1. Two-State Chain with Linear Algebra

Consider the Markov chain \((X_n, \ n \in \mathbb{N})\), shown in Figure 1, where \(\alpha, \beta \in (0, 1)\).

![Markov chain for Problem 1.](image)

(a) Find the probability transition matrix \(P\).

(b) Find two real numbers \(\lambda_1\) and \(\lambda_2\) such that there exists two non-zero vectors \(u_1\) and \(u_2\) such that \(Pu_i = \lambda_i u_i\) for \(i = 1, 2\). Further, show that \(P\) can be written as \(P = U\Lambda U^{-1}\), where \(U\) and \(\Lambda\) are \(2 \times 2\) matrices and \(\Lambda\) is a diagonal matrix.

*Hint:* This is called the eigendecomposition of a matrix.

(c) Find \(P^n\) in terms of \(U\) and \(\Lambda\) for each \(n \in \mathbb{N}\).

(d) Assume that \(X_0 = 0\). Use the result in part (c) to compute the PMF of \(X_n\) for all \(n \in \mathbb{N}\).

(e) What does the fraction of time spent in state 0, \(n^{-1} \sum_{i=1}^{n} 1\{X_i = 0\}\), converge to (almost surely) as \(n \to \infty\)?

**Solution:**

(a) The probability transition matrix is

\[
P = \begin{bmatrix}
1 - \alpha & \alpha \\
\beta & 1 - \beta
\end{bmatrix}.
\]
(b) Since \((P - \lambda I)x = 0\) has non-zero solution \(u_i\), we have \(\det(P - \lambda I) = 0\), i.e., \(\lambda_1\) and \(\lambda_2\) are solutions to
\[
\det \begin{bmatrix}
1 - \alpha & -\lambda \\
\beta & 1 - \beta
\end{bmatrix} = \lambda^2 - (2 - \alpha - \beta)\lambda + 1 - \alpha - \beta.
\]
Then we get \(\lambda_1 = 1\), and \(\lambda_2 = 1 - \alpha - \beta\). Then we can get \(u_1\) and \(u_2\):
\[
u_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}^T \text{ and } u_2 = \begin{bmatrix} \alpha & -\beta \end{bmatrix}^T.
\]
Further, we can see that if we let
\[
U = \begin{bmatrix} u_1 & u_2 \end{bmatrix} = \begin{bmatrix} 1 & \alpha \\
1 & 1 - \beta \end{bmatrix},
\]
and
\[
\Lambda = \begin{bmatrix} 1 & 0 \\
0 & 1 - \alpha - \beta \end{bmatrix},
\]
we have \(PU = U\Lambda\), which is equivalent to \(P = U\Lambda U^{-1}\).

(c) We have
\[
P^n = U\Lambda U^{-1} \ldots U\Lambda U^{-1} = U\Lambda^n U^{-1}.
\]

(d) Let \(\pi(n) = [\Pr(X_n = 0) \ Pr(X_n = 1)]\) be the PMF of \(X_n\). Then we have
\[
\pi(n) = \pi(0)P^n = \pi(0)U\Lambda^n U^{-1}.
\]
Since we have \(\pi(0) = \begin{bmatrix} 1 & 0 \end{bmatrix}\), by some computation, we get
\[
\pi(n) = \frac{1}{\alpha + \beta}\left[ \beta + \alpha(1 - \alpha - \beta)^n \alpha - \alpha(1 - \alpha - \beta)^n \right].
\]

(e) By the Big Theorem, the fraction of time spent in state 0 converges to the stationary distribution at state 0, \(\pi(0)\). The stationary distribution is
\[
\pi = \frac{1}{\alpha + \beta}\begin{bmatrix} \beta & \alpha \end{bmatrix},
\]
so \(\pi(0) = \beta/(\alpha + \beta)\).

2. Reducible Markov Chain
Consider the following Markov chain, for \(\alpha, \beta, p, q \in (0, 1)\).

(a) What are all of the communicating classes? (Two nodes \(x\) and \(y\) are said to belong to the same communicating class if \(x\) can reach \(y\) and \(y\) can reach \(x\) through paths of positive probability.) For each communicating class, classify it as recurrent or transient.
(b) Given that we start in state 2, what is the probability that we will reach state 0 before state 5?

(c) What are all of the possible stationary distributions of this chain? (Note that there is more than one.)

(d) Suppose we start in the initial distribution \( \pi_0 := \begin{bmatrix} 0 & 0 & \gamma & 1-\gamma & 0 & 0 \end{bmatrix} \) for some \( \gamma \in [0,1] \). Does the distribution of the chain converge, and if so, to what?

