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

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

See that the goal is to transform the stationary distribution 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 s.t. $(i,j) \in E$,

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

Since $\pi(i) = \pi(j) = \frac{1}{|V|}$ in a uniform distribution, the acceptance probabilities become

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

In order to ensure that the rows of the resulting transition matrix sum to 1, we define the self-transition probability as

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

Thus,

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

In order to confirm that π is the stationary distribution of the modified chain, we can check the detailed balance condition. See that $\forall i, j \in V$ s.t. $(i,j) \in E$,

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

Note 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 $\forall i, j$ s.t. $(i,j) \in E$ (and this chain is a time-reversible MC). Therefore, π is the stationary distribution.

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 state space as the set of feasible solutions,

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

Trivially, is the state is not within this set, the sum of the weights will exceed the budget. The chain is initialized at the empty knapsack $x = (0, 0, \ldots, 0)$. Given the current state $x \in C$, the chain proceeds by

1. Choosing index j uniformly at random from $\{1, 2, \ldots, m\}$, i.e. with probability $1/m$)
2. Define x^j as the state obtained from x by flipping the $j-$th coordinate such that

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

3. If and only if $x^j \in C$, then update x to x^J. Otherwise, do not change x.

Formally, if

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

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

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

Note that if $x, y \in C$, define x and y as two states that only differ in one coordinate, i.e. $y = x^J$ for some J. From the definition of the transition probabilities,

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

since the move $x \to y$ (and $y \to x$) is feasible. Thus, for such neighboring states,

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

For moves in which the chain stays in the same state due to the proposed move being infeasible, the balance is trivial, as $P(x, x) = P(x, x)$. Hence, the detailed balance condition holds for all transitions, and the uniform distribution over C is 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.

Modifying the base chain from part (a) using the Metropolis algorithm, we can start from a current state $x \in C$ and do the following:

1. First, propose a candidate y using an index $j \in \{1, \ldots, m\}$ picked uniformly at random and setting $y = x^j$, conditioning on that if $x^j \notin C$, then we set $y = x$.
2. If $y \neq x$, or the proposed state is inviable, accept the move with probability

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

otherwise, do not update the state and stay at x.

Thus, for any $x \in C$ and for any $j \in \{1, \ldots, m\}$, let x^j denote the state obtained by flipping the $j-$th coordinate. This yields a modified transition probability

$$P(x, y) = \begin{cases} \frac{1}{m} \min\Big\{1, \exp\big(\beta(f(y) - f(x))\big)\Big\}, & \text{if } x \neq y, y = x^j \in C \text{ for some } j \\ 1 - \sum\limits_{j:\, x^j \in C,\, x^j \neq x} \frac{1}{m} \min\Big\{1, \exp\big(\beta(f(x^j) - f(x))\big)\Big\} - \sum\limits_{j:\, x^j \notin C} \frac{1}{m}, & \text{if } x = y \\ 0, \quad \text{otherwise} \end{cases}$$

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.

Recall an alternative characterization of the total variation distance,

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

where the supremum is taken over all functions $f : \mathcal{S} \to \mathbb{R}$ with $\|f\|_\infty \leq 1$, or $|f(x)| \leq 1 \; \forall x \in \mathcal{S}$.
For any function f with $\|f\|_\infty \leq 1$,

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

Interchanging the sums,

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

In order to bound the right side, define a function g as

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

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

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

Therefore, $\|g\|_\infty \leq 1$.
Using the definition of total variation distance for the pair (π_0, ρ_0),

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

Since this inequality holds for every function f with $\|f\|_\infty \leq 1$ and the corresponding g satisfies $\|g\|_\infty \leq 1$),

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

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.

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

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

Since G_1 is the number of offspring of the initial individual, trivially

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

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

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

Note that each individual in generation t produces offspring independently given 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 $\psi(z)$,

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

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, so

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

Therefore, by induction, $\forall t \geq 1$,

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

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.

The offspring distribution is Poisson(2), whose probability generating function is given by

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

where $\lambda = 2$. Hence,

$$f(s) \;=\; \exp\big(2(s-1)\big)$$

By standard branching process theory, the extinction probability p is the smallest nonnegative root of the equation

$$p \;=\; f(p)$$
$$p \;=\; \exp\big(2(p-1)\big)$$
$$p = \exp\left(2p-2\right)$$
$$\ln p = 2p \;-\; 2$$
$$2p \;-\; \ln p = 2$$

Since $\lambda > 1$, there must exist a unique solution to this equation within the interval $(0,1)$. p can be approximated to four decimal places as

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

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?

Consider the 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,

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

In order to satisfy the property of time-reversibility for Q, define

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

Assuming $\mu(i) = c_A\,\pi(i)$ for $i \in A$ for some c_A, then

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

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

- Case 2: $i, j \notin A$. The transition is trivially $Q(i,j) = P(i,j)$. A similar argument can be constructed to show that any constant c_B can be used for $\mu(i) = c_B\,\pi(i)$ for $i \notin A$ to represent these transitions, as long as $\{X_t\}$ is reversible w.r.t. π.

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

which trivially must be valid from π's detailed balance with P.

- Case 3: $i \in A$, $j \notin A$. See that $Q(i,j) = \alpha\,P(i,j)$ and $Q(j,i) = P(j,i)$, since $j \notin A$. To show time-reversibility, we need to find a constant c_A and c_B such that

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

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

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

So,

$$c_A\,\alpha = c_B \quad \big(\text{assuming } \pi(j)\,P(j,i) \neq 0\big)$$

- 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 $c_B = \alpha\,c_A$.

