1 - x_0, X_0 = x_o) \\ \overset{E_s \perp E_t}{=} \quad & \mathbb{P}(E_t = x_{t+1} - x_t) = \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t) \end{aligned}$$

Thus $\{X_t : t \in \mathbb{N}_0\}$ is a Markov chain. ◁

The distribution of a Markov chain is fully specified by its *initial distribution* $\mathbb{P}(X_0 = x_0)$ and the *transition probabilities* $\mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t)$, as the following proposition shows.

Proposition 1.6. *For a discrete Markov chain $\{X_t : t \in \mathbb{N}_0\}$ we have that*

$$\mathbb{P}(X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_0 = x_0) \cdot \prod_{\tau=0}^{t-1} \mathbb{P}(X_{\tau+1} = x_{\tau+1} | X_\tau = x_\tau).$$

Proof. From the definition of conditional probabilities we can derive that

$$\begin{aligned}\mathbb{P}(X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) &= \mathbb{P}(X_0 = x_0) \\ &\quad \cdot \mathbb{P}(X_1 = x_1 | X_0 = x_0) \\ &\quad \cdot \underbrace{\mathbb{P}(X_2 = x_2 | X_1 = x_1, X_0 = x_0)}_{=\mathbb{P}(X_2 = x_2 | X_1 = x_1)} \\ &\quad \cdots \\ &\quad \cdot \underbrace{\mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_0 = x_0)}_{=\mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1})} \\ &= \prod_{\tau=0}^{t-1} \mathbb{P}(X_{\tau+1} = x_{\tau+1} | X_\tau = x_\tau). \qquad \square\end{aligned}$$

Comparing the equation in proposition 1.6 to the first equation of the proof (which holds for any sequence of random variables) illustrates how powerful the Markovian assumption is.

To simplify things even further we will introduce the concept of a *homogeneous Markov chain*, which is a Markov chains whose behaviour does not change over time.

Definition 1.7 (Homogeneous Markov Chain). *A Markov chain* $\{X_t : t \in \mathbb{N}_0\}$ *is said to be* homogeneous *if*

$$\mathbb{P}(X_{t+1} = j | X_t = i) = p_{ij}$$

for all $i, j \in S$*, and independent of* $t \in \mathbb{N}_0$.

In the following we will assume that all Markov chains are homogeneous.

Definition 1.8 (Transition kernel). *The matrix* $\mathbf{K} = (k_{ij})_{ij}$ *with* $k_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$ *is called the* transition kernel *(or* transition matrix*) of the homogeneous Markov chain* X.

We will see that together with the initial distribution, which we might write as a vector $\boldsymbol{\lambda}_0 = (\mathbb{P}(X_0 = i))_{(i \in S)}$, the transition kernel $\mathbf{K}$ fully specifies the distribution of a homogeneous Markov chain.

However, we start by stating two basic properties of the transition kernel $\mathbf{K}$:

– The entries of the transition kernel are non-negative (they are probabilities).
– Each row of the transition kernel sums to 1, as

$$\sum_j k_{ij} = \sum_j \mathbb{P}(X_{t+1} = j | X_t = i) = \mathbb{P}(X_{t+1} \in S | X_t = i) = 1$$

Example 1.3 (Phone line (continued)). Suppose that in the example of the phone line the probability that someone makes a new call (if the phone is currently unused) is 10% and the probability that someone terminates an active phone call is 30%. If we denote the states by 0 (phone not in use) and 1 (phone in use). Then

$$\begin{aligned}\mathbb{P}(X_{t+1} = 0 | X_t = 0) &= 0.9 & \mathbb{P}(X_{t+1} = 1 | X_t = 0) &= 0.1 \\ \mathbb{P}(X_{t+1} = 0 | X_t = 1) &= 0.3 & \mathbb{P}(X_{t+1} = 1 | X_t = 1) &= 0.7,\end{aligned}$$

and the transition kernel is

$$\mathbf{K} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}.$$

The transition probabilities are often illustrated using a so-called Markov graph. The Markov graph for this example is shown in figure 1.4. Note that knowing $\mathbf{K}$ alone is not enough to find the distribution of the states: for this we also need to know the initial distribution $\boldsymbol{\lambda}_0$. $\triangleleft$

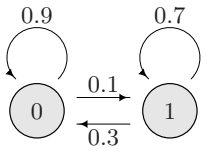


Figure 1.4. Markov graph for the phone line example.

Example 1.4 (Random walk on $\mathbb{Z}$ (continued)). The transition kernel for the random walk on $\mathbb{Z}$ is a Toeplitz matrix with an infinite number of rows and columns:

$$\mathbf{K} = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \alpha & 1-\alpha-\beta & \beta & 0 & 0 & 0 & \ddots \\ \ddots & 0 & \alpha & 1-\alpha-\beta & \beta & 0 & 0 & \ddots \\ \ddots & 0 & 0 & \alpha & 1-\alpha-\beta & \beta & 0 & \ddots \\ \ddots & 0 & 0 & 0 & \alpha & 1-\alpha-\beta & \beta & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

The Markov graph for this Markov chain was given in figure 1.3. ◁

We will now generalise the concept of the transition kernel, which contains the probabilities of moving from state i to step j in one step, to the m-step transition kernel, which contains the probabilities of moving from state i to step j in m steps:

Definition 1.9 (m-step transition kernel). *The matrix $\mathbf{K}^{(m)} = (k_{ij}^{(m)})_{ij}$ with $k_{ij}^{(m)} = \mathbb{P}(X_{t+m} = j | X_t = i)$ is called the* m-step transition kernel *of the homogeneous Markov chain X.*

We will now show that the m-step transition kernel is nothing other than the m-power of the transition kernel.

Proposition 1.10. *Let X be a homogeneous Markov chain, then*

i. $\mathbf{K}^{(m)} = \mathbf{K}^m$, *and*
ii. $\mathbb{P}(X_m = j) = (\boldsymbol{\lambda}_0' \mathbf{K}^{(m)})_j$.

