Chapter 7

Markov Chains

The Markov property is satisfied by any random process for which *the future is conditionally independent from the past given the present.*

Definition 7.0.1 (Markov property)**.** *A random process satisfies the Markov property if $\widetilde{X}(t_{i+1})$ is conditionally independent of $\widetilde{X}(t_1), \ldots, \widetilde{X}(t_{i-1})$ given $\widetilde{X}(t_i)$ for any $t_1 < t_2 < \ldots < t_i < t_{i+1}$. If the state space of the random process is discrete, then for any $x_1, x_2, \ldots, x_{i+1}$*

$$p_{\widetilde{X}(t_{n+1})|\widetilde{X}(t_1),\widetilde{X}(t_2),\ldots,\widetilde{X}(t_i)}(x_{n+1}|x_1, x_2, \ldots, x_n) = p_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_i)}(x_{i+1}|x_i). \tag{7.1}$$

If the state space of the random process is continuous (and the distribution has a joint pdf),

$$f_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_1),\widetilde{X}(t_2),\ldots,\widetilde{X}(t_i)}(x_{i+1}|x_1, x_2, \ldots, x_i) = f_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_i)}(x_{i+1}|x_i). \tag{7.2}$$

Figure 7.1 shows the directed graphical model that corresponds to the dependence assumptions implied by the Markov property. Any iid sequence satisfies the Markov property, since all conditional pmfs or pdfs are just equal to the marginals (in this case there would be no edges in the directed acyclic graph of Figure 7.1). The random walk also satisfies the property, since once we fix where the walk is at a certain time i the path that it took before i has no influence in its next steps.

Lemma 7.0.2. *The random walk satisfies the Markov property.*

Proof. Let $\widetilde{X}$ denote the random walk defined in Section 5.6. Conditioned on $\widetilde{X}(j) = x_i$ for $j \leq i$, $\widetilde{X}(i+1)$ equals $x_i + \widetilde{S}(i+1)$. This does not depend on $x_1, \ldots, x_{i-1}$, which implies (7.1). □

7.1 Time-homogeneous discrete-time Markov chains

A Markov chain is a random process that satisfies the Markov property. Here we consider **discrete-time** Markov chains with a **finite state space**, which means that the process can only take a finite number of values at any given time point. To specify a Markov chain, we only need to define the pmf of the random process at its starting point (which we will assume is at $i = 0$) and its transition probabilities. This follows from the Markov property, since for any

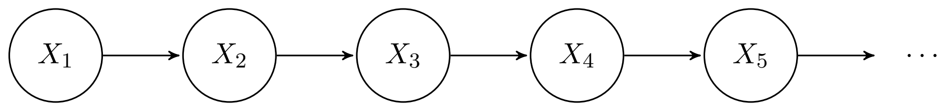

Figure 7.1: Directed graphical model describing the dependence assumptions implied by the Markov property.

$n \geq 0$

$$p_{\widetilde{X}(0),\widetilde{X}(1),\ldots,\widetilde{X}(n)}(x_0, x_1, \ldots, x_n) := \prod_{i=0}^{n} p_{\widetilde{X}(i)|\widetilde{X}(0),\ldots,\widetilde{X}(i-1)}(x_i|x_0, \ldots, x_{i-1}) \tag{7.3}$$

$$= \prod_{i=0}^{n} p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_i|x_{i-1}) \,. \tag{7.4}$$

If these transition probabilities are the same at every time step (i.e. they are constant and do not depend on i), then the Markov chain is said to be **time homogeneous**. In this case, we can store the probability of each possible transition in an $s \times s$ matrix $T_{\widetilde{X}}$, where s is the number of states.

$$\left(T_{\widetilde{X}}\right)_{jk} := p_{\widetilde{X}(i+1)|\widetilde{X}(i)}(x_j|x_k) \,. \tag{7.5}$$

In this chapter we focus on time-homogeneous finite-state Markov chains. The transition probabilities of these chains can be visualized using a state diagram, which shows each state and the probability of every possible transition. See Figure 7.2 below for an example. The state diagram should not be confused with the directed acyclic graph (DAG) that represents the dependence structure of the model, illustrated in Figure 7.1. In the state diagram, each node corresponds to a state and the edges to transition probabilities between states, whereas the DAG just indicates the dependence structure of the random process in time and is usually the same for all Markov chains.

To simplify notation we define an s-dimensional vector $\vec{p}_{\widetilde{X}(i)}$ called the **state vector**, which contains the marginal pmf of the Markov chain at each time i,

$$\vec{p}_{\widetilde{X}(i)} := \begin{bmatrix} p_{\widetilde{X}(i)}(x_1) \\ p_{\widetilde{X}(i)}(x_2) \\ \cdots \\ p_{\widetilde{X}(i)}(x_s) \end{bmatrix} . \tag{7.6}$$

Each entry in the state vector contains the probability that the Markov chain is in that particular state at time i. It is *not* the value of the Markov chain, which is a random variable.

The initial state space $\vec{p}_{\widetilde{X}(0)}$ and the transition matrix $T_{\widetilde{X}}$ suffice to completely specify a time-homogeneous finite-state Markov chain. Indeed, we can compute the joint distribution of the chain at any n time points i_1, i_2, …, i_n for any $n \geq 1$ from $\vec{p}_{\widetilde{X}(0)}$ and $T_{\widetilde{X}}$ by applying (7.4) and marginalizing over any times that we are not interested in. We illustrate this in the following example.

