Solutions to Problems

Stochastic Learning and Optimization

– A Sensitivity-Based Approach

published by Springer 2007

Xi-Ren Cao

The Hong Kong University of Science and Technology

email: eecao@ust.hk

Co-authored with (in alphabetical order)

Fang Cao, Xianping Guo, Yanjie Li, Li Xia, Yankai Xu, and Junyu Zhang

January 22, 2008

Some problems are meant to stimulate new research ideas. Solutions provided in this manuscript are only for your references. Comments and suggestions are welcome.

Xi-Ren Cao

December 2007

Contents

1

Solutions to Chapter 1

1.1 Give an example of a real world problem that fits the general model of learning and optimization illustrated in Figure 1.1.

[Solution]

In the Internet routing protocols, the routing problem can be viewed as a good example that fits the general model of learning and optimization.

In the routing problem, the goal of protocols is to find an optimal route from the source computer to the destination computer. Between the source and destination node, there are many routers which can relay the data packets. Routing protocol is to choose a set of routers to relay the packets efficiently.

The input action is the relay probabilities of each router. It is supposed that with high relay probability, the router would more likely relay these packets. We can adjust these relay probabilities to get a good routes.

The destination computer can use the lost-packet rate and transmission delay to qualify the routes' performance. The performance can be observed through destination computer. These are the output variables.

The optimization problem is to adjust the relay probabilities of each router to get a good routes performance. The detailed construction and information in the internal network may be very complicated. We can use the learning and optimization method to learn the network behaviors and optimize the relay routes.

1.2 A person travels from the star point shown in Figure 1.20 to one of the seven destinations indicated as the circles in the figure. The person may receive a reward shown as the number in the corresponding circle when she/he reaches a destination. There are three time steps, $l = 0, 1, 2$, in this problem. The letters $\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3}$ and $\alpha_{2,1}, \ldots,$ near the arrows represent the actions. Develop an optimal policy for the person to receive the biggest reward. Note that there is more than one optimal policy.

[Solution]

At first we consider the open-loop policy, we can find the best reward is 10. Action sequences $\{\alpha_{1,1}, \alpha_{2,2}, \alpha_{3,2}\}$ $\{\alpha_{1,2}, \alpha_{2,2}, \alpha_{3,1}\}$ can reach this optimal reward.

Next, we consider the policy depending on the action history. We can know the optimal policy has three sections as below.

Step1: $d_0(\emptyset) = \{\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3}\}$;

Step2: $d_1(\alpha_{1,1}) = \{\alpha_{2,2}\}, d_1(\alpha_{1,2}) = \{\alpha_{2,2}\}, d_1(\alpha_{1,3}) = \{\alpha_{2,2}\} \Rightarrow d_1 = \{\alpha_{2,2}\}$;

Step3: $d_2(\alpha_{1,1}, \alpha_{2,1}) = \{\alpha_{3,2}\}, d_2(\alpha_{1,1}, \alpha_{2,2}) = \{\alpha_{3,2}\}, d_2(\alpha_{1,2}, \alpha_{2,1}) = \{\alpha_{3,1}\}, d_2(\alpha_{1,2}, \alpha_{2,2}) = \{\alpha_{3,1}\}, d_2(\alpha_{1,3}, \alpha_{2,1}) = \{\alpha_{3,1}\}, d_2(\alpha_{1,3}, \alpha_{2,2}) = \{\alpha_{3,1}\} \Rightarrow d_2(\alpha_{1,1}, \alpha_{2,i}) = \{\alpha_{3,2}\}, d_2(\alpha_{1,2}, \alpha_{2,i}) = \{\alpha_{3,1}\}, d_2(\alpha_{1,3}, \alpha_{2,i}) = \{\alpha_{3,1}\},\ i = 1, 2$

So the optimal policy is derived: $d = \{d_0, d_1, d_2\}$.

1.3 In Example 1.2, at $l = 0$, there are two possible observations y_0 and y_1. Thus, the number of possible sub-policy $d_0 : \{y_0, y_1\} \rightarrow \{\alpha_0, \alpha_1\}$ is $2^2 = 4$. Next, if we do not follow any policy at $l = 0$, then at $l = 1$, there are eight possible different histories $\{Y_0, A_0, Y_1\}$. In this case, at time $l = 1$ every policy d_1 needs to specify an action for every one of these eight different action-observation histories. Thus, there are $2^{2^3} = 2^8 = 256$ possible sub-policies d_1 at $l = 1$. However, if we follow any sub-policy at $l = 0$, because $A_0 = d_0(Y_0)$,

we only have four (instead of eight) possible different histories for each d_0. Therefore, if we follow any sub-policy at $l = 0$, each sub-policy at $l = 1$ needs to specify actions for these four different action-observation histories. That is, for each sub-policy d_0, there are only 2^4 different sub-policies d_1 at $l = 1$. Thus, there are altogether $2^2 \times 2^4 = 64$ different combined policies $\{d_0, d_1\}$. Convince yourself about the above argument, and continue to calculate how many policies there are for $\boldsymbol{d} = \{d_0, d_1, d_2\}$.

[solution]

At time l=1 the history sequence is $\{Y_0, A_0, Y_1\}$, thus the number of histories is indeed 2^3=8. But in fact the action A_0 is decided by policy d_0. Therefore, for fixed policy d_0, at time $l = 1$, there are only 2^2 possible different histories.

From this point, we can know that at time l=0, the number of policies is $|d_0| = 2^2$. At time l=1, the policy number is $|d_1| = 2^{2^2}$ if we follow the policy d_0.

At time l=2 the history sequence is $\{Y_0, A_0, Y_1, A_1, Y_2\}$, the number of histories is indeed 2^5=32. But in fact the actions A_0 and A_1 are decided by policy d_1. Therefore, for fixed policy d_1, at time $l = 2$, there are only 2^3 possible different histories. Thus, at time l=2, the policy number is $|d_2| = 2^{2^3}$.

So the total policies space size is $|d| = |d_0| \times |d_1| \times |d_2| = 2^2 \times 2^{2^2} \times 2^{2^3} = 2^{14}$ if we follow policy $\boldsymbol{d}$.

1.4 Prove that the optimal feedback policy based on observations performs better than the optimal open-loop policy on average (cf. Example 1.2).

[solution]

As mentioned in Table 1.2, we assume that the observations of the system at time $l = 0$ and $l = 1$ are fixed as y_0 and y_1, respectively, and the probabilities of $Y_2 = y_0$ and $Y_2 = y_1$ are both 0.5. From Table 1.2, we can know that if we use the optimal open-loop action sequence $(\alpha_1, \alpha_0, \alpha_1)$, then the optimal reward is $\frac{1}{2}(12 + 8) = 10$. However, from Table 1.2, it is clear that given the history (up to $l = 1$) $\{y_0, \alpha_1, y_1, \alpha_0\}$, if we observe $Y_2 = y_0$, we definitely should take action α_1 at $l = 2$ to receive a reward of 12. But, if we observe $Y_2 = y_1$ at $l = 2$, we should take α_0 to receive a reward of 10 (instead of taking α_1 to get 8). Thus, the optimal feedback policy based on observations can obtain a better

performance, which is $\frac{1}{2}(12+10) = 11$ on average.

1.5 Consider an MDP with state space $\mathcal{S} = \{1, 2, \ldots, S\}$. Let the action space be $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_S\}$; suppose that when action α_j is taken in any state, the system will, with probability one, move to state j, $j = 1, 2, \ldots, S$.

a. For any $i \in \mathcal{S}$, define a distribution on $\mathcal{A}$ as $\nu_i = \{p(1|i), p(2|i), \ldots, p(S|i)\}$. Let $\nu_i = d(i)$, $i \in \mathcal{S}$, be a randomized policy defined as follows: In any state i, $i \in \mathcal{S}$, action α_j is taken with probability $p(j|i)$, $j \in \mathcal{S}$. What is the Markov chain under this policy $\nu_i = d(i)$, $i \in \mathcal{S}$?

b. Let $\alpha^{(1)}$ and $\alpha^{(2)}$ represent another two actions: If $\alpha^{(k)}$ is taken at state i, then the system moves according to the probability distribution $\nu_i^{(k)} = \{p^{(k)}(1|i), p^{(k)}(2|i), \ldots, p^{(k)}(S|i)\}$, $k = 1, 2$. Let $\nu_i = d(i)$, $i \in \mathcal{S}$, be a randomized policy defined as follows: At any state i, action $\alpha^{(1)}$ is taken with probability p_i, and action $\alpha^{(2)}$ is taken with probability q_i, $p_i + q_i = 1$, $i = 1, 2, \ldots, S$. What is the Markov chain under this policy $\nu_i = d(i)$, $i \in \mathcal{S}$?

[solution]

a. Since the random policy takes the action α_j with probability $p(j|i)$ and the system will move to state j when action α_j is taken, we can know the system will move to state j with probability $p(j|i)$. So, the transition probability matrix of the Markov chain under the policy $\nu_i = d(i), i \in \mathcal{S}$, is $P = [p(j|i)]_{i,j=1}^{S}$.

b. The transition probability from state i to state j is $p^{\nu}(j|i) = p_i p^{(1)}(j|i) + (1 - p_i)p^{(2)}(j|i)$. So, the transition probability matrix of the Markov chain under the policy ν is $P = [p^{\nu}(j|i)]_{i,j=1}^{S}$.

1.6 Consider a two-state process $\widetilde{\boldsymbol{X}}$ with history-dependent transition probabilities $p[1|(1,1)] = 0$, $p[0|(1,1)] = 1$; $p[1|(0,0)] = 1$, $p[0|(0,0)] = 0$; $p[0|(1,0)] = 1$, $p[1|(1,0)] = 0$; and $p[1|(0,1)] = 1$, $p[0|(0,1)] = 0$.

a. Draw a sample path of $\widetilde{\boldsymbol{X}}$. What property does it have?

b. Derive the equivalent Markov chain $\boldsymbol{X}$ as shown in Example 1.3.

c. Suppose that the reward function depends on three consecutive states (X_l, X_{l+1}, X_{l+2}) and is defined as $f(1,1,1) = f(0,0,0) = 100$ and $f(i,j,k) = 0$ otherwise. Explain that the steady-state performance measures for both $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}$ defined as $\widetilde{\eta} = \sum_{i,j,k} \widetilde{\pi}(i,j,k) f(i,j,k)$ and $\eta = \sum_{i,j,k} \pi(i,j,k) f(i,j,k)$, respectively, are different.

[**solution**]

a. A sample path of $\widetilde{\boldsymbol{X}}$ is $\{0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,\ldots\}$. The process is periodic and its period is 4.

b. Define $Y_l = \{\widetilde{X}_{l-1}, \widetilde{X}_l\}, l = 1,2,\ldots$. Then the process $\mathbf{Y} = \{Y_1, Y_2, \ldots\}$ is a Markov chain with state space $\mathcal{S} = \{(0,0),(0,1),(1,1),(1,0)\}$ and its transition probability matrix is

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

The steady state distribution of $\mathbf{Y}$ is $(1/4, 1/4, 1/4, 1/4)$. Similarly to Example 1.3, let $\pi(0) = \sum_{k'=0,1} \pi(k',0) = 1/2, \pi(1) = \sum_{k'=0,1} \pi(k',1) = 1/2$. we can obtain the transition probability matrix of an equivalent Markov chain $\boldsymbol{X}$ with state space $\mathcal{S} = \{0,1\}$

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}.$$

by $p(k|j) = \sum_{i \in \mathcal{S}} \left\{ \frac{\pi(i,j)}{\pi(j)} p[k|(i,j)] \right\}, j,k = 0,1.$

c. For process $\widetilde{\boldsymbol{X}}$, the case $(\widetilde{X}_l = 1, \widetilde{X}_{l+1} = 1, \widetilde{X}_{l+2} = 1)$ or $(\widetilde{X}_1 = 0, \widetilde{X}_{l+1} = 0, \widetilde{X}_{l+2} = 0)$ does not occur, so the steady-state performance $\eta = 0$. However, for Markov chain $\boldsymbol{X}$, we know the steady-state probability that $\widetilde{X}_l = i, \widetilde{X}_{l+1} = j, \widetilde{X}_{l+2} = k$ is $\pi(i,j,k) = \pi(i) * p(j|i) * p(k|j)$, so, we have $\pi(1,1,1) = 1/8$ and $\pi(0,0,0) = 1/8$, then the steady-state performance $\eta = 100 * 1/8 + 100 * 1/8 = 25$. Thus, the steady-state performance measures are different for both $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}$.

1.7 The exhaustive search algorithm presented in Section 1.1.3 is very "robust". Suppose that because of the estimation error, the relationship $\eta^{d_i} > \widetilde{\eta}$ cannot be accurately verified.

a. If d_M is an optimal policy, then the algorithm outputs a correct optimal policy if only the last comparison is correctly made.

b. Explain that the algorithm outputs the optimal policy as long as the comparisons $\eta^{d_i} > \widetilde{\eta}$ are correctly made when η^{d_i} or $\widetilde{\eta}$ is the optimal performance.

c. Suppose η^* is the best performance and η^*_- is the next to the best performance, and set $\delta = \eta^* - \eta^*_-$. Then the algorithm outputs the correct optimal policy if the estimation error for the performance is always smaller than $\delta/2$.

[solution]

a. Since d_M is the optimal policy, that is, $\eta^{d_M} > \eta^{d_i}, i = 1, \ldots, M-1$. If the last comparison is correctly made, i.e., $\widetilde{\eta} < \eta^{d_M}$, then, the algorithm must output $\widetilde{d} = d_M$. That is, the algorithm outputs a correct optimal policy.

b. When the optimal performance is η^{d_i}, since the comparison between η^{d_i} and $\widetilde{\eta}$ is correctly made, then, we have $\widetilde{d} = d_i$ and $\widetilde{\eta} = \eta^{d_i}$. That means $\widetilde{\eta}$ is the optimal performance after this comparison. Because the comparisons between η^{d_i} and $\widetilde{\eta}$ are correctly made when $\widetilde{\eta}$ is the optimal, thus, until the algorithm ends, the $\widetilde{\eta}$ is always equal to the optimal performance η^{d_i}. If $\widetilde{\eta}$ has been the optimal performance, similarly, since the comparisons between η^{d_i} and $\widetilde{\eta}$ are correctly made, then $\widetilde{\eta}$ is always the optimal performance. Thus, the algorithm can output the correct optimal policy.

c. Since $\delta = \eta^* - \eta^*_-$, we have $\eta^* - \eta^{d_i} \geq \delta$ for $\eta^{d_i} \neq \eta^*, i = 1, 2, \ldots, M$. Suppose the estimations of η^* and η^{d_i} are $\hat{\eta}^*$ and $\hat{\eta}^{d_i}$, respectively, if

$$|\hat{\eta}^* - \eta^*| < \delta/2$$

and

$$|\hat{\eta}^{d_i} - \eta^{d_i}| < \delta/2,$$

then,

$$\hat{\eta}^* - \hat{\eta}^{d_i} > \eta^* - \delta/2 - (\eta^{d_i} + \delta/2) = \eta^* - \eta^{d_i} - \delta \geq 0.$$

So, the comparisons can be correctly made when η^{d_i} or $\widetilde{\eta}$ is the optimal performance. By using the result in part b), the algorithm outputs the correct optimal policy.

1.8 Derive Equation (1.12) by the Poisson equation (1.9) and derive (1.10) by (1.12).

[solution]

We consider the Poisson equation (1.9)

$$(I - P^d)g^d + \eta^d e = f^d$$

under policy d. Left-multiplying the both sides of the above Poisson equation by π^h, which is the steady-state probability of the Markov chain under policy h, we get

$$\pi^h(I - P^d)g^d + \eta^d \pi^h e = \pi^h f.$$

By $\pi^h P^h = \pi^h$, $\pi^h e = 1$ and $\eta^h = \pi^h f$, we obtain

$$\eta^h - \eta^d = \pi^h(P^h - P^d)g^d.$$

This is Equation (1.12).

Define $P_\delta^{d,h} = P^d + \delta(\Delta P) = (1-\delta)P^d + \delta P^h$, where $\Delta P = P^h - P^d$ and $0 \le \delta \le 1$. Let π_δ and η_δ be the steady-state probability and performance measure associated with $P_\delta^{d,h}$. We have $P_0^{d,h} = P^d$, and $P_1^{d,h} = P^h$. We can easily prove π_δ and η_δ are continuous with respect to δ and we have $\pi_0 = \pi^d$ and $\eta_0 = \eta^d$. By $\eta_\delta - \eta^d = \pi^h(P_\delta^{d,h} - P^d)g^d$, we have

$$\eta_\delta - \eta^d = \delta \pi_\delta(\Delta P)g^d.$$

It is equivalent to

$$\frac{\eta_\delta - \eta^d}{\delta} = \pi_\delta(\Delta P)g^d.$$

Let $\delta \to 0$, we can obtain the performance gradient at policy d along the direction ΔP is $\frac{d\eta_\delta}{d\delta}|_{\delta=0} = \lim_{\delta\to 0}\frac{\eta_\delta - \eta}{\delta} = \pi^d(\Delta P)g^d$, which is Equation (1.10).

1.9 In the MDP problem, the reward function may depend on the next state; i.e., it may take the form $f(X_l, X_{l+1}, \alpha)$, $\alpha \in \mathcal{A}(X_l)$. Prove that this problem is equivalent to the standard MDP with $f(i,\alpha)$ replaced by $\bar{f}(i,\alpha) = \sum_{j\in\mathcal{S}}[f(i,j,\alpha)p^\alpha(j|i)]$.

[solution]

A policy of MDP is a mapping from $\mathcal{S}$ to $\mathcal{A}$, i.e., $d : d(i) \to \alpha \in \mathcal{A}(i)$, for all $i \in \mathcal{S}$.

$$\begin{aligned} & \eta^d \\ = & \lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1} E\left[f(X_l, X_{l+1}, d(X_l))\right] \end{aligned}$$

$$
\begin{aligned}
&= \lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}\sum_{i\in\mathcal{S}} E\left[f(X_l, X_{l+1}, d(X_l))|X_l = i)\right] p^d(X_l = i) \\
&= \lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}} f(i,j,d(i))p^{d(i)}(j|i)p^d(X_l = i) \\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}} f(i,j,d(i))p^{d(i)}(j|i)\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1} p^d(X_l = i) \\
&= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}} f(i,j,d(i))p^{d(i)}(j|i)\pi^d(i) \\
&= \sum_{i\in\mathcal{S}}\pi^d(i)\sum_{j\in\mathcal{S}} f(i,j,d(i))p^{d(i)}(j|i) \\
&= \sum_{i\in\mathcal{S}}\pi^d(i)\bar{f}(i,d(i)).
\end{aligned}
$$

Thus, the average performance with performance function $f(i,j,\alpha)$ is equivalent to that with performance function $\bar{f}(i,\alpha)$.

1.10 Consider a Markov chain $\{X_0, X_1, \ldots\}$ defined on a finite state space $\mathcal{S}$. In any state $i \in \mathcal{S}$, an action $\alpha \in \mathcal{A}(i)$ can be taken, which determines the transition probability as $p^\alpha(j|i)$, $j \in \mathcal{S}$. Now, let us assume that the action chosen at X_l depends on both X_{l-1}, and X_l. Thus, if $X_{l-1} = k$ and $X_l = i$, the action is denoted as $\alpha = d(k,i)$ and the transition probabilities at X_l are $p^{d(k,i)}(j|i)$, $j \in \mathcal{S}$, where $d(k,i)$ is the policy.

a. Prove that this problem is equivalent to the standard MDP with an enlarged state space.

b. Can you find an equivalent standard MDP in state space $\mathcal{S}$.

[solution] a. If we make $Y_l = (X_{l-1}, X_l)$ as a state at time l, then we can easily prove the process $\mathbf{Y} = \{Y_1, Y_2, \ldots\}$ is also a Markov process. The policy d chooses an action at time l according to the state $(X_{l-1} = k, X_l = i)$. Thus, this is a standard MDP problem with state space $\mathcal{S} \times \mathcal{S}$.

b. Yes, we can find an equivalent standard MDP in state space $\mathcal{S}$. We only need to find an equivalent policy that depends only on the current state and under this policy the performance is equal to that under policy d.

We assume that the initial state $X_0 = x$ is fixed. Define a randomized policy $\mathcal{L}$ depending only on X_l by

$$\mathcal{L}_l(\alpha|i) := \mathcal{P}^d\{A_l = \alpha | X_l = i, X_0 = x\}, \alpha \in \mathcal{A}(i).$$

Next, we show the equivalence between policy $\mathcal{L}$ and policy d, that is, we should show

$$\mathcal{P}^{\mathcal{L}}\{X_l = j, A_l = \alpha | X_0 = x\} = \mathcal{P}^{d}\{X_l = j, A_l = \alpha | X_0 = x\}, l = 1, 2, \ldots. \tag{1.1}$$

By using induction, we show this result holds. Clearly it holds with $l = 1$. Assume (1.1) holds for $l = 2, 3, \ldots, l-1$. Then

$$\begin{aligned}
\mathcal{P}^{d}\{X_l = j | X_0 = x\} &= \sum_{k\in\mathcal{S}} \sum_{\alpha\in\mathcal{A}(k)} \mathcal{P}^{d}\{X_{l-1} = k, A_{l-1} = \alpha | X_0 = x\} p(j|k,\alpha) \\
&= \sum_{k\in\mathcal{S}} \sum_{\alpha\in\mathcal{A}(k)} \mathcal{P}^{\mathcal{L}}\{X_{l-1} = k, A_{l-1} = \alpha | X_0 = x\} p(j|k,\alpha) \\
&= \mathcal{P}^{\mathcal{L}}\{X_l = j | X_0 = x\}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathcal{P}^{\mathcal{L}}\{X_l = j, A_l = \alpha | X_0 = x\} &= \mathcal{P}^{\mathcal{L}}\{X_l = j | X_0 = x\} \mathcal{P}^{\mathcal{L}}\{A_l = \alpha | X_l = j, X_0 = x\} \\
&= \mathcal{P}^{d}\{X_l = j | X_0 = x\} \mathcal{P}^{d}\{A_l = \alpha | X_l = j, X_0 = x\} \\
&= \mathcal{P}^{d}\{X_l = j, A_l = \alpha | X_0 = x\}.
\end{aligned}$$

We have proved the equivalence between policies $\mathcal{L}$ and d.

If we only consider the equivalence under the long run average performance criteria, we can also show it as follows. Define a randomized stationary policy $\mathcal{L}$ depending only on $X_l = i$ by

$$\mathcal{L}(\alpha|i) := \sum_{k:d(k,i)=\alpha} \frac{\pi^d(k,i)}{\pi^d(i)}.$$

where $\pi^d(k,i)$ is the steady state probability of the Markov chain $\{Y_1, Y_2, \cdots\}, Y_l = (X_{l-1}, X_l)$, under policy d, and $\pi^d(i) = \sum_{k\in\mathcal{S}} \pi^d(k,i)$. Under the randomized policy $\mathcal{L}$, the transition probability from state i to state j is $\sum_{\alpha\in\mathcal{A}(i)} \mathcal{L}(\alpha|i) p(j|i,\alpha)$. Then, we have

$$\begin{aligned}
&\sum_{i\in\mathcal{S}} \pi^d(i) \sum_{\alpha\in\mathcal{A}(i)} \mathcal{L}(\alpha|i) p(j|i,\alpha) \\
= &\sum_{i\in\mathcal{S}} \pi^d(i) \sum_{\alpha\in\mathcal{A}(i)} \sum_{k:d(k,i)=\alpha} \frac{\pi^d(k,i)}{\pi^d(i)} p(j|i,\alpha) \\
= &\sum_{i\in\mathcal{S}} \pi^d(i) \sum_{k\in\mathcal{S}} \frac{\pi^d(k,i)}{\pi^d(i)} p(j|i,d(k,i))
\end{aligned}$$

$$\begin{aligned} &= \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S}} \pi^d(k,i)p(j|i,d(k,i)) \qquad \text{(cf. (1.5))} \\ &= \pi^d(j). \end{aligned}$$

That is to say, the steady state probability $\pi^{\mathcal{L}}(i)$ under policy $\mathcal{L}$ is equal to $\pi^d(i)$. Under the randomized policy $\mathcal{L}$, the performance function is $f^{\mathcal{L}}(i) = \sum_{\alpha\in\mathcal{A}_{(i)}} \mathcal{L}(\alpha|i)f(i,\alpha)$. Thus, we have

$$\begin{aligned} \eta^{\mathcal{L}} &= \sum_{i\in\mathcal{S}} \pi^{\mathcal{L}}(i)f^{\mathcal{L}}(i) = \sum_{i\in\mathcal{S}} \pi^d(i) \sum_{\alpha\in\mathcal{A}(i)} \sum_{k:d(k,i)=\alpha} \frac{\pi^d(k,i)}{\pi^d(i)} f(i,d(k,i)) \\ &= \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S}} \pi^d(k,i)f(i,d(k,i)) = \eta^d. \end{aligned}$$

Therefore, we have proved the equivalence between the MDP under policy d and a MDP under randomized policy $\mathcal{L}$.

1.11 Consider the optimization problem for a discrete time M/M/1 queue. When a customer arrives at the server, the number of customers in the system increases by one. The server serves one customer at a time. Other customers have to wait in a queue. When a customer finishes its service, s/he leaves the server, and the number of customers in the system decreases by 1. Let X_l be the number of customers in the server at time $l = 0, 1, \ldots$. If $X_l = n$, then the probability that a customer arrives in the lth period (i.e., $X_{l+1} = X_l + 1$) is $a(n)$, and the probability that a customer leaves (i.e., $X_{l+1} = X_l - 1$) is $b(n)$, and X_l stays the same with probability $1 - [a(n) + b(n)]$. If $X_l = 0$, then $b(0) = 0$. The system has a capacity of N; i.e., an arrival customer will be rejected if there are N customers in the system, or equivalently, $a(N) = 0$. Suppose that $a(n), n = 0, 1, 2, \ldots, N-1$, can take M different values: $a_1, a_2, \cdots, a_M \in [0, 1]$. We wish to maximize

$$\eta = \kappa_1\eta_1 - \kappa_2\eta_2,$$

where η_1 is the average number of customers accepted to the system, η_2 is the average of $w(X_l)$, with w being a function of the number of customers in the system, and $\kappa_1, \kappa_2 > 0$ are two weighting factors.

Formulate this problem as a standard MDP with random policies.

[Solution]

The state space is: $S = \{0, 1, 2, ..., N\}$.

The action space is: $\mathcal{A} = \{a, l, s\}$, where a denotes that a customer arrives, l denotes that a customer leaves and s denotes that a customer stays the same.

The policy is a randomized policy with tunable parameters, which chooses action a with probability $a(n)$ at state $n = 0, 1, \cdots, N-1$, and $a(N) = 0$; chooses action l with probability $b(n)$ in state $n = 1, 2, \cdots, N$, and $b(0) = 0$ when state is 0; and chooses action s with probability $1 - a(n) - b(n)$. In this policy, $a(n), n = 0, 1, 2, \ldots, N-1$, are the tunable parameters, which can take M different values: $a_1, a_2, \cdots, a_M \in [0, 1]$.

The transition probability matrices under actions a, l, s are as follows, respectively,

$$P(a) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ * & * & * & \cdots & * \end{bmatrix}, \quad P(l) = \begin{bmatrix} * & * & * & \cdots & * \\ 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

$$P(s) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

where $*$ denotes the transition does not occur.

The performance function is: $f(X_l, A_l) = \kappa_1 I_a(X_l, A_l) + \kappa_2 w(X_l)$, where $I_a(X_l, A_l) = 1$ for any X_l when $A_l = a$, otherwise, $I_a(X_l, A_l) = 0$.

The average performance is: $\eta = \lim\limits_{L\to\infty} \frac{1}{L} \sum\limits_{l=0}^{L-1} f(X_l, A_l) = \lim\limits_{L\to\infty} \frac{1}{L} \sum\limits_{l=0}^{L-1} \Big[\kappa_1 I_a(X_l, A_l) + \kappa_2 w(X_l)\Big]$.

The optimization problem is to choose the proper arrival rate $a(n)$ to get the maximum average performance.

1.12 For an ergodic Markov chain, we have

$$\eta = \lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} f(X_l), \qquad \text{w.p.1.}$$

Develop a "learning" algorithm which updates iteratively the estimates of η at every transition of the Markov chain using the reward observed at the transition. That is, find

an algorithm

$$\hat{\eta}_l = \kappa_l \hat{\eta}_{l-1} + (1-\kappa_l) f(X_l),$$

with $\hat{\eta}_{-1} = 0$ and $0 < \kappa_l < 1$, such that $\lim_{l\to\infty} \hat{\eta}_l = \eta$. Determine κ_l for $l = 0, 1, \ldots$.

[Solution]

Define $\hat{\eta}_l = \frac{1}{l+1} \sum_{k=0}^{l} f(X_k)$, then

$$\begin{aligned} \hat{\eta}_l &= \frac{1}{l+1} \sum_{k=0}^{l} f(X_k) = \frac{1}{l+1} \left[\sum_{k=0}^{l-1} f(X_k) + f(X_l) \right] \\ &= \frac{1}{l+1} \left[l\hat{\eta}_{l-1} + f(X_l) \right] = \frac{l}{l+1} \hat{\eta}_{l-1} + \frac{1}{l+1} f(X_l) \\ &= \frac{l}{l+1} \hat{\eta}_{l-1} + (1 - \frac{l}{l+1}) f(X_l). \end{aligned} \tag{1.2}$$

Therefore, we can let $\kappa_l = \frac{l}{l+1}$.

On the other hand, formula (1.2) can also be written as follows,

$$\hat{\eta}_l = \hat{\eta}_{l-1} + \frac{1}{l+1} \left(f(X_l) - \hat{\eta}_{l-1} \right) = \hat{\eta}_{l-1} + \mu_l \left(f(X_l) - \hat{\eta}_{l-1} \right).$$

We know that $\mu_l = 1 - \kappa_l = \frac{1}{l+1}$ which satisfy $\sum_{l=0}^{\infty} \mu_l = \infty$ and $\sum_{l=0}^{\infty} (\mu_l)^2 < \infty$. In fact it is the derivation of the stochastic approximation method. The factor $\mu_l = \frac{1}{l+1}$ is one of the most classical step size in stochastic approximation algorithm.

1.13 Consider a Markov chain under a deterministic policy $\alpha_i = d(i)$, $i \in \mathcal{S}$. Drive the equation for Q-factors:

$$Q^d(i, \alpha_i) - \sum_{j \in \mathcal{S}} p^{\alpha_i}(j|i) Q^d(j, \alpha_j) + \eta^d = f(i, \alpha_i).$$

[Solution]

From the definition of Q-factor, we have

$$Q^d(i, \alpha_i) = \sum_{j=1}^{S} p^{\alpha}(j|i) g^d(j) + f(i, \alpha_i) - \eta^d.$$

For deterministic policy, we have

$$g^d(j) = E \left\{ \sum_{l=0}^{\infty} \left[f(X_l, A_l) - \eta \right] \Big| X_0 = j \right\}$$

$$
\begin{aligned}
&= E\left\{\sum_{l=0}^{\infty}[f(X_l, A_l) - \eta]\Big| X_0 = j, A_0 = \alpha_j\right\} \\
&= \sum_{j=1}^{S} p^{\alpha_i}(j|i) g^d(j) + f(i, \alpha_i) - \eta^d \\
&= Q^d(j, \alpha_j).
\end{aligned}
$$

Thus,

$$
Q^d(i, \alpha_i) - \sum_{j \in \mathcal{S}} p^{\alpha_i}(j|i) Q^d(j, \alpha_j) + \eta^d = f(i, \alpha_i).
$$

1.14 Consider a Markov chain with state space $\mathcal{S}$. At each state $i \in \mathcal{S}$, there are two available actions denoted as $\alpha_{1,i}$ and $\alpha_{2,i}$. Let d be a randomized policy with $d(i) = \nu_i = \{p_{1,i}, p_{2,i}\}$, $p_{1,i}, p_{2,i} > 0$, $p_{1,i} + p_{2.i} = 1$, representing the probabilities of taking actions $\alpha_{1,i}$ and $\alpha_{2,i}$, respectively, $i \in \mathcal{S}$. We also can view ν_i as an action, which determines the transition probabilities of state i (see Problem 1.5). Therefore, we have three actions for each state: $\alpha_{1,i}$, $\alpha_{2,i}$, and ν_i, $i \in \mathcal{S}$. Observe a sample path of the system under randomized policy d. Overall, when the system visits state i, it takes action ν_i. This is equivalent to a system which takes action $\alpha_{1,i}$ sometimes when the system visits state i, and takes action $\alpha_{2,i}$ other times when it visits i, with probabilities $p_{1,i}$ and $p_{2,i}$, respectively. Thus, a sample path of the system under policy d contains the information about $Q^d(i, \alpha_{1,i})$, $Q^d(i, \alpha_{2,i})$, and $Q^d(i, \nu_i)$, $i \in \mathcal{S}$.

a. Prove $Q^d(i, \nu_i) = p_{1,i} Q^d(i, \alpha_{1,i}) + p_{2,i} Q^d(i, \alpha_{2,i})$.

b. If $Q^d(i, \alpha_{1,i}) \geq Q^d(i, \alpha_{2,i})$, then $Q^d(i, \alpha_{1,i}) \geq Q^d(i, \nu_i)$.

c. Prove that for every randomized policy d there is always a deterministic policy which is at least as good as d.

[Solution]

a.

$$
\begin{aligned}
Q^d(i, \nu_i) &= \sum_{j \in \mathcal{S}} \sum_{l=1}^{2} p_{l,i} p^{\alpha_{l,i}}(j|i) g^d(j) + \sum_{l=1}^{2} p_{l,i} f(i, \alpha_{l,i}) - \eta^d \\
&= \sum_{l=1}^{2} p_{li} \{\sum_{j \in \mathcal{S}} p^{\alpha_{l,i}}(j|i) g^d(j) + f(i, \alpha_{l,i}) - \eta^d\} \\
&= p_{1,i} Q^d(i, \alpha_{1,i}) + p_{2,i} Q^d(i, \alpha_{2,i}),
\end{aligned}
$$

where $p^{\alpha_{l,i}}(j|i), l = 1, 2, i, j \in \mathcal{S}$ are the transition probability under the action $\alpha_{l,i}$ and $g^d(j)$ is the potential at state j under policy d.

b. From Part a),

$$Q^d(i, \nu_i) = p_{1,i}Q^d(i, \alpha_{1,i}) + p_{2,i}Q^d(i, \alpha_{2,i}) \leq p_{1,i}Q^d(i, \alpha_{1,i}) + p_{2,i}Q^d(i, \alpha_{1,i}) = Q^d(i, \alpha_{1,i})$$

c. The transition probability from state i to state j under randomized policy d is $p^d(j|i) = \sum_{l=1}^{2} p_{l,i}p^{\alpha_{l,i}}(j|i)$. From the performance difference formula $\eta^h - \eta^d = \pi^h[(P^h - P^d)g^d + f^h - f^d]$, where h is another randomized policy with $h(i) = \{p'_{1,i}, p'_{2,i}\}$, we can obtain the following performance difference formula based on Q-factor,

$$\eta^h - \eta^d = \sum_{i \in \mathcal{S}} \pi^h(i) \sum_{l=1,2} [p'_{l,i} - p_{l,i}]Q^d(i, \alpha_{l,i}). \tag{1.3}$$

From part b), we have $\max_{\alpha_{l,i}, l=1,2} Q^d(i, \alpha_{l,i}) \geq Q^d(i, \nu_i)$. If policy h chooses action

$$\alpha^*_{l,i} = arg \max_{\alpha_{l,i}} Q^d(i, \alpha_{l,i}), \tag{1.4}$$

at state i with probability 1, we have

$$\sum_{l=1,2} [p'_{l,i} - p_{l,i}]Q^d(i, \alpha_{l,i}) = Q^d(i, \alpha^*_{l,i}) - Q^d(i, \nu_i) \geq 0.$$

Thus, from difference formula (1.3), we have $\eta^h \geq \eta^d$. That is to say, for every randomized policy d, there is always a deterministic policy h choosing actions as (1.4), which is at least as good as d.

1.15 Consider a linear control system defined as

$$X_{l+1} = X_l + u_l + \xi_l, \qquad l = 0, 1, \ldots.$$

The state space is the set of integers $\mathcal{S} := \{\ldots, -1, 0, 1, \ldots\}$, the control variable u can take two values -2 and 2, the random noise ξ takes values from the integer set $\{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ with probabilities $p(\xi = 0) = 0.2$ and $p(\xi = i) = 0.1$ if $i \neq 0$. Describe the system in the MDP formulation.

[Solution]

The state space of the system is $\mathcal{S} = \{0, 1, -1, 2, -2, 3, -3, 4, -4, \ldots\}$. The action space is $\mathcal{A} = \{-2, 2\}$. If the current state is 0, then the probability that the system

transits to state 0 under the action $a = -2$ is $\mathcal{P}\{\xi = 2\} = 0.1$. In the same way, we can obtain the transition probability matrix as follows:

$$P(a=-2) = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.2 & 0 & 0.1 & 0 & 0.1 & 0 & 0.1 & 0 & 0.1 & \cdots \\ 0.1 & 0.1 & 0.2 & 0.1 & 0.1 & 0.1 & 0.1 & 0 & 0.1 & 0 & 0.1 & 0 & 0 & \cdots \\ 0.1 & 0.1 & 0.1 & 0 & 0.1 & 0 & 0.2 & 0 & 0.1 & 0 & 0.1 & 0 & 0.1 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$P(a=2) = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.2 & 0.1 & 0.1 & 0 & 0.1 & 0 & 0.1 & 0 & 0.1 & 0 & \cdots \\ 0.1 & 0.1 & 0.1 & 0.1 & 0 & 0.2 & 0 & 0.1 & 0 & 0.1 & 0 & 0.1 & 0 & \cdots \\ 0.1 & 0.2 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0 & 0.1 & 0 & 0 & 0 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

1.16 Consider an admission control problem of a communication system consisting of three servers. The system is Jackson type and hence its state can be denoted as $\boldsymbol{n} = (n_1, n_2, n_3)$ with n_i being the number of customers in server i, $i = 1, 2, 3$. Define an event a_{+4} as: a customer arrives at the network and finds that there are 4 customers already in the network. Clearly define this event by a set of state transitions. (Denote the transition from state $\boldsymbol{n}$ to $\boldsymbol{n}'$ as $\langle \boldsymbol{n}, \boldsymbol{n}' \rangle$.)

[Solution]

Denote $a_{+4,i}$ as the event that a customer arrives at the network and finds that there are 4 customers already in the network and this customer will be accepted to server i, $i = 1, 2, 3$.

Denote $a_{+4,0}$ as the event that a customer arrives at the network and finds that there are 4 customers already in the network and this customer will be rejected.

We know that $a_{+4} = \cup_{i=0}^{3} a_{+4,i}$.

$$\begin{aligned} & a_{+4,0} \\ = \; & \{\langle (0,0,4),(0,0,4)\rangle, \langle (0,4,0),(0,4,0)\rangle, \langle (4,0,0),(4,0,0)\rangle, \langle (1,1,2),(1,1,2)\rangle, \\ & \langle (1,2,1),(1,2,1)\rangle, \langle (2,1,1),(2,1,1)\rangle, , \langle (3,1,0),(3,1,0)\rangle, \langle (3,0,1),(3,0,1)\rangle, \\ & \langle (1,0,3),(1,0,3)\rangle, \langle (1,3,0),(1,3,0)\rangle, \langle (0,1,3),(0,1,3)\rangle, \langle (0,3,1),(0,3,1)\rangle, \\ & \langle (2,2,0),(2,2,0)\rangle, \langle (2,0,2),(2,0,2)\rangle, \langle (0,2,2),(0,2,2)\rangle\} \end{aligned}$$

$$
\begin{aligned}
& a_{+4,1} \\
= \; & \{\langle(0,0,4),(1,0,4)\rangle, \langle(0,4,0),(1,4,0)\rangle, \langle(4,0,0),(5,0,0)\rangle, \langle(1,1,2),(2,1,2)\rangle, \\
& \langle(1,2,1),(2,2,1)\rangle, \langle(2,1,1),(3,1,1)\rangle, , \langle(3,1,0),(4,1,0)\rangle, \langle(3,0,1),(4,0,1)\rangle, \\
& \langle(1,0,3),(2,0,3)\rangle, \langle(1,3,0),(2,3,0)\rangle, \langle(0,1,3),(1,1,3)\rangle, \langle(0,3,1),(1,3,1)\rangle, \\
& \langle(2,2,0),(3,2,0)\rangle, \langle(2,0,2),(3,0,2)\rangle, \langle(0,2,2),(1,2,2)\rangle\}
\end{aligned}
$$

$$
\begin{aligned}
& a_{+4,2} \\
= \; & \{\langle(0,0,4),(0,1,4)\rangle, \langle(0,4,0),(0,5,0)\rangle, \langle(4,0,0),(4,1,0)\rangle, \langle(1,1,2),(1,2,2)\rangle, \\
& \langle(1,2,1),(1,3,1)\rangle, \langle(2,1,1),(2,2,1)\rangle, , \langle(3,1,0),(3,2,0)\rangle, \langle(3,0,1),(3,1,1)\rangle, \\
& \langle(1,0,3),(1,1,3)\rangle, \langle(1,3,0),(1,4,0)\rangle, \langle(0,1,3),(0,2,3)\rangle, \langle(0,3,1),(0,4,1)\rangle, \\
& \langle(2,2,0),(2,3,0)\rangle, \langle(2,0,2),(2,1,2)\rangle, \langle(0,2,2),(0,3,2)\rangle\}
\end{aligned}
$$

$$
\begin{aligned}
& a_{+4,3} \\
= \; & \{\langle(0,0,4),(0,0,5)\rangle, \langle(0,4,0),(0,4,1)\rangle, \langle(4,0,0),(4,0,1)\rangle, \langle(1,1,2),(1,1,3)\rangle, \\
& \langle(1,2,1),(1,2,2)\rangle, \langle(2,1,1),(2,1,2)\rangle, , \langle(3,1,0),(3,1,1)\rangle, \langle(3,0,1),(3,0,2)\rangle, \\
& \langle(1,0,3),(1,0,4)\rangle, \langle(1,3,0),(1,3,1)\rangle, \langle(0,1,3),(0,1,4)\rangle, \langle(0,3,1),(0,3,2)\rangle, \\
& \langle(2,2,0),(2,2,1)\rangle, \langle(2,0,2),(2,0,3)\rangle, \langle(0,2,2),(0,2,3)\rangle\}
\end{aligned}
$$

2

Solutions to Chapter 2

2.1 In Figure 2.2 , the three points P_0, P_1, and P_2 represent three policies. Every point P in the triangle with the three points as vertices represents a randomized policy denoted as $P(\delta_1, \delta_1, \delta_2) = \delta_0 P_0 + \delta_1 P_1 + \delta_2 P_2$, $\delta_0 + \delta_1 + \delta_2 = 1$, with $P_0 = P(1, 0, 0)$, $P_1 = P(0, 1, 0)$, and $P_2 = P(0, 0, 1)$.

a. Determine the values of δ_0, δ_1, and δ_2 by the lengths of the segments shown in the figure.

b. Along the line from P_0 to P_1, we have randomized policies $P_\delta = (1-\delta)P_0 + \delta P_1, 0 < \delta < 1$, and we can obtain the directional derivative in this direction, denoted as $\frac{d\eta_\delta}{d\delta}|_{P_0 - P_1}$. Similarly, we can obtain the directional derivative in the direction from P_0 to P_2, denoted as $\frac{d\eta_\delta}{d\delta}|_{P_0 - P_2}$. What is the directional derivative from P_0 to P? Express it in terms of $\frac{d\eta_\delta}{d\delta}|_{P_0 - P_1}$ and $\frac{d\eta_\delta}{d\delta}|_{P_0 - P_2}$. (*Hint: Along this direction* δ_1/δ_2 *is fixed.*)

Solution:

a. Firstly, since P' is a point in the line segment P_1P_2, we have

$$P' = \frac{|P_2P'|}{|P_1P_2|}P_1 + \frac{|P_1P'|}{|P_1P_2|}P_2, \tag{2.1}$$

where $|\cdot|$ denotes the length of a line segment. Similarly, since P is in the line segment P_0P', we have

$$P = \frac{|P'P|}{|P_0P'|}P_0 + \frac{|P_0P|}{|P_0P'|}P'. \tag{2.2}$$

Putting (2.1) into (2.2), we have

$$\begin{aligned} P &= \frac{|P'P|}{|P_0P'|}P_0 + \frac{|P_0P|}{|P_0P'|}(\frac{|P_2P'|}{|P_1P_2|}P_1 + \frac{|P_1P'|}{|P_1P_2|}P_2) \\ &= \frac{|P'P|}{|P_0P'|}P_0 + \frac{|P_0P|}{|P_0P'|}\frac{|P_2P'|}{|P_1P_2|}P_1 + \frac{|P_0P|}{|P_0P'|}\frac{|P_1P'|}{|P_1P_2|}P_2. \end{aligned}$$

Thus $\delta_0 = \frac{|P'P|}{|P_0P'|}$, $\delta_1 = \frac{|P_0P|}{|P_0P'|}\frac{|P_2P'|}{|P_1P_2|}$ and $\delta_2 = \frac{|P_0P|}{|P_0P'|}\frac{|P_1P'|}{|P_1P_2|}$.

b. Since

$$\begin{aligned} P_\delta &= P_0 + \delta(P - P_0) \\ &= P_0 + \delta(\delta_0P_0 + \delta_1P_1 + \delta_2P_2 - (\delta_0 + \delta_1 + \delta_2)P_0) \\ &= P_0 + \delta(\delta_1(P_1 - P_0) + \delta_2(P_2 - P_0)) \\ &= P_0 + \delta\Delta P, \end{aligned}$$

where $\Delta P = \delta_1(P_1 - P_0) + \delta_2(P_2 - P_0)$, the directional derivative from P_0 to P is

$$\begin{aligned} \frac{d\eta_\delta}{d\delta}\Big|_{P_0-P} &= \pi\Delta P g \\ &= \pi[\delta_1(P_1 - P_0) + \delta_2(P_2 - P_0)]g \\ &= \delta_1\frac{d\eta_\delta}{d\delta}\Big|_{P_0-P_1} + \delta_2\frac{d\eta_\delta}{d\delta}\Big|_{P_0-P_2}. \end{aligned}$$

2.2 (Random walk) A random walker moves among five positions $i = 1, 2, 3, 4, 5$. At position $i = 2, 3, 4$, s/he moves to positions $i - 1$ and $i + 1$ with an equal probability $p(i-1|i) = p(i+1|i) = 0.5$; at the boundary positions $i = 1$ and $i = 5$, s/he bounces back with probability one $p(4|5) = p(2|1) = 1$. We are given a sequence of 20 $[0, 1)$-uniformly and independently distributed random variables as follows.