Proof. i. We will first show that for $m_1, m_2 \in \mathbb{N}$ we have that $\mathbf{K}^{(m_1+m_2)} = \mathbf{K}^{(m_1)} \cdot \mathbf{K}^{(m_2)}$:

$$\begin{aligned} \mathbb{P}(X_{t+m_1+m_2} = k | X_t = i) &= \sum_j \mathbb{P}(X_{t+m_1+m_2} = k, X_{t+m_1} = j | X_t = i) \\ &= \sum_j \underbrace{\mathbb{P}(X_{t+m_1+m_2} = k | X_{t+m_1} = j, X_t = i)}_{=\mathbb{P}(X_{t+m_1+m_2} = k | X_{t+m_1} = j) = \mathbb{P}(X_{t+m_2} = k | X_t = j)} \mathbb{P}(X_{t+m_1} = j | X_t = i) \\ &= \sum_j \mathbb{P}(X_{t+m_2} = k | X_t = j) \mathbb{P}(X_{t+m_1} = j | X_t = i) \\ &= \sum_j \mathbf{K}_{ij}^{(m_1)} \mathbf{K}_{jk}^{(m_2)} = \left(\mathbf{K}^{(m_1)} \mathbf{K}^{(m_2)} \right)_{i,k} \end{aligned}$$

Thus $\mathbf{K}^{(2)} = \mathbf{K} \cdot \mathbf{K} = \mathbf{K}^2$, and by induction $\mathbf{K}^{(m)} = \mathbf{K}^m$.

ii. $\mathbb{P}(X_m = j) = \sum_i \mathbb{P}(X_m = j, X_0 = i) = \sum_i \underbrace{\mathbb{P}(X_m = j | X_0 = i)}_{=\mathbf{K}_{ij}^{(m)}} \underbrace{\mathbb{P}(X_0 = i)}_{=(\boldsymbol{\lambda}_0)_i} = (\boldsymbol{\lambda}_0' \mathbf{K}^m)_j$ □

Example 1.5 (Phone line (continued)). In the phone-line example, the transition kernel is

$$\mathbf{K} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}.$$

The m-step transition kernel is

$$\mathbf{K}^{(m)} = \mathbf{K}^m = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}^m = \begin{pmatrix} \frac{3+\left(\frac{3}{5}\right)^m}{4} & \frac{1-\left(\frac{3}{5}\right)^m}{4} \\ \frac{1+\left(\frac{3}{5}\right)^m}{4} & \frac{3-\left(\frac{3}{5}\right)^m}{4} \end{pmatrix}.$$

Thus the probability that the phone is free given that it was free 10 hours ago is $\mathbb{P}(X_{t+10} = 0 | X_t = 0) = K_{0,0}^{(10)} = \frac{3+\left(\frac{3}{5}\right)^{10}}{4} = \frac{7338981}{9765625} = 0.7515.$ ◁

1.2.2 Classification of states

Definition 1.11 (Classification of states). *(a) A state i is said to* lead to *a state j ("$i \rightsquigarrow j$") if there is an $m \geq 0$ such that there is a positive probability of getting from state i to state j in m steps, i.e.*

$$k_{ij}^{(m)} = \mathbb{P}(X_{t+m} = j | X_t = i) > 0.$$

(b) Two states i and j are said to communicate *("$i \sim j$") if $i \rightsquigarrow j$ and $j \rightsquigarrow i$.*

From the definition we can derive for states $i, j, k \in S$:

- $i \rightsquigarrow i$ (as $k_{ii}^{(0)} = \mathbb{P}(X_{t+0} = i | X_t = i) = 1 > 0$), thus $i \sim i$.
- If $i \sim j$, then also $j \sim i$.
- If $i \rightsquigarrow j$ and $j \rightsquigarrow k$, then there exist $m_{ij}, m_2 \geq 0$ such that the probability of getting from state i to state j in m_{ij} steps is positive, i.e. $k_{ij}^{(m_{ij})} = \mathbb{P}(X_{t+m_{ij}} = j | X_t = i) > 0$, as well as the probability of getting from state j to state k in m_{jk} steps, i.e. $k_{jk}^{(m_{jk})} = \mathbb{P}(X_{t+m_{jk}} = k | X_t = j) = \mathbb{P}(X_{t+m_{ij}+m_{jk}} = k | X_{t+m_{ij}} = j) > 0$. Thus we can get (with positive probability) from state i to state k in $m_{ij} + m_{jk}$ steps:

$$\begin{aligned} k_{ik}^{(m_{ij}+m_{jk})} &= \mathbb{P}(X_{t+m_{ij}+m_{jk}} = k | X_t = i) = \sum_{\iota} \mathbb{P}(X_{t+m_{ij}+m_{jk}} = k | X_{t+m_{ij}} = \iota) \mathbb{P}(X_{t+m_{ij}} = \iota | X_t = i) \\ &\geq \underbrace{\mathbb{P}(X_{t+m_{ij}+m_{jk}} = k | X_{t+m_{ij}} = j)}_{>0} \underbrace{\mathbb{P}(X_{t+m_{ij}} = j | X_t = i)}_{>0} > 0 \end{aligned}$$

Thus $i \rightsquigarrow j$ and $j \rightsquigarrow k$ imply $i \rightsquigarrow k$. Thus $i \sim j$ and $j \sim k$ also imply $i \sim k$

Thus $\sim$ is an equivalence relation and we can partition the state space S into *communicating classes*, such that all states in one class communicate and no larger classes can be formed. A class C is called *closed* if there are no paths going out of C, i.e. for all $i \in C$ we have that $i \rightsquigarrow j$ implies that $j \in C$.

We will see that states within one class have many properties in common.

Example 1.6. Consider a Markov chain with transition kernel

$$\mathbf{K} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{3}{4} & 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

The Markov graph is shown in figure 1.5. We have that $2 \sim 4$, $2 \sim 5$, $3 \sim 6$, $4 \sim 5$. Thus the communicating classes are $\{1\}$, $\{2, 4, 5\}$, and $\{3, 6\}$. Only the class $\{3, 6\}$ is closed. ◁

Finally, we will introduce the notion of an *irreducible chain*. This concept will become important when we analyse the limiting behaviour of the Markov chain.

Definition 1.12 (Irreducibility). *A Markov chain is called* irreducible *if it only consists of a single class, i.e. all states communicate.*

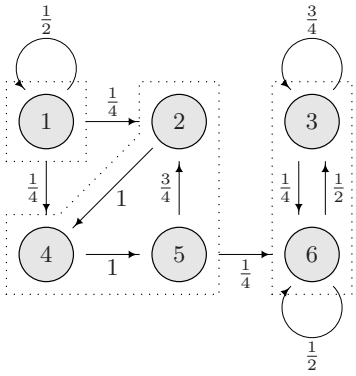


Figure 1.5. Markov graph of the chain of example 1.6. The communicating classes are $\{1\}$, $\{2,4,5\}$, and $\{3,6\}$.