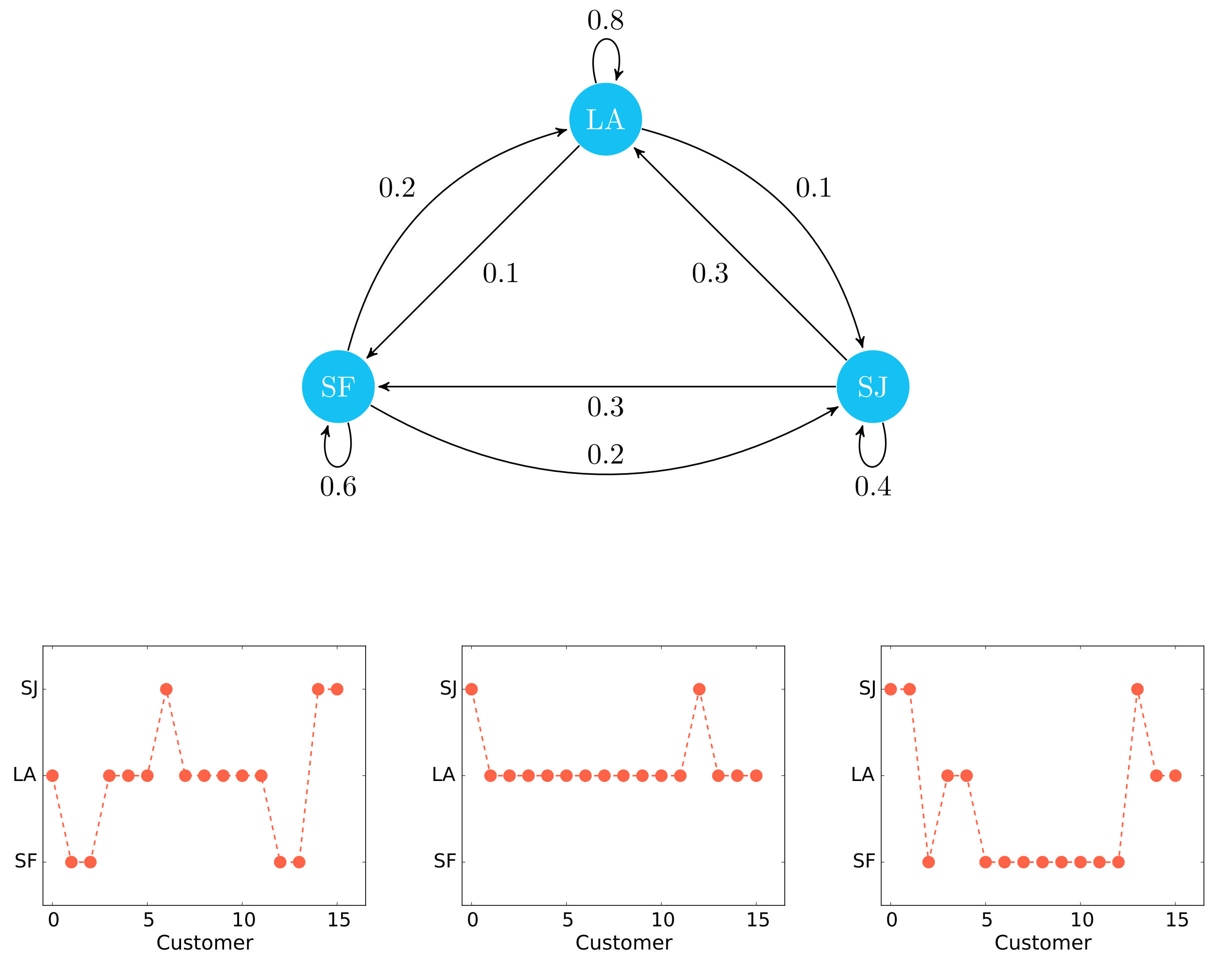

Figure 7.2: State diagram of the Markov chain described in Example (7.1.1) (top). Each arrow shows the probability of a transition between the two states. Below we show three realizations of the Markov chain.

Example 7.1.1 (Car rental)**.** A car-rental company hires you to model the location of their cars. The company operates in Los Angeles, San Francisco and San Jose. Customers regularly take a car in a city and drop it off in another. It would be very useful for the company to be able to compute how likely it is for a car to end up in a given city. You decide to model the location of the car as a Markov chain, where each time step corresponds to a new customer taking the car. The company allocates new cars evenly between the three cities. The transition probabilities, obtained from past data, are given by

$$
\begin{array}{cc}
\begin{array}{ccc} \text{San Francisco} & \text{Los Angeles} & \text{San Jose} \end{array} & \\
\left(\begin{array}{ccc} 0.6 & 0.1 & 0.3 \\ 0.2 & 0.8 & 0.3 \\ 0.2 & 0.1 & 0.4 \end{array}\right) & \begin{array}{c} \text{San Francisco} \\ \text{Los Angeles} \\ \text{San Jose} \end{array}
\end{array}
$$

To be clear, the probability that a customer moves the car from San Francisco to LA is 0.2, the probability that the car stays in San Francisco is 0.6, and so on.

The initial state vector and the transition matrix of the Markov chain are

$$\vec{p}_{\widetilde{X}(0)} := \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \qquad T_{\widetilde{X}} := \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.2 & 0.8 & 0.3 \\ 0.2 & 0.1 & 0.4 \end{bmatrix}. \tag{7.7}$$

State 1 is assigned to *San Francisco*, state 2 to *Los Angeles* and state 3 to *San Jose*. Figure 7.2 shows a state diagram of the Markov chain. Figure 7.2 shows some realizations of the Markov chain.

The company wants to find out the probability that the car starts in San Francisco, but is in San Jose right after the second customer. This is given by

$$p_{\widetilde{X}(0),\widetilde{X}(2)}(1,3) = \sum_{i=1}^{3} p_{\widetilde{X}(0),\widetilde{X}(1),\widetilde{X}(2)}(1,i,3) \tag{7.8}$$

$$= \sum_{i=1}^{3} p_{\widetilde{X}(0)}(1)\, p_{\widetilde{X}(1)|\widetilde{X}(0)}(i|1)\, p_{\widetilde{X}(2)|\widetilde{X}(1)}(3|i) \tag{7.9}$$

$$= \left(\vec{p}_{\widetilde{X}(0)}\right)_1 \sum_{i=1}^{3} \left(T_{\widetilde{X}}\right)_{i1} \left(T_{\widetilde{X}}\right)_{3i} \tag{7.10}$$

$$= \frac{0.6 \cdot 0.2 + 0.2 \cdot 0.1 + 0.2 \cdot 0.4}{3} \approx 7.33\, 10^{-2}. \tag{7.11}$$

The probability is 7.33%.

△

The following lemma provides a simple expression for the state vector at time i $\vec{p}_{\widetilde{X}(i)}$ in terms of $T_{\widetilde{X}}$ and the previous state vector.

Lemma 7.1.2 (State vector and transition matrix)**.** *For a Markov chain $\widetilde{X}$ with transition matrix $T_{\widetilde{X}}$*

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}\, \vec{p}_{\widetilde{X}(i-1)}. \tag{7.12}$$

If the Markov chain starts at time 0 then

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}^{i}\, \vec{p}_{\widetilde{X}(0)}, \tag{7.13}$$

where $T_{\widetilde{X}}^{i}$ denotes multiplying i times by matrix $T_{\widetilde{X}}$.

Proof. The proof follows directly from the definitions,

$$\vec{p}_{\widetilde{X}(i)} := \begin{bmatrix} p_{\widetilde{X}(i)}(x_1) \\ p_{\widetilde{X}(i)}(x_2) \\ \cdots \\ p_{\widetilde{X}(i)}(x_s) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_j) \\ \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_j) \\ \cdots \\ \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_j) \end{bmatrix} \tag{7.14}$$

$$= \begin{bmatrix} p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_s) \\ p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_s) \\ & & \cdots & \\ p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_s) \end{bmatrix} \begin{bmatrix} p_{\widetilde{X}(i-1)}(x_1) \\ p_{\widetilde{X}(i-1)}(x_2) \\ \cdots \\ p_{\widetilde{X}(i-1)}(x_s) \end{bmatrix}$$

$$= T_{\widetilde{X}}\, \vec{p}_{\widetilde{X}(i-1)} \tag{7.15}$$

Equation (7.13) is obtained by applying (7.12) i times and taking into account the Markov property. □

Example 7.1.3 (Car rental (continued)). The company wants to estimate the distribution of locations right after the 5th customer has used a car. Applying Lemma 7.1.2 we obtain

$$\vec{p}_{\widetilde{X}(5)} = T_{\widetilde{X}}^{5}\, \vec{p}_{\widetilde{X}(0)} \tag{7.16}$$

$$= \begin{bmatrix} 0.281 \\ 0.534 \\ 0.185 \end{bmatrix}. \tag{7.17}$$

The model estimates that after 5 customers more than half of the cars are in Los Angeles.

△

7.2 Recurrence

The states of a Markov chain can be classified depending on whether the Markov chain is guaranteed to always return to them or whether it may eventually stop visiting those states.