In order for $\sum_{i\in\mathcal{S}} \mu(i) = 1$,

$$\sum_{i\in A}\mu(i) \;+\; \sum_{j\notin A}\mu(j) \;=\; c_A\sum_{i\in A}\pi(i) \;+\; c_B\sum_{j\notin A}\pi(j) = 1$$

Define $s_A = \sum_{i\in A}\pi(i)$ and $s_B = \sum_{j\notin A}\pi(j)$. Note that $s_A + s_B = \sum_{i\in\mathcal{S}}\pi(i) = 1$. As such,

$$c_A s_A + c_B s_B = c_A\,s_A \;+\; (\alpha\,c_A)\,s_B \;=\; c_A\left[s_A + \alpha\,s_B\right] = 1$$

$$c_A s_A + c_B s_B = \frac{c_B}{\alpha}\,s_A \;+\; c_B\,s_B \;=\; c_B\left(\frac{s_A + \alpha s_B}{\alpha}\right) = 1$$

$$c_A \;=\; \frac{1}{s_A + \alpha\,s_B} \qquad \text{and} \qquad c_B \;=\; \frac{\alpha}{s_A + \alpha\,s_B}$$

Thus the stationary distribution for the chain $\{Y_t\}$ is

$$\mu(i) \;=\; \begin{cases} \dfrac{\pi(i)}{s_A + \alpha\,s_B}, & i \in A \\[2ex] \dfrac{\alpha\,\pi(i)}{s_A + \alpha\,s_B}, & i \notin A \end{cases}$$

Since $s_A = \sum_{i \in A} \pi(i)$ and $s_B = 1 - s_A$,

$$\mu(i) \;=\; \begin{cases} \dfrac{\pi(i)}{\sum_{k \in A} \pi(k) \;+\; \alpha \sum_{\ell \notin A} \pi(\ell)}, & i \in A \\ \dfrac{\alpha\, \pi(i)}{\sum_{k \in A} \pi(k) \;+\; \alpha \sum_{\ell \notin A} \pi(\ell)}, & i \notin A \end{cases}$$

(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 satisfies the detailed balance condition

$$\mu(i)\, Q(i,j) \;=\; \mu(j)\, Q(j,i) \quad \text{for all } i, j$$

However, recall that this was already defined,

- If the step is 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.
- If the step is 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 trivally maintains the equal product form $\mu(i)\, Q(i,j) = \mu(j)\, Q(j,i)$ in those cross transitions as well.

Therefore,

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

As such, $\{Y_t\}$ is time-reversible w.r.t. μ.

(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,

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

This chain is a directed cycle $1 \to 2 \to 3 \to 1$ with a unique stationary distribution is $\pi = (1/3,\, 1/3,\, 1/3)$. However, note that the chain is not reversible. Consider $\pi(1)\, P(1,2) = \frac{1}{3} \cdot 1 = \frac{1}{3} \neq \pi(2)\, P(2,1) = \frac{1}{3} \cdot 0 = 0$. Let $A = \{1\}$ and $\alpha \in (0,1)$. Define a modified chain Q by

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

Using the simple relationship defined above, the stationary distribution of Q should be $\mu(1) = c_A\, \pi(1)$, $\mu(2) = c_B\, \pi(2)$, $\mu(3) = c_B\, \pi(3)$

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

However, note that $P(2,1) = 0$. As such, this quickly leads to contradictions or trivial case where $\alpha = 0$. Hence, the simple relationship from part (a) breaks down when we cannot assume X to be reversible.

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,

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

Substituting yields,

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

Analyzing the ratio, $\frac{\pi(j)\,Q(j,i)}{\pi(i)\,Q(i,j)}$, first see that this ratio, r, must be $r > 0$, since π and Q are strictly positive. Therefore,

$$\min(1,r) \;=\; \begin{cases} r, & r \leq 1 \\ 1, & r \geq 1 \end{cases}$$

Thus,

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

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

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

In either case, 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 $\forall i \neq j$, the detailed balance condition is satisfied. It follows that π is indeed a stationary and reversible distribution for P, s.t.

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

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

In the Metropolis method, we typically choose a symmetric proposal distribution $Q(i,j)$,

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

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

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

becomes

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

However, see that this is exactly the classical Metropolis acceptance probability

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

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.

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

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

For $t = 1, 2, \ldots, n$, define

$$\beta_t(x_t) \;=\; p\big(y_{t+1}, y_{t+2}, \ldots, y_n \;\big|\; x_t\big)$$

where $\{x_t\}$ denotes the underlying, or 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. To derive the recursion for $\beta_{t-1}(x_{t-1})$, see that

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

Note that

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

Factoring the joint probability inside the sum,

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

However, note that the transition from x_{t-1} to x_t is given by $A(x_{t-1}, x_t)$ under the Markov property, s.t.

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

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,

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

By the definition of the emission probabilities,

$$p(y_t \mid x_t) \;=\; B(x_t, y_t)$$

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

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

Therefore,

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

Thus, backwards recursion is derived.
Solving for the initial conditions for the backward probabilities, see by the definition,

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

However, note that for $t = n$, the expression $y_{n+1}, \ldots, y_n$ corresponds to an empty set of observations, since there are no future observations after time n. As such,

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

Therefore, in order to initialize the backward recursion, we set

$$\beta_n(x_n) = 1$$