The Markov chain considered in the phone-line example (examples 1.1,1.3, and 1.5) and the random walk on $\mathbb{Z}$ (examples 1.2 and 1.4) are irreducible chains. The chain of example 1.6 is not irreducible.

In example 1.6 the states 2, 4 and 5 can only be visited in this order: if we are currently in state 2 (i.e. $X_t = 2$), then we can only visit this state again at time $t+3$, $t+6$, ... Such a behaviour is referred to as periodicity.

Definition 1.13 (Period). *(a) A state $i \in S$ is said to have* period

$$d(i) = \gcd\{m \geq 1 : \ K_{ii}^{(m)} > 0\},$$

where gcd *denotes the greatest common denominator.*
(b) If $d(i) = 1$ the state i is called aperiodic.
(c) If $d(i) > 1$ the state i is called periodic.

For a periodic state i, the number of steps required to possibly get back to this state must be a multiple of the period $d(i)$.

To analyse the periodicity of a state i we must check the existence of paths of positive probability and of length m going from the state i back to i. If no path of length m exists, then $K_{ii}^{(m)} = 0$. If there exists a single path of positive probability of length m, then $K_{ii}^{(m)} > 0$.

Example 1.7 (Example 1.6 continued). In example 1.6 the state 2 has period $d(2) = 3$, as all paths from 2 back to 2 have a length which is a multiple of 3, thus

$$K_{22}^{(3)} > 0, \qquad K_{22}^{(6)} > 0, \qquad K_{22}^{(9)} > 0, \qquad \ldots$$

All other $K_{22}^{(m)} = 0$ ($\frac{m}{3} \notin \mathbb{N}_0$), thus the period is $d(2) = 3$ (3 being the greatest common denominator of $3, 6, 9, \ldots$). Similarly $d(4) = 3$ and $d(5) = 3$.

The states 3 and 6 are aperiodic, as there is a positive probability of remaining in these states, thus $K_{33}^{(m)} > 0$ and $K_{66}^{(m)} > 0$ for all m, thus $d(3) = d(6) = 1$. ◁

In example 1.6 all states within one communicating class had the same period. This holds in general, as the following proposition shows:

Proposition 1.14. *(a) All states within a communicating class have the same period.*
(b) In an irreducible chain all states have the same period.

Proof. (a) Suppose $i \sim j$. Thus there are paths of positive probability between these two states. Suppose we can get from i to j in m_{ij} steps and from j to i in m_{ji} steps. Suppose also that we can get from j back to j in m_{jj} steps. Then we can get from i back to i in $m_{ij} + m_{ji}$ steps as well as in $m_{ij} + m_{jj} + m_{ji}$ steps. Thus $m_{ij} + m_{ji}$ and $m_{ij} + m_{jj} + m_{ji}$ must be divisible by the period $d(i)$ of state i. Thus m_{jj} is also divisible by $d(i)$ (being the difference of two numbers divisible by $d(i)$).

The above argument holds for any path between j and j, thus the length of any path from j back to j is divisible by $d(i)$. Thus $d(i) \leq d(j)$ ($d(j)$ being the greatest common denominator).

Repeating the same argument with the rôles of i and j swapped gives us $d(j) \leq d(i)$, thus $d(i) = d(j)$.

(b) An irreducible chain consists of a single communicating class, thus (b) is implied by (a). □

1.2.3 Recurrence and transience

If we follow the Markov chain of example 1.6 long enough, we will eventually end up switching between state 3 and 6 without ever coming back to the other states Whilst the states 3 and 6 will be visited infinitely often, the other states will eventually be left forever.

In order to formalise these notions we will introduce the *number of visits* in state i:

$$V_i = \sum_{t=0}^{+\infty} 1_{\{X_t = i\}}$$

The expected number of visits in state i given that we start the chain in i is

$$\mathbb{E}(V_i | X_0 = i) = \mathbb{E}\left(\sum_{t=0}^{+\infty} 1_{\{X_t=i\}} \Big| X_0 = i \right) = \sum_{t=0}^{+\infty} \mathbb{E}(1_{\{X_t=i\}} | X_0 = i) = \sum_{t=0}^{+\infty} \mathbb{P}(X_t = i | X_o = i) = \sum_{t=0}^{+\infty} k_{ii}^{(t)}$$

Based on whether the expected number of visits in a state is infinite or not, we will classify states as recurrent or transient:

Definition 1.15 (Recurrence and transience). *(a) A state i is called* recurrent *if* $\mathbb{E}(V_i | X_0 = i) = +\infty$.
(b) A state i is called transient *if* $\mathbb{E}(V_i | X_0 = i) < +\infty$.

One can show that a recurrent state will (almost surely) be visited infinitely often, whereas a transient state will (almost surely) be visited only a finite number of times.

In proposition 1.14 we have seen that within a communicating class either all states are aperiodic, or all states are periodic. A similar dichotomy holds for recurrence and transience.

Proposition 1.16. *Within a communicating class, either all states are transient or all states are recurrent.*

Proof. Suppose $i \sim j$. Then there exists a path of length m_{ij} leading from i to j and a path of length m_{ji} from j back to i, i.e. $k_{ij}^{(m_{ij})} > 0$ and $k_{ji}^{(m_{ji})} > 0$.

Suppose furthermore that the state i is transient, i.e. $\mathbb{E}(V_i | X_0 = i) = \sum_{t=0}^{+\infty} k_{ii}^{(t)} < +\infty$.

This implies

$$\begin{aligned}
\mathbb{E}(V_j | X_0 = j) = \sum_{t=0}^{+\infty} k_{jj}^{(t)} &= \frac{1}{k_{ij}^{(m_{ij})} k_{ji}^{(m_{ji})}} \sum_{t=0}^{+\infty} \underbrace{k_{ij}^{(m_{ij})} k_{jj}^{(t)} k_{ji}^{(m_{ji})}}_{\leq k_{ii}^{(m+t+n)}} \leq \frac{1}{k_{ij}^{(m_{ij})} k_{ji}^{(m_{ji})}} \sum_{t=0}^{+\infty} k^{(m_{ij}+t+m_{ji})} \\
&\leq \frac{1}{k_{ij}^{(m_{ij})} k_{ji}^{(m_{ji})}} \sum_{s=0}^{+\infty} k_{ii}^{(s)} < +\infty,
\end{aligned}$$

thus state j is be transient as well. □

Finally we state without proof two simple criteria for determining recurrence and transience.

Proposition 1.17. *(a) Every class which is not closed is transient.*
(b) Every finite closed class is recurrent.