Definition 7.2.1 (Recurrent and transient states). *Let $\widetilde{X}$ be a time-homogeneous finite-state Markov chain. We consider a particular state x. If*

$$\mathrm{P}\left(\widetilde{X}(j) = s \text{ for some } j > i \mid \widetilde{X}(i) = s\right) = 1 \tag{7.18}$$

*then the state is **recurrent**. In words, given that the Markov chain is at x, the probability that it returns to x is one. In contrast, if*

$$\mathrm{P}\left(\widetilde{X}(j) \neq s \text{ for all } j > i \mid \widetilde{X}(i) = s\right) > 0 \tag{7.19}$$

*the state is **transient**. Given that the Markov chain is at x, there is nonzero probability that it will never return.*

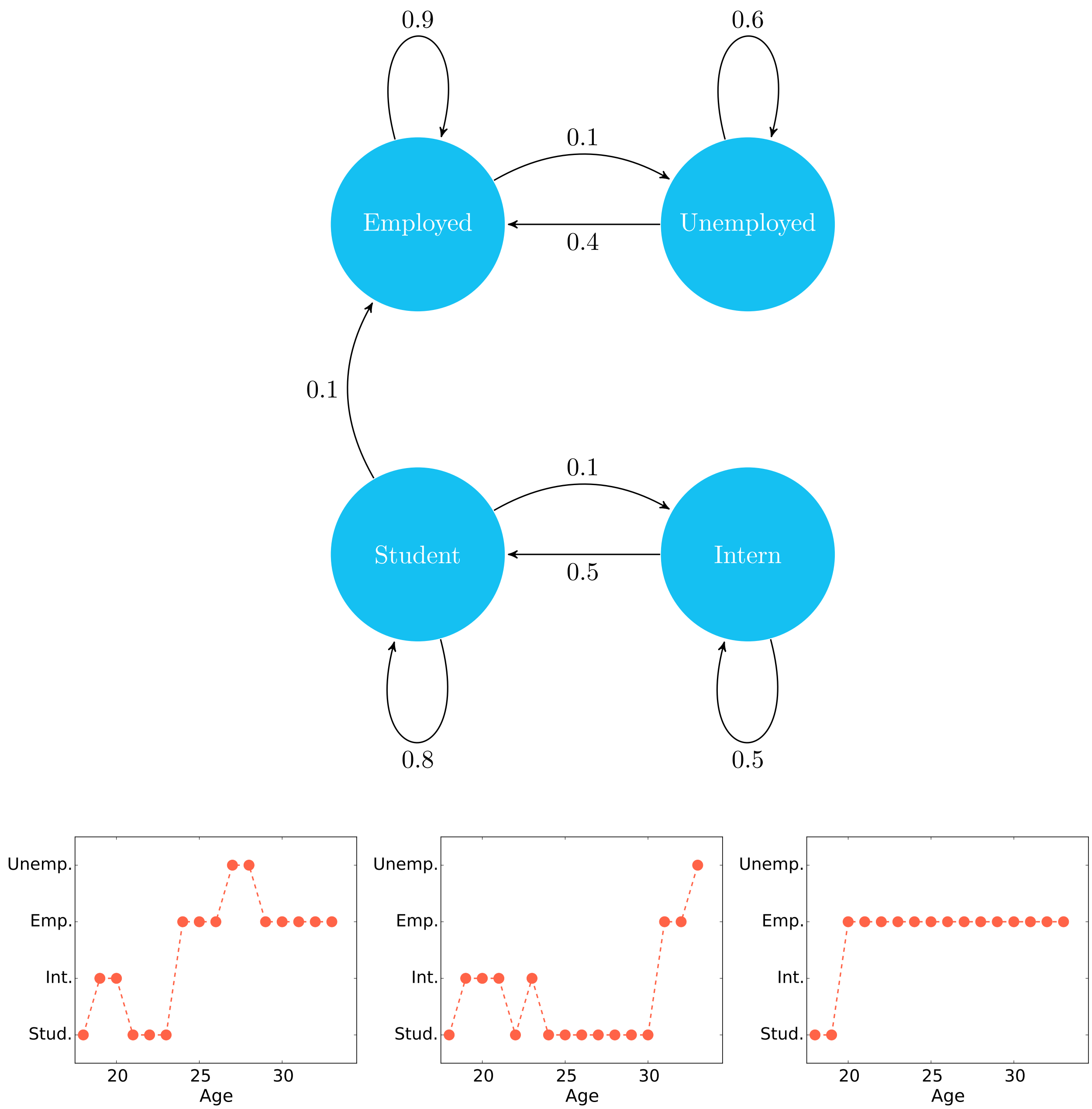

Figure 7.3: State diagram of the Markov chain described in Example (7.2.2) (top). Below we show three realizations of the Markov chain.

The following example illustrates the difference between recurrent and transient states.

Example 7.2.2 (Employment dynamics)**.** A researcher is interested in modeling the employment dynamics of young people using a Markov chain.

She determines that at age 18 a person is either a student with probability 0.9 or an intern with probability 0.1. After that she estimates the following transition probabilities:

$$
\begin{array}{cccccc}
 & \text{Student} & \text{Intern} & \text{Employed} & \text{Unemployed} & \\
\left(\begin{array}{c} \\ \\ \\ \\ \end{array}\right.\!\!\! &
\begin{array}{c} 0.8 \\ 0.1 \\ 0.1 \\ 0 \end{array} &
\begin{array}{c} 0.5 \\ 0.5 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 0 \\ 0 \\ 0.9 \\ 0.1 \end{array} &
\begin{array}{c} 0 \\ 0 \\ 0.4 \\ 0.6 \end{array} \!\!\!\left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right) &
\begin{array}{c} \text{Student} \\ \text{Intern} \\ \text{Employed} \\ \text{Unemployed} \end{array}
\end{array}
$$

The Markov assumption is obviously not completely precise, someone who has been a student for longer is probably less likely to remain a student, but such Markov models are easier to fit (we only need to estimate the transition probabilities) and often yield useful insights.

The initial state vector and the transition matrix of the Markov chain are

$$
\vec{p}_{\widetilde{X}(0)} := \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \\ 0 \end{bmatrix}, \qquad T_{\widetilde{X}} := \begin{bmatrix} 0.8 & 0.5 & 0 & 0 \\ 0.1 & 0.5 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0.4 \\ 0 & 0 & 0.1 & 0.6 \end{bmatrix}. \tag{7.20}
$$

Figure 7.3 shows the state diagram and some realizations of the Markov chain.