**Solution:**

(a) The communicating classes are \( \{0,1\} \) (recurrent), \( \{4,5\} \) (recurrent), and \( \{2,3\} \) (transient).

(b) Let \( T_0 \) and \( T_5 \) denote the time it takes to reach states 0 and 5 respectively. (Note that exactly one of \( T_0 \) and \( T_5 \) will be finite.) We are looking to compute \( \mathbb{P}_2(T_0 < T_5) \), and we can set up hitting equations:

\[
\begin{align*}
\mathbb{P}_2(T_0 < T_5) &= \frac{1}{2} + \frac{1}{2} \mathbb{P}_3(T_0 < T_5), \\
\mathbb{P}_3(T_0 < T_5) &= \frac{1}{2} \mathbb{P}_2(T_0 < T_5).
\end{align*}
\]

Thus, \( \mathbb{P}_2(T_0 < T_5) = 2/3 \).

(c) First we observe that no stationary distribution can put positive probability on a transient state, so the stationary distribution is supported on the states \( \{0,1,4,5\} \). Next, if we restrict our attention to only the states \( \{0,1\} \), then we have an irreducible Markov chain with stationary distribution

\[ \pi_1 := \frac{1}{\alpha + \beta} \begin{bmatrix} \beta & \alpha \end{bmatrix}, \]

and similarly, if we restrict our attention to only the states \( \{4,5\} \), then again we have an irreducible Markov chain with stationary distribution

\[ \pi_2 := \frac{1}{p + q} \begin{bmatrix} q & p \end{bmatrix}. \]

Any stationary distribution for the entire chain must be some convex combination of these two stationary distributions. Explicitly, the stationary distributions are of the form

\[ \pi = \begin{bmatrix} \frac{c\beta}{\alpha + \beta} & \frac{c\alpha}{\alpha + \beta} & 0 & 0 & \frac{(1-c)q}{p+q} & \frac{(1-c)p}{p+q} \end{bmatrix} \]

for some \( c \in [0,1] \).

(d) Indeed the distribution will converge, even though we do not have irreducibility. The intuition is as follows. The probability will leak out of the transient states \( \{2,3\} \) until all of the probability mass is supported on the recurrent states. The two recurrent classes can each be considered to be an irreducible aperiodic Markov chain and so the probability mass which
enters a recurrent class will settle into equilibrium. To aid us in finding the limiting distribution, we can use the results of Part (b). With probability \( \gamma \), we start in state 2, and with a further probability \( 2/3 \) we end up in the recurrent class \( \{0, 1\} \). By symmetry, the probability that we end up in \( \{0, 1\} \) starting form state 3 is \( 1/3 \). Thus, the total probability mass which settles into the recurrent class \( \{0, 1\} \) is \( 2\gamma/3 + (1-\gamma)/3 = 1/3 + \gamma/3 \). Then, the probability mass settling in the recurrent class \( \{4, 5\} \) is \( 2/3 - \gamma/3 \). Therefore, the chain converges to the stationary distribution in (1) with \( c = 1/3 + \gamma/3 \).

3. Product of Rolls of a Die
A fair die with labels (1 to 6) is rolled until the product of the last two rolls is 12. What is the expected number of rolls?

**Solution:**
We model this process as a Markov chain with 3 states. The states correspond to the outcome of the last roll. If the last outcome is 1 or 5, it is useless for getting a product of 12, and we say that the Markov chain is in state \( s_1 \). If the last outcome is one of 2, 3, 4, or 6, the outcome is useful, and we say that the Markov chain is in state \( s_2 \). If the product of the last two rolls is 12, we say that the Markov chain is in state \( s_3 \). Then the probability transition matrix is

\[
P = \begin{bmatrix}
\frac{1}{3} & \frac{2}{3} & 0 \\
\frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\
0 & 0 & 1
\end{bmatrix}.
\]

Let \( T_i \) be the expected number of rolls that is needed to get to state \( s_3 \), starting from state \( s_i \), \( i = 1, 2 \). Then we have

\[
T_1 = 1 + \frac{1}{3}T_1 + \frac{2}{3}T_2,
\]
\[
T_2 = 1 + \frac{1}{3}T_1 + \frac{1}{2}T_2.
\]

Solving the equations, we get \( T_1 = 10.5 \) and \( T_2 = 9 \). Then the expected number of rolls is

\[
T = 1 + \frac{1}{3}T_1 + \frac{2}{3}T_2 = 10.5.
\]

4. Metropolis-Hastings Algorithm
In this problem we introduce the **Metropolis-Hastings Algorithm**, which is an example of **Markov Chain Monte Carlo (MCMC)** sampling. In the lab this week, you will implement Metropolis-Hastings and explore its performance.