Proof. For a proof see (Norris, 1997, sect. 1.5). □

Example 1.8 (Examples 1.6 and 1.7 continued). The chain of example 1.6 had three classes: $\{1\}$, $\{2,4,5\}$, and $\{3,6\}$. The classes $\{1\}$ and $\{2,4,5\}$ are not closed, so they are transient. The class $\{3,6\}$ is closed and finite, thus recurrent. ◁

Note that an infinite closed class is not necessarily recurrent. The random walk on $\mathbb{Z}$ studied in examples 1.2 and 1.4 is only recurrent if it is symmetric, i.e. $\alpha = \beta$, otherwise it drifts off to $-\infty$ or $+\infty$. An interesting result is that a symmetric random walk on $\mathbb{Z}^p$ is only recurrent if $p \leq 2$ (see e.g. Norris, 1997, sect. 1.6).

1.2.4 Invariant distribution and equilibrium

In this section we will study the long-term behaviour of Markov chains. A key concept for this is the invariant distribution.

Definition 1.18 (Invariant distribution). *Let $\boldsymbol{\mu} = (\mu_i)_{i\in S}$ be a probability distribution on the state space S, and let X be a Markov chain with transition kernel $\mathbf{K}$. Then $\boldsymbol{\mu}$ is called the* invariant distribution *(or* stationary distribution*) of the Markov chain X if*[3]

$$\boldsymbol{\mu}'\mathbf{K} = \boldsymbol{\mu}'.$$

If $\boldsymbol{\mu}$ is the stationary distribution of a chain with transition kernel $\mathbf{K}$, then

$$\boldsymbol{\mu}' = \underbrace{\boldsymbol{\mu}'}_{=\boldsymbol{\mu}'\mathbf{K}}\mathbf{K} = \boldsymbol{\mu}'\mathbf{K}^2 = \ldots = \boldsymbol{\mu}'\mathbf{K}^m = \boldsymbol{\mu}'\mathbf{K}^{(m)}$$

for all $m \in \mathbb{N}$. Thus if X_0 in drawn from $\boldsymbol{\mu}$, then all X_m have distribution $\boldsymbol{\mu}$: according to proposition 1.10

$$\mathbb{P}(X_m = j) = (\boldsymbol{\mu}'\mathbf{K}^{(m)})_j = (\boldsymbol{\mu})_j$$

for all m. Thus, if the chain has $\boldsymbol{\mu}$ as initial distribution, the distribution of X will not change over time.

Example 1.9 (Phone line (continued)). In example 1.1,1.3, and 1.5 we studied a Markov chain with the two states 0 ("free") and 1 ("in use") and which modeled whether a phone is free or not. Its transition kernel was

$$\mathbf{K} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}.$$

To find the invariant distribution, we need to solve $\boldsymbol{\mu}'\mathbf{K} = \boldsymbol{\mu}'$ for $\boldsymbol{\mu}$, which is equivalent to solving the following system of linear equations:

$$(\mathbf{K}' - \mathbf{I})\boldsymbol{\mu} = \mathbf{0}, \qquad \text{i.e.} \qquad \begin{pmatrix} -0.1 & 0.3 \\ 0.1 & -0.3 \end{pmatrix} \cdot \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

It is easy to see that the corresponding system is under-determined and that $-\mu_0 + 3\mu_1 = 0$, i.e. $\boldsymbol{\mu} = (\mu_0, \mu_1)' \propto (3,1)$, i.e. $\boldsymbol{\mu} = \left(\frac{3}{4}, \frac{1}{4}\right)'$ (as $\boldsymbol{\mu}$ has to be a probability distribution, thus $\mu_0 + \mu_1 = 1$). ◁

Not every Markov chain has an invariant distribution. The random walk on $\mathbb{Z}$ (studied in examples 1.2 and 1.4) for example does not have an invariant distribution, as the following example shows:

Example 1.10 (Random walk on $\mathbb{Z}$ (continued)). The random walk on $\mathbb{Z}$ had the transition kernel (see example 1.4)

$$\mathbf{K} = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \alpha & 1-\alpha-\beta & \beta & 0 & 0 & 0 & \ddots \\ \ddots & 0 & \alpha & 1-\alpha-\beta & \beta & 0 & 0 & \ddots \\ \ddots & 0 & 0 & \alpha & 1-\alpha-\beta & \beta & 0 & \ddots \\ \ddots & 0 & 0 & 0 & \alpha & 1-\alpha-\beta & \beta & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

[3] i.e. $\boldsymbol{\mu}$ is the left eigenvector of $\mathbf{K}$ corresponding to the eigenvalue 1.

As $\alpha + (1-\alpha-\beta) + \beta = 1$ we have for $\boldsymbol{\mu} = (\ldots, 1, 1, 1, \ldots)'$ that $\boldsymbol{\mu}'\mathbf{K} = \boldsymbol{\mu}'$, however $\boldsymbol{\mu}$ cannot be renormalised to become a probability distribution. ◁

We will now show that if a Markov chain is irreducible and aperiodic, its distribution will in the long run tend to the invariant distribution.

Theorem 1.19 (Convergence to equilibrium). *Let X be an irreducible and aperiodic Markov chain with invariant distribution $\boldsymbol{\mu}$. Then*

$$\mathbb{P}(X_t = i) \overset{t\to+\infty}{\longrightarrow} \mu_i$$

for all states i.

Outline of the proof. We will explain the outline of the proof using the idea of coupling.

Suppose that X has initial distribution $\boldsymbol{\lambda}$ and transition kernel $\mathbf{K}$. Define a new Markov chain Y with initial distribution $\boldsymbol{\mu}$ and same transition kernel $\mathbf{K}$. Let T be the first time the two chains "meet" in the state i, i.e.

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

Then one can show that $\mathbb{P}(T < \infty) = 1$ and define a new process Z by

$$Z_t = \begin{cases} X_t & \text{if } t \leq T \\ Y_t & \text{if } t > T \end{cases}$$

Figure 1.6 illustrates this new chain Z. One can show that Z is a Markov chain with initial distribution $\boldsymbol{\lambda}$ (as

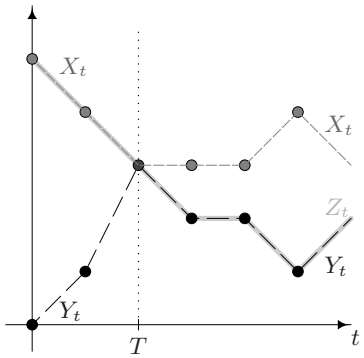


Figure 1.6. Illustration of the chains X (– – –), Y (— —) and Z (thick line) used in the proof of theorem 1.19.

