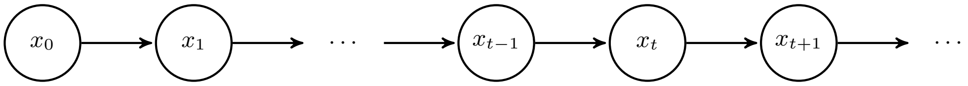
Chapter 14

Markov Chains

14.1 Introduction

A **Markov chain (MC)** is a stochastic process whose future is independent from its past and can be represented as the following Bayesian network:



The value of x_t is called the **state** of the Markov chain at time t. The set of all possible states is the **state space**. For example,

- We may represent daily weather with the state space: {sunny, cloudy, rainy}
- The state of the disease in a patient may be represented by a MC with two states: {remission, relapse}.
- The number of animals of a certain species can be represented with states $\{0, 1, 2, \dots\}$.

Uncountable state spaces are also possible (e.g., temperature) and we will rely on them for sampling later. But for simplicity, we focus on finite-state MCs. Also, note that a MC is usually an approximation of the world since we like to have a small number of states.

To complete the characterization of a MC, we also need to know the CPDs,

$$p(x_0 = i), \quad p(x_{t+1} = j | x_t = i).$$

We refer to $p(x_0)$ as the **initial distribution** and to the CPD $p(x_{t+1} = j|x_t = i)$ as **transition probabilities**. We are interested in **time-homogeneous** MCs only, in which $p(x_{t+1} = j|x_t = i)$ is independent of t, i.e., the same for all time instances. In such MCs, we can represent the transition probabilities as a transition matrix A with

$$P_{ij} = p(x_{t+1} = j|x_t = i),$$

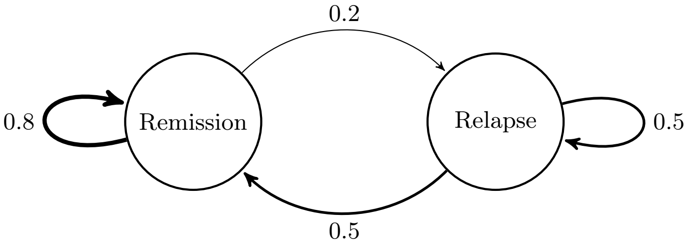
which is particularly useful if the state space is a finite set.

Example 14.1. In a Markov chain representing the health of a patient, if we let 1 represent 'remission' and 2 represent 'relapse,' we may have

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{pmatrix}.$$

△

Given that the important features of a time-homogeneous MC are its state space and transition probabilities, it is useful to represent the chain as graph, called the **state-transition graph**, whose nodes are the states and edges represent transitions and their probabilities. For example, for a disease, we may have

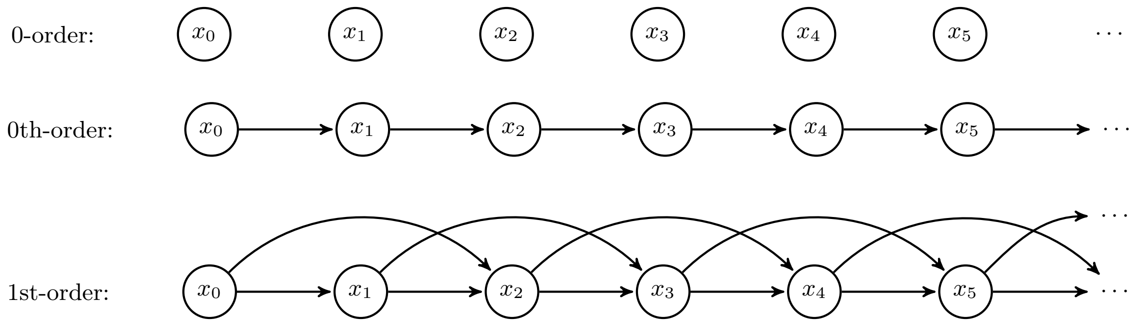


Here are some other examples of common MCs:

- **Random walk** on a grid (1D, 2D, ...). For example, in the 1-dimensional case, we can move left or right at random. This extends to n dimensions. In this context, "a drunk man will find his way home, but a drunk bird may get lost forever."
- **Page-rank**. This is closely related to the previous chain, except that this time the states are webpages, and we click on a link in the current page to transition to another one. This was the main idea behind Google search's ranking of web pages, using stationary probabilities (more on these below).
- **DNA mutations**. There are four states $\{A, C, G, T\}$ and due to mutations, a position in the genome may change from one state to another. Several variations are used in phylogenetics.

