# Chapter 10
# Markov chain Monte Carlo I

Chapter 8 presents methods to draw i.i.d. samples from an arbitrary distribution and Chapter 9 shows that i.i.d. samples can be used to estimate the moments (the mean and the covariance, in particular). The most robust i.i.d sampling method that works in any dimension is the rejection-acceptance sampling algorithm though it may be slow in practice. In this chapter, we introduce the Markov chain Monte Carlo (MCMC) method which is the most popular sampling approach. It is in general more effective than any methods discussed so far, particularly for complex target density in high dimensions, though it has its own problems. One of them is that we no longer have i.i.d. samples but correlated ones. Next, let us introduce some notations to study MCMC methods.

**Definition 10.1.** A collection $\{m_0, m_1, \ldots, m_N, \ldots\}$ is called *Markov chain* if the distribution of $m_k$ depends only on the immediate previous state $m_{k-1}$.

**Definition 10.2.** We call the probability of $m_k$ in $A$ starting from $m_{k-1}$ as the *transition probability* and denote it as $P(m_{k-1}, A)$. With an abuse of notation, we introduce the *transition kernel* $P(m_{k-1}, m)$ such that

*What does $P(m_{k-1}, dm)$ mean?*

$$P(m_{k-1}, A) \stackrel{\text{def}}{=} \int_A P(m_{k-1}, m)\, dm = \int_A P(m_{k-1}, dm).$$

Clearly $P(m, S) = \int_S P(m, p)\, dp = 1$.

*Example 10.1.* Assume that we have a set of Internet websites that may be linked to the others. We represent these sites as nodes and mutual linkings by directed arrows connecting nodes such as in Figure 10.1. We assign the network of nodes a *transition matrix* $P$ as

$$P = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 1/3 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}. \tag{10.1}$$

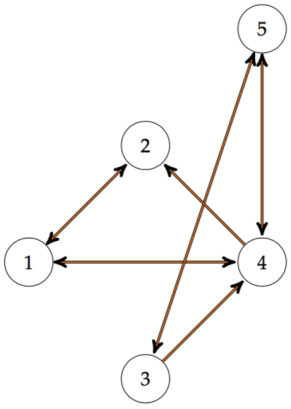


**Fig. 10.1** Five internet websites and their connections.

The $j$th row of $P$ is the probability of moving from the $j$th node to the rest. For example, the first row says that if we start from node 1, we can move to either node 2 or node 4, each with probability $\frac{1}{2}$. Note that we have treated all the nodes equally, that is, the transition probability from one node to other linked nodes is the same (a node is not linked to itself in this model).

Let $m_{k-1} = 4$, then the probability kernel $P(m_{k-1} = 4, m)$ is exactly the fourth row of $P$, i.e.,

$$P(m_{k-1} = 4, m) = [1/3, 1/3, 0, 0, 1/3].$$

The transition probability $P(m_{k-1} = 4, m_k = 1)$ is thus given by

$$\begin{aligned} P(m_{k-1} = 4, m_k = 1) &= \int_S P(m_{k-1} = 4, m)\, \delta(m-1)\, dm \\ &= \sum_{k=1}^{5} P(4,k)\, \delta(k-1) = P(4,1) = 1/3. \end{aligned}$$

△

**Definition 10.3.** We call $\mu(dm) = \pi(m)\, dm$ the invariant distribution and $\pi(m)$ invariant density of the transition probability $P(m_{k-1}, dm)$ if

$$\mu(dm) = \pi(m)\, dm = \int_S P(p, dm)\, \pi(p)\, dp. \tag{10.2}$$

*Example 10.2.* The discrete version of (10.2), applying to our website Example 10.1 and denoting the invariant measure as $\pi_\infty$, reads

$$\pi_\infty(j) = \sum_{k=1}^{5} P(k,j)\, \pi_\infty(k) = \pi_\infty P(:, j), \quad \forall j = 1, \ldots, 5,$$

which shows that the invariant measure $\pi_\infty$ is the left eignvector of the transition matrix $P$ corresponding to the unity eigenvalue. $\triangle$

**Exercise 10.1.** Assume we are initially at node 4, and we represent the initial probability density as

$$\pi_0 = \big[\,0,0,0,1,0\,\big],$$

that is, we are initially at node 4 with certainty. In order to know the next node to visit, we first compute the probability density of the next state by

$$\pi_1 = \pi_0 P,$$

then randomly move to a node by drawing a sample from the (discrete) probability density $\pi_1$. In general, the probability density after $k$ steps is given by

$$\pi_k = \pi_{k-1} P = \ldots = \pi_0 P^k. \tag{10.3}$$

Observing (10.3) you may wonder what happens if $k$ approaches infinity. Assume, on credit, the limit probability density $\pi_\infty$ exists, then it ought to satisfy

$$\pi_\infty = \pi_\infty P. \tag{10.4}$$

It follows that $\pi_\infty$ is, if exists, nothing more than the invariant measure!