Suppose that \( \pi \) is a probability distribution on a finite set \( \mathcal{X} \). Assume that we can compute \( \pi \) up to a normalizing constant. Specifically, assume that we can efficiently calculate \( \tilde{\pi}(x) \) for any \( x \in \mathcal{X} \), where \( \pi(x) = \tilde{\pi}(x)/\sum_{x' \in \mathcal{X}} \tilde{\pi}(x') \). The normalizing constant \( 1/\sum_{x' \in \mathcal{X}} \tilde{\pi}(x') \) is called the **partition function** in some contexts, and it can be difficult to compute if \( \mathcal{X} \) is very large.
Instead of computing $\pi$ directly, we will use $\tilde{\pi}$ to design an algorithm to sample from the distribution $\pi$. We can then approximate $\pi$ if we take enough samples. The idea behind MCMC methods is to design a Markov chain whose stationary distribution is $\pi$; then, we can “run” the chain until it is close to stationarity, and then collect samples from the chain.

Initialize the chain with $X_0 = x_0$, where $x_0$ is picked arbitrarily from $\mathcal{X}$. Let $f : \mathcal{X} \times \mathcal{X} \to [0, 1]$ be a proposal distribution: for each $x \in \mathcal{X}$, $f(x, \cdot)$ is a probability distribution on $\mathcal{X}$. (In the lab, you will look at what the desirable properties of a proposal distribution are.) If the chain is at state $x \in \mathcal{X}$, the chain makes a transition according to the following rule:

- Propose the next state $y$ according to the distribution $f(x, \cdot)$.
- Accept the proposal with probability

$$A(x, y) = \min\left\{ 1, \frac{\pi(y)}{\pi(x)} \frac{f(y, x)}{f(x, y)} \right\}.$$

- If the proposal is accepted, then move the chain to $y$; otherwise, stay at $x$.

Assume that the proposal distribution $f$ is chosen to make the chain irreducible.

(a) Explain why the Markov chain can be simulated efficiently, even though $\pi$ cannot be computed efficiently.

(b) The key to showing why Metropolis-Hastings works is to look at the detailed balance equations. Suppose we have a finite irreducible Markov chain on a state space $\mathcal{X}$ with transition matrix $P$. Show that if there exists a distribution $\pi$ on $\mathcal{X}$ such that for all $x, y \in \mathcal{X}$,

$$\pi(x)P(x, y) = \pi(y)P(y, x),$$

then $\pi$ is the stationary distribution of the chain. If these equations hold, then the Markov chain is called reversible because it turns out that the equations imply that the chain looks the same going forwards as backwards.

(c) Now return to the Metropolis-Hastings chain. Use detailed balance to argue that $\pi$ is the stationary distribution of the chain.

(d) If the chain is aperiodic, then the chain will converge to the stationary distribution. If the chain is not aperiodic, we can force it to be aperiodic by considering the lazy chain: on each transition, the chain decides not to move with probability $1/2$ (independently of the propose-accept step). Explain why the lazy chain is aperiodic, and explain why the stationary distribution is the same as before.

Solution:

(a) In order to simulate the Metropolis-Hastings chain, we need to draw samples from the proposal distribution, so the proposal distribution must
be chosen to be efficiently computable. To compute the acceptance probability $A(x, y)$, observe that we only need the ratio
\[
\frac{\pi(y)}{\pi(x)} = \frac{\tilde{\pi}(y)}{\tilde{\pi}(x)},
\]
since the normalizing constant cancels out. Thus, we only need to compute $\tilde{\pi}$ in order to simulate the chain.

(b) Suppose that detailed balance holds. Then,
\[
\sum_{x \in \mathcal{X}} \pi(x) P(x, y) = \sum_{x \in \mathcal{X}} \pi(y) P(y, x) = \pi(y) \sum_{x \in \mathcal{X}} P(y, x) = \pi(y),
\]
so the balance equations hold, i.e., $\pi$ is the stationary distribution.

