Say we would like to model the USD price of bitcoin. We could observe the actual price at every hour and record it by a sequence of real numbers $x_1, x_2, \cdots$. However, it is more interesting to build a ‘model’ that could predict the price of bitcoin at time $t$, or at least give some meaningful insight how the actual bitcoin price behaves over time. Since there are so many factors affecting the price at every time, it might be reasonable that the price at time $t$ should be given by a certain RV, say $X_t$. Then our sequence of predictions would be a sequence of RVs, $(X_t)_{t\geq 0}$. This is an example of stochastic processes. Here ‘process’ means that we are not interest in just a single RV, that their sequence as a whole: ‘stochastic’ means that the way the RVs evolve in time might be random.

In this note, we will be studying a very important class of stochastic processes called Markov chains. The importance of Markov chains lies two places: 1) They are applicable for a wide range of physical, biological, social, and economical phenomena, and 2) the theory is well-established and we can actually compute and make predictions using the models.

1. Definition and examples

Roughly speaking, Markov processes are used to model temporally changing systems where future state only depends on the current state. For instance, if the price of bitcoin tomorrow depends only on its price today, then bitcoin price can be modeled as a Markov process. (Of course, the entire history of price often affects decisions of buyers/sellers so it may not be a realistic assumption.)

Even through Markov processes can be defined in vast generality, we concentrate on the simplest setting where the state and time are both discrete. Let $\Omega = \{1, 2, \cdots, m\}$ be a finite set, which we call the state space. Consider a sequence $(X_t)_{t\geq 0}$ of $\Omega$-valued RVs, which we call a chain. We call the value of $X_t$ the state of the chain at time $t$. In order to narrow down the way the chain $(X_t)_{t\geq 0}$ behaves, we introduce the following properties:

(i) (Markov property) The distribution of $X_{t+1}$ given the history $X_0, X_1, \cdots, X_t$ depends only on $X_t$. That is, for any values of $j_0, j_1, \cdots, j_t, k \in \Omega$,
\[
P(X_{t+1} = k | X_t = j_t, X_{t-1} = j_{t-1}, \cdots, X_1 = j_1, X_0 = j_0) = P(X_{t+1} = k | X_t = j_t).
\]

(ii) (Time-homogeneity) The transition probabilities
\[
p_{ij} = P(X_{t+1} = j | X_t = i) \quad i, j \in \Omega
\]
do not depend on $t$.

When the chain $(X_t)_{t\geq 0}$ satisfies the above two properties, we say it is a (discrete-time and time-homogeneous) Markov chain. We define the transition matrix $P$ to be the $m \times m$ matrix of transition probabilities:

\[
P = (p_{ij})_{1 \leq i,j \leq m} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}.
\]

Finally, since the state $X_t$ of the chain is a RV, we represent its probability mass function (PMF) via a row vector
\[
r_t = [P(X_t = 1), P(X_t = 2), \cdots, P(X_t = m)].
\]

**Example 1.1** (Gambler’s ruin). Suppose a gambler has fortune of $k$ dolors initially and starts gambling. At each time he wins or loses 1 dolor independently with probability $p$ and $1 - p$, respectively. The game ends when his fortune reaches either 0 or $N$ dolors. What is the probability that he wins $N$ dolors and goes home happy?
We use Markov chains to model his fortune after betting $t$ times. Namely, let $\Omega = \{0, 1, 2, \cdots, N\}$ be the state space. Let $(X_t)_{t \geq 0}$ be a sequence of RVs where $X_t$ is the gambler’s fortune after betting $t$ times. We first draw the state space diagram for $N = 4$ below: Next, we can write down its transition probabilities as

\[
\begin{align*}
\mathbb{P}(X_{t+1} = k + 1 | X_t = k) &= p & \forall 1 \leq k < N \\
\mathbb{P}(X_{t+1} = k | X_t = k - 1) &= 1 - p & \forall 1 \leq k < N \\
\mathbb{P}(X_{t+1} = 0 | X_t = 0) &= 1 \\
\mathbb{P}(X_{t+1} = N | X_t = N) &= 1.
\end{align*}
\]  

(5)

For example, the transition matrix $P$ for $N = 5$ is given by

\[
P = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
1 - p & 0 & p & 0 & 0 \\
0 & 1 - p & 0 & p & 0 \\
0 & 0 & 1 - p & 0 & p \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}.
\]  

(6)

We call the resulting Markov chain $(X_t)_{t \geq 0}$ the gambler’s chain. ▲

Example 1.2 (Ehrenfest Chain). This chain is originated from the physics literature as a model for two cubical volumes of air connected by a thin tunnel. Suppose there are total $N$ indistinguishable balls split into two “urns” $A$ and $B$. At each step, we pick up one of the $N$ balls uniformly at random, and move it to the other urn. Let $X_t$ denote the number of balls in urn $A$ after $t$ steps. This is a Markov chain called the Ehrenfest chain. (See the state space diagram in Figure 2.)

![Figure 2. State space diagram of the Ehrenfest chain with 4 balls](image)

It is easy to figure out the transition probabilities by considering different cases. If $X_t = k$, then urn $B$ has $N - k$ balls at time $t$. If $0 < k < N$, then with probability $k/N$ we move one ball from $A$ to $B$ and with probability $(N - k)/N$ we move one from $B$ to $A$. If $k = 0$, then we must pick up a ball from urn $B$ so $X_{t+1} = 1$ with probability 1. If $k = N$, then we must move one from $A$ to $B$ and $X_{t+1} = N - 1$ with probability 1. Hence, the transition kernel is given by

\[
\begin{align*}
\mathbb{P}(X_{t+1} = k + 1 | X_t = k) &= (N - k)/N & \forall 0 \leq k < N \\
\mathbb{P}(X_{t+1} = k - 1 | X_t = k) &= k/N & \forall 0 < k \leq N \\
\mathbb{P}(X_{t+1} = 1 | X_t = 0) &= 1 \\
\mathbb{P}(X_{t+1} = N - 1 | X_t = N) &= 1.
\end{align*}
\]  

(7)
For example, the transition matrix $P$ for $N = 5$ is given by

$$
P = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
1/5 & 0 & 4/5 & 0 & 0 \\
0 & 2/5 & 0 & 3/5 & 0 \\
0 & 0 & 3/5 & 0 & 2/5 \\
0 & 0 & 0 & 4/5 & 0 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}.
$$

(8)

Exercise 1.3. Repeat rolling two four sided dices with numbers 1, 2, 3, and 4 on them. Let $Y_k$ be the sum of the two dice at the $k$th roll. Let $S_n = Y_1 + Y_2 + \cdots + Y_n$ be the total of the first $n$ rolls, and define $X_t = S_t \pmod{6}$. Show that $(X_t)_{t \geq 0}$ is a Markov chain on the state space $\Omega = \{0,1,2,3,4,5\}$. Furthermore, identify its transition matrix.

Example 1.4. Let $\Omega = \{1,2\}$ and let $(X_t)_{t \geq 0}$ be a Markov chain on $\Omega$ with the following transition matrix

$$
P = \begin{bmatrix}
p_{11} & p_{12} \\
p_{21} & p_{22}
\end{bmatrix}.
$$

(9)

We can also represent this Markov chain pictorially as in Figure 4, which is called the 'state space diagram' of the chain $(X_t)_{t \geq 0}$.