$X_0 = Z_0$) and transition kernel $\mathbf{K}$ (as both X and Y have the transition kernel $\mathbf{K}$). Thus X and Z have the same distribution and for all $t \in \mathbb{N}_0$ we have that $\mathbb{P}(X_t = j) = \mathbb{P}(Z_t = j)$ for all states $j \in S$.
The chain Y has its invariant distribution as initial distribution, thus $\mathbb{P}(Y_t = j) = \mu_j$ for all $t \in \mathbb{N}_0$ and $j \in S$.
As $t \to +\infty$ the probability of $\{Y_t = Z_t\}$ tends to 1, thus

$$\mathbb{P}(X_t = j) = \mathbb{P}(Z_t = j) \to \mathbb{P}(Y_t = j) = \mu_j.$$

A more detailed proof of this theorem can be found in (Norris, 1997, sec. 1.8).

Example 1.11 (Phone line (continued)). We have shown in example 1.9 that the invariant distribution of the Markov chain modeling the phone line is $\boldsymbol{\mu} = \left(\frac{3}{4}, \frac{1}{4}\right)$, thus according to theorem 1.19 $\mathbb{P}(X_t = 0) \to \frac{3}{4}$ and $\mathbb{P}(X_t = 1) \to \frac{1}{4}$. Thus, in the long run, the phone will be free 75% of the time. ◁

Example 1.12. This example illustrates that the aperiodicity condition in theorem 1.19 is necessary.
Consider a Markov chain X with two states $S = \{1, 2\}$ and transition kernel

$$\mathbf{K} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This Markov chains switches deterministically, thus goes either $1,\ 0,\ 1,\ 0,\ \ldots$ or $0,\ 1,\ 0,\ 1,\ \ldots$. Thus it is periodic with period 2.
Its invariant distribution is $\boldsymbol{\mu}' = \left(\frac{1}{2}, \frac{1}{2}\right)$, as

$$\boldsymbol{\mu}'\mathbf{K} = \left(\frac{1}{2}, \frac{1}{2}\right)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \left(\frac{1}{2}, \frac{1}{2}\right) = \boldsymbol{\mu}'.$$

However if the chain is started in $X_0 = 1$, i.e. $\boldsymbol{\lambda} = (1, 0)$, then

$$\mathbb{P}(X_t = 0) = \begin{cases} 1 & \text{if } t \text{ is odd} \\ 0 & \text{if } t \text{ is even} \end{cases}, \qquad \mathbb{P}(X_t = 1) = \begin{cases} 0 & \text{if } t \text{ is odd} \\ 1 & \text{if } t \text{ is even} \end{cases},$$

which is different from the invariant distribution, under which all these probabilities would be $\frac{1}{2}$. ◁

1.2.5 Reversibility and detailed balance

In our study of Markov chains we have so far focused on conditioning on the past. For example, we have defined the transition kernel to consist of $k_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$. What happens if we analyse the distribution of X_t conditional on the future, i.e we turn the universal clock backwards?

$$\mathbb{P}(X_t = j | X_{t+1} = i) = \frac{\mathbb{P}(X_t = j, X_{t+1} = i)}{\mathbb{P}(X_{t+1} = i)} = \mathbb{P}(X_{t+1} = i | X_t = j) \cdot \frac{\mathbb{P}(X_t = j)}{\mathbb{P}(X_{t+1} = i)}$$

This suggests defining a new Markov chain which goes back in time. As the defining property of a Markov chain was that the past and future are conditionally independent given the present, the same should hold for the "backward chain", just with the rôles of past and future swapped.

Definition 1.20 (Time-reversed chain). *For $\tau \in \mathbb{N}$ let $\{X_t :\ t = 0, \ldots, \tau\}$ be a Markov chain. Then $\{Y_t :\ t = 0, \ldots, \tau\}$ defined by $Y_t = X_{\tau - t}$ is called the* time-reversed chain *corresponding to X.*

We have that

$$\begin{aligned}\mathbb{P}(Y_t = j | Y_{t-1} = i) &= \mathbb{P}(X_{\tau-t} = j | X_{\tau-t+1} = i) = \mathbb{P}(X_s = j | X_{s+1} = i) = \frac{\mathbb{P}(X_s = j, X_{s+1} = i)}{\mathbb{P}(X_{s+1} = i)} \\ &= \mathbb{P}(X_{s+1} = i | X_s = j) \cdot \frac{\mathbb{P}(X_s = j)}{\mathbb{P}(X_{s+1} = i)} = k_{ji} \cdot \frac{\mathbb{P}(X_s = j)}{\mathbb{P}(X_{s+1} = i)},\end{aligned}$$

thus the time-reversed chain is in general not homogeneous, even if the forward chain X is homogeneous.

This changes however if the forward chain X is initialised according to its invariant distribution $\boldsymbol{\mu}$. In this case $\mathbb{P}(X_{s+1} = i) = \mu_i$ and $\mathbb{P}(X_s = j) = \mu_j$ for all s, and thus Y is a homogeneous Markov chain with transition probabilities

$$\mathbb{P}(Y_t = j | Y_{t-1} = i) = k_{ji} \cdot \frac{\mu_j}{\mu_i}. \tag{1.2}$$

In general, the transition probabilities for the time-reversed chain will thus be different from the forward chain.

Example 1.13 (Phone line (continued)). In the example of the phone line (examples 1.1, 1.3, 1.5, 1.9, and 1.11) the transition matrix was

$$\mathbf{K} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}.$$

The invariant distribution was $\boldsymbol{\mu} = \left(\frac{3}{4}, \frac{1}{4}\right)'$.
If we use the invariant distribution $\boldsymbol{\mu}$ as initial distribution for X_0, then using (1.2)

$$\mathbb{P}(Y_t = 0|Y_{t-1} = 0) = k_{00} \cdot \frac{\mu_0}{\mu_0} = k_{00} = \mathbb{P}(X_t = 0|X_{t-1} = 1)$$
$$\mathbb{P}(Y_t = 0|Y_{t-1} = 1) = k_{01} \cdot \frac{\mu_0}{\mu_1} = 0.1 \cdot \frac{\frac{3}{4}}{\frac{1}{4}} = 0.3 = k_{10} = \mathbb{P}(X_t = 0|X_{t-1} = 1)$$
$$\mathbb{P}(Y_t = 1|Y_{t-1} = 0) = k_{10} \cdot \frac{\mu_1}{\mu_0} = 0.3 \cdot \frac{\frac{1}{4}}{\frac{3}{4}} = 0.1 = k_{01} = \mathbb{P}(X_t = 1|X_{t-1} = 0)$$
$$\mathbb{P}(Y_t = 1|Y_{t-1} = 1) = k_{11} \cdot \frac{\mu_1}{\mu_1} = k_{11} = \mathbb{P}(X_t = 1|X_{t-1} = 1)$$

Thus in this case both the forward chain X and the time-reversed chain Y have the same transition probabilities. We will call such chains *time-reversible*, as their dynamics do not change when time is reversed. ◁

We will now introduce a criterion for checking whether a chain is time-reversible.