0.740, 0.605, 0.234, 0.342, 0.629, 0.965, 0.364, 0.230, 0.599, 0.079,
0.782, 0.219, 0.475, 0.051, 0.596, 0.850, 0.865, 0.434, 0.617, 0.969.

a. With this sequence, construct a sample path $\boldsymbol{X}$ of the random walk from X_0 to X_{20} according to (2.3). Set $X_0 = 3$.

b. Suppose that the perturbed transition probabilities are $p'(i-1|i) = 0.3$, $p'(i+1|i) = 0.7$, $i = 2, 3, 4$, and $p'(4|5) = p'(2|1) = 1$. Set $p_\delta(j|i) = p(j|i) + \delta[p'(j|i) - p(j|i)]$. By using the original sample path obtained in (a), construct a perturbed sample path $\boldsymbol{X}_\delta$, $\delta = 1$, following Figure 2.5. Use the following independently distributed $[0, 1)$ random variable when $\boldsymbol{X}_\delta$ is different than $\boldsymbol{X}$ (use the lth number to determine the lth transition of X_δ, if $X_{\delta,l} \neq X_l$):

0.173, 0.086, 0.393, 0.804, 0.011, 0.233, 0.934, 0.230, 0.786, 0.410,
0.119, 0.634, 0.862, 0.418, 0.601, 0.118, 0.626, 0.835, 0.361, 0.336.

c. Repeat b) for $\delta = 0.7, 0.5, 0.3, 0.2, 0.1$.

d. Observe the trend of the perturbed paths $\boldsymbol{X}_\delta$. In particular, when δ is small, most likely the perturbed parts from the jumping point to the merging point are the same as if they follow the original transition probabilities $p(j|i)$, $i, j = 1, 2, \cdots, \mathcal{S}$.

Solution:

a. We assume that the initial position is $X_0 = 3$. According to (2.2), since $\sum_{k=1}^{3} p(k|3) \leq 0.740 < \sum_{k=1}^{4} p(k|3)$, the next state is 4. Similarly, the other subsequent states can be generated by using (2.2) and the sample path is

$$3, 4, 5, 4, 3, 4, 5, 4, 3, 4, 3, 4, 3, 2, 1, 2, 3, 4, 3, 4, 5.$$

The sample path is described in Figure 2.1.

b. We assume that the perturbed parts from the jumping point are generated by using corresponding random number in the given sequence of 20 $[0, 1)$-uniformly and independently distributed random variables. That is, if there is a jump at i-th time, then next state of perturbed sample path is generated by using i-th $[0, 1)$-uniformly random

number in the given sequence. According to (2.2), the perturbed sample path when $\delta = 1$ is

$$3, 4, 5, 4, \bar{\mathbf{5}}, 4, 5, 4, 3, 4, 3, 4, 3, \bar{\mathbf{4}}, \bar{\mathbf{5}}, \bar{\mathbf{4}}, 3, 4, \bar{\mathbf{5}}, 4, 5.$$

The sample path is described in Figure 2.1.

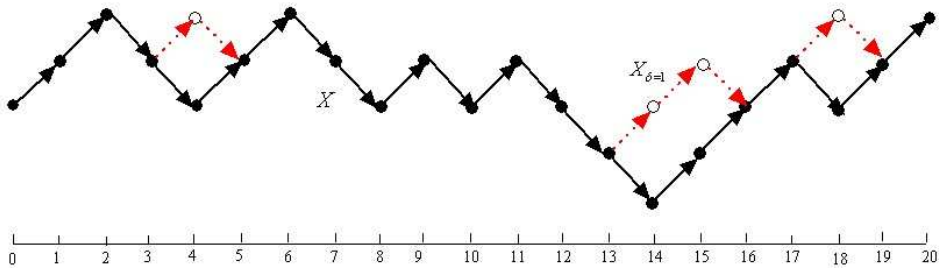

Figure 2.1: The original sample path and perturbed sample path with $\delta = 1$

c. The perturbed sample path when $\delta = 0.7$ is

$$3, 4, 5, 4, 3, 4, 5, 4, 3, 4, 3, 4, 3, \bar{\mathbf{4}}, \bar{\mathbf{5}}, \bar{\mathbf{4}}, 3, 4, \bar{\mathbf{5}}, 4, 5.$$

The sample path is described in Figure 2.3. The perturbed sample path when $\delta = 0.5$ is the same as the one when $\delta = 0.7$. The perturbed sample path when $\delta = 0.3$ is

$$3, 4, 5, 4, 3, 4, 5, 4, 3, 4, 3, 4, 3, \bar{\mathbf{4}}, \bar{\mathbf{3}}, \bar{\mathbf{4}}, 3, 4, 3, 4, 5,$$

which is described in Figure 2.3. When $\delta = 0.2$, the perturbed sample path is the same as the one when $\delta = 0.3$. The perturbed sample path when $\delta = 0.1$ is the same as the original one.

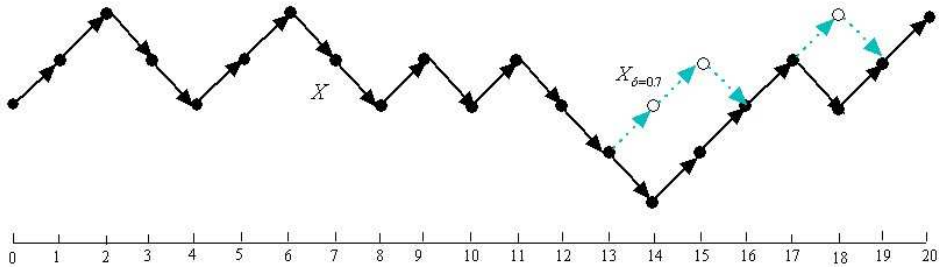

Figure 2.2: The original sample path and perturbed sample path with $\delta = 0.7$ and $\delta = 0.5$

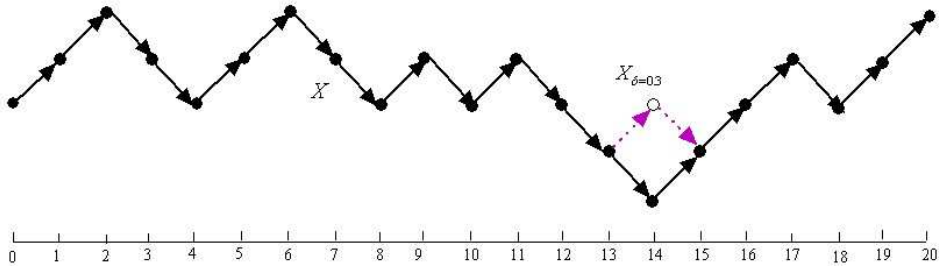

Figure 2.3: The original sample path and perturbed sample path with $\delta = 0.3$

d. We may observe the trend of the perturbed paths $\boldsymbol{X}_\delta$. When δ is small, there are fewer perturbations on the perturbed sample paths. Moreover, most likely the perturbed parts from the jumping point to the merging point are the same as if they follow the original transition probabilities $p(j|i)$, $i, j = 1, 2, \cdots, S$.

2.3 Let $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ be two independent ergodic Markov chain with the same transition probability matrix P on the same state space $\mathcal{S}$. Define $\mathbf{Y} = (\boldsymbol{X}, \widetilde{\boldsymbol{X}})$.

a. Prove that $\mathbf{Y}$ is ergodic.

b. Express L^*_{ij} in Figure 2.6 in terms of the Markov chain $\mathbf{Y}$.

Solution:

a. Proving that $\mathbf{Y}$ is ergodic means proving that $\mathbf{Y}$ is irreducible and aperiodic under the condition that $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are ergodic. Firstly, we prove that $\mathbf{Y}$ is irreducible. Since $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ are ergodic, we know that for any states $i, j \in \mathcal{S}$, there is a $N > 0$ such that when $n > N$, $p^n(j|i) > 0$, where $p^n(j|i)$ denotes the probability that Markov chain moves from state i to j at n-th step. Thus, for any states (i, j) and (k, l) of Markov chain $\mathbf{Y}$, if $m > N$, then $p^m((k, l)|(i, j)) = p^m(k|i)p^m(l|j) > 0$, where we have used the independence of $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$. That is, $\mathbf{Y}$ is irreducible. Moreover, since for any $m > N$, $p^m((i, j)|(i, j)) > 0$, the great common divisor of $\{k|p^k((i, j)|(i, j)) > 0\}$ is 1. Thus, $\mathbf{Y}$ is aperiodic. Irreducible and aperiodic Markov chain $\mathbf{Y}$ is ergodic.

b. Define $M = \{(i,i), i \in \mathcal{S}\}$, then, $L^*_{ij} = \min\{l > 0, Y_l \in M | Y_0 = (i,j)\}$, which is the first hitting time of $\mathbf{Y}$ to reach set M from state (i,j).

2.4 Consider a three-state Markov chain with

$$P = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.1 & 0.6 & 0.3 \\ 0.7 & 0.1 & 0.2 \end{bmatrix}, \qquad f = \begin{bmatrix} 10 \\ 5 \\ 8 \end{bmatrix}.$$

a. Solve the Poisson equation (2.12) $(I-P)g + \eta e = f$ for g and η (by e.g. setting $g(0) = 0$).

b. Solve $\pi = \pi P$ and $\pi e = 1$ for π first, then solve $(I - P + e\pi)g = f$ for g.

c. Compare both methods in a) and b).

Solution:

a. Since the solution g to the Poisson equation is unique up to a constant, we can first let $g(0)$ equal to a constant and solve the Poisson equation to obtain a special solution. The general solution g is equal to $g + ce$. For example, setting $g(0) = 0$, we have

$$\begin{bmatrix} 1 & -0.5 & -0.5 \\ -0.1 & 0.4 & -0.3 \\ -0.7 & -0.1 & 0.8 \end{bmatrix} \begin{bmatrix} 0 \\ g(1) \\ g(2) \end{bmatrix} + \begin{bmatrix} \eta \\ \eta \\ \eta \end{bmatrix} = \begin{bmatrix} 10 \\ 5 \\ 8 \end{bmatrix}.$$

Arranging the equation, we have

$$\begin{bmatrix} -0.5 & -0.5 & 1 \\ 0.4 & -0.3 & 1 \\ -0.1 & 0.8 & 1 \end{bmatrix} \begin{bmatrix} g(1) \\ g(2) \\ \eta \end{bmatrix} = \begin{bmatrix} 10 \\ 5 \\ 8 \end{bmatrix}.$$

Solving this equation, we can obtain $g(1) = -5.5963, g(2) = 0.1835, \eta = 7.2936$. Thus, the general solution is $g(0) = c, g(1) = -5.5963 + c, g(2) = 0.1835 + c$ and $\eta = 7.2936$.

b. Solving the balance equation $\pi = \pi P$ and $\pi e = 1$, we obtain $\pi = [0.2661, 0.4128, 0.3211]$. Putting π into $(I - P + e\pi)g = f$ and calculating the inverse of $I - P + e\pi$, we may obtain $g = (I - P + e\pi)^{-1} f = [9.5451, 3.9487, 9.7286]^T$.

c. In a), we do not need to compute the steady-state probability π. However, in b), the steady-state probability should be computed firstly.

2.5 For an ergodic Markov chain $\boldsymbol{X} = \{X_l, l = 0, 1, \ldots\}$, derive the Poisson equation using

$$g(i) = \lim_{L\to\infty} \sum_{l=0}^{L-1} E\left\{[f(X_l) - \eta]\,|X_0 = i\right\}.$$

Solution:

$$\begin{aligned}
g(i) &= \lim_{L\to\infty} \sum_{l=0}^{L-1} \mathrm{E}\left\{[f(X_l) - \eta]|X_0 = i\right\} \\
&= \lim_{L\to\infty} \left\{ f(i) - \eta + \sum_{l=1}^{L-1} \mathrm{E}\{[f(X_l) - \eta]|X_0 = i\} \right\} \\
&= f(i) - \eta + \lim_{L\to\infty} \sum_{l=1}^{L-1} \mathrm{E}[f(X_l) - \eta|X_0 = i] \\
&= f(i) - \eta + \sum_{j\in\mathcal{S}} p(j|i) \lim_{L\to\infty} \sum_{l=1}^{L-1} \mathrm{E}[f(X_l) - \eta|X_1 = j] \\
&= f(i) - \eta + \sum_{j} p(j|i) g(j).
\end{aligned}$$

Rewriting it as a matrix form, we have

$$g = f - \eta e + Pg. \qquad \Longrightarrow \qquad (I - P)g + \eta e = f.$$

Then we have obtained the Poisson equation.

2.6 The Poisson equation for the perturbed Markov chain is

$$(I - P_\delta)g_\delta + \eta_\delta e = f_\delta,$$

where $P_\delta = P + \delta\Delta P$ and $f_\delta = f + \delta\Delta f$. Derive the performance derivative formula (2.26) from the above equation.

Solution: Taking derivative of both sides of the Poisson equation for the perturbed Markov chain, we have

$$-\Delta P g_\delta + (I - P_\delta)\frac{dg_\delta}{d\delta} + \frac{d\eta_\delta}{d\delta} e = \Delta f.$$

Right multiplying this equation by π_δ, we have

$$\frac{d\eta_\delta}{d\delta} = \pi_\delta(\Delta P g_\delta + \Delta f).$$

According to the continuity of π_δ and g_δ with respect to δ, the derivative at $\delta = 0$ is

$$\frac{d\eta_\delta}{d\delta}\Big|_{\delta=0} = \pi(\Delta P g + \Delta f),$$

which is the performance derivative formula (2.26).

2.7 Prove the following results:

a. If $f = ce$ with c being a constant, then $g = ce$ is a constant vector.

b. If $p(j|i) = p_j$ for all $i \in \mathcal{S}$; i.e., every row in the transition probability matrix is the same, then $g = f$.

c. If $p(j|i) = p(i|j)$, for all $i, j \in \mathcal{S}$; i.e., the transition probability matrix P is symmetric, then $\sum_{i=1}^{S} g(i) = \sum_{i=1}^{S} f(i)$.

Solution:

a). If $f = ce$, we have

$$\eta = \pi f = \pi ce = c.$$

g can be written as

$$g(i) = \lim_{L\to\infty} \mathrm{E}\{\sum_{l=0}^{L-1}[f(X_l) - \eta]|X_0 = i\} = \lim_{L\to\infty} \mathrm{E}\{\sum_{l=0}^{L-1}[c - c]|X_0 = i\} = 0$$

Since potential g can be added by any constant vector, we have $g = ce$.

b). Because $p(j|i) = p_j$, from $\pi = \pi P$ we have

$$\pi_j = \sum_i \pi_i p(j|i) = \sum_i \pi_i p_j = p_j.$$

Thus we have $P = e\pi$. From Poisson equation

$$f = (I - P + e\pi)g = (I - e\pi + e\pi)g = g,$$

then $g = f$ is proved.

c). Because $P^{\mathrm{T}} = P$, we have $(Pe)^{\mathrm{T}} = e^{\mathrm{T}} \Longrightarrow e^{\mathrm{T}}P = e^{\mathrm{T}}$. And we have $\pi P = \pi$ and this equation has unique solution with $\pi e = 1$. Compare these two equations, we have $\pi = \frac{1}{S}e^{\mathrm{T}}$. For a special potential g such that $\pi g = \pi f$, we have $\frac{1}{S}e^{\mathrm{T}}g = \frac{1}{S}e^{\mathrm{T}}f$, thus $\sum_{i=1}^{S} g(i) = \sum_{i=1}^{S} f(i)$.

2.8 Prove $e\frac{d\eta_\delta}{d\delta} = \lim_{\beta\uparrow 1}\frac{d\eta_{\beta,\delta}}{d\delta}$, in other words,

$$\frac{d}{d\delta}[\lim_{\beta\uparrow 1}\eta_{\beta,\delta}] = \lim_{\beta\uparrow 1}\frac{d\eta_{\beta,\delta}}{d\delta}.$$

Solution: From equation (2.44)

$$\frac{\mathrm{d}\eta_{\beta,\delta}}{\mathrm{d}\delta} = (1-\beta)(I-\beta P)^{-1}[\beta\Delta P g_\beta + \Delta f],$$

and from (2.38) and (2.39), i.e.,

$$\lim_{\beta\uparrow 1} g_\beta = g,$$

$$\lim_{\beta\uparrow 1}(1-\beta)(I-\beta P)^{-1} = e\pi,$$

then, we have

$$\lim_{\beta\uparrow 1}\frac{\mathrm{d}\eta_{\beta,\delta}}{\mathrm{d}\delta} = e\pi[\Delta P g + \Delta f] = e\frac{\mathrm{d}\eta_\delta}{\mathrm{d}\delta}.$$

2.9 Assume that P changes to $P_\delta = P + \delta(\Delta P)$, $\Delta P e = 0$, and $f_\delta \equiv f$. Derive the second-order derivative of the discounted performance $\eta_{\beta,\delta}$ with respect to δ, $\frac{d^2\eta_{\beta,\delta}}{d\delta^2}$.

Solution:

From equation (2.31),

$$(I-\beta P_\delta)\eta_{\beta,\delta} = (1-\beta)f. \tag{2.3}$$

Taking derivative of both sides of (2.3), we have

$$-\beta\Delta P\eta_{\beta,\delta} + (I-\beta P_\delta)\frac{d\eta_{\beta,\delta}}{d\delta} = 0. \tag{2.4}$$

Thus, we have $\frac{d\eta_{\beta,\delta}}{d\delta} = (I-\beta P_\delta)^{-1}\beta\Delta P\eta_{\beta,\delta}$. By using (2.37) and (2.31), we have $\eta_{\beta,\delta} = (1-\beta)g_{\beta,\delta} + \beta e\eta_\delta$. Thus,

$$\frac{d\eta_{\beta,\delta}}{d\delta} = (1-\beta)(I-\beta P_\delta)^{-1}\beta\Delta P g_{\beta,\delta}. \tag{2.5}$$

Taking derivative of both sides of (2.4), we have

$$-2\beta\Delta P\frac{d\eta_{\beta,\delta}}{d\delta}+(I-\beta P_\delta)\frac{d^2\eta_{\beta,\delta}}{d\delta^2}=0. \tag{2.6}$$

Thus, the second order derivative $\frac{d^2\eta_{\beta,\delta}}{d\delta^2}=2(I-\beta P_\delta)^{-1}\beta\Delta P\frac{d\eta_{\beta,\delta}}{d\delta}$. Putting (2.5) into it, we have

$$\frac{d^2\eta_{\beta,\delta}}{d\delta^2}=2(1-\beta)[(I-\beta P_\delta)^{-1}\beta\Delta P]^2 g_{\beta,\delta}.$$

and the derivative at $\delta=0$ is

$$\frac{d^2\eta_{\beta,\delta}}{d\delta^2}\Big|_{\delta=0}=2(1-\beta)[(I-\beta P)^{-1}\beta\Delta P]^2 g_\beta.$$

2.10 In Example 2.2 , we have

$$G_1:=\Delta P(I-P+e\pi)^{-1}=\begin{bmatrix}-3.2 & 3.2\\ 3.2 & -3.2\end{bmatrix}.$$

a. Find the eigenvalues and eigenvectors of G_1.

b. Verify that

$$\begin{bmatrix}-3.2 & 3.2\\ 3.2 & -3.2\end{bmatrix}=\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & -6.4\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}^{-1}.$$

c. Prove

$$G_1^n=\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & (-6.4)^n\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}^{-1},$$

and

$$\pi_\delta=\pi\sum_{n=0}^{\infty}G_\delta^n=\pi\sum_{n=0}^{\infty}(\delta G_1)^n=\pi\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & \sum_{n=0}^{\infty}(-6.4\delta)^n\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}^{-1}.$$

d. Determine the convergence region of π_δ. Extend the discussion to more general case.

Solution:

a. The eigenvalues of G are -6.4 and 0. The corresponding eigenvectors are $[1,-1]^T$ and $[1,1]^T$.

b. The inverse of matrix $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ is $\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}$. Then, computing the left side by using matrix multiplication, it can be found that the both sides are equal.

c. According to b),

$$\begin{aligned} G_1^n &= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -6.4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \cdots \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -6.4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & (-6.4)^n \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1}. \end{aligned}$$

By using (2.54) and the above result, we can obtain

$$\pi_\delta = \pi \sum_{n=0}^{\infty} G_\delta^n = \pi \sum_{n=0}^{\infty} (\delta G_1)^n = \pi \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \sum_{n=0}^{\infty}(-6.4\delta)^n \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1}.$$

d. We can find if $0 \le \delta < \frac{1}{6.4}$, then $6.4\delta < 1$, so the series $\sum_{n=0}^{\infty}(-6.4\delta)^n$ converges. For more general case, the convergence domain of π_δ is $0 \le \delta < r := \frac{1}{\rho[\Delta P(I-P+e\pi)^{-1}]}$.

2.11 A group is a nonempty set G, together with a binary operation on G, denoted as juxtaposition ab, $a, b \in G$, and $ab \in G$, with the following properties: (i) *(Associativity)* $(ab)c = a(bc)$, for all $a, b, c \in G$; (ii) *(Identity)* There exists an element $e \in G$ for which $ea = ae = a$ for all $a \in G$; and (iii) *(Inverse)* For each $a \in G$, there is an element denoted a^{-1}, for which $aa^{-1} = a^{-1}a = e$, [220].

a. Verify that the set of matrices defined in (2.50) with matrix multiplication satisfies the above properties.

b. In Example 2.2, we have

$$B = P - I = \begin{bmatrix} -0.10 & 0.10 \\ 0.15 & -0.15 \end{bmatrix},$$

what is its group inverse? Is the inverse an infinitesimal generator?

Solution:

a. For any $B_1, B_2 \in \mathcal{B}$, we have $\pi B_1 B_2 = (\pi B_1)B_2 = 0$ and $B_1 B_2 e = B_1(B_2 e) = 0$, so $B_1 B_2 \in \mathcal{B}$. Since the binary operation on $\mathcal{B}$ is matrix multiplication, the *Associativity*

holds. We can easily verify that $I-e\pi \in \mathcal{B}$. For element $I-e\pi \in \mathcal{B}$, we have $B(I-e\pi) = B$ by using $Be = 0$ and $(I-e\pi)B = B$ by using $\pi B = 0$, Thus, the *Identity* holds. By using (2.49) and $B^{\#}e = 0$ and $\pi B^{\#} = 0$, we know the *Inverse* holds.

b. By using $\pi e = 1$ and $\pi B = 0$, we firstly obtain $\pi = [0.6, 0.4]$. Putting π into (2.48), the group inverse of B is

$$B^{\#} = \begin{bmatrix} -1.6000 & 1.6000 \\ 2.4000 & -2.4000 \end{bmatrix}.$$

$B^{\#}$ is also an infinitesimal generator.

2.12 Assume that the Maclaurin series of P_δ exists in $[0, \delta]$. Equation (2.57) can be derived directly by the following procedure: Taking the derivatives of the both sides of $\pi_\delta[I - P_\delta] = 0$ n times, we can obtain $\frac{d^n\pi}{d\delta^n}$ at $\delta = 0$. Then we can construct the Maclaurin series of π. Work out the details of this approach and derive the Maclaurin series of η_δ at $\delta = 0$.

Solution: Taking derivative of both sides of $\pi_\delta[I - P_\delta] = 0$, we have

$$\frac{d\pi_\delta}{d\delta}[I - P_\delta] = \pi_\delta \frac{dP_\delta}{d\delta}. \tag{2.7}$$

Multiplying the both sides of the above equation on the right with $-B^{\#}$, which is the group inverse of $I - P_\delta$, and noting that $(I - P_\delta)(-B^{\#}) = I - e\pi$ and $\pi e = 1$, we get

$$\frac{d\pi_\delta}{d\delta} = \pi_\delta \frac{dP_\delta}{d\delta}(-B^{\#}). \tag{2.8}$$

Thus $\frac{d\eta_\delta}{d\delta}|_{\delta=0} = \pi\frac{dP}{d\delta}(-B^{\#})f$. Continuing taking derivative of both sides of (2.7), we obtain,

$$\frac{d^2\pi_\delta}{d\delta^2}(I - P_\delta) = 2\frac{d\pi_\delta}{d\delta}\frac{dP_\delta}{d\delta} + \pi_\delta \frac{d^2P_\delta}{d\delta^2}. \tag{2.9}$$

Similarly, multiplying the both sides of the above equation on the right with $-B^{\#}$ and putting (2.8) into it, we get

$$\frac{d^2\pi_\delta}{d\delta^2} = 2\pi_\delta(\frac{dP_\delta}{d\delta}(-B^{\#}))^2 + \pi_\delta \frac{d^2P_\delta}{d\delta^2}(-B^{\#}).$$

Thus $\frac{d^2\eta_\delta}{d\delta^2}|_{\delta=0} = 2\pi(\frac{dP}{d\delta}(-B^{\#}))^2 f + \pi\frac{d^2P}{d\delta^2}(-B^{\#})f$. We can continue the computation of $\frac{d^n\eta_\delta}{d\delta^n}, n \geq 3$. Putting these derivatives into the Maclaurin expansion of η_δ at $\delta = 0$, we

have

$$\begin{aligned}\eta_\delta &= \eta + \sum_{n=1}^{\infty} \frac{1}{n!} \frac{d^n \eta}{d\delta^n}|_{\delta=0} \delta^n \\ &= \pi f + \pi \frac{dP}{d\delta}(-B^{\#})\delta f + \pi(\frac{dP}{d\delta}(-B^{\#}))^2 \delta^2 f + \frac{1}{2}\pi \frac{d^2P}{d\delta^2}(-B^{\#})\delta^2 f + \cdots \\ &= \pi\Big\{I + \frac{dP}{d\delta}(-B^{\#})\delta + \Big[(\frac{dP}{d\delta}(-B^{\#}))^2 + \frac{1}{2}\frac{d^2P}{d\delta^2}(-B^{\#})\Big]\delta^2 + \cdots\Big\} f.\end{aligned}$$

2.13 Prove the continuous version of the PRF equation (2.62) from its discrete version (2.7) by setting $B = P - I$, and vice versa.

Solution:

The continuous version of Lyapunov equation is $B\Gamma + \Gamma B^{\mathrm{T}} = -F$ and discrete version is $\Gamma - P\Gamma P^{\mathrm{T}} = F$. Because $B = P - I$, we have $Be = 0$.

From discrete to continuous: Replacing P with $B + I$, we get

$$\begin{aligned}&\Gamma - (B+I)\Gamma(B+I)^{\mathrm{T}} = F \\ &\Gamma - B\Gamma B^{\mathrm{T}} - B\Gamma - \Gamma B^{\mathrm{T}} - \Gamma = F \\ &B\Gamma + \Gamma B^{\mathrm{T}} = -F - B\Gamma B^{\mathrm{T}}\end{aligned}$$

Since $B\Gamma B^{\mathrm{T}} = B(eg^{\mathrm{T}} - ge^{\mathrm{T}})B^{\mathrm{T}} = Beg^{\mathrm{T}}B^{\mathrm{T}} - Bg(Be)^{\mathrm{T}} = 0$, we have

$$B\Gamma + \Gamma B^{\mathrm{T}} = -F,$$

which is the continuous version of the PRF equation (2.62).

From continuous to discrete: Replacing B with $P - I$, we get

$$\begin{aligned}&(P-I)\Gamma + \Gamma(P-I)^{\mathrm{T}} = -F \\ &(P-I)\Gamma + \Gamma(P-I)^{\mathrm{T}} = -F - (P-I)\Gamma(P-I)^{\mathrm{T}}\end{aligned}$$

where we have used $(P-I)\Gamma(P-I)^{\mathrm{T}} = 0$. Then, arranging the above equation, we have $\Gamma - P\Gamma P^{\mathrm{T}} = F$, which is the discrete version of the PRF equation (2.7).

2.14 Consider a Markov chain $\boldsymbol{X}$ with transition probabilities $p(j|i)$, $i, j \in \mathcal{S}$ and reward function f. For any $0 < p < 1$, we define an equivalent Markov chain $\boldsymbol{X}'$ with transition probabilities $p'(j|i) = (1-p)p(j|i)$, $j \neq i$, and $p'(i|i) = p + (1-p)p(i|i)$, $i \in \mathcal{S}$. Set $f' = f$. Prove that $\eta' = \eta$ and $g' = \frac{g}{1-p}$.

Solution: Let π and π' be the steady-state probabilities of Markov chain $\boldsymbol{X}$ and $\boldsymbol{X}'$, respectively. Define $P = [p(j|i)]$ and $P' = [p'(j|i)]$. From the definition of Markov chain $\boldsymbol{X}'$, we have $P' = pI + (1-p)P$. Thus, we have

$$\pi P' = \pi(pI + (1-p)P) = \pi.$$

This means the steady-state probability of Markov chain $\boldsymbol{X}$ is equal to that of Markov chain $\boldsymbol{X}'$, i.e. $\pi' = \pi$. Since the average performance $\eta' = \pi' f$, we have $\eta' = \pi' f = \pi f = \eta$.

Considering the Poisson equation

$$(I - P')g' + \eta' e = f. \tag{2.10}$$

Since the solution to Poisson equation (2.10) is up to a constant, we can choose a solution satisfying $\pi' g' = \frac{\eta}{1-p}$. In this case, putting $P' = pI + (1-p)P$ into (2.10), we can obtain

$$(1-p)(I - P)g' + (1-p)e\pi' g' = f.$$

Thus, we have

$$g' = \frac{1}{1-p}(I - P + e\pi')^{-1} f.$$

Since $\pi' = \pi$ and $g = (I - P + e\pi)^{-1} f$ is the potential of Markov chain satisfying $\pi g = \eta$, we have $g' = \frac{g}{1-p}$.

2.15 Consider a Markov process $\boldsymbol{X}$ with transition rates $\lambda(i)$, and transition probabilities $p(j|i)$, $i, j \in \mathcal{S}$, and reward function f. For any $\lambda > \lambda(i)$, $i \in \mathcal{S}$, we define an equivalent Markov process $\boldsymbol{X}'$ with transition rates $\lambda'(i) \equiv \lambda$, and transition probabilities $p'(j|i) = \frac{\lambda(i)}{\lambda} p(j|i)$, $j \neq i$, and $p'(i|i) = (1 - \frac{\lambda(i)}{\lambda}) + \frac{\lambda(i)}{\lambda} p(i|i)$. Set $f' = f$.

a. Prove that $\eta' = \eta$ and $g' = g$.

b. Let the discrete time Markov chain embedded at the transition epoches of $\boldsymbol{X}'$ as $\boldsymbol{X}^\dagger$. Find the steady state probability $\pi^\dagger$ and the potential $g^\dagger$ of $\boldsymbol{X}^\dagger$.

c. Suppose that $1 = \lambda > \lambda(i)$, $i \in \mathcal{S}$, prove that $g^\dagger = g$.

d. For any $\kappa > 0$, we define a Markov process $\widetilde{\boldsymbol{X}}$ with transition rates $\widetilde{\lambda}(i) = \kappa\lambda(i)$, $i \in \mathcal{S}$, transition probabilities $\widetilde{p}(j|i) = p(j|i)$, $i, j \in \mathcal{S}$, and reward function $\widetilde{f} = f$. Prove that $\widetilde{\pi} = \pi$ and $\widetilde{g} = \frac{g}{\kappa}$.

e. Given any Markov process $\boldsymbol{X}$, can you find a Markov chain that has the same steady-state probability π and potential g as $\boldsymbol{X}$? (Hint: use the results in b) - d).)

Solution:

a. From the definition of Markov process $\boldsymbol{X}$, we have the infinitesimal matrix of $\boldsymbol{X}$ is $A = \Lambda[P - I]$, where $P = [p(j|i)]$, I is the unit matrix and $\Lambda = diag\{\lambda(1), \cdots, \lambda(S)\}$. Similarly, the infinitesimal matrix of $\boldsymbol{X}'$ is $A' = \lambda[P' - I]$, where $P' = [p'(j|i)]$. From the definition of P', we have $P' = I + \frac{\Lambda}{\lambda}[P - I] = I + \frac{A}{\lambda}$. Thus, we have $A' = \lambda[P' - I] = \lambda * \frac{A}{\lambda} = A$. That is to say, Markov chains $\boldsymbol{X}$ and $\boldsymbol{X}'$ have the same infinitesimal matrix. Thus, they have the same steady-state distribution, i.e., $\pi' = \pi$. Then, by using $\eta = \pi f$ and Poisson equation (2.66), they have the same average performance and potentials, i.e. $\eta' = \eta$ and $g' = g$.

b. From the definition of $\boldsymbol{X}'$, we know the transition probability matrix of embedded Markov chain $\boldsymbol{X}^\dagger$ is $P' = I + \frac{A}{\lambda}$. We assume the steady-state probability of Markov process $\boldsymbol{X}$ is π, that is, π satisfies $\pi A = 0$ and $\pi e = 1$. Then, we have $\pi P' = \pi(I + \frac{A}{\lambda}) = \pi$. Thus, π is also the steady-state probability of $\boldsymbol{X}^\dagger$, which means the embedded Markov chain $\boldsymbol{X}^\dagger$ has the same steady-state probability as Markov process $\boldsymbol{X}$. For the potentials of $\boldsymbol{X}^\dagger$, we have

$$\begin{aligned} g^\dagger &= (I - P' + e\pi)^{-1} f \\ &= [I - (I + \frac{A}{\lambda}) + e\pi]^{-1} f \\ &= \lambda(-A + \lambda e\pi)^{-1} f. \end{aligned}$$

We can test $g := (-A + \lambda e\pi)^{-1} f$ is the potential of Markov process $\boldsymbol{X}$ satisfying $\pi g = \frac{\eta}{\lambda}$. Thus, we have the following relationship between the potentials of Markov chain $\boldsymbol{X}^\dagger$ and Markov process $\boldsymbol{X}$

$$g^\dagger = \lambda g.$$

c. From b), we have $g^\dagger = g$ if $\lambda = 1 > \lambda(i), i \in \mathcal{S}$.

d. From the definition of $\widetilde{\boldsymbol{X}}$, we can obtain the infinitesimal matrix of $\widetilde{\boldsymbol{X}}$,

$$\widetilde{A} = diag\{\widetilde{\lambda}(1), \cdots, \widetilde{\lambda}(S)\}[P - I] = \kappa \; diag\{\lambda(1), \cdots, \lambda(S)\}[P - I] = \kappa A.$$

Thus, we have $\pi\widetilde{A} = \kappa\pi A = 0$. That is to say, π is the steady-state probability of $\widetilde{\boldsymbol{X}}$, i.e. $\widetilde{\pi} = \pi$. For the potential, considering the Poisson equation

$$-\widetilde{A}\widetilde{g} + \widetilde{\eta}e = f,$$

Since $\widetilde{A} = \kappa A$ and $\widetilde{\eta} = \widetilde{\pi}f = \pi f = \eta$, we have

$$-\kappa A\widetilde{g} + \eta e = f. \tag{2.11}$$

The solution to equation (2.11) is up to a constant, in particular, we can choose $\widetilde{g}$ satisfying $\pi\widetilde{g} = \frac{\eta}{\kappa}$. In this case, the Poisson equation (2.11) becomes

$$\kappa(-A + e\pi)\widetilde{g} = f.$$

Thus, we have

$$\widetilde{g} = \frac{1}{\kappa}(-A + e\pi)^{-1}f.$$

We can verify $g := (-A + e\pi)^{-1}f$ is the potential of Markov process $\boldsymbol{X}$ satisfying $\pi g = \eta$. Thus, we have $\widetilde{g} = \frac{g}{\kappa}$.

e. From the the above discussions, for any Markov process $\boldsymbol{X}$ with infinitesimal matrix A satisfying the transition rate $\lambda(i) \leq 1, i \in \mathcal{S}$, we can find the Markov chain with transition probability matrix $P = A + I$, i.e. let $\lambda = 1$ as part c), has the same steady-state probability and potential as Markov process $\boldsymbol{X}$. If $\lambda(i) > 1$ for some i, it is difficult to find a Markov chain that has the same steady-state probability π and potential g as $\boldsymbol{X}$.

2.16 For semi-Markov processes with the discounted performance defined in (2.93), set $\eta_\beta := (\eta_\beta(1), \cdots, \eta_\beta(S))^T$ and $g_\beta := (g_\beta(1), \cdots, g_\beta(S))^T$. Prove that (cf.[57])

$$\lim_{\beta \to 0+} g_\beta = g,$$

$$\lim_{\beta \to 0+} \eta_\beta = \eta e,$$

and

$$\eta_\beta = \beta g_\beta + \eta e.$$

Solution:

$$\begin{aligned}
\eta_\beta(i) &= \lim_{T\to\infty} E\Big[\int_0^T \beta e^{-\beta t} f(X_t, Y_t)dt | X_0 = i\Big] \\
&= E\Big[\int_0^{T_1} \beta e^{-\beta t} f(i, Y_0)dt | X_0 = i\Big] + \lim_{N\to\infty} E\Big[\int_{T_1}^{T_N} \beta e^{-\beta t} f(X_t, Y_t)dt | X_0 = i\Big] \\
&= \sum_{j\in\mathcal{S}} \int_0^\infty \int_0^{T_1=\tau} \beta e^{-\beta t} f(i,j)dtdQ(i,j,d\tau) \\
&\quad + \lim_{N\to\infty} \sum_{j\in\mathcal{S}} \Big\{ \int_0^\infty e^{-\beta\tau} \int_{T_1=\tau}^{T_N} e^{-\beta(t-\tau)} f(X_t, Y_t)dtdQ(i,j,d\tau)\Big\} \\
&= \sum_{j\in\mathcal{S}} \Big\{ f(i,j) \int_0^\infty (1 - e^{-\beta\tau})Q(i,j,d\tau)\Big\} + \sum_{j\in\mathcal{S}} \Big\{ \int_0^\infty e^{-\beta\tau} Q(i,j,d\tau)\eta_\beta(j)\Big\} \\
&= f_\beta(i) \int_0^\infty (1 - e^{-\beta\tau})Q(i,d\tau) + \sum_{j\in\mathcal{S}} \Big\{ \int_0^\infty e^{-\beta\tau} Q(i,j,d\tau)\eta_\beta(j)\Big\}, \qquad (2.12)
\end{aligned}$$

where $f_\beta(i) = \frac{\sum_{j\in\mathcal{S}}\{f(i,j)\int_0^\infty (1-e^{-\beta\tau})Q(i,j,d\tau)\}}{\int_0^\infty (1-e^{-\beta\tau})Q(i,d\tau)}$. Dividing both sides of (2.12) by $\int_0^\infty (1 - e^{-\beta\tau})Q(i,d\tau)$, we have

$$\eta_\beta(i) = f_\beta(i) - \frac{1}{\beta}\lambda_\beta(i)\eta_\beta(i) + \frac{1}{\beta}\sum_{j\in\mathcal{S}} \lambda_\beta(i)Q_\beta(i,j)\eta_\beta(j). \qquad (2.13)$$

where $\lambda_\beta(i) = \frac{\beta\int_0^\infty e^{-\beta\tau}Q(i,d\tau)}{\int_0^\infty (1-e^{-\beta\tau})Q(i,d\tau)}$ and $Q_\beta(i,j) = \frac{\int_0^\infty e^{-\beta\tau}Q(i,j,d\tau)}{\int_0^\infty e^{-\beta\tau}Q(i,d\tau)}$. Thus, $(\beta I - A_\beta)\eta_\beta = f_\beta$, where

$$A_\beta = \begin{cases} \lambda_\beta(i)Q_\beta(i,j), & if \quad i \neq j \\ -\lambda_\beta(i)[1 - Q_\beta(i,i)], & if \quad i = j. \end{cases} \qquad (2.14)$$

So,

$$\eta_\beta = \beta(\beta I - A_\beta)^{-1} f_\beta. \qquad (2.15)$$

We can easily prove that

$$\lim_{\beta\to 0} Q_\beta(i,j) = Q(i,j).$$
$$\lim_{\beta\to 0} \lambda_\beta(i) = \lambda(i).$$
$$\lim_{\beta\to 0} f_\beta(i) = f(i).$$

Thus, we have $A_\beta \to A$. Next, we prove that $\beta(\beta I - A_\beta)^{-1} \to ep$, where p satisfies $pA = 0$ and $pe = 1$. From (2.14), we can find A_β is a infinitesimal matrix. Thus, the balance equation $p_\beta A_\beta = 0$ and $p_\beta e = 1$ has a unique solution p_β. With the continuity of p_β, we have $p_\beta \to p$. Similarly to (2.43), we have the following continuous-version equation,

$$(\beta I - A_\beta + ep_\beta)^{-1} = (\beta I - A_\beta)^{-1} - \frac{ep_\beta}{\beta(1+\beta)}. \tag{2.16}$$

Multiplying the both sides of (2.16) with β and letting $\beta \to 0$, we can easily prove $\beta(\beta I - A_\beta) \to ep$. On the bases, using (2.15), we have $\lim_{\beta\to 0} \eta_\beta = \eta e$. By using the definition of g_β and η_β, we can directly obtain $\eta_\beta = \beta g_\beta + \eta e$. Thus, putting (2.15) into $\eta_\beta = \beta g_\beta + \eta e$, we have

$$(\beta I - A_\beta) g_\beta = f_\beta - \eta e. \tag{2.17}$$

we know g_β is a unique solution up to a constant of (2.18). Let $\beta \to 0$, we know $g_0 := \lim_{\beta\to 0} g_\beta$ satisfies

$$-Ag_0 = f + \eta e. \tag{2.18}$$

From the uniqueness, $g = g_0$. Thus, we know that $\lim_{\beta\to 0} g_\beta = g$.

Reference: Cao Xi-Ren, "Semi-Markov Decision Problems and Performance sensitivity Analysis", *IEEE Transactions on Automatic Control*, vol. 48, no. 5, 758-769, 2003.

2.17 Consider a two-server cyclic Jackson queueing network with service rates μ and λ for servers 1 and 2, respectively. There are N customers in the network. The system's state $\boldsymbol{n} = n$ is the number of customers at server 1. The state process is Markov. Let the performance be the average response time of the customers at server 1, denoted as $\bar{\tau}$. Calculate the performance potentials $g(i)$, $i = 1, 2, \cdots, S$, and the performance measure $\bar{\tau}$, and derive the derivative of $\bar{\tau}$ with respect to λ and μ.

Solution:

The average response time of a customer at server 1 is

$$\bar{\tau} = \lim_{L\to\infty} \frac{1}{L} \int_0^{T_L} n(t)dt.$$

where $n(t)$ denotes the number of customers at server 1 at time t and L denotes the number of service completions at server 1 in $[0, T_L]$. We have

$$\bar{\tau} = \lim_{L\to\infty} \frac{T_L}{L} \frac{\int_0^{T_L} n(t)dt}{T_L} = \frac{\eta_T^{(f)}}{\eta_{th}},$$

where $\eta_{th} = \lim_{L\to\infty} \frac{L}{T_L}$ is the throughput of server 1 and $\eta_T^{(f)} = \lim_{L\to\infty} \frac{\int_0^{T_L} n(t)dt}{T_L} = \pi f$. We have $\eta_{th} = \sum_{n=1}^{N} \pi(n)\mu = \mu(1 - \pi(0))$. We can find

$$\bar{\tau} = \lim_{L\to\infty} \frac{\int_0^{T_L} \frac{n(t)}{\eta_{th}} dt}{T_L},$$

which is a time-average performance. Thus, this problem has become a sensitivity problem of Markov process with performance function $\tilde{f}(n(t)) = \frac{n(t)}{\eta_{th}}$ and time-average performance $\bar{\tau}$. For this process, the infinitesimal matrix is

$$B = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & \cdots & \mu & -(\mu+\lambda) & \lambda \\ 0 & 0 & \cdots & \mu & -\mu \end{bmatrix}. \tag{2.19}$$

We can compute π by $\pi B = 0$ and $\pi e = 1$. Then we can compute the potentials by using $g^{\tilde{f}} = (-B + e\pi)^{-1}\tilde{f}$, which is the potential corresponding to performance function $\tilde{f}$, and compute the performance measure $\bar{\tau}$ by using $\bar{\tau} = \pi\tilde{f}$. The derivative of $\bar{\tau}$ with respect to μ is

$$\begin{aligned} \frac{d\bar{\tau}}{d\mu} &= \pi[\frac{dB}{d\mu} g^{\tilde{f}} + \frac{d\tilde{f}}{d\mu}] = \pi[\frac{dB}{d\mu} g^{\tilde{f}} - \frac{f d\eta_{th}}{\eta_{th}^2}] \\ &= \pi\frac{dB}{d\mu} g^{\tilde{f}} - \frac{\bar{\tau}}{\eta_{th}} \frac{d\eta_{th}}{d\mu} = \pi\frac{dB}{d\mu} g^{\tilde{f}} - \frac{\bar{\tau}}{\eta_{th}} \pi\frac{dB}{d\mu} g^{\mu}, \end{aligned} \tag{2.20}$$

where we have used the derivative of throughput η_{th} with respect to μ and g^{μ} is the potentials corresponding to the performance function $\mu = [0, \mu, \mu, \cdots, \mu]^T$. By using this derivative formula, we can compute the derivative.

If we consider the potential g^f corresponding to the performance function $f(n(t)) = n(t)$, the potential g^f can be also computed by $g^f = (-B + e\pi)^{-1} f$ and $\bar{\tau}$ can be computed by $\bar{\tau} = \frac{\eta_T^{(f)}}{\eta_{th}}$, and the derivative $\frac{d\bar{\tau}}{d\mu}$ can be computed by taking derivative of quotient of

$\eta_T^{(f)}$ and η_{th} as follows:

$$\begin{aligned}\frac{d\bar{\tau}}{d\mu} &= \frac{\eta_{th}\frac{d\eta_T^{(f)}}{d\mu} + \eta_T^{(f)}\frac{d\eta_{th}}{d\mu}}{\eta_{th}^2} = \frac{\eta_{th}\pi\frac{dB}{d\mu}g^f + \eta_T^{(f)}\pi\frac{dB}{d\mu}g^\mu}{\eta_{th}^2} \\ &= \frac{\pi\frac{dB}{d\mu}g^f + \bar{\tau}\pi\frac{dB}{d\mu}g^\mu}{\eta_{th}} = \frac{1}{\eta_{th}}\pi\frac{dB}{d\mu}(g^f + \bar{\tau}g^\mu). \end{aligned} \tag{2.21}$$

In fact, derivative formula (2.20) is equal to (2.21) because $g^{\tilde{f}} = \frac{g^f}{\eta_{th}}$. Similarly, we can also compute the derivative with respect to λ by using the following derivative formula,

$$\frac{d\bar{\tau}}{d\lambda} = \frac{1}{\eta_{th}}\pi\frac{dB}{d\lambda}(g^f + \bar{\tau}g^\mu). \tag{2.22}$$

2.18 The two-server N-customer cyclic Jackson queueing network studied in Problem 2.17 is equivalent to an $M/M/1/N$ queue with arrival rate λ, service rate μ, and a finite buffer size N. (When the number of customers in the queue $n = N$, an arriving customer is simply lost.)

a. Suppose the arrival rate only changes when $n = 0$; i.e., when $n = 0$, λ changes to $\lambda + \Delta\lambda$, and when $n > 0$, λ keeps unchanged. What is the derivative of the average response time $\bar{\tau}$ with respect to this change?

b. Suppose the arrival rate only changes when $n = n^*$, with $0 < n^* < N$. What is the derivative of $\bar{\tau}$ with respect to this change?

c. Suppose the arrival rate only changes when $n = N$. What is the derivative of $\bar{\tau}$ with respect to this change? (You may view the $M/M/1/N$ queue as the two-server cyclic queue again to verify your result.)