States 1 (student) and 2 (intern) are transient states. Note that the probability that the Markov chain returns to those states after visiting state 3 (employed) is zero, so

$$
\begin{aligned}
\mathrm{P}\left(\widetilde{X}(j) \neq 1 \text{ for all } j > i \mid \widetilde{X}(i) = 1\right) &\geq \mathrm{P}\left(\widetilde{X}(i+1) = 3 \,|\widetilde{X}(i) = 1\right) && (7.21)\\
&= 0.1 > 0, && (7.22)\\
\mathrm{P}\left(\widetilde{X}(j) \neq 2 \text{ for all } j > i \mid \widetilde{X}(i) = 2\right) &\geq \mathrm{P}\left(\widetilde{X}(i+2) = 3 \,|\widetilde{X}(i) = 2\right) && (7.23)\\
&= 0.5 \cdot 0.1 > 0. && (7.24)
\end{aligned}
$$

In contrast, states 3 and 4 (unemployed) are recurrent. We prove this for state 3 (the argument for state 4 is exactly the same):

$$
\begin{aligned}
&\mathrm{P}\left(\widetilde{X}(j) \neq 3 \text{ for all } j > i \mid \widetilde{X}(i) = 3\right) && (7.25)\\
&= \mathrm{P}\left(\widetilde{X}(j) = 4 \text{ for all } j > i \,|\widetilde{X}(i) = 3\right) && (7.26)\\
&= \lim_{k\to\infty} \mathrm{P}\left(\widetilde{X}(i+1) = 4 \,|\widetilde{X}(i) = 3\right) \prod_{j=1}^{k} \mathrm{P}\left(\widetilde{X}(i+j+1) = 4 \,|\widetilde{X}(i+j) = 4\right) && (7.27)\\
&= \lim_{k\to\infty} 0.1 \cdot 0.6^{k} && (7.28)\\
&= 0. && (7.29)
\end{aligned}
$$

△

In this example, it is not possible to reach the states *student* and *intern* from the states *employed* or *unemployed*. Markov chains for which there is a possible transition between any two states (even if it is not direct) are called irreducible.

Definition 7.2.3 (Irreducible Markov chain)**.** *A time-homogeneous finite-state Markov chain is irreducible if for any state x, the probability of reaching every other state $y \neq x$ in a finite number of steps is nonzero, i.e. there exists $m \geq 0$ such that*

$$\mathrm{P}\left(\widetilde{X}\,(i+m) = y \mid \widetilde{X}\,(i) = x\right) > 0. \tag{7.30}$$

One can easily check that the Markov chain in Example 7.1.1 is irreducible, whereas the one in Example 7.2.2. An important result is that all states in an irreducible Markov chain are recurrent.

Theorem 7.2.4 (Irreducible Markov chains)**.** *All states in an irreducible Markov chain are recurrent.*

Proof. In any finite-state Markov chain there must be at least one state that is recurrent. If all the states are transient there is a nonzero probability that it leaves all of the states forever, which is not possible. Without loss of generality let us assume that state x is recurrent. We now provide a sketch of a proof that another arbitrary state y must also be recurrent. To alleviate notation let

$$p_{x,x} := \mathrm{P}\left(\widetilde{X}\,(j) = x \text{ for some } j > i \mid \widetilde{X}\,(i) = x\right), \tag{7.31}$$

$$p_{x,y} := \mathrm{P}\left(\widetilde{X}\,(j) = y \text{ for some } j > i \mid \widetilde{X}\,(i) = x\right), \tag{7.32}$$

$$p_{y,x} := \mathrm{P}\left(\widetilde{X}\,(k) = x \text{ for some } j > i \mid \widetilde{X}\,(i) = y\right). \tag{7.33}$$

The chain is irreducible so there is a nonzero probability $p_m > 0$ of reaching y from x in at most m steps for some $m > 0$. The probability that the chain goes from x to y and never goes back to x is consequently at least $p_m\,(1 - p_{y,x})$. However, x is recurrent, so this probability must be zero! Since $p_m > 0$ this implies $p_{y,x} = 1$.

Consider the following event:

1. $\widetilde{X}$ goes from y to x.
2. $\widetilde{X}$ does not return to y in m steps after reaching x.
3. $\widetilde{X}$ eventually reaches x again at a time $m' > m$.

The probability of this event is equal to $p_{y,x}\,(1 - p_m)\,p_{x,x} = 1 - p_m$ (recall that x is recurrent so $p_{x,x} = 1$). Now imagine that steps 2 and 3 repeat k times, i.e. that $\widetilde{X}$ fails to go from x to y in m steps k times. The probability of this event is $p_{y,x}\,(1 - p_m)^k\,p_{x,x}^k = (1 - p_m)^k$. Taking $k \to \infty$ this is equal to zero for any m so the probability that $\widetilde{X}$ does not eventually return to x must be zero (this can be made rigorous, but the details are beyond the scope of these notes). □

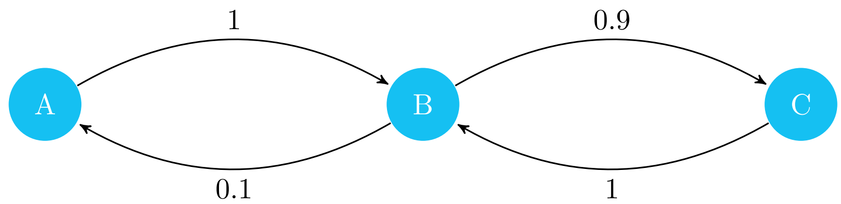

Figure 7.4: State diagram of a Markov chain where states the states have period two.

7.3 Periodicity

Another important consideration is whether the Markov chain always visits a given state at regular intervals. If this is the case, then the state has a period greater than one.

Definition 7.3.1 (Period of a state)**.** *Let $\widetilde{X}$ be a time-homogeneous finite-state Markov chain and x a state of the Markov chain. The period m of x is the largest integer such that it is only possible to return to x in a number of steps that is a multiple of m, i.e. we can only return in km steps with nonzero probability where k is a positive integer.*

Figure 7.4 shows a Markov chain where the states have a period equal to two. Aperiodic Markov chains do not contain states with periods greater than one.

Definition 7.3.2 (Aperiodic Markov chain)**.** *A time-homogeneous finite-state Markov chain $\widetilde{X}$ is aperiodic if all states have period equal to one.*

The Markov chains in Examples 7.1.1 and 7.2.2 are both aperiodic.

7.4 Convergence

In this section we study under what conditions a finite-state time-homogeneous Markov chain $\widetilde{X}$ converges in distribution. If a Markov chain converges in distribution, then its state vector $\vec{p}_{\widetilde{X}(i)}$, which contains the first order pmf of $\widetilde{X}$, converges to a fixed vector $\vec{p}_\infty$,

$$\vec{p}_\infty := \lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)}. \tag{7.34}$$

In that case the probability of the Markov chain being in each state eventually tends to a fixed value (which does *not* imply that the Markov chain will stay at a given state!).

By Lemma 7.1.2 we can express (7.34) in terms of the initial state vector and the transition matrix of the Markov chain

$$\vec{p}_\infty = \lim_{i\to\infty} T_{\widetilde{X}}^{\,i}\, \vec{p}_{\widetilde{X}(0)}. \tag{7.35}$$

Computing this limit analytically for a particular $T_{\widetilde{X}}$ and $\vec{p}_{\widetilde{X}(0)}$ may seem challenging at first sight. However, it is often possible to leverage the eigendecomposition of the transition matrix (if it exists) to find $\vec{p}_\infty$. This is illustrated in the following example.