Definition 1.21 (Detailed balance). *A transition kernel $\mathbf{K}$ is said to be in* detailed balance *with a distribution $\boldsymbol{\mu}$ if for all $i, j \in S$*

$$\mu_i k_{ij} = \mu_j k_{ji}.$$

It is easy to see that Markov chain studied in the phone line example (see example 1.13) satisfies the detailed-balance condition.

The detailed-balance condition is a very important concept that we will require when studying Markov Chain Monte Carlo (MCMC) algorithms later. The reason for its relevance is the following theorem, which says that if a Markov chain is in detailed balance with a distribution $\boldsymbol{\mu}$, then the chain is time-reversible, and, more importantly, $\boldsymbol{\mu}$ is the invariant distribution. The advantage of the detailed-balance condition over the condition of definition 1.18 is that the detailed-balance condition is often simpler to check, as it does not involve a sum (or a vector-matrix product).

Theorem 1.22. *Let X be a Markov chain with transition kernel $\mathbf{K}$ which is in detailed balance with some distribution $\boldsymbol{\mu}$ on the states of the chain. Then*

i. $\boldsymbol{\mu}$ is the invariant distribution of X.

ii. If initialised according to $\boldsymbol{\mu}$, X is time-reversible, i.e. both X and its time reversal have the same transition kernel.

Proof. i. We have that

$$(\boldsymbol{\mu}'\mathbf{K})_i = \sum_j \underbrace{\mu_j k_{ji}}_{=\mu_i k_{ij}} = \mu_i \underbrace{\sum_j k_{ij}}_{=1} = \mu_i,$$

thus $\boldsymbol{\mu}'\mathbf{K} = \boldsymbol{\mu}'$, i.e. $\boldsymbol{\mu}$ is the invariant distribution.

ii. Let Y be the time-reversal of X, then using (1.2)

$$\mathbb{P}(Y_t = j|Y_{t-1} = i) = \frac{\overbrace{\mu_j k_{ji}}^{\mu_i k_{ij}}}{\mu_i} = k_{ij} = \mathbb{P}(X_t = j|X_{t-1} = i),$$

thus X and Y have the same transition probabilities. □

Note that not every chain which has an invariant distribution is time-reversible, as the following example shows:

Example 1.14. Consider the following Markov chain on $S = \{1, 2, 3\}$ with transition matrix

$$\mathbf{K} = \begin{pmatrix} 0 & 0.8 & 0.2 \\ 0.2 & 0 & 0.8 \\ 0.8 & 0.2 & 0 \end{pmatrix}$$

The corresponding Markov graph is shown in figure 1.7: The stationary distribution of the chain is $\boldsymbol{\mu} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

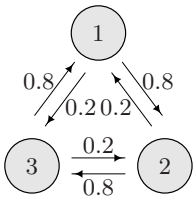


Figure 1.7. Markov graph for the Markov chain of example 1.14.

However the distribution is not time-reversible. Using equation (1.2) we can find the transition matrix of the time-reversed chain Y, which is

$$\begin{pmatrix} 0 & 0.2 & 0.8 \\ 0.8 & 0 & 0.2 \\ 0.2 & 0.8 & 0 \end{pmatrix},$$

which is equal to $\mathbf{K}'$, rather than $\mathbf{K}$. Thus the chains X and its time reversal Y have different transition kernels. When going forward in time, the chain is much more likely to go clockwise in figure 1.7; when going backwards in time however, the chain is much more likely to go counter-clockwise. ◁

1.3 General state space Markov chains

So far, we have restricted our attention to Markov chains with a discrete (i.e. at most countable) state space S. The main reason for this was that this kind of Markov chain is much easier to analyse than Markov chains having a more general state space.

However, most applications of Markov Chain Monte Carlo algorithms are concerned with continuous random variables, i.e. the corresponding Markov chain has a continuous state space S, thus the theory studied in the preceding section does not directly apply. Largely, we defined most concepts for discrete state spaces by looking at events of the type $\{X_t = j\}$, which is only meaningful if the state space is discrete.

In this section we will give a brief overview of the theory underlying Markov chains with general state spaces. Although the basic principles are not entirely different from the ones we have derived in the discrete case, the study of general state space Markov chains involves many more technicalities and subtleties, so that we will not present any proofs here. The interested reader can find a more rigorous treatment in (Meyn and Tweedie, 1993), (Nummelin, 1984), or (Robert and Casella, 2004, chapter 6).

Though this section is concerned with general state spaces we will notationally assume that the state space is $S = \mathbb{R}^d$.

First of all, we need to generalise our definition of a Markov chain (definition 1.4). We defined a Markov chain to be a stochastic process in which, conditionally on the present, the past and the future are independent. In the discrete case we formalised this idea using the conditional probability of $\{X_t = j\}$ given different collections of past events.

In a general state space it can be that all events of the type $\{X_t = j\}$ have probability 0, as it is the case for a process with a continuous state space. A process with a continuous state space spreads the probability so thinly that the probability of exactly hitting one given state is 0 for all states. Thus we have to work with conditional probabilities of sets of states, rather than individual states.

Definition 1.23 (Markov chain). *Let X be a stochastic process in discrete time with general state space S. X is called a* Markov chain *if X satisfies the* Markov property

$$\mathbb{P}(X_{t+1} \in A | X_0 = x_0, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} \in A | X_t = x_t)$$

for all measurable sets $A \subset S$.

If S is at most countable, this definition is equivalent to definition 1.4.

In the following we will assume that the Markov chain is *homogeneous*, i.e. the probabilities $\mathbb{P}(X_{t+1} \in A | X_t = x_t)$ are independent of t. For the remainder of this section we shall also assume that we can express the probability from definition 1.23 using a *transition kernel* $K : S \times S \to \mathbb{R}_0^+$:

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t) = \int_A K(x_t, x_{t+1})\, dx_{t+1} \tag{1.3}$$

where the integration is with respect to a suitable dominating measure, i.e. for example with respect to the Lebesgue measure if $S = \mathbb{R}^d$.[4] The transition kernel $K(x, y)$ is thus just the conditional probability density of X_{t+1} given $X_t = x_t$.

