S&DS 351: Stochastic Processes - Homework 5

Bryan SebaRaj

Professor Ilias Zadik

March 3, 2025

1. (20 points) Let $G = (V, E)$ be a connected simple graph. Let $d(i)$ denote the degree of vertex i, which varies for different vertices. Let π be the uniform distribution on the vertex set V. Let the base chain be the random walk on G. Apply the Metropolis method to modify the chain so that the stationary distribution is the uniform distribution π. Find the resulting transition matrix.

Denote by $d(i)$ the degree of vertex i and by $N(i)$ the set of its neighbors. Consider the base chain corresponding to the simple random walk on G. Its transition probabilities are given by

$$Q(i,j) = \begin{cases} \frac{1}{d(i)} & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The goal is to transform the stationary distribution of the chain into a uniform distribution π on V, where

$$\pi(i) = \frac{1}{|V|}, \quad \forall\, i \in V$$

The Metropolis algorithm adjusts the proposal $Q(i,j)$ by accepting moves with probability

$$\alpha(i,j) = \min\left\{1, \frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}\right\}$$

For i, j such that $(i,j) \in E$,

$$Q(i,j) = \frac{1}{d(i)} \quad \text{and} \quad Q(j,i) = \frac{1}{d(j)}$$

Since $\pi(i) = \pi(j) = \frac{1}{|V|}$, the acceptance probability becomes

$$\alpha(i,j) = \min\left\{1, \frac{(1/|V|)(1/d(j))}{(1/|V|)(1/d(i))}\right\} = \min\left\{1, \frac{d(i)}{d(j)}\right\}$$

The self-transition probability is defined to ensure that the rows of the transition matrix sum to 1,

$$P(i,i) = 1 - \sum_{k \in N(i)} P(i,k)$$

Thus, every value in the transition matrix is given by

$$P(i,j) = Q(i,j)\,\alpha(i,j) = \begin{cases} \frac{1}{d(i)} \min\left\{1, \frac{d(i)}{d(j)}\right\} & \text{if } (i,j) \in E \\ 1 - \sum_{k \in N(i)} P(i,k) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

To confirm that π is the stationary distribution of the modified chain, we verify the detailed balance condition. For any $i, j \in V$ with $(i,j) \in E$,

$$\pi(i)P(i,j) = \frac{1}{|V|}\frac{1}{d(i)} \min\left\{1, \frac{d(i)}{d(j)}\right\} \quad \text{and} \quad \pi(j)P(j,i) = \frac{1}{|V|}\frac{1}{d(j)} \min\left\{1, \frac{d(j)}{d(i)}\right\}$$

Observe that

$$\min\left\{1, \frac{d(i)}{d(j)}\right\} = \frac{d(i)}{d(j)} \min\left\{1, \frac{d(j)}{d(i)}\right\}$$

Therefore,

$$\pi(i)P(i,j) = \frac{1}{|V|}\frac{1}{d(i)} \min\left\{1, \frac{d(i)}{d(j)}\right\} = \frac{1}{|V|}\frac{1}{d(j)} \min\left\{1, \frac{d(j)}{d(i)}\right\} = \pi(j)P(j,i)$$

Thus, the detailed balance condition,

$$\pi(i)P(i,j) = \pi(j)P(j,i)$$

holds for all i, j with $(i,j) \in E$ (and this chain is a time-reversible MC). Therefore, π is the stationary distribution of the modified chain.

2. (Metropolis for optimization) Consider the knapsack problem: Given m items with weights $w_1, \ldots, w_m$ and values $v_1, \ldots, v_m$, and a total weight budget W, the goal is to find the subset of items with maximal value subject to a weight constraint. This can be formulated as a constrained optimization problem:

$$\max \sum_{i=1}^{m} x_i v_i$$

$$\text{s.t.} \quad \sum_{i=1}^{m} x_i w_i \leq W$$

$$x_i \in \{0,1\}.$$

Here the maximization is over the decision variable $x = (x_1, \ldots, x_m) \in \{0,1\}^m$, where x_i indicates the ith item is included or not. This is a hard problem to solve fast.

(a) (5 points) Consider the following Markov chain. Starting from the initial state $(0, 0, \ldots, 0)$ (an empty knapsack), if the current state is $x = (x_1, \ldots, x_m)$, in the next step update it as follows: Choose an item J uniformly at random and replace x_J by $1 - x_J$. If this satisfies the constraint, update x accordingly; otherwise, do not update x. Identify the state space of this Markov chain and its transition rule.