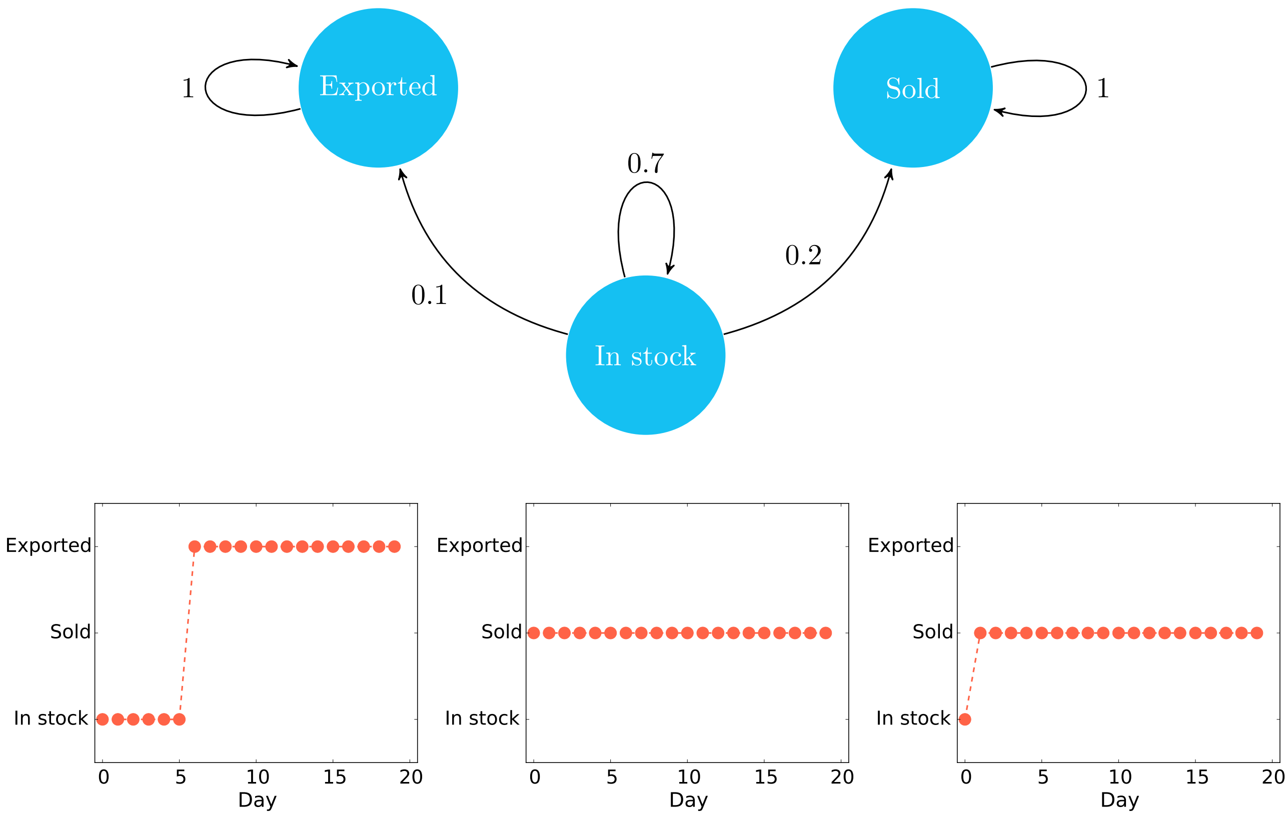

Figure 7.5: State diagram of the Markov chain described in Example (7.4.1) (top). Below we show three realizations of the Markov chain.

Example 7.4.1 (Mobile phones)**.** A company that makes mobile phones wants to model the sales of a new model they have just released. At the moment 90% of the phones are in stock, 10% have been sold locally and none have been exported. Based on past data, the company determines that each day a phone is sold with probability 0.2 and exported with probability 0.1. The initial state vector and the transition matrix of the Markov chain are

$$\vec{a} := \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \end{bmatrix}, \qquad T_{\widetilde{X}} = \begin{bmatrix} 0.7 & 0 & 0 \\ 0.2 & 1 & 0 \\ 0.1 & 0 & 1 \end{bmatrix}. \tag{7.36}$$

We have used $\vec{a}$ to denote $\vec{p}_{\widetilde{X}(0)}$ because later we will consider other possible initial state vectors. Figure 7.6 shows the state diagram and some realizations of the Markov chain.

The company is interested in the fate of the new model. In particular, it would like to compute what fraction of mobile phones will end up exported and what fraction will be sold locally. This

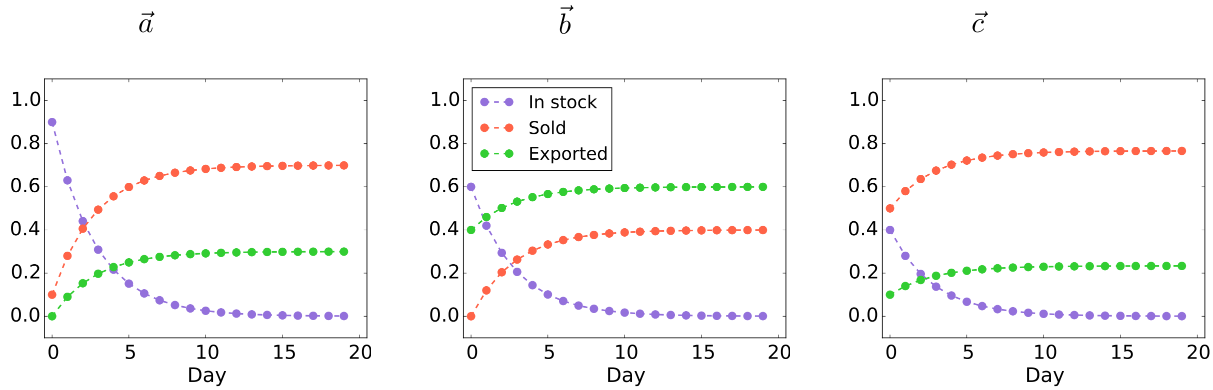

Figure 7.6: Evolution of the state vector of the Markov chain in Example (7.4.1) for different values of the initial state vector $\vec{p}_{\widetilde{X}(0)}$.

is equivalent to computing

$$\lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)} = \lim_{i\to\infty} T_{\widetilde{X}}^{\,i}\, \vec{p}_{\widetilde{X}(0)} \tag{7.37}$$

$$= \lim_{i\to\infty} T_{\widetilde{X}}^{\,i}\, \vec{a}. \tag{7.38}$$

The transition matrix $T_{\widetilde{X}}$ has three eigenvectors

$$\vec{q}_1 := \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad \vec{q}_2 := \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \vec{q}_3 := \begin{bmatrix} 0.80 \\ -0.53 \\ -0.27 \end{bmatrix}. \tag{7.39}$$

The corresponding eigenvalues are $\lambda_1 := 1$, $\lambda_2 := 1$ and $\lambda_3 := 0.7$. We gather the eigenvectors and eigenvalues into two matrices

$$Q := \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & \vec{q}_3 \end{bmatrix}, \qquad \Lambda := \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \tag{7.40}$$

so that the eigendecomposition of $T_{\widetilde{X}}$ is

$$T_{\widetilde{X}} := Q\Lambda Q^{-1}. \tag{7.41}$$

It will be useful to express the initial state vector $\vec{a}$ in terms of the different eigenvectors. This is achieved by computing

$$Q^{-1}\vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0.3 \\ 0.7 \\ 1.122 \end{bmatrix}, \tag{7.42}$$

so that

$$\vec{a} = 0.3\,\vec{q}_1 + 0.7\,\vec{q}_2 + 1.122\,\vec{q}_3. \tag{7.43}$$

We conclude that

$$\lim_{i \to \infty} T_{\widetilde{X}}^{\,i}\, \vec{a} = \lim_{i \to \infty} T_{\widetilde{X}}^{\,i}\, (0.3\, \vec{q}_1 + 0.7\, \vec{q}_2 + 1.122\, \vec{q}_3) \tag{7.44}$$