Solution:

a. Suppose the arrival rate only changes when $n = 0$, then the infinitesimal matrix B in Problem 2.17 changes to

$$B = \begin{bmatrix} -(\lambda+\Delta\lambda) & \lambda+\Delta\lambda & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & \cdots & \mu & -(\mu+\lambda) & \lambda \\ 0 & 0 & \cdots & \mu & -\mu \end{bmatrix} \tag{2.23}$$

Then, the derivative B with respect to this change is

$$\frac{dB}{d\lambda} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Substituting it into (2.22), we have the derivative of $\bar{\tau}$ with respect to this change:

$$\frac{d\bar{\tau}}{d\lambda} = \frac{\pi(0)}{\eta_{th}}\Big((g^f(1) - g^f(0)) + \bar{\tau}(g^\mu(1) - g^\mu(0))\Big)$$

b. Similarly, if the arrival rate only changes when $n = n^*$, the derivative of $\bar{\tau}$ with respect to this change is

$$\frac{d\bar{\tau}}{d\lambda} = \frac{\pi(n^*)}{\eta_{th}}\Big[(g^f(n^*+1) - g^f(n^*)) + \bar{\tau}(g^\mu(n^*+1) - g^\mu(n^*))\Big].$$

c. Suppose the arrival rate only changes when $n = N$, this change does not affect the performance measure and the infinitesimal matrix is still the original one, so the derivative with respect to this change is zero.

2.19 Consider a Markov chain with one closed recurrent state set $\mathcal{S}_1$ and one transient state set $\mathcal{S}_2$ (a uni-chain). Let the transition probability matrix be

$$P = \begin{bmatrix} P_1 & 0 \\ P_{21} & P_{22} \end{bmatrix},$$

with P_1 corresponding to $\mathcal{S}_1$ and $P_{21} \neq 0$, P_{22} to $\mathcal{S}_2$, and 0 being a matrix with all zero components. Denote the potential vector as $g = (g_1^T, g_2^T)^T$ with $g_1 = (g(1), \cdots, g(S_1))^T$ and $g_2 = (g(S_1+1), \cdots, g(S))^T$, $S_1 = |\mathcal{S}_1|$, $S_2 = |\mathcal{S}_2|$, $S_1 + S_2 = S$.

Derive an equation for g_1 and express g_2 in terms of g_1 and P_{21}, P_{22}.

Solution:

S_2 is a transient state set, then the steady-state probability of this Markov chain is

$$\pi = (\pi_1, \mathbf{0})$$

where π_1 is steady state probability of S_1 recurrent states and $\mathbf{0}$ is an S_2 dimensional row vector whose components are all zeros. Thus from balance equation, we have $\pi_1 P_1 = \pi_1$.

Let $f = (f_1^{\mathrm{T}}, f_2^{\mathrm{T}})^{\mathrm{T}}$. From Poisson equation,

$$(I - P + e\pi)g = f,$$

we have

$$\left(\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} P_1 & 0 \\ P_{21} & P_{22} \end{pmatrix} + \begin{pmatrix} e_1\pi_1 & 0 \\ e_2\pi_1 & 0 \end{pmatrix} \right) \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix},$$

where e_1 and e_2 are S_1-dimension and S_2-dimension column vector with all elements equal 1. That is,

$$\begin{pmatrix} I - P_1 + e_1\pi_1 & 0 \\ -P_{21} + e_2\pi_1 & I - P_{22} \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

Therefore, we have

$$\begin{aligned} g_1 &= (I - P_1 + e_1\pi_1)^{-1} f_1, \\ (-P_{21} + e_2\pi_1)g_1 &+ (I - P_{22})g_2 = f_2. \end{aligned}$$

Therefore

$$g_2 = (I - P_{22})^{-1} \left[f_2 + (P_{21} - e_2\pi_1)g_1 \right].$$

2.20 Consider a Markov chain with transition probability matrix

$$P = \begin{bmatrix} B & b \\ 0 & 1 \end{bmatrix},$$

where B is an $(S-1) \times (S-1)$ irreducible matrix, $b > 0$ is an $(S-1)$ dimensional column vector, 0 represents an $(S-1)$ dimensional row vector whose components are all zero. The last state S is an absorbing state. Clearly, the long-run average performance for this Markov chain is $\eta = f(S)$, independent of B, b, and the initial state. Thus, the long-run average does not reflect the transient behavior. Now, we set $f(S) = 0$. Define

$$g(i) = E\{\sum_{l=0}^{\infty} f(X_l) | X_0 = i\}.$$

Let $L_{i,S} = \min\{l \geq 0, X_l = S | X_0 = i\}$ be the first passage time from i to S. Then

$$g(i) = E\{ \sum_{l=0}^{L_{i,S}-1} f(X_l) | X_0 = i\}.$$

a. Derive an equation for $g = (g(1), \cdots, g(S))^T$.

b. Derive an equation for the average first passage times $E[L_{i,S}]$, $i \in \mathcal{S}$.

Solution:

a). Let $f = (f_1^{\mathrm{T}}, 0)^{\mathrm{T}}$ be the performance function and $g = (g_1^{\mathrm{T}}, g(S))^{\mathrm{T}}$ be the performance potential. Obviously, we have $g(S) = 0$ and $\pi = (0, \cdots, 0, 1)$. From Poisson equation, we have

$$\left(\begin{pmatrix} I & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} B & b \\ 0 & 1 \end{pmatrix} + e_S(0, \cdots, 0, 1) \right) \begin{pmatrix} g_1 \\ 0 \end{pmatrix} = \begin{pmatrix} f_1 \\ 0 \end{pmatrix}$$

Then we get $g_1 = (I - B)^{-1} f_1$.

b). Let $f = (1, \cdots, 1, 0)^{\mathrm{T}}$, then $f_1 = e_{S-1}$, which is a $(S-1)$ dimensional column vector whose components are all 1. $g(i) = \mathrm{E}\{\sum_{l=0}^{L_{i,S}-1} f(X_l)|X_0 = i\} = \mathrm{E}\{L_{i,S}\}$. From above results, we have

$$\mathrm{E}\{L_{i,S}\} = [(I - B)^{-1} e_{S-1}]_i.$$

where $[\ \cdot\ ]_i$ denotes the i-th component of vector.

2.21* (This problem helps in understanding the difference between the discounted performance criteria for both the discrete-time and continuous-time models.) Consider a Markov chain $\boldsymbol{X}$ with transition probability matrix $P = [p(j|i)]_{i,j=1}^S$ and reward function $f(i)$, $i = 1, 2, \cdots, S$. For simplicity, we assume that $p(i|i) = 0$ for all $i = 1, 2, \cdots, S$. Let $\widetilde{\boldsymbol{X}}$ be a Markov chain with reward function $\widetilde{f}(i) = f(i)$, $i = 1, 2, \cdots, S$, and transition probability matrix $\widetilde{P}$ defined as $\widetilde{p}(i|i) = q$, $0 < q < 1$, and $\widetilde{p}(j|i) = (1-q)p(j|i)$, $j \neq i$, $i, j = 1, 2, \ldots, S$.

a. Prove that $\widetilde{\boldsymbol{X}}$ is equivalent to $\boldsymbol{X}$ in the sense that they have the same steady-state probabilities: $\widetilde{\pi}(i) = \pi(i)$ for all $i = 1, 2, \ldots, S$.

b. The discounted reward of $\boldsymbol{X}$ is defined as (2.35):

$$\eta_\beta(i) = (1-\beta) E\left\{ \sum_{l=0}^{\infty} \beta^l f(X_l) \Big| X_0 = i \right\},$$

where $0 < \beta < 1$ is a discount factor. Similarly, the discounted reward of $\widetilde{\boldsymbol{X}}$ is defined with a discount factor $0 < \widetilde{\beta} < 1$ as

$$\widetilde{\eta}_{\widetilde{\beta}}(i) = (1-\widetilde{\beta}) E\left\{ \sum_{l=0}^{\infty} \widetilde{\beta}^l f(\widetilde{X}_l) \Big| \widetilde{X}_0 = i \right\}.$$

Find a value for $\widetilde{\beta}$ such that $\widetilde{\eta}_{\widetilde{\beta}}(i) = \eta_\beta(i)$ for all $i = 1, 2, \cdots, S$.

c. Let $\Delta > 0$ be a positive number. Consider a continuous-time (not Markov) process $\hat{\boldsymbol{X}} := \{\hat{X}_t, t \in [0, \infty)\}$, where $\hat{X}_t = X_l$ if $l\Delta \le t < (l+1)\Delta$, $l = 0, 1, \cdots$, with $\boldsymbol{X} = \{X_l, l = 0, 1, \cdots\}$ being the Markov chain considered in a). The discounted reward of $\hat{\boldsymbol{X}}$ is defined by an exponential weighting factor (cf. (2.93)):

$$\eta_\alpha(i) = \lim_{T\to\infty} E\left[\int_0^T \alpha e^{-\alpha t} f(\hat{X}_t)dt \Big| X_0 = i\right], \qquad T_0 = 0.$$

What is the equivalent β such that $\eta_\beta(i) = \eta_\alpha(i)$ for all $i = 1, 2, \cdots, S$?

d. Repeat c) for continuous-time process $\hat{\boldsymbol{X}} := \{\hat{X}_t, t \in [0, \infty)\}$, with $\hat{X}_t = \widetilde{X}_l$ if $l\Delta \le t < (l+1)\Delta$, $l = 0, 1, \cdots$.

e. How about in d) we let $\Delta \to 0$ while keeping $\frac{1-q}{\Delta} = \lambda$? (where λ is a constant).

(Hint: If $\boldsymbol{X} = \{X_0 = i_0, X_1 = i_1, \cdots, \}$, then we have $\widetilde{\boldsymbol{X}} = \{\widetilde{X}_0 = \widetilde{X}_1 = \cdots = \widetilde{X}_{n_0-1} = i_0, \widetilde{X}_{n_0} = \widetilde{X}_{n_0+1} = \cdots = \widetilde{X}_{n_0+n_1-1} = i_1, \cdots, \}$, where n_l is the numbers of consecutive visits to state i_l, $l = 0, 1, \cdots$. Note that n_l is geometrically distributed with parameter q. Therefore,

$$\widetilde{\eta}_{\widetilde{\beta}}(i) = (1-\widetilde{\beta})E\{(1 + \widetilde{\beta} + \cdots + \widetilde{\beta}^{n_0-1})f(i_0) + (\widetilde{\beta}^{n_0} + \cdots + \widetilde{\beta}^{n_0+n_1-1})f(i_1) + \cdots\}.$$

We conclude that $\widetilde{\eta}_{\widetilde{\beta}}(i) = \eta_\beta(i)$ if $\beta = \frac{(1-q)\widetilde{\beta}}{1-q\widetilde{\beta}}$.)

Solution:

a. The steady state probability vector of Markov chain $\widetilde{X}$ satisfies the following flow balance equation:

$$\widetilde{\pi}(\widetilde{P} - I) = 0, \tag{2.24}$$

$$\widetilde{\pi}e = 1. \tag{2.25}$$

From (2.24), we have $(1-q)\widetilde{\pi}(P - I) = 0, 0 < q < 1$. So $\widetilde{\pi}(P - I) = 0$. Combining with (2.25), we know $\widetilde{\pi}$ is also the steady state probability vector of Markov chain X. Thus, $\widetilde{\pi}(i) = \pi(i), i \in \mathcal{S}$.

b. Firstly, we give an intuitive explanation. For Markov chain $\widetilde{X}$ with discount factor $\widetilde{\beta}$, since the transition probability matrix is defined as $\widetilde{p}(i|i) = q$, $0 < q < 1$, and

$\widetilde{p}(j|i) = (1-q)p(j|i)$, $j \neq i$, $i, j = 1, 2, \cdots, S$, thus, we know the dwell time in each state follows a geometrical distribution with parameter q. Then, the probability that Markov chain $\widetilde{X}$ transits to one state n times consecutively is $q^n(1-q)$. So, the total expected discount factor is $\sum_{n=0}^{\infty}(1-q)q^n\widetilde{\beta}^{n+1} = \frac{(1-q)\widetilde{\beta}}{1-q\widetilde{\beta}}$. Therefore, if let $\beta = \frac{(1-q)\widetilde{\beta}}{1-q\widetilde{\beta}}$, i.e. $\widetilde{\beta} = \frac{\beta}{1-q+q\beta}$, then $\widetilde{\eta}_{\widetilde{\beta}}(i) = \eta_{\beta}(i)$.

Next, we give a rigorous proof. From (2.31),

$$\begin{aligned}
\eta_{\widetilde{\beta}} &= (1-\widetilde{\beta})(1-\widetilde{\beta}\widetilde{P})^{-1}f \\
&= (1-\widetilde{\beta})\sum_{l=0}^{\infty}\widetilde{\beta}^l\widetilde{P}^l f \\
&= (1-\widetilde{\beta})\sum_{l=0}^{\infty}\widetilde{\beta}^l[qI+(1-q)P]^l f \\
&= (1-\widetilde{\beta})\sum_{l=0}^{\infty}\widetilde{\beta}^l\sum_{n=0}^{l}\binom{l}{n}(1-q)^n q^{l-n}P^n f \\
&= (1-\widetilde{\beta})\sum_{n=0}^{\infty}\sum_{l=n}^{\infty}\widetilde{\beta}^l\binom{l}{n}(1-q)^n q^{l-n}P^n f \\
&= (1-\widetilde{\beta})\sum_{n=0}^{\infty}(1-q)^n\widetilde{\beta}^n\sum_{l=0}^{\infty}\widetilde{\beta}^l\binom{l+n}{n}q^l P^n f \\
&= (1-\widetilde{\beta})\sum_{n=0}^{\infty}(1-q)^n\widetilde{\beta}^n\frac{1}{(1-q\widetilde{\beta})^{n+1}}P^n f \\
&= (1-\frac{(1-q)\widetilde{\beta}}{1-q\widetilde{\beta}})\sum_{n=0}^{\infty}\Big(\frac{(1-q)\widetilde{\beta}}{1-q\widetilde{\beta}}\Big)^n P^n f \\
(if\ \beta := \frac{(1-q)\widetilde{\beta}}{1-q\widetilde{\beta}} \Rightarrow) \quad &= (1-\beta)\sum_{n=0}^{\infty}\beta^n P^n f \\
&= (1-\beta)(I-\beta P)^{-1}f = \eta_{\beta}
\end{aligned}$$

where we have used the Binomial formula $\frac{1}{(1-x)^r} = \sum_{k=0}^{\infty}\binom{r+k-1}{k}x^k$ in the seventh equation.

c. Since

$$\begin{aligned}
\eta_{\alpha}(i) &= \lim_{T\to\infty} E\left[\int_0^T \alpha e^{-\alpha t}f(\hat{X}_t)dt|X_0 = i\right] \\
&= E\{\int_0^{\Delta}\alpha e^{-\alpha t}dt f(X_0) + \int_{\Delta}^{2\Delta}\alpha e^{-\alpha t}dt f(X_1) + \cdots\} \\
&= (1-e^{-\alpha\Delta})E\{f(X_0) + e^{-\alpha\Delta}f(X_1) + e^{-2\alpha\Delta}f(X_2)\cdots\}
\end{aligned}$$

Thus, if $\beta = e^{-\alpha\Delta}$, $\eta_{\beta}(i) = \eta_{\alpha}(i)$ for all i.

d. Similarly to c), we have equivalent $\widetilde{\beta} = e^{-\alpha\Delta}$ such that $\eta_{\widetilde{\beta}}(i) = \eta_\alpha(i)$ for all $i \in \mathcal{S}$. Then by using b), if $\beta = \frac{(1-q)e^{-\alpha\Delta}}{1-qe^{-\alpha\Delta}}$, we have $\eta_\beta(i) = \eta_\alpha(i)$ for all i.

e. If let $\Delta \to 0$ while keeping $\frac{1-q}{\Delta} = \lambda$, then $q \to 1$. Since

$$\begin{aligned}\beta &= \frac{(1-q)e^{-\alpha\Delta}}{1-qe^{-\alpha\Delta}} = \frac{(1-q)(1-\alpha\Delta+o(\Delta^2))}{1-q(1-\alpha\Delta+o(\Delta^2))} \\ &= \frac{\frac{1-q}{\Delta} - (1-q)\alpha + o(\Delta)}{\frac{1-q}{\Delta} + q\alpha + o(\Delta)} \\ &\to \frac{\lambda}{\lambda+\alpha}.\end{aligned}$$

2.22 Prove that the random variable s generated according to (2.96) is indeed exponentially distributed.

Solution: We have $x = -\bar{s}\ln(1-\xi)$, then the distribution function is

$$F(s) = P(x < s) = P(-\bar{s}\ln(1-\xi) < s) = P(\xi < 1 - e^{-\frac{s}{\bar{s}}}),$$

for $s > 0$. It's obviously that $1 - e^{-\frac{s}{\bar{s}}} < 1$, and ξ is uniform distribution on $[0, 1)$, so

$$F(s) = 1 - e^{-\frac{s}{\bar{s}}}, \qquad s > 0.$$

If $s \le 0$, since $-\bar{s}\ln(1-\xi) > 0$, we have $F(s) = P(-\bar{s}\ln(1-\xi) < s) = 0$. Then,

$$F(s) = \begin{cases} 1 - e^{-\frac{s}{\bar{s}}}, & s > 0, \\ 0, & s \le 0. \end{cases}$$

Thus, he random variable s generated according to (2.96) is indeed exponentially distributed.

2.23 Develop a PA algorithm to determine a perturbed path for an open Jackson network consisting of M servers, with mean service time $\bar{s}_i$, $i = 1, 2, \cdots, M$. The customers arrive in a Poisson process with mean inter-arrival time $a = \frac{1}{\lambda}$. Both a and $\bar{s}_i$, $i = 1, 2, \cdots, M$, may be perturbed.

Solution: Algorithm:

Given an original sample path for an open Jackson network.

i. Initialize: Set $\Delta_i := 0$, $i = 0, 1, 2, \ldots, M$.

ii. Perturbation generation: At the k-th service completion time of server i, set $\Delta_i := \Delta_i + s_{i,k}$, *k=1,2,..., i=1,2,...,M.* $s_{i,k}$ *is the service time of the customer. At the k-th outside customer arrival, set* $\Delta_0 := \Delta_0 + a_k$, *k=1,2,....* a_k *is the inter-arrival time of the customer.*

iii. Perturbation propagation: When a customer from server i terminates an idle period of server j, set $\Delta_j := \Delta_i$. *When an outside customer terminates an idle period of server j, set* $\Delta_j := \Delta_0$.

2.24 Suppose that at some time the perturbations of the servers in a closed network are $\Delta_1, \Delta_2, \cdots, \Delta_M$ determined by Algorithm 2.1. What is the perturbation that has been realized by the network at that time? As we know, if a perturbation is realized, then the future perturbed sample path looks the same as the original one except shifted to the right by the amount equal to the perturbation. Can we use this fact to simplify the calculation in Algorithm 2.2?

Solution: The perturbation that has been realized by the network at that time is $\Delta = \min(\Delta_1, \Delta_2, \cdots, \Delta_M)$.

we can simplify the Algorithm 2.2 as follows: Since the perturbation Δ has been realized at some time m, then each of $\Delta_1, \Delta_2, \cdots, \Delta_M$ contains this perturbation. That is to say, the perturbed sample path is the same as the original one except that the entire sample path is shifted to the right by the amount Δ. Then, at the transition times after m, the update of ΔF can be set as $\Delta F := \Delta F + [f(\mathbf{n}) - f(\mathbf{n}')]\Delta T_l'$, where $\Delta T_l' = \Delta T_l - \Delta$, $\mathbf{n} = N(T_{l-})$, and $\mathbf{n}' = N(T_l)$. This is because $\Delta F_L = \Delta F_{m-1} + \sum_{l=m}^{L}[f(N(T_{l-})) - f((N(T_l))]\Delta T_l = \Delta F_{m-1} + \sum_{l=m}^{L}[f(N(T_{l-})) - f((N(T_l))]\Delta T_l' + [f(N(T_{m-})) - f((N(T_L))]\Delta$. Since performance derivative is $\frac{\bar{s}_v}{\eta^{(I)}}\frac{\partial \eta^{(f)}}{\partial \bar{s}_v} \approx \frac{\Delta F_L}{T_L}$, when T_L is sufficiently large, we can omit $[f(N(T_{m-})) - f((N(T_L))]\Delta$. Thus, we can update ΔF as $\Delta F := \Delta F + [f(\mathbf{n}) - f(\mathbf{n}')]\Delta T_l'$.

2.25 Using the 0-1 vector array (2.105), discuss the situation of the propagation of M perturbations with the same size, each at one server, along a sample path. Prove $\sum_{i=1}^{M} c(\boldsymbol{n}, i) = 1$.

Solution: In the same way, we have M row vectors, each of which represents the propagation of one perturbation. Then we have a $M \times M$ unit matrix. From perturbation

propagation rule, if a customer from server i terminates an idle period of server j, then all perturbations of server i is propagated to server j. We just need to copy the ith column of the matrix to the jth column. Obviously, no matter how the perturbations propagate, each column of the matrix has one and only one 1, other components are all 0. That means, each server will have one and only one realized perturbation. Eventually, the matrix may reach a matrix in which one row is all 1. That is, only one of these perturbations is realized and the others are lost. Then the probability that perturbations are realized is 1. So the summation of realization probabilities is 1. That is $\sum_{i=1}^{M} c(\boldsymbol{n}, i) = 1$.

2.26 We further study the propagations of two equal perturbations $\Delta_1 = \Delta$ at server 1 and $\Delta_2 = \Delta$ at server 2 simultaneously on the same sample path. Consider the array in (2.105). Set $w(t) = w_1(t) + w_2(t)$.

1. What is the meaning of $w(t)$?

2. What does it mean when $w(t) = (1, 1, \ldots, 1)$ or $w(t) = (0, 0, \ldots, 0)$?

3. How does $w(t)$ evolve?

Solution:

a. $w(t)$ denotes which server has the perturbation Δ_1 or Δ_2 at time t.

b. $w(t) = (1, 1, \ldots, 1)$ denotes all severs have a perturbation Δ, which can be either Δ_1 or Δ_2. $w(t) = (0, 0, \ldots, 0)$ denotes the perturbations Δ_1 and Δ_2 have been lost.

c. When server i terminates an idle period of server j, the perturbation (either 0 or Δ) will be propagated to server j. This is equivalent to simply set $w_j = w_i$. The initial value of $w(t)$ is $w(0) = (1, 1, 0, \ldots, 0)$. Eventually, the array may reach $(1, 1, \ldots, 1)$ or $(0, 0, \ldots, 0)$.

2.27 In addition to (2.94), we may define the system performance as the long-run time average

$$\eta_T^{(f)} = \lim_{L\to\infty} \frac{1}{T_L} \int_0^{T_L} f(\mathbf{N}(t))dt.$$

We have $\eta_T^{(f)} = \frac{\eta^{(f)}}{\eta^{(I)}}$.

a. Derive the derivative of $\eta_T^{(f)}$ with respect to $\bar{s}_i$, $i = 1, 2, \cdots, M$.

b. Define the reward function f corresponding to the steady-state probability $\pi(\boldsymbol{n})$, with $\boldsymbol{n}$ being any state, and derive $\frac{d\pi(\boldsymbol{n})}{d\bar{s}_i}$, $i = 1, 2 \cdots, M$.

Solution:

a. We have $\eta_T^{(f)} = \eta^{(f)}/\eta^{(I)}$, thus

$$\begin{aligned}\frac{\partial \eta_T^{(f)}}{\partial \bar{s}_i} &= \frac{1}{\eta^{(I)}}\frac{\partial \eta^{(f)}}{\partial \bar{s}_i} - \eta^{(f)}(\frac{1}{\eta^{(I)}})^2\frac{\partial \eta^{(I)}}{\partial \bar{s}_i} \\ &= \frac{1}{\bar{s}_i}\sum_{all\ \mathbf{n}} \pi(\mathbf{n})(c^{(f)}(\mathbf{n}, i) - c(\mathbf{n}, i)\frac{\eta^{(f)}}{\eta^{(I)}}) \\ &= \frac{1}{\bar{s}_i}\sum_{all\ \mathbf{n}} \pi(\mathbf{n})(c^{(f)}(\mathbf{n}, i) - c(\mathbf{n}, i)\eta_T^{(f)})\end{aligned}$$

b. For any state $\mathbf{n}$, let $f(\mathbf{n}) = 1$ and 0 otherwise. Since $\eta_T^f = \sum_{\mathbf{n}\in\mathcal{S}} \pi(\mathbf{n})f(\mathbf{n})$, we have $\eta_T^{(f)} = \pi(\mathbf{n})$. We can apply the equation in a) to get $\frac{\partial \pi(\mathbf{n})}{\partial \bar{s}_i}$.

2.28 Prove that in a closed Jackson network the sample function $T_L(\xi, \bar{s}_v)$ (with ξ fixed) is a piecewise linear function of $\bar{s}_v$, $v = 1, 2, \cdots, M$ (see [46]).

Solution:

For a closed Jackson network, the service time of server v follows an exponential distribution with mean $\bar{s}_v$, then the service time of the k-th customer at server v is

$$s_{v,k} = -\bar{s}_v \ln(1 - \xi_{v,k}), \quad k = 1, 2, \ldots.$$

If $\bar{s}_v$ is changed to $\bar{s}_v + \Delta\bar{s}_v$, the service time of the k-th customer at server v is

$$s_{v,k} = -(\bar{s}_v + \Delta\bar{s}_v) \ln(1 - \xi_{v,k}), \quad k = 1, 2, \ldots.$$

Let $t_{i,k}$ be the time of service completion of k-th customer at server $i, i = 1, 2, \ldots, M, k = 1, 2, \ldots$. Then, we have the following recursive formula for $t_{i,k}$:

$$t_{i,k} = \begin{cases} t_{i,k-1} + s'_{i,k} & if\ N(t_{i,k-1}+) \neq 0 \\ t_{j,h} + s'_{i,k} & if\ N(t_{i,k-1}+) = 0 \end{cases},$$

where

$$s'_{i,k} = \begin{cases} s_{i,k} & if\ i \neq v, \\ -(\bar{s}_v + \Delta\bar{s}_v) \ln(1 - \xi_{v,k}) & if\ i = v. \end{cases}$$

and $t_{j,h}$ denotes the time of service completion of the h-th customer that has completed its service at server j moves into server i. If the sample path for different $\bar{s}_v$ is similar, that is, the embedded Markov chain of closed Jackson network for these different $\bar{s}_v$ is the same (or the order of events in the nominal and perturbed paths remains the same), then it is clear that $t_{i,k}, i = 1, 2, \ldots, M$, are linear with respect to $\bar{s}_v$. To guarantee the similarity of the sample path under different $\bar{s}_v$, we need that $\bar{s}_v$ cannot be changed largely. That is, if $\bar{s}_v$ changes in a interval $(\bar{s}_v^{min}, \bar{s}_v^{max})$, which depends on the sample path, the similarity can be guaranteed. Thus $t_{i,k}, i = 1, 2, \cdots, M$, are a linear function on this interval. For any $\bar{s}_v$, there is a interval to make $t_{i,k}, i = 1, 2 \cdots, M$, linear with respect to $\bar{s}_v$. Therefore, $t_{i,k}, i = 1, 2 \cdots, M$, are piecewise linear functions of $\bar{s}_v$. Since there always exists a w and a k such that $T_L(\xi, \bar{s}_v) = t_{w,k}$, $T_L(\xi, \bar{s}_v)$ is piecewise linear with respect to these $\bar{s}_v$, $v = 1, 2, \ldots, M$.

2.29 Consider a closed Jackson network in which $\mu_i q_{i,j} = \mu_j q_{j,i}$, $i, j = 1, 2, \cdots, M$. Prove that

$$c(\boldsymbol{n}, k) = \frac{n_k}{N}, \qquad k = 1, 2, \cdots, M;$$

and

$$\frac{\bar{s}_k}{\eta} \frac{\partial \eta}{\partial \bar{s}_k} = -\frac{1}{M},$$

where $k = 1, 2, \cdots, M$, denote any server in the network.

Solution:

In order to prove $c(\mathbf{n}, k) = n_k/N$, we can substitute it into the equations which the realization probabilities satisfy. If the equations still hold, then $c(\mathbf{n}, k) = n_k/N$ holds because the equations have a unique solution for irreducible closed Jackson networks.

1. If $n_k = 0$, then $c(\mathbf{n}, k) = 0 = n_k/N$.
2. $\sum_{k=1}^{M} c(\mathbf{n}, k) = \sum_{k=1}^{M} n_k/N = 1$.
3. If $n_k > 0$, then $\epsilon(n_k) = 1$. The first term on the right side is

$$\begin{aligned} & \sum_{i=1}^{M} \sum_{j=1}^{M} \epsilon(n_i) \mu_i q_{ij} c(\mathbf{n}_{ij}, k) \\ = & \sum_{i \neq k} \sum_{j \neq k} \epsilon(n_i) \mu_i q_{ij} \frac{n_k}{N} + \sum_{i \neq k} \epsilon(n_i) \mu_i q_{ik} \frac{(n_k + 1)}{N} + \sum_{j \neq k} \epsilon(n_k) \mu_k q_{kj} \frac{(n_k - 1)}{N} + \epsilon(n_k) \mu_k q_{kk} \frac{n_k}{N} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\neq k}\sum_{j}\epsilon(n_i)\mu_i q_{ij}\frac{n_k}{N} + \sum_{i\neq k}\epsilon(n_i)\mu_i q_{ik}\frac{1}{N} + \sum_{j}\epsilon(n_k)\mu_k q_{kj}\frac{(n_k-1)}{N} + \epsilon(n_k)\mu_k q_{kk}\frac{1}{N} \\
&= \sum_{i\neq k}\epsilon(n_i)\mu_i\frac{n_k}{N} + \sum_{i\neq k}\epsilon(n_i)\mu_i q_{ik}\frac{1}{N} + \epsilon(n_k)\mu_k\frac{(n_k-1)}{N} + \epsilon(n_k)\mu_k q_{kk}\frac{1}{N} \\
&= \sum_{i}\epsilon(n_i)\mu_i\frac{n_k}{N} + \sum_{i\neq k}\epsilon(n_i)\mu_i q_{ik}\frac{1}{N} - \mu_k\frac{1}{N} + \mu_k q_{kk}\frac{1}{N} \qquad (\epsilon(n_k)=1)
\end{aligned}
$$

The second term on the right side is

$$
\begin{aligned}
&\sum_{i=1}^{M}\mu_k q_{ki}(1-\epsilon(n_i))c(\mathbf{n}_{ki},i) \\
&= \sum_{i\neq k}\mu_k q_{ki}(1-\epsilon(n_i))\frac{n_i+1}{N} + \mu_k q_{kk}(1-\epsilon(n_k))\frac{n_k}{N} \\
&= \sum_{i\neq k}\mu_k q_{ki}(1-\epsilon(n_i))\frac{n_i+1}{N} \qquad (\epsilon(n_k)=1) \\
&= \sum_{i\neq k}\mu_k q_{ki}(1-\epsilon(n_i))\frac{1}{N} \qquad (n_i(1-\epsilon(n_i))=0)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&right \\
&= \sum_{i\neq k}\mu_k q_{ki}(1-\epsilon(n_i))\frac{1}{N} + \sum_{i}\epsilon(n_i)\mu_i\frac{n_k}{N} + \sum_{i\neq k}\epsilon(n_i)\mu_i q_{ik}\frac{1}{N} - \mu_k\frac{1}{N} + \mu_k q_{kk}\frac{1}{N} \\
&= \sum_{i}\epsilon(n_i)\mu_i\frac{n_k}{N} + \sum_{i\neq k}\mu_k q_{ki}\frac{1}{N} + \mu_k q_{kk}\frac{1}{N} - \mu_k\frac{1}{N} + \sum_{i\neq k}\epsilon(n_i)\frac{1}{N}(\mu_i q_{ik} - \mu_k q_{ki}) \\
&= \sum_{i}\epsilon(n_i)\mu_i\frac{n_k}{N} + \sum_{i}\mu_k q_{ki}\frac{1}{N} - \mu_k\frac{1}{N} \qquad (\mu_i q_{ik} = \mu_k q_{ki}) \\
&= \sum_{i}\epsilon(n_i)\mu_i\frac{n_k}{N} \\
&= \sum_{i=1}^{M}\epsilon(n_i)\mu_i c(\mathbf{n},k) \\
&= left
\end{aligned}
$$

From 1,2 and 3, $c(\mathbf{n},k) = n_k/N$ is proved.

$\mu_i = \sum_{j=1}^{M}\mu_i q_{ij} = \sum_{j=1}^{M}\mu_j q_{ji}$, so μ_i is one solution of equation (C.5). Let visit ratio $v_i = \mu_i$, then $x_i = v_i/\mu_i = 1$.

$$
\frac{\bar{s}_k}{\eta}\frac{\partial\eta}{\partial\bar{s}_k} = -\sum_{\text{all }\mathbf{n}} p(\mathbf{n})c(\mathbf{n},i) = -\sum_{n=1}^{N} p(n_k=n)\frac{n}{N} = -\frac{\bar{n}_k}{N}
$$

$\bar{n}_k$ is the mean queueing length of server k, and

$$\bar{n}_k = \sum_{n=1}^{N} x_k^n \frac{G_M(N-n)}{G_M(N)} = \sum_{n=1}^{N} \frac{G_M(N-n)}{G_M(N)}$$

$\frac{G_M(N-n)}{G_M(N)}$ is the same for all server. Thus $\bar{n}_k$ is the same for all server. And $\sum_{k=1}^{M} \bar{n}_k = N$, so $\bar{n}_k = N/M$. Therefore $\frac{\bar{s}_k}{\eta}\frac{\partial \eta}{\partial \bar{s}_k} = -\frac{\bar{n}_k}{N} = -\frac{1}{M}$.

2.30 *(This problem requires a good knowledge of queueing theory)* Consider an M/M/1 queue with arrival rate λ and service rate μ. The system state is simply the number of customers in the queue; i.e., $\mathbf{n} = n$. The performance measure is the average response time $\bar{\tau} = lim_{L\to\infty}\frac{1}{L}\int_0^{T_L} n(t)dt$. Thus $f(n) = n$. For the M/M/1 queue, there is a source sending customers to the queue with rate λ. Denote the source as server 0, and the server as server 1. Server 0 can be viewed as always having infinitely many customers.

a. Prove that the realization factors $c^{(f)}(n,0)$ and $c^{(f)}(n,1)$, $n = 0, 1, \cdots$, satisfy the following equations:

$$c^{(f)}(0,0) = 0,\ c^{(f)}(0,1) = 0,$$

$$c^{(f)}(n,0) + c^{(f)}(n,1) = n, \qquad n \geq 0,$$

$$(\lambda+\mu)c^{(f)}(n,0) = \mu c^{(f)}(n-1,0) + \lambda c^{(f)}(n+1,0) - \lambda, \qquad n > 0,$$

and

$$(\lambda+\mu)c^{(f)}(n,1) = \lambda c^{(f)}(n+1,1) + \mu c^{(f)}(n-1,1) + \mu, \qquad n > 0.$$

b. To solve for $c^{(f)}(n,i)$, for $i = 0, 1$, we need a boundary condition. Using the physical meaning of perturbation realization, prove that $c^{(f)}(1,1)$ equals the average number of customers served in a busy period of the M/M/1 queue; i.e., (see, e.g., [169])

$$c^{(f)}(1,1) = \frac{\mu}{\mu-\lambda} = \frac{1}{1-\rho}, \qquad \rho = \frac{\lambda}{\mu}.$$

c. Prove

$$c^{(f)}(n,1) = \frac{n}{1-\rho},$$

and

$$c^{(f)}(n,0) = -\frac{n\rho}{1-\rho}.$$

d. By the same argument as in the closed networks, explain and derive

$$\frac{\mu}{\eta^{(I)}}\frac{d\bar{\tau}}{d\mu} = -\frac{\lambda\mu}{(\mu-\lambda)^2} = -\frac{\rho}{(1-\rho)^2},$$

and

$$\frac{\lambda}{\eta^{(I)}}\frac{d\bar{\tau}}{d\lambda} = \frac{\lambda^2}{(\mu-\lambda)^2} = \frac{\rho^2}{(1-\rho)^2}.$$

Solution:

a. When system state is $n = 0$, a perturbation on server 0(perturbation in the mean arrival interval) will propagate thought the system, and the sample path is the same as original one except that it is shift to the right by the amount Δ. So the performance difference between the two sample pathes is $\Delta F_L = f(0)\Delta = 0$. Then $c^{(f)}(0,0) = 0$.

When system state is $n = 0$, a perturbation on server 1 contribute nothing to performance of system, thus $c^{(f)}(0,1) = 0$.

When system state is n, both server 0 and server 1 are perturbed by the amount Δ. Then the next arriving customer and the next leaving customer both delay for Δ. Therefore, the perturbed sampled path is shift to the right by Δ. We have the performance difference $\Delta F_L = f(n)\Delta$. Then the total perburbation realization factor is

$$c^{(f)}(n,0) + c^{(f)}(n,1) = \Delta F_L/\Delta = f(n) = n$$

When system state is n and server 0 is perturbed, the system will transit to next state $n-1$ with probability $\mu/(\mu+\lambda)$. Then, the perturbation is wholly inherited by the new state. The system will transit to the next state $n+1$ with probability $\lambda/(\mu+\lambda)$. In this situation, the system performance will have a decreased amount $(f(n)-f(n+1))\Delta = -\Delta$, besides influence on the new state. Thus the realization factor satisfies

$$c^{(f)}(n,0) = \mu/(\lambda+\mu)c^{(f)}(n-1,0) + \lambda/(\lambda+\mu)[c^{(f)}(n+1,0) - 1]$$

That is $(\lambda+\mu)c^{(f)}(n,0) = \mu c^{(f)}(n-1,0) + \lambda c^{(f)}(n+1,0) - \lambda$.

When system state is n and server 1 is perturbed, the system will transit to the next state $n-1$ with probability $\mu/(\mu+\lambda)$. In this situation, the system performance will have increased amount $(f(n)-f(n-1))\Delta = \Delta$, besides influence on the new state. The system will transit to the next state $n+1$ with probability $\lambda/(\mu+\lambda)$. Then the perturbation is

wholly inherited by the new state. Thus the realization factor satisfies

$$c^{(f)}(n,1) = \mu/(\lambda+\mu)[c^{(f)}(n-1,1)+1] + \lambda/(\lambda+\mu)c^{(f)}(n+1,1)$$

That is $(\lambda+\mu)c^{(f)}(n,1) = \lambda c^{(f)}(n+1,1) + \mu c^{(f)}(n-1,0) + \mu$. The results in a. is proved.

b. When system state is $n = 1$ and server 1 has a perturbation Δ, we can know this perturbation will only exist in the current busy period. Every customer's departure in the busy period is delayed by Δ. So we can know the total perturbation propagation number equals to the number of served customers in the busy period. Denote the number of customers served in the busy period as N_B, then $c^{(f)}(1,1) = \mathrm{E}\{N_B\Delta/\Delta\} = \mathrm{E}\{N_B\} = \bar{N}_B$, where $\bar{N}_B$ is average number of served customers in a busy period. We use the sub-busy period concept to compute $\bar{N}_B$. The period from the time that the system enters state $n, n > 0$, to the first time that the system state is $n-1$ behave statistically similar to a busy period. Such a period is called a sub-busy period. Sub-busy period has the similar statistic properties as busy period, e.g, average number of served customers in a sub-busy period equals to that in a busy period. If $n = 1$, the sub-busy period is a busy period because of memoryless of M/M/1 system. From the physical meaning of sub-busy period, we can get following equation between busy period and sub-busy period:

$$\bar{N}_B = \frac{\mu}{(\lambda+\mu)} \times 1 + \frac{\lambda}{(\lambda+\mu)}[\bar{N}_B + \bar{N}_B]$$

When current state is $n = 1$, the next event is a customer departure with probability $\mu/(\lambda+\mu)$, so it means only one customer in the busy period. The next event is a customer arrival with probability $\lambda/(\lambda+\mu)$. Then the next system state is 2. From now on, the busy period can be divided into two sub-busy periods: the first sub-busy period is from state 2 to 1, the second is from state 1 to 0. The first sub-busy period serves $\bar{N}_B$ customers on average. The second sub-busy period also has average customer number $\bar{N}_B$. From above equation, we can get the average served customer number in a busy period is:

$$\bar{N}_B = \frac{\mu}{\mu-\lambda} = \frac{1}{1-\rho}$$

Therefore the perturbation realization factor is $c^{(f)}(1,1) = \bar{N}_B = \frac{1}{1-\rho}$. b) is proved.

c. When the system state is n, with the similar analysis we have $c^{(f)}(n,1)$ equals the average number of served customers which is counted from state n in a busy period. We

denote it as $\bar{N}_B^n$, with $\bar{N}_B^1 = \bar{N}_B$. From sub-busy period concept, we can also get the following equation:

$$\bar{N}_B^n = \frac{\mu}{\lambda+\mu}(1+\bar{N}_B^{n-1}) + \frac{\lambda}{\lambda+\mu}[\bar{N}_B^n + \bar{N}_B^1]$$

So we get: $\bar{N}_B^n = \bar{N}_B^{n-1} + \frac{\mu}{\mu-\lambda}$. Therefore $\bar{N}_B^n = \frac{n\mu}{\mu-\lambda} = \frac{n}{1-\rho}$. The perturbation realization factor is

$$c^{(f)}(n,1) = \bar{N}_B^n = \frac{n}{1-\rho}$$

From $c^{(f)}(n,0) + c^{(f)}(n,1) = n$, we can know

$$c^{(f)}(n,0) = n - c^{(f)}(n,1) = -\frac{n\rho}{1-\rho}$$

d. The steady state probability $p(n) = (1-\rho)\rho^n$. From the equation of performance derivative, we have

$$\frac{\mu}{\eta^{(I)}}\frac{\mathrm{d}\bar{\tau}}{\mathrm{d}\mu} = -\sum_{n=0}^{\infty} p(n)c^{(f)}(n,1) = -\sum_{n=0}^{\infty}(1-\rho)\rho^n\frac{n}{(1-\rho)} = -\frac{\rho}{(1-\rho)^2}$$

Similarly, we have

$$\frac{\lambda}{\eta^{(I)}}\frac{\mathrm{d}\bar{\tau}}{\mathrm{d}\lambda} = -\sum_{n=0}^{\infty} p(n)c^{(f)}(n,0) = \sum_{n=0}^{\infty}(1-\rho)\rho^n\frac{n\rho}{(1-\rho)} = \frac{\rho^2}{(1-\rho)^2}$$

Reference:

L. Kleinrock, *Queueing Systems*, Volume I, John Wiley, New York,1975.

X. R. Cao, *Realization Probabilities, The Dynamics of Queueing Systems*, Springer-Verg, 1994.

2.31 The head-processing time of a packet in a communication system, or the machine tool set-up time in manufacturing, is usually a fixed amount of time. Consider a two-server cyclic queueing network in which the service times of the two servers are exponentially distributed with mean $\bar{s}_1$ and $\bar{s}_2$, respectively. Suppose that every service time of server 1 increased by a fixed amount of time Δ. Derive the derivative of performance $\eta^{(f)}$ with respect to Δ using performance realization factors $c^{(f)}(\boldsymbol{n}, 1)$.

Solution:

Let $\pi(n)$ be the steady-state probability that there are n customers at server 1. From the PSATA theorem, the probability that the departing customer leaves behind n customers in the system is equal to $\pi(n)$. Let L_1 be the number of service completions at

server 1. The number of the service completions when the state is in $n+1$ is $L_1\pi(n)$. The total time perturbation generated from state $n+1$ at server 1 is $L_1\pi(n)\Delta$. Since each perturbation on average has an effect of $c^{(f)}(n+1,1)$ on F_L. Thus the total effect on F_L of all the perturbations is

$$\begin{aligned}\Delta F_L &\approx \sum_n L_1\pi(n)\Delta c^{(f)}(n+1,1)\\ &= \sum_n L_1\pi(n)\Delta c^{(f)}(n+1,1).\end{aligned}$$

So, we have

$$\frac{L}{L_1}\frac{\Delta F_L/L}{\Delta} \approx \sum_n \pi(n)c^{(f)}(n+1,1).$$

Letting $L\to\infty$ and $\Delta\to 0$, we have

$$\frac{1}{v_1}\frac{\partial\eta^f}{\partial\Delta} = \sum_n \pi(n)c^{(f)}(n+1,1),$$

where v_1 is the visit ratio of server 1. v_1 can be easily obtained by (C.5) for cyclic network, i.e. $v_1 = v_2$. Since $v_1+v_2=1$, we have $v_1 = v_2 = 1/2$. Thus, we can obtain the derivative with respect to fixed change Δ as follows:

$$\frac{\partial\eta^f}{\partial\Delta} = \frac{1}{2}\sum_n \pi(n)c^{(f)}(n+1,1).$$

2.32 Prove that Algorithm 2.2 yields a strongly consistent estimate for the sensitivity of the mean response time in an M/G/1 queue; i.e., in (2.134) we have

$$\frac{\mu}{\eta^{(I)}}\frac{\partial\bar{\tau}}{\partial\mu} = -\lim_{K\to\infty}\frac{1}{T_L}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{i} s_{k,l} \quad a.s.$$

Solution: Firstly, we consider the problem for a general $M/G/1$ queueing system. Then we consider the problem for the special $M/G/1$ queueing system with service time $s = \frac{C}{\mu}$ in the above problem, where C denotes a fixed "length" of a customer, such as a packet length in communication systems, and μ denotes the service rate.

The mean response time of a customer is given by Pollaczek-Khinchin, or P-K formula in Kleinrock (1975),

$$\bar{\tau} = E(s) + \frac{\lambda E(s^2)}{2(1-\lambda E(s))}$$

where s denotes the service time of a customer, which follows a general distribution $F(s,\mu)$. The sensitivity of the average response time to μ is then obtained by straightforward differentiation to be

$$\frac{d\bar{\tau}}{d\mu} = \frac{dE(s)}{d\mu} + \frac{\lambda}{2(1-\lambda E(s))}\frac{dE(s^2)}{d\mu} + \frac{\lambda^2 E(s^2)}{2(1-\lambda E(s))^2}\frac{dE(s)}{d\mu}.$$

For the $M/G/1$ queue with determined service time, we have $E(s) = \frac{C}{\mu}$ and $E(s^2) = \frac{C^2}{\mu^2}$, so, the sensitivity of the average response time to μ is

$$\frac{d\bar{\tau}}{d\mu} = -\frac{C}{\mu^2} - \frac{\lambda}{1-\lambda E(s)}\frac{C^2}{\mu^3} - \frac{\lambda^2 E(s^2)}{2(1-\lambda E(s))^2}\frac{C}{\mu^2}$$

Let $C_{k,i}$ denote the i-th customer in the k-th busy period and $s_{k,i}(\omega,\mu)$ denote the service time of the i-th customer in the k-th busy period. For the general $M/G/1$ queueing system, we assume that the derivative $\frac{ds_i}{d\mu}$ of sample function $s_{k,i}(\omega,\mu)$ with respect to μ depends only on the value of the service time $s_{k,i}$, i.e. $\frac{ds_{k,i}}{d\mu} = \phi(s_{k,i})$. For the special $M/G/1$ queue with service time $s_{k,i} = \frac{C}{\mu}$, we have $\frac{ds_{k,i}}{d\mu} = -\frac{s_{k,i}}{\mu}$, which satisfies our assumption.