Define the set of *feasible* solutions by

$$C = \{\, x = (x_1, x_2, \ldots, x_m) \in \{0,1\}^m : \sum_{i=1}^{m} w_i x_i \leq W \,\}.$$

Thus, the state space of the chain is the set C.
The chain is initialized at the empty knapsack $x = (0, 0, \ldots, 0)$. Given the current state $x \in C$, the chain evolves as follows:

1. Choose an index J uniformly at random from $\{1, 2, \ldots, m\}$ (with probability $1/m$).
2. Let x^J be the state obtained from x by flipping the Jth coordinate; that is,

$$x_i^J = \begin{cases} 1 - x_i, & \text{if } i = J, \\ x_i, & \text{if } i \neq J. \end{cases}$$

3. If $x^J \in C$ (i.e., the weight constraint is still satisfied), then update x to x^J; otherwise, leave x unchanged.

Formally, if we define

$$A(x) = \{\, J \in \{1, \ldots, m\} : x^J \in C \,\} \quad \text{and} \quad B(x) = \{\, J \in \{1, \ldots, m\} : x^J \notin C \,\},$$

then the one-step transition probability $P(x, y)$ is given by

$$P(x, y) = \begin{cases} \dfrac{1}{m}, & \text{if } y = x^J \text{ for some } J \in A(x), \\ \dfrac{|B(x)|}{m}, & \text{if } y = x, \\ 0, & \text{otherwise.} \end{cases}$$

(b) (5 points) Show that the stationary distribution of this chain is the uniform distribution over the feasible set

$$C = \{(x_1, \dots, x_m) : \sum_{i=1}^{m} x_i w_i \le W, x_i \in \{0, 1\}\}.$$

We claim that the stationary distribution of the chain is the uniform distribution over C, i.e.,

$$\pi(x) = \frac{1}{|C|} \quad \text{for all } x \in C.$$

To see this, note that if $x, y \in C$ differ in exactly one coordinate (say, $y = x^J$ for some J), then by the definition of the chain,

$$P(x, y) = \frac{1}{m} \quad \text{and} \quad P(y, x) = \frac{1}{m},$$

since the move $x \to y$ (or $y \to x$) is accepted if feasible. Thus, for such neighboring states we have

$$\pi(x)P(x, y) = \frac{1}{|C|} \cdot \frac{1}{m} = \frac{1}{|C|} \cdot \frac{1}{m} = \pi(y)P(y, x).$$

For moves in which the chain stays in the same state (either because the proposed move is infeasible or because it is rejected by construction), the balance is trivial. Hence, the detailed balance condition holds for all transitions, and the uniform distribution over C is indeed stationary.

(c) (5 points) Recall the goal is to maximize the value of the selected items. Fix some parameter $\beta > 0$. Define a distribution π over the feasible set C such that

$$\pi(x) \propto \exp(\beta f(x)), \quad x \in C$$

where $f(x) = \sum_{i=1}^{m} x_i v_i$ is the objective function. If we choose a large β, π is close to the uniform distribution over the maximizers. Use the chain in part (a) as the base chain and apply Metropolis method to produce a modified chain with stationary distribution π. Find the transition rule.

When β is large, π concentrates most of its mass on the maximizers of f.
We now modify the base chain (from part (a)) using the Metropolis algorithm. Starting from a current state $x \in C$, we proceed as follows:

1. Propose a candidate y by choosing an index $J \in \{1, \dots, m\}$ uniformly at random and setting $y = x^J$ (with the stipulation that if $x^J \notin C$, then we set $y = x$).
2. If $y \ne x$ (i.e., if the proposed flip yields a feasible new state), accept the move with probability

$$\alpha(x, y) = \min\Big\{1, \frac{\pi(y)}{\pi(x)}\Big\} = \min\Big\{1, \exp\big(\beta(f(y) - f(x))\big)\Big\}.$$

   If the move is rejected, remain at x.