We obtain the special case of definition 1.8 by setting $K(i, j) = k_{ij}$, where k_{ij} is the (i, j)-th element of the transition matrix $\mathbf{K}$. For a discrete state space the dominating measure is the counting measure, so integration just corresponds to summation, i.e. equation (1.3) is equivalent to

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t) = \sum_{x_{t+1} \in A} k_{x_t x_{t+1}}.$$

We have for measurable set $A \subset S$ that

$$\mathbb{P}(X_{t+m} \in A | X_t = x_t) = \int_A \int_S \cdots \int_S K(x_t, x_{t+1}) K(x_{t+1}, x_{t+2}) \cdots K(x_{t+m-1}, x_{t+m})\, dx_{t+1} \cdots dx_{t+m-1} dx_{t+m},$$

thus the m-step transition kernel is

$$K^{(m)}(x_0, x_m) = \int_S \cdots \int_S K(x_0, x_1) \cdots K(x_{m-1}, x_m)\, dx_{m-1} \cdots dx_1$$

The m-step transition kernel allows for expressing the m-step transition probabilities more conveniently:

$$\mathbb{P}(X_{t+m} \in A | X_t = x_t) = \int_A K^{(m)}(x_t, x_{t+m})\, dx_{t+m}$$

Example 1.15 (Gaussian random walk on $\mathbb{R}$). Consider the random walk on $\mathbb{R}$ defined by

$$X_{t+1} = X_t + E_t,$$

where $E_t \sim \mathsf{N}(0, 1)$, i.e. the probability density function of E_t is $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$. This is equivalent to assuming that

$$X_{t+1} | X_t = x_t \sim \mathsf{N}(x_t, 1).$$

We also assume that E_t is independent of $X_0, E_1, \ldots, E_{t-1}$. Suppose that $X_0 \sim \mathsf{N}(0, 1)$. In contrast to the random walk on $\mathbb{Z}$ (introduced in example 1.2) the state space of the Gaussian random walk is $\mathbb{R}$. In complete analogy with example 1.2 we have that

$$\begin{aligned}\mathbb{P}(X_{t+1} \in A | X_t = x_t, \ldots, X_0 = x_0) &= \mathbb{P}(E_t \in A - x_t | X_t = x_t, \ldots, X_0 = x_0) \\ &= \mathbb{P}(E_t \in A - x_t) = \mathbb{P}(X_{t+1} \in A | X_t = x_t),\end{aligned}$$

where $A - x_t = \{a - x_t : \ a \in A\}$. Thus X is indeed a Markov chain. Furthermore we have that

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t) = \mathbb{P}(E_t \in A - x_t) = \int_A \phi(x_{t+1} - x_t)\, dx_{t+1}$$

Thus the transition kernel (which is nothing other than the conditional density of $X_{t+1} | X_t = x_t$) is thus

$$K(x_t, x_{t+1}) = \phi(x_{t+1} - x_t)$$

To find the m-step transition kernel we could use equation (1.3). However, the resulting integral is difficult to compute. Rather we exploit the fact that

[4] A more correct way of stating this would be $\mathbb{P}(X_{t+1} \in A | X_t = x_t) = \int_A K(x_t, dx_{t+1})$.

$$X_{t+m} = X_t + \underbrace{E_t + \ldots + E_{t+m-1}}_{\sim \mathsf{N}(0,m)},$$

thus $X_{t+m}|X_t = x_t \sim \mathsf{N}(x_t, m)$.

$$\mathbb{P}(X_{t+m} \in A|X_t = x_t) = \mathbb{P}(X_{t+m} - X_t \in A - x_t) = \int_A \frac{1}{\sqrt{m}}\phi\left(\frac{x_{t+m} - x_t}{\sqrt{m}}\right) dx_{t+m}$$

Comparing this with (1.3) we can identify

$$K^{(m)}(x_t, x_{t+m}) = \frac{1}{\sqrt{m}}\phi\left(\frac{x_{t+m} - x_t}{\sqrt{m}}\right)$$

as m-step transition kernel. ◁

In section 1.2.2 we defined a Markov chain to be irreducible if there is a positive probability of getting from any state $i \in S$ to any other state $j \in S$, possibly via intermediate steps.

Again, we cannot directly apply definition 1.12 to Markov chains with general state spaces: it might be — as it is the case for a continuous state space — that the probability of hitting a given state is 0 for all states. We will again resolve this by looking at sets of states rather than individual states.

Definition 1.24 (Irreducibility). *Given a distribution μ on the states S, a Markov chain is said to be* μ-irreducible *if for all sets A with $\mu(A) > 0$ and for all $x \in S$, there exists an $m \in \mathbb{N}_0$ such that*

$$\mathbb{P}(X_{t+m} \in A|X_t = x) = \int_A K^{(m)}(x, y)\, dy > 0.$$

If the number of steps $m = 1$ for all A, then the chain is said to be strongly μ-irreducible.

Example 1.16 (Gaussian random walk (continued)). In example 1.15 we had that $X_{t+1}|X_t = x_t \sim \mathsf{N}(x_t, 1)$. As the range of the Gaussian distribution is $\mathbb{R}$, we have that $\mathbb{P}(X_{t+1} \in A|X_t = x_t) > 0$ for all sets A of non-zero Lebesgue measure. Thus the chain is strongly irreducible with the respect to any continuous distribution. ◁

Extending the concepts of periodicity, recurrence, and transience studied in sections 1.2.2 and 1.2.3 from the discrete case to the general case requires additional technical concepts like *atoms* and *small sets*, which are beyond the scope of this course (for a more rigorous treatment of these concepts see e.g. Robert and Casella, 2004, sections 6.3 and 6.4). Thus we will only generalise the concept of recurrence.

In section 1.2.3 we defined a discrete Markov chain to be recurrent, if all states are (on average) visited infinitely often. For more general state spaces, we need to consider the number of visits to a set of states rather than single states. Let $V_A = \sum_{t=0}^{+\infty} 1_{\{X_t \in A\}}$ be the number of visits the chain makes to states in the set $A \subset S$. We then define the expected number of visits in $A \subset S$, when we start the chain in $x \in S$:

$$\mathbb{E}(V_A|X_0 = x) = \mathbb{E}\left(\sum_{t=0}^{+\infty} 1_{\{X_t \in A\}} \Big| X_0 = x\right) = \sum_{t=0}^{+\infty} \mathbb{E}(1_{\{X_t \in A\}}|X_0 = x) = \sum_{t=0}^{+\infty} \int_A K^{(t)}(x, y)\, dy$$

This allows us to define recurrence for general state spaces. We start with defining recurrence of sets before extending the definition of recurrence of an entire Markov chain.