Next, we prove the following results about a strongly consistent estimate of the above sensitivity.

$$\frac{\partial\bar{\tau}}{\partial\mu} = \lim_{K\to\infty}\frac{1}{L}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{i}\phi(s_{k,l}) =: \frac{d\bar{\tau}}{d\mu}|_{est} \quad a.s. \tag{2.26}$$

where $s_{k,l}$ denotes the service time of the lth customer in the k-th busy period, L denotes the number of customer completions in the period of $[0,T_L]$, and n_k denotes the number of customers served in the kth busy period. Let $h_k = \sum_{i=1}^{n_k}\sum_{l=1}^{i}\phi(s_{k,l})$, then the right hand side of (2.26) can be written as

$$\frac{d\bar{\tau}}{d\mu}\Big|_{est} = \lim_{K\to\infty}\frac{1}{L}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{i}\phi(s_{k,l}) = \lim_{K\to\infty}\frac{\frac{1}{K}\sum_{k=1}^{K}h_k}{\frac{1}{K}\sum_{k=1}^{K}n_k},$$

From the strong law of large number, we have

$$\frac{d\bar{\tau}}{d\mu}\Big|_{est} = \lim_{K\to\infty}\frac{\frac{1}{K}\sum_{k=1}^{K}h_k}{\frac{1}{K}\sum_{k=1}^{K}n_k} = \frac{E[h_1]}{E[n_1]}, \quad w.p.1. \tag{2.27}$$

From Kleinrock (1975, page 217) (cf. Problem C. 2), we have

$$E[n_1] = \frac{1}{1-\lambda E(s)}.$$

Thus,

$$\frac{d\bar{\tau}}{d\mu}|_{est} = [1 - \lambda E(s)]E[h_1]. \tag{2.28}$$

Following the approach in Kleinrock (1975) for busy period analysis, we decompose the busy period into sub-busy periods.

From (2.27), we only need to analyze the first busy period, therefore, we will omit the subscript k about busy period. We assume m customers arrive during the service time of C_1. As explained in Kleinrock (1975), each of C_2 through C_{m+1} initiates a sub-busy period statistically identical to the busy period initiated by C_1. Furthermore, these sub-busy periods are statistically identical. Let us define the quantity

$$g = \sum_{l=1}^{n_1} \phi(s_l),$$

Now, we will number the sub-busy periods in the order that they occur, and we will number the customers in the order that they are served using the LCFS discipline. Let m_r be the number of customers in the r-th sub-busy period and define $m^{(r)} = 1+m_1+\ldots+m_r$ with $m^{(0)} = 1$. Then $C_{m^{(r-1)}+1}$ through $C_{m^{(r)}}$ are the customers that belong to the r-th sub-busy period. We consider the quantities

$$g^{(r)} = \sum_{i=1}^{m_r} \phi(s_{m(r-1)+i}) \quad \text{with} \quad g^{(0)} = \phi(s_1),$$

$$h^{(r)} = \sum_{i=1}^{m_r} \sum_{j=1}^{i} \phi(s_{m(r-1)+j}).$$

where s_i denote the ith customer in the first busy period. Then, we have

$$g = \sum_{l=1}^{n_1} \phi(s_l) = \phi(s_1) + \sum_{r=1}^{m} \sum_{i=1}^{m_r} \phi(s_{m^{(r-1)}} + i) = \sum_{r=0}^{m} g^{(r)}, \tag{2.29}$$

$$h_1 = \phi(s_1) + \sum_{r=1}^{m} [h^{(r)} + m_r \sum_{s=0}^{r-1} g^{(s)}]. \tag{2.30}$$

Next, we wish to derive the expected values of g and h_1. Taking expected values on both sides of (2.29), we have

$$E(g) = E(\phi(s_l)) + E(m)E(g).$$

Noting that $E(m)$ equals $\lambda E(s)$ and solving for $E(g)$ gives

$$E(g) = \frac{E(\phi(s_1))}{1 - \lambda E(s)}. \tag{2.31}$$

Taking expected values in (2.30) conditioned on s_1 and m, we have

$$E(h_1|s_1,m) = \phi(s_1) + \sum_{r=1}^{m} E[h^{(r)}|s_1,m] + \sum_{r=1}^{m} E[m_r \sum_{s=0}^{r-1} g^{(s)}|s_1,m].$$

Taking into account the fact that quantities referring to different sub-busy periods are independent from each other and from s_1 and m and identically distributed to the parent busy period and also that m_r is independent of $g^{(s)}$ for $s < r$, we get

$$\begin{aligned} E(h_1|s_1,m) &= \phi(s_1) + mE(h_1) + kE(m_1)\phi(s_1) \\ &\quad + \Big(E[m_2] + 2E[m_3] + \ldots + (m-1)E[m_k]\Big)E(g) \\ &= \phi(s_1) + m[E(h) + E(m_1)\phi(x_1)] + E(m_r)E(g)(m^2-m)/2. \quad (2.32) \end{aligned}$$

Taking expectations with respect to m in the above equation conditioned on s_1. Then $E(m|s_1)$ is the average number of Poisson arrivals in an interval of length s_1 and so it is equal to λs_1. Similarly, $E(m^2|s_1)$ is equal to $\lambda s_1 + (\lambda s_1 - 1)^2$. Taking also into account that m_r is identically distributed to n_1 (the number of customers in the parent busy period) and thus $E(m_r) = E(n_1) = 1/(1-\lambda E(s))$, and using (2.31), we get

$$E(h_1|s_1) = \phi(s_1) + \lambda s_1[E(h_1) + \phi(s_1)/(1-\lambda E(s))] + (\lambda s_1)^2 E(\phi(s_1))/2(1-\lambda E(s))^2 \quad (2.33)$$

Taking expectation with respect to s_1, we get

$$E(h_1) = E(\phi(s))/(1-\lambda E(s)) + \lambda E(s\phi(s))/(1-\lambda E(s))^2 + \lambda^2 E(s^2)E(\phi(s))/2(1-\lambda E(s))^3 \quad (2.34)$$

Using the exchangeability of expectation and differentiation and (2.28), we get

$$\frac{d\bar{\tau}}{d\mu}|_{est} = \frac{dE(s)}{d\mu} + \frac{\lambda}{2(1-\lambda E(s))}\frac{dE(s^2)}{d\mu} + \frac{\lambda^2 E(s^2)}{2(1-\lambda E(s))^2}\frac{dE(s)}{d\mu}, \quad a.s.$$

Therefore, (2.26) has been proved.

For the special $M/G/1$ queueing system in problem 2.31, the proof is simple. Letting $ds_i = -s_i/\mu$, the result in problem 2.31 can be obtained.

Reference:

1. L. Kleinrock, *Queueing Systems*, Volume I, John Wiley, New York,1975.

2. R. Suri and M. A. Zananis, Perturbation Analysis Gives Strongly Consistent Sensitivity Estimates for the M/G/1 Queue, *Management Science*, Vol. 34, No. 1, 39-64, 1988.

2.33 Consider a closed Jackson network with M servers and N customers. The throughput of server i is $\eta_i = \breve{\eta} v_i$ where $\breve{\eta}$ is the "un-normalized system throughput":

$$\breve{\eta} = \frac{G_M(N-1)}{G_M(N)},$$

where v_i is server i's visiting ratio: the solution to

$$v_i = \sum_{j=1}^{M} q_{j,i} v_j, \qquad j = 1, 2, \cdots, M,$$

and (see (A.55) in the Appendix)

$$G_m(n) = \sum_{n_1+\cdots+n_M=n} \prod_{i=1}^{m} x_i^{n_i},$$

where $x_i = v_i \bar{s}_i$, $i = 1, 2 \cdots, M$. We have

$$dx_i = dv_i \bar{s}_i + v_i d\bar{s}_i. \tag{2.35}$$

Now we consider the derivative of $\breve{\eta}$ with respect to the routing probability matrix $Q = [q_{i,j}]_{i,j=1}^{M}$. It is clear that $\breve{\eta}$ depends on the routing probabilities only through x_i, $i = 1, 2, \cdots, M$. Suppose that v_i changes to $v_i + dv_i$, $i = 1, 2, \cdots, M$. From (2.35), we observe that in terms of the changes in x_i, dx_i, $i = 1, 2, \ldots, M$, this is equivalent to setting $dv_i = 0$ and $d\bar{s}_i = \bar{s}_i \frac{dv_i}{v_i}$ for all $i = 1, 2, \cdots, M$.

a. Explain that for closed Jackson networks, the derivative of any steady-state performance $\sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) f(\boldsymbol{n})$ with respect to the changes in routing probabilities can be obtained through the derivatives of the performance with respect to mean service times.

b. Derive the performance derivative formula $\frac{d\eta_i}{dQ}$, by using performance realization factors $c(\boldsymbol{n}, i)$, $i = 1, 2, \cdots$.

Solution:

a. For the closed Jackson networks, from (A.44), the steady state probability $\pi(\boldsymbol{n})$ depends on the routing probabilities and $\bar{s}_i$ only through $x_i = v_i \bar{s}_i$, $i = 1, 2, \cdots, M$. The changes of routing probability matrix Q will lead to the change of v_i. From (2.35), suppose that v_i changes to $v_i + dv_i$, $i = 1, 2, \cdots, M$, and $\bar{s}_i$ does not change, we know the change

of x_i is $dx_i = \bar{s}_i dv_i$. The change can be equivalent to setting $dv_i = 0$ and $d\bar{s}_i = \bar{s}_i \frac{dv_i}{v_i}$ for all $i = 1, 2, \cdots, M$. That is, the effect on x_i of the change of v_i is equivalent to that of the change of the service rate s_i. In other words, the changes of routing probability can be equivalent to the changes of mean service times. Therefore, the derivative of any steady-state performance $\sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) f(\boldsymbol{n})$ with respect to the changes in routing probabilities can be obtained through the derivatives of the performance with respect to mean service times.

b. Suppose that Q changes to $Q_\delta = Q + \delta \Delta Q$, where $\Delta Q = Q' - Q$. From $v_\delta Q_\delta = v_\delta$, we have $v_\delta(I - Q_\delta) = 0$, which is similar to $\pi_\delta P_\delta = \pi_\delta$. Taking derivative of both sides with respect to δ, we have

$$\frac{dv_\delta}{d\delta}[I - Q_\delta] = v_\delta \Delta Q.$$

Then,

$$\frac{dv}{d\delta}|_{\delta=0} = v \Delta Q Q^{\#},$$

where $Q^{\#}$ is the group inverse of $I - Q$. Then we can compute the value of $\frac{dv_i}{d\delta}$.

Similarly to the discussion in Part a), since $\breve{\eta}$ depends on the routing probabilities and $\bar{s}_i$ only through $x_i = v_i \bar{s}_i$, $i = 1, 2, \cdots, M$, we know the changes of routing probabilities are equivalent to the changes of service rates. Thus, we consider the changes of service rates.

$$\frac{d\breve{\eta}}{d\delta} = \sum_{i=1}^{M} \frac{d\breve{\eta}}{d\bar{s}_i} \frac{d\bar{s}_i}{d\delta} = \sum_{i=1}^{M} \frac{d\breve{\eta}}{d\bar{s}_i} \frac{\bar{s}_i}{v_i} \frac{d\bar{v}_i}{d\delta},$$

where we have used $d\bar{s}_i = \bar{s}_i \frac{dv_i}{v_i}$. $\frac{d\breve{\eta}}{d\bar{s}_i}$ is the derivative with respect to service rate, then we can compute it by using perturbation analysis. Since the throughput $\eta_i = v_i \breve{\eta}, i = 1, 2, \cdots$, where η_i is the throughput of server i, we have $\eta := \sum_{i=1}^{M} \eta_i = \breve{\eta} \sum_{i=1}^{M} v_i$, where η is the throughput of the network. Then, $\frac{d\breve{\eta}}{d\bar{s}_i} = \frac{1}{\sum_{i=1}^{M} v_i} \frac{d\eta}{d\bar{s}_i} = -\frac{1}{\sum_{i=1}^{M} v_i} \frac{\eta}{\bar{s}_i} \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) c(\boldsymbol{n}, i)$. Thus, we have

$$\frac{d\breve{\eta}}{d\delta} = -\frac{1}{\sum_{i=1}^{M} v_i} \sum_{i=1}^{M} \frac{\eta}{v_i} \frac{dv_i}{d\delta} \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) c(\boldsymbol{n}, i).$$

From $\eta_i = v_i \breve{\eta}$, we have $\frac{d\eta_i}{dQ} = \frac{dv_i}{d\delta} \breve{\eta} + v_i \frac{d\breve{\eta}}{d\delta}$. Therefore, we have

$$\frac{d\eta_i}{dQ} = \frac{dv_i}{d\delta} \breve{\eta} - \frac{\eta}{\sum_{i=1}^{M} v_i} \sum_{i=1}^{M} \frac{dv_i}{d\delta} \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) c(\boldsymbol{n}, i).$$

2.34 Consider the same two-server cyclic Jackson queueing network studied in Problem 2.17. Let $\eta_T^{(f)} = \lim_{L\to\infty} \frac{\int_0^{T_L} f(n(t))dt}{T_L}$ denote the time-average performance, where $n(t)$ is the number of customers at time t at server 1, L denote the transition numbers and performance function $f(n) = n$. Suppose the arrival rate λ and the service rate μ change only when the state is n.

a. Derive $\frac{d\eta_T^{(f)}}{d\lambda}$ and $\frac{d\eta_T^{(f)}}{d\mu}$ in terms of the realization factors $c^{(f)}(n,1), c^{(f)}(n,2)$ and realization probability $c(n,1), c(n,2)$.

b. Express $\frac{d\eta_T^{(f)}}{d\lambda}$ and $\frac{d\eta_T^{(f)}}{d\mu}$ in terms of the performance potentials $g(n)$.

c. Compare both results in a) and b) and derive a relation between the realization factors and the potentials. Give an intuitive explanation for this relation. (cf. [260])

Solution:

a. Since

$$\begin{aligned}\eta_T^{(f)} &= \lim_{L\to\infty} \frac{\int_0^{T_L} f(n(t))dt}{T_L} \\ &= \lim_{L\to\infty} \frac{L}{T_L} \frac{\int_0^{T_L} f(n(t))dt}{L} \\ &= \bar{\tau}/\eta^{(I)},\end{aligned}$$

where $\eta^{(I)} = \lim_{L\to\infty} \frac{T_L}{L}$ and $\bar{\tau} = \lim_{L\to\infty} \frac{\int_0^{T_L} n(t)dt}{L}$, which is the average response time of each customer at server 1, we have

$$\begin{aligned}\frac{d\eta_T^{(f)}}{d\mu} &= \frac{\frac{d\bar{\tau}}{d\mu}\eta^{(I)} - \bar{\tau}\frac{d\eta^{(I)}}{d\mu}}{(\eta^{(I)})^2} \\ &= \frac{1}{\eta^{(I)}}\left[\frac{d\bar{\tau}}{d\mu} - \frac{d\eta^{(I)}}{d\mu}\eta_T^{(f)}\right]. \qquad (2.36)\end{aligned}$$

Next, we obtain the derivatives $\frac{d\bar{\tau}}{d\mu}$ and $\frac{d\eta^{(I)}}{d\mu}$ in term of realization factors $c^{(f)}(n,1)$ and $c(n,1)$, respectively, by using the method in in Section 2.4.3. (These derivatives can be directly from the results of Perturbation Analysis.)

Let $\pi(n)$ be the steady-state probability of state n. Consider a time period $[0, T_L]$ with $L \gg 1$. The length of the total time that the system is in state n in $[0, T_L]$ is $T_L\pi(n)$. The

total perturbation generated in this period at serve 1 due to the change of μ is $T_L\pi(n)\frac{\Delta s_1}{s_1}$, where s_1 is the average service time of server 1, i.e., $s_1 = \frac{1}{\mu}$, and thus $\Delta s_1 \approx -\frac{1}{\mu^2}\Delta\mu$. Since such perturbation on average has an effect of $c^{(f)}(n,1)$ on F_L, the overall effect on F_L of the perturbation is $-T_L\pi(n)\frac{\Delta\mu}{\mu}c^{(f)}(n,1)$, thus we have

$$\Delta F \approx -T_L\pi(n)\frac{\Delta\mu}{\mu}c^{(f)}(n,1).$$

From this, we have

$$\frac{\Delta F_L/L}{\Delta\mu} = \frac{T_L}{\mu L}\pi(n)c^{(f)}(n,1).$$

Letting $L \to \infty$ and $\Delta\mu \to 0$, we obtain

$$\frac{d\bar{\tau}}{d\mu} = -\frac{\eta^{(I)}}{\mu}\pi(n)c^{(f)}(n,1). \tag{2.37}$$

If performance function $f \equiv 1$, we have

$$\frac{d\eta^{(I)}}{d\mu} = -\frac{\eta^{(I)}}{\mu}\pi(n)c(n,1). \tag{2.38}$$

Putting (2.38) and (2.37) into (2.36), we have

$$\frac{d\eta_T^{(f)}}{d\mu} = -\frac{1}{\mu}\pi(n)\left[c^{(f)}(n,1) - c(n,1)\eta_T^{(f)}\right]. \tag{2.39}$$

Similarly, we can obtain

$$\frac{d\eta_T^{(f)}}{d\lambda} = -\frac{1}{\lambda}\pi(n)\left[c^{(f)}(n,2) - c(n,2)\eta_T^{(f)}\right]. \tag{2.40}$$

b. From the potential theory, we have

$$\frac{d\eta_T^{(f)}}{d\mu} = \pi\frac{dB}{d\mu}g,$$

where

$$B = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & \cdots & \mu & -(\mu+\lambda) & \lambda \\ 0 & 0 & \cdots & \mu & -\mu \end{bmatrix}.$$

If μ changes only when the state is $n, n = 1, 2 \ldots, N$, we have

$$\frac{dB}{d\mu} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 0 & \cdots & 1 & -1 & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Thus, we have

$$\frac{d\eta_T^{(f)}}{d\mu} = \pi(n)\left[g(n-1) - g(n)\right], \qquad n = 1, 2, \ldots, N. \tag{2.41}$$

Similarly, we can obtain

$$\frac{d\eta_T^{(f)}}{d\lambda} = \pi(n)\left[g(n+1) - g(n)\right], \qquad n = 0, 1, \ldots, N-1. \tag{2.42}$$

c. Comparing (2.39) and (2.41), we have

$$-\frac{1}{\mu}\pi(n)\left[c^{(f)}(n,1) - c(n,1)\eta_T^{(f)}\right] = \pi(n)\left[g(n-1) - g(n)\right].$$

That is,

$$c^{(f)}(n,1) - c(n,1)\eta_T^{(f)} = \mu\left[g(n) - g(n-1)\right], \qquad n = 1, 2, \ldots, N.$$

Comparing (2.40) and (2.42), we have

$$-\frac{1}{\lambda}\pi(n)\left[c^{(f)}(n,2) - c(n,2)\eta_T^{(f)}\right] = \pi(n)\left[g(n+1) - g(n)\right].$$

That is

$$c^{(f)}(n,2) - c(n,2)\eta_T^{(f)} = \lambda\left[g(n) - g(n+1)\right], \qquad n = 0, 1, \ldots, N-1.$$

An intuitive explanation:

$$\begin{aligned} c^f(n,1) &= \lim_{L\to\infty,\Delta\to 0}\frac{1}{\Delta}E\{\int_0^{T_L'} f(n'(t))dt - \int_0^{T_L} f(n(t))dt\} \\ &= \lim_{L\to\infty,\Delta\to 0}\frac{1}{\Delta}E\{\int_0^{T_L}[f(n'(t)) - f(n(t))]dt\} \\ &\quad + \lim_{L\to\infty,\Delta\to 0}\frac{1}{\Delta}E\int_{T_L}^{T_L'} f(n'(t))dt. \end{aligned} \tag{2.43}$$

Firstly, we consider the first term.

$$\begin{aligned} & \lim_{L\to\infty,\Delta\to 0}\frac{1}{\Delta}E\{\int_0^{T_L}[f(n'(t))-f(n(t))]dt\} \\ = & \lim_{L\to\infty,\Delta\to 0}\frac{1}{\Delta}E\{\int_0^{\Delta}[f(n'(t))-f(n(t))]dt \\ & +\lim_{L\to\infty,\Delta\to 0}\frac{1}{\Delta}E\{\int_{\Delta}^{T_L}[f(n'(t))-f(n(t))]dt. \end{aligned} \tag{2.44}$$

Since the $n(t)$ is a Markov process with infinitesimal matrix $A=[a_{n,m}]$, where

$$a_{n,m}=\begin{cases} \epsilon(N-n)\lambda, & m=n+1, \\ \epsilon(n)\mu, & m=n-1, \\ -\epsilon(N-n)\lambda-\epsilon(n)\mu, & m=n, \\ 0, & others. \end{cases} \qquad i,j=1,2.$$

From Kolmogorov theorem, we have $P(t)=e^{At}$. Thus, we can obtain the probability that the original process moves from state n to state $(n+1)$ at time Δ is approximately equal to $\epsilon(N-n)\lambda\Delta$, the probability that the original process moves from state n to state $(n-1)$ is approximately equal to $\epsilon(n)\mu\Delta$ and the probability that the original process moves from state n to state n is approximately equal to $1-\epsilon(N-n)\lambda\Delta-\epsilon(n)\mu\Delta$, where we have omitted the higher-order terms of Δ . For the perturbed Markov process, since the service time was delayed Δ, we know the probability that the perturbed Markov process moves from state n to state $n-1$ at time Δ is zero, the probability that the perturbed Markov process moves from state n to state $n+1$ at time Δ is $\epsilon(N-n)\lambda\Delta$ and the probability that the perturbed Markov process moves from state n to state n at time Δ is $1-\epsilon(N-n)\lambda\Delta$.

So, we know that the probability that the original process and the perturbed process transit to different state at time Δ is the same order infinitesimal of Δ. On this basis, since $|f(n'(t)-f(n(t))|$ is bounded, we have

$$\lim_{L\to\infty,\Delta\to 0}\frac{1}{\Delta}E\{\int_0^{\Delta}[f(n'(t))-f(n(t))]dt=0. \tag{2.45}$$

For the second term in (2.44), we consider $\int_{\Delta}^{T_L}[f(n'(t))-f(n(t))]dt$ from the point view of perturbation realization factor of Markov process.

$$\int_{\Delta}^{T_L}[f(n'(t))-f(n(t))]dt$$

$$\begin{aligned} &\approx \epsilon(N-n)\lambda\Delta d(n+1,n) + \epsilon(n)\mu\Delta d(n-1,n) + \epsilon(N-n)\lambda\Delta d(n,n+1) \\ &= \epsilon(n)\mu\Delta(g^f(n) - g^f(n-1)). \end{aligned}$$

where we have used $d(n+1,n) = -d(n,n+1)$. Thus, we have

$$\lim_{L\to\infty,\Delta\to 0} \frac{1}{\Delta} E\{\int_{\Delta}^{T_L} [f(n'(t)) - f(n(t))]dt = \epsilon(n)\mu(g^f(n) - g^f(n-1)). \tag{2.46}$$

From (2.45) and (2.46), we know the first term in (2.43) is equal to $\epsilon(n)\mu(g^f(n)-g^f(n-1))$. For the second term in (2.43), when L is large enough, we have $E[f(n'(t))] = \eta_T^{(f)}$, thus,

$$\lim_{L\to\infty,\Delta\to 0} \frac{1}{\Delta} E\int_{T_L}^{T'_L} f(n'(t))dt = \eta \lim_{\Delta\to 0} \frac{T'_L - T_L}{\Delta} = c(n,1)\eta_T^{(f)}. \tag{2.47}$$

Thus, we have $c^f(n,1) = \epsilon(n)\mu(g^f(n) - g^f(n-1)) + c(n,1)\eta_T^{(f)}$. Similarly, we can intuitively obtain $c^f(n,2) = \epsilon(N-n)\lambda(g^f(n) - g^f(n+1)) + c(n,2)\eta_T^{(f)}$.

2.35 In weak derivative expression (2.125), we may choose $P^+ = P'$ and $P^- = P$.

a. Derive (2.126) and express its meaning based on sample paths.

b. Derive (2.127).

Solution:

a. If we choose $P^+ = P'$ and $P^- = P$, then $c(i) = 1$. (2.124) becomes

$$\begin{aligned} \frac{d\eta_\delta}{d\delta} &= \pi(P' - P)\sum_{l=0}^{\infty} P^l f \\ &= \sum_{i\in\mathcal{S}} \pi(i)\sum_{l=0}^{\infty}(p'_i P^l f - p_i P^l f). \end{aligned}$$

Thus we have

$$\begin{aligned} \frac{d\eta_\delta}{d\delta} &= \sum_{i=1}^{\mathcal{S}} \pi(i)\sum_{l=0}^{\infty} E[f(X'_l) - f(X_l)|X'_0 = i, X_0 = i] \\ &= \sum_{i=1}^{\mathcal{S}} \pi(i)\sum_{l=0}^{L^*} E[f(X'_l) - f(X_l)|X'_0 = i, X_0 = i]. \end{aligned}$$

The meaning based on the sample path: On the sample path, the state is state i, at which the first jump from X'_0 to X'_1 follows transition probability vector p'_i and the first

jump from X_0 to X_1 follows transition probability p_i. The rest transitions of $X'_l, l \geq 1$ and $X_l, l \geq 1$ all follow transition probability matrix P. In fact, this is similar to the perturbation at state i.

b. Since

$$\begin{aligned}
&\sum_{l=0}^{L^*} E[f(X'_l) - f(X_l)|X'_0 = i, X_0 = i] \\
= \; & E\{\sum_{l=1}^{L^*} E[f(X'_l) - f(X_l)|X'_1, X_1, X'_0 = i, X_0 = i]|X'_0 = i, X_0 = i\} \\
= \; & \sum_{j' \in \mathcal{S}, j \in \mathcal{S}} p(X'_1 = j', X_1 = j|X'_0 = i, X_0 = i) \sum_{l=1}^{L^*} E[f(X'_l) - f(X_l)|X'_1 = j', X_1 = j] \\
= \; & \sum_{j' \in \mathcal{S}, j \in \mathcal{S}} p'(j'|i)p(j|i)\gamma(j, j'),
\end{aligned}$$

thus we have

$$\frac{d\eta_\delta}{d\delta} = \sum_{i=1}^{\mathcal{S}} \pi(i) \sum_{j' \in \mathcal{S}, j \in \mathcal{S}} p'(j'|i)p(j|i)\gamma(j, j').$$

2.36 Derive (2.23) from (2.127).

Solution: Since

$$p^+(j|i) = \begin{cases} \frac{1}{c(i)} \max\{\Delta p(j|i), 0\} & if \; c(i) > 0 \\ 0 & if \; c(i) = 0. \end{cases}$$

$$p^-(j|i) = \begin{cases} \frac{1}{c(i)} \max\{-\Delta p(j|i), 0\} & if \; c(i) > 0 \\ 0 & if \; c(i) = 0. \end{cases}$$

and $\sum_{j \in \mathcal{S}} \Delta p(j|i) = 0$, we have $\sum_{j \in \mathcal{S}} p^+(j|i) = 1$ and $\sum_{j \in \mathcal{S}} p^-(j|i) = 1$. From (2.127), we have

$$\begin{aligned}
\frac{d\eta_\delta}{d\delta} &= \sum_{i=1}^{S} \pi(i)c(i) \sum_{j_1, j_2 = 1}^{S} \gamma(j_1, j_2)p^-(j_1|i)p^+(j_2|i) \\
&= \sum_{i=1}^{S} \pi(i)c(i) \sum_{j_1, j_2 = 1}^{S} p^-(j_1|i)p^+(j_2|i)[g(j_2) - g(j_1)] \\
&= \sum_{i=1}^{S} \pi(i)c(i) \sum_{j \in \mathcal{S}} [p^+(j|i) - p^-(j|i)]g(j)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{S} \pi(i) \sum_{j \in \mathcal{S}} [p'(j|i) - p(j|i)] g(j) \\
&= \pi \Delta P g.
\end{aligned}
$$

2.37 Consider a (continuous-time) Markov process with transition rates $\lambda(i)$ and transition probabilities $p(j|i)$, $i, j = 1, 2, \ldots, S$. Suppose that the transition probability matrix $P := [p(j|i]_{i,j \in \mathcal{S}}$ changes to $P + \delta \Delta P$ and the transition rates $\lambda(i)$, $i = 1, 2, \ldots, S$ remain unchanged. Let η be the average reward with reward function f. Derive the performance derivative formula for $\frac{d\eta_\delta}{d\delta}$ using the construction approach illustrated in Section 2.1.3.

Solution: To derive the performance derivative formula $\frac{d\eta_\delta}{d\delta}$, we consider a sample path $\boldsymbol{X}$ with infinitesimal generator $B = [b(i,j)]$ consisting of $L >> 1$ transitions, where

$$
b(i,j) = \begin{cases} -\lambda(i) & if\ i = j \\ \lambda(i) p(j|i) & if\ i \neq j \end{cases}
$$

Among these transitions, on the average the time that the process stays at state i is $T\pi(k)$, where $\pi = (\pi(1), \cdots, \pi(S))$ is the steady-state probability of continuous-time Markov process. Since the average holding time at state is $\frac{1}{\lambda(k)}$, then there are $T\pi(k)\lambda(k)$ times from state k on the average. Each time when $\boldsymbol{X}$ visits state i after visiting state k, because of the change from P to $P_\delta = p + \delta \Delta P$. the perturbed path $\boldsymbol{X}_\delta$ may have a jump, denoted as from state i to j. Denote the probability of a jump from i to j after visiting state k as $p(i,j|k)$. Then, we have

$$
\sum_{j=1}^{S} p(i,j|k) = p(i|k), \tag{2.48}
$$

$$
\sum_{i=1}^{S} p(i,j|k) = p_\delta(j|k). \tag{2.49}
$$

On the average, in the time interval $[0, T)$ there are $T\pi(k)\lambda(k)p(i,j|k)$ jumps from i to j on the sample path. Each such jump has on the average an effect of $\gamma(i,j)$ on F_L. Thus, on the average the total effect on F_L due to the change in P to P_δ is

$$
\begin{aligned}
& E(F_{\delta,T} - F_T) \\
:= \ & E\{\int_0^T f(X_{\delta,t}) dt - \int_0^T f(X_t) dt\}
\end{aligned}
$$

$$\begin{aligned} &\approx \sum_{k=1}^{S}\{\sum_{i,j=1}^{S} T\pi(k)\lambda(k)p(i,j|k)\gamma(i,j)\} \\ &= \sum_{k=1}^{S}\{\sum_{i,j=1}^{S} T\pi(k)\lambda(k)p(i,j|k)[g(j)-g(i)]\}. \end{aligned}$$

From (2.48) and (2.49), we have

$$\begin{aligned} & E(F_{\delta,T}-F_T) \\ \approx\ & T\sum_{k=1}^{S}\pi(k)\lambda(k)\sum_{j=1}^{S}[p_\delta(j|k)-p(j|k)]g(j) \\ =\ & T\pi\Lambda[P_\delta-P]g = T\pi\Lambda(\Delta P)\delta g, \end{aligned}$$

where $\Lambda = diag\{\lambda(1),\cdots,\lambda(S)\}$. Thus,

$$\eta_\delta - \eta = \frac{1}{T}E(F_{\delta,T}-F_T) \approx \pi\Lambda(\Delta P)\delta g.$$

Finally, we obtain the performance derivative formula

$$\frac{d\eta_\delta}{d\delta} = \pi\Lambda(\Delta P)g.$$

3

Solutions to Chapter 3

3.1 Study the potential with $g(S) = 0$:

a. Prove that the solution to (3.4) satisfies $p_{S*}g = \eta - f(S)$.

b. Derive (3.4) from the Poisson equation $(I - P)g + \eta e = f$ with the normalization condition $p_{S*}g = \eta - f(S)$.

Solution:

a. Putting $P_{-} = P - ep_{S*}$ into (3.4), we have

$$g = Pg - ep_{S*}g + f_{-}.$$

Multiplying the both sides of the above equation with π, we have

$$\pi g = \pi Pg - \pi ep_{S*}g + \pi f_{-}.$$

Using $\pi P = \pi$ and $\pi e = e$, we have

$$p_{S*}g = \pi f_{\_} = \sum_{i=1}^{S} \pi(i)[f(i) - f(S)] = \eta - f(S).$$

b. Since $p_{S*}g = \eta - f(S)$, we have $ep_{S*}g = \eta e - f(S)e$. That is, $\eta e - f(S)e - ep_{S*}g = 0$. From the Poisson equation and the above equation, we have

$$g = Pg - \eta e + f = Pg - \eta e + f + \eta e - f(S)e - ep_{S*}g = P_{\_}g + f_{\_},$$

which is Equation (3.4).

3.2 Let P be an $S \times S$ ergodic stochastic transition matrix and ν be an S dimensional (row) vector with $\nu e = 1$. Set $P_{-\nu} = P - e\nu$.

a. Suppose that there is a potential g such that $\nu g = \eta$, prove $g = P_{-\nu}g + f$.

b. Prove that the eigenvalues of $P - e\nu$ are 0 and $\lambda_i, i = 1, 2, \cdots, S-1$, where λ_i, with $|\lambda_i| < 1, i = 1, 2, \cdots, S-1$, are the eigenvalues of P.

c. Develop an iterative algorithm similar to (3.7).

d. For any vecotr ν with $\nu e = 1$, we can develop the algorithm in c) without presenting $\nu g = \eta$. Prove that the potential obtained by the algorithm indeed satisfies $\nu g = \eta$.

e. Prove that the algorithm (3.4)-(3.7) is a special case of the above algorithm and verify $p_{s*}g = \eta$.

Solution:

a. From Poisson equation and $\nu g = \eta$, we have

$$g - Pg + e\nu g = f.$$

That is, $g = P_{-\nu}g + f$.

b. Since $(P - e\nu)e = 0$, we know 0 is an eigenvalue of $P - e\nu$. Let x_i be the eigenvector of P corresponding to eigenvalue $\lambda_i \neq 0, 1$, i.e., $Px_i = \lambda_i x_i$. Define $x_i' = x_i - \frac{1}{\lambda_i} e\nu x_i$, since the eigenvalue of ev is 1 and 0 and 0 is an eigenvalue of $S-1$ multiplicity, we have $x_i' \neq 0$. Moreover,

$$(P - e\nu)x_i' = \lambda_i(x_i - \frac{1}{\lambda_i} e\nu x_i) = \lambda_i x_i'.$$

Thus, the eigenvalues of P, $\lambda_i \neq 0, 1$, are the eigenvalues of $P - e\nu$. Since P is ergodic, from Lemma B.1 in Appendix B, $|\lambda_i| < 1$. Suppose 0 is an m-multiplicity eigenvalue of P and $y_j, j = 1, 2, \cdots, m$ are the corresponding eigenvectors, we have $Py_j = 0$. Next we prove y_j are also the eigenvector of $P - ev$ corresponding to eigenvalue 0.

$$(P - ev)y_j = -evy_j.$$

Since ev have a unique nonzero eigenvalue 1 and the corresponding eigenvector is e, we know for any vector $x \neq ce$, where c is an arbitrary constant, $evx = 0$. From $Pe = 1$, we have $y_j \neq ce$, thus $evy_j = 0$. Therefore, $(P - ev)y_j = 0$. From the above discussion, we know the eigenvalues of $P - ev$ are 0 and $\lambda_i, i = 1, 2, \cdots, S - 1$, where λ_i, with $|\lambda_i| < 1, i = 1, 2, \cdots, S - 1$ are the eigenvalues of P, in which λ_i may be zero. If $\lambda_i = 0$ is the m-multiplicity eigenvalue of P, 0 is the $(m + 1)$-multiplicity eigenvalue of $P - ev$.

c. Similarly to (3.7), we have the following iterative algorithm:

$$g_0 = f, \quad g_k = P_{-\nu}g_{k-1} + f, \quad k \geq 1.$$

d. From the algorithm in c), we know the algorithm converges to $g = \sum_{n=0}^{\infty}(P_{-\nu})^n f = \sum_{n=0}^{\infty}(P - e\nu)^n f = f + \sum_{n=1}^{\infty}(P^n - evP^{n-1})f$, where we have used $\nu e = 1$ and $Pe = 1$ but we have not preset $\nu g = \eta$. Then, we have

$$\begin{aligned}
\nu g &= \lim_{N\to\infty} \nu[f + \sum_{n=1}^{N}(P^n - e\nu P^{n-1})f] \\
&= \lim_{N\to\infty} \Big\{\nu f + \sum_{n=1}^{N}(\nu P^n - \nu P^{n-1})f\Big\} \\
&= \lim_{N\to\infty} \nu P^N f \\
&= \nu e \pi f \\
&= \eta.
\end{aligned}$$

e. Firstly, we have $p_{S*}e = e$, so p_{S*} is a special ν and $P\_$ is equivalent to $P_{-\nu}$. From the result in d), we have $p_{S*}g = \eta$. Next, we prove the equivalence between $f\_$ and f. From $P\_e = 0$, we know $P\_f\_ = P\_(f - f(S)e) = P\_f$. Thus, the potential $\sum_{n=0}^{\infty}(P\_)^n f\_$ obtained from (3.7) is equivalent to potential $\sum_{n=0}^{\infty}(P\_)^n f$, which is the result of the above algorithm. Therefore, the algorithm (3.4)-(3.7) is a special case of the above algorithm.

3.3 For any vector ν with $\nu e = 1$,

a. Prove $g = (I-P+e\nu)^{-1}f$ is a potential vector with normalization condition $\nu g = \eta$.

b. Can you derive a sample path based algorithm similar to (2.16) based on a)?

Solution:

a. We only need to prove $g = (I - P + e\nu)^{-1}f$ with normalization condition $\nu g = \eta$ is a solution of Poisson equation. From Poisson equation and normalization condition $\nu g = \eta$, we have

$$(I - P + e\nu)g = f.$$

Moreover, since the eigenvalues of $P-e\nu$ are all less than 1, matrix $I-P+e\nu$ is invertible. Thus, $g = (I - P + e\nu)^{-1}f$ is a solution of Poisson equation.

b.

$$\begin{aligned} g &= (I - P + e\nu)^{-1}f \\ &= \sum_{n=0}^{\infty}(P - e\nu)^n f \\ &= f + \sum_{n=1}^{\infty}(P^n - e\nu P^{n-1})f. \end{aligned}$$

Writing it in its components, similarly to (2.18), we have

$$g(i) = \lim_{L\to\infty}\Big\{E_i[\sum_{l=0}^{L-1} f(X_l)] - E_\nu[\sum_{l=0}^{L-1} f(X_l)]\Big\} \tag{3.1}$$

$$= \lim_{L\to\infty}\Big\{E_i[\sum_{l=0}^{L-1} f(X_l) - \eta] - E_\nu[\sum_{l=0}^{L-1} f(X_l) - \eta]\Big\}, \tag{3.2}$$

where E_i denotes the conditional expectation with respect to initial state $X_0 = i$, E_ν denotes the conditional expectation with respect to initial distribution ν. Since a sample path with initial distribution ν cannot be obtained, it is difficult to design an sample-path-based algorithm to estimate $\lim_{L\to\infty} E_\nu[\sum_{l=0}^{L-1} f(X_l) - \eta]$. Thus, we cannot design an sample-path-based algorithm similar to (2.16) to estimate the potential $g = (I - P + e\nu)^{-1}f$.

3.4 Consider

$$P = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.7 & 0 & 0.3 \\ 0.4 & 0.6 & 0 \end{bmatrix}, \qquad f = \begin{bmatrix} 10 \\ 2 \\ 7 \end{bmatrix}.$$

a. Calculate the potential vector using algorithm (3.1).

b. Calculate the potential vector using algorithm (3.3).

c. Calculate the potential vector using algorithm (3.7).

d. Calculate the potential vector using algorithm proposed in Problem 3.2.

Observe the convergence speeds and compare them with that of of $\lim_{k\to\infty} P^k = e\pi$.

Solution:

a. Using algorithm (3.1), we obtain the potential vector $g = [8.7600, 3.7380, 6.4252]$ if the algorithm is stopped when the norm of g_k and g_{k+1} is less than 0.001. The number of iterations is 18.

b. Using algorithm (3.3), the potential vector obtained is same as that in a) and the algorithm is stopped when the norm of g_k and g_{k+1} is less than 0.001. The number of iterations is 18.

c. Using algorithm (3.7), we obtain the potential vector $g = [2.3348, -2.6872, 0]$ if the algorithm is stopped when the norm of g_k and g_{k+1} is less than 0.001. The number of iterations is 18.

d. If we assume $\nu = [1, 0, 0]$, using the algorithm in problem 3.2, we obtain the potential vector $g = [6.312, 1.2996, 3.9868]$ if the algorithm is stopped when the norm of g_k and g_{k+1} is less than 0.001. The number of iterations is 18.

Computing P^n, we find P^n is approximately equal to $e\pi$ at $n = 18$. Thus, the convergence speed is same as that of the above algorithms.

3.5 Suppose a Markov chain starts from state i and we use the consecutive visits to the state i as the regenerative points (cf.(3.18)). That is, we set

$$i_0 = 0, \quad \text{with} \quad X_0 = i$$

$$i_k = \text{the epoch that} \quad X_l \quad \text{first visits state} \quad i \quad \text{after} \quad i_{k-1}, k \geq 1.$$

Then we denote the first visit epoch to state j in the kth regenerative period as j_k; i.e., $j_k = \min\{i_{k-1} < l \leq i_k : X_l = j\}$. We note that in some periods, such a point may not exit. Can we use the average of the sum of $\sum_{l=i_{k-1}}^{j_k-1} f(X_l)$ as the estimate of $\gamma(j,i)$? If not, why?

Solution:

We cannot use the average of the sum of $\sum_{l=i_{k-1}}^{j_k-1} f(X_l)$ as the estimate of $\gamma(j,i)$. From (2.17) in Chapter 2, we know $\gamma(j,i) = E\{\sum_0^{L(j|i)-1}[f(X_l) - \eta]|X_0 = i\}$, thus, we may use the average of $\sum_{l=i_{k-1}}^{j_k-1}[f(X_l) - \eta]$ as the estimate of $\gamma(j,i)$. If we directly omit η and use $\sum_{l=i_{k-1}}^{j_k-1} f(X_l)$ to estimate $\gamma(j,i)$, then the estimate will generate an estimate bias $\eta E(j_k - i_{k-1}) = \eta E[L(j|i)]$, where $L(j|i)$ denotes the time that the process moves to state j firstly from state i.

3.6 Let $p(1|1) = 0.5, p(2|1) = 0.2$ and $p(3|1) = 0.3$; $p(1|2) = 0.3$ and $p(2|2) = 0.5$, and $p(3|2) = 0.2$. Suppose $X = 1$ and $\tilde{X} = 2$, and we use the same uniformly distributed random variable $\xi \in [0,1)$ to determine the transition from both $X = 1$ and $\tilde{X} = 2$, according to (2.2). In this case, what are the conditional transition probabilities $\tilde{p}_{1|1}(*|2)$, $\tilde{p}_{2|1}(*|2)$ and $\tilde{p}_{3|1}(*|2)$?

Solution: Firstly, we consider $\tilde{p}_{1|1}(*|2)$. Given that the Markov chain $\boldsymbol{X}$ moves from state 1 to state 1, we know ξ is in $[0, 0.5)$. According to (2.2), we know the Markov chain $\tilde{\boldsymbol{X}}$ can only transit from state 2 to state 1 or state 2. If ξ is in $[0, 0.3)$, $\tilde{\boldsymbol{X}}$ transits from state 2 to state 1. If ξ is in $[0.3, 0.5)$, $\tilde{\boldsymbol{X}}$ transits from state 2 to state 2. Thus, we have $\tilde{p}_{1|1}(1|2) = \frac{0.3}{0.5} = 0.6$, $\tilde{p}_{1|1}(2|2) = \frac{0.5-0.3}{0.5} = 0.4$ and $\tilde{p}_{1|1}(3|2) = 0$. Similarly, Given that the Markov chain $\boldsymbol{X}$ transits from state 1 to state 2, we know ξ is in $[0.5, 0.7)$. According to (2.2), we know $\tilde{\boldsymbol{X}}$ can only transit from state 2 to state 2. Thus, $\tilde{p}_{2|1}(1|2) = 0$, $\tilde{p}_{2|1}(2|2) = 1$ and $\tilde{p}_{2|1}(3|2) = 0$. Given that the Markov chain $\boldsymbol{X}$ transits from state 1 to state 3, we know ξ is in $[0.7, 1)$. If ξ is in $[0.7, 0.8)$, $\tilde{\boldsymbol{X}}$ will transit from state 2 to state 2. If ξ is in $[0.8, 1)$, $\tilde{\boldsymbol{X}}$ will transit from state 2 to state 3. Thus, we have $\tilde{p}_{3|1}(1|2) = 0$, $\tilde{p}_{3|1}(2|2) = \frac{0.8-0.7}{1-0.7} = 1/3$ and $\tilde{p}_{3|1}(3|2) = \frac{1-0.8}{1-0.7} = 2/3$.

3.7 Let X and Y be two random variables with probability distributions $F(x)$ and $G(y)$, respectively. Their means are denoted as $\bar{x} = E(X)$ and $\bar{y} = E(Y)$. We wish to estimate

$\bar{x} - \bar{y} = E(X - Y)$ by simulation. We generate random variables X and Y using the inverse transformation method. Thus, we have $X = F^{-1}(\xi_1)$ and $Y = G^{-1}(\xi_2)$, where ξ_1 and ξ_2 are two uniformly distributed random variables in $[0, 1)$. Prove that if we choose $\xi_1 = \xi_2$, then the variance of $X - Y$, $Var[X - Y]$, is the smallest among all possible pairs of ξ_1 and ξ_2.

Solution: This problem is same as Problem A.4.

3.8 In the coupling approach, Prove the following statement:

a. Let $\widehat{\pi}$ be the S^2 dimensional steady-state probability (row) vector of $\widehat{P}$, i.e., $\widehat{\pi}\widehat{P} = \widehat{\pi}$, and π be the steady-state probability vector of P, i.e., $\pi P = \pi$. Then, $\widehat{\pi}(e_S \otimes I) = \widehat{\pi}(I \otimes e_S) = \pi$, and $\widehat{\pi}\widehat{g} = \widehat{\pi}\widehat{f} = 0$.

b. Equation (3.22) can take the form

$$(I - \widehat{P} + e_{S^2}\widehat{\pi})\widehat{g} = \widehat{f},$$

with $\widehat{\pi}\widehat{g} = 0$. Therefore, we have

$$\widehat{g} = \sum_{l=0}^{\infty} \widehat{P}^l \widehat{f}.$$

Solution:

b. Since $\widehat{\pi}(I \otimes e_S) = \widehat{\pi}\widehat{P}(I \otimes e_S) = \widehat{\pi}(P \otimes e_S) = \widehat{\pi}(I \otimes e_S)P$ and $\widehat{\pi}(I \otimes e_S)e_S = \widehat{\pi}(e_S \otimes e_S) = 1$, we have $\widehat{\pi}(I \otimes e_S) = \pi$ from the uniqueness of the solution of $\pi P = \pi$ and $\pi e = 1$. Similarly, since $\widehat{\pi}(e_S \otimes I) = \widehat{\pi}\widehat{P}(e_S \otimes I) = \widehat{\pi}(e_S \otimes P) = \widehat{\pi}(e_S \otimes I)P$ and $\widehat{\pi}(e_S \otimes I)e_S = \widehat{\pi}(e_S \otimes e_S) = 1$, we have $\widehat{\pi}(e_S \otimes I) = \pi$. From $\widehat{\pi}(e_S \otimes I) = \widehat{\pi}(I \otimes e_S) = \pi$, $\widehat{f} = (e_S \otimes f - f \otimes e_S) = (e_S \otimes I - I \otimes e_S)f$ and $\widehat{g} = (e_S \otimes I - I \otimes e_S)g$, we can easily obtain $\widehat{\pi}\widehat{g} = 0$ and $\widehat{\pi}\widehat{f} = 0$.

c. From (3.22), $\widehat{\eta} = 0$, and $\widehat{\pi}\widehat{g} = 0$, we have

$$(I - \widehat{P} + e_{S^2}\widehat{\pi})\widehat{g} = \widehat{f}.$$

That is, Equation (3.22) can take the form $(I - \widehat{P} + e_{S^2}\widehat{\pi})\widehat{g} = \widehat{f}$. From the result of Problem 2.3, we know $\widehat{P}$ is ergodic, then we know the fundamental matrix $I - \widehat{P} + e_{S^2}\widehat{\pi}$ is invertible, and $(I - \widehat{P} + e_{S^2}\widehat{\pi})^{-1} = \sum_{l=0}^{\infty}(\widehat{P} - e_{S^2}\widehat{\pi})^l = I + \sum_{l=1}^{\infty}(\widehat{P}^l - e_{S^2}\widehat{\pi})$. Thus, using $\widehat{\pi}\widehat{f} = 0$, we have $\widehat{\gamma} = \widehat{f} + \sum_{l=1}^{\infty}(\widehat{P}^l\widehat{f} - e_{S^2}\widehat{\pi}\widehat{f}) = \sum_{l=0}^{\infty}\widehat{P}^l\widehat{f}$.