Thus, for any $x \in C$ and for any $J \in \{1, \dots, m\}$, let x^J denote the state obtained by flipping the Jth coordinate. Then the modified transition probability is given by:

$$P(x, y) = \begin{cases} \dfrac{1}{m} \min\Big\{1, \exp\big(\beta(f(y) - f(x))\big)\Big\}, & \text{if } y = x^J \in C \text{ for some } J \text{ and } y \ne x, \\ 1 - \displaystyle\sum_{J:\, x^J \in C,\, x^J \ne x} \dfrac{1}{m} \min\Big\{1, \exp\big(\beta(f(x^J) - f(x))\big)\Big\} - \displaystyle\sum_{J:\, x^J \notin C} \dfrac{1}{m}, & \text{if } y = x, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to verify that with this transition rule, the detailed balance condition

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

is satisfied for all $x, y \in C$, and hence $\pi(x) \propto \exp(\beta f(x))$ is the stationary distribution.

Chang Problems

1.26. (15 points) Let π_0 and ρ_0 be probability mass functions on $\mathcal{S}$, and define $\pi_1 = \pi_0 P$ and $\rho_1 = \rho_0 P$, where P is a probability transition matrix. Show that $||\pi_1 - \rho_1|| \leq ||\pi_0 - \rho_0||$. That is, in terms of total variation distance, π_1 and ρ_1 are closer to each other than π_0 and ρ_0 were.

An alternative (dual) characterization of the total variation distance is

$$\|\mu - \nu\|_{TV} = \sup_{\|f\|_\infty \leq 1} \left| \sum_{x\in\mathcal{S}} f(x)(\mu(x) - \nu(x)) \right|,$$

where the supremum is taken over all functions $f : \mathcal{S} \to \mathbb{R}$ with $\|f\|_\infty \leq 1$ (that is, $|f(x)| \leq 1$ for all $x \in \mathcal{S}$). **Step 2. Expressing $\pi_1 - \rho_1$:** For any function f with $\|f\|_\infty \leq 1$, we have

$$\sum_{y\in\mathcal{S}} f(y)\left(\pi_1(y) - \rho_1(y)\right) = \sum_{y\in\mathcal{S}} f(y) \left(\sum_{x\in\mathcal{S}} (\pi_0(x) - \rho_0(x))P(x,y) \right).$$

Interchanging the sums, we obtain

$$\sum_{y\in\mathcal{S}} f(y)\left(\pi_1(y) - \rho_1(y)\right) = \sum_{x\in\mathcal{S}} (\pi_0(x) - \rho_0(x)) \left(\sum_{y\in\mathcal{S}} f(y)P(x,y) \right).$$

Define

$$g(x) = \sum_{y\in\mathcal{S}} f(y)P(x,y).$$

Step 3. Bounding $g(x)$: Since $P(x,\cdot)$ is a probability mass function and $|f(y)| \leq 1$ for all y, it follows that

$$|g(x)| \leq \sum_{y\in\mathcal{S}} |f(y)|P(x,y) \leq \sum_{y\in\mathcal{S}} P(x,y) = 1.$$

Thus, $\|g\|_\infty \leq 1$. **Step 4. Applying the Dual Representation:** Using the definition of total variation distance for the pair (π_0, ρ_0), we have

$$\left| \sum_{x\in\mathcal{S}} (\pi_0(x) - \rho_0(x))g(x) \right| \leq \|\pi_0 - \rho_0\|_{TV}.$$

Since this inequality holds for every function f with $\|f\|_\infty \leq 1$ (and the corresponding function g defined above also satisfies $\|g\|_\infty \leq 1$), we obtain

$$\sup_{\|f\|_\infty \leq 1} \left| \sum_{y\in\mathcal{S}} f(y)\left(\pi_1(y) - \rho_1(y)\right) \right| \leq \|\pi_0 - \rho_0\|_{TV}.$$

By the dual representation, the left-hand side is exactly $\|\pi_1 - \rho_1\|_{TV}$. Hence,

$$\|\pi_1 - \rho_1\|_{TV} \leq \|\pi_0 - \rho_0\|_{TV}.$$

2.1. (5 points) For a branching process $\{G_t\}$ with $G_0 = 1$, define the probability generating function of G_t to be ψ_t, that is,

$$\psi_t(z) = \mathbb{E}[z^{G_t}] = \sum_{k=0}^{\infty} z^k P(G_t = k).$$

With ψ defined as $\rho = \sum_{k=0}^{\infty} f(k)\rho^k =: \psi(\rho)$, show that $\psi_1(z) = \psi(z)$, $\psi_2(z) = \psi(\psi(z))$, $\psi_3(z) = \psi(\psi(\psi(z)))$, and so on.