Figure 10.2 shows the visiting frequency (blue) for each node after $N = 1500$ moves. Here, visiting frequency of a node is the number of visits to that node divided by $N$. We expect that numerical visiting frequencies approximate the visiting probabilities in the limit. We confirm this expectation by also plotting the components of $\pi_\infty$ (red) in Figure 10.2. By the way, $\pi_{1500}$ is equal to $\pi_\infty$ up to machine zero, meaning that a draw from $\pi_N$, $N \geq 1500$, is distributed by the limit distribution $\pi_\infty$.

Suppose you are seeking sites that contains a keyword of interest for which all the nodes, and hence websites, contain. A good search engine will show you all these websites. The question is now which website should be ranked first, second, and so on? You may guess that node 4 should be the first one in the list. However, Figure 10.2 shows that node 1 is the most visited one, and hence should appear at the top of the website list coming from the search engine!

Use the CDF-based sampling algorithm, namely Algorithm 1, to reproduce Figure 10.2. Compare the probability density $\pi_{1500}$ with the limit density, are they the same? Generate 5 figures corresponding to starting nodes $1,\ldots,5$, what do you observe?

**Exercise 10.2.** Using the above probabilistic method to determine the probability that the economy, as shown in Figure 10.3, is in recession. •

**Definition 10.4.** A Markov chain $\mathscr{M} = \{m_0, m_1, \ldots, m_N, \ldots\}$ is *reversible* with respect to $\pi(\cdot)$ if

$$\pi(m)\,P(m,p) = \pi(p)\,P(p,m), \quad \forall m,p \in \mathscr{M}. \tag{10.5}$$

*What do you conclude if you integrate both side of* (10.5) *with respect to p (or m)?*

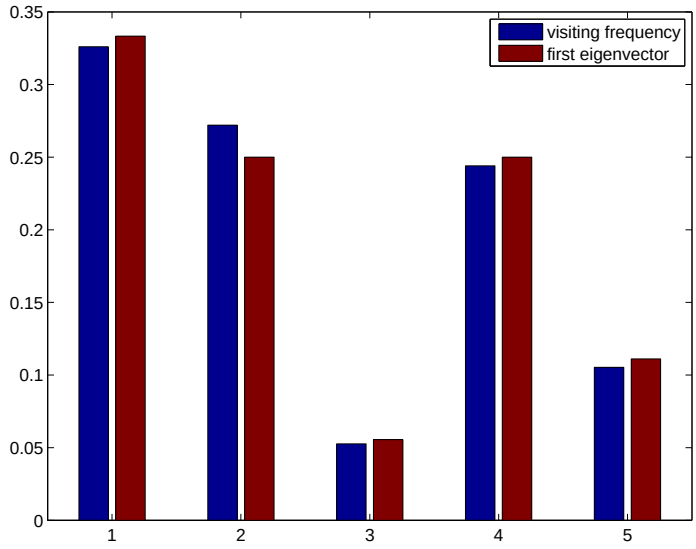


**Fig. 10.2** Visiting frequency after $N = 1500$ moves and the first eigenvector.

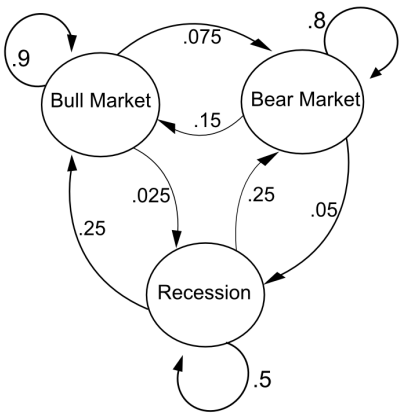


**Fig. 10.3** An example of economy states (source *Wikipedia*).

The reversibility relation (10.5) is also known as *detailed balanced equation*. You can think of the reversibility saying that the likelihood of moving from $m$ to $p$ is equal to the likelihood of moving from $p$ to $m$.