As stated before, MCs are usually approximations of real phenomena because we cannot include all relevant information in the state. As an example, consider a MC for weather. Suppose our chain represents a short period where seasonal effects are negligible and so we can assume the chain to be time-homogeneous. Each state of the MC could be the total amount of precipitation. This is already useful since a rainy day is more likely after a rainy day than after a sunny day. But if we add information about temperature, cloud cover, air pressure, etc., the model becomes more accurate and useful.

Another way that MCs can be extended is by allowing dependence on more than previous state, i.e., allowing the **order** to be larger than 1. Graphical examples of zeroth-order, first-order, and second-order MCs are shown below:



Example 14.2 ([1])**.** More accurate models can produce more realistic data, as shown in the following example from Shannon on modeling English text as a MC.

1. Zero-order approximation with uniform distribution (symbols are independent and equally probable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

2. Zero-order approximation (symbols independent but their probability is the same as English text).

   OCRO IlLI RGWR NMIELWIS EU LL NBNESEBYA TH EEl ALHENHTTPA OOBTTVA NAH BRL

3. First-order approximation (digram structure; the conditional probability of each symbol given the previous is like English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

4. Second-order approximation (trigram structure as in English).

   IN NO 1ST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

5. Zero-Order Word approximation; words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. First-Order Word approximation; the word transition probabilities are as in English text.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

△

14.2 State distribution as a function of time

Consider a MC with m states. Let $\boldsymbol{\pi}_t = (\pi_{t1}, \pi_{t2}, \ldots, \pi_{tm})$ denote the probability distribution over the states at time t, where $\pi_{tj} = p(x_t = j)$. Usually, $\boldsymbol{\pi}_0$, or equivalently, $p(x_0)$ is given. We have the following recursion,

$$\pi_{tj} = \sum_{i=1}^{m} p(x_{t-1} = i)p(x_t = j|x_{t-1} = i) = \sum_{i=1}^{m} \pi_{t-1,i} P_{ij},$$

or more compactly

$$\boldsymbol{\pi}_t = \boldsymbol{\pi}_{t-1} P \qquad \text{and} \qquad \boldsymbol{\pi}_t = \boldsymbol{\pi}_0 P^t.$$

Furthermore, the ijth element of P^t, shown as $(P^t)_{ij}$, is the probability of ending up in state j in t steps if we start from state i.

Example 14.3 (Example 14.1 continued)**.** Suppose $\pi_0 = (1, 0)^T$, i.e., the patient starts in remission. Then,

$$\boldsymbol{\pi}_1 = (1,0)\begin{pmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{pmatrix} = (0.8, 0.2), \qquad \boldsymbol{\pi}_2 = \boldsymbol{\pi}_1 \begin{pmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{pmatrix} = (0.74, 0.26)$$

$$\boldsymbol{\pi}_5 = \boldsymbol{\pi}_0 P^5 = (0.71498, 0.28502), \qquad \boldsymbol{\pi}_{10} = \boldsymbol{\pi}_0 P^{10} = (0.71429, 0.28571)$$

So after 10 days, the probability of being in remission is about 71%.

Now suppose the patient starts in relapse. Then

$$\boldsymbol{\pi}_1 = (0.5, 0.5), \qquad \boldsymbol{\pi}_2 = (0.65, 0.35)$$

$$\boldsymbol{\pi}_5 = (0.71255, 0.28745), \qquad \boldsymbol{\pi}_{10} = (0.71428, 0.28572)$$

We can see that, interestingly, $\boldsymbol{\pi}_5$ and $\boldsymbol{\pi}_{10}$ are very close to each other and almost independent of π_0. We will study this further in the next section. △

14.3 Long-term Behavior of Markov Chains

What happens to a MC if we let it run for a long time? This problem is of interest in a variety of contexts, e.g., the Page-rank algorithm above and sampling methods discussed later. We saw in the previous example that as t grows the distribution over the states appears to settle down on a certain distribution, which is called the **limiting distribution**. In the example, the limiting distribution was independent of the initial distribution. In this section, we will study when and why this happens.