We prove by induction on the generation number t that

$$\psi_t(z) = \underbrace{\psi(\psi(\cdots\psi(z)\cdots))}_{t \text{ times}}.$$

Base Case: For $t = 1$, the process starts with one individual, i.e., $G_0 = 1$. By definition, the number of offspring produced by this single individual is distributed with probability generating function

$$\psi(z) = \sum_{k=0}^{\infty} f(k)z^k.$$

Since G_1 is exactly the number of offspring of the initial individual, we have

$$\psi_1(z) = \mathbb{E}[z^{G_1}] = \psi(z).$$

Inductive Step: Assume that for some $t \geq 1$,

$$\psi_t(z) = \underbrace{\psi(\psi(\cdots\psi(z)\cdots))}_{t \text{ times}}.$$

We now show that

$$\psi_{t+1}(z) = \psi(\psi_t(z)).$$

Each individual in generation t produces offspring independently according to the generating function $\psi(z)$. Conditioning on the number of individuals G_t in generation t, the generating function for generation $t+1$ is given by

$$\psi_{t+1}(z) = \mathbb{E}[z^{G_{t+1}}] = \mathbb{E}\left[z^{\sum_{i=1}^{G_t} X_i}\right],$$

where $\{X_i\}$ are i.i.d. random variables representing the number of offspring of each individual in generation t. Since the X_i's are independent and each has generating function $\psi(z)$, we have

$$\mathbb{E}\left[z^{\sum_{i=1}^{G_t} X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{G_t} z^{X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{G_t} \psi(z)\right] = \mathbb{E}\left[\psi(z)^{G_t}\right].$$

But by the definition of the generating function for G_t,

$$\mathbb{E}\left[\psi(z)^{G_t}\right] = \psi_t(\psi(z)).$$

Thus,

$$\psi_{t+1}(z) = \psi_t(\psi(z)).$$

By the inductive hypothesis, $\psi_t(z)$ is the t-fold composition of ψ with itself, so we conclude that

$$\psi_{t+1}(z) = \underbrace{\psi(\psi(\cdots\psi(z)\cdots))}_{t+1 \text{ times}}.$$

Therefore, by induction, for all $t \geq 1$,

$$\psi_t(z) = \underbrace{\psi(\psi(\cdots\psi(z)\cdots))}_{t \text{ times}},$$

which implies in particular that

$$\psi_1(z) = \psi(z), \quad \psi_2(z) = \psi(\psi(z)), \quad \psi_3(z) = \psi(\psi(\psi(z))),$$

and so on.

2.3. (5 points) Consider a branching process with offspring distribution Poisson(2), that is, Poisson with mean 2. Calculate the extinction probability p to four decimal places.

Step 1. Determine the probability generating function (PGF).
The offspring distribution is Poisson(2), whose PGF is given by

$$f(s) \;=\; \mathbb{E}[s^X] \;=\; \exp\bigl(\lambda(s-1)\bigr)$$

where $\lambda = 2$ in our case. Hence

$$f(s) \;=\; \exp\bigl(2(s-1)\bigr).$$

Step 2. Find the fixed-point equation for the extinction probability.
By standard branching process theory, the extinction probability p is the smallest nonnegative root of the equation

$$p \;=\; f(p).$$

Substituting the PGF, we obtain

$$p \;=\; \exp\bigl(2(p-1)\bigr).$$

Step 3. Solve for p.
Rewriting:

$$p = e^{2p-2} \quad\Longleftrightarrow\quad \ln p = 2p \;-\; 2 \quad\Longleftrightarrow\quad 2p \;-\; \ln p = 2.$$

Since $\lambda = 2 > 1$, there exists a unique solution to this equation in the interval $(0, 1)$. This solution can be obtained either via iterative methods or direct numerical algorithms (e.g., the `nsolve` routine in a computer algebra system). *Step 4. Numerical approximation.*
A simple numerical method shows that

$$p \;\approx\; 0.2032.$$

Thus, to four decimal places, the extinction probability is

$$p \approx 0.2032.$$

2.7. Consider an irreducible, time-reversible Markov chain $\{X_t\}$ with $X_t \sim \pi$, where the distribution π is stationary. Let A be a subset of the state space. Let $0 < \alpha < 1$, and define on the same state space a Markov chain $\{Y_t\}$ having probability transition matrix Q satisfying, for $i \neq j$,

$$Q(i,j) = \begin{cases} \alpha P(i,j) & \text{if } i \in A \text{ and } j \notin A, \\ P(i,j) & \text{otherwise.} \end{cases}$$

Define the diagonal elements $Q(i,i)$ so that the rows of Q sum to 1.