For some concrete example, suppose $p_{11} = 0.2$, $p_{12} = 0.8$, $p_{21} = 0.6$, $p_{22} = 0.4$. (10)

If the initial state of the chain $X_0$ is 1, then

$$
P(X_1 = 1) = P(X_1 = 1 | X_0 = 1)P(X_0 = 1) + P(X_1 = 1 | X_0 = 2)P(X_0 = 2)
$$

$$
= P(X_1 = 1 | X_0 = 1) = p_{11} = 0.2
$$

(11)

(12)

and similarly,

$$
P(X_1 = 2) = P(X_1 = 2 | X_0 = 1)P(X_0 = 1) + P(X_1 = 2 | X_0 = 2)P(X_0 = 2)
$$

$$
= P(X_1 = 2 | X_0 = 1) = p_{12} = 0.8.
$$

(13)

(14)

Also we can compute the distribution of $X_2$. For example,

$$
P(X_2 = 1) = P(X_2 = 1 | X_1 = 1)P(X_1 = 1) + P(X_2 = 1 | X_1 = 2)P(X_1 = 2)
$$

$$
= p_{11}P(X_1 = 1) + p_{21}P(X_1 = 2)
$$

$$
= 0.2 \cdot 0.2 + 0.6 \cdot 0.8 = 0.04 + 0.48 = 0.52.
$$

(15)

(16)

(17)

In general, the distribution of $X_{t+1}$ can be computed from that of $X_t$ via a simple linear algebra. Note that for $i = 1, 2$,

$$
P(X_{t+1} = i) = P(X_{t+1} = i | X_t = 1)P(X_t = 1) + P(X_{t+1} = i | X_t = 2)P(X_t = 2)
$$

$$
= p_{1i}P(X_t = 1) + p_{2i}P(X_t = 2).
$$

(18)

(19)
This can be written as
\[ \begin{bmatrix} P(X_{t+1} = 2), P(X_{t+1} = 2) \end{bmatrix} = \begin{bmatrix} P(X_{t+1} = 2), P(X_{t+1} = 2) \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}. \] (20)

That is, if we represent the distribution of \( X_t \) as a row vector, then the distribution of \( X_{t+1} \) is given by multiplying the transition matrix \( P \) to the left. ▲

We generalize this observation in the following exercise.

Exercise 1.5. Let \( (X_t)_{t \geq 0} \) be a Markov chain on state space \( \Omega = \{1, 2, \cdots, m\} \) with transition matrix \( P = (p_{ij})_{1 \leq i, j \leq m} \). Let \( r_t = [P(X_t = 1), \cdots, P(X_t = m)] \) denote the row vector of the distribution of \( X_t \).

(i) Show that for each \( i \in \Omega \),
\[ P(X_{t+1} = i) = \sum_{j=1}^{m} p_{ji} P(X_t = j). \] (21)

(ii) Show that for each \( t \geq 0 \),
\[ r_{t+1} = r_t P. \] (22)

(iii) Show by induction that for each \( t \geq 0 \),
\[ r_t = r_0 P^t. \] (23)

While right-multiplication of \( P \) advances a given row vector of distribution one step forward in time, left-multiplication of \( P \) on a column vector computes the expectation of a given function with respect to the future distribution. This point is clarified in the following exercise.

Exercise 1.6. Let \( (X_t)_{t \geq 0} \) be a Markov chain on a state space \( \Omega = \{1, 2, \cdots, m\} \) with transition matrix \( P \). Let \( f : \Omega \to \mathbb{R} \) be a function. Suppose that if the chain \( X_t \) has state \( x \) at time \( t \), then we get a ‘reward’ of \( f(x) \). Let \( r_t = [P(X_t = 1), \cdots, P(X_t = m)] \) be the distribution of \( X_t \). Let \( v = [f(1), f(2), \cdots, f(m)]^T \) be the column vector representing the reward function \( f \).

(i) Show that the expected reward at time \( t \) is given by
\[ \mathbb{E}[f(X_t)] = \sum_{i=1}^{m} f(i) P(X_t = i) = r_t v. \] (24)

(ii) Use part (i) and Exercise 1.5 to show that
\[ \mathbb{E}[f(X_t)] = r_0 P^t v. \] (25)

(iii) The total reward up to time \( t \) is a RV given by \( R_t = \sum_{k=0}^{t} f(X_k) \). Show that
\[ \mathbb{E}[R_t] = r_0 (I + P + P^2 + \cdots + P^t) v. \] (26)

Exercise 1.7. Suppose that the probability it rains today is 0.4 if neither of the last two days was rainy, but 0.5 if at least one of the last two days was rainy. Let \( \Omega = \{S, R\} \), where \( S = \) sunny and \( R = \) rainy. Let \( W_t \) be the weather of day \( t \).

(i) Show that \( (W_t)_{t \geq 0} \) is not a Markov chain.

(ii) Expand the state space into the set of pairs \( \Sigma := \Omega^2 \). For each \( t \geq 0 \), define \( X_t = (W_{t-1}, W_t) \in \Sigma \). Show that \( (X_t)_{t \geq 0} \) is a Markov chain on \( \Sigma \). Identify its transition matrix.

(iii) What is the two-step transition matrix?

(iv) What is the probability that it will rain on Wednesday if it didn’t rain on Sunday and Monday?
2. Stationary Distribution and Examples

Let $(X_t)_{t \geq 0}$ be a Markov chain on state space $\Omega = \{1, 2, \cdots, m\}$ with transition matrix $P$. If $\pi$ is a distribution on $\Omega$ such that

$$\pi = \pi P,$$

then we say $\pi$ is a stationary distribution of the Markov chain $(X_t)_{t \geq 0}$.

In Exercise 1.5, we observed that we can simply multiply the transition matrix $P$ to a given row vector $r_t$ of distribution on the state space $\Omega$ in order to get the next distribution $r_{t+1}$. Hence if the initial distribution of the chain is $\pi$, then its distribution is invariant in time.

**Example 2.1.** Consider the 2-state Markov chain $(X_t)_{t \geq 0}$ with transition matrix (as in Exercise 1.4)

$$P = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix}. \tag{28}$$

Then $\pi = [3/7, 4/7]$ is a stationary distribution of $X_t$. Indeed,

$$[3/7, 4/7] = [3/7, 4/7] \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix}. \tag{29}$$

Furthermore, this is the unique stationary distribution. To see this, let $\pi = [\pi_1, \pi_2]$ be a stationary distribution of $X_t$. Then $\pi = \pi P$ gives

$$0.2\pi_1 + 0.6\pi_2 = \pi_1 \tag{30}$$
$$0.8\pi_1 + 0.4\pi_2 = \pi_2. \tag{31}$$

These equations lead to

$$4\pi_1 = 3\pi_2. \tag{32}$$

Since $\pi$ is a probability distribution, $\pi_1 + \pi_2 = 1$ so $\pi = [3/7, 4/7]$ is the only solution. This shows the uniqueness of the stationary distribution for $X_t$. ▲

A Markov chain may have multiple stationary distributions, as the following example illustrates.

**Example 2.2.** Let $(X_t)_{t \geq 0}$ be a 2-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{33}$$

Then any distribution $\pi = [p, 1-p]$ is a stationary distribution for the chain $(X_t)_{t \geq 0}$.

Stationary distributions are closely related with eigenvectors and eigenvalues of the transition matrix $P$. Namely, by taking transpose,

$$\pi^T \pi^T = P^T \pi^T. \tag{34}$$

Hence, the transpose of any stationary distribution is an eigenvector of $P^T$ associated with eigenvalue 1. More properties of stationary distribution in this line of thought are given by the following exercise.

**Exercise 2.3.** Given a matrix $A$, a row vector $v$, and a real number $\lambda$, we say $v$ is a left eigenvector of $A$ associated with left eigenvalue $\lambda$ if

$$vA = \lambda v. \tag{35}$$

If $v$ is a column vector and if $Av = \lambda v$, then we say $v$ is a (right) eigenvector associated with (right) eigenvalue $\lambda$. Let $(X_t)_{t \geq 0}$ be a Markov chain on state space $\Omega = \{1, 2, \cdots, m\}$ with transition matrix $P = (p_{ij})_{1 \leq i, j \leq m}$.