$$= \lim_{i \to \infty} 0.3\, T_{\widetilde{X}}^{\,i}\, \vec{q}_1 + 0.7\, T_{\widetilde{X}}^{\,i}\, \vec{q}_2 + 1.122\, T_{\widetilde{X}}^{\,i}\, \vec{q}_3 \tag{7.45}$$

$$= \lim_{i \to \infty} 0.3\, \lambda_1^{\,i}\, \vec{q}_1 + 0.7\, \lambda_2^{\,i}\, \vec{q}_2 + 1.122\, \lambda_3^{\,i}\, \vec{q}_3 \tag{7.46}$$

$$= \lim_{i \to \infty} 0.3\, \vec{q}_1 + 0.7\, \vec{q}_2 + 1.122\, 0.5^{\,i}\, \vec{q}_3 \tag{7.47}$$

$$= 0.3\, \vec{q}_1 + 0.7\, \vec{q}_2 \tag{7.48}$$

$$= \begin{bmatrix} 0 \\ 0.7 \\ 0.3 \end{bmatrix}. \tag{7.49}$$

This means that eventually the probability that each phone has been sold locally is 0.7 and the probability that it has been exported is 0.3. The left graph in Figure 7.6 shows the evolution of the state vector. As predicted, it eventually converges to the vector in equation (7.49).

In general, because of the special structure of the two eigenvectors with eigenvalues equal to one in this example, we have

$$\lim_{i \to \infty} T_{\widetilde{X}}^{\,i}\, \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0 \\ \left(Q^{-1} \vec{p}_{\widetilde{X}(0)}\right)_2 \\ \left(Q^{-1} \vec{p}_{\widetilde{X}(0)}\right)_1 \end{bmatrix}. \tag{7.50}$$

This is illustrated in Figure 7.6 where you can see the evolution of the state vector if it is initialized to these other two distributions:

$$\vec{b} := \begin{bmatrix} 0.6 \\ 0 \\ 0.4 \end{bmatrix}, \qquad Q^{-1} \vec{b} = \begin{bmatrix} 0.6 \\ 0.4 \\ 0.75 \end{bmatrix}, \tag{7.51}$$

$$\vec{c} := \begin{bmatrix} 0.4 \\ 0.5 \\ 0.1 \end{bmatrix}, \qquad Q^{-1} \vec{c} = \begin{bmatrix} 0.23 \\ 0.77 \\ 0.50 \end{bmatrix}. \tag{7.52}$$

△

The transition matrix of the Markov chain in Example 7.4.1 has two eigenvectors with eigenvalue equal to one. If we set the initial state vector to equal either of these eigenvectors (note that we must make sure to normalize them so that the state vector contains a valid pmf) then

$$T_{\widetilde{X}}\, \vec{p}_{\widetilde{X}(0)} = \vec{p}_{\widetilde{X}(0)}, \tag{7.53}$$

so that

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}^{\,i}\, \vec{p}_{\widetilde{X}(0)} \tag{7.54}$$

$$= \vec{p}_{\widetilde{X}(0)} \tag{7.55}$$

for all i. In particular,

$$\lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)} = \vec{p}_{\widetilde{X}(0)}, \tag{7.56}$$

so $\vec{X}$ converges to a random variable with pmf $\vec{p}_{\widetilde{X}(0)}$ in distribution. A distribution that satisfies (7.56) is called a *stationary* distribution of the Markov chain.

Definition 7.4.2 (Stationary distribution)**.** *Let $\widetilde{X}$ be a finite-state time-homogeneous Markov chain and let $\vec{p}_{stat}$ be a state vector containing a valid pmf over the possible states of $\widetilde{X}$. If $\vec{p}_{stat}$ is an eigenvector associated to an eigenvalue equal to one, so that*

$$T_{\widetilde{X}}\, \vec{p}_{stat} = \vec{p}_{stat}, \tag{7.57}$$

then the distribution corresponding to $\vec{p}_{stat}$ is a stationary or steady-state distribution of $\widetilde{X}$.

Establishing whether a distribution is stationary by checking whether (7.57) holds may be challenging computationally if the state space is very large. We now derive an alternative condition that implies stationarity. Let us first define reversibility of Markov chains.

Definition 7.4.3 (Reversibility)**.** *Let $\widetilde{X}$ be a finite-state time-homogeneous Markov chain with s states and transition matrix $T_{\widetilde{X}}$. Assume that $\widetilde{X}(i)$ is distributed according to the state vector $\vec{p} \in \mathbb{R}^s$. If*

$$\mathrm{P}\left(\widetilde{X}(i) = x_j, \widetilde{X}(i+1) = x_k\right) = \mathrm{P}\left(\widetilde{X}(i) = x_k, \widetilde{X}(i+1) = x_j\right), \quad \text{for all } 1 \le j,k \le s, \tag{7.58}$$

then we say that $\widetilde{X}$ is reversible with respect to $\vec{p}$. This is equivalent to the detailed-balance condition

$$\left(T_{\widetilde{X}}\right)_{kj} \vec{p}_j = \left(T_{\widetilde{X}}\right)_{jk} \vec{p}_k, \quad \text{for all } 1 \le j,k \le s. \tag{7.59}$$

As proved in the following theorem, reversibility implies stationarity, but the converse does not hold. A Markov chain is not necessarily reversible with respect to a stationary distribution (and often will not be). The detailed-balance condition therefore only provides a sufficient condition for stationarity.

Theorem 7.4.4 (Reversibility implies stationarity)**.** *If a time-homogeneous Markov chain $\widetilde{X}$ is reversible with respect to a distribution p_X, then p_X is a stationary distribution of $\widetilde{X}$.*

Proof. Let $\vec{p}$ be the state vector containing p_X. By assumption $T_{\widetilde{X}}$ and $\vec{p}$ satisfy (7.59), so for $1 \le j \le s$

$$\left(T_{\widetilde{X}}\vec{p}\right)_j = \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{jk} \vec{p}_k \tag{7.60}$$

$$= \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{kj} \vec{p}_j \tag{7.61}$$

$$= \vec{p}_j \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{kj} \tag{7.62}$$

$$= \vec{p}_j. \tag{7.63}$$

$$\vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \qquad \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix}$$

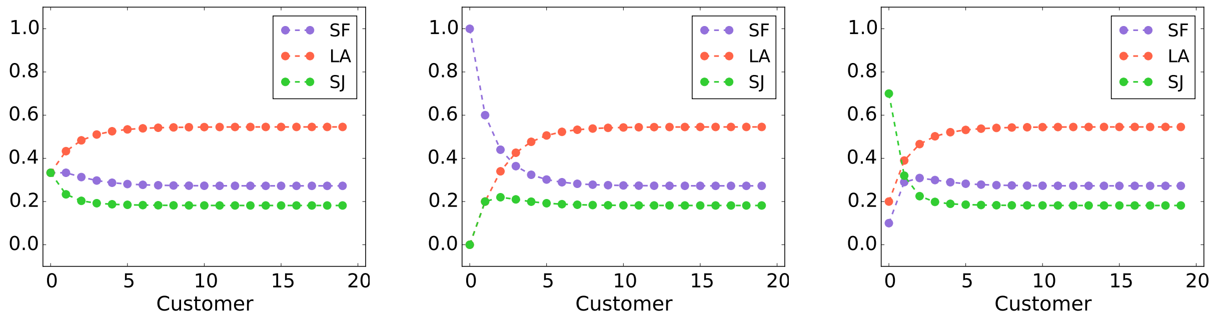

Figure 7.7: Evolution of the state vector of the Markov chain in Example (7.4.7).