A **stationary distribution** of a MC is a distribution $\boldsymbol{\sigma}$ that satisfies

$$\boldsymbol{\sigma} = \boldsymbol{\sigma} P.$$

Any finite-state Markov chain has at least one stationary distribution [2]. The limiting distribution, if it exists, must be a stationary distribution.

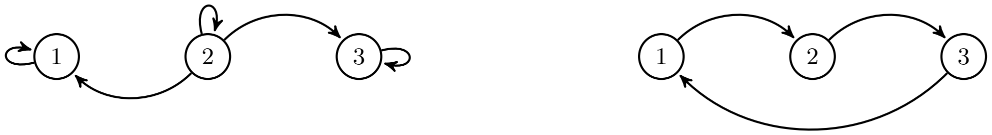
Example 14.4 (Example 14.3 continued)**.** The stationary distribution $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ is obtained by solving $(\sigma_1, \sigma_2) = (\sigma_1, \sigma_2)P$ and $\sigma_1 + \sigma_2 = 1$. It can be shown that the unique solution to these equations is

$$\boldsymbol{\sigma} = (5/7, 2/7) = (0.71429, 0.28571),$$

which indeed appears to be the limiting distribution regardless of the initial distribution. △

Graph vs. transition matrix. Whether or not a MC converges to a unique limiting distribution is determined by P. This dependence is only on P_{ij} being zero or nonzero but not how large the values are otherwise. The zero/positive status of each transition probability is given by the MC graph—an edge from states i to state j exists if and only if $P_{ij} > 0$. So the graph is sufficient to decide whether the MC will converge to a unique stationary distribution.

First, let us see some examples when the stationary distribution is not unique:



On the left, the limiting distribution depends on the initial distribution. This arises because of a lack of connectivity between the states. On the right a limiting distribution does not exist because the chain is *periodic* in a certain sense.

We can eliminate both of these possibilities by defining regular Markov chains. A Markov chain is **regular** if there is a positive integer k such that for all i and j it is possible to go from state i to state j in k steps. This is equivalent to $(P^k)_{ij} > 0$ for all i, j and also equivalent to the existence of a path of length k between any two states. In Example 14.1, we have $k = 1$.

Theorem 14.5. *If a MC with transition matrix P is regular, then there exists a unique distribution $\boldsymbol{\sigma}$ such that $\boldsymbol{\sigma} = \boldsymbol{\sigma}P$ and for any $\boldsymbol{\pi}_0$, we have $\boldsymbol{\pi}_t = \boldsymbol{\pi}_0 P^t \to \boldsymbol{\sigma}$ as $t \to \infty$.*

The above theorem guarantees that regular MCs converge to their unique stationary distributions. Furthermore, since we can choose $\boldsymbol{\pi}_0$ to have a 1 in any position, the theorem also implies that each row of P^t converges to $\boldsymbol{\sigma}$.

Example 14.6 (Example 14.4 continued)**.** Indeed, $\boldsymbol{\sigma} = (5/7, 2/7) = (0.71429, 0.28571)$ is the stationary distribution of

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{pmatrix}$$

and $\boldsymbol{\pi}_t \to \boldsymbol{\sigma}$ regardless of $\boldsymbol{\pi}_0$ as we saw in Example 14.1. Furthermore,

$$P^2 = \begin{pmatrix} 0.74 & 0.26 \\ 0.65 & 0.35 \end{pmatrix}, \qquad P^5 = \begin{pmatrix} 0.71498 & 0.28502 \\ 0.71255 & 0.28745 \end{pmatrix}, \qquad P^{10} = \begin{pmatrix} 0.71429 & 0.28571 \\ 0.71428 & 0.28572 \end{pmatrix}$$

△

14.3.1 How often does the Markov Chain visit each state?

For a regular MC with stationary distribution $\boldsymbol{\sigma}$, we know if t is large, at time t, the probability of being in state j is σ_j. But in a time period of length N, how many times state j is visited? The answer is approximately $N\sigma_j$ if N is large. (While this seems natural, similar statements do not necessarily hold for other random processes.)