3.9 To illustrate the coupling approach used in simulation for speeding up the estimation of $\gamma(i, j)$, let us consider a simple Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.2 & 0.3 & 0.5 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}.$$

a. Suppose that we generate two independent Markov chains with initial states $X_0 = 1$ and $X'_0 = 2$, respectively. What is the average length from $l = 0$ to L^*_{12}, $E(L^*_{12})$?

b. If we use the same $[0, 1)$ uniformly distributed random variable ξ to determine the state transitions for both Markov chain, what is $E(L^*_{12})$?

c. Answer the question in a) and b), if

$$P = \begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 0.4 & 0.2 & 0.4 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}.$$

Solution:

a. Define stochastic process $\widetilde{\boldsymbol{X}} = \{(X_n, X'_n), n \geq 0\}$, where X_n and X'_n are independent Markov chain with same transition matrix P, We know $L^*_{12} = \min\{n \geq 0, \widetilde{\boldsymbol{X}}_n \in A | \widetilde{\boldsymbol{X}}_0 = (1, 2)\}$, where $A = \{(1, 1), (2, 2), (3, 3)\}$. Therefore, this problem can be transformed into a problem about computing the expectation value of first passage time of $\widetilde{\boldsymbol{X}}$ reaching the set A from initial state $(1, 2)$. The transition matrix of $\widetilde{\boldsymbol{X}}$ is

$$\widetilde{P} = P \otimes P =$$

$$\begin{bmatrix} 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \\ 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \\ 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \\ 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \\ 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \\ 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \\ 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \\ 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \\ 0.0400 & 0.0600 & 0.1000 & 0.0600 & 0.0900 & 0.1500 & 0.1000 & 0.1500 & 0.2500 \end{bmatrix}.$$

since

$$\begin{aligned} E(L^*_{1,2}) &= \sum_{n=0}^{\infty} \mathcal{P}(N_A = n)n = \sum_{n=1}^{\infty}\sum_{l=1}^{n} \mathcal{P}(N_A = n) \\ &= \sum_{l=1}^{\infty}\sum_{n=l}^{\infty} \mathcal{P}(N_A = n) = \sum_{l=1}^{\infty} \mathcal{P}(N_A \geq l), \end{aligned}$$

where N_A denotes the step numbers that $\widetilde{\boldsymbol{X}}$ moves to set A firstly. Since $\mathcal{P}(N_A \geq l) = \mathcal{P}(\text{the first } l-1 \text{ transitions stay at } B = A^c)$, where B is the complement set of A, i.e. $B = A^c$. Thus, in a vector way, the expectation of first passage time from any state $i \in B$ to set A is the corresponding component in $(I - P_B)^{-1}e = \sum_{n=0}^{\infty} P_B^n e$, where P_B is a matrix that deletes the columns and rows corresponding to the states in set A in transition matrix P. This result is a generalization of the result of b) in Problem 2.20. $(I-P_B)^{-1}e = (2.6316, 2.6316, 2.6316, 2.6316, 2.6316, 2.6316)^T$, so $E(L^*_{12}) = [(I - P_B)^{-1}e]_1 = 2.6316$.

b. If we use the same $[0, 1)$ uniformly distributed random variable ξ to determine the state transitions for both Markov chains, since the transition probabilities from state 1 and state 2 to any state are the same, two Markov chain will reach the same state in one step. Thus, $E(L^*_{12}) = 1$.

c. Similarly to a), we can obtain $E(L^*_{12}) = [(I - P_B)^{-1}e]_1 = 3.125$. If we use the same $[0, 1)$ uniformly distributed random variable ξ to determine the state transitions for both Markov chain, two Markov chain transit to the same states in one step when ξ falls in $[0, 0.2)$ or $[0.4, 1)$. That is, two Markov chain transit to the same states in one step with probability 0.8 and transit to different states with probability 0.2, thus $E(L^*_{12}) = \sum_{n=1}^{\infty} 0.8 * (0.2)^{n-1}n = \frac{0.8}{(1-0.2)^2} = 1.25$.

From this example, we can find the coupling approach can reduce the the time that two Markov chains merge. Thus, this approach can estimate $\gamma(i, j)$ with less variance.

3.10 The realization factor $\gamma(i, j)$ can be obtained by simulating two sample path initiating with i and j, respectively, up to its merging point L_{ij}:

$$\gamma(i, j) = E\{ \sum_{l=0}^{L_{i,j}-1} [f(X'_l) - f(X_l)] | X_0 = i, X'_0 = j\}.$$

If the two sample paths are independent, as shown in the text, we can obtain the perturbation realization factor equation. However, in simulation, we may use coupling to

reduce the variance in estimating the difference of the mean values of two random variables ($\gamma(i,j) = g(j) - g(i)$). In our case, we wish to let the two sample paths, initiating with i and j, merge as early as possible.

To this end, in simulation we can force the two sample paths to jump to the same state, from i to j respectively, with a probability as large as possible. We may use the same random variable to determine the state transitions in the two paths. For example, if $p(k|i) = 0.3$ and $p(k|j) = 0.2$, instead of using two independent random numbers in $[0,1)$ to determine the state transitions for $X_0 = i$ and $X_0' = j$, respectively, we generate one uniformly distributed random number $\xi \in [0,1)$, if $\xi \in [0, 0.2)$, we let both $X_1 = X_1' = k$. We use an example to show this coupling method: Let $p(1|2) = 0.5$, $p(2|2) = 0.3, p(3|2) = 0.2$, and $p(1|3) = 0.2, p(2|3) = 0.7, p(3|3) = 0.1$. The largest probabilities for the two paths starting from $X_0 = 2$ and $X_0' = 3$ to merge at $X_1 = X_1' = 1$ is $\min\{p(1|2), p(1|3)\} = 0.2$, to merge at $X_1 = X_1' = 2$ is $\min\{p(2|2), p(2|3)\} = 0.3$, and to merge at $X_1 = X_1' = 3$ is $\min\{p(3|2), p(3|3)\} = 0.1$. Thus, the largest probability that the two sample paths merge at $X_1 = X_1'$ with the coupling technique is $0.2 + 0.3 + 0.1 = 0.6$. We simulate the two sample paths in two steps. In the first step, we generate a uniformly distributed random variable $\xi \in [0,1)$. If $\xi \in [0, 0.2)$, we set $X_1 = X_1' = 1$; if $\xi \in [0.2, 0.5)$, we set $X_1 = X_1' = 2$; if $\xi \in [0.5, 0.6)$, we set $X_1 = X_1' = 3$. If $\xi \in [0.6, 1)$, we go to the second step: using another two independent random numbers determine the transitions for the two sample paths.

Continue the above reasoning and mathematically formulate it. Work on $\gamma(i, S)$ for all state $i \in S$ and derive the following equation

$$g(i) - g(S) = f(i) - f(S) + \sum_{j=1}^{S} [p(j|i) - p(j|S)]g(j), \qquad i \in S.$$

Prove it is the same as (3.4).

Solution:

From the above reasoning, our objective is to maximize the probability that the two sample paths starting from different states i and j merge. This problem can be transformed into a linear programming problem:

Linear Programming: For $i \neq j$,

$$\begin{aligned} & \max \sum_{k\in\mathcal{S}} p[(k,k)|(i,j)], \\ s.t. \quad & \sum_{l\in S} p[(k,l)|(i,j)] = p(k|i), \\ & \sum_{k\in\mathcal{S}} p[(k,l)|(i,j)] = p(l|j), \\ & p[(k,l)|(i,j)] \geq 0, \quad k,l \in \mathcal{S}. \end{aligned}$$

For $i = j$, we can choose the transition probabilities $p[(k,l)|(i,i)], k,l \in \mathcal{S}$ to satisfy the following equations:

$$\begin{aligned} & \sum_{l\in S} p[(k,l)|(i,i)] = p(k|i), \\ & \sum_{k\in S} p[(k,l)|(i,i)] = p(l|i), \\ & p[(k,l)|(i,i)] \geq 0, k,l \in \mathcal{S}. \end{aligned}$$

Since

$$\begin{aligned} & \gamma(i,S) \\ = \; & g(S) - g(i) \\ = \; & \lim_{L\to\infty} E\left\{\sum_{l=0}^{L-1}[f(X_l)-\eta]|X_0 = S\right\} - \lim_{L\to\infty} E\left\{\sum_{l=0}^{L-1}[f(X_l)-\eta]|X_0 = i\right\} \\ = \; & f(S) + \sum_{j\in\mathcal{S}} p(j|S) \lim_{L\to\infty} E\left\{\sum_{l=1}^{L-1}[f(X_l)-\eta]|X_1 = j\right\} \\ & -f(i) - \sum_{j\in\mathcal{S}} p(j|i) \lim_{L\to\infty} E\left\{\sum_{l=1}^{L-1}[f(X_l)-\eta]|X_1 = j\right\} \\ = \; & f(S) - f(i) + \sum_{j\in\mathcal{S}} (p(j|S) - p(j|i))g(j). \end{aligned}$$

thus, we have $g(i) - g(S) = f(i) - f(S) + \sum_{j\in\mathcal{S}}(p(j|i) - p(j|S))g(j),\ \ i \in \mathcal{S}$. This equation is the same as that obtained by subtracting the last row of the Poisson equation from all the rows.

3.11 One of the restriction of the basic formula (3.32) is that it requires $p(j|i) > 0$ if $\Delta p(j|i) > 0$ for all $i, j \in \mathcal{S}$. This condition can be relaxed. For example, we may assume

that if $\Delta p(j|i) > 0$ then there exists a state, denoted as $k_{i,j}$, such that $p(k_{i,j}|i)p(j|k_{i,j}) > 0$. Under this assumption, we have

$$\frac{d\eta_\delta}{d\delta} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}\{\pi(i)[p(k_{ij}|i)p(j|k_{i,j})\frac{\Delta p(j|i)}{p(k_{i,j}|i)p(j|k_{i,j})}g(j)]\}.$$

Furthermore, we have

$$\frac{d\eta_\delta}{d\delta} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}\{\pi(i)[\sum_{k\in\mathcal{S}}p(k|i)p(j|k)\frac{\Delta p(j|i)}{\sum_{k\in\mathcal{S}}p(k|i)p(j|k)}g(j)]\}.$$

a. Continue the analysis and develop the direct learning algorithms for the performance derivatives,

b. Compared with (3.32), what are the disadvantages of this "improved" approach, if any?

c. Extend this analysis to the more general case of irreducible Markov chains.

Solution:

a. We consider the approximation by truncation similar to Algorithm 3.1. Since

$$\begin{aligned}
\frac{d\eta_\delta}{d\delta} &= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}\left\{\pi(i)\Big[\sum_{k\in\mathcal{S}}p(k|i)p(j|k)\frac{\Delta p(j|i)}{\sum_{k\in\mathcal{S}}p(k|i)p(j|k)}g(j)\Big]\right\} \\
&= E\Big\{\frac{\Delta p(X_{l+2}|X_l)}{\sum_{k\in\mathcal{S}}p(X_{l+2}|k)p(k|X_l)}g(X_{l+2})\Big\} \\
&\approx \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\frac{\Delta p(X_{n+2}|X_n)}{\sum_{k\in\mathcal{S}}p(X_{n+2}|k)p(k|X_n)}\sum_{l=n+2}^{n+L+1}f(X_l) \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\left\{f(X_{n+L+1})\sum_{l=0}^{L-1}\Big[\frac{\Delta P(X_{n+l+2}|X_{n+l})}{\sum_{k\in\mathcal{S}}p(k|X_{n+l})p(X_{n+l+2}|k)}\Big]\right\}.
\end{aligned}$$

Similarly, we can obtain the approximation by discount factor.

$$\frac{d\eta_\delta}{d\delta} = \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\Big\{f(X_{n+1})\sum_{l=0}^{n-1}[\beta^{n-l-1}\frac{\Delta p(X_{l+2}|X_l)}{\sum_{k\in\mathcal{S}}p(k|X_l)p(X_{l+2}|k)}]\Big\}$$

b. In the "improved" method, the summation $\sum_{k\in\mathcal{S}}p(k|X_l)p(X_{l+2}|k)$ will lead to the increment of computation.

c. For general irreducible Markov chains, we know there is $k_{i,j} > 0$ such that $p^{k_{i,j}}(j|i) > 0$ for any two states i and j. Define $K = \max_{i,j\in\mathcal{S}}\{k_{i,j}\}$, we have $p^K(j|i) > 0$ for any states $i \in \mathcal{S}$ and $j \in \mathcal{S}$. Thus,

$$\frac{d\eta_\delta}{d\delta} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}}\left\{\pi(i)p^K(j|i)\Big[\frac{\Delta p(j|i)}{p^K(j|i)}g(j)\Big]\right\}.$$

Then, similarly to a), we can develop the direct learning algorithms.

3.12 In the gradient estimates (3.34), we have ignored the constant term η in the expression of g. A more accurate estimate should be

$$\frac{d\eta_\delta}{d\delta} \approx \lim_{N\to\infty}\frac{1}{N}\Big\{\sum_{n=0}^{N-1}\{\frac{\Delta p(X_{n+1}|X_n)}{p(X_{n+1}|X_n)}\}\sum_{l=0}^{L-1}[f(X_{n+l+1}) - \eta]\Big\}, \quad w.p.1.$$

Prove

$$\frac{d\eta_\delta}{d\delta} \approx \lim_{N\to\infty}\frac{1}{N}\Big\{\sum_{n=0}^{N-1}\{\frac{\Delta p(X_{n+1}|X_n)}{p(X_{n+1}|X_n)}\}\sum_{l=0}^{L-1}f(X_{n+l+1})\Big\}, \quad w.p.1.$$

and discuss the estimation error caused by a finite $L\eta$.

Solution: Since

$$\begin{aligned}
&\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\{\frac{\Delta p(X_{n+1}|X_n)}{p(X_{n+1}|X_n)}\}L\eta \\
= \;& L\eta E\Big[\frac{\Delta p(X_{n+1}|X_n)}{p(X_{n+1}|X_n)}\Big] \\
= \;& L\eta\sum_{i\in\mathcal{S}}\pi(i)\sum_{j\in\mathcal{S}}p(j|i)\frac{\Delta p(j|i)}{p(j|i)} = 0,
\end{aligned}$$

Thus, we have

$$\frac{d\eta_\delta}{d\delta} \approx \lim_{N\to\infty}\frac{1}{N}\Big\{\sum_{n=0}^{N-1}\{\frac{\Delta p(X_{n+1}|X_n)}{p(X_{n+1}|X_n)}\}\sum_{l=0}^{L-1}f(X_{n+l+1})\Big\}, \quad \text{w.p.1.}$$

Although the omittance of $L\eta$ does not result in the bias, it will result in a large variance. This is because the omittance makes the sum $\sum_{l=0}^{L-1}f(X_{n+l+1})$ larger, which results in a large fluctuation of the estimate.

3.13 Discuss the error in the gradient estimate (3.41) caused by ignoring the second term of (3.40) for a finite N. You may set $f \equiv 1$.

Solution: The error is

$$\text{error} = \frac{1}{N}\sum_{n=0}^{N-1}\{\frac{\Delta p(X_{n+1}|X_n)}{p(X_{n+1}|X_n)}\}\sum_{l=N-n}^{\infty}\beta^l f(X_{n+l+1}).$$

If we set $f \equiv 1$ and $\left\|\frac{\Delta p(X_{n+1}|X_n)}{p(X_{n+1}|X_n)}\right\| \leq B$, we have

$$\text{error} \leq B\frac{1}{N}\sum_{n=0}^{N-1}\sum_{l=N-n}^{\infty}\beta^l = B\frac{1}{N}\sum_{n=0}^{N-1}\frac{\beta^{N-n}}{1-\beta} = \frac{B\beta(1-\beta^N)}{N(1-\beta)^2}.$$

When $N \to \infty$, the error tends to zero.

3.14 Let η_r be the average performance of a Markov chain with transition probability matrix P_r defined as $p_r(i|i) = r$ for all $i \in \mathcal{S}$ and $p_r(j|i) = (1-r)q_{ij}, j \neq i, i, j \in \mathcal{S}$, with $\sum_{j\in\mathcal{S}} q_{i,j} = 1$ for all $i \in \mathcal{S}$. Please prove $\frac{d\eta_r}{dr} = 0$ for all $0 < r < 1$ using performance derivative formula (3.30).

Solution:

Let $\Delta P_r = P_{r'} - P_r$, then $\Delta p_r(i|i) = r' - r$ and $\Delta p_r(j|i) = -(r'-r)q_{ij}, i, j \in \mathcal{S}$, thus, $\frac{\Delta p_r(X_{l+1}|X_l)}{p_r(X_{l+1}|X_l)}$ is equal to $\frac{r'-r}{r}$ when X_l transits to the same state at time $l+1$ and is equal to $\frac{-(r'-r)}{1-r}$ when X_l transits to different state at time $l+1$. No matter what is X_l, X_l transits to the same state at time $l+1$ with probability r and transits to different state at time $l+1$ with probability $1-r$, thus, $\frac{\Delta p_r(X_{l+1}|X_l)}{p_r(X_{l+1}|X_l)} = \frac{r'-r}{r}$ with probability r and $\frac{\Delta p_r(X_{l+1}|X_l)}{p_r(X_{l+1}|X_l)} = \frac{-(r'-r)}{1-r}$ with probability $1-r$. From performance derivative formula (3.30), we know

$$\begin{aligned}
\frac{d\eta_r}{dr} &= E\Big\{\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}g(X_{l+1})\Big\} \\
&= E\left\{E\Big[\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}g(X_{l+1})\Big|X_{l+1}\Big]\right\} \\
&= E\left\{E\Big[\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}\Big|X_{l+1}\Big]g(X_{l+1})\right\} \\
&= E\left\{\Big[r * \frac{r'-r}{r} + (1-r)\frac{-(r'-r)}{1-r}\Big]g(X_{l+1})\right\} \\
&= 0.
\end{aligned}$$

3.15 In Algorithm 3.1, prove that the following equation holds

$$\lim_{L\to\infty}\{\sum_{l=0}^{L-1}P^l(\Delta P)P^{L-l-1}\} = e\pi(\Delta P)(I - P + e\pi)^{-1}.$$

In addition, prove that at the steady state, we have

$$\pi(i)\rho_L(i) = E\Big\{ I_i(X_L) \sum_{l=0}^{L-1} \frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)} \Big\} = \pi \sum_{l=0}^{L-1} \{P^l(\Delta P)P^{L-l-1}\} e_{\cdot i},$$

where $e_{\cdot i}$ is the ith column vector of the identity matrix I. Equation (3.38) and the convergence of (3.37) follow directly from these two equations.

Solution: For ergodic Markov chain, we have $P^l \to e\pi$ when $l \to \infty$. Therefore, there is a N, when $l \geq N$, we have $-\epsilon E \leq P^l - e\pi \leq \epsilon E$, where E is a $\mathcal{S} \times \mathcal{S}$ matrix with all components equal to 1. Moreover,

$$\begin{aligned} & \sum_{l=0}^{L-1} P^l(\Delta P)P^{L-l-1} \\ = & \sum_{m=0}^{L-1} P^{L-m-1}(\Delta P)P^m \quad (Let \quad m = L-l-1) \\ = & \sum_{m=0}^{L-N-1} P^{L-m-1}(\Delta P)P^m + \sum_{m=L-N-1}^{L-1} P^{L-m-1}(\Delta P)P^m. \end{aligned} \tag{3.3}$$

When L is large enough, for example $L > 2N+1$, we have $L-N-1 > N$. For the second item in equation (3.3), we have

$$\begin{aligned} \sum_{m=L-N-1}^{L-1} P^{L-m-1}(\Delta P)(e\pi - \epsilon E) & \leq \sum_{m=L-N-1}^{L-1} P^{L-m-1}(\Delta P)P^m \\ & \leq \sum_{m=L-N-1}^{L-1} P^{L-m-1}(\Delta P)(e\pi + \epsilon E). \end{aligned}$$

From $\Delta P e = 0$, we know the second item $\sum_{m=L-N-1}^{L-1} P^{L-m-1}(\Delta P)P^m = 0$. For the first item of (3.3), since $L-m-1 \geq N$ for $0 \leq m \leq L-N-1$, we have

$$\sum_{m=0}^{L-N-1} (e\pi - \epsilon E)(\Delta P)P^m \leq \sum_{m=0}^{L-N-1} P^{L-m-1}(\Delta P)P^m \leq \sum_{m=0}^{L-N-1} (e\pi + \epsilon E)(\Delta P)P^m.$$

Let $L \to \infty$, we have

$$(e\pi - \epsilon E)(\Delta P) \sum_{m=0}^{\infty} P^m \leq \lim_{L\to\infty} \sum_{m=0}^{L-N-1} P^{L-m-1}(\Delta P)P^m \leq (e\pi + \epsilon E)(\Delta P) \sum_{m=0}^{\infty} P^m.$$

From the arbitrary property of ϵ, we have

$$\lim_{L\to\infty} \{\sum_{l=0}^{L-1} P^l(\Delta P)P^{L-l-1}\}$$

$$
\begin{aligned}
&= \lim_{L\to\infty} \sum_{m=0}^{L-N-1} P^{L-m-1}(\Delta P)P^m \\
&= e\pi(\Delta P)\sum_{m=0}^{\infty} P^m.
\end{aligned}
$$

Because $(I-P+e\pi)^{-1} = \sum_{m=0}^{\infty} P^m - e\pi$ and $\Delta P e = 0$, we have

$$
\lim_{L\to\infty}\{\sum_{l=0}^{L-1} P^l(\Delta P)P^{L-l-1}\} = e\pi(\Delta P)(I-P+e\pi)^{-1}.
$$

Next, we prove $\pi(i)\rho_L(i) = E\Big\{I_i(X_L)\sum_{l=0}^{L-1}\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}\Big\} = \pi\sum_{l=0}^{L-1}\{P^l(\Delta P)P^{L-l-1}\}e_{\cdot i}$.

$$
\begin{aligned}
& E\Big\{I_i(X_L)\sum_{l=0}^{L-1}\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}\Big\} \\
= & \sum_{l=0}^{L-1} E\Big\{I_i(X_L)\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}\Big\} \\
= & \sum_{l=0}^{L-1}\sum_{i_0}\pi(i_0)\sum_{i_1}p(i_1|i_0)\sum_{i_2}p(i_2|i_1)\cdots\sum_{i_{l+1}}p(i_{l+1}|i_l)\frac{\Delta p(i_{l+1}|i_l)}{p(i_{l+1}|i_l)} \\
& \sum_{i_{l+2}}p(i_{l+2}|i_{l+1})\cdots\sum_{i_{L-1}}p(i_{l-1}|i_{L-2})p(i|i_{L-1}) \\
= & \sum_{l=0}^{L-1}\sum_{i_0}\pi(i_0)\sum_{i_1}p(i_1|i_0)\sum_{i_2}p(i_2|i_1)\cdots\sum_{i_{l+1}}\Delta p(i_{l+1}|i_l)\sum_{i_{l+2}}p(i_{l+2}|i_{l+1})\cdots p(i|i_{L-1}) \\
= & \pi\{\sum_{l=0}^{L-1} P^l(\Delta P)P^{L-l-1}\}e_{\cdot i}.
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
& E\Big\{I_i(X_L)\sum_{l=0}^{L-1}\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}\Big\} \\
= & E\left\{E\left[I_i(X_L)\sum_{l=0}^{L-1}\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}\Big|X_L\right]\right\} \\
= & \pi(i)\rho_L(i).
\end{aligned}
$$

Thus, we have

$$
\pi(i)\rho_L(i) = E\Big\{I_i(X_L)\sum_{l=0}^{L-1}\frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)}\Big\} = \pi\left\{\sum_{l=0}^{L-1} P^l(\Delta P)P^{L-l-1}\right\}e_{\cdot i}.
$$

From the above equation, we know the limit of $\rho_L(i)$ exists when $L \to \infty$ and

$$\begin{aligned}\lim_{L\to\infty}\sum_{i\in\mathcal{S}}\pi(i)\rho_L(i)f(i) &= \lim_{L\to\infty}\pi\sum_{l=0}^{L-1}P^l(\Delta P)P^{L-l-1}f \\ &= \pi e\pi(\Delta P)(I-P+e\pi)^{-1}f \\ &= \frac{d\eta_\delta}{d\delta}.\end{aligned}$$

3.16 In Problem 3.15, we set $G_L = \sum_{l=0}^{L-1} P^l(\Delta P)P^{L-l-1}$. Prove

$$G_{L+1} = PG_L + G_LP - PG_{L-1}P.$$

with $G_0 = 0, G_1 = \Delta P$. Set $G = \lim_{L\to\infty} G_L$. Explain the meaning of G. Finally, letting $L \to \infty$ on both sides of the above equation, we obtain $G = PG + GP - PGP$. Is this equation useful in any sense?

Solution: When $L = 1$, it is obvious that

$$G_2 = P\Delta P + \Delta PP = PG_1 + G_1P.$$

When $L \geq 2$, we have

$$\begin{aligned}& PG_L + G_LP - PG_{L-1}P \\ =\ & \sum_{l=0}^{L-1}P^{l+1}(\Delta P)P^{L-l-1} + \sum_{l=0}^{L-1}P^l(\Delta P)P^{L-l} - \sum_{l=0}^{L-2}P^{l+1}(\Delta P)P^{L-l-1} \\ =\ & P^L\Delta P + \sum_{l=0}^{L-2}P^{l+1}(\Delta P)P^{L-l-1} + P^{L-1}\Delta PP + \sum_{l=0}^{L-2}P^l(\Delta P)P^{L-l} \\ & - \sum_{l=0}^{L-2}P^{l+1}(\Delta P)P^{L-l-1} \\ =\ & P^L\Delta P + P^{L-1}\Delta PP + \sum_{l=0}^{L-2}P^l(P\Delta P + \Delta PP - P\Delta P)P^{L-l-1} \\ =\ & P^L\Delta P + P^{L-1}\Delta PP + \sum_{l=0}^{L-2}P^l\Delta PP^{L-l} \\ =\ & \sum_{l=0}^{L}P^l\Delta PP^{L-l} = G_{L+1}.\end{aligned}$$

G is the limit point of the iteration $G_{L+1} = PG_L + G_LP - PG_{L-1}P$. We can see that G_L denotes the perturbation effect on L-step transition matrix P^L due to the parameter

change ΔP . So the physical meaning of G is the perturbation effect on steady state $P^{\infty} = e\pi$ due to the parameter change ΔP, i.e. $G = e\frac{d\pi}{d\delta}$.

The equation have infinite solution, for example, for any row vector v, ev is a solution of this equation. Thus, from the equation, we cannot obtain the solution we need. However, the iteration from this equation can be used to compute the performance derivative. By using the iteration, we obtain G, then $Gf = \frac{d\eta}{d\delta}e$, which avoid the computation of the inverse.

3.17 Write a computer simulation program

a. to estimate potentials by using (3.15) and (3.19)

b. to estimate the performance derivative by using (3.35), (3.41), and (3.43).

Solution:

a. **The algorithm by using (3.15):**

Given arrays: **StateNum**, **StatePerf** and **Statequeue**; (**StateNum** records the number of visiting state, which is $1 \times S$ dimension; **StatePerf** records the total performance,i.e. $\textbf{StatePerf}(X_n) = \sum_{l=0}^{L-1} f(X_{n+l})$ and **Statequeue** records L continuous states, that is from X_n to X_{n+L-1}. We do 10000 transitions.

for $k = 1$ to 10000 **do**
  if $k <= L$ **then**
    Statequeue(k)=X_k
  else
    StateNum(Statequeue(1))=StateNum(Statequeue(1))+1
    for $l = 1$ to L **do**
      StatePerf(Statequeue(1))=StatePerf(Statequeue(1))+f(Statequeue(l))
    end for
    for $l = 1$ to $L - 1$ **do**
      Statequeue(l)= Statequeue($l + 1$)
    end for
  end if
  Statequeue(L)=X_k

end for

for $i = 1$ to S **do**

Potential(i)=StatePerf(i)/StateNum(i)

end for

The algorithm by using (3.19):

Given arrays:

1. $S \times S$ matrix: **StateTrNum**, whose (m, n)th component denotes "the number from state m to firstly visit state n";
2. $S \times S$ matrix: **SumStateTrNum**, whose (m, n)th component denotes "the sum of the number from state m to firstly visit state n", i.e. $\sum_k L_k(n|m)$;
3. $S \times S$ matrix: **StatePerf**, whose (m, n)th component denotes "the sum of performance from state m to state n," i.e. $\sum_k R_k(m, n)$;
4. $S \times S$ matrix: **Flag**, which is indicator matrix, and its initial value is zero matrix.

We do 10000 transitions.

for $j = 1$ to S **do**

Flag(X_0, j)=1;

end for

for $k = 0$ to 9999 **do**

StateNum(X_k)=StateNum(X_k)+1;

for $i = 1$ to S **do**

for $j = 1$ to S **do**

StateTrTemp(i, j)=StateTrTemp(i, j)+Flag(i, j);

StatePerfTemp(i, j)=StatePerfTemp(i, j)+f(X_k)*Flag(i, j)

end for

end for

Generate the next state X_{k+1}

for $i = 1$ to S **do**

StateTrNum(i, X_{k+1})=StateNum(i, X_{k+1})+Flag(i, X_{k+1});

SumStateTrNum(i, X_{k+1})=SumStateTrNum(i, X_{k+1})+StateTrTemp(i, X_{k+1});

StatePerf(i, X_{k+1})=StatePerf(i, X_{k+1})+StatePerfTemp(i, X_{k+1});

StateTrTemp(i, X_{k+1})=0; StatePerfTemp(i, X_{k+1})=0;Flag(i, X_{k+1})=0

end for

for $j = 1$ to S **do**

Flag(X_{k+1}, j)=1

end for

end for

AllNum=$\sum_k$StateNum(k);

for $i = 1$ to S **do**

$\hat{\pi}(i)$= StateNum(i)/AlltNum;

end for

$\hat{\eta} = \hat{\pi} f$;

for $i = 1$ to S **do**

for $j = 1$ to S **do**

$$\hat{\gamma}(i, j) = \frac{\text{StatePerf}_{(i,j)}}{\text{StateTrNum}_{(i,j)}} - \frac{\text{SumStateTrNum}_{(i,j)}}{\text{StateTrNum}_{(i,j)}} \hat{\eta}$$

end for

end for

$\hat{g} = \hat{\Gamma}^T \hat{\pi}^T$.

b. **Algorithm by using (3.35):**

Set **ImportSampQueue** be a $1 \times L$-dimensional matrix, $k = 0$ and $\Delta_0 = 0$

for $k = 1$ to 10000 **do**

if $k <= L$ **then**

ImportSampQueue(k) = $\frac{\Delta p(X_k|X_{k-1})}{p(X_k|X_{k-1})}$

if $k = L$ **then**

$\Delta_{k-L+1} = \Delta_{k-L} + \frac{1}{k-L+1}[f(X_k) \sum_l \text{ImportSampQueue}(l) - \Delta_{k-L}]$

end if

else

for $l = 1$ to $L - 1$ **do**

ImportSampQueue(l) = ImportSampQueue($l + 1$);

end for

ImportSampQueue$(L) = \frac{\Delta p(X_{k+1}|X_k)}{p(X_{k+1}|X_k)}$

$\Delta_{k-L+1} = \Delta_{k-L} + \frac{1}{k-L+1}[f(X_k)\sum_k \text{ImportSampQueue}(k) - \Delta_{k-L}]$;

end if

end for

Δ_k is the value of the derivative.

Algorithm by using (3.41):

Set $Z_0 = 0$, $k = 0$ and $\Delta_0 = 0$

for each state X_{k+1} visited **do**

$Z_{k+1} = \beta Z_k + \frac{\Delta p(X_{k+1}|X_k)}{p(X_{k+1}|X_k)}$;

$\Delta_{k+1} = \Delta_k + \frac{1}{k+1}(f(X_{k+1})Z_k - \Delta_k)$;

end for

Δ_k is the value of the derivative.

Algorithm by using (3.43):

Set $Z_0 = 0$, $k = 0$ and $\Delta_0 = 0$

for each state X_{k+1} visited **do**

$$Z_{k+1} = \begin{cases} Z_k + \frac{\Delta p(X_{k+1}|X_k)}{p(X_{k+1}|X_k)}, & \text{if } X_{k+1} \neq i^* \\ 0, & \text{if } X_{k+1} = i^* \end{cases}$$

$\Delta_{k+1} = \Delta_k + \frac{1}{k+1}(f(X_{k+1})Z_k - \Delta_k)$;

end for

Δ_k is the value of the derivative.

3.18 The group inverse (2.48) $B^{\#} = -[(I - P + e\pi)^{-1} - e\pi]$ (for ergodic chains) plays an important role in performance sensitivity analysis. Let $b^{\#}(i,j)$ be the (i,j)th component of $B^{\#}$. Consider a Markov chain starting from state $i \in S$. Let $N_{ij}^{(L)}$ be the expected number of times that the Markov chain visits state $j \in S$ in the first L stages. Prove

$$\lim_{L\to\infty}(N_{ji}^{(L)} - N_{ki}^{(L)}) = b^{\#}(k,i) - b^{\#}(j,i).$$

Solution: Because $N_{ij}^{(L)} = \sum_{n=0}^{L} p(X_n = j|X_0 = i)$, we have $N_{ij}^{(L)} = [\sum_{n=0}^{L} P^n]_{ij}$, where $[\cdot]_{ij}$ denotes the (i,j) component of matrix. Since $B^{\#} = -[(I - P + e\pi)^{-1} - e\pi] =$

$-\sum_{n=0}^{\infty}(P-e\pi)^n + e\pi = -I - \sum_{n=1}^{\infty}(P^n + e\pi) - e\pi = \lim_{L\to\infty}\sum_{n=0}^{L} -P^n + (L+1)e\pi,$

$$\begin{aligned} & b^{\#}(k,i) - b^{\#}(j,i) \\ = & \lim_{L\to\infty}\left\{ [\sum_{n=0}^{L} -P^n + (L+1)e\pi]_{k,i} - [\sum_{n=0}^{L} -P^n + (L+1)e\pi]_{j,i} \right\} \\ = & \lim_{L\to\infty}\left\{ [\sum_{n=0}^{L} P^n]_{j,i} - [\sum_{n=0}^{L} P^n]_{k,i} \right\} \\ = & \lim_{L\to\infty}(N_{j,i}^{L} - N_{k,i}^{L}). \end{aligned}$$

3.19 Given a direction defined by ΔP, is it possible to estimate the second order derivative $\frac{d^2\eta_\delta}{d\delta^2}$ using a sample path of the Markov chain with transition probability matrix P (cf. Section 2.1.5)? How about the second order performance derivative of any given reward function $f(\theta)$?

Solution: From Section 2.1.5 in Chapter 2, we have

$$\frac{d^2\eta_\delta}{d\delta^2} = 2\pi(\Delta P)(I-P+e\pi)^{-1}(\Delta P)(I-P+e\pi)^{-1}f.$$

From Problem 3.15, we know we can use

$$\hat{\omega}_i := \frac{\sum_{n=0}^{N-1} I_i(X_{n+L}) \sum_{l=0}^{L-1} \frac{\Delta p(X_{n+l+1}|X_{n+l})}{p(X_{n+l+1}|X_{n+l})}}{\sum_{n=0}^{N-1} I_i(X_{n+L})}$$

as an estimate of $\pi(\Delta P)(I-P+e\pi)^{-1}e_i$. Since $(I-P+e\pi)^{-1}f$ is the potential, we can estimate it by using a sample path of Markov chain. We use the methods in Section 3.1.2 to estimate the potential $(I-P+e\pi)^{-1}f$ and get potential estimates $\hat{g}$, then compute the value $\Delta P\hat{g}$, whose ith component is ν_i. Finally we use $2\sum_{i=1}^{S}\hat{\omega}_i\nu_i$ to estimate the second order derivative.

Moreover, we can also firstly use one part sample path of Markov chain to get the potential estimate $\hat{g}$. Then making $\Delta P\hat{g}$ as the performance function, we utilize another part sample path to estimate $(I-P+e\pi)^{-1}\Delta P\hat{g}$. Finally, we use $2\pi\Delta P(I-P+e\pi)^{-1}\Delta P\hat{g}$ to estimate the second order derivative. In this method, we need to repeat using one sample path or make two simulations.

When the reward function is related with parameters, we have $\eta(\theta) = \pi(\theta)f(\theta)$, thus, the second order derivative of $\eta(\theta)$ is

$$\frac{d^2\eta(\theta)}{d\theta^2} = \frac{d^2\pi(\theta)}{d\theta^2}f + 2\frac{d\pi(\theta)}{d\theta}\frac{df(\theta)}{d\theta} + \pi(\theta)\frac{d^2f(\theta)}{d\theta^2}. \tag{3.4}$$

For the first item in the (3.4), we can use the above estimate method of the second order derivative to estimate it. For the second item, we make $\frac{df(\theta)}{d\theta}$ as a reward function, using the estimate methods of the derivative in book, we can obtain its estimate. For the last item, we view it as an average reward performance, where the performance function is $\frac{d^2f(\theta)}{d\theta^2}$, and can also obtain its estimation.

3.20 Consider a continuous-time Markov process with transition rates $\lambda(i)$ and transition probabilities $p(j|i), i, j = 1, 2, 3, \cdots, S$. Suppose that the transition probability matrix $P := [p(j|i)]_{i\in\mathcal{S}, j\in\mathcal{S}}$ changes to $P + \delta\Delta P$, and the transition rates $\lambda(i), i = 1, 2, \ldots, S$, remain unchanged. Let η be the average reward with reward function f. Develop a direct learning algorithm for $\frac{d\eta_\delta}{d\delta}$.

Solution: Suppose that the transition probability matrix $P := [p(j|i)]_{i\in\mathcal{S}, j\in\mathcal{S}}$ changes to $P + \delta\Delta P$, and the transition rate $\lambda(i), i = 1, 2, \ldots, S$, remain unchanged, we can obtain $B_\delta = \Lambda(P + \delta\Delta P - I) = B + \delta\Lambda\Delta P$. From the derivative formula $\frac{d\eta_\delta}{d\delta} = \pi(\Delta B)g$, we have

$$\frac{d\eta_\delta}{d\delta} = \pi\Lambda\Delta P g = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}} \pi(i)\lambda(i)\Delta p(j|i)g(j).$$

We consider the importance sampling technique.

$$\begin{aligned}
\frac{d\eta_\delta}{d\delta} &= \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}} \pi(i)p(j|i)\frac{\lambda(i)\Delta p(j|i)}{p(j|i)}g(j) \\
&\approx \lim_{N\to\infty}\frac{1}{T_N}\sum_{n=0}^{N-1}\frac{\lambda(X_n)\Delta p(X_{n+1}|X_n)}{p(X_{n+1}|X_n)}S_n\int_{T_n}^{T_n+T} f(X_t)dt, \qquad \text{w.p.1,}
\end{aligned}$$

where S_n is the sojourn time that the process stays at state X_n.

3.21 Consider a closed Jackson network consisting of M servers and N customers with mean service times $\bar{s}_i, i = 1, 2, \ldots, S$, and routing probabilities $q_{i,j}, i, j = 1, 2, \ldots, M$. let

$$\eta_T^f = \lim_{L\to\infty}\frac{1}{T_L}\int_0^{T_L} f(\mathbf{N}(t))dt$$

be the time-average performance. Suppose that the routing probabilities change to $q_{i,j} + \delta\Delta q_{i,j}, i, j = 1, 2, \ldots, M$. Develop a direct learning algorithm for the derivative of the time-average reward using performance potentials. Use the intuition explained in Section 2.1.3 to develop the performance derivative formula.

Solution: For the closed Jackson network, the state is the number of customers at each server, which is denoted as $\boldsymbol{n} = (n_1, \ldots, n_M)$. We assume $q_{ii} = 0$ and $\Delta q_{ii} = 0$. Define $\mu_i = \frac{1}{\bar{s}_i}$ and $\mu(\boldsymbol{n}) = \sum_{i=1}^{M} \epsilon(n_i)\mu_i$. The infinitesimal matrix of the closed Jackson network is

$$a_{\boldsymbol{nm}} = \begin{cases} \epsilon(n_i)\mu_i q_{ij}, & \boldsymbol{m} = \boldsymbol{n}_{i,j}, i \neq j; \\ -\mu(\boldsymbol{n}), & \boldsymbol{m} = \boldsymbol{n}; \\ 0, & \text{otherwise.} \end{cases}$$

where $\boldsymbol{n}_{i,j} = (n_1, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_M)$. Thus, we can easily obtain the elements of ΔB, $\Delta B(\boldsymbol{n}, \boldsymbol{m})$, when the routing probabilities change to $q_{i,j} + \delta \Delta q_{i,j}, i, j = 1, 2, \ldots, M$,

$$\Delta B(\boldsymbol{n}, \boldsymbol{m}) = \begin{cases} \epsilon(n_i)\mu_i \Delta q_{ij}, & \boldsymbol{m} = \boldsymbol{n}_{i,j}, i \neq j; \\ 0, & \boldsymbol{m} = \boldsymbol{n}; \\ 0, & \text{otherwise.} \end{cases}$$

and the transition probability of embedded Markov chain

$$p(\boldsymbol{m}|\boldsymbol{n}) = \begin{cases} \frac{\epsilon(n_i)\mu_i q_{ij}}{\mu(\boldsymbol{n})}, & \boldsymbol{m} = \boldsymbol{n}_{i,j}, i \neq j; \\ 0, & \text{otherwise.} \end{cases}$$

We assume the station which has a service completion at the k-th transition is denoted by c_k and the station which has an arrival right after the k-th transition is denoted by a_k. Then according to the derivative formula and using importance sampling similarly to (**??**), we have

$$\begin{aligned} \frac{d\eta_\delta}{d\delta} &= \pi \Delta B g \\ &= \sum_{\boldsymbol{n} \in \mathcal{S}} \sum_{\boldsymbol{m} \in \mathcal{S}} \pi(\boldsymbol{n}) \Delta B(\boldsymbol{n}, \boldsymbol{m}) g(\boldsymbol{m}) \\ &= \sum_{\boldsymbol{n} \in \mathcal{S}} \sum_{\boldsymbol{m} \in \mathcal{S}} \pi(\boldsymbol{n}) p(\boldsymbol{m}|\boldsymbol{n}) \frac{\Delta B(\boldsymbol{n}, \boldsymbol{m})}{p(\boldsymbol{m}|\boldsymbol{n})} g(\boldsymbol{m}) \\ &\approx \lim_{K \to \infty} \frac{1}{T_K} \sum_{k=0}^{K-1} \frac{\mu(\boldsymbol{n}_k) \Delta q(a_k|c_k)}{q(a_k|c_k)} S_k(\boldsymbol{n}_k) \int_{T_k}^{T_k + T} f(\mathbf{N}(t)) dt, \qquad \text{w.p.1.} \end{aligned}$$

The intuitive explanation of the performance derivative formula is the same as the solution of Problem 9.14.

4

Solutions to Chapter 4

4.1 Consider a discrete-time $M/M/1$ queue. The system state at time $l \geq 0$ is denoted as $X_l = n, l = 0, 1, \cdots$, with n being the number of customers in the server. The arrival rate is reflected by the transition probabilities $p(X_{l+1} = n+1|X_l = n) = r, 0 < r < 1, n = 0, 1, \cdots$ and $l = 0, 1, \ldots$. The service rate depends on the number of customers in the server and is reflected by $p(X_{l+1} = n-1|X_l = n) = \mu_n, 0 < \mu_n < 1-r, n = 1, 2, \ldots$. When the system is at state n and with service rate μ_n, the cost is $\alpha n + \beta\mu_n$, in which αn represents the cost for waiting time, and $\beta\mu_n$ represents the cost for the service. We wish to minimize the average cost by choosing the right service rates $\mu_n, n = 1, 2, \ldots$, among all the available choices. Model this problem as a Markov decision process.

Solution:

Markov decision process contains five parts: the state space, the (available) action space, the transition probability, the cost (reward, gain) and the criterion. For this problem, the state of MDP is the number of customers n in the server, thus, the state

space is $\mathcal{S} = \{0, 1, 2, \cdots\}$. The available action space $\mathcal{A}(n)$ at state n is the real space $\Re$. An action μ_n can be taken from $\Re$ when the state is n, i.e. $d(n) = \mu_n$. The transition probability from state n to $n+1$ under policy d is $p(X_{l+1} = n+1|X_l = n) = r$, which is not related with action $d(n)$. The transition probability from state n to $n-1$ under policy d is $p(X_{l+1} = n-1|X_l = n, d(n)) = \mu_n$. Since $\mu_n < 1 - r$, the system can transit from state n to itself with probability $1 - r - \mu_n$. The other transition probabilities are 0. The reward is $\alpha n + \beta\mu_n$ when the state is n. The optimization objective is the average cost vector, whose ith component is defined as

$$\eta(i) = \lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} E\left\{\alpha X_l + \beta d(X_l)|X_0 = i\right\}. \tag{4.1}$$

4.2 A retailer orders N pieces of merchandize every evening based on the stock left on that day. The every day's demand on the merchandize can be described by an integer random variable with distribution $p_n, n = 0, 1, \ldots$. The retailer earns c_1 dollars for every piece sold, and s/he suffers a penalty of c_2 dollars for each piece left in every evening. The retailer wishes to make the right order to maximize his/her earnings in a long term. Model the problem as an MDP.