(c) We will check that detailed balance holds for a pair of states $(x, y)$. Without loss of generality, assume that $\pi(y)f(y, x) \geq \pi(x)f(x, y)$. Then,
\[
\pi(x)P(x, y) = \pi(x)f(x, y),
\]
since our assumption means that the proposal $x \to y$ is always accepted. On the other hand,
\[
\pi(y)P(y, x) = \pi(y)f(y, x) \frac{\pi(x)f(x, y)}{\pi(y)f(y, x)} = \pi(x)f(x, y),
\]
because the proposal $y \to x$ is accepted with probability
\[
A(y, x) = \frac{\pi(x)f(x, y)}{\pi(y)f(y, x)}.
\]

(d) Now the chain has self-loops so it is aperiodic. The fact that the stationary distribution is unchanged can be argued because detailed balance still holds. Alternatively, the stationary distribution satisfies $\pi P = \pi$. The lazy chain is equivalent to replacing $P$ with $(P + I)/2$, where $I$ is the identity matrix, but then $\pi (P + I)/2 = \pi$, so $\pi$ is still stationary for the lazy chain.

5. Reversible Markov Chains

Let $(X_n)_{n \in \mathbb{N}}$ be an irreducible Markov chain on a finite set $\mathcal{X}$, with stationary distribution $\pi$ and transition matrix $P$. The graph associated with the Markov chain is formed by taking the transition diagram of the Markov chain, removing the directions on the edges (making the graph undirected), removing self-loops, and removing duplicate edges. Show that if the graph associated with the Markov chain is a tree, then the Markov chain is reversible.

**Hint:** To solve this problem, try induction on the size of $\mathcal{X}$:

(a) Use the fact that every tree has a leaf node $x$ connected to only one neighbor $y$, and show that detailed balance holds for the edge $(x, y)$ connecting the leaf with its single neighbor.
(b) Then, argue that if you remove the leaf \( x \) from the Markov chain and increase the probability of a self-transition at state \( y \) by \( P(y, x) \), then the stationary distribution of the original chain (when restricted to \( \mathcal{X} \setminus \{x\} \)) is the stationary distribution for the new chain, and use this to conclude the inductive proof.

**Solution:**

If \( |\mathcal{X}| = 1 \), the claim is trivial. Otherwise, the transition diagram has some leaf \( i \) which is connected to some other vertex \( j \) (this is because every tree has a leaf, and the leaf must be connected to the rest of the graph because the chain is irreducible). The balance equation for \( i \) says \( \pi(i) = \pi(j) P(j, i) + \pi(i) P(i, i) \), and rearranging gives \( \pi(i) P(i, j) = \pi(j) P(j, i) \), which shows that the pair \( (i, j) \) satisfies detailed balance.

Next, remove \( i \) from the Markov chain. Formally, consider the Markov chain on the state space \( \mathcal{X} \setminus \{i\} \), with transitions as before except the transition \( P(j, j) \) is increased to \( P(j, i) + P(j, j) \); this is a valid Markov chain. We claim that the distribution \( \pi \) is stationary for this new chain as well (when restricted to \( \mathcal{X} \setminus \{i\} \) and renormalized). For \( k \in \mathcal{X} \setminus \{i, j\} \), then we still have \( \pi(k) = \sum_{k' \in \mathcal{X} \setminus \{i\}} \pi(k') P(k', k) \) since none of these transition probabilities were modified, and for vertex \( j \) we get

\[
\pi(j) = \sum_{k \in \mathcal{X} \setminus \{i\}} \pi(k) P(k, j) + \pi(i) P(i, j)
= \sum_{k \in \mathcal{X} \setminus \{i, j\}} \pi(k) P(k, j) + \pi(j) (P(j, i) + P(j, j))
\]

where we have used detailed balance of the pair \( (i, j) \). Thus, \( \pi \) satisfies the balance equations of the new chain, so \( \pi/[1 - \pi(i)] \) is the stationary distribution of the new chain. By the inductive hypothesis applied to the chain on \( \mathcal{X} \setminus \{i\} \), then \( \pi \) satisfies detailed balance for all pairs \( (k, k') \) where \( k, k' \in \mathcal{X} \setminus \{i\} \), which finishes the proof.

6. **[Bonus] Entropy Rate of a Markov Chain**

The bonus question is just for fun. You are not required to submit the bonus question, but do give it a try and write down your progress.

Consider an irreducible Markov chain \((X_n)_{n \in \mathbb{N}}\) with state space \( \mathcal{X} \), transition matrix \( P \), and stationary distribution \( \pi \). Although Markov chains are generally not i.i.d., there is also an AEP for Markov chains.

(a) Compute \( H := \lim_{n \to \infty} H(X_1, \ldots, X_n)/n \). For this, you will want to use the Chain Rule, \( H(X, Y) = H(X) + H(Y | X) \), where

\[
H(Y | X) = -\sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X}(y | x) \log_2 p_{Y|X}(y | x).
\]

[Hint: First show that \( H(X_1, \ldots, X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i | X_{i-1}) \).]
(b) The quantity $\mathcal{H}$ defined above is called the **entropy rate** of the Markov chain. It turns out that $-n^{-1}\log_2 p_{X_1,\ldots,X_n}(X_1,\ldots,X_n) \to \mathcal{H}$ a.s., although this is much harder to prove than the i.i.d. case. Taking this for granted, argue that it requires $\mathcal{H}$ bits per symbol to describe the Markov chain.