(i) Show that a distribution $\pi$ on $\Omega$ is a stationary distribution for the chain $(X_t)_{t \geq 0}$ if and only if it is a left eigenvector of $P$ associated with left eigenvalue 1.
(ii) Show that 1 is a right eigenvalue of $P$ with right eigenvector $[1, 1, \cdots, 1]^T$.

(iii) Recall that a square matrix and its transpose have the same (right) eigenvalues and corresponding (right) eigenspaces have the same dimension. Show that the Markov chain $(X_t)_{t \geq 0}$ has a unique stationary distribution if and only if $[1, 1, \cdots, 1]^T$ spans the (right) eigenspace of $P$ associated with (right) eigenvalue 1.

In the following exercise, we compute the stationary distribution of the so-called birth-death chain.

**Exercise 2.4** (Birth-Death chain). Let $\Omega = \{0, 1, 2, \cdots, N\}$ be the state space. Let $(X_t)_{t \geq 0}$ be a Markov chain on $\Omega$ with transition probabilities

$$
\begin{align*}
\mathbb{P}(X_{t+1} = k+1 | X_t = k) &= p & 0 \leq k < N \\
\mathbb{P}(X_{t+1} = k-1 | X_t = k) &= 1 - p & 1 \leq k \leq N \\
\mathbb{P}(X_{t+1} = 0 | X_t = 0) &= 1 - p \\
\mathbb{P}(X_{t+1} = N | X_t = N) &= p.
\end{align*}
$$

This is called a Birth-Death chain. Its state space diagram is as below.

![State space diagram of a 5-state Birth-Death chain](image)

**Figure 4.** State space diagram of a 5-state Birth-Death chain

(i) Let $\pi = [\pi_0, \pi_1, \cdots, \pi_N]$ be a distribution on $\Omega$. Show that $\pi$ is a stationary distribution of the Birth-Death chain if and only if it satisfy the following ‘balance equation’

$$p \pi_k = (1 - p) \pi_{k+1} \quad 0 \leq k < N. \tag{37}$$

(ii) Let $\rho = p/(1 - p)$. From (ii), deduce that $\pi_k = \rho^k \pi_0$ for all $0 \leq k < N$.

(iii) Using the normalization condition $\pi_0 + \pi_1 + \cdots + \pi_N = 1$, show that $\pi_0 = 1/(1 + \rho + \rho^2 + \cdots + \rho^N)$. Conclude that

$$\pi_k = \frac{\rho^k}{1 + \rho + \rho^2 + \cdots + \rho^N} \quad 0 \leq k \leq N. \tag{38}$$

Conclude that the Birth-Death chain has a unique stationary distribution given by (38), which becomes the uniform distribution on $\Omega$ when $p = 1/2$.

Next, we take a look at an example of an important class of Markov chains, which is called the random walk on graphs. This is the basis of many algorithms involving machine learning on networks (e.g., Google's PageRank).

**Example 2.5** (Random walk on graphs). We first introduce some notions in graph theory. A graph $G$ consists of a pair $(V,E)$ of sets of nodes $V$ and edges $E \subseteq V^2$. A graph $G$ can be concisely represented as a $|V| \times |V|$ matrix $A_G$, which is called the adjacency matrix of $G$. Namely, the $(i,j)$ entry of $A_G$ is defined by

$$A_G(i,j) = 1 \text{ (nodes } i \text{ and } j \text{ are adjacent in } G). \tag{39}$$

We say $G$ is simple if $(i,j) \in E$ implies $(j,i) \in E$ and $(i,i) \notin E$ for all $i \in V$. For a simple graph $G = (V,E)$, we say a node $j$ is adjacent to $i$ if $(i,j) \in E$. We denote $\deg_G(i)$ the number of neighbors of $i$ in $G$, which we call the degree of $i$.

Consider we hop around the nodes of a given simple graph $G = (V,E)$: at each time, we jump from one node to one of the neighbors with equal probability. For instance, if we are currently at
node 2 and if 2 is adjacent to 3, 5, and 6, then we jump to one of the three neighbors with probability 1/3. The location of this jump process at time $t$ can be described as a Markov chain. Namely, a Markov chain $(X_t)_{t \geq 0}$ on the node set $V$ is called a random walk on $G$ if

$$P(X_{t+1} = j | X_t = i) = \frac{A_G(i, j)}{\deg_G(i)}.$$  \hspace{1cm} (40)

Note that its transition matrix $P$ is obtained by normalizing each row of the adjacency matrix $A_G$ by the corresponding degree.

![Figure 5](image.jpg)

**Figure 5.** A 4-node simple graph $G$, its adjacency matrix $A_G$, and associated random walk transition matrix $P$

What is the stationary distribution of random walk on $G$? There could be many (see Exercise 4.1), but here is a typical one that always works. Define a probability distribution $\pi$ on $V$ by

$$\pi(i) = \frac{\deg_G(i)}{\sum_{j \in V} \deg_G(i)} = \frac{\deg_G(i)}{2|E|}. \hspace{1cm} (41)$$

Namely, the probability that $X_t = i$ is proportional to the degree of node $i$. Then $\pi$ is a stationary distribution for $P$. Indeed,

$$\sum_{i \in V} \pi(j) P(j, i) = \sum_{i \in V} \frac{\deg_G(i)}{2|E|} A_G(i, j) \frac{\deg_G(j)}{\deg_G(i)} = \frac{1}{2|E|} \sum_{i \in V} A_G(i, j) = \frac{\deg_G(j)}{2|E|} = \pi(j). \hspace{1cm} (43)$$

Hence if we perform a random walk on facebook network, then we are about ten times more likely to visit a person of 100 friends than to visit a person of 10 friends.

## 3. Existence of stationary distribution

Does a finite-state Markov chain always have a stationary distribution? If yes, when there is a unique stationary distribution? In Exercise 2.3, we have exploited the relation between stationary distribution and eigenvectors associated with eigenvalue 1 to study existence and uniqueness of stationary distribution. Namely, the all-one column vector is a right-eigenvector associated with eigenvalue 1 of a given transition matrix $P$. Hence its transpose $P^T$ has an eigenvector $v$ associated with eigenvalue 1. That is, $P^T v = v$. Then taking transpose, we get $v^T P = v^T$, so $\pi = v^T / D$, where $D$ is the sum of all entries in $v$, is a stationary distribution of $P$. For the uniqueness, see part (iii) of Exercise 2.3.

However, the linear algebra approach do not provide us a useful intuition for the behavior of the Markov chain itself. In Theorem 3.1, we give an alternative and constructive argument to show that every Markov chain has a stationary distribution. This is valuable since we can write down what the stationary distribution is, whereas the linear algebra argument doesn't give this information.

**Theorem 3.1 (Existence of stationary distribution).** Let $(X_t)_{t \geq 0}$ be a Markov chain on a finite state space $\Omega$ with transition matrix $P$. 

(i) For each \(x, y \in \Omega\), let \(V_t(x, y)\) be the number of visits to \(y\) in the first \(t\) steps given that \(X_0 = x\),

\[
V_t(x, y) = \sum_{k=1}^{t} 1(X_k = y | X_0 = x).
\] (44)

Define

\[
\pi_x(y) = \lim_{t \to \infty} \frac{1}{t} E[V_t(x, y)],
\] (45)

which is the expected proportion of times spending at \(y\) starting at \(x\). Then \(\pi_x\) is a probability distribution on \(\Omega\).

(ii) For each \(x \in \Omega\), \(\pi_x\) is a stationary distribution for \(P\).