Solution:

The state of MDP is the number of merchandize on the stock left every day, then the state space is $\mathcal{S} = \{0, 1, 2, \cdots\}$. The action is how much merchandize the retailer orders. So, the available action space at state n is $\mathcal{A}(n) = \{0, 1, 2, \cdots\}$. The policy is that the retailer order $d(X_l)$ pieces of merchandize for tomorrow when there are X_l pieces of merchandize on the stock left at time l. The transition probability under the policy is $p[X_{l+1} = n|X_l = m, d(X_l)] = p_{m+d(m)-n}$. The reward is

$$f(X_l, X_{l+1}, d(X_l)) = c_1[X_l + d(X_l) - X_{l+1}] - c_2 X_{l+1}. \tag{4.2}$$

The optimization objective is his/her earning in a long term. We can use the discounted reward to measure his/her earning in a long term. That is, the optimization objective is the discounted reward vector, whose ith component is

$$\eta_\alpha(i) = \lim_{L\to\infty} E\{\sum_{l=0}^{L} \alpha^l f(X_l, X_{l+1}, d(X_l))|X_0 = i\}, \quad 0 < \alpha < 1. \tag{4.3}$$

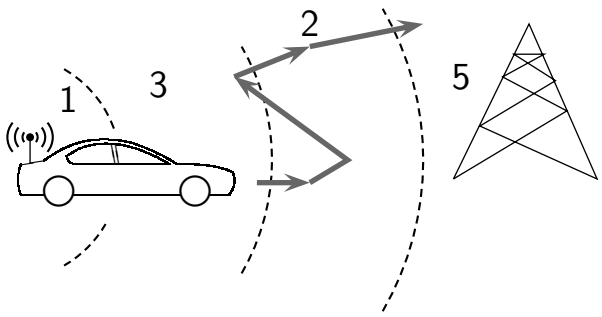

Figure 4.1: A Wireless Communication System

4.3 A mobile phone user travels through different regions shown in Figure 4.9; each region is characterized into one of the M classes according to the transmission condition in the region. In a region with a "bad" condition, the transmission of the signals requires a high power and the bit error rate is also high; therefore, the mobile phone user may prefer to delay the transition, by transmitting fewer bits, until s/he reaches a better region. On the other hand, the transmission can not be postponed for too long. In a class i region, $i = 1, 2, \cdots, M$, if the mobile phone has n bits in its buffer, the user may choose different level of powers, denoted as $d(i, n), i = 1, 2, \cdots, M$, and $n = 0, 1, \cdots$. Time is discrete and is denoted as $l = 1, 2, \cdots$. When the mobile phone is in a class i region and there are n bits in its buffer, if power $d(i, n)$ is used, then the number of the correctly transmitted bit in the time slot, k has a distribution $q_k^{d(i,n)}, k = 0, 1, \cdots, n$, $\sum_{k=0}^{n} q_k^{d(i,n)} = 1$. When the user is in class i region in one time slot, s/he will travel to class j region in the next time slot with probability p_{ij}, $i, j = 1, 2, \cdots, M$. In each time slot, the user generates r bits with probability of p_r, $\sum_{r=0}^{\infty} p_r = 1$. The cost function is $f(i, n) = \alpha n + \beta_i d(i, n)$, where β_i is the cost per unit of power in a class i region and α represents a weighting factor between the cost of power and the queue length. Model the problem as a discrete MDP.

Solution: The state of this problem is the region that the user stays and the number of bits in the buffer. Thus the state space is $S = \{(i, n)|i = 1, 2, \ldots, M; n = 0, 1, \ldots, \}$, where i denotes the region and n denotes the number of bits in the buffer. The action is using the different levels of powers. The user can choose different levels of powers $d(i, n)$ when the state is (i, n), which is the policy of MDP. The transition probability form state (i, n) to state (j, m) is $p[X_{l+1} = (j, m)|X_l = (i, n), d(i, n)] = p_{ij} \sum_{r,k:r-k=m-n} p_r q_k^{d(i,n)}$, $k = 0, 1, 2, \ldots, n, r = 0, 1, 2, \ldots$. The cost function is $f[(i, n), d(i, n)] = \alpha n + \beta_i d(i, n)$. We can make the average cost performance as the optimization criterion, whose (i, n)-th

component is

$$\eta^d(i,n) = \lim_{L\to\infty} \frac{1}{L} E\left\{ \sum_{l=0}^{L-1} f[X_l, d(X_l)] | X_0 = (i,n) \right\}. \tag{4.4}$$

4.4 Consider a closed network consisting of M single-server stations and N customers. Let n_i be the number of customers in the server $i, i = 1, 2, \cdots, M$, and $\mathbf{n} := (n_1, n_2, \cdots, n_M)$. The service rate of server $i, i = 1, 2, \cdots, M$, depends on the system "state" $\mathbf{n}$ and is denoted as $\mu_{i,\mathbf{n}}$. That is, if at time $t \in [0, \infty)$ the system is state $\mathbf{n}$. then server i completes its service to its customer in $[t, t+\Delta t)$ with probability $\mu_{i,\mathbf{n}}\Delta t$. After a customer completes its service at server i, the customer will transit to server j with probability $q_{ij}, i, j = 1, 2, \ldots, M$. We may control the service rates $\mu_{i,\mathbf{n}}$, $i = 1, 2, \cdots, M$, $\mathbf{n} \in \mathcal{S} := \{(n_1, \cdots, n_M) : \sum_{k=1}^{M} n_k = N\}$, to optimize a properly defined average reward η. We assume that the reward function f is independent of $\mu_{i,\mathbf{n}}$.

a. Model the problem as a Markov decision process.

b. Suppose that the service rate of server i, $i = 1, 2, \ldots, M$, depends on the number of customers in server i, n_i, and is denoted as μ_{i,n_i}, and we may control the load-dependent service rates μ_{i,n_i}, $n_i = 1, 2, \ldots, N, i = 1, 2, \ldots, M$, to optimize an average reward. Can we model this problem as a standard MDP? Why?

Solution:

a. For this problem, we need use continuous time Markov decision process to model it. The state space is $\mathcal{S} = \{\mathbf{n} = (n_1, n_2, \ldots, n_M) | \sum_{k=1}^{M} n_k = N\}$. The available action space $\mathcal{A}(\mathbf{n})$ at state $\mathbf{n}$ is the real space R^M. An action $\mu_{i,\mathbf{n}}$ at server i can be taken from R when the state is $\mathbf{n}$, i.e. $d(\mathbf{n}) = (\mu_{1,\mathbf{n}}, \mu_{2,\mathbf{n}}, \cdots, \mu_{M,\mathbf{n}})$. The service rate at server i when system stays at state $\mathbf{n}$ is $\mu_{i,\mathbf{n}}$ and the routing probability is q_{ij}. From the results of closed network, we can easily obtain the infinitesimal generator as follows:

$$b_{\mathbf{nm}} = \begin{cases} \epsilon(n_i)\mu_{i,\mathbf{n}} q_{ij}, & \mathbf{m} = \mathbf{n}_{i,j}, i \neq j; \\ \sum_{i=1}^{M} \epsilon(n_i)\mu_{i,\mathbf{n}} q_{i,i}, & \mathbf{m} = \mathbf{n}; \\ 0, & \text{otherwise.} \end{cases}$$

where $\mathbf{n}_{ij} = (n_1, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_M)$. The reward function is $f(\mathbf{n})$. The optimization criterion is the average reward defined as follows:

$$\eta^d(\mathbf{n}_0) = \lim_{T\to\infty} \frac{1}{T} E\left\{ \int_{t=0}^{T} f(X_t) \Big| X_0 = \mathbf{n}_0 \right\}.$$

When the system is ergodic, $\eta(\mathbf{n}_0)$ is independent of the initial state $\mathbf{n}_0$.

b. We cannot model this problem as a standard MDP. Since the service rate at server i depends only on the number of customers in server i, the action choice at the different states is not independent. For example, We consider the case of 3 servers and 4 customers. The service rates μ_{1,n_1} of server 1 at states $(1, 2, 1)$ and $(1, 1, 2)$ are the same. This point is different from the standard Markov decision processes.

4.5 Derive the average-reward difference formula for continuous-time ergodic Markov processes with a finite state space and a finite number of actions, and derive the policy iteration algorithm from it.

Solution:

For the continuous time Markov chain, we have the Poisson equation as follows:

$$Bg = -f + \eta e.$$

Left-multiplying on the both sides of Poisson equation by π', using $\pi' e = 1$, we get

$$\pi' Bg = -\pi' f + \pi' e\eta = -\pi' f + \eta.$$

That is,

$$\eta = \pi' Bg + \pi' f.$$

By $\pi' B' = 0$ and $\pi' f' = \eta'$, we have

$$\eta' - \eta = \pi'(f' - Bg - f) = \pi'[(f' + B'g) - (f + Bg)].$$

From the aforementioned average performance difference formula, it is natural to propose the following *Policy Iteration Algorithm.*

1. Guess an initial policy d_0, set $k = 0$.

2. (Policy evaluation) Obtain the potential g^{d_k} by solving the continuous time Poisson equation $B^{d_k} g^{d_k} = -f^{d_k} + \eta^{d_k} e$.

3. (Policy improvement) Choose

$$d_{k+1} = arg\{\max_{d\in\mathcal{E}}[f^d + B^d g^{d_k}]\}, \tag{4.5}$$

component-wisely (i.e., to determine an action for each state). If at a state i, action $d_k(i)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$.

4. If $d_{k+1} = d_k$, stop; otherwise set $k := k + 1$ and go to step 2.

4.6 Derive the bias-difference formula for continuous-time ergodic Markov processes with a finite state space and a finite number of actions, and derive the policy iteration algorithm from it.

Solution:

On the condition of $\eta' = \eta = \eta^*$, by the average performance difference formula in Problem 4.5 and $\pi' > 0$, similarly to the proof of Lemma 4.1, we can obtain (in fact, (4.6) is the sufficient and necessary condition such that $\eta' = \eta = \eta^*$)

$$B'g + f' = Bg + f. \tag{4.6}$$

By Poisson equation $Bg = -f + \eta e$ and $B'g' = -f' + \eta e$, we get

$$Bg + f = B'g' + f'. \tag{4.7}$$

Combining (4.6) and (4.7), we get $B'(g' - g) = 0$. Since the continuous-time Markov process is ergodic, we obtain $g' - g = ce$ for any constant c.

Next we need to specify the constant c. Since g and g' are the biases, we have $\pi'g' = 0$. By $\pi'g' = \pi'(g + ce) = 0$, we get $c = -\pi'g$. By replacing f by the bias $-g$ in the Poisson equation, we have the Poisson equation for the 2nd bias

$$Bw = g.$$

By using $\pi'B' = 0$ and the above Poisson equation, we have the following bias difference formula:

$$g' - g = ce = \pi'(B' - B)we.$$

From the aforementioned bias difference formula, we can derive *the policy iteration algorithm for a bias-optimal policy*:

1. Starting with any gain-optimal policy d_0, which may be obtained from the gain-optimal policy iteration algorithm, set $k = 0$.

2. Determine $\mathcal{D}_0$ by

$$\mathcal{D}_0(i) = \left\{ a \in \mathcal{A}(i) : f(i,a) + \sum_{j=1}^{S} B^a(j|i) g^{d_0}(j) = f^{d_0}(i) + \sum_{j=1}^{S} B^{d_0}(j|i) g^{d_0}(j) \right\}.$$

3. Obtain the bias g^{d_k} by solving $B^{d_k} g^{d_k} = -f^{d_k} + \eta^{d_k} e$ and $\pi^{d_k} g^{d_k} = 0$, and bias-potential w^{d_k} by solving $B^{d_k} w^{d_k} = g^{d_k}$.

4. Choose

$$d_{k+1} = arg\{\max_{d \in \mathcal{E}_0} [B^d w^{d_k}]\},$$

component-wisely (i.e., to determine an action for each state). If at a state i, action $d_k(i)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$.

5. If $d_{k+1} = d_k$, stop; otherwise set $k := k + 1$ and go to step 3.

4.7 Policy iteration requires the actions at different states should be chosen independently. Consider the following optimization problem. The state space consists of $2S$ states denoted as (i, j), $i = 1, 2, \cdots, S$, $j = 1, 2$. The same action has to be taken when the system is at state $(i, 1)$ or $(i, 2)$ for the same i, $i = 1, 2, \cdots$. Thus, if action α is taken at both $(i, 1)$ and $(i, 2)$, then the transition probabilities from both state $(i, 1)$ and $(i, 2)$, $p^\alpha(\cdot|(i, 1))$ and $p^\alpha(\cdot|(i, 2))$ are determined simultaneously.

a. Explain why the standard policy iteration algorithm does not apply to this problem.

b. Let $\pi(i) := \pi(i, 1) + \pi(i, 2)$ be the steady-state marginal distribution and $\pi(j|i) = \frac{\pi(i,j)}{\pi(i)}$ be the steady-state conditional probabilities, $i = 1, 2, \cdots, S$, $j = 1, 2$. In this problem, a policy determines an action based on the first component of the state, i. Consider any two policies $h(i)$ and $d(i)$. We assume that these conditional probabilities are the same for all policies. Thus, $\pi^d(j|i) = \pi^h(j|i)$ for all $i = 1, 2, \ldots, S$ and $j = 1, 2$. Now we have the average performance difference formula

$$\eta^h - \eta^d = \sum_{i=1}^{S} \pi^h(i) \left\{ \sum_{j=1}^{2} \pi^d(j|i) \Big\{ \Big[f^h(i,j) + \sum_{i'=1}^{S} \sum_{j'=1}^{2} p^h[(i,j),(i'.j')] g^d(i',j') \Big]$$

$$-\Big[f^d(i,j)+\sum_{i'=1}^{S}\sum_{j'=1}^{2}p^d[(i,j),(i'.j')]g^d(i',j')\Big]\Big\}.\Bigg\}$$

The $\pi^d(j|i)$ and $g^d(i,j)$ in the big bracket do not depend on P^h. Derive a policy iteration optimization algorithm for the "aggregated" state i.

c. Can you derive a sample path based optimization algorithm for the problem in b)?

solution:

a. From the process of the policy iteration, we can find that the action choices at different states are requested to be independent, but in this problem, the action choices at different states are not independent.

b. From the aforementioned average performance difference formula, it is natural to propose the following *Policy Iteration Algorithm.*

1. Guess an initial policy d_0, set $k=0$.

2. (Policy evaluation) Obtain the potential g^{d_k} by solving the Poisson equation $(I-P^{d_k})g^{d_k}+\eta^{d_k}e=f^{d_k}$ and compute the steady-state probability π^{d_k} by $\pi^{d_k}P^{d_k}=\pi^{d_k}$ and $\pi^{d_k}e=e$, then obtain $\pi^{d_k}(j|i)$ by $\pi^{d_k}(j|i)=\frac{\pi^{d_k}(i,j)}{\pi^{d_k}(i)}$.

3. (Policy improvement) For $i=1,2,\cdots,S$, choose

$$d_{k+1}(i)=arg\max_{a\in\mathcal{A}(i)}\left\{\sum_{j=1}^{2}\pi^{d_k}(j|i)\Big[f^a(i,j)+\sum_{i'=1}^{S}\sum_{j'=1}^{2}p^a[(i,j),(i'.j')]g^{d_k}(i',j')\Big]\right\}\quad(4.8)$$

If at a state i, action $d_k(i)$ attains the maximum, then set $d_{k+1}(i)=d_k(i)$.

4. If $d_{k+1}=d_k$, stop; otherwise set $k:=k+1$ and go to step 2.

c. We can derive a sample path based policy iteration algorithm. Based on a sample path, we can estimate the potential by using the methods in Section 3.1.2 and obtain the estimation of potential $\hat{g}(i,j)$. The estimation $\hat{\pi}^{d_k}(i,j)$ of $\pi^{d_k}(i,j)$ can be obtained by $\lim_{L\to\infty}\frac{\sum_{l=0}^{L-1}I_{(i,j)}(X_l)}{L}$, then $\hat{\pi}^{d_k}(j|i)=\frac{\hat{\pi}^{d_k}(i,j)}{\sum_{j=1}^{2}\hat{\pi}^{d_k}(i,j)}$. Putting the estimations $\hat{g}^{d_k}(i,j)$ and $\hat{\pi}^{d_k}(j|i)$ into (4.8), we can complete the policy improvement by using these estimates. This method does not need the second step in the above policy iteration.

4.8 Are the following statements true?

a. When the average reward policy iteration algorithm stops at a policy $\hat{d}$, the directional performance derivative from $\hat{d}$ to any other policy in $\mathcal{D}$ is non-positive.

b. If $\hat{d}$ is a gain optimal policy, then another policy d is gain optimal, if the directional performance derivative from $\hat{d}$ to d is zero.

c. If $\hat{d}$ is a gain optimal policy, then the directional performance derivative from $\hat{d}$ to any other gain optimal policy d is zero.

d. The bias optimal policy has the largest bias in the policy space $\mathcal{D}$.

e. The difference of the biases of any two policies is a constant vector (i.e. all its components are equal).

Solution:

a. For ergodic Markov decision processes, this statement is true. The performance derivative along the direction from d_k to any policy $d \in \mathcal{D}$ is

$$\frac{d\eta_\delta}{d\delta} = \pi^{\hat{d}}[(f^d + P^d g^{\hat{d}}) - (f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}})]. \tag{4.9}$$

Since the policy iteration algorithm stops at the policy $\hat{d}$, for any policy d, we have $f^d + P^d g^{\hat{d}} \leq f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}}$. Otherwise, the algorithm cannot be stopped. Thus, from $\pi^{d_k}(i) > 0, \forall i \in \mathcal{S}$, the directional performance derivative from $\hat{d}$ to any other policy in $\mathcal{D}$ is non-positive. For the case of Multiple Markov chain, this statement is also true. The performance derivative along the direction from policy $\hat{d}$ to any policy $d \in \mathcal{D}$ is

$$\frac{d\eta_\delta}{d\delta} = (P^{\hat{d}})^*[(f^d + P^d g^{\hat{d}}) - (f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}})] + \sum_{l=0}^{\infty} P^{\hat{d}^l}(P^d - I)\eta^{\hat{d}}. \tag{4.10}$$

When the policy iteration stops at policy $\hat{d}$, then $f^d(i) + P^d g^{\hat{d}}(i) \leq f^{\hat{d}}(i) + P^{\hat{d}} g^{\hat{d}}(i)$ for all recurrent states i and $P^d \eta^{\hat{d}} \leq \eta^{\hat{d}}$. Thus, we have $\frac{d\eta_\delta}{d\delta} \leq 0$.

b. For ergodic Markov decision processes, this statement is not true. From (4.9), although $\pi^{\hat{d}} > 0$, but we can not guarantee

$$f^d + P^d g^{\hat{d}} = f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}}, \tag{4.11}$$

which is the sufficient and necessary condition that d is also the gain optimal policy. Maybe policy d makes $f^d(i) + P^d g^{\hat{d}}(i) < f^{\hat{d}}(i) + P^{\hat{d}} g^{\hat{d}}(i)$ for some i and $f^d(j) + P^d g^{\hat{d}}(j) > f^{\hat{d}}(j) + P^{\hat{d}} g^{\hat{d}}(j)$ for some j, but $\frac{d\eta_\delta}{d\delta} = 0$ still holds. Thus, this statement is not true. For the multiple Markov chain, this statement is also not true.

c. This statement is true for ergodic Markov chain. From Lemma 4.1, we know if d and $\hat{d}$ are the gain optimal policies, then

$$f^d + P^d g^{\hat{d}} = f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}}. \tag{4.12}$$

So the directional performance derivative from $\hat{d}$ to any other gain optimal policy d is zero. Thus, this statement is true for ergodic Markov chain. For the case of multiple Markov chain, this statement is not true. Since $\eta^{\hat{d}} = \eta^d$, then, $P^d \eta^{\hat{d}} = \eta^{\hat{d}}$ and $P^{d^*} \eta^{\hat{d}} = \eta^{\hat{d}}$. From the average performance difference formula, we have

$$0 = \eta^d - \eta^{\hat{d}} = \left(P^d\right)^* [(f^d + P^d g^{\hat{d}}) - (f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}})]. \tag{4.13}$$

For different polices d and $\hat{d}$, the classes of recurrent states may be different. We cannot draw a conclusion that $\left(P^{\hat{d}}\right)^* [(f^d + P^d g^{\hat{d}}) - (f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}})] = 0$ from $\left(P^d\right)^* [(f^d + P^d g^{\hat{d}}) - (f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}})] = 0$. Thus, the directional performance derivative from $\hat{d}$ to any other gain optimal policy d may not be zero.

d. This statement is not true. The bias optimal policy has the largest bias only in the set of gain-optimal policies $\mathcal{D}_0$.

e. This statement is not true. We know the difference of the biases of two policies in the set of gain-optimal policies $\mathcal{D}_0$ is a constant. However, if one of two policies is not the gain optimal policy, this conclusion cannot hold.

4.9 Let $\hat{d}$ and d be two ergodic gain optimal policies in Lemma 1. We define a randomized policy d_δ by setting $P^{d_\delta} = P^d + \delta(P^{\hat{d}} - P^d)$, $f^{d_\delta} = f^d + \delta(f^{\hat{d}} - f^d)$.

a. Let η^{d_δ} be the average reward of d_δ, prove $\eta_{d_\delta} = \eta^*$.

b. Derive a directional bias-derivative equation from d to $\hat{d}$, denoted as $\frac{dg_\delta}{d\delta}$.

c. When the bias policy iteration algorithm stops at a policy $\hat{h}$, what are the directional derivatives from this policy to other policies in $\mathcal{D}_0$?

d. Calculate the bias derivative between various policies in Example 4.1.

Solution:

a. From the performance difference formula, we have

$$\begin{aligned} \eta^{d_\delta} - \eta^d &= \pi^{d_\delta}(f^{d_\delta} + P^{d_\delta}g^d - (f^d + P^d g^d)) \\ &= \pi^{d_\delta}\delta[(f^{\hat{d}} - f^d) + (P^{\hat{d}} - P^d)g^d]. \end{aligned} \tag{4.14}$$

Since $\hat{d}$ and d are two ergodic gain optimal policies, we have $f^{\hat{d}} + P^{\hat{d}}g^d = f^d + P^d g^d$ from Lemma 1. Thus, we have $\eta^{d_\delta} = \eta^d$. That is, d_δ is also a gain optimal policy.

b. Similarly to the method in Section 4.1.2, we can obtain the following difference formula:

$$\begin{aligned} g^{d_\delta} - g^d &= \{\pi^{d_\delta}(P^{d_\delta} - P^d)w^d\}e \\ &= \{\pi^{d_\delta}\delta(P^{\hat{d}} - P^d)w^d\}e, \end{aligned}$$

where π^{d_δ} is the steady-state distribution of P^{d_δ}. Dividing by δ on both sides and letting $\delta \to 0$, we have

$$\frac{dg_\delta}{d\delta} = \{\pi^d(\hat{P} - P^d)w^d\}e.$$

c. When the bias policy iteration algorithm stops at a policy $\hat{h}$, the directional derivatives from $\hat{h}$ to other policies in $\mathcal{D}_0$ is non-positive.

d. Using Poisson equation, we have $w^{d_1} = (-1, 1)^T$, $w^{d_2} = (-0.64, 0.96)^T$, $w^{d_3} = (-0.8889, 1.7778)^T$ and $w^{d_4} = (-0.5, 1.5)^T$. The bias derivative along the direction from d_1 to d_2 is

$$(0.5, 0.5)\left\{ \begin{bmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{bmatrix} - \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \right\} \begin{bmatrix} -1 \\ 1 \end{bmatrix} e = -0.25e,$$

where $e = (1, 1)^T$.

Similarly, we can obtain the bias derivative $\frac{dg_\delta}{d\delta}|_{d_1 \to d_3} = -0.25e$, $\frac{dg_\delta}{d\delta}|_{d_1 \to d_4} = -0.5e$, $\frac{dg_\delta}{d\delta}|_{d_2 \to d_1} = 0.16e$, $\frac{dg_\delta}{d\delta}|_{d_2 \to d_3} = -0.08e$, $\frac{dg_\delta}{d\delta}|_{d_2 \to d_4} = -0.24e$, $\frac{dg_\delta}{d\delta}|_{d_3 \to d_1} = 0.4445e$, $\frac{dg_\delta}{d\delta}|_{d_3 \to d_2} = 0.2222e$, $\frac{dg_\delta}{d\delta}|_{d_3 \to d_4} = -0.2222e$, $\frac{dg_\delta}{d\delta}|_{d_4 \to d_1} = 0.5e$, $\frac{dg_\delta}{d\delta}|_{d_4 \to d_2} = 0.375e$, $\frac{dg_\delta}{d\delta}|_{d_4 \to d_3} = 0.125e$.

4.10 In Section 4.1.1, we proved that at an optimal policy the performance derivatives along the directions to all other policies are non-positive.

a. Suppose $\frac{d\eta^{d_\delta}}{d\delta} > 0$ at policy d along a direction defined by d_δ: $P^{d_\delta} = P^d + \delta\Delta P$, $f^\delta = f^d + \delta\Delta f$ with $\Delta P = P^h - P^d$, $\Delta f = f^h - f^d$. Can we claim $\eta^h > \eta^d$? If not, give a counter example. If yes, what does this imply in terms of policy iteration?

b. Prove that a policy $d \in \mathcal{D}$ is average-reward optimal if and only if at this policy the performance derivative along the directions to all other policies are non-positive.

Solution:

a. If $\frac{d\eta^{d_\delta}}{d\delta} > 0$ along the direction defined by d_δ: $P^{d_\delta} = P^d + \delta\Delta P$, $f^\delta = f^d + \delta\Delta f$ with $\Delta P = P^h - P^d$, $\Delta f = f^h - f^d$, we cannot claim $\eta^h > \eta^d$. If there exist some state i such that $(\Delta P g^d + \Delta f)(i) > 0$ and some state j such that $(\Delta P g^d + \Delta f)(j) < 0$, but $\frac{d\eta^{d_\delta}}{d\delta} = \pi^d(\Delta P g^d + \Delta f) > 0$. For this case, we cannot claim $\eta^h > \eta^d$. For example,

$$P^d = \begin{bmatrix} 0.2 & 0.1 & 0.3 & 0.4 \\ 0.5 & 0.2 & 0.1 & 0.2 \\ 0.2 & 0.3 & 0.1 & 0.4 \\ 0.4 & 0.2 & 0.2 & 0.2 \end{bmatrix}, f^d = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, P^h = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.4 \\ 0.1 & 0.4 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.2 & 0.2 \\ 0.3 & 0.1 & 0.4 & 0.2 \end{bmatrix}, f^h = \begin{bmatrix} 1 \\ 2 \\ 2.9 \\ 4 \end{bmatrix}.$$

We have

$$g^d = \begin{bmatrix} 1.3957 \\ 1.6010 \\ 3.0966 \\ 3.7711 \end{bmatrix}, (P^h - P^d)g^d + (f^h - f^d) = \begin{bmatrix} -0.1496 \\ 0.3138 \\ -0.4050 \\ 0.3196 \end{bmatrix}.$$

and $\frac{d\eta^{d_\delta}}{d\delta} = 0.0297 > 0$, but we have $\eta^h = 2.4789, \eta^d = 2.4809$, and $\eta^h - \eta^d = -0.002$.

b. We firstly prove the necessary condition ("⇒"). We use the contradiction method. If there exists a policy $\hat{d}$ such that the directional derivative $\frac{d\eta^{d_\delta}}{d\delta}$ is positive. From $\pi^d > 0$ and $\frac{d\eta^{d_\delta}}{d\delta}\Big|_{d\to\hat{d}} = \pi^d\left[P^{\hat{d}}g^d + f^{\hat{d}} - (P^d g^d + f^d)\right] > 0$, there must exist a state i such that $[P^{\hat{d}}g^d + f^{\hat{d}}](i) > [P^d g^d + f^d](i)$. Then, we create a policy d^* by setting $d^*(i) = \hat{d}(i)$ and $d^*(j) = d(j)$ for all $j \neq i$. We have

$$P^{d^*}g^d + f^{d^*} \succeq P^d g^d + f^d.$$

By using the performance difference formula, we have $\eta^{d^*} > \eta^d$. This contradicts the fact that d is an optimal policy. Thus, the necessary condition is proved.

Next, we prove the sufficient condition with the contradiction ("⇐"). If d is not the average-reward optimal, then there exists an average-reward optimal policy d^* such that

$$P^{d^*}g^d + f^{d^*} \succeq P^d g^d + f^d.$$

From $\pi^d > 0$. Thus, the performance derivative from P to P^* is positive. This is a contradiction. Therefore, d is average-reward optimal.

4.11 Suppose that $\widehat{d}$ is the gain-optimal policy with potential $g^{\widehat{d}}$ in Lemma 4.1. Then for any policy $d \in \mathcal{D}_0$, we have $f^d + P^d g^{\widehat{d}} = f^{\widehat{d}} + P^{\widehat{d}} g^{\widehat{d}}$. From this, prove that for any other policy $d' \in \mathcal{D}_0$, we have $f^d + P^d g^{d'} = f^{d'} + P^{d'} g^{d'}$, for all $d \in \mathcal{D}_0$.

Solution: Since $f^d + P^d g^{\widehat{d}} = f^{\widehat{d}} + P^{\widehat{d}} g^{\widehat{d}}$ holds for for any policy $d \in \mathcal{D}_0$, we have

$$f^d + P^d g^{\widehat{d}} = f^{d'} + P^{d'} g^{\widehat{d}} = f^{\widehat{d}} + P^{\widehat{d}} g^{\widehat{d}}, \tag{4.15}$$

for any policy $d' \in \mathcal{D}_0$. From (4.13), $g^{d'} - g^{\widehat{d}} = ce$. Putting $g^{\widehat{d}} = g^{d'} - ce$ into (4.15), we have $f^d + P^d g^{d'} = f^{d'} + P^{d'} g^{d'}$.

4.12 Prove that the second policy iteration algorithm for bias optimality in Section 4.1.2 converges to a bias-optimal policy in a finite number of iterations.

Solution:

In the process of policy iteration, $d_{k+1} \in \widetilde{\mathcal{D}}$. By the gain-optimal policy iteration algorithm, we know $\eta^{d_{k+1}} > \eta^{d_k}$ before d_k becomes a gain-optimal policy. Since the number of policies is finite, we know d_k must be a gain-optimal policy in a finite number of iterations. After that, $\tilde{\mathcal{D}}$ is the set of gain-optimal policies $\mathcal{D}_0$. According to the bias difference formula (4.15), we know the bias increases at each iteration before it stops because of $d_{k+1} \in \arg\left\{\max_{d \in \widetilde{\mathcal{D}}} P^d w^{d_k}\right\}$. Since the number of gain optimal policies is finite, the iteration procedure has to stop after a finite number of iterations. Suppose it stops at a policy denoted as $\hat{d}$. Then $\hat{d}$ must satisfy the optimality conditions $f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}} = f^d + P^d g^{\hat{d}}$ and $P^{\hat{d}} w^{\hat{d}} \geq P^d w^{\hat{d}}$, for all $d \in \mathcal{D}_0$, because otherwise for some i, we can find the next improved policy in the policy iteration. Thus, by gain difference formula and bias difference formula, we have $g^{\hat{d}} \geq g^d$ for any $d \in \mathcal{D}_0$, that is, policy $\hat{d}$ is bias optimal.

4.13 Calculate the bias-potential w in Example 4.1 for policy d_2 and then find the bias-

optimal policy by policy iteration.

Solution: The bias-potential in Example 4.1 for policy d_2 is $w^{d_2} = (-0.64, 0.96)^T$. Thus, from $[p^{\alpha_2}(\cdot|1) - p^{\alpha_1}(\cdot|1)]w^{d_2} = -0.4 < 0$ and $[p^{\beta_2}(\cdot|2) - p^{\beta_1}(\cdot|2)]w^{d_2} = -0.4 < 0$, we conclude that $d_1 = (\alpha_1, \beta_1)$ is a bias-optimal policy.

4.14 Consider a two-state Markov chain. There are two actions at state 1, corresponding to transition probabilities $(0.5, 0.5)$, and $(0.25, 0.75)$ and rewards 1 and 1.5, respectively; and there are three actions at state 2, corresponding to transition probabilities $(0.5, 0.5)$, $(0.25, 0.75)$, and $(0.75, 0.25)$ and rewards $-1, -0.5$, and -1.5, respectively. Apply policy iteration to obtain the set of gain-optimal policies and a bias-optimal policy.

Solution: From the problem, we know there are 6 policies in the policy space, which is denoted as $\{d_1, \cdots, d_6\}$.

1. Start the policy iteration from an initial policy
$$d_1 \models \left\{ P_1 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, f_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\}.$$
2. Obtain the potential $g_1 = (I - P_1 + e\pi_1)^{-1} f_1 = (1, -1)^T$.
3. Since $[(0.5, 0.5) - (0.25, 0.75)] * (1, -1)^T + (1 - 1.5) = 0$, $[(0.5, 0.5) - (0.75, 0.25)] * (1, -1)^T + (-1 + 1.5) = 0$, and $[(0.5, 0.5) - (0.25, 0.75)] * (1, -1)^T + (-1 + 0.5) = 0$, we know d_1 is a gain-optimal policy.

From the third step in the above policy iteration algorithm, we can find any policy d in the policy space satisfy $f^d + P^d g^{d_1} = f^{d_1} + P^{d_1} g^{d_1}$. Thus, the set of optimal gain policies is the whole policy space, in which there are 6 policies.

By using (4.14), we obtain the 2nd potential of policy d_1, $w^{d_1} = -(I - P_1 + e\pi_1)^{-1} g_1 = (-1, 1)^T$. Since $[(0.5, 0.5) - (0.25, 0.75)] * (-1, 1)^T = -0.5 < 0$, and $[(0.5, 0.5) - (0.75, 0.25)] * (-1, 1)^T = 0.5 > 0$, we know policy $\left\{ \begin{bmatrix} 0.25 & 0.75 \\ 0.25 & 0.75 \end{bmatrix}, \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \right\}$ has a better bias than d_1, whose 2nd bias is $(-1.5, 0.5)^T$. Since $[(0.5, 0.5) - (0.25, 0.75)] * (-1.5, 0.5)^T = -0.5 < 0$, and $[(0.75, 0.25) - (0.75, 0.25)] * (-1.5, 0.5)^T = -1 < 0$, the bias-optimal policy is $\left\{ \begin{bmatrix} 0.25 & 0.75 \\ 0.25 & 0.75 \end{bmatrix}, \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \right\}$.

4.15 For multi-chains, prove

a. There are more than one solution to $(I - P)u = 0$.

b. The Poisson equation $(I - P)g + \eta = f$ and the normalization condition $P^*g = 0$ uniquely determine the bias of the Markov chain.

Solution:

a. Suppose that P is in a canonical form. Then P can be writhen as

$$P = \begin{bmatrix} P_1 & 0 & 0 & \cdots & \cdot & 0 \\ 0 & P_2 & 0 & \cdots & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & P_m & 0 \\ R_1 & R_2 & R_3 & \cdots & R_m & R_{m+1} \end{bmatrix}.$$

For any constant group of $c_1, \ldots, c_m$, then $u = (c_1 e_1^T, \ldots, c_m e_m^T, -(\sum_{i=1}^m c_i (I - R_{m+1})^{-1} R_i e_i)^T)^T$ is one of the solutions to $(I - P)u = 0$, where $e_i = (1, \ldots, 1)^T$ whose dimension is the same as P_i, $i = 1, \ldots, m$.

b. By $P^*g = 0$ and the Poisson equation, we obtain $(I - P + P^*)g = f - \eta$. Since $I - P + P^*$ is invertible (cf. (B.12) in Appendix B.3), then the bias of the Markov chain is uniquely determined by $g = (I - P + P^*)^{-1}(f - \eta)$.

4.16 Suppose d and h are the two policies satisfying conditions (a) and (b) in Comparison Lemma (4.41). Prove

a. If in addition to (a) and (b), we have $v(i) = [f^h(i) + (P^h g^d)(i)] - [f^d(i) + (P^d g^d)(i)] > 0$ for some recurrent state i of P^h, then $\eta^h \succeq \eta^d$.

b. If in addition to (a) and (b) , we have $P^h \eta^d \neq \eta^d$, then $\eta^h \succeq \eta^d$.

[**solution**]

a. From condition (a) in Lemma (4.41), we have $u = P^h \eta^d - \eta^d \geq 0$. Because $P^{h^*} P^h = P^{h^*}$, we have $P^{h^*} u = 0$. Thus, from Lemma (4.41), $u(i) = 0$ for all recurrent states i of P^h. Next, it follows from condition (b) that $v(i) = [f'(i) + (P'g)(i)] - [f(i) + (Pg)(i)] \geq 0$ for all recurrent states of P^h. If in addition to (a) and (b), we have $v(i) = [f^h(i) +$

$(P^h g^d)(i)] - [f^d(i) + (P^d g^d)(i)] > 0$ for some recurrent state i of P^h. From the canonical form of P^{h^*}, we have $P^{h^*} v \succeq 0$. On the other hand, since $P^h \eta^d \geq \eta^d$, and so $P^{h^k} \eta^d \geq \eta^d$ for all $k \geq 1$. Therefore, by (4.27) we get $P^{h^*} \eta^d \geq \eta^d$. Finally, by the average performance difference formula, we have $\eta^h - \eta^d = P^{h^*} v + (P^{h^*} - I)\eta^d \geq P^{h^*} v \succeq 0$.

b. From Lemma (4.41), we know $\eta^h \geq \eta^d$. Now we just prove that $\eta^h \neq \eta^d$. Suppose $\eta^h = \eta^d$, then $P^h \eta^d = P^h \eta^h = \eta^h = \eta^d$, which conflicts with $P^h \eta^d \neq \eta^d$. Moreover, we can also prove this problem as follows:

From (a), we know $P^h \eta^d \geq \eta^d$. If $P^h \eta^d \neq \eta^d$, we know $P^h \eta^d \succeq \eta^d$. Because $(P^{h^*} - I)\eta^d = \sum_{l=0}^{\infty} P^{h^l}(P^h - I)\eta^d$, we have $(P^{h^*} - I)\eta^d \geq (P^h - I)\eta^d \succeq 0$. From (b), we can prove $P^{h^*} v \geq 0$. Thus, we have $\eta^h \succeq \eta^d$ from the average-reward difference formula (4.36).

4.17 Find both the gain- and bias- optimal policies using policy iteration for the multi-chain MDP in Example 4.6.

Solution:

Denote policy P^{α_1} if we choose α_1 at state 1, policy P^{α_2} if we choose α_2 at state 1. Then we have

$$P^{\alpha_2} = \begin{bmatrix} 0.1 & 0.9 \\ 0 & 1 \end{bmatrix}, f^{\alpha_2} = \begin{bmatrix} 100 \\ 0 \end{bmatrix}, (P^{\alpha_2})^* = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \eta^{\alpha_2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, g^{\alpha_2} = \begin{bmatrix} \frac{1000}{9} \\ 0 \end{bmatrix}.$$

$$P^{\alpha_1} = \begin{bmatrix} 0.99 & 0.01 \\ 0 & 1 \end{bmatrix}, f^{\alpha_1} = \begin{bmatrix} 100 \\ 0 \end{bmatrix}, (P^{\alpha_1})^* = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \eta^{\alpha_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, g^{\alpha_1} = \begin{bmatrix} 10000 \\ 0 \end{bmatrix}.$$

1. Suppose we start from policy P^{α_2} in the policy iteration.

2. solve Poisson equation, we get $\eta^{\alpha_2} = (P^{\alpha_2})^* f^{\alpha_2} = (0, 0)^T$ and $g^{\alpha_2} = (I - P^{\alpha_2} + (P^{\alpha_2})*)^{-1}(f^{\alpha_2} - \eta^{\alpha_2}) = (\frac{1000}{9}, 0)^T$.

3. Since $f^{\alpha_1} + P^{\alpha_1} g^{\alpha_2} \succeq f^{\alpha_2} + P^{\alpha_2} g^{\alpha_2}$ and there are only two policies, then we get the gain-optimal policy P^{α_1}, which is also the bias-optimal policy.

4.18 Consider a Markov chain studied in Problem 2.20 with transition probability matrix

$$P = \begin{bmatrix} B & b \\ 0 & 1 \end{bmatrix},$$

where B is an $(S-1)\times(S-1)$ irreducible matrix, $b > 0$ is an $(S-1)$ dimensional column vector, 0 represents an $(S-1)$ dimensional row vector whose all components are zero. The last state S is an absorbing state. Set $f(S) = 0$. Clearly, the long-run average reward for this Markov chain is $\eta = 0$. The total reward obtained before reaching the absorbing state, $E\{\sum_{l=0}^{\infty} f(X_l)|X_0 = i\}$, can be viewed as the bias for the problem:

$$g(i) = E\left\{\sum_{l=0}^{\infty} f(X_l)\Big|X_0 = i\right\}.$$

The Poisson equation for $g = (g(1), \ldots, g(S))^T$ has been derived in the problem 2.19.

a. Derive the bias-difference equation for any two policies h and d.

b. Derive a policy iteration algorithm for the bias-optimal policy.

This problem indicates that optimization of the total reward of Markov chains with absorbing states can be solved by the policy iteration for bias optimal policies.

Solution:

a. For the Markov chain with an absorbing state in the problem, the steady state probability under any policy is $\pi = (0, 0, \ldots, 0, 1)$. Thus we have $g^d(S) = 0$ for the bias g^d under any policy d from $\pi^d g^d = 0$. Denote $g^d = ((g_1^d)^T, 0)^T$ and $f = ((f_1^d)^T, 0)^T$. From Problem 2.19, we have $(I - B^d)g_1^d = f_1^d$. Next, we derive the difference equation for $g_1^h - g_1^d$.

$$\begin{aligned} g_1^h - g_1^d &= (B^h g_1^h + f^h) - (B^d g_1^d + f^d) \\ &= (B^h g_1^d + f_1^h) - (B^d g_1^d + f_1^d) + B^h(g_1^h - g_1^d). \end{aligned}$$

Thus, we have the following bias difference formula

$$g_1^h - g_1^d = (I - B^h)^{-1}[(B^h g_1^d + f_1^h) - (B^d g_1^d + f_1^d)]. \tag{4.16}$$

b. **Policy iteration algorithm:**

1. Guess an initial policy d_0, set $k = 0$.

2. (Policy evaluation) Obtain the potential $g_1^{d_k}$ by solving $(I - B^{d_k})g_1^{d_k} = f_1^{d_k}$.

3. (Policy improvement) Choose

$$d_{k+1} \in arg\{\max_{d\in\mathcal{D}}[f_1^d + B^d g_1^{d_k}]\},$$

component-wisely (i.e., to determine an action for each state). If at a state i, action $d_k(i)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$.

4. If $d_{k+1} = d_k$, stop; otherwise, set $k = k + 1$ and go to step 2.

4.19 For the MDPs with discounted performance criterion,

a. Prove the performance difference formula (4.73) and (4.74),

b. Prove that in (4.77), if $d' \neq d$, then $\eta_\beta^{d'} \succeq \eta_\beta^d$.

c. Prove the convergence of the policy iteration algorithm.

Solution:

a. We have $\eta_\beta^d = (1-\beta)(I - \beta P^d)^{-1} f^d$. That is, $\eta_\beta^d - \beta P^d \eta_\beta^d = (1-\beta) f^d$. We obtain

$$\begin{aligned}
\eta_\beta^h - \eta_\beta^d &= (1-\beta)f^h + \beta P^h \eta_\beta^h - [(1-\beta)f^d + \beta P^d \eta_\beta^d] \\
&= (1-\beta)f^h + \beta P^h \eta_\beta^d - [(1-\beta)f^d + \beta P^d \eta_\beta^d] + \beta P^h(\eta_\beta^h - \eta_\beta^d).
\end{aligned}$$

$$\Longrightarrow (I - \beta P^h)(\eta_\beta^h - \eta_\beta^d) = (1-\beta)(f^h - f^d) + \beta(P^h - P^d)\eta_\beta^d.$$

Since $I - \beta P^h$ is invertible, we obtain

$$\eta_\beta^h - \eta_\beta^d = (I - \beta P^h)^{-1}[(1-\beta)(f^h - f^d) + \beta(P^h - P^d)\eta_\beta^d]. \tag{4.17}$$

This is (4.73).

Since we also have $\eta_\beta^d = (1-\beta)g_\beta^d + \beta\eta^d$ (similar to (2.41)), then

$$\begin{aligned}
&(\eta_\beta^h - \eta_\beta^d) \\
=\ & (I - \beta P^h)^{-1}\{(1-\beta)(f^h - f^d) + \beta(P^h - P^d)[(1-\beta)g_\beta^d + \beta\eta^d]\} \\
=\ & (1-\beta)(I - \beta P^h)^{-1}\{(f^h - f^d) + \beta(P^h - P^d)g_\beta^d\} + \beta^2(I - \beta P^h)^{-1}(P^h - P^d)\eta^d \\
=\ & (1-\beta)(I - \beta P^h)^{-1}[(f^h + \beta P^h g_\beta^d) - (f^d + \beta P^d g_\beta^d)] + \beta^2(I - \beta P^h)^{-1}(P^h - I)\eta^d.
\end{aligned}$$

We have obtained (4.74).

b. Since $d' \neq d$, from (4.77), we have $Q_\beta^d(i, d'(i)) \geq Q_\beta^d(i, d(i)), i \in \mathcal{S}$ and for at least one state i, $Q_\beta^d(i, d'(i)) > Q_\beta^d(i, d(i))$. In a vector form, we have $Q_\beta^d(d') \succeq Q_\beta^d(d)$, where $Q_\beta^d(d') = (Q_\beta^d(1, d'(1)), \ldots, Q_\beta^d(S, d'(S)))^T$. From the discounted reward difference formula (4.73), we have

$$\eta_\beta^{d'} - \eta_\beta^d = (I - \beta P^{d'})^{-1}(Q_\beta^d(d') - Q_\beta^d(d)).$$

Since $(I - \beta P^{d'})^{-1} = I + \sum_{k=1}^\infty \beta^k (P^{d'})^k \geq I$, we obtain $\eta_\beta^{d'} \succeq \eta_\beta^d$.

c. From the result of Part b), we know $\eta_\beta^{d_{k+1}} \succeq \eta_\beta^{d_k}$ if $d_{k+1} \neq d_k$. That is, the discounted reward strictly increase during the policy iteration procedure. Since the policy is finite, we know the policy iteration algorithm must stop in a finite number of steps.

4.20 In (4.53), the bias potential w is defined as the potential of the bias g satisfying $P^* g = 0$. We can also define a potential of potential by using the potential g, which is only up to an additive vector u satisfying $(I - P)u = 0$, as follows:

$$(I - P)w - P^* g = -g.$$

a. Prove that the potential of potential defined in this way is the same as the bias potential defined in (4.53).

b. Define the nth potential by using the $(n-1)$th potential g_{n-1}, and prove that this definition is the same as (4.78).

Solution:

a. The potential is up to an additive vector u satisfying $(I - P)u = 0$. From $u = Pu$, we have $u = P^* u$. We assume $\tilde{g}$ is a bias. Then, any potential $g = \tilde{g} + u$. If we define a potential of potential by

$$(I - P)w - P^* g = -g$$

which can also be rewritten as

$$(I - P)w - P^*(\tilde{g} + u) = -(\tilde{g} + u).$$

From $P^* u = u$ and $P^* \tilde{g} = 0$ (from the definition of the bias), we have

$$(I - P)w = -\tilde{g},$$

which is the same as the definition in (4.53).

b. We can also define the nth potential by using the $(n-1)$th potential g_{n-1}, which is only up to an additive vector u with $(I-P)u=0$, as follows:

$$(I-P)g_n - P^* g_{n-1} = -g_{n-1}.$$

The potential is up to an additive vector u satisfying $(I-P)u=0$. From $u=Pu$, we have $u=P^*u$. We assume $\tilde{g}_{n-1}$ is a bias. Then, any $(n-1)$th potential $g_{n-1}=\tilde{g}_{n-1}+u$. If we define a potential of potential by

$$(I-P)g_n - P^* g_{n-1} = -g_{n-1}$$

which can also be rewritten as

$$(I-P)g_n - P^*(\tilde{g}_{n-1}+u) = -(\tilde{g}_{n-1}+u).$$

From $P^*u=u$ and $P^*\tilde{g}_{n-1}=0$ (from the definition of the bias), we have

$$(I-P)g_n = -\tilde{g}_{n-1},$$

which is the same as the definition in (4.78).