**Solution:**

(a) Because of the Chain Rule,

$$H(X_1,\ldots,X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_{i-1},\ldots,X_1).$$

Now, we argue that $H(X_i \mid X_{i-1},\ldots,X_1) = H(X_i \mid X_{i-1})$. Intuitively this is clear from the Markov property, but we will verify this now.

$$H(X_i \mid X_{i-1},\ldots,X_1) = -\sum_{x_{i-1},\ldots,x_1} p_{X_{i-1,\ldots,X_1}}(x_{i-1},\ldots,x_1) \times \sum_{x_i} p_{X_i \mid X_{i-1,\ldots,X_1}}(x_i \mid x_{i-1},\ldots,x_1) \log_2 p_{X_i \mid X_{i-1,\ldots,X_1}}(x_i \mid x_{i-1},\ldots,x_1) = -\sum_{x_{i-1},\ldots,x_1} p_{X_{i-1,\ldots,X_1}}(x_{i-1},\ldots,x_1) \times \sum_{x_i} p_{X_i \mid X_{i-1}}(x_i \mid x_{i-1}) \log_2 p_{X_i \mid X_{i-1}}(x_i \mid x_{i-1}) = -\sum_{x_{i-1} \in \mathcal{X}} p_{X_{i-1}}(x_{i-1}) \times \sum_{x_i} p_{X_i \mid X_{i-1}}(x_i \mid x_{i-1}) \log_2 p_{X_i \mid X_{i-1}}(x_i \mid x_{i-1}) = H(X_i \mid X_{i-1}).$$

Now,

$$\sum_{i=2}^{n} H(X_i \mid X_{i-1})$$

$$= \sum_{i=2}^{n} \sum_{x \in \mathcal{X}} p_{X_{i-1}}(x) \sum_{y \in \mathcal{X}} p_{X_i \mid X_{i-1}}(y \mid x) \log_2 p_{X_i \mid X_{i-1}}(y \mid x)$$

$$= \sum_{i=2}^{n} \sum_{x \in \mathcal{X}} p_{X_{i-1}}(x) \sum_{y \in \mathcal{X}} P(x,y) \log_2 P(x,y)$$

$$= \sum_{x \in \mathcal{X}} \sum_{i=2}^{n} p_{X_{i-1}}(x) \sum_{y \in \mathcal{X}} P(x,y) \log_2 P(x,y).$$

Thus,

$$\frac{1}{n} H(X_1,\ldots,X_n)$$
\[
\frac{H(X_1)}{n} + \sum_{x \in \mathcal{X}} \left[ \sum_{y \in \mathcal{X}} P(x, y) \log_2 P(x, y) \right] \frac{1}{n} \sum_{i=2}^{n} p_{X_{i-1}}(x) \\
\rightarrow \sum_{x \in \mathcal{X}} \pi(x) \sum_{y \in \mathcal{X}} P(x, y) \log_2 P(x, y).
\]

To see why the last line holds, we know that \( n^{-1} \sum_{i=1}^{n} \mathbb{1}\{X_i = x\} \) (the fraction of time spent in \( x \)) converges a.s. to \( \pi(x) \), and this does not require the chain to be aperiodic. Then, we can take expectations to deduce that \( \mathbb{E}[n^{-1} \sum_{i=1}^{n} \mathbb{1}\{X_i = x\}] = n^{-1} \sum_{i=1}^{n} p_{X_i}(x) \rightarrow \pi(x) \).

(b) The idea is very much the same as the i.i.d. case. For \( \epsilon > 0 \), define the typical set
\[
A^{(n)}_{\epsilon} := \{ x \in \mathcal{X}^n : 2^{-n(H+\epsilon)} \leq p_{X_1,\ldots,X_n}(x) \leq 2^{-n(H-\epsilon)} \}.
\]

As before, \( \mathbb{P}((X_1, \ldots, X_n) \in A^{(n)}_{\epsilon}) \rightarrow 1 \) as \( n \rightarrow \infty \), and we can again prove \( |A^{(n)}_{\epsilon}| \leq 2^{n(H+\epsilon)} \), so it requires \( \approx nH \) bits to describe sequences of length \( n \), or \( H \) bits per symbol.