The last step follows from the fact that the columns of a valid transition matrix must add to one (the chain always has to go somewhere). □

In Example 7.4.1 the Markov chain has two stationary distributions. It turns out that this is not possible for irreducible Markov chains.

Theorem 7.4.5. *Irreducible Markov chains have a single stationary distribution.*

Proof. This follows from the Perron-Frobenius theorem, which states that the transition matrix of an irreducible Markov chain has a single eigenvector with eigenvalue equal to one and nonnegative entries. □

If in addition, the Markov chain is aperiodic, then it is guaranteed to converge in distribution to a random variable with its stationary distribution *for any initial state vector*. Such Markov chains are called **ergodic**.

Theorem 7.4.6 (Convergence of Markov chains)**.** *If a discrete-time time-homogeneous Markov chain $\widetilde{X}$ is irreducible and aperiodic its state vector converges to the stationary distribution $\vec{p}_{stat}$ of $\widetilde{X}$ for any initial state vector $\vec{p}_{\widetilde{X}(0)}$. This implies that $\widetilde{X}$ converges in distribution to a random variable with pmf given by $\vec{p}_{stat}$.*

The proof of this result is beyond the scope of these notes.

Example 7.4.7 (Car rental (continued))**.** The Markov chain in the car rental example is irreducible and aperiodic. We will now check that it indeed converges in distribution. Its transition matrix has the following eigenvectors

$$\vec{q}_1 := \begin{bmatrix} 0.273 \\ 0.545 \\ 0.182 \end{bmatrix}, \qquad \vec{q}_2 := \begin{bmatrix} -0.577 \\ 0.789 \\ -0.211 \end{bmatrix}, \qquad \vec{q}_3 := \begin{bmatrix} -0.577 \\ -0.211 \\ 0.789 \end{bmatrix}. \tag{7.64}$$

The corresponding eigenvalues are $\lambda_1 := 1$, $\lambda_2 := 0.573$ and $\lambda_3 := 0.227$. As predicted by Theorem 7.4.5 the Markov chain has a single stationary distribution.

For any initial state vector, the component that is collinear with $\vec{q}_1$ will be preserved by the transitions of the Markov chain, but the other two components will become negligible after a while. The chain consequently converges in distribution to a random variable with pmf $\vec{q}_1$ (note that $\vec{q}_1$ has been normalized to be a valid pmf), as predicted by Theorem 7.4.6. This is illustrated in Figure 7.7. No matter how the company allocates the new cars, eventually 27.3% will end up in San Francisco, 54.5% in LA and 18.2% in San Jose. $\triangle$

7.5 Markov-chain Monte Carlo

The convergence of Markov chains to a stationary distribution is very useful for simulating random variables. Markov-chain Monte Carlo (MCMC) methods generate samples from a target distribution by constructing a Markov chain in such a way that the stationary distribution equals the desired distribution. These techniques are of huge importance in modern statistics and in particular in Bayesian modeling. In this section we describe one of the most popular MCMC methods and illustrate it with a simple example.

The key challenge in MCMC methods is to design an irreducible aperiodic Markov chain for which the target distribution is stationary. The Metropolis-Hastings algorithm uses an auxiliary Markov chain to achieve this.

Algorithm 7.5.1 (Metropolis-Hastings algorithm)**.** *We store the pmf p_X of the target distribution in a vector $\vec{p} \in \mathbb{R}^s$, such that*

$$\vec{p}_j := p_X(x_j), \qquad 1 \leq j \leq s. \tag{7.65}$$

Let T denote the transition matrix of an irreducible Markov chain with the same state space $\{x_1, \ldots, x_s\}$ as $\vec{p}$.

Initialize $\widetilde{X}(0)$ randomly or to a fixed state, then repeat the following steps for $i = 1, 2, 3, \ldots$.

1. *Generate a candidate random variable C from $\widetilde{X}(i-1)$ by using the transition matrix T, i.e.*

$$\mathrm{P}\left(C = k \,|\, \widetilde{X}(i-1) = j\right) = T_{kj}, \qquad 1 \leq j, k \leq s. \tag{7.66}$$

2. *Set*

$$\widetilde{X}(i) := \begin{cases} C & \text{with probability } p_{\text{acc}}\left(\widetilde{X}(i-1), C\right), \\ \widetilde{X}(i-1) & \text{otherwise,} \end{cases} \tag{7.67}$$

where the acceptance probability *is defined as*

$$p_{\text{acc}}(j,k) := \min\left\{\frac{T_{jk}\,\vec{p}_k}{T_{kj}\,\vec{p}_j}, 1\right\} \qquad 1 \leq j, k \leq s. \tag{7.68}$$

It turns out that this algorithm yields a Markov chain that is reversible with respect to the distribution of interest, which ensures that the distribution is stationary.

Theorem 7.5.2. *The pmf in $\vec{p}$ corresponds to a stationary distribution of the Markov chain $\widetilde{X}$ obtained by the Metropolis-Hastings algorithm.*

Proof. We show that the Markov chain $\widetilde{X}$ is reversible with respect to $\vec{p}$, i.e. that

$$\left(T_{\widetilde{X}}\right)_{kj}\vec{p}_j = \left(T_{\widetilde{X}}\right)_{jk}\vec{p}_k, \tag{7.69}$$

holds for all $1 \leq j, k \leq s$. This establishes the result by Theorem 7.4.4. The detailed-balanced condition holds trivially if $j = k$. If $j \neq k$ we have

$$\begin{aligned}
\left(T_{\widetilde{X}}\right)_{kj} &:= \mathrm{P}\left(\widetilde{X}(i) = k \,|\, \widetilde{X}(i-1) = j\right) && (7.70)\\
&= \mathrm{P}\left(\widetilde{X}(i) = C, C = k \,|\, \widetilde{X}(i-1) = j\right) && (7.71)\\
&= \mathrm{P}\left(\widetilde{X}(i) = C \,|\, C = k, \widetilde{X}(i-1) = j\right) P\left(C = k \,|\, \widetilde{X}(i-1) = j\right) && (7.72)\\
&= p_{\mathrm{acc}}(j,k)\, T_{kj} && (7.73)
\end{aligned}$$

and by exactly the same argument $\left(T_{\widetilde{X}}\right)_{jk} = p_{\mathrm{acc}}(k,j)\, T_{jk}$. We conclude that

$$\begin{aligned}
\left(T_{\widetilde{X}}\right)_{kj}\vec{p}_j &= p_{\mathrm{acc}}(j,k)\, T_{kj}\,\vec{p}_j && (7.74)\\
&= T_{kj}\,\vec{p}_j \min\left\{\frac{T_{jk}\,\vec{p}_k}{T_{kj}\,\vec{p}_j}, 1\right\} && (7.75)\\
&= \min\left\{T_{jk}\,\vec{p}_k, T_{kj}\,\vec{p}_j\right\} && (7.76)\\
&= T_{jk}\,\vec{p}_k \min\left\{1, \frac{T_{kj}\,\vec{p}_j}{T_{jk}\,\vec{p}_k}\right\} && (7.77)\\
&= p_{\mathrm{acc}}(k,j)\, T_{jk}\,\vec{p}_k && (7.78)\\
&= \left(T_{\widetilde{X}}\right)_{jk}\vec{p}_k. && (7.79)
\end{aligned}$$