Proof. For (i), note that \(0 \leq V_t(x, y) \leq t\), so \(\pi_x(y) \in [0, 1]\) for all \(y \in \Omega\). To show \(\sum_y \pi_x(y) = 1\), since the chain has to be at some unique state at each time, we have

\[
\sum_{y \in \Omega} V_t(x, y) = \sum_{y \in \Omega} \sum_{k=1}^{t} 1(X_k = y | X_0 = x)
\] (46)

\[= \sum_{k=1}^{t} \sum_{y \in \Omega} 1(X_k = y | X_0 = x) = \sum_{k=1}^{t} 1 = t.\] (47)

Hence \(r^{-1} \sum_y V_t(x, y) = 1\), so by taking expectation and letting \(t \to \infty\), we get

\[
1 = \lim_{t \to \infty} \frac{1}{t} \sum_{y \in \Omega} E[V_t(x, y)] = \sum_{y \in \Omega} \lim_{t \to \infty} \frac{1}{t} E[V_t(x, y)] = \sum_{y \in \Omega} \pi_x(y).
\] (48)

This shows (i).

Next, we show (ii). It amounts to show the matrix equation \(\pi_x P = \pi_x\), viewing \(\pi_x\) as a row vector. Fix \(y \in \Omega\). By considering the \(y\)-th entry of \(\pi_x P\), our goal is to show

\[
\sum_{z \in \Omega} \pi_x(z) P(z, y) = \pi_x(y).
\] (49)

Indeed, first observe that

\[
P(X_k = z | X_0 = x)P(z, y) = P(X_k = z | X_0 = x)P(X_{k+1} = y | X_k = z)
\] (50)

\[= P(X_k = z | X_0 = x)P(X_{k+1} = y | X_k = z, X_0 = x)\] (51)

\[= P(X_{k+1} = y, X_k = z | X_0 = x).\] (52)

Hence we have

\[
\sum_{z \in \Omega} \pi_x(z) P(z, y) = \sum_{z \in \Omega} \lim_{t \to \infty} \frac{1}{t} E[V_t(x, z)] P(z, y)
\] (53)

\[= \lim_{t \to \infty} \frac{1}{t} \sum_{z \in \Omega} E[V_t(x, z)] P(z, y)\] (54)

\[= \lim_{t \to \infty} \frac{1}{t} \sum_{z \in \Omega} \sum_{k=1}^{t} P(X_k = z | X_0 = x) P(z, y)\] (55)

\[= \lim_{t \to \infty} \frac{1}{t} \sum_{k=1}^{t} \sum_{z \in \Omega} P(X_{k+1} = y, X_k = z | X_0 = x)\] (56)

\[= \lim_{t \to \infty} \frac{1}{t} \sum_{k=2}^{t+1} P(X_k = y | X_0 = x) = \pi_x(y).\] (57)

This shows (ii). \(\square\)
4. Uniqueness of Stationary Distribution

Next, we consider the uniqueness of stationary distribution. Before we state our main result, we first take a look at random walk on disconnected graphs and that it must have multiple stationary distribution.

**Exercise 4.1** (Random walk on disconnected graphs). Let \( G = (V, E) \) be a graph with two disjoint components \( G_1 \) and \( G_2 \). Let \( (X_t)_{t \geq 0} \) be a random walk on \( G \) and denote by \( P \) its transition matrix.

(i) Let \( P_1 \) and \( P_2 \) be the transition matrices for random walks on \( G_1 \) and \( G_2 \), respectively. Let \( V_1 = \{1, 2, \cdots, n\} \) and \( V_2 = \{n + 1, n + 2, \cdots, n + m\} \) be the set of nodes in \( G_1 \) and \( G_2 \). Then show that \( P \) is of the following block diagonal form

\[
P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}.
\]

(ii) Let \( \pi_1 = [\pi_1^{(1)}, \cdots, \pi_1^{(m)}] \) and \( \pi_2 = [\pi_2^{(1)}, \cdots, \pi_2^{(m)}] \) be any stationary distributions for \( P_1 \) and \( P_2 \). Let

\[
\tilde{\pi}_1 = [\pi_1^{(1)}, \cdots, \pi_1^{(m)}, 0, \cdots, 0]
\]

\[
\tilde{\pi}_2 = [0, \cdots, 0, \pi_2^{(1)}, \cdots, \pi_2^{(m)}]
\]

be distributions on \( V = V_1 \cup V_2 = \{1, 2, \cdots, n, n + 1, \cdots, n + m\} \). Show that for any \( \lambda \in [0, 1] \), \( \pi := \lambda \tilde{\pi}_1 + (1 - \lambda)\tilde{\pi}_2 \) is a stationary distribution for \( P \).

(iii) Recall that random walk on arbitrary graph has a stationary distribution, where probability of each node is proportional to its degree. Conclude that the random walk \( (X_t)_{t \geq 0} \) on \( G \) has infinitely many stationary distribution.

A graph \( G \) is said to be connected if there exists a path between any given two nodes. A path in \( G \) is a sequence of distinct nodes \( v_0, v_1, \cdots, v_k \) such that consecutive nodes are adjacent in \( G \). In Exercise 4.1, we have seen that the random walk on a disconnected graph has infinitely many stationary distribution. How about random walk on connected graphs? We will see that they have a unique stationary distribution, which must be the typical one that we know.

**Proposition 4.2** (RW on connected graphs). Let \( G = (V, E) \) be a connected graph. Then the random walk on \( G \) has a unique stationary distribution \( \pi \), which is given by

\[
\pi(i) = \frac{\deg_G(i)}{2|E|}.
\]

Proof. Let \( P \) denote the transition matrix of random walk on \( G \). Let \( V = \{1, 2, \cdots, N\} \) and let \( E_1 \) be the column eigenspace of \( P \) associated with eigenvalue 1. That is,

\[
E_1 = \{v \in \mathbb{R}^N | Pv = v\}.
\]

Note that \([1, 1, \cdots, 1]^T \in E_1\), so \( E_1 \) has dimension \( \geq 1 \). According to Exercise 2.3, it suffice to show that \( E_1 \) has dimension 1. This is equivalent to say that \( E_1 \) is spanned by \([1, \cdots, 1]^T\).

The argument uses the ‘maximum principle’ for ‘harmonic functions’. Let \( x \in E_1 \) be arbitrary. We will view \( x \) as a function on the node set \( V \). So we want to show that \( x \) is a constant function on \( V \). Since \( V \) is finite, there is some node \( k \) where \( x \) attains its global maximum. Suppose node \( k \) has some neighbor \( j \) such that \( f(j) < f(k) \). Then from the eigenvector condition \( Px = x \),

\[
x(k) = \sum_{i=1}^{N} P(k, i)x(i)
\]

\[
= P(k, j)x(j) + \sum_{i \neq j} P(k, i)x(i)
\]

\[
< P(k, j)x(k) + \sum_{i \neq j} P(k, i)x(i)
\]

which contradicts the assumption that \( x(k) \) is the maximum.
\[ P(k, j) x(k) + \sum_{i \neq j} P(k, i) x(k) = x(k), \]  

which is a contradiction. This implies that \( x \) also attains its global maximum. Apply the similar argument to the neighbors of \( k \). Repeating the same argument, we see that \( x \) must be constant on a ‘connected component’ of \( G \) containing \( k \). But since \( G \) is connected, it follows that \( x \) is constant on \( V \). This shows the assertion. \( \square \)

If we carefully examine the proof of above result, we find that we haven’t really used the value of entries of the transition matrix \( P \). Rather, we only used an abstract property following from the connectivity of \( G \): We can reach any node from any other node. We extract this as a general property of Markov chains.

**Definition 4.3** (Irreducibility). Let \( P \) be the transition matrix of a Markov chain \((X_t)_{t \geq 0}\) on a finite state space \( \Omega \). We say the chain (or \( P \)) is irreducible if for any \( i, j \in \Omega \),

\[ \mathbb{P}(X_k = j \mid X_0 = i) > 0 \]  

for some integer \( k \geq 0 \).

In words, a Markov chain is irreducible if every state is ‘accessible’ from any other state.