(a) (8 points) What is the stationary distribution of $\{Y_t\}$, in terms of π and α?

Step 1: Detailed balance in separate "blocks". Consider three main cases of transitions from i to j:

- Case 1: $i, j \in A$. Then $Q(i,j) = P(i,j)$ (for $i \neq j$). By reversibility of P w.r.t. π, we have

$$\pi(i)\,P(i,j) = \pi(j)\,P(j,i).$$

  Hence, to satisfy time-reversal in this block, we want

$$\mu(i)\,P(i,j) \;=\; \mu(j)\,P(j,i).$$

  If we assume $\mu(i) = c_A\,\pi(i)$ for $i \in A$, then

$$c_A\,\pi(i)\,P(i,j) \;=\; c_A\,\pi(j)\,P(j,i),$$

  which holds if and only if $\pi(i)\,P(i,j) = \pi(j)\,P(j,i)$. But that is true by hypothesis. Hence any constant factor c_A works for transitions strictly within A.

- Case 2: $i, j \notin A$. The transition again is $Q(i,j) = P(i,j)$. A similar argument shows that setting $\mu(i) = c_B\,\pi(i)$ for $i \notin A$ works for these transitions, as long as $\{X_t\}$ is reversible w.r.t. π. Indeed:

$$c_B\,\pi(i)\,P(i,j) \;=\; c_B\,\pi(j)\,P(j,i) \quad\Longleftrightarrow\quad \pi(i)\,P(i,j) \;=\; \pi(j)\,P(j,i),$$

  which is again valid by π's detailed balance with P.

- Case 3: $i \in A$, $j \notin A$. Here $Q(i,j) = \alpha\, P(i,j)$ and $Q(j,i) = P(j,i)$ (since $j \notin A$). We want
$$\mu(i)\, \alpha\, P(i,j) \;=\; \mu(j)\, P(j,i).$$
Since $\mu(i) = c_A\, \pi(i)$ for $i \in A$ and $\mu(j) = c_B\, \pi(j)$ for $j \notin A$, the above becomes
$$c_A\, \pi(i)\, \alpha\, P(i,j) \;=\; c_B\, \pi(j)\, P(j,i).$$
But by the original chain's time-reversibility, $\pi(i)\, P(i,j) \;=\; \pi(j)\, P(j,i)$. Thus the left side is $c_A\, \alpha\, \pi(j)\, P(j,i)$, and matching both sides yields
$$c_A\, \alpha\, \pi(j)\, P(j,i) \;=\; c_B\, \pi(j)\, P(j,i) \quad\Longrightarrow\quad c_A\, \alpha = c_B \quad \bigl(\text{assuming } \pi(j)\, P(j,i) \neq 0\bigr).$$
Hence we get a single equation connecting c_A and c_B.
- Case 4: $i \notin A$, $j \in A$. By symmetry, $Q(i,j) = P(i,j)$ and $Q(j,i) = \alpha\, P(j,i)$. We similarly obtain $c_B = \alpha\, c_A$.