4.21 Derive a general bias difference equation for $g^h - g^d$, when $\eta^h \neq \eta^d$, for ergodic chains. Discuss whether we can use this equation to derive policy iteration algorithm.

Solution: From the Poisson equation (4.9), we have

$$\begin{aligned} g^h - g^d &= (f^h + P^h g^h - \eta^h e) - (f^d + P^d g^d - \eta^d e) \\ &= (f^h + P^h g^d) - (f^d + P^d g^d) + P^h(g^h - g^d) - (\eta^h - \eta^d)e. \end{aligned}$$

Thus, we have

$$(I-P^h)(g^h - g^d) = (f^h + P^h g^d) - (f^d + P^d g^d) - (\eta^h - \eta^d)e. \tag{4.18}$$

From the definition of the bias, we know $\pi^h(g^h - g^d) = -\pi^h g^d = \pi^h(P^h - P^d)w^d$. Combining with (4.18), we have

$$\begin{aligned} & (I - P^h + e\pi^h)(g^h - g^d) \\ = \; & (f^h + P^h g^d) - (f^d + P^d g^d) - (\eta^h - \eta^d)e + e\pi^h(P^h - P^d)w^d. \end{aligned}$$

From $(I - P^h + e\pi^h)^{-1}e = e$, we have

$$\begin{aligned} & g^h - g^d \\ = \; & (I - P^h + e\pi^h)^{-1}[(f^h + P^h g^d) - (f^d + P^d g^d)] + [\pi^h(P^h - P^d)w^d + \eta^h - \eta^d] \end{aligned} \tag{4.19}$$

We cannot use this equation to derive policy iteration algorithm. On one hand, we cannot determine whether $(I - P^h + e\pi^h)^{-1}$ is non-negative, so we cannot determine whether the first item in this equation is larger than 0. On the other hand, we cannot decouple the effect of two terms on the right hand side of (4.19).

4.22 This problem helps to understand the bias optimality. First, if $\widehat{d}$ and its gain and potential (not necessary bias) $\eta^{\widehat{d}}$ and $g^{\widehat{d}}$ satisfy (4.60) and (4.61), then $\eta^{\widehat{d}} = \eta^*$ is the optimal gain (and $g^{\widehat{d}}$ may not be optimal), and

$$\widehat{\mathcal{A}}_0(i) = \mathcal{A}_0^*(i) := \left\{ a \in \mathcal{A}(i) : \; \sum_{j \in \mathcal{S}} p^a(j|i)\eta^*(j) = \eta^*(i) \right\},$$

and

$$\widehat{\mathcal{A}}_1(i) := \left\{ a \in \widehat{\mathcal{A}}_0(i) : \; \eta^*(i) + g^{\widehat{d}}(i) = f(i,a) + \sum_{j \in \mathcal{S}} p^a(j|i) g^{\widehat{d}}(j) \right\}.$$

Now let $d \in \times_{i \in \mathcal{S}} \widehat{\mathcal{A}}_1(i)$. Then by definition we have

$$\begin{aligned} P^d \eta^* &= \eta^*, \\ f^d + P^d g^{\widehat{d}} &= \eta^* + g^{\widehat{d}}. \end{aligned}$$

a. Let g^d be the potential of d. Prove $\eta^d = \eta^*$ and $g^d = g^{\widehat{d}} + u$ with $(I - P^d)u = 0$.

b. Let g^d and $g^{\widehat{d}}$ be the biases of d and $\widehat{d}$, respectively. Prove $g^d - g^{\widehat{d}} = -(P^d)^* g^{\widehat{d}}$.

c. From b), the bias can be improved by optimizing $P^{d^*}(-g^{\widehat{d}})$ (cf. (4.13) for the ergodic case). Can we develop the policy iteration algorithm for bias optimality by using this property? What, if any, are the problems with this approach?

Solution:

a. Pre-multiplying the both sides of $f^d + P^d g^{\widehat{d}} = \eta^* + g^{\widehat{d}}$ by $(P^d)^*$, we have

$$(P^d)^* f^d + (P^d)^* P^d g^{\widehat{d}} = (P^d)^* \eta^* + (P^d)^* g^{\widehat{d}}.$$

By $(P^d)^*P^d = (P^d)^*$ and $P^d\eta^* = \eta^*$, we obtain $\eta^d = (P^d)^*f^d = (P^d)^*\eta^* = \eta^*$.

We prove $g^d = g^{\widehat{d}} + u$ with $(I - P^d)u = 0$. From the Poisson equation and $f^d + P^d g^{\widehat{d}} = \eta^* + g^{\widehat{d}}$, we have $g^d - g^{\widehat{d}} = f^d + P^d g^d - \eta^* - g^{\widehat{d}} = P^d(g^d - g^{\widehat{d}})$. Denote $u = g^d - g^{\widehat{d}}$, then $(I - P^d)u = 0$.

b. From part a), we have $u = P^d u$. Then, we get that $u = (P^d)^* u$. That is, $g^d - g^{\widehat{d}} = (P^d)^*(g^d - g^{\widehat{d}}) = -(P^d)^* g^{\widehat{d}}$ since $(P^d)^* g^d = 0$.

c. Let $g^{\widehat{d}}$ also denote the bias of policy $\widehat{d}$ with $(P^{\widehat{d}})^* g^{\widehat{d}} = 0$. Pre-multiplying on the both sides of $(I - P^{\widehat{d}})w^{\widehat{d}} = -g^{\widehat{d}}$ (Poisson Equation) by $(P^d)^*$, we get $-(P^d)^* g^{\widehat{d}} = (P^d)^*(I - P^{\widehat{d}})\widehat{w} = (P^d)^*(P^d - P^{\widehat{d}})\widehat{w}$. Combining with b), we get $g^d - g^{\widehat{d}} = (P^d)^*(P^d - P^{\widehat{d}})\widehat{w}$. We can develop the policy iteration algorithm for the bias optimality but this algorithm may not converge to the bias-optimal policy. This is because in this algorithm d is chosen only from $\widehat{\mathcal{A}}_1$. That is, we only search the bias optimal policy in $\{d | \eta^* + g^{\widehat{d}} = f^d + P^d g^{\widehat{d}}\}$, which will lose the policy improvement by choosing action α satisfying $f(i,\alpha) + \sum_{j\in\mathcal{S}} p^\alpha(j|i) g^{\widehat{d}}(j) > \eta^* + g^{\widehat{d}}(j)$. That is, in (4.69) we only consider the policies satisfying $\sum_j p^\alpha(j|i) w^{\widehat{d}}(j) > \sum_j p^{\widehat{d}}(j|i) w^{\widehat{d}}(j)$ when $Q^{\widehat{d}}(i,\alpha) = Q^{\widehat{d}}(i, \widehat{d}(i))$ but do not consider the policies satisfying $Q^{\widehat{d}}(i,\alpha) > Q^{\widehat{d}}(i, \widehat{d}(i))$. Under this iteration, the policy iteration may stop before it reaches the bias-optimal policy.

4.23 Prove $(I - P)(I - P + P^*)^{-n}\eta = 0$, and therefore from (4.80) $g_n = (-1)^{-1}(I - P + P^*)^{-1} f$ is a solution to (4.78) with $P^* g_n = (-1)^{n-1}\eta$.

Solution:

By $\eta = P^* f$ and $(I - P + P^*)^{-1} P^* = P^*$, we get $(I - P)(I - P + P^*)^{-n}\eta = (I - P)(I - P + P^*)^{-n} P^* f = (I - P)P^* f = 0$ noting $PP^* = P^*$. Denote the nth bias g_n^b. Then $g_n^b = (-1)^{n-1}(I - P + P^*)^{-n}(f - \eta) = (-1)^{n-1}(I - P + P^*)^{-n} f - (-1)^{n-1}(I - P + P^*)^{-n}\eta = (-1)^{n-1}(I - P + P^*)^{-n} f - (-1)^{n-1}\eta = g_n - (-1)^{n-1}\eta$. Since $(I - P)(-1)^{n-1}\eta = 0$ and g_n^b satisfies (4.78) with $P^* g_n^b = 0$, we know that g_n is a solution to (4.78) with $P^* g_n = P^*(-1)^{n-1}\eta = (-1)^{n-1}\eta$.

4.24 Derive (4.81) recursively.

Solution:

$$\begin{aligned} g_{n+1} &= (-1)^n (I - P + P^*)^{-(n+1)} (f - \eta) \\ &= (-1)^n [I + \sum_{n=1}^{\infty} (P^n - P^*)]^{n+1} (f - \eta) \\ &= (-1)^n \sum_{k=0}^{\infty} C_k (P^k - P^*)(f - \eta) \\ &= (-1)^n \sum_{k=0}^{\infty} C_k (P^k f - \eta). \end{aligned}$$

Where C_k is the coefficient of $P^k - P^*$ in the expansion of $[I + \sum_{n=1}^{\infty}(P^n - P^*)]^{n+1}$. The computation of C_k is equivalent to the number of solutions of $x_1 + x_2 + \cdots + x_{n+1} = k, x_i = 0, 1, 2, \cdots, i = 1, 2, \ldots, n+1$. Thus, from the results in combination mathematics, we know $C_k = \begin{pmatrix} (n+1) + k - 1 \\ (n+1) - 1 \end{pmatrix} = \begin{pmatrix} n + k \\ n \end{pmatrix}$.

Next, we prove it by induction.

For $n = 0$, we know that

$$g_1 = \sum_{k=0}^{\infty} (P^k f - \eta) = \sum_{k=0}^{\infty} \begin{pmatrix} k \\ 0 \end{pmatrix} (P^k f - \eta).$$

For $n = 1$, we obtain

$$\begin{aligned} g_2 &= -\sum_{l=0}^{\infty} P^l g_1 = -\sum_{l=0}^{\infty} P^l \sum_{k=0}^{\infty} (P^k f - \eta) = -\sum_{l=0}^{\infty} \sum_{k=0}^{\infty} (P^{k+l} f - \eta) \\ &= -\sum_{l=0}^{\infty} (l+1)(P^l f - \eta) = (-1)^1 \sum_{k=0}^{\infty} \begin{pmatrix} 1 + k \\ 1 \end{pmatrix} (P^k f - \eta). \end{aligned}$$

We can see that (4.87) holds for $n = 0, 1$. Now we assume that (4.87) holds for $n = m$, that is

$$g_{m+1} = (-1)^m \sum_{k=0}^{\infty} \begin{pmatrix} m + k \\ m \end{pmatrix} (P^k f - \eta).$$

For $n = m + 1$,

$$g_{m+2} = -\sum_{l=0}^{\infty} P^l g_{m+1}$$

$$
\begin{aligned}
&= -\sum_{l=0}^{\infty} P^l (-1)^m \sum_{k=0}^{\infty} \begin{pmatrix} m+k \\ m \end{pmatrix} (P^k f - \eta) \\
&= (-1)^{m+1} \sum_{l=0}^{\infty} \sum_{k=0}^{\infty} \begin{pmatrix} m+k \\ m \end{pmatrix} (P^{k+l} f - \eta) \\
&= (-1)^{m+1} \sum_{l=0}^{\infty} \begin{pmatrix} m+1+l \\ m+1 \end{pmatrix} (P^l f - \eta).
\end{aligned}
$$

4.25 Suppose that a sequence of vectors $g_0^{\hat{d}}$, $g_1^{\hat{d}}$, ..., $g_n^{\hat{d}}$, and $g_{n+1}^{\hat{d}}$ satisfies the optimality equations (4.90)-(4.92). Find a policy that has $g_0^{\hat{d}}$, $g_1^{\hat{d}}$, ..., $g_n^{\hat{d}}$, and $g_{n+1}^{\hat{d}}$ as its kth biases $k = 0, 1, \cdots, n+1$, respectively. Then by the sufficient optimality equations (4.90)-(4.92), $g_k^{\hat{d}}$ is the optimal kth biases, $k = 0, 1, \cdots, n$, respectively. Therefore, in the sufficient optimality condition (4.90)-(4.92), we may replace the sentence "A policy $\widehat{d}$ is nth optimal if ..." by " If a sequence of vectors $g_0^{\hat{d}}$, $g_1^{\hat{d}}$, ..., $g_n^{\hat{d}}$, and $g_{n+1}^{\hat{d}}$ satisfies (4.90)-(4.92), then $g_k^{\hat{d}}$ are the optimal kth bias, $k = 0, 1, \ldots, n$."

Solution:

We prove that if policy $d \in \mathcal{D}_{n+2}(g_0^{\hat{d}}, g_1^{\hat{d}}, \ldots, g_n^{\hat{d}}, g_{n+2}^{\hat{d}})$, then policy d has $g_k^{\hat{d}}$ as its kth biases, $k = 0, 1, \cdots, n+1$, respectively.

Recall that

$$
\begin{aligned}
&\mathcal{D}_{n+2}(g_0^{\hat{d}}, g_0^{\hat{d}}, \cdots, g_{n+2}^{\hat{d}}) \\
= \; &\{\text{all } d : P^d g_0^{\hat{d}} = g_0^{\hat{d}}, f^d + P^d g_1^{\hat{d}} = g_0^{\hat{d}} + g_1^{\hat{d}}, P^d g_{l+1}^{\hat{d}} = g_l^{\hat{d}} + g_{l+1}^{\hat{d}},\ l = 1, \cdots, n+1\}.
\end{aligned}
$$

Pre-multiplying the both sides of $f^d + P^d g_1^{\hat{d}} = g_0^{\hat{d}} + g_1^{\hat{d}}$ by $(P^d)^*$, we get $g_0^d = (P^d)^* f^d = (P^d)^* g_0^{\hat{d}} = g_0^{\hat{d}}$ since $(P^d)^* P^d = (P^d)^*$ and $P^d g_0^{\hat{d}} = g_0^{\hat{d}}$.

In the similar way, pre-multiplying the both sides of $P^d g_2^{\hat{d}} = g_1^{\hat{d}} + g_2^{\hat{d}}$ by $(P^d)^*$, we get $(P^d)^* g_1^{\hat{d}} = 0$ since $(P^d)^* P^d = (P^d)^*$. Combining with $f^d + P^d g_1^{\hat{d}} = g_0^{\hat{d}} + g_1^{\hat{d}}$, we obtain $g_1^d = [I - P^d + (P^d)^*]^{-1}(f^d - g_0^{\hat{d}}) = g_1^{\hat{d}}$.

Suppose $g_l^d = g_l^{\hat{d}}$, $1 \le l \le n$. Pre-multiplying the both sides of $P^d g_{n+2}^{\hat{d}} = g_{n+1}^{\hat{d}} + g_{n+2}^{\hat{d}}$ by $(P^d)^*$, we get $(P^d)^* g_{n+1}^{\hat{d}} = 0$ since $(P^d)^* P^d = (P^d)^*$. Combining with $P^d g_{n+1}^{\hat{d}} = g_n^{\hat{d}} + g_{n+1}^{\hat{d}}$, we obtain $g_{n+1}^d = -[I - P^d + (P^d)^*]^{-1} g_n^{\hat{d}} = g_{n+1}^{\hat{d}}$.

From the aforementioned, in the sufficient optimality condition (4.90)-(4.92), we may replace the sentence "A policy $\hat{d}$ is nth optimal if ..." by " If a sequence of vectors $g_0^{\hat{d}}$, $g_1^{\hat{d}}$,

..., $g_n^{\widehat{d}}$, and $g_{n+1}^{\widehat{d}}$ satisfies (4.90)-(4.92), then $g_k^{\widehat{d}}$ are the optimal kth bias, $k = 0, 1, \ldots, n$."

4.26 Develop a policy iteration algorithms that myopically maximizes the expected mth potentials, $m = 1, \cdots, n$, of the actions at each iteration, as illustrated on the right-hand side of Figure 4.8. Prove its convergence.

Solution:

This is stated as the *the second policy iteration algorithm for an nth bias optimal policy*:

1. Starting with any policy $d_0 \in \mathcal{D}$, and set $k = 0$.

2. Obtain the bias $g_l^{d_k}$, $l = 0, 1, \cdots, n$ and $(n+1)$th potential $g_{n+1}^{d_k}$ by solving

$$\begin{aligned} P^{d_k} g_0^{d_k} &= g_0^{d_k} \\ (I - P^{d_k}) g_1^{d_k} &= f^{d_k} - g_0^{d_k} \\ (I - P^{d_k}) g_l^{d_k} &= -g_{l-1}^{d_k}, \quad l = 2, 3, \cdots, n+1, \end{aligned}$$

   subject to $(P^{d_k})^* g_m^{d_k} = 0, m = 1, 2, \ldots, n$.

3. Set (component-wisely)

$$\tilde{\mathcal{D}}_0 := \left\{ d = arg\{\max_{d \in \mathcal{D}} [P^d g_0^{d_k}]\} \right\},$$

$$\tilde{\mathcal{D}}_1 := \left\{ d = arg\{\max_{d \in \tilde{\mathcal{D}}_0} [f^d + P^d g_1^{d_k}]\} \right\},$$

$$\tilde{\mathcal{D}}_l := \left\{ d = arg\{ \max_{d \in \tilde{\mathcal{D}}_{l-1}} [P^d g_l^{d_k}]\} \right\}, \quad l = 2, 3, \cdots, n,$$

   and choose

$$d_{k+1} = arg\{\max_{d \in \tilde{\mathcal{D}}_n} [P^d g_{n+1}^{d_k}]\},$$

   If at a state i, action $d_k(i)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$.

4. If $d_{k+1} = d_k$, stop; otherwise set $k := k + 1$ and go to step 2.

Firstly, since $d_{k+1} \in \tilde{\mathcal{D}}_0$ and $d_{k+1} \in \tilde{\mathcal{D}}_1$, according to the policy iteration algorithm for the gain-optimal policy, we know policy sequence $\{d_k\}$ must converge to a gain-optimal policy in a finite number of iterations. After that, $\tilde{\mathcal{D}}_0$ is the set of the optimal polices

$\mathcal{D}_0$. Since $d_{k+1} \in \tilde{\mathcal{D}}_2 \subseteq \tilde{\mathcal{D}}_1 \subseteq \mathcal{D}_0$, according to the policy iteration algorithm for the bias-optimal policy, we know $\{d_k\}$ must converge to a bias-optimal policy. Similarly, after that, $\tilde{\mathcal{D}}_1 = \mathcal{D}_1$. Going on this process, since $d_{k+1} \in \tilde{\mathcal{D}}_n \subseteq \mathcal{D}_{n-1}$ and $d_{k+1} = arg\{\max_{d\in\tilde{\mathcal{D}}_n}[P^d g_{n+1}^{d_k}]\}$, we can find the second policy iteration algorithm can converge to an nth-bias optimal policy.

4.27 A weak version of Lemma 4.7 can be easily established by the well-known Cayley-Hamilton theorem [154]: *For any $n \times n$ matrix A, define its characteristic polynomial as $r(s) = det(sI - A)$. We have $r(A) = 0$.* Use the Cayley-Hamilton theorem to prove that if policy (P^d, f^d) is an $(S+1)$th bias optimal policy, then it is also an n-bias optimal policy for all $n \geq 0$. (Hint: set $A = (I - P^d + (P^d)^*)^{-1}$ in the Cayley-Hamilton theorem.) **[solution]**

Denote $A = (I - P^d + (P^d)^*)^{-1}$ and $r(s) = \sum_{k=0}^{S} b_k s^k$, $b_S = 1$. We know that $r(A) = \sum_{k=0}^{S} b_k A^k = 0$, where $A^0 = I$. That is, $A^S = -\sum_{k=0}^{S-1} b_k A^k$. Then

$$A^{S+2} = -\sum_{k=0}^{S-1} b_k A^{k+2} = -\sum_{k=2}^{S+1} b_{k-2} A^k. \tag{4.20}$$

Similar to (4.109), we have

$$(P^h - P^d)[I - P^d + P^{d*}]^{-k}(f^d - g_0^*) = 0, \qquad \forall\ 1 < k \leq S+1. \tag{4.21}$$

Combining with (4.20), we get

$$(P^h - P^d)[I - P^d + P^{d*}]^{-(S+2)}(f^d - g_0^*) = 0.$$

Further, we obtain

$$(P^h - P^d)[I - P^d + P^{d*}]^{-k}(f^d - g_0^*) = 0, \qquad \forall\ k \geq S+2.$$

That is, $(P^h - P^d)g_n^d = 0$ for all $n \geq S+2$.

Finally, from the nth-bias difference equation (4.90) and by induction on n, we can prove

$$g_n^h - g_n^d = [I - P^h + P^{h*}]^{-1}(P^h - P^d)g_n^d + P^{h*}(P^h - P^d)g_{n+1}^d = 0,$$

for all $n \geq S+2$. That is, the nth-biases of the policies in $\mathcal{D}_{S+1}$ are all the same for all $n \geq S+2$. Since $(S+1)$th bias-optimal policy must be nth bias optimal policy,

$0 \leq n \leq S+1$, then an $(S+1)$th-bias optimal policy is also an n-bias optimal policy for all $n \geq 0$.

4.28 Let $d, h \in \mathcal{D}$ be two policies.

a. Prove that the following expansion holds for any $N \geq 1$:

$$\eta^h - \eta^d = f^h - f^d + \sum_{k=1}^{N} (P^h - I)^{k-1}(P^h - P^d)g_k^d + (P^h - I)^N (g_N^h - g_N^d).$$

b. Give the conditions under which $(P^h - I)^N (g_N^h - g_N^d)$ converges to zero as $N \to \infty$.

c. What do a) and b) indicate?

Solution:

a. For any policy d, we have the following Poisson equations.

$$\begin{aligned}
g_1^d &= f^d - \eta^d + P^d g_1^d \\
g_2^d &= -g_1^d + P^d g_2^d \\
g_3^d &= -g_2^d + P^d g_3^d \\
g_4^d &= -g_3^d + P^d g_4^d \\
&\vdots \\
g_n^d &= -g_{n-1}^d + P^d g_n^d
\end{aligned}$$

Then we can get

$$\begin{aligned}
& \eta^h - \eta^d \\
= \; & f^h + P^h g_1^h - g_1^h - (f^d + P^d g_1^d - g_1^d) \\
= \; & f^h + P^h g_1^d - f^d - P^d g_1^d + P^h(g_1^h - g_1^d) - (g_1^h - g_1^d) \\
= \; & f^h + P^h g_1^d - f^d - P^d g_1^d + (P^h - I)(g_1^h - g_1^d) \\
= \; & f^h - f^d + (P^h - P^d) g_1^d + (P^h - I)(g_1^h - g_1^d) \\
= \; & f^h - f^d + (P^h - P^d) g_1^d + (P^h - I)[P^h g_2^h - g_2^h - (P^d g_2^d - g_2^d)] \\
= \; & f^h - f^d + (P^h - P^d) g_1^d + (P^h - I)[(P^h - P^d) g_2^d - (g_2^h - g_2^d) + P^h(g_2^h - g_2^d)] \\
= \; & f^h - f^d + (P^h - P^d) g_1 + (P^h - I)[(P^h - P^d) g_2^d + (P^h - I)(g_2^h - g_2^d)] \\
= \; & f^h - f^d + (P^h - P^d) g_1^d + (P^h - I)(P^h - P^d) g_2^d + (P^h - I)^2 (g_2^h - g_2^d)
\end{aligned}$$

$$
\begin{aligned}
&= f^h - f^d + (P^h - P^d)g_1^d + (P^h - I)(P^h - P^d)g_2^d \\
&\quad +(P^h - I)^2[P^h g_3^h - g_3^h - (P^d g_3^d - g_3^d)] \\
&= f^h - f^d + (P^h - P^d)g_1^d + (P^h - I)(P^h - P^d)g_2^d \\
&\quad +(P^h - I)^2[(P^h - P^d)g_3^d - (g_3^h - g_3^d) + P^h(g_3^h - g_3^d)] \\
&= f^h - f^d + (P^h - P^d)g_1^h + (P^h - I)(P^h - P^d)g_2^d \\
&\quad +(P^h - I)^2[(P^h - P^d)g_3^d + (P^h - I)(g_3^h - g_3^d)] \\
&= f^h - f^d + (P^h - P^d)g_1^d + (P^h - I)(P^h - P^d)g_2^d \\
&\quad +(P^h - I)^2(P^h - P^d)g_3^d + (P^h - I)^3(g_3^h - g_3^d) \\
&= f^h - f^d + (P^h - P^d)g_1^d + (P^h - I)(P^h - P^d)g_2^d \\
&\quad +(P^h - I)^2(P^h - P^d)g_3^d + (P^h - I)^3[P^h g_4^h - g_4^h - (P^d g_4^d - g_4^d)] \\
&= f^h - f^d + (P^h - P^d)g_1^d + (P^h - I)(P^h - P^d)g_2^d \\
&\quad +(P^h - I)^2(P^h - P^d)g_3^d + (P^h - I)^3[(P^h - P^d)g_4^d - (g_4^h - g_4^d) + P^h(g_4^h - g_4^d)] \\
&= f^h - f^d + (P^h - P^d)g_1^d + (P^h - I)(P^h - P^d)g_2^d \\
&\quad +(P^h - I)^2(P^h - P^d)g_3^d + (P^h - I)^3[(P^h - P^d)g_4^d + (P^h - I)(g_4^h - g_4^d)] \\
&= f^h - f^d + (P^h - P^d)g_1^d + (P^h - I)(P^h - P^d)g_2^d \\
&\quad +(P^h - I)^2(P^h - P^d)g_3^d + (P^h - I)^3(P^h - P^d)g_4^d + (P^h - I)^4(g_4^h - g_4^d) \\
&\quad \vdots \\
&= f^h - f^d + \sum_{k=1}^{N}(P^h - I)^{k-1}(P^h - P^d)g_k^d + (P^h - I)^N(g_N^h - g_N^d),
\end{aligned}
$$

for any $N \geq 1$.

b. If all the eigenvalues of $P^h - I$ are within the unit circle, then $(P^h - I)^N$ will converge to 0 matrix. Assume that the eigenvalues of P^h are 1 and λ. Then eigenvalues of $P^h - I$ are 0 and $\lambda - 1$. If we would like $(P^h - I)^N$ converge, then we need $|\lambda - 1| < 1$.

c. By b), we see that the convergence of $(P^h - I)^N(g_N^h - g_N^d)$ does not depend on P^d. Based on a) and b), we know the difference of the gains under two different policies can be expressed by all nth biases under one policy, $n = 1, 2, \ldots$.

4.29 The results presented in this chapter are strongly related to the sensitive discount optimality (n-discount optimality and Blackwell optimality), see [194, 216, 248, 249].

For any Markov chain with transition prbability matrix P and reward function f, the discounted reward is defined as (cf. (4.72)):

$$\upsilon_\beta(i) := E\left\{\sum_{l=0}^{\infty}\beta^l f(X_l)\Big|X_0 = i\right\}, \quad 0<\beta<1.$$

Denote $\upsilon_\beta = (\upsilon_\beta(1),\ldots,\upsilon_\beta(S))^T$. Set $\beta = (1+\rho)^{-1}$, or $\rho = (1-\beta)/\beta$. $0<\beta<1$ implies $\rho > 0$. Let ρ_0 be the non-zero eigenvalue of $I-P$ with the smallest absolute value. We have the Laurent series expansion:

$$\upsilon_\beta = (1+\rho)\sum_{n=-1}^{\infty}\rho^n y_n, \quad 0<\rho<\rho_0,$$

where $y_{-1} = P^* f$ and $y_n = (-1)^n H_P^{n+1} f$, $n = 0,1,\ldots$, $H_P = (I-P+P^*)^{-1}(I-P^*)$.

a. Explain the meaning of ρ. (Hint: inflation rate)

b. Prove the Laurent series expansion (cf. Theorem 8.2.3 of [216]).

c. Prove $y_n = g_{n+1}$ be the $(n+1)$th bias of (P,f), $n=-1,0,1,\ldots$. Thus, we have

$$\upsilon_\beta = (1+\rho)\sum_{n=0}^{\infty}\rho^{n-1} g_n, \quad 0<\rho<\rho_0.$$

Solution:

a. ρ can be viewed as the inflation rate (or the interest rate). One dollar today will become $1+\rho$ dollar tomorrow. Contrarily, one dollar tomorrow is equal to $(1+\rho)^{-1}$ dollar today. If the rewards in the future are all converted into the current rewards, the reward $f(X_n)$ at time n is equal to $(1+\rho)^{-n} f(X_n)$ at current time. Thus the total reward is

$$\upsilon_\beta(i) = E\Big\{\sum_{l=0}^{\infty}\beta^l f(X_l)|X_0 = i\Big\},$$

if the initial state is i, where $\beta = (1+\rho)^{-1}$.

b. From (2.31), we have

$$\upsilon_\beta \;=\; (I-\beta P)^{-1} f.$$

Putting $\beta = (1+\rho)^{-1}$ into the above equation, we have

$$\begin{aligned}
\upsilon_\beta &= (I-\beta P)^{-1} f \\
&= (I-\beta P)^{-1}(f-\eta) + (I-\beta P)^{-1}\eta \\
&= (I-\beta P)^{-1}(f-\eta) + \sum_{n=0}^{\infty} \beta^n P^n \eta \\
&= (I-\beta P)^{-1}(I-P^*) f + (1+\rho)\frac{\eta}{\rho} \\
&= (I-\beta P+\beta P^*)^{-1}(I-P^*) f + (1+\rho)\frac{\eta}{\rho} \\
&= (1+\rho)(\rho I + I - P + P^*)^{-1}(I-P^*) f + (1+\rho)\frac{\eta}{\rho}.
\end{aligned}$$

Define $H_P(\rho) = (\rho I + I - P + P^*)^{-1}(I-P^*)$, then, $H_P = H_P(0) = (I-P+P^*)^{-1}(I-P^*)$ and

$$\upsilon_\beta = (1+\rho)\left[H_P(\rho) f + \frac{\eta}{\rho}\right].$$

Since

$$(\rho I + I - P + P^*)H_P = \rho H_P + (I-P)H_P = \rho(I-P^*)H_P + (I-P^*) = (I-P^*)(I+\rho H_P),$$

where we have used $P^* H_P = 0$ and $(I-P)H_P = I - P^*$. Left-multiplying $(\rho I + I - P + P^*)^{-1}$ and right-multiplying $(I+\rho H_P)^{-1}$ on both sides of the above equation, we have

$$H_P(\rho) = H_P(I+\rho H_P)^{-1}.$$

Thus,

$$\upsilon_\beta = (1+\rho)\left[H_P(I+\rho H_P)^{-1} f + \frac{\eta}{\rho}\right].$$

If the spectral radius of ρH_P, i.e., $\sigma(\rho H_P) < 1$, we have

$$\begin{aligned}
\upsilon_\beta &= (1+\rho)\left[\frac{\eta}{\rho} + H_P(I+\rho H_P)^{-1} f\right] \\
&= (1+\rho)\left[\frac{\eta}{\rho} + \sum_{n=0}^{\infty}(-\rho)^n H_P^{n+1} f\right].
\end{aligned}$$

The Laurent series is obtained. Next, we need to consider the eigenvalues of H_P to find the condition such that $\sigma(\rho H_P) < 1$. We can prove that if the eigenvalues of P are

$\{1, 1, \ldots, 1, \lambda_{m+1}, \ldots, \lambda_S\}$, then the eigenvalues of H_P is $\{0, 0, \ldots, 0, \frac{1}{1-\lambda_{m+1}}, \ldots, \frac{1}{1-\lambda_S}\}$. Thus, to guarantee $\sigma(\rho H_P) < 1$, we need $0 \leq \rho < \min\{|1 - \lambda_i|, i = m+1, \ldots, S\}$.

Moreover, we can obtain Laurent series as follows:

From Theorem A.5 in [216], for any transition probability matrix P with m recurrent classes, there exists a nonsingular matrix W for which

$$P = W^{-1} \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} W,$$

where I is an $m \times m$ identity matrix and Q is an $(|S| - m) \times (|S| - m)$ matrix with the following properties:

1. 1 is not an eigenvalue of Q.
2. The spectral radius of Q, $\sigma(Q)$ is smaller than or equal to 1 and if all recurrent sub-chains of P are aperiodic, $\sigma(Q) < 1$.
3. $(I - Q)^{-1}$ exists.

and W satisfies

$$W^{-1} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W = P^*, \quad W^{-1} \begin{bmatrix} (I-Q)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W = H_P. \tag{4.22}$$

Next, define the resolvent of $P - I$ by

$$R^\rho = (\rho I + [I - P])^{-1}.$$

From (2.31), we have

$$\upsilon_\beta = (1 + \rho) R^\rho f. \tag{4.23}$$

Let $B = I - Q$. Then

$$\rho I + I - P = W^{-1} \begin{bmatrix} \rho I + B & 0 \\ 0 & \rho I \end{bmatrix} W,$$

so that

$$\begin{aligned} R^\rho &= W^{-1} \begin{bmatrix} (\rho I + B)^{-1} & 0 \\ 0 & \rho^{-1} I \end{bmatrix} W \\ &= \rho^{-1} W^{-1} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W + W^{-1} \begin{bmatrix} (\rho I + B)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W. \end{aligned} \tag{4.24}$$

Since

$$(\rho I + B)^{-1} = (I + \rho B^{-1})^{-1} B^{-1},$$

and whenever $\sigma(\rho B^{-1}) < 1$ or ρ is smaller than the non-zero eigenvalue of $I - Q$ or $I - P$ with the smallest absolute value,

$$(\rho I + B)^{-1} = \sum_{n=0}^{\infty} (-\rho)^n (B^{-1})^{n+1}, \tag{4.25}$$

Putting (4.22) and (4.25) into (4.24), we have

$$R^{\rho} = \rho^{-1} P^* + \sum_{n=0}^{\infty} (-\rho)^n H_P^{n+1}. \tag{4.26}$$

Putting (4.26) into (4.23), we can obtain the Laurent series expansion.

Reference:

1. B. L. Miller and A. F. Veinott, Discrete Dynamic Programming with a Small Interest Rate, *The Annals of Mathematical Statistics*, vol. 40, no. 2, 366-370, 1969.

2. M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, New York, 1994.

c.

$$\begin{aligned} y_n &= (-1)^n H_P^{n+1} f = (-1)^n (I - P + P^*)^{-(n+1)} (I - P^*)^{(n+1)} f \\ &= (-1)^n (I - P + P^*)^{-(n+1)} (I - P^*) f \\ &= (-1)^n (I - P + P^*)^{-(n+1)} (f - \eta) \\ &= g_{n+1}. \end{aligned}$$

Thus, we have

$$v_\beta = (1 + \rho) \sum_{n=0}^{\infty} \rho^{n-1} g_n, \quad 0 < \rho < \rho_0.$$

4.30 A policy $d_b \in \mathcal{D}$ is called a (stationary and deterministic) Blackwell policy if there exists a β^*, $0 \le \beta^* < 1$, such that

$$v_\beta^{d_b} \ge v_\beta^d, \quad \text{for all } d \in \mathcal{D} \text{ and all } \beta \in [\beta^*, 1).$$

a. Prove that if $d \in D_S$, then d is a Blackwell optimal policy.

b. Prove $d_b \in D_n$ for all $n \geq 0$.

Solution:

a. Since $\rho \to 0$ when $\beta \to 1$, the definition of Blackwell optimality is equivalent to the following: A policy $d_b \in \mathcal{D}$ is a Blackwell optimal policy if there exists a ρ^* for which $v_\beta^{d_b} \geq v_\beta^d$ for $0 < \rho \leq \rho^*$ and any policy d.

From Lemma 4.7, if $d^* \in D_S$, we know d^* is an nth bias for all $n \geq 0$. This is, d^* maximizes all $g_n, n \geq 0$. Let $d \in \mathcal{D}$ and suppose that for some $n = -1, 0, 1, \ldots$, $d \bar{\in} \mathcal{D}_n$. Let n' be the minimal n for which this holds. $n = -1$ means $d \in \mathcal{D}$.

$$(1+\rho)^{-1}\rho^{-(n'-1)}[v_\beta^{d^*} - v_\beta^d] = g_{n'}^{d^*} - g_n^d + \sum_{k=n'+1}^{\infty} \rho^{k-n'}[g_k^{d^*} - g_k^d], \tag{4.27}$$

Since d is not the n'th bias optimal, for some state i, $x = g_{n'}^{d^*}(i) - g_{n'}^d(i) > 0$. From (4.27), it follows that

$$(1+\rho)^{-1}(v_\beta^{d^*}(i) - v_\beta^d(i)) = \rho^{n'-1}x + \sum_{k=n'+1}^{\infty} \rho^{k-1}[g_k^{d^*}(i) - g_k^d(i)].$$

We can find a ρ_d for which the above expression is positive for $0 < \rho < \rho_d$.

Repeating the above argument, we obtain a ρ_d for each $d \in \mathcal{D}$. Set $\rho^* = \min_{d \in \mathcal{D}} \rho_d$. Since $\mathcal{D}$ is finite, $\rho^* > 0$. Therefore,

$$v_\beta^{d^*} \geq v_\beta^d$$

for all $d \in \mathcal{D}$ and $0 \leq \rho < \rho^*$. Thus, d^* is a Blackwell optimal policy

b. If d_b is a Blackwell optimal, we have

$$(1+\rho)^{-1}\rho^{-(n-1)}[v_\beta^{d_b} - v_\beta^d] \geq 0,$$

for all $0 \leq \rho < \rho^*$, and all $d \in \mathcal{D}$ for $n = 0, 1 \ldots$, From (4.27) and let $\rho \to 0$, we have $g_n^{d_b} \geq g_n^d$ for all $n = 0, 1 \ldots$. That is, $d_b \in \mathcal{D}_n$, for all $n \geq 0$.

5

Solutions to Chapter 5

5.1 Repeat Example 5.1 by using the continuous-time Markov model.

Solution: We consider the infinitesimal generator of continuous time Markov model. The infinitesimal generator can be expressed by the transition probability matrix as follows:

$$A = \Lambda(P - I),$$

where Λ is a diagonal matrix, whose (i, i)th component is the equivalent service rate at state i. From the transition probability of embedded Markov chain given in the textbook, we can obtain

$$\begin{aligned}
a[(n,1),(n+1,1)] &= (\lambda_1 + \lambda_4) * \frac{\lambda_4}{\lambda_1 + \lambda_4} = \lambda_4, \\
a[(n,1),(n,2)] &= (\lambda_1 + \lambda_4) * \frac{\lambda_1}{\lambda_1 + \lambda_4} = \lambda_1, \\
a[(n,1),(n,1)] &= -(\lambda_1 + \lambda_4), \\
a[(n,2),(n+1,2)] &= (\lambda_2 + \lambda_4) * \frac{\lambda_4}{\lambda_2 + \lambda_4} = \lambda_4,
\end{aligned}$$

$$
\begin{aligned}
&a[(n,2),(n,3)] = (\lambda_2+\lambda_4) * \frac{\lambda_2}{\lambda_2+\lambda_4} = \lambda_2,\\
&a[(n,2),(n,2)] = -(\lambda_2+\lambda_4),\\
&a[(n,3),(n+1,3)] = (\lambda_3+\lambda_4) * \frac{\lambda_4}{\lambda_3+\lambda_4} = \lambda_4,\\
&a[(n,3),(n-1,1)] = (\lambda_3+\lambda_4) * \frac{\lambda_3}{\lambda_3+\lambda_4} b^a(n) = \lambda_3 b^a(n),\\
&a[(n,3),(n,1)] = (\lambda_3+\lambda_4) * \frac{\lambda_3}{\lambda_3+\lambda_4}[1-b^a(n)] = \lambda_3[1-b^a(n)],\\
&a[(n,3),(n,3)] = -(\lambda_3+\lambda_4),
\end{aligned}
$$

for $0 < n < N$; and

$$
\begin{aligned}
&a[0,(1,1)] = \lambda_4, a[0,0] = -\lambda_4,\\
&a[(N,1),(N,2)] = \lambda_1, a[(N,1),(N,1)] = -\lambda_1,\\
&a[(N,2),(N,3)] = \lambda_2, a[(N,2),(N,2)] = -\lambda_2,\\
&a[(N,3),(N,1)] = \lambda_3[1-b^a(N)], a[(N,3),(N-1,1)] = \lambda_3 b^a(N), a[(N,3),(N,3)] = -\lambda_3.
\end{aligned}
$$

The transitions from states $(n,1)$ and $(n,2)$ also do not depend on the actions. The comparison of actions in the policy improvement step for state $(n,3)$, $0 < n < N$ is the same as (5.1).

5.2 A machine produces M different products, denoted as $1, 2, \cdots, M$. To process product i, the machine has to take N_i different operations, denoted as $(i,1),(i,2),\cdots,(i,N_i)$. We use discrete time model. At each time l, $l = 0, 1, \cdots$, the machine can only process one product and undertake one operation. If at time instant l the machine is producing product i and is at operation (i,j), $j \neq N_i$, then at time instant $l+1$ the machine will take operation (i,j') with probability $p_i(j'|j)$, $i = 1,2,\cdots,M$, $j = 1,\cdots,N_i-1$, and $j' = 1,\cdots,N_i$. If the machine is at operation (i,N_i), then it will pick up a new product i' and start to process it at operation $(i',1)$ at the next time instant with probability $p^a(i'|i)$, $i,i' = 1,2,\cdots,M$, where $a \in \mathcal{A}(i)$ represents an action. The operation $(i,1)$ is called an *entrance operation* and (i,N_i) is called an *exit operation*. The system can be modelled as a Markov chain with state space $\mathcal{S} := \{(i,j), i = 1,2,\cdots,M; j = 1,\cdots,N_i\}$. Let f be the properly defined performance function. Derive the policy iteration condition (similar to (5.1)in Example 5.1) for this problem and show that with the sample-path-based

approach we do not need to estimate the potentials for all the states.

Solution: The transition probabilities are

$$p[(i,j')|(i,j)] = p_i(j'|j),$$

when $i = 1, 2, \cdots, M$, $j = 1, \cdots, N_i - 1, j' = 1, 2, \cdots, N_i$, and

$$p[(i',1)|(i,N_i)] = p^a(i'|i), \quad i, i' = 1, 2, \cdots, M.$$

The other transition probabilities are zeros. For simplicity, we assume the performance function f depends only on the states and does not depend on the actions. From the above transitions probabilities, we can find the transition from states $(i,j), i = 1, 2, \cdots, M, j = 1, 2, \cdots, N_i - 1$ do not depend on the actions. The comparison of actions in the policy improvement step for state $(i, N_i), i = 1, \cdots, M$, is

$$\begin{aligned} & p^{a'}(1|i)g(1,1) + p^{a'}(2|i)g(2,1) + \cdots + p^{a'}(M|i)g(M,1) \\ \geq \; & p^a(1|i)g(1,1) + p^a(2|i)g(2,1) + \cdots + p^a(M|i)g(M,1). \end{aligned} \tag{5.1}$$

From (5.1), with the sample-path-based approach, we do not need to estimate the potentials for all the states and only need to estimate the potentials of $g(i,1), i = 1, 2, \cdots, M$.

5.3 In Problem 4.1, prove that if we use the sample path based approach then we do not need to know the value of r.

Solution: The comparison of actions in the policy improvement step for state n is

$$\begin{aligned} & cn + \beta\mu'_n + rg(n+1) + (1 - \mu'_n - r)g(n) + \mu'_n g(n-1) \\ \geq \; & cn + \beta\mu'_n + rg(n+1) + (1 - \mu'_n - r)g(n) + \mu_n g(n-1). \end{aligned}$$

This is equivalent to

$$\beta\mu'_n + \mu'_n[g(n-1) - g(n)] \geq \beta\mu'_n + \mu_n[g(n-1) - g(n)].$$

Thus, we do not need to know the value of r.

5.4 As discussed in Section 5.1 , to save memory and computation at each iteration, we may partition the state space $\mathcal{S} = \{1, 2, \cdots, S\}$ into N subsets and at each iteration we

may only update the actions for the states in one of the subsets. In the extreme case, at each iteration we may update the action for only one state. That is, at the first iteration, we update $d(1)$; at the second iteration, we update $d(2)$, ... , and at the Sth iteration, we update $d(S)$. Then at the $(S+1)$th iteration, we update $d(1)$ again, and so on in a round robin manner. In such an iteration procedure we cannot stop if at some iteration there is no improvement in performance. We let the iteration algorithm stop after the performance does not improve in S consecutive iterations.

a. Formally state this policy iteration algorithm,

b. Prove that the algorithm stops after a finite number of iterations,

c. Prove that the algorithm stops at a gain-optimal policy, and

d. Extend this algorithm to the general case where $\mathcal{S}$ is partitioned into N subsets.

Solution:

a. **Policy Iteration Algorithm:**

1. Select an initial policy d_0 and set $i = 1$, $c = 0$, and $k = 0$.

2. (Policy evaluation) Obtain the potential g^{d_k} by solving the Poisson equation $(I - P^{d_k})g^{d_k} + \eta^{d_k} e = f^{d_k}$.

3. (Policy improvement) If $i \equiv S + 1$, then set $i = 1$; otherwise, choose

$$d_{k+1}(i) = arg\{\max_{d(i)\in\mathcal{A}(i)} [f(i, d(i)) + p^{d(i)}(i, j)g^{d_k}(j)]\}.$$

   If action $d_k(i)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$. Set $i = i + 1$.

4. If $d_{k+1} = d_k$, set $c = c + 1$; otherwise, $c = 0$. If $c = S$, stop; otherwise, set $k = k + 1$ and go to step 2.

b. The convergence of the above policy iteration algorithm: In S consecutive steps, if there is at least one step such that the policy is different, then we have $f_{k+S}^{d_{k+S}} + P^{d_{k+S}} g^{d_k} \succeq f_k^{d_k} + P^{d_k} g^{d_k}$. Thus, we have $\eta^{d_{k+S}} > \eta^{d_k}$. This is to say, the policy can be improved if

the algorithm does not stop. Since the policies are finite, thus the algorithm must stop after a finite number of iterations.

c. When the algorithm stops at step k, we set $\hat{d} := d_{k+1} = d_k = \cdots = d_{k-S+2}$. From the above algorithm, we have

$$\hat{d} \in arg\{\max_d [f^d + P^d g^{\hat{d}}]\}.$$

or

$$f^{\hat{d}} + P^{\hat{d}} g^{\hat{d}} \geq f^d + P^d g^{\hat{d}}, \qquad \text{for all } d \in \mathcal{D}.$$

By the optimality condition (4.5), $\hat{d}$ is the optimal policy.

d. If the state space $\mathcal{S}$ is partitioned into N subsets, we may only update the actions for the states in one of the subsets. For example, we assume the state space $\mathcal{S}$ is partitioned into N subsets defined as $\mathcal{S}_1 = \{1, \cdots, n_1\}, \mathcal{S}_2 = \{n_1 + 1, \cdots, n_2\}, \cdots, \mathcal{S}_m = \{n_{m-1} + 1, \cdots, n_m\}, \cdots, \mathcal{S}_N = \{n_{N-1} + 1, \cdots, S\}$. At the first iteration, we update $d(1), \cdots, d(n_1)$; at the second iteration, we update $d(n_1 + 1), \cdots, d(n_2)$, and at the Nth iteration, we update $d(n_{N-1} + 1), \cdots, d(S)$. Then, at the $N + 1$th iteration, we update $d(1), \cdots, d(n_1)$ again, and so on in a round robin manner. We let the iteration algorithm stop after the performance does not improve in N consecutive iterations.