**Exercise 4.4** (RW on connected graphs is irreducible). Let \( G = (V, E) \) be a connected graph. Let \((X_t)_{t \geq 0}\) be a random walk on \( G \) and denote by \( P \) its transition matrix. Let \( v_0, v_1, \ldots, v_k \) such that consecutive nodes are adjacent in \( G \). Show that

\[ P^k(v_0, v_k) = \mathbb{P}(X_k = v_k \mid X_0 = v_0) \geq P(v_0, v_1)P(v_1, v_2) \cdots P(v_{k-1}, v_k) \]

\[ = \frac{1}{\deg_G(v_0)} \frac{1}{\deg_G(v_1)} \cdots \frac{1}{\deg_G(v_k)} > 0. \]

Conclude that random walks on connected graphs are irreducible.

Now we state the general uniqueness theorem of stationary distribution.

**Theorem 4.5** (Uniqueness of stationary distribution). Let \((X_t)_{t \geq 0}\) be an irreducible Markov chain on a finite state space \( \Omega \). Then it has a unique stationary distribution, which is given by (45).

**Proof.** Exercise. Mimic the proof of Proposition 4.2. \( \square \)

**Exercise 4.6** (Hitting times). Let \((X_t)_{t \geq 0}\) be an irreducible Markov chain on a finite state space \( \Omega \) with transition matrix \( P \). Let \( X_0 = x \in \Omega \). For any \( y \in \Omega \), define the first hitting time of \( y \) by

\[ \tau_y = \min\{k \geq 1 \mid X_k = y\}. \]  

**i)** Recall that by the irreducibility, for any \( z, y \in \Omega \), \( P^r(z, y) > 0 \) for some \( r \geq 1 \). Let

\[ R = \max_{z, y \in \Omega} \min\{r \geq 1 \mid P^r(z, y) > 0\}. \]

Show that there exists some constant \( \delta > 0 \) such that

\[ \mathbb{P}(X_{t+R} = y \mid X_t = z) \geq \delta > 0 \]  

for all \( t \geq 0 \) and \( y, z \in \Omega \).

**ii)** Show that for any \( k \geq 1 \),

\[ \mathbb{P}(\tau_y > kR) \leq (1 - \delta)\mathbb{P}(\tau_y > (k - 1)R). \]  

Use induction to conclude that

\[ \mathbb{P}(\tau_y > kR) \leq (1 - \delta)^k. \]
(iii) Show that
\[
E[\tau_y] = \sum_{t=1}^{\infty} P(\tau_y \geq t) \leq R \sum_{k=0}^{\infty} P(\tau_y \geq kR) \leq R \sum_{k=0}^{\infty} (1-\delta)^k = R/\delta < \infty.
\] (77)

5. CONVERGENCE TO THE STATIONARY DISTRIBUTION

Let \((X_t)_{t\geq0}\) be an irreducible Markov chain on a finite state space \(\Omega\). By Theorems 3.1 and 4.5, we know that the chain has unique stationary distribution \(\pi\). Hence if we denote by \(\pi_t\) the distribution of \(X_t\), then we should expect that \(\pi_t\) 'converges' to \(\pi\) as \(t \to \infty\) in some sense. We will prove a precise version of this claim in this section.

We start with an example, where we show the convergence of a 2-state chain using a diagonalization of its transition matrix.

Exercise 5.1. Let \((X_t)_{t\geq0}\) be a Markov chain on \(\Omega = \{1,2\}\) with the following transition matrix
\[
P = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix}. \tag{78}
\]

(i) Show that \(P\) admits the following diagonalization
\[
P = \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -2/5 \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1}. \tag{79}
\]

(ii) Show that \(P^t\) admits the following diagonalization
\[
P^t = \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (-2/5)^t \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1}. \tag{80}
\]

(iii) Let \(r_t\) denote the row vector of distribution of \(X_t\). Show that
\[
r_t = r_0 \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (-2/5)^t \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1}. \tag{81}
\]

Also show that
\[
\lim_{t \to \infty} r_t = r_0 \begin{bmatrix} 3/7 & 4/7 \\ 3/7 & 4/7 \end{bmatrix} = [3/7, 4/7]. \tag{82}
\]

(iv) Show that \(\pi = [3/7, 4/7]\) is the unique stationary distribution for \(P\). Conclude that regardless of the initial distribution \(r_0\), the distribution of the Markov chain \((X_t)_{t\geq0}\) converges to \([3/7, 4/7]\).

Next, we observe that an irreducible MC may not always converge to the stationary distribution. The key issue there is the notion of 'periodicity'.

Example 5.2. Let \((X_t)_{t\geq0}\) be a 2-state MC on state space \(\Omega = \{0,1\}\) with transition matrix
\[
P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{83}
\]

Namely, the chain deterministically alternates between the two states. Note that it is irreducible and has a unique stationary distribution
\[
\pi = [1/2, 1/2]. \tag{84}
\]
Let \( \pi_t \) be the distribution of \( X_t \), where the initial distribution is given by \( \pi_0 = [1, 0] \). Then we have
\[
\begin{align*}
\pi_1 &= [0, 1] & (85) \\
\pi_2 &= [1, 0] & (86) \\
\pi_3 &= [0, 1] & (87) \\
\pi_4 &= [1, 0], & (88)
\end{align*}
\]
and so on. Hence the distributions \( \pi_t \) do not converge to the stationary distribution \( \pi \). ▲

**Example 5.3 (RW on torus).** Let \( \mathbb{Z}_n \) be the set of integers modulo \( n \). Let \( G = (V, E) \) be a graph where \( V = \mathbb{Z}_n \times \mathbb{Z}_n \) and two nodes \( u = (u_1, u_2) \) and \( v = (v_1, v_2) \) are adjacent if and only if
\[
|u_1 - v_1| + |v_1 - v_2| = 1.
\]
Such a graph \( G \) is called the \( n \times n \) **torus** and we write \( G = \mathbb{Z}_n \times \mathbb{Z}_n \). Intuitively, it is obtained from the \( n \times n \) square grid by adding boundary edges to wrap around (see Figure 6 left).

![Figure 6](image-url)

**Figure 6.** (Left) Torus graph (Figure excerpted from [LP17]). (Right) RW on torus \( G = \mathbb{Z}_6 \times \mathbb{Z}_6 \) has period 2.

Now let \((X_t)_{t \geq 0}\) be a random walk on \( G \). Since \( G \) is connected, \( X_t \) is irreducible. Since all nodes in \( G \) have degree 4, the uniform distribution on \( \mathbb{Z}_n \times \mathbb{Z}_n \), which we denote by \( \pi \), is the unique stationary distribution of \( X_t \). Let \( \pi_t \) denote the distribution of \( X_t \).

For instance, consider \( G = \mathbb{Z}_6 \times \mathbb{Z}_6 \). As illustrated in Figure 6 below, observe that if \( X_0 \) is one of the red nodes (where sum of coordinates is even), then \( X_t \) is at a red node for any \( t = \text{even} \) and at a black node (where sum of coordinates is odd) at \( t = \text{odd} \). Hence, \( \pi_t \) is supported only on the ‘even’ nodes for even times and on the ‘odd’ nodes for the odd times. Hence \( \pi_t \) does not converge in any sense to the uniform distribution \( \pi \). ▲

The key issue in the 2-periodicity of RW on \( G = \mathbb{Z}_6 \times \mathbb{Z}_6 \) is that it takes even number of steps to return to any given node. Generalizing this observation, we introduce the following notion of periodicity.

**Definition 5.4.** Let \( P \) be the transition matrix of a Markov chain \((X_t)_{t \geq 0}\) on a finite state space \( \Omega \). For each state \( x \in \Omega \), let \( \mathcal{T}(x) = \{t \geq 1 \mid P^t(x, x) > 0\} \) be the set of times when it is possible for the chain to return to starting state \( x \). We define the **period** of \( x \) by the greatest common divisor of \( \mathcal{T}(x) \). We say the chain \( X_t \) is **aperiodic** if all states have period 1.