Step 2: Normalization. From Case 3 (and 4) we have
$$c_B \;=\; \alpha\, c_A.$$
We also need $\sum_{i \in \mathcal{S}} \mu(i) = 1$, which gives
$$1 \;=\; \sum_{i \in A} \mu(i) \;+\; \sum_{j \notin A} \mu(j) \;=\; c_A \sum_{i \in A} \pi(i) \;+\; c_B \sum_{j \notin A} \pi(j).$$
Let $s_A = \sum_{i \in A} \pi(i)$ and $s_B = \sum_{j \notin A} \pi(j)$. Note $s_A + s_B = \sum_{i \in \mathcal{S}} \pi(i) = 1$. Hence
$$1 \;=\; c_A\, s_A \;+\; (\alpha\, c_A)\, s_B \;=\; c_A\, \bigl[s_A + \alpha\, s_B\bigr],$$
so
$$c_A \;=\; \frac{1}{s_A + \alpha\, s_B}, \qquad c_B \;=\; \frac{\alpha}{s_A + \alpha\, s_B}.$$
Thus the stationary distribution for the chain $\{Y_t\}$ is
$$\mu(i) \;=\; \begin{cases} \dfrac{\pi(i)}{s_A + \alpha\, s_B}, & i \in A, \\ \dfrac{\alpha\, \pi(i)}{s_A + \alpha\, s_B}, & i \notin A. \end{cases}$$
Since $s_A = \sum_{i \in A} \pi(i)$ and $s_B = 1 - s_A$, we can rewrite the denominator if desired. Either way, the answer is an explicit function of (π, α):
$$\boxed{\mu(i) \;=\; \begin{cases} \dfrac{\pi(i)}{\sum_{k \in A} \pi(k) \;+\; \alpha\, \sum_{\ell \notin A} \pi(\ell)}, & i \in A, \\ \dfrac{\alpha\, \pi(i)}{\sum_{k \in A} \pi(k) \;+\; \alpha\, \sum_{\ell \notin A} \pi(\ell)}, & i \notin A. \end{cases}}$$

(b) (2 points) Show that the chain $\{Y_t\}$ is also time-reversible.

To prove $\{Y_t\}$ is time-reversible, it suffices to show that μ (just derived) satisfies the detailed balance condition
$$\mu(i)\, Q(i,j) \;=\; \mu(j)\, Q(j,i) \quad \text{for all } i, j.$$
But we have actually verified this already in the course of deriving μ.

- Inside A or inside $\mathcal{S} \setminus A$, Q is identical to P and μ is proportional to π. So detailed balance holds by the original reversibility of π and P.
- Across the boundary between A and $\mathcal{S} \setminus A$, the factor α appears on one side but not the other. This mismatch is corrected by the relation $c_B = \alpha\, c_A$ in μ. Hence it maintains the equal product form $\mu(i)\, Q(i,j) = \mu(j)\, Q(j,i)$ in those cross transitions as well.

Therefore,

$$\mu(i)\,Q(i,j) \;=\; \mu(j)\,Q(j,i) \quad \forall i,j \quad\Longrightarrow\quad \{Y_t\} \text{ is time-reversible w.r.t. } \mu.$$

Since time-reversibility implies stationarity as well, μ is indeed the stationary distribution.

(c) (5 points) Show by example that the simple relationship of part (1) need not hold if we drop the assumption that X is reversible.

Consider a 3-state Markov chain with states $\{1,2,3\}$ and transition matrix P given by

$$P \;=\; \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

This chain is a directed cycle $1 \to 2 \to 3 \to 1$. Its unique stationary distribution is $\pi = (1/3,\, 1/3,\, 1/3)$. But the chain is not reversible because for instance, $\pi(1)\,P(1,2) = \frac{1}{3}\cdot 1 = \frac{1}{3}$ while $\pi(2)\,P(2,1) = \frac{1}{3}\cdot 0 = 0$, which are not equal. Let $A = \{1\}$ and $\alpha \in (0,1)$. Define a modified chain Q by

$$Q(1,2) \;=\; \alpha, \quad Q(1,3) = 0, \quad Q(1,1) = 1-\alpha, \quad \text{and} \quad Q(i,j) = P(i,j) \text{ for } i \neq 1.$$

If we try to "guess" a stationary distribution of the simple form $\mu(1) = c_A\,\pi(1)$, $\mu(2) = c_B\,\pi(2)$, $\mu(3) = c_B\,\pi(3)$ as if the chain were reversible, we would end up imposing relations akin to

$$\mu(1)\,\alpha \;=\; \mu(2)\,P(2,1), \quad \mu(1)\,\alpha\,P(1,2) \;=\; \mu(2)\,P(2,1), \quad \ldots$$

but $P(2,1) = 0$, so these quickly lead to contradictions (or trivialize $\alpha = 0$). Hence, the neat factorization from part (a) breaks down in the nonreversible case.

2.12. [Metropolis-Hastings method] For simplicity, let us assume that π is positive, so that we won't have to worry about dividing by 0. Choose any probability transition matrix $Q = (Q(i,j))$ [again, suppose it is positive], and define $P(i,j)$ for $i \neq j$ by

$$P(i,j) = Q(i,j)\min\left(1, \frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}\right),$$

and of course define $P(i,i) = 1 - \sum_{j\neq i} P(i,j)$.