□

The following example is taken from Hastings's seminal paper *Monte Carlo Sampling Methods Using Markov Chains and Their Applications.*

Example 7.5.3 (Generating a Poisson random variable)**.** Our aim is to generate a Poisson random variable X. Note that we don't need to know the normalizing constant in the Poisson pmf, which equals to e^{λ}, as long as we know that it is proportional to

$$p_X(x) \propto \frac{\lambda^x}{x!} \tag{7.80}$$

The auxiliary Markov chain must be able to reach any possible value of X, i.e. all positive integers. We will use a modified random walk that takes steps upwards and downwards with probability $1/2$, but never goes below 0. Its transition matrix equals

$$T_{kj} := \begin{cases} \frac{1}{2} & \text{if } j = 0 \text{ and } k = 0, \\ \frac{1}{2} & \text{if } k = j+1, \\ \frac{1}{2} & \text{if } j > 0 \text{ and } k = j-1, \\ 0 & \text{otherwise.} \end{cases} \tag{7.81}$$

T is symmetric so the acceptance probability is equal to the ratio of the pmfs:

$$p_{\text{acc}}(j,k) := \min\left\{\frac{T_{jk}\, p_X(k)}{T_{kj}\, p_X(j)}, 1\right\} \tag{7.82}$$

$$= \min\left\{\frac{p_X(k)}{p_X(j)}, 1\right\}. \tag{7.83}$$

To compute the acceptance probability, we only consider transitions that are possible under the random walk. If $j = 0$ and $k = 0$

$$p_{\text{acc}}(j,k) = 1. \tag{7.84}$$

If $k = j + 1$

$$p_{\text{acc}}(j,j+1) = \min\left\{\frac{\frac{\lambda^{j+1}}{(j+1)!}}{\frac{\lambda^{j}}{j!}}, 1\right\} \tag{7.85}$$

$$= \min\left\{\frac{\lambda}{j+1}, 1\right\}. \tag{7.86}$$

If $k = j - 1$

$$p_{\text{acc}}(j,j-1) = \min\left\{\frac{\frac{\lambda^{j-1}}{(j-1)!}}{\frac{\lambda^{j}}{j!}}, 1\right\} \tag{7.87}$$

$$= \min\left\{\frac{j}{\lambda}, 1\right\}. \tag{7.88}$$

We now spell out the steps of the Metropolis-Hastings method. To simulate the auxiliary random walk we use a sequence of Bernoulli random variables that indicate whether the random walk is trying to go up or down (or stay at zero). We initialize the chain at $x_0 = 0$. Then, for $i = 1, 2, \ldots$, we

- Generate a sample b from a Bernoulli distribution with parameter $1/2$ and a sample u uniformly distributed in $[0,1]$.
- If $b = 0$:
  - If $x_{i-1} = 0$, $x_i := 0$.
  - If $x_{i-1} > 0$:
    - If $u < \frac{x_{i-1}}{\lambda}$, $x_i := x_{i-1} - 1$.
    - Otherwise $x_i := x_{i-1}$.
- If $b = 1$:
  - If $u < \frac{\lambda}{x_{i-1}+1}$, $x_i := x_{i-1} + 1$.
  - Otherwise $x_i := x_{i-1}$.

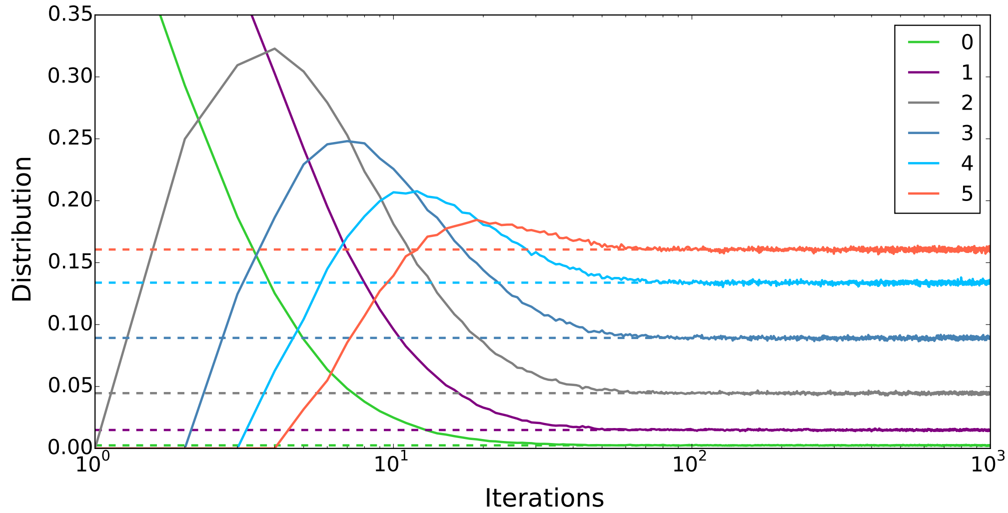

Figure 7.8: Convergence in distribution of the Markov chain constructed in Example 7.8 for $\lambda := 6$. To prevent clutter we only plot the empirical distribution of 6 states, computed by running the Markov chain 10^4 times.

The Markov chain that we have built is irreducible: there is nonzero probability of going from any nonnegative integer to any other nonnegative integer (although it could take a while!). We have not really proved that the chain should converge to the desired distribution, since we have not discussed convergence of Markov chains with infinite state spaces, but Figure 7.8 shows that the method indeed allows to sample from a Poisson distribution with $\lambda := 6$.

$\triangle$

For the example in Figure 7.8, approximate convergence in distribution occurs after around 100 iterations. This is called the **mixing time** of the Markov chain. To account for it, MCMC methods usually discard the samples from the chain over an initial period known as *burn-in* time.

The careful reader might be wondering about the point of using MCMC methods if we already have access to the desired distribution. It seems much simpler to just apply the method described in Section 2.6.1 instead. However, the Metropolis-Hastings method can be applied to discrete distributions with infinite supports and also to continuous distributions (justifying this is beyond the scope of these notes). Crucially, in contrast with inverse-transform and rejection sampling, Metropolis-Hastings does not require having access to the pmf p_X or pdf f_X of the target distribution, but rather to the ratio $p_X(x)/p_X(y)$ or $f_X(x)/f_X(y)$ for every $x \neq y$. This is very useful when computing conditional distributions within probabilistic models.

Imagine that we have access to the marginal distribution of a continuous random variable A and the conditional distribution of another continuous random variable B given A. Computing the conditional pdf

$$f_{A|B}(a|b) = \frac{f_A(a) f_{B|A}(b|a)}{\int_{u=-\infty}^{\infty} f_A(u) f_{B|A}(b|u) \, \mathrm{d}u} \tag{7.89}$$

is not necessary feasible due to the integral in the denominator. However, if we apply Metropolis-Hastings to sample from $f_{A|B}$ we don't need to compute the normalizing factor since for any $a_1 \neq a_2$

$$\frac{f_{A|B}(a_1|b)}{f_{A|B}(a_2|b)} = \frac{f_A(a_1) f_{B|A}(b|a_1)}{f_A(a_2) f_{B|A}(b|a_2)}. \tag{7.90}$$