Definition 1.25 (Recurrence). *(a) A set $A \subset S$ is said to be* recurrent *for a Markov chain X if for all $x \in A$*

$$\mathbb{E}(V_A|X_0 = x) = +\infty,$$

(b) A Markov chain is said to be recurrent, *if*

i. The chain is μ-irreducible for some distribution μ.

ii. Every measurable set $A \subset S$ with $\mu(A) > 0$ is recurrent.

According to the definition a set is recurrent if on average it is visited infinitely often. This is already the case if there is a non-zero probability of visiting the set infinitely often. A stronger concept of recurrence can be obtained if we require that the set is visited infinitely often with probability 1. This type of recurrence is referred to as *Harris recurrence*.

Definition 1.26 (Harris Recurrence). *(a) A set $A \subset S$ is said to be* Harris-recurrent *for a Markov chain X if for all $x \in A$*

$$\mathbb{P}(V_A = +\infty | X_0 = x) = 1,$$

(b) A Markov chain is said to be Harris-recurrent*, if*

i. The chain is μ-irreducible for some distribution μ.

ii. Every measurable set $A \subset S$ with $\mu(A) > 0$ is Harris-recurrent.

It is easy to see that Harris recurrence implies recurrence. For discrete state spaces the two concepts are equivalent.

Checking recurrence or Harris recurrence can be very difficult. We will state (without) proof a proposition which establishes that if a Markov chain is irreducible and has a unique invariant distribution, then the chain is also recurrent.

However, before we can state this proposition, we need to define invariant distributions for general state spaces.

Definition 1.27 (Invariant Distribution). *A distribution μ with density function f_μ is said to be the* invariant distribution *of a Markov chain X with transition kernel K if*

$$f_\mu(y) = \int_S f_\mu(x) K(x, y)\, dx$$

for almost all $y \in S$.

Proposition 1.28. *Suppose that X is a μ-irreducible Markov chain having μ as unique invariant distribution. Then X is also recurrent.*

Proof. see (Tierney, 1994, theorem 1) or (Athreya et al., 1992) □

Checking the invariance condition of definition 1.27 requires computing an integral, which can be quite cumbersome. A simpler (sufficient, but not necessary) condition is, just like in the case discrete case, detailed balance.

Definition 1.29 (Detailed balance). *A transition kernel K is said to be in* detailed balance *with a distribution μ with density f_μ if for almost all $x, y \in S$*

$$f_\mu(x) K(x, y) = f_\mu(y) K(y, x).$$

In complete analogy with theorem 1.22 one can also show in the general case that if the transition kernel of a Markov chain is in detailed balance with a distribution μ, then the chain is time-reversible and has μ as its invariant distribution. Thus theorem 1.22 also holds in the general case.

1.4 Ergodic theorems

In this section we will study the question whether we can use observations from a Markov chain to make inferences about its invariant distribution. We will see that under some regularity conditions it is even enough to follow a single sample path of the Markov chain.

For independent identically distributed data the Law of Large Numbers is used to justify estimating the expected value of a functional using empirical averages. A similar result can be obtained for Markov chains. This result is the reason why Markov Chain Monte Carlo methods work: it allows us to set up simulation algorithms to generate a Markov chain, whose sample path we can then use for estimating various quantities of interest.

Theorem 1.30 (Ergodic Theorem). *Let X be a μ-irreducible, recurrent $\mathbb{R}^d$-valued Markov chain with invariant distribution μ. Then we have for any integrable function $g : \mathbb{R}^d \to \mathbb{R}$ that with probability 1*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} g(X_i) \to \mathbb{E}_\mu(g(X)) = \int_S g(x) f_\mu(x)\, dx$$

for almost every starting value $X_0 = x$. If X is Harris-recurrent this holds for every starting value x.

Proof. For a proof see (Roberts and Rosenthal, 2004, fact 5), (Robert and Casella, 2004, theorem 6.63), or (Meyn and Tweedie, 1993, theorem 17.3.2). □

Under additional regularity conditions one can also derive a Central Limit Theorem which can be used to justify Gaussian approximations for ergodic averages of Markov chains. This would however be beyond the scope of this course.

We conclude by giving an example that illustrates that the conditions of irreducibility and recurrence are necessary in theorem 1.30. These conditions ensure that the chain is permanently exploring the entire state space, which is a necessary condition for the convergence of ergodic averages.

Example 1.17. Consider a discrete chain with two states $S = \{1, 2\}$ and transition matrix

$$\mathbf{K} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The corresponding Markov graph is shown in figure 1.8. This chain will remain in its intial state forever. Any

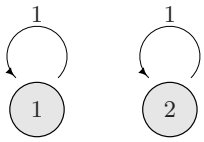


Figure 1.8. Markov graph of the chain of example 1.17

distribution $\boldsymbol{\mu}$ on $\{1, 2\}$ is an invariant distribution, as

$$\boldsymbol{\mu}'\mathbf{K} = \boldsymbol{\mu}'\mathbf{I} = \boldsymbol{\mu}'$$

for all $\boldsymbol{\mu}$. However, the chain is not irreducible (or recurrent): we cannot get from state 1 to state 2 and vice versa. If the initial distribution is $\boldsymbol{\mu} = (\alpha, 1-\alpha)'$ with $\alpha \in [0, 1]$ then for every $t \in \mathbb{N}_0$ we have that

$$\mathbb{P}(X_t = 1) = \alpha \qquad \mathbb{P}(X_t = 2) = 1 - \alpha.$$

By observing one sample path (which is either $1, 1, 1, \ldots$ or $2, 2, 2, \ldots$) we can make no inference about the distribution of X_t or the parameter α. The reason for this is that the chain fails to explore the space (i.e. switch between the states 1 and 2). In order to estimate the parameter α we would need to look at more than one sample path. ◁

Note that theorem 1.30 does not require the chain to the aperiodic. In example 1.12 we studied a periodic chain. Due to the periodicity we could not apply theorem 1.19. We can however apply theorem 1.30 to this chain. The reason for this is that whilst theorem 1.19 was about the distribution of states at a given time t, theorem 1.30 is about averages, and the periodic behaviour does not affect averages.