(a) (5 points) Show that the probability transition matrix P has stationary distribution π.

Fix two distinct states $i \neq j$. By the definition of P, we have

$$P(i,j) \;=\; Q(i,j)\;\min\!\Big(1, \frac{\pi(j)\,Q(j,i)}{\pi(i)\,Q(i,j)}\Big), \quad P(j,i) \;=\; Q(j,i)\;\min\!\Big(1, \frac{\pi(i)\,Q(i,j)}{\pi(j)\,Q(j,i)}\Big).$$

Consider the product $\pi(i)\,P(i,j)$. Substituting from above,

$$\pi(i)\,P(i,j) \;=\; \pi(i)\,Q(i,j)\;\min\!\Big(1, \frac{\pi(j)\,Q(j,i)}{\pi(i)\,Q(i,j)}\Big).$$

We analyze the ratio

$$\frac{\pi(j)\,Q(j,i)}{\pi(i)\,Q(i,j)}.$$

Let us denote this ratio by $r > 0$ (since π and Q are strictly positive). Note that

$$\min(1,r) \;=\; \begin{cases} r, & r \le 1, \\ 1, & r \ge 1. \end{cases}$$

Hence,

$$\pi(i)\,P(i,j) \;=\; \begin{cases} \pi(i)\,Q(i,j)\,r, & \text{if } r \le 1, \\ \pi(i)\,Q(i,j), & \text{if } r \ge 1. \end{cases} \;=\; \begin{cases} \pi(j)\,Q(j,i), & \text{if } r \le 1, \\ \pi(i)\,Q(i,j), & \text{if } r \ge 1. \end{cases}$$

But precisely the same reasoning applies to $\pi(j)\,P(j,i)$, with the roles of i and j reversed. In particular,

$$\pi(j)\,P(j,i) \;=\; \begin{cases} \pi(j)\,Q(j,i), & \text{if } r \geq 1, \\ \pi(i)\,Q(i,j), & \text{if } r \leq 1. \end{cases}$$

Hence, in either case ($r \leq 1$ or $r \geq 1$), the two products $\pi(i)\,P(i,j)$ and $\pi(j)\,P(j,i)$ coincide. Therefore

$$\pi(i)\,P(i,j) \;=\; \pi(j)\,P(j,i), \quad \text{for each } i \neq j.$$

Since this holds for all $i \neq j$, the detailed balance condition is satisfied. It follows that π is indeed a stationary (and actually *reversible*) distribution for P:

$$\sum_i \pi(i)\,P(i,j) \;=\; \sum_i \pi(j)\,P(j,i) \;=\; \pi(j) \sum_i P(j,i) \;=\; \pi(j).$$

(b) (5 points) Show how the Metropolis method we have discussed is a special case of this Metropolis-Hastings method.

In the *Metropolis* algorithm (sometimes called the *independence sampler*), we typically choose a symmetric proposal distribution $Q(i,j)$, i.e.

$$Q(i,j) \;=\; Q(j,i) \quad \forall i,j.$$

In the simplest classic setting, $Q(i,j)$ might be

$$Q(i,j) \;=\; \begin{cases} \frac{1}{(\text{degree of } i)}, & \text{if } j \text{ is a neighbor of } i, \\ 0, & \text{otherwise.} \end{cases}$$

In any case, if Q is symmetric then

$$\frac{\pi(j)\,Q(j,i)}{\pi(i)\,Q(i,j)} \;=\; \frac{\pi(j)}{\pi(i)}.$$

Thus the Metropolis–Hastings update rule

$$P(i,j) \;=\; Q(i,j)\,\min\Bigl(1, \frac{\pi(j)\,Q(j,i)}{\pi(i)\,Q(i,j)}\Bigr)$$

becomes

$$P(i,j) \;=\; Q(i,j)\,\min\Bigl(1, \frac{\pi(j)}{\pi(i)}\Bigr).$$

This is exactly the classical Metropolis acceptance probability

$$a(i,j) \;=\; \min\bigl(1, \tfrac{\pi(j)}{\pi(i)}\bigr),$$

multiplied by $Q(i,j)$ to decide how often we propose a move from i to j in the first place. Hence, when Q is symmetric, the Metropolis–Hastings chain reduces to the usual Metropolis algorithm. Therefore, the Metropolis method is indeed a *special case* of Metropolis–Hastings.