For example, for a chain with transition matrix,

$$P = \frac{1}{5}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 4 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 0 \\ 0 & 1 & 0 & 1 & 3 \end{pmatrix},$$

whose graph is shown in Figure 14.1 (left), a simulation of length 1000 time units produced an empirical distribution close to the stationary distribution. The first 20 samples are as follows: 32244322242222244122.

14.4 Balance Properties and Finding the Stationary Distribution

14.4.1 Global Balance

A distribution $\boldsymbol{\pi}$ over the states of the MC satisfies the **Global Balance Property (GBP)** if for any partition[1] $\{R, L\}$ of the states of the MC, we have

$$\sum_{i\in L}\pi_i \sum_{j\in R} P_{ij} = \sum_{j\in R}\pi_j \sum_{i\in L} P_{ji}.$$

In particular, for any node i,

$$\pi_i \sum_{j\neq i} P_{ij} = \sum_{j\neq i}\pi_j P_{ji}.$$

It is not difficult to show mathematically that any stationary distribution $\boldsymbol{\sigma}$ of the Markov chain satisfies the global balance property. To see this intuitively, imagine Alice performs a random walk over the state-transition graph, going from state to state according to the transition probabilities P. Assume that $\boldsymbol{\pi}_0 = \boldsymbol{\sigma}$, i.e., Alice chooses her initial position according to $\boldsymbol{\sigma}$. It follows that $\boldsymbol{\pi}_t = \boldsymbol{\sigma}$. During N steps, where N is large, the number of times that Alice goes from a state in L to a state in R is approximately $N\sum_{i\in L}\pi_i\sum_{j\in R}P_{ij}$. Similarly, the number of times that Alice goes from R to L is about $\sum_{j\in R}\pi_j\sum_{i\in L}P_{ji}$. Sinece Alice cannot disapparate, we must have $\sum_{i\in L}\pi_i\sum_{j\in R}P_{ij} = \sum_{j\in R}\pi_j\sum_{i\in L}P_{ji}$.

We can use the GBP to find the stationary distribution as shown in the next example.

Example 14.7 (Example 14.6 continued)**.** For this chain we can set $L =$\{Remission\} and $R =$\{Relapse\}. Then the GBP says

$$\sigma_1 \times 0.2 = (1-\sigma_1)\times 0.5 \Rightarrow 7\sigma_1 = 5 \Rightarrow \sigma_1 = 5/7, \sigma_2 = 2/7 \Rightarrow \boldsymbol{\sigma} = (0.71429, 0.28571),$$

[1] A partition of a set S is a collection of disjoint sets whose union is equal to S.