**Exercise 10.3.** What is the discrete version of (10.5)? Does the transition matrix in the website ranking problem satisfy the reversibility? If not, why? How about the transition matrix in Exercise 10.2? •

The reversibility of a Markov chain is useful since we can immediately conclude that $\pi(m)$ is its invariant density.

**Proposition 10.1.** *If the Markov chain $\mathscr{M} = \{m_0, m_1, \ldots, m_N, \ldots\}$ is reversible with respect to $\pi(m)$, then $\pi(m)$ is the invariant density.*

*Proof.* We need to prove (10.2), but it is straightforward since

$$\int_S \pi(p)\, P(p, dm)\, dp \overset{\text{reversibility}}{=} \pi(m)\, dm \int_S P(m, p)\, dp = \pi(m)\, dm.$$

□

The above discussion seems to indicate that if a Markov chain is reversible then eventually the states in the chain are distributed by the underlying invariant distribution. *It is important to point out that this is only a sufficient for the Markov chain to converges to the desired stationary distribution*. Indeed, one can construct non-reversible chains that converge (sometimes faster than the reversible counterparts) a distribution, but this is beyond the scope of this book. A question you may ask is how to construct a transition kernel such that reversibility holds. This is exactly the question Markov chain Monte Carlo methods are designed to answer. Let us now present the Metropolis-Hastings MCMC method in Algorithm 3.

**Algorithm 3** Metropolis-Hastings MCMC Algorithm

Choose initial $m_0$
**for** $k = 0, \ldots, N$ **do**

1. Draw a sample $p$ from the proposal density $q(m_k, p)$
2. Compute $\pi(p)$, $q(m_k, p)$, and $q(p, m_k)$
3. Compute the acceptance probability

$$\alpha(m_k, p) = \min\left\{1, \frac{\pi(p) q(p, m_k)}{\pi(m_k) q(m_k, p)}\right\}$$

4. **Accept** and set $m_{k+1} = p$ with probability $\alpha(m_k, p)$. Otherwise, **reject** and set $m_{k+1} = m_k$

**end for**

The idea behind the Metropolis-Hastings Algorithm 3 is very similar to that of rejection-acceptance sampling algorithm. That is, we first draw a sample from an "easy" distribution $q(m_k, p)$, then make correction so that it is distributed more like the target density $\pi(p)$. However, there are two main differences. First, the proposal distribution $q(m_k, p)$ is a function of the last state $m_k$. Second, the acceptance probability involves both the last state $m_k$ and the proposal move $p$. As a result, a chain generated from Algorithm 3 is in fact a Markov chain.

What remains is to show that the transition kernel of Algorithm 3 indeed satisfies the reversibility condition (10.5). This is the focus of the next proposition.

**Proposition 10.2.** *Markov chains generated by Algorithm 3 are reversible.*

*Proof.* We proceed in two steps. In the first step, we consider the case in which the proposal $p$ is accepted. Denote $B$ as the event of accepting a draw $q$ (or the acceptance event). Following the same proof of Proposition 8.1, we have

$$\mathbb{P}[B|p] = \alpha(m_k, p),$$

*Make sure you see this!*

leading to

$$\pi(B, p) = \mathbb{P}[B|p]\, \pi_{\text{prior}}(p) = \alpha(m_k, p) q(m_k, p),$$

which is exactly $P(m_k, p)$, the probability density of the joint event of drawing $p$ from $q(m_k, p)$ and accept it, starting from $m_k$. It follows that the reversibility holds since

$$\begin{aligned}
\pi(m_k) P(m_k, p) &= \pi(m_k)\, q(m_k, p) \min\left\{1, \frac{\pi(p) q(p, m_k)}{\pi(m_k) q(m_k, p)}\right\} \\
&= \min\{\pi(m_k) q(m_k, p), \pi(p) q(p, m_k)\} \\
&= \min\left\{\frac{\pi(m_k) q(m_k, p)}{\pi(p) q(p, m_k)}, 1\right\} \pi(p) q(p, m_k) \\
&= \pi(p) P(p, m_k).
\end{aligned}$$

In the second step, we remain at $m_k$, i.e., $m_{k+1} = m_k$, then the reversibility is trivially satisfied no matter what the transition kernel $P(m_k, p)$ is. This is the end of the proof. □

*Explain why*

**Exercise 10.4.** What is the probability of staying put at $m_k$?