**Example 5.5.** Let \( \mathcal{T}(x) = \{4, 6, 8, 10, \cdots\} \). Then the period of \( x \) is 2, even through it is not possible to go from \( x \) to \( x \) in 2 steps. If \( \mathcal{T}(x) = \{4, 6, 8, 10, \cdots\} \cup [3, 6, 9, 12, \cdots] \), then the period of \( x \) is 1. This means the return times to \( x \) is irregular. For the RW on \( G = \mathbb{Z}_6 \times \mathbb{Z}_6 \) in Example 5.3, all nodes have period 2. ▲

**Exercise 5.6 (Aperiodicity of RW on graphs).** Let \((X_t)_{t \geq 0}\) be a random walk on a connected graph \( G = (V, E) \).
(i) Show that all nodes have the same period.

(ii) If \( G \) contains an odd cycle \( C \) (e.g., triangle), show that all nodes in \( C \) have period 1.

(iii) Show that \( X_t \) is aperiodic if and only if \( G \) contains an odd cycle.

(iv)* Show that \( X_t \) is periodic if and only if \( G \) is bipartite. (A graph \( G \) is bipartite if there exists a partition \( V = A \cup B \) of nodes such that all edges are between \( A \) and \( B \).

Remark 5.7. If \( (X_t)_{t \geq 0} \) is an irreducible Markov chain on a finite state space \( \Omega \), then all states \( x \in \Omega \) must have the same period. The argument is similar to that for Exercise 5.6 (i).

We now state the convergence theorem for random walks on graphs.

**Theorem 5.8** (Convergence of RW on graphs). Let \( (X_t)_{t \geq 0} \) be a random walk on a connected graph \( G = (V,E) \) with an odd cycle. Let \( \pi \) denote the unique stationary distribution \( \pi \). Then for each \( x, y \in V \),

\[
\lim_{t \to \infty} \mathbb{P}(X_t = y \mid X_0 = x) = \pi(y). \tag{90}
\]

**Proof.** Since \( G \) contains an odd cycle, random walk on \( G \) is aperiodic according to Exercise 5.6. Let \( (Y_t)_{t \geq 0} \) be another RW on \( G \). We evolve \( X_t \) and \( Y_t \) simultaneously independently until they meet, after which we evolve them in union. Namely, let \( \tau \) be the ‘meeting time’ of these two walkers:

\[
\tau = \min \{ t \geq 0 \mid X_t = Y_t \}. \tag{91}
\]

Then we let \( X_t = Y_t \) for all \( t \geq \tau \). Still, if we disguise one random walk, the other behaves exactly as it should do. (This is a ‘coupling’ between the two random walks.)

Now suppose \( X_0 = x \) for some \( x \in V \) and \( Y_0 \) is drawn from the stationary distribution \( \pi \). In particular, the distribution of \( Y_t \) is \( \pi \) for all \( t \). Let \( \tau_t \) denote the distribution of \( X_t \). Then for any \( y \in V \),

\[
|\pi_t(y) - \pi(y)| \leq |\mathbb{P}(X_t = y) - \mathbb{P}(Y_t = y)| \tag{92}
\]

\[
\leq \mathbb{P}(X_t = y, Y_t \neq y) + \mathbb{P}(X_t \neq y, Y_t = y) \tag{93}
\]

\[
\leq \mathbb{P}(X_t \neq Y_t) \tag{94}
\]

\[
= \mathbb{P}(\tau > t). \tag{95}
\]

Hence if suffices to show that \( \mathbb{P}(\tau > t) \to 0 \) as \( t \to \infty \), as this will yield \( |\pi_t(y) - \pi(y)| \to 0 \) as \( t \to \infty \). This follows from the fact that \( \mathbb{P}(\tau < \infty) = 1 \), that is, the two independent random walks on \( G \) eventually meet with probability 1 (see Exercise 5.11).

The following exercise shows that if a RW on \( G \) is irreducible and aperiodic, then it is possible to reach any node from any other node in a fixed number of steps.

**Exercise 5.9.** Let \( (X_t)_{t \geq 0} \) be a RW on a connected graph \( G = (V,E) \). Let \( P \) denote its transition matrix. Suppose \( G \) contains an odd cycle, so that the walk is irreducible and aperiodic. For each \( x \in V \), let \( \mathcal{T}(x) \) denote the set of all possible return times to the state \( x \).

(i) For any \( x \in V \), show that \( a, b \in \mathcal{T}(x) \) implies \( a + b \in \mathcal{T}(x) \).

(ii) For any \( x \in V \), show that \( \mathcal{T}(x) \) contains 2 and some odd integer \( b \).

(iii) For each \( x \in V \), let \( b_x \) be the smallest odd integer contained in \( \mathcal{T}(x) \). Show that \( m \in \mathcal{T}(x) \) whenever \( m \geq b_x \).

(iv) Let \( b_* = \max_{x \in V} b_x \). Show that \( m \in \mathcal{T}(x) \) for all \( x \in V \) whenever \( m \geq b_* \).

(v) Let \( r = |V| + b_* \). Show that \( P^r(x, y) > 0 \) for all \( x, y \in V \).

With a little more work, one can also show a similar statement for general irreducible and aperiodic Markov chains.

**Exercise 5.10.** Let \( P \) be the transition matrix of a Markov chain \( (X_t)_{t \geq 0} \) on a finite state space \( \Omega \). Show that the following statements are equivalent:
(i) \( P \) is irreducible and aperiodic.
(ii) There exists an integer \( r \geq 0 \) such that for all \( i, j \in \Omega \),
\[
\mathbb{P}(X_r = j \mid X_0 = i) > 0.
\] (96)

(iii) There exists an integer \( r \geq 0 \) such that every entry of \( P^r \) is positive.

**Exercise 5.11.** Let \((X_t)\) and \((Y_t)\) be independent random walks on a connected graph \( G = (V, E) \). Suppose that \( G \) contains an odd cycle. Let \( P \) be the transition matrix of the random walk on \( G \).

(i) Let \( t \geq 0 \) be arbitrary. Use Exercise 5.9 to deduce that for some integer \( r \geq 1 \), we have
\[
\mathbb{P}(X_{t+r} = Y_{t+r} \mid X_t = x, Y_t = y) = \sum_{z \in V} \mathbb{P}(X_{t+r} = z = Y_{t+r} \mid X_t = x, Y_t = y)
\] (97)
\[
= \sum_{z \in V} \mathbb{P}(X_{t+r} = z \mid X_t = x) \mathbb{P}(Y_{t+r} = z \mid Y_t = y)
\] (98)
\[
= \sum_{z \in V} P^r(x, z) P^r(y, z) > 0.
\] (99)

(ii) Let \( r \geq 1 \) be as in (i) and let
\[
\delta = \min_{x, y \in V} \sum_{z \in V} P^r(x, z) P^r(y, z) > 0.
\] (100)

Use (i) and Markov property that in every \( r \) steps, \( X_t \) and \( Y_t \) meet with probability \( \geq \delta > 0 \).

(iii) Let \( \tau \) be the first time that \( X_t \) and \( Y_t \) meet. From (ii) and Markov property, deduce that
\[
\mathbb{P}(\tau \geq kr) = \mathbb{P}(\tau \geq r) \mathbb{P}(X_t \text{ and } Y_t \text{ never meet during } [r, kr) \mid X_{r-1} \neq Y_{r-1})
\]
\[
\leq (1 - \delta) \mathbb{P}(X_t \text{ and } Y_t \text{ never meet during } [r, kr) \mid X_{r-1} \neq Y_{r-1}).
\] (101)

By an induction on \( k \), conclude that
\[
\mathbb{P}(\tau \geq kr) \leq (1 - \delta)^k \to 0 \quad \text{as } k \to \infty.
\] (103)

The general convergence theorem is stated below.