3.9. (15 points) Derive the recursion,

$$\beta_{t-1}(x_{t-1}) = \sum_{x_t} A(x_{t-1}, x_t) B(x_t, y_t) \beta_t(x_t).$$

for the "backward" probabilities. Show that it is appropriate to start the calculations by setting

$$\beta_n(x_n) = 1 \quad \text{for all } x_n \in \mathcal{X}.$$

1. Definition of backward probabilities. Backward probability:
For $t = 1, 2, \ldots, n$, define

$$\beta_t(x_t) \;:=\; p\bigl(y_{t+1}, y_{t+2}, \ldots, y_n \mid x_t\bigr),$$

where $\{x_t\}$ denotes the underlying (hidden) state sequence and $\{y_t\}$ the observed emission sequence. Thus $\beta_t(x_t)$ is the conditional probability of "all future observations" $y_{t+1}, \ldots, y_n$ given that the hidden chain is in state x_t at time t. **2. Derivation of the backward recursion.** To derive the recursion for

$\beta_{t-1}(x_{t-1})$, we write:

$$\beta_{t-1}(x_{t-1}) \;=\; p\big(y_t, y_{t+1}, \ldots, y_n \;\big|\; x_{t-1}\big).$$

We note that

$$p(y_t, y_{t+1}, \ldots, y_n \mid x_{t-1}) \;=\; \sum_{x_t \in \mathcal{X}} p\big(y_t, y_{t+1}, \ldots, y_n, x_t \;\big|\; x_{t-1}\big),$$

where we sum over all possible states x_t that the chain might occupy at time t. Next, we factor the joint probability inside the sum as

$$p\big(y_t, y_{t+1}, \ldots, y_n, x_t \;\big|\; x_{t-1}\big) \;=\; p\big(x_t \;\big|\; x_{t-1}\big)\, p\big(y_t, y_{t+1}, \ldots, y_n \;\big|\; x_{t-1}, x_t\big).$$

Since the transition from x_{t-1} to x_t is given by $A(x_{t-1}, x_t)$ under the Markov property, we identify

$$p\big(x_t \;\big|\; x_{t-1}\big) \;=\; A(x_{t-1}, x_t).$$

Furthermore, the probability of the future observations $y_t, y_{t+1}, \ldots, y_n$ given x_{t-1}, x_t can be split, by conditional independence assumptions of the HMM, as follows:

$$p\big(y_t, y_{t+1}, \ldots, y_n \;\big|\; x_{t-1}, x_t\big) \;=\; p\big(y_t \;\big|\; x_t\big)\, p\big(y_{t+1}, \ldots, y_n \;\big|\; x_t\big).$$

By the definition of the emission probabilities, we have

$$p\big(y_t \;\big|\; x_t\big) \;=\; B(x_t, y_t),$$

and by the definition of $\beta_t(x_t)$ as the probability of all future observations given the current state x_t, we identify

$$p\big(y_{t+1}, \ldots, y_n \;\big|\; x_t\big) \;=\; \beta_t(x_t).$$

Putting all these pieces together, we get

$$p\big(y_t, y_{t+1}, \ldots, y_n \mid x_{t-1}\big) \;=\; \sum_{x_t \in \mathcal{X}} A(x_{t-1}, x_t)\; B(x_t, y_t)\; \beta_t(x_t).$$

Hence,

$$\beta_{t-1}(x_{t-1}) \;=\; \sum_{x_t \in \mathcal{X}} A(x_{t-1}, x_t)\; B(x_t, y_t)\; \beta_t(x_t).$$

This proves the backward recursion. **3. Initial condition for backward probabilities.** Finally, we

justify why we set $\beta_n(x_n) = 1$ for all x_n. By the definition,

$$\beta_n(x_n) \;=\; p\big(y_{n+1}, \ldots, y_n \;\big|\; x_n\big).$$

However, for $t = n$, the expression $y_{n+1}, \ldots, y_n$ corresponds to an *empty set* of observations (i.e. there are no future observations after time n). The probability of observing an empty set of future outcomes is conventionally 1, since we have

$$p\big(\varnothing \mid x_n\big) \;=\; 1.$$

Therefore, we set

$$\beta_n(x_n) = 1,$$

which initializes our backward recursion.