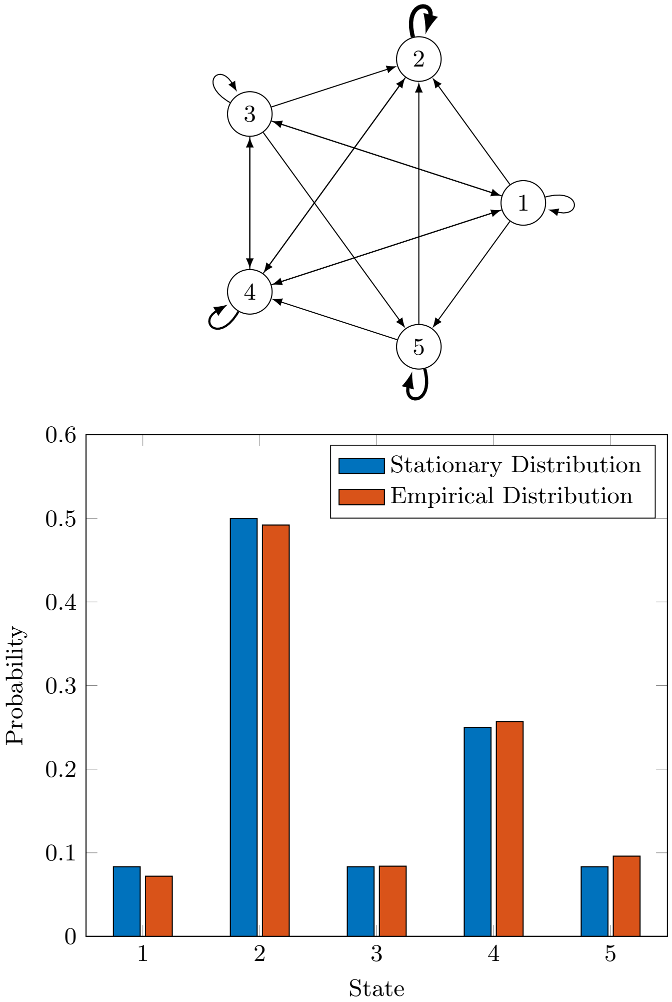


Figure 14.1: In the Markov chain (left), edges between different nodes have probability 1/5 and the probability of self loops is such that the outgoing probabilities sum to 1. The stationary distribution and an empirical (time-averaged) distribution are given on the right.

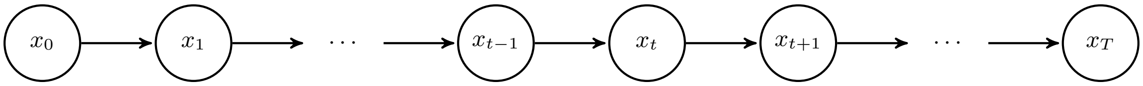
which is indeed the stationary distribution. △

14.4.2 Detailed Balance

A distribution $\boldsymbol{\pi}$ satisfies the **Detailed Balance Property (DBP)** if

$$\pi_i P_{ij} = \pi_j P_{ji}.$$

Time-reversibility and DBP. Consider the Markov chain



and assume that $\boldsymbol{\pi}_t = \boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is a stationary distribution. suppose that we run the chain backward in time (or play a movie of it backward). Note that the Markov property still holds as

$$p(x_t|x_{t+1}, \ldots, x_T) = p(x_t|x_{t+1})$$

So what are the transition probabilities P^- for the reversed MC? We have

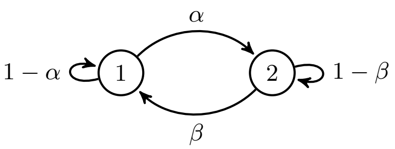
$$P^-_{ij} = p(x_t = j|x_{t+1} = i) = \frac{p(x_t = j, x_{t+1} = i)}{p(x_{t+1} = i)} = \frac{\pi_j P_{ji}}{\pi_i}.$$

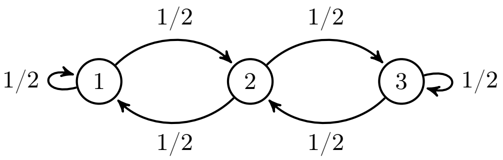
The MC is called **time-reversible** if $P^- = P$, which is equivalent to $\pi_i P_{ij} = \pi_j P_{ji}$ for all i, j, which are the detailed balance equations.

Based on the following theorem, it is easy to find the stationary distribution for Time reversible MCs, and for this reason, they are commonly used in Markov Chain Monte Carlo (MCMC) methods which we discuss later.

Theorem 14.8. *For a regular MC, if a vector $\boldsymbol{\pi}$ satisfies the detailed balance property, then π is the unique stationary distribution ($\boldsymbol{\pi} = \boldsymbol{\sigma}$) and the MC is time-reversible.*

Exercise 14.9. Using DBP, find the stationary distribution for the following MCs.





△

Bibliography

[1] C. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[2] A. Furman, “WHAT IS . . . a Stationary Measure?,” *Notices of the AMS*, vol. 58, no. 9.