**Theorem 5.12 (Convergence Thm).** Let \((X_t)_{t \geq 0}\) be an irreducible aperiodic Markov chain on a finite state space \( \Omega \). Let \( \pi \) denote the unique stationary distribution of \( X_t \). Then for each \( x, y \in V \),
\[
\lim_{t \to \infty} \mathbb{P}(X_t = y \mid X_0 = x) = \pi(y).
\] (104)

**Proof.** As always, there is a linear algebra proof to this result (see, e.g., [LP17, Thm. 4.9]) Mimic the argument for Theorem 5.8 for a coupling based proof. \( \square \)

**Exercise 5.13** (Convergence of empirical distribution). Let \((X_t)_{t \geq 0}\) be an irreducible Markov chain on a finite state space \( \Omega \). Fix a state \( x \in \Omega \), and let \( T_k \) be the \( k \)th return time to \( x \).

(i) Let \( \tau_k = T_k - T_{k-1} \) for \( k \geq 2 \). Show that \( \tau_k \)'s are i.i.d. and \( 0 < \mathbb{E}[\tau_k] < \infty \). (See Exercise 4.6.)

(ii) Use Markov property and strong law of large numbers to show
\[
\mathbb{P} \left( \lim_{n \to \infty} \frac{T_2 + \cdots + T_n}{n} = \mathbb{E}[\tau_2] \right) = 1.
\] (105)

(iii) Noting that \( T_n = T_1 + \tau_2 + \cdots + \tau_n \), show that
\[
\mathbb{P} \left( \lim_{n \to \infty} \frac{T_n}{n} = \mathbb{E}[\tau_2] \right) = 1.
\] (106)

(iv) Let \( V(x)(t) = V(t) \) be the number of visits to \( x \) up to time \( t \). Using the fact that \( \mathbb{E}[\tau_k] < \infty \), show that \( V(t) \to \infty \) as \( t \to \infty \) a.s. Also, noting that \( T_{V(t)} \leq t < T_{V(t)+1} \), use (iii) and (iv) to deduce
\[
\mathbb{P} \left( \lim_{t \to \infty} \frac{V(x)(t)}{t} = \frac{1}{\mathbb{E}[\tau_2]} \right) = 1.
\] (107)
(v) Suppose \((X_t)_{t \geq 0}\) is also aperiodic and let \(\pi\) denote the unique stationary distribution. Use Theorem 3.1 to deduce
\[
\pi(x) = \lim_{t \to \infty} \frac{E[V_x(t)]}{t} = \frac{1}{E[\tau_2]}.
\] (108)

Conclude that
\[
P\left(\lim_{t \to \infty} \frac{V_x(t)}{t} = \pi(x)\right) = 1. \tag{109}
\]

6. Markov chain Monte Carlo

So far, we were given a Markov chain \((X_t)_{t \geq 0}\) on a finite state space \(\Omega\) and studied existence and uniqueness of its stationary distribution and convergence to it. In this section, we will consider the reverse problem. Namely, given a distribution \(\pi\) on a sample space \(\Omega\), can we construct a Markov chain \((X_t)_{t \geq 0}\) such that \(\pi\) is a stationary distribution? If in addition the chain is irreducible and aperiodic, then by the convergence theorem (Theorem 5.12), we know that the distribution \(\pi_t\) of \(X_t\) converges to \(\pi\). Hence if we run the chain for long enough, the state of the chain is asymptotically distributed as \(\pi\). In other words, we can sample a random element of \(\Omega\) according to the prescribed distribution \(\pi\) by emulating it through a suitable Markov chain. This method of sampling is called Markov chain Monte Carlo (MCMC).

\[\text{Figure 7. MCMC simulation of Ising model on 200 by 200 torus at temperature } T = 1 \text{ (left), } 2 \text{ (middle), and } 5 \text{ (right).}\]

**Example 6.1** (Uniform distribution on regular graphs). Let \(G = (V,E)\) be a connected regular graph, meaning that all nodes have the same degree. Let \(\mu\) be the uniform distribution on the node set \(V\). How can we sample a random node according to \(\mu\)? If we have a list of all nodes, then we can label them from 1 to \(N = |V|\), choose a random number between 1 and \(N\), and find corresponding node. But often times, we do not have the full list of nodes, especially when we want to sample a random node from a social network. Given only local information (set of neighbors for each given node), can we still sample a uniform random node from \(G\)?

One answer is given by random walk. Indeed, random walks on graphs are defined by only using local information of the underlying graph: Choose a random neighbor and move there. Moreover, since \(G\) is connected, there is a unique stationary distribution \(\pi\) for the walk, which is given by
\[
\pi(x) = \frac{\deg_G(x)}{2|E|}.
\] (110)

Since \(G\) is regular, any two nodes have the same degree, so \(\pi(x) = \pi(y)\) for all \(x, y \in V\). This means \(\pi\) equals the uniform distribution \(\mu\) on \(V\). Hence the sampling algorithm we propose is as follows:

(*) Run a random walk on \(G\) for \(t \gg 1\) steps, and take the random node that the walk sits on.
However, there is a possible issue of convergence. Namely, if the graph $G$ does not contain any odd cycle, then random walk on $G$ is periodic (see Exercise 5.6), so we are not guaranteed to have convergence. We can overcome this by using a lazy random walk instead, which is introduced in Exercise 6.2. We know that the lazy RW on $G$ is irreducible, aperiodic, and has the same set of stationary distribution as the ordinary RW on $G$. Hence we can apply the sampling algorithm (∗) above for lazy random walk on $G$ to sample a uniform random node in $G$.

**Exercise 6.2** (Lazy RW on graphs). Let $G = (V, E)$ be a graph. Let $(X_t)_{t \geq 0}$ be a Markov chain on the node set $V$ with transition probabilities

$$\mathbb{P}(X_{t+1} = j \mid X_t = i) = \begin{cases} 1/2 & \text{if } j = i \\ 1/(2 \deg_G(i)) & \text{if } j \text{ is adjacent to } i \\ 0 & \text{otherwise.} \end{cases} \quad (111)$$

This chain is called the lazy random walk on $G$. In words, the usual random walker on $G$ now flips a fair coin to decide whether it stays at the same node or make a move to one of its neighbors.

(i) Show that for any connected graph $G$, the lazy random walk on $G$ is irreducible and aperiodic.

(ii) Let $P$ be the transition matrix for the usual random walk on $G$. Show that the following matrix

$$Q = \frac{1}{2} (P + I) \quad (112)$$

is the transition matrix for the lazy random walk on $G$.

(iii) For any distribution $\pi$ on $V$, show that $\pi Q = \pi$ if and only if $\pi P = \pi$. Conclude that the usual and lazy random walks on $G$ have the same set of stationary distribution.

**Example 6.3** (Finding local minima). Let $G = (V, E)$ be a connected graph and let $f : V \to [0, \infty)$ be a ‘cost’ function. The objective is to find a node $x^* \in V$ such that $f$ takes global minimum at $x^*$. This problem has a lot of application in machine learning, for example. Note that if the domain $V$ is very large, then an exhaustive search is too expensive to use.

Here is simple form of the popular algorithm of stochastic gradient descent, which lies at the heart of most of the important machine learning algorithms.

(i) Initialize the first guess $X_0 = x_0 \in V$.

(ii) Suppose $X_t = x \in V$ is chosen. Let $D_t = \{ y \text{ a neighbor of } x \mid f(y) \leq f(x) \}$. (113)

Define $X_{t+1}$ to be a uniform random node from $D_t$.

(iii) The algorithm terminates if it finds a local minima.

In words, at each step we move to a random neighbor which could possibly decrease the current value of $f$. It is easy to see that one always converges to a local minima, which may not a global minimum. In a very complex machine learning task (e.g., training a deep neural network), this is often good enough. Is this a Markov chain? Irreducible? Aperiodic? Stationary distribution? ▲

There is a neat solution to finding global minimum. The idea is to allow that we go uphill with a small probability.

**Example 6.4** (Finding global minimum). Let $G = (V, E)$ be a connected regular graph and let $f : V \to [0, \infty)$ be a cost function. Let

$$V^* = \left\{ x \in V \mid f(x) = \min_{y \in V} f(y) \right\} \quad (114)$$

be the set of all nodes where $f$ attains global minimum.

Fix a parameter $\lambda \in (0, 1]$, and define a probability distribution $\pi_\lambda$ on $V$ by

$$\pi_\lambda(x) = \frac{\lambda f(x)}{Z_\lambda}, \quad (115)$$
where $Z_{\lambda} = \sum_{x \in V} \lambda f(x)$ is the normalizing constant. Since $\pi_{\lambda}(x)$ is decreasing in $f(x)$, it favors nodes $x$ for which $f(x)$ is small.

Let $(X_t)_{t \geq 0}$ be Markov chain on $V$, whose transition is defined as follows. If $X_t = x$, then let $y$ be a uniform random neighbor of $x$. If $f(y) \leq f(x)$, then move to $y$; If $f(y) > f(x)$, then move to $y$ with probability $\lambda f(y) / f(x)$ and stay at $x$ with probability $1 - \lambda f(y) / f(x)$. We analyze this MC below:

(i) (Irreducibility) Since $G$ is connected and we are allowing any move (either downhill or uphill) we can go from one node to any other in some number of steps. Hence the chain $(X_t)_{t \geq 0}$ is irreducible.

(ii) (Aperiodicity) By (i) and Remark 5.7, all nodes have the same period. Moreover, let $x \in V^*$ be an arbitrary node where $f$ takes global minimum. Then all return times are possible, so $x$ has period 1. Hence all nodes have period 1, so the chain is aperiodic.

(iii) (Stationarity) Here we show that $\pi_{\lambda}$ is a stationary distribution of the chain. To do this, we first need to write down the transition matrix $P$. Namely, if we let $A_G(y, z)$ denote the indicator that $y$ and $z$ are adjacent, then

$$P(x, y) = \begin{cases} \frac{A_G(x, y)}{\deg_G(x)} & \text{if } x \neq y \\ 1 - \sum_{y \neq x} P(x, y) & \text{if } y = x. \end{cases}$$

To show $\pi_{\lambda} P = \pi_{\lambda}$, it suffices to show for any $y \in V$ that

$$\sum_{z \in V} \pi_{\lambda}(z) P(z, y) = \pi_{\lambda}(y).$$

Note that

$$\sum_{z \in V} \pi_{\lambda}(z) P(z, y) = \pi_{\lambda}(y) P(y, y) + \sum_{z \neq y} \pi_{\lambda}(z) P(z, y) = \pi_{\lambda}(y) - \sum_{z \neq y} \pi_{\lambda}(y) P(y, z) + \sum_{z \neq y} \pi_{\lambda}(z) P(z, y).$$

Hence it suffices to show that

$$\pi_{\lambda}(y) P(y, z) = \pi_{\lambda}(z) P(z, y)$$

for any $z \neq y$. Indeed, considering the two cases $f(z) \leq f(y)$ and $f(z) > f(y)$, we have

$$\pi_{\lambda}(y) P(y, z) = \frac{\lambda f(y)}{Z_{\lambda}} \frac{A_G(y, z)}{\deg_G(y)} \min(1, \lambda f(z) / f(y)) = \frac{1}{Z_{\lambda}} \frac{A_G(y, z)}{\deg_G(y)} \lambda^{\max(f(y), f(z))},$$

$$\pi_{\lambda}(z) P(z, y) = \frac{\lambda f(z)}{Z_{\lambda}} \frac{A_G(z, y)}{\deg_G(z)} \min(1, \lambda f(y) / f(z)) = \frac{1}{Z_{\lambda}} \frac{A_G(z, y)}{\deg_G(y)} \lambda^{\max(f(y), f(z))}. $$

Now since $A_G(z, y) = A_G(y, z)$ and we are assuming $G$ is a regular graph, this yields (120), as desired. Hence $\pi_{\lambda}$ is a stationary distribution for the chain $(X_t)_{t \geq 0}$.

(iv) (Convergence) By (i), (iii), Theorems 4.5, we see that $\pi_{\lambda}$ is the unique stationary distribution for the chain $X_t$. Furthermore, by (i)-(iii) and Theorem 5.12, we conclude that the distribution of $X_t$ converges to $\pi_{\lambda}$.

(v) (Global minima) Let $f_* = \min_{x \in V} f(x)$ be the global minimum of $f$. Note that by definition of $V^*$, we have $f(x) = f_*$ for any $x \in V^*$. Then observe that

$$\lim_{\lambda \to 0} \pi_{\lambda}(x) = \lim_{\lambda \to 0} \frac{\lambda f(x)}{Z_{\lambda}} \sum_{y \in V} \frac{A_G(x, y)}{\deg_G(y)} \min(1, \lambda f(y) / f(x)) = \frac{\lambda f(x) / f_*}{\sum_{y \in V} \lambda f(y) / f_*} \frac{1}{|V^*|},$$

$$= \frac{\lambda f(x) / f_*}{|V^*|} \frac{1}{|V^*|} \frac{1}{|V^*|}. $$

Thus for $\lambda$ very small, $\pi_{\lambda}$ is approximately the uniform distribution on the set of all nodes $V^*$ where $f$ attains global minimum. ▲
7. ADDITIONAL EXERCISES

In the following example, we will encounter a new concept of ‘absorption’ of Markov chains.

**Exercise 7.1** (Gambler’s ruin). Let \((X_t)_{t \geq 0}\) be the Gambler’s chain on state space \(\Omega = \{0, 1, 2, \ldots, N\}\) introduced in Example 1.1. (see Lecture note 1)

(i) Show that any distribution \(\pi = [a, 0, 0, \cdots, 0, b]\) on \(\Omega\) is stationary with respect to the gambler’s chain. Also show that any stationary distribution of this chain should be of this form.

(ii) Clearly the gambler’s chain eventually visits state 0 or \(N\), and stays at that boundary state thereafter. This is called absorption. Let \(\tau_i\) denote the time until absorption starting from state \(i\):

\[
\tau_i = \min\{t \geq 0 : X_t \in \{0, N\} | X_0 = i\}. \tag{125}
\]

We are going to compute the ‘winning probabilities’: \(q_i := P(X_{\tau_i} = N)\).

By considering what happens in one step, show that they satisfy the following recursion

\[
\begin{cases}
    q_i = pq_{i+1} + (1-p)q_{i-1} & \forall 1 \leq i < N \\
    q_0 = 0, \quad q_N = 1
\end{cases} \tag{126}
\]

(iii) Denote \(\rho = (1-p)/p\). Show that

\[q_{i+1} - q_i = \rho(q_i - q_{i-1}) \quad \forall 1 \leq i < N. \tag{127}\]

Deduce that

\[q_{i+1} - q_i = \rho^i(q_1 - q_0) = \rho^i q_1 \quad \forall 1 \leq i < N, \tag{128}\]

and that

\[q_i = q_1(1 + \rho + \cdots + \rho^{i-1}) \quad \forall 1 \leq i \leq N. \tag{129}\]

(iv)* Use \(q_N = 1\) to deduce

\[q_1 = \frac{1}{1 + \rho + \cdots + \rho^{N-1}}. \tag{130}\]

Conclude that

\[q_i = \frac{1 + \rho + \cdots + \rho^{i-1}}{1 + \rho + \cdots + \rho^{N-1}} = \begin{cases}
    \frac{1 - \rho^i}{1 - \rho} & \text{if } p \neq 1/2 \\
    \frac{i}{N} & \text{if } p = 1/2.
\end{cases} \tag{131}\]

(Remark: Unlike the Birth-Death chain problem, \(q_i\’s\) do not have to add up to 1)

REFERENCES