5.5 To illustrate the idea behind Lemma 5.2 , we consider the following simple problem. There are N different balls with identical appearance but different weights, denoted as $m_1, m_2, \cdots, m_N$, respectively, $m_i \neq m_j$, $i \neq j$. These weights are known to us. You have a scale at your hand which is in-accurate with a maximal absolute error $r > 0$. Under what condition you may accurately identify these balls using this scale?

Solution: Let $m = \min\{|m_i - m_j| : i \neq j, i, j = 1, 2, \cdots, N\}$. If $r < m$, we can accurately identify these balls by using this scale.

5.6 Suppose that when the sample-path-based policy iteration algorithm 5.2 stops the estimation error of the potentials satisfies $|r| = |\bar{g} - g| < \delta/2$, where $\delta > 0$ is any positive number. Let $\bar{\eta}$ be the optimal average reward thus obtained. Prove

$$|\bar{\eta} - \eta^*| < \delta,$$

where η^* is the true optimal average reward.

Solution: We assume that policy d^* is the optimal policy, then $\eta^{d^*} = \eta^*$; and d is the policy obtained by the sample-path-based algorithm. From the definition of $\phi(g)$, we have $f^d + P^d\bar{g} \geq f^{d^*} + P^{d^*}\bar{g}$. From this equation, we have

$$f^d + P^d g + (P^d - P^{d^*})(\bar{g} - g) \geq f^{d^*} + P^{d^*} g.$$

Therefore,

$$(f^{d^*} + P^{d^*} g) - (f^d + P^d g) \leq (P^d - P^{d^*})(\bar{g} - g).$$

According to the difference formula, we have

$$\eta^* - \bar{\eta} = \pi^*[(f^{d^*} + P^{d^*} g) - (f^d + P^d g)] \leq \pi^*(P^d - P^{d^*})(\bar{g} - g) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Since $\eta^* \geq \bar{\eta}$, we have $|\bar{\eta} - \eta^*| \leq \delta$.

5.7 If we use

$$g_{L,N}(i) = \frac{\sum_{n=0}^{N-L+1}\{I_i(X_n)[\sum_{l=0}^{L-1} f(X_{n+l}) - \eta]\}}{\sum_{n=0}^{N-L+1} I_i(X_n)}.$$

to estimate the potentials,

a. Convince yourself that the results in Section 5.2.2 still hold, and

b. Revise the proofs in Section 5.2.2 for the sample-path-based policy iteration with the above potential estimates

Solution:

a. We consider the biased estimate

$$g_{L,N}(i) = \frac{\sum_{n=0}^{N-L+1}\{I_i(X_n)[\sum_{l=0}^{L-1} f(X_{n+l}) - \eta]\}}{\sum_{n=0}^{N-L+1} I_i(X_n)}.$$

From the proofs of the results in Section 5.2, these proofs do not need to know the estimate methods of the potentials. We only assume the estimate error should satisfy some conditions with probability 1. Using the fundamental ergodicity theorem in Chapter 3, we have

$$\lim_{N\to\infty, L\to\infty} g_{L,N}(i) = g(i) = E\{\sum_{l=0}^{\infty}[f(X_{n+l}) - \eta]|X_0 = i\}, \quad a.s.$$

Thus, as long as N and L are large enough, the estimation error can satisfy the conditions in Section 5.2.2. Thus, the results in Section 5.2.2 still hold.

b. In the proof of the sample-path-based policy iteration in Section 5.2.2, we also only require that the estimate error should satisfy some conditions with probability 1. Thus, the proof under the new potential estimate is the same as the original proof.

5.8 With the sample-path-based policy iteration algorithm 5.1. Suppose that the Markov chain is ergodic with a finite state space under all policies, and the number of policies is finite. If $|r| = |\bar{g}^d - g^d| < (\kappa/2)e$, where g^d and $\bar{g}^d$ are the potential of policy d and its estimate. Following the same argument as that in Lemma 5.3 , prove

$$\phi(\bar{g}^d) \subseteq \phi(g^d).$$

Solution: Let $h \in \phi(g^d)$ and $h' \in \phi(\bar{g}^d)$. By the definition of $\phi(g)$ in (5.11), we have $f^h + P^h g^d \geq f^{h'} + P^{h'} g^d$ and $f^{h'} + P^{h'} \bar{g}^d \geq f^h + P^h \bar{g}^d$. From the latter equation, we have

$$f^{h'} + P^{h'} g^d + (P^{h'} - P^h)(\bar{g}^d - g^d) \geq f^h + P^h g^d.$$

Therefore,

$$(f^h + P^h g^d) - (f^{h'} + P^{h'} g^d) \leq (P^{h'} - P^h)(\bar{g}^d - g^d).$$

This, together with $f^h + P^h g^d \geq f^{h'} + P^{h'} g^d$, leads to

$$|(f^h + P^h g^d) - (f^{h'} + P^{h'} g^d)| \leq |(P^{h'} - P^h)(\bar{g}^d - g^d)|. \tag{5.2}$$

From (5.2), if $|r| = |\bar{g}^d - g^d| < (\kappa/2)e$, then $|(f^h + P^h g^d) - (f^{h'} + P^{h'} g^d)| < \kappa e$. By the definition of κ, we must have $f^h + P^h g^d = f^{h'} + P^{h'} g^d$. In other words, $h' \in \phi(g^d)$. Thus, $\phi(\bar{g}^d) \subseteq \phi(g^d)$.

5.9 In Problem 5.8, we proved that $\phi(\bar{g}^d) \subseteq \phi(g^d)$.

a. On the surface, it looks like that the same method as that in Lemma 5.3 can be used to prove $\phi(g^d) \subseteq \phi(\bar{g}^d)$. Give a try.

b. If you cannot prove the result in a), explain why; if you feel that you did prove it, find out what's wrong in your proof.

c. Suppose that $h, h' \in \phi(g^d)$, and thus $f^h + P^h g^d = f^{h'} + P^{h'} g^d$. Because of the error in $\bar{g}^d$, we may have $f^h + P^h \bar{g}^d \neq f^{h'} + P^{h'} \bar{g}^d$. Therefore, one of them cannot be in $\phi(\bar{g}^d)$. Give an example to show that no matter how small the error $r = g^d - \bar{g}^d$ is, this fact is true.

Solution:

a) and b) On the surface, similarly to the method in Lemma 3, we can also obtain

$$|(f^h + P^h \bar{g}^d) - (f^{h'} + P^{h'} \bar{g}^d)| \leq |(P^{h'} - P^h)(g^d - \bar{g}^d)| + \nu e.$$

In the textbook, κ can be defined because g^d is determined. However, since $\bar{g}^d$ is a random variable, we cannot define a constant similar to κ. Thus, we cannot prove $\phi(g^d) \subseteq \phi(\bar{g}^d)$.

c. We consider Example 4.1 in Chapter 4. Under policy d_1, the potential $g_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. We can easily verify that $f_2 + P_2 g_1 = f_1 + P_1 g_1 = \max_d\{f_d + P_d g_1\}$, that is, d_1 and d_2 are all in $\phi(g_1)$. However, when we consider the estimate $\bar{g}_1$, Because of the error in $\bar{g}_1$, we may have $f_2 + P_2 \bar{g}_1 \neq f_1 + P_1 \bar{g}_1$. For example, for any $\delta > 0$, if the estimate $\bar{g}_1 = \begin{bmatrix} 1+\delta \\ -1+\delta/2 \end{bmatrix}$, we have $|r| = |g_1 - \bar{g}_1| < \delta$, but we can easily obtain

$$f_2 + P_2 \bar{g}_1 = \begin{bmatrix} 1 + 3\delta/4 \\ -1 + 7\delta/8 \end{bmatrix} \neq f_1 + P_1 \bar{g}_1 = \begin{bmatrix} 1 + 3\delta/4 \\ -1 + 3\delta/4 \end{bmatrix}.$$

Since δ is arbitrary, no matter how small the error $r = |g_1 - \bar{g}_1|$ is, we have $f_2 + P_2 \bar{g}_1 \neq f_1 + P_1 \bar{g}_1$. Therefore, one of them cannot be in $\phi(\bar{g}^d)$

5.10 Are the following statements true? If so, please explain the reasons:

a. Suppose we use $d_{k+1} \in \phi(\bar{g}^{d_k})$ to replace (5.14) in step 3 of Algorithm 5.1 (i.e., set $\nu = 0$ in (5.12) . Then the algorithm may not stop even if $\phi(\bar{g}^{d_k}) \subseteq \phi(g^{d_k})$ for $K' > K$ consecutive iterations $k = n, n+1, \cdots, n+K-1$, where K is the number of policies in $\mathcal{D}$.

b. Algorithm 5.2 may not stay in $\mathcal{D}_0$ even after $\phi(\bar{g}^{d_k}_{N_k}) = \phi(g^{d_k})$ for K consecutive iterations, where K is any large integer.

c. The above statement b) is true even if we add the following sentence to step 3 of Algorithm 5.2 : "If at a state i, action $d_k(i)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$."

Solution:

a. The performance increases every iteration and the policy iteration must reach the optimal policies if $\phi(\bar{g}^{d_k}) \subseteq \phi(g^{d_k})$ for $K' > K$ consecutive iterations. but after it reaches the optimal policy set, it may happen that $d_k \notin \phi(\bar{g}^{d_k})$. In this case, the policy may oscillates and never stops.

b. Although the performance increases every iteration and the policy iteration must reach the optimal policies when $\phi(\bar{g}_{N_k}^{d_k}) = \phi(g^{d_k})$ in K consecutive iterations, after that it may happen that $\phi(\bar{g}_{N_k}^{d_k}) \neq \phi(g^{d_k})$. At this case, the policy may go out of $\mathcal{D}_0$.

c. Even if we add the following sentence to step 3 of Algorithm 2 : "If at a state i, action $d_k(i)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$", we cannot guarantee that $\phi(\bar{g}_{N_k}^{d_k}) \neq \phi(g^{d_k})$ does not happen after $\phi(\bar{g}_{N_k}^{d_k}) = \phi(g^{d_k})$ for K consecutive iterations. Thus, the policy can still go out of $\mathcal{D}_0$.

5.11 Can you propose any stopping criteria for the sample-path-based algorithms to stop at an optimal policy in a finite number of iterations with probability 1?

Solution: The answer is depressed. Because we can only guarantee the algorithms stop in a finite number of iterations with a certain probability p_0 defined in (5.18), but not with probability 1.

5.12 In Lemma 5.4, $\sum_{k=0}^{\infty}(1 - y_k) < \infty$ implies $\lim_{k\to\infty} y_k = 1$, which, however, is not enough for $\lim_{n\to\infty} \prod_{k\geq n} y_k = 1$. For the latter to hold, y_k has to approach 1 fast enough.

a. For $y_k = 1 - \frac{1}{k}$, $k = 1, 2, \cdots$, we have $\lim_{k\to\infty} y_k = 1$. What is $\lim_{n\to\infty} \prod_{k\geq n} y_k$?

b. Verify the lemma for $y_k = 1 - \frac{1}{k^2}$, $k = 1, 2, \cdots$. What is $\lim_{n\to\infty} \prod_{k\geq n} y_k$?

c. For a sequence y_k, $0 \leq y_k \leq 1$, $k = 1, 2, \cdots$, if $\sum_{k=0}^{\infty}(1 - y_k) < \infty$ then we have $\sum_{k=0}^{\infty}(1 - y_k^c) < \infty$ for any $c < 1$ and we can apply this lemma. How about $c > 1$?

Solution:

a.

$$\prod_{k\geq n} y_k = \prod_{k\geq n}(1-\frac{1}{k}) = \frac{n-1}{n}\frac{n}{n+1}\frac{n+1}{n+2}\cdots = 0.$$

Thus, $\lim_{n\to\infty}\prod_{k\geq n} y_k = 0$.

b.

$$\begin{aligned}\prod_{k\geq n} y_k &= \prod_{k\geq n}(1-\frac{1}{k^2}) \\ &= \prod_{k\geq n}(1-\frac{1}{k})(1+\frac{1}{k}) \\ &= \frac{n-1}{n}\frac{n+1}{n}\frac{n}{n+1}\frac{n+2}{n+1}\frac{n+1}{n+2}\frac{n+3}{n+2}\cdots = \frac{n-1}{n}.\end{aligned}$$

Thus, $\lim_{n\to\infty}\prod_{k\geq n} y_k = 1$.

c. Since $0 \leq y_k \leq 1$, then y_k^c is a decreasing function with respect to c. Thus, if $c < 1$, we have $y_k^c > y_k$ and $1 - y_k^c < 1 - y_k$. On this basis, if $\sum_{k=0}^{\infty}(1-y_k) < \infty$, $\sum_{k=0}^{\infty}(1-y_k^c) \leq \sum_{k=0}^{\infty}(1-y_k) < \infty$ holds. When $c > 1$, since $1 - y_k^c > 1 - y_k$, we cannot determine the convergence of $\sum_{k=0}^{\infty}(1-y_k^c)$.

5.13 Write a simulation program for the "fast" Algorithm 5.3. Run it for a simple example with, say, $S = 3$, and each $\mathcal{A}(i)$, $i \in \mathcal{S}$, containing three to five actions. Record the sequence of d_k, $k = 0, 1, 2, \cdots$, and observe its behavior, e.g.; how it changes from one policy to another one. Run it for a few times with different N's.

Solution: We consider a simple example: There are 3 states in $\mathcal{S}$ and 4 actions in each $\mathcal{A}(i)$, $i \in \mathcal{S}$. The transition probability matrix under action i is P_i defined as:

$$P_1 = \begin{bmatrix} 0.3 & 0.3 & 0.4 \\ 0.2 & 0.5 & 0.3 \\ 0.5 & 0.1 & 0.4 \end{bmatrix}, P_2 = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.5 & 0.2 & 0.3 \\ 0.5 & 0.2 & 0.3 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.5 & 0.2 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}, P_4 = \begin{bmatrix} 0.3 & 0.4 & 0.3 \\ 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \end{bmatrix}.$$

The reward function is $f(i) = i, i = 1, 2, 3 \in \mathcal{S}$. For this example, the optimal policy is $d(1) = 2, d(2) = 4, d(3) = 3$. We simulate this example and have the simulation results as follows:

a. When $N = 10$, the policy sequence is $d_0 = (1,1,1), d_1 = (1,4,3), d_2 = (1,4,1), d_3 = (1,4,3), d_4 = (1,4,3), d_5 = (1,4,3), d_6 = (1,4,3), d_7 = (2,4,3), d_8 = (2,1,3), d_9 = (2,4,3), d_{10} = (2,1,3), d_{11} = (2,1,3), d_{12} = (2,1,3), d_{13} = (2,1,3), d_{14} = (2,1,3), d_{15} = (2,1,3), d_{16} = (2,4,3), d_{17} = (2,1,3), d_{18} = (2,4,3), d_{19} = (2,4,3), d_{20} = (2,4,3), d_{21} = (2,4,3), d_{22} = (2,4,3), d_{23} = (2,4,3), d_{24} = (2,4,3), d_{25} = (2,4,3), d_{26} = (2,4,3), d_{27} = (2,4,3), d_{28} = (2,4,3), \cdots$.

b. When $N = 100$, the policy sequence is $d_0 = (1,1,1), d_1 = (1,4,3), d_2 = \cdots = d_{38} = (1,4,3), d_{39} = d_{40} = \cdots = d_{138} = (2,4,3)$.

From the above simulation results, we can find that the policy always keeps invariable for several iterations, which can collect more information under this policy. Moreover, when N is small, there are more oscillations, this is because the estimates of potentials are not very accurate. A large N will give us more accurate potential estimate but bring more computation simultaneously.

5.14 This problem is designed to help to understand the remark on the proofs in Section 5.3.1. Consider an ergodic Markov chain $\boldsymbol{X} = \{X_0, X_1, \cdots, X_l, \cdots\}$ with state space $\mathcal{S}$ and reward function $f(i)$, $i \in \mathcal{S}$. Let $i^* \in \mathcal{S}$ be a special state. Suppose that we let the Markov chain stop when $X_l = X_{l+2} = i^*$, and when it stops, the total reward is $f(X_{l+1})$. Prove

a. The expected total reward is $\bar{r} = \sum_{k \in \mathcal{S}} p(k|i^*) f(k)$.

b. We may prove that the Markov chain stops w.p.1 under the special condition $p(i^*|i^*) \neq 0$.

Obviously, $p(i^*|i^*) \neq 0$ is not a necessary condition, and this special condition does not change the expected total reward $\bar{r}$ in part a.

Solution:

a. When the Markov chain stops, we have a total reward $f(X_{l+1})$, where $X_{l+1} = k$ with a probability $p(k|i^*)$. Thus, the expected total reward is $\sum_{k \in \mathcal{S}} p(k|i^*) f(k)$.

b. Since the Markov chain is ergodic, if the Markov chain does not stop, when l is large enough, there is a constant $p > 0$ such that the probability that $X_l = X_{l+1} = X_{l+2} = i^*$

is $\pi_l(i^*)p(i^*|i^*)p(i^*|i^*) \geq p > 0$, where $\pi_l(i^*)$ is the probability that the state stays at i^* at time l. Now, we divide the sample path into many intervals. Each consists of three consecutive states. Therefore, the probability that in the first k intervals there is no such interval that three consecutive states are all i^* is less than $(1-p)^k$. As $k \to \infty$, this probability goes to zero. That is, the case that 3 consecutive states are all i^* can occur with probability 1. Thus, the Markov chain stops with probability 1.

Obviously, the condition that three consecutive states are all i^* is not a necessary condition because if there exists a state j such that $p(j|i^*) > 0$ and $p(i^*|j) > 0$, the Markov chain can also stop with probability 1.

5.15 If we implement Algorithm 5.3 for a few reference states i^* in parallel, then we can update the policy whenever the system reaches one of these states. In the extreme case, if we implement the algorithm for every state, we may update the policy at every state transition.

We need to study the convergence of such algorithms. Consider, for example, the case where we have two reference states i^* and j^*. Whenever we meet states i^* or j^*, we will update the policy. Therefore, if in a period starting from one i^* to the next i^*, the sample path visits state j^*, then the policy used in this period before visiting j^* is different from that used after the visit. Does this cause a major problem in the convergence of the algorithm? How about the algorithm in which we use all states as reference states?

Solution: For simplicity, we consider two reference states i^* and j^*. Firstly, We cannot simultaneously use i^* and j^* as reference states to estimate the potential. Different reference states may lead to different potentials. If we use i^* as reference state, then we can obtain the estimation of potential g such that $g(i^*) = 0$. However, if we use j^* as the reference state, we obtain the estimation of potential g such that $g(j^*) = 0$. Two potentials are up to a constant. we cannot mix these two different potentials to carry out the policy improvement. Of course, we may only choose i^* (or j^*) as reference state. When j^* is met, the estimation of potential is not updated.

Secondly, when the policy used in a period before visiting j^* is different from that used after the visit, the new policy generated after the visit generally would not make a big impact on the estimations of potentials, thus, the subsequent policies are most likely same

as the new policy until the potential estimates under the new policy are more accurate. The policy gets update once enough data under this policy is collected. The case that all states as reference states is the same as the above case.

6

Solutions to Chapter 6

6.1 Let us revisit the stochastic approximation algorithm (6.1) when the function $f(x)$ is known. In the proof of convergence, we have assumed that the function is convex and $\frac{df(x)}{dx} > 0$. Consider the convex function $f(x) = x^2$ with a zero at $x = 0$ at which $\frac{dx^2}{dx} = 0$. Modify the proof in the text to fit this case.

Solution: Suppose $x_0 > 0$, we have $f(x_0) > 0$. Since $(\frac{df(x)}{dx})_{x=x_0} = 2x_0 > 0$, we have $x_0 > x_1$. By the same argument, we have $x_k > x_{k+1}$. Because the function is convex, the curve of $f(x)$ always lies on the same side of the tangent lines. Therefore, we have $x_k > 0$ for all $k = 0, 1, 2, \cdots$. Since $f(x)$ is increasing when $x > 0$, we also have $f(x_k) > f(x_{k+1}), k = 0, 1, 2, \ldots$. Thus, we have two decreasing sequences

$$x_0 > x_1 > \cdots > x_k > x_{k+1} > \cdots > 0$$
$$f(x_0) > f(x_1) > \cdots > f(x_k) > f(x_{k+1}) > \cdots > f(0) = 0.$$

Suppose $x_0 < 0$, we have $f(x_0) > 0$. Since $(\frac{df(x)}{dx})_{x=x_0} = 2x_0 < 0$, we have $x_0 < x_1$. By

the same argument, we have two sequences

$$x_0 < x_1 < \cdots < x_k < x_{k+1} < \cdots < 0$$
$$f(x_0) > f(x_1) > \cdots > f(x_k) > f(x_{k+1}) > \cdots > f(0) = 0.$$

Next, we continue the argument of the case that $x_0 > 0$. The same argument can be carried out for the case that $x_0 < 0$. The decreasing sequence $x_0, x_1, \cdots, x_k \cdots$ monotonously converges to a point denoted as $\hat{x}$. Then, $\hat{x} \geq 0$. Next, we prove $\hat{x} = 0$. Otherwise, $\hat{x} > 0$, thus we have $f(\hat{x}) > 0$ and $(\frac{df(x)}{dx})_{x=\hat{x}} = 2\hat{x} > 0$. Similarly to the argument in the text, we can define $\epsilon = \frac{1}{(\frac{df(x)}{dx})_{x=\hat{x}}} f(\hat{x}) > 0$. Since $\lim_{k\to\infty} x_k = \hat{x}$ and

$$\lim_{k\to\infty} \frac{1}{(\frac{df(x)}{dx})_{x=x_k}} f(x_k) = \frac{1}{(\frac{df(x)}{dx})_{x=\hat{x}}} f(\hat{x}) = \epsilon > 0,$$

There must be a point denoted as $x_{\tilde{k}}$ such that $x_{\tilde{k}} - \hat{x} < \epsilon/2$ and $\frac{1}{(\frac{df(x)}{dx})_{x=x_{\tilde{k}}}} f(x_{\tilde{k}}) > \epsilon/2$. Thus, we have

$$x_{\tilde{k}+1} = x_{\tilde{k}} - \frac{1}{(\frac{df(x)}{dx})_{x=x_{\tilde{k}}}} f(x_{\tilde{k}}) < x_{\tilde{k}} - \epsilon/2 < \hat{x}.$$

This is impossible. Thus we have $\hat{x} = 0$.

6.2 Study the convergence property of the sequence x_k, $k = 0, 1, \ldots$, in Example 6.1, for the following cases $1 > \kappa > 0, 2 > \kappa > 1, \kappa = 2$, and $\kappa > 2$, respectively, by using the figure illustrated in Figure 6.1.

Solution: Firstly, we consider the case that $1 > \kappa > 0$. Since κ is equivalent to $\frac{1}{\frac{df(x)}{dx}|_{x=x_k}}$, from $1 > \kappa > 0$, we have $\frac{df(x)}{dx}|_{x=x_k} > 1$. We can approximately draw Figure 6.1 similarly to Figure 6.1. Thus, we can find the sequence $x_k, k = 0, 1, 2, \cdots$ converges.

We can use contract mapping theorem to strictly prove the convergence of sequence x_k. Defining a mapping $g(x) = x - \kappa(x - b) = (1 - \kappa)x + \kappa b$, we can find $|g(x) - g(y)| \leq (1 - \kappa)|x - y|$. Thus, $g(x)$ is a contract mapping when $0 < \kappa < 1$. From the contract mapping theorem, we can prove the convergence of sequence $x_k, k = 0, 1, \cdots$.

Next, we consider the case that $1 < \kappa < 2$. From $1 < \kappa < 2$, we have $\frac{1}{2} < \frac{df(x)}{dx}|_{x=x_k} < 1$. We can approximately draw Figure 6.2. Thus, the sequence also converges when $1 < \kappa < 2$. Similarly, we can draw Figure 6.3 and Figure 6.4 when $\kappa = 2$ and $\kappa > 2$, respectively. We can find the sequence circles when $\kappa = 2$ and diverges when $\kappa > 2$.

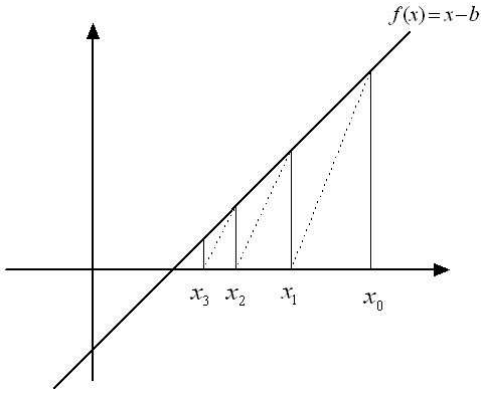

Figure 6.1: When $1 > \kappa > 0$

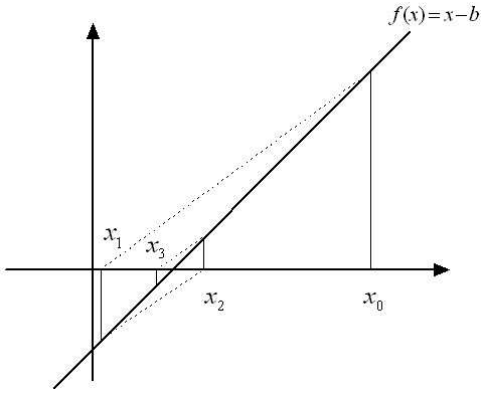

Figure 6.2: When $1 < \kappa < 2$

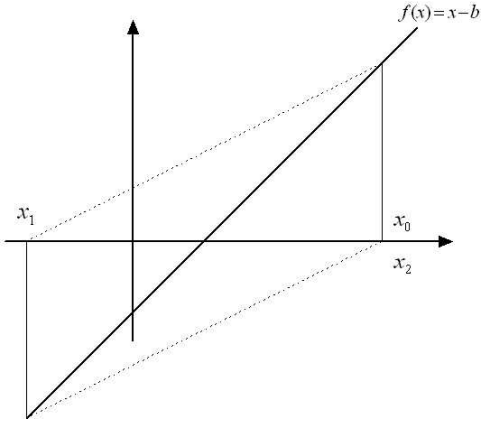

Figure 6.3: When $\kappa = 2$

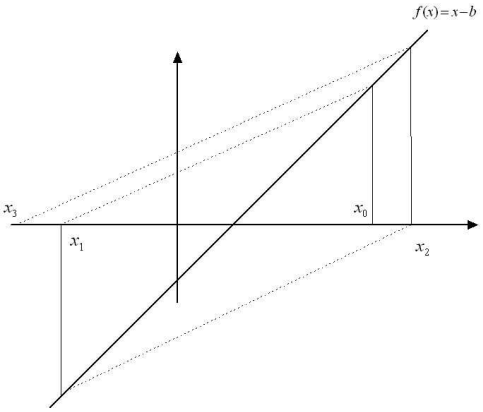

Figure 6.4: When $1 < \kappa < 2$

6.3 The algorithm in (6.12) can be used to estimate the mean of a random variable ω. This has been verified for step sizes $\kappa_k = \frac{1}{k+1}, k = 0, 1, \cdots$, in Section 6.1.2.

a. Study the case for step sizes $\kappa_k = \frac{1}{2(k+1)}, k = 0, 1, \cdots$.

b. Choose a few sequence of κ_k, $k = 0, 1, \cdots$, that satisfy conditions (6.11) or (6.10) and run simulation to see whether the sequences of $x_k, k = 0, 1, \cdots$, converge and compare their convergence speeds, if possible.

Solution:

a. If we take $\kappa_k = \frac{1}{2(k+1)}, k = 0, 1, \cdots$, then we have

$$\begin{aligned} & x_{k+1} \\ = \; & (1 - \frac{1}{2(k+1)})x_k + \frac{1}{2(k+1)}w_k \end{aligned}$$

$$
\begin{aligned}
&= \frac{2k+1}{2(k+1)}x_k + \frac{1}{2(k+1)}w_k \\
&= \frac{2k+1}{2(k+1)}\frac{2k-1}{2k}\frac{2k-3}{2(k-1)}\cdots\frac{1}{2}x_0 + \{\frac{1}{2(k+1)}w_k + \\
&\quad \frac{2k+1}{2(k+1)}\frac{1}{2k}w_{k-1} + \cdots + \frac{2k+1}{2(k+1)}\frac{2k-1}{2k}\frac{2k-3}{2(k-1)}\cdots\frac{1}{2}w_0\}. \qquad (6.1)
\end{aligned}
$$

Firstly, we prove $\frac{2k+1}{2(k+1)}\frac{2k-1}{2k}\frac{2k-3}{2(k-1)}\cdots\frac{1}{2} = \prod_{n=0}^{k}\frac{2n+1}{2(n+1)} \to 0$ when $k \to \infty$. Since

$$
\ln\prod_{n=0}^{k}\frac{2n+1}{2(n+1)} = \ln\prod_{n=0}^{k}(1-\frac{1}{2(n+1)}) = \sum_{n=0}^{k}\ln(1-\frac{1}{2(n+1)}).
$$

and $\ln(1-\frac{1}{2(n+1)}) < -\frac{1}{2(n+1)}$, we have $\ln\prod_{n=0}^{k}\frac{2n+1}{2(n+1)} < -\sum_{n=0}^{k}\frac{1}{2(n+1)} \to -\infty$, Thus, we have $\ln\prod_{n=0}^{k}\frac{2n+1}{2(n+1)} \to -\infty$ and $\prod_{n=0}^{k}\frac{2n+1}{2(n+1)} \to 0$.

Next, if we assume w with a finite mean $E(w) < c$ and a finite variance $E[w-E(w)]^2 = \sigma^2$, we prove the convergence of $x_n, n = 0, 1, \cdots$, under the more general condition (6.10). Here, we need a lemma from Loeve [1]:

Lemma 1 *Let $\{v_n\}$ be a sequence of random variables such that $\sum_{n=1}^{\infty} Ev_n^2 < \infty$. Then $\sum_{j=1}^{n}[v_j - E(v_j|v_1, \cdots, v_{j-1})]$ converges to a random variable with probability one.*

Firstly, we can prove the sequence $\{x_{n+1} + \sum_{k=1}^{n}\kappa_k[x_k - E(w)]\}$ converges to a random variable with probability one.

Let $v_k = \kappa_k[w_k - E(w)]$, then $E\{v_k^2\} = \kappa_k^2 E\{[w_k - E(w)]^2\} = \kappa_k^2\sigma^2$. Then, we have $\sum_{k=1}^{\infty} E\{v_k^2\} = \sigma^2\sum_{k=1}^{\infty}\kappa_k^2 < \infty$. Moreover, since $w_k, k = 1, 2, \cdots$ are i.i.d, we have $E[w_k - E(w)|w_1 - E(w), \cdots, w_{k-1} - E(w)] = 0$. Thus, from the above lemma, we obtain that $\sum_{k=1}^{\infty}\kappa_k[w_k - E(w)]$ converges with probability one. Since $x_{n+1} = x_1 - \sum_{k=1}^{n}\kappa_k(x_k - w_k)$, then we obtain $x_{n+1} + \sum_{k=1}^{n}\kappa_k[x_k - E(w)] = x_1 + \sum_{k=1}^{n}\kappa_k[w_k - E(w)]$ converges with probability one.

Next, we prove the convergence of x_n. Firstly, we prove $P\{\lim_{n\to\infty} x_n = +\infty\} + P\{\lim_{n\to\infty} x_n = -\infty\} = 0$. Suppose $\{x_n\}$ is a sequence with $\lim_{n\to\infty} x_n = \infty$. Then for n sufficiently large we have $x_n - E(w) > 0$, Then $\lim_{n\to\infty}\{x_{n+1} + \sum_{k=1}^{n}\kappa_k[x_k - E(w)]\} = +\infty$, but this can only happen with probability zero from the above argument. Thus, we have proved $P\{\lim_{n\to\infty} x_n = +\infty\} + P\{\lim_{n\to\infty} x_n = -\infty\} = 0$. Now, we suppose that x_n does not converge. Then, there exists a set of sequences of positive probability with the

following properties:

$$\begin{cases} (a) & x_{n+1} + \sum_{k=1}^{n} \kappa[x_k - E(w)] \text{ converges to a finite number} \\ (b) & \liminf_{n\to\infty} x_n < \limsup_{n\to\infty} x_n. \end{cases}$$

for every sequence in the set. Let $\{x_n\}$ be such a sequence and assume $\limsup_{n\to\infty} x_n > E(w)$. (A similar argument handles the situation $\limsup_{n\to\infty} x_n \le E(w)$.) Then we can choose numbers a and b satisfying

$$a > E(w), \quad \liminf x_n < a < b < \limsup x_n.$$

Since $\lim_{n\to\infty} \kappa_n = 0$, we may choose N so large that $N \le n < m$ implies

$$\begin{cases} (a) & \kappa_n \le min\left\{\frac{1}{3}, \frac{b-a}{3[c+|E(w)|]}\right\} \\ (b) & |x_m - x_n + \sum_{k=n}^{m-1} \kappa_k (x_k - E(w))| \le \frac{b-a}{3}. \end{cases} \tag{6.2}$$

Now choose m and n such that

$$\begin{cases} (a) & N \le n < m \\ (b) & x_n < a,\ x_m > b \\ (c) & n < j < m \text{ implies } a \le x_j \le b. \end{cases} \tag{6.3}$$

Then, we obtain

$$x_m - x_n \le \frac{(b-a)}{3} + \sum_{k=n}^{m-1} \kappa_l (E(w) - x_k) \le \frac{(b-a)}{3} + \kappa_n (E(w) - x_n). \tag{6.4}$$

If $E(w) < x_n$, we obtain

$$x_m - x_n \le (b-a)/3$$

which is a contradiction to (b) in (6.3). Suppose then $E(w) \ge x_n$, we have

$$|x_n - E(w)| \le c + |x_n| \le c + |E(w)| + |E(w) - x_n| \le c + |E(w)| + (x_m - x_n).$$

Hence, by applying (6.4), we have

$$x_m - x_n \le (b-a)/3 + \kappa_n[c + |E(w)|] + \kappa_n (x_m - x_n).$$

Thus, $x_m - x_n \le 2(b-a)/3(1-\kappa_n) \le b - a$ by (6.2). But this is again a contradiction to (6.3), we prove the convergence of x_k. Combined with the convergence of $\{x_{n+1} +$

$\sum_{k=1}^{n} \kappa_k[x_k - E(w)]\}$, we know $\sum_{k=1}^{n} \kappa_k[x_k - E(w)]$ converges with probability one. Then by using $\sum_{k=1}^{\infty} \kappa_k = \infty$, we have $x_k \to E(w)$ with probability one.

References:

[1] M. Loève, "On Almost Sure Convergence", Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1951, pp. 279-303.

[2] J. R. Blum, "Approximation Methods which Converge with Probability one", The Annals of Mathematical Statistics, vol. 25, no. 2, pp. 382-386, 1954.

b. Considering a $[0, 1]$-uniformly distributed random variable ω and letting $\kappa_k = \frac{1}{k+1}, \frac{1}{2(k+1)}, \frac{1}{\sqrt{k+1}}$, respectively, we simulate them respectively and get the simulation results as Figure 6.5 and Figure 6.6.

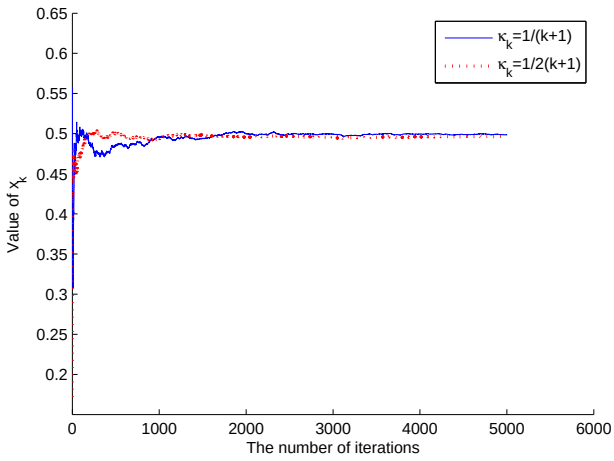

Figure 6.5: The comparison of x_k when $\kappa = \frac{1}{k+1}$ and $\kappa = \frac{1}{2(k+1)}$

6.4 Let us revisit Section 6.1.2 "Estimate Mean Values". Assume that the step sizes satisfy $\sum_{k=0}^{\infty} \kappa_k = \infty$ and $\sum_{k=0}^{\infty} \kappa_k^2 < \infty$. Working on (6.12) recursively, we may obtain

$$x_{k+1} = a_k x_0 + \xi_k,$$

a. Derive an expression of a_k and ξ_k in term of $\kappa_0, \cdots, \kappa_k$ and $\omega_0, \cdots, \omega_k$.

b. Prove $\lim_{k\to\infty} a_k = 0$, $\lim_{k\to\infty} E(\xi_k) = E(\omega)$, and $\lim_{k\to\infty} var(\xi_k) = 0$.

Solution:

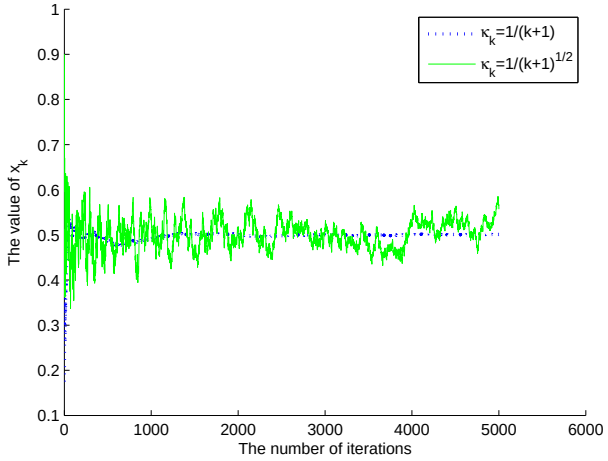

Figure 6.6: The comparison of x_k when $\kappa = \frac{1}{k+1}$ and $\kappa = \frac{1}{\sqrt{k+1}}$

a. Going on the iteration (6.12), we have

$$
\begin{aligned}
& x_{k+1} \\
= \; & x_k - \kappa_k(x_k - \omega_k) \\
= \; & (1-\kappa_k)x_k + \kappa_k\omega_k \\
= \; & (1-\kappa_k)[(1-\kappa_{k-1})x_{k-1} + \kappa_{k-1}\omega_{k-1}] + \kappa_k\omega_k \\
= \; & (1-\kappa_k)(1-\kappa_{k-1})x_{k-1} + (1-\kappa_k)\kappa_{k-1}\omega_{k-1} + \kappa_k\omega_k \\
= \; & (1-\kappa_k)(1-\kappa_{k-1})[(1-\kappa_{k-2})x_{k-2} + \kappa_{k-2}\omega_{k-2}] + (1-\kappa_k)\kappa_{k-1}\omega_{k-1} + \kappa_k\omega_k \\
= \; & (1-\kappa_k)(1-\kappa_{k-1})(1-\kappa_{k-2})x_{k-2} + (1-\kappa_k)(1-\kappa_{k-1})\kappa_{k-2}\omega_{k-2} \\
& +(1-\kappa_k)\kappa_{k-1}\omega_{k-1} + \kappa_k\omega_k \\
\vdots & \\
= \; & (1-\kappa_k)(1-\kappa_{k-1})\cdots(1-\kappa_0)x_0 + (1-\kappa_k)\cdots(1-\kappa_1)\kappa_0\omega_0 + (1-\kappa_k)\cdots(1-\kappa_{l+1})\kappa_l\omega_l \\
& + \cdots + \kappa_k\omega_k
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
a_k &= (1-\kappa_k)(1-\kappa_{k-1})\cdots(1-\kappa_0) \\
\xi_k &= (1-\kappa_k)\cdots(1-\kappa_1)\kappa_0\omega_0 + (1-\kappa_k)\cdots(1-\kappa_{l+1})\kappa_l\omega_l + \cdots + \kappa_k\omega_k
\end{aligned}
$$

b. Set $f(x) = e^{-x} - 1 + x$, we have $f'(x) = 1 - e^{-x} \geq 0$ when $x \geq 0$. Thus, $f(x)$ is increasing and we have $f(x) \geq f(0) = 0$ when $x \geq 0$. Therefore, $1 - x \leq e^{-x}$ for $x \geq 0$.

Taking the logarithm on both sides, we have $\ln(1-x) \leq -x$. Since

$$\ln a_k = \sum_{l=0}^{k} \ln(1-\kappa_l) \leq -\sum_{l=0}^{k} \kappa_l,$$

we have

$$0 \leq a_k \leq e^{-\sum_{l=0}^{k} \kappa_l}.$$

From $\sum_{k=0}^{\infty} \kappa_k = \infty$, we have $a_k \to 0$.

Next, if we assume w with a finite mean $E(w)$ and a finite variance $E[w-E(w)]^2$, we prove $\lim_{k\to\infty} var(\xi_k) = 0$. Firstly, we consider $E[x_{k+1} - E(\omega)]^2$.

$$\begin{aligned}
b_{k+1} &:= E[x_{k+1} - E(\omega)]^2 \qquad (6.5)\\
&= E\{E[(x_{k+1} - E(\omega))^2 | x_k]\}\\
&= E\Big\{E\Big[\Big(x_k + \kappa_k(x_k - \omega_k) - E(\omega)\Big)^2 | x_k\Big]\Big\}\\
&= E(x_k - E(\omega))^2 + \kappa_k^2 E\{E[(x_k - \omega_k)^2 | x_k]\} - 2\kappa_k E\{E[(x_k - E(\omega))(x_k - \omega_k)|x_k]\}\\
&= b_k + \kappa_k^2 E\{E[(x_k - \omega_k)^2 | x_k]\} - 2\kappa_k b_k. \qquad (6.6)
\end{aligned}$$

By using the result in the proof of Problem 6.3, we know x_k converges with probability one. Thus, we can prove $E\{E[(x_k - \omega_k)^2|x_k]$ is bounded for any x_k. We assume $E\{E[(x_k - \omega_k)^2|x_k] \leq C$. Summing (6.6), we obtain

$$b_{n+1} = b_1 + C\sum_{k=1}^{n} \kappa_k^2 - 2\sum_{k=1}^{n} \kappa_k b_k.$$

Since $b_{n+1} \geq 0$, it follows that

$$\sum_{k=1}^{n} \kappa_k b_k \leq \frac{1}{2} b_1 + C\sum_{k=1}^{\infty} \kappa_k^2 < \infty.$$

Hence the series $\sum_{k=1}^{\infty} \kappa_k b_k$ converges. Thus, we have $b_k \to 0$; otherwise, this is a contradiction with $\sum_{k=1}^{\infty} \kappa_k = \infty$. This means x_k converges to $E(w)$ in the mean square sense. From $x_{k+1} = a_k x_0 + \xi_k$, we have

$$E[a_k x_0 + \xi_k - E(w)]^2 = a_k^2 x_0^2 + 2a_k x_0 E[\xi_k - E(w)] + E[\xi_k - E(w)]^2 \to 0. \qquad (6.7)$$

Since x_k converges to a random variable with probability one, then $E[\xi_k - E(w)]$ is bounded. From (6.7), we have $E[\xi_k - E(w)]^2 \to 0$, which means $\lim_{k\to\infty} var(\xi_k) = 0$.

From Cauchy's inequality, we have $E[\xi_k - E(w)] \leq \{E[\xi_k - E(w)]^2\}^{1/2} \to 0$, which means $\lim_{k\to\infty} E(\xi_k) = E(w)$.

6.5 Consider the estimation of a continuous time average reward. Let $\{T_0, T_1, \cdots, T_l, \cdots\}$ be the sequence of transition times of a continuous Markov process with $T_0 = 0$. The state in the time period $[T_l, T_{l+1})$ is $X_l, l = 0, 1, \cdots$, and set $\tau_l = T_{l+1} - T_l, l = 0, 1, \cdots$. The reward rate function is $f(X_l)$ and the average reward is defined as

$$\eta = \lim_{l\to\infty} \eta_l, \quad \text{w.p.1}, \qquad \eta_l = \frac{1}{T_l}\int_0^{T_l} f[X(t)]dt.$$

We wish to develop a recursive formula for η_l as follows:

$$\eta_{l+1} = \eta_l + \kappa_l[f(X_l) - \eta_l],\ l = 0, 1, \cdots, \text{with } \eta_0 = 0.$$

Please find the value of $\kappa_l, l = 0, 1, \cdots$, in term of T_l, etc.

Solution:

$$\begin{aligned}
\eta_{l+1} &= \frac{1}{T_{l+1}}\int_0^{T_{l+1}} f[X(t)]dt \\
&= \frac{1}{T_{l+1}}\Big\{\int_0^{T_l} f[X(t)]dt + f(X_l)(T_{l+1} - T_l)\Big\} \\
&= \frac{T_l}{T_{l+1}}\eta_l + \frac{T_{l+1} - T_l}{T_{l+1}} f(X_l) \\
&= \eta_l + \frac{T_{l+1} - T_l}{T_{l+1}}[f(X_l) - \eta_l].
\end{aligned}$$

Thus we get $\kappa_l = \frac{T_{l+1}-T_l}{T_{l+1}}$.

6.6 Derive the $TD(0)$ algorithm for the discounted performance criterion:

$$\eta_\beta(i) = E\{\sum_{l=0}^{\infty} \beta^l f(X_l)|X_0 = i\}, \quad 1 > \beta > 0.$$

Solution: Denote $X_l = i$, we have

$$\begin{aligned}
\eta_\beta(i) &= E\{\sum_{k=0}^{\infty} \beta^k f(X_{l+k})|X_l = i\} \\
&= f(i) + E\Big\{E\{\sum_{k=1}^{\infty} \beta^k f(X_{l+k})|X_{l+1}\}|X_l = i\Big\} \\
&= f(i) + \beta E[\eta_\beta(X_{l+1})|X_l = i].
\end{aligned}$$

From this, we have

$$\eta_\beta(X_l) = E\{f(X_l) + \beta\eta_\beta(X_{l+1})|X_l\}.$$

Therefore, we can use $f(X_l)+\beta\eta_\beta(X_{l+1})$ as an estimate of $\eta_\beta(X_l)$. Thus, by the stochastic approximation algorithm (6.12), we can obtain the $TD(0)$ algorithm for the discounted performance criterion:

$$\begin{aligned}\eta_\beta(X_l) &:= \eta_\beta(X_l) - \kappa_l\{\eta_\beta(X_l) - [f(X_l) + \beta\eta_\beta(X_{l+1})]\} \\ &= \eta_\beta(X_l) + \kappa_l[f(X_l) + \beta\eta_\beta(X_{l+1}) - \eta_\beta(X_l)].\end{aligned}$$

6.7 $TD(0)$ *with random steps*: For any two states $i, j \in \mathcal{S}$, set $\mathcal{S}_0 = \{i, j\}$. Consider a sample path of a Markov chain $\{X_0, \cdots, X_l, \cdots\}$. Denote the time sequence at which the Markov chain is in $\mathcal{S}_0$ as $l_0, l_1, \cdots, l_k, \cdots$. We may set $g(i) = 0$.

a. Develop a $TD(0)$ algorithm for estimating $g(j)$, by using the temporal difference obtained in the periods from $l_k + 1$ to l_{k+1}, $k = 0, 1, \cdots$.

b. Explain that the algorithm converges to the right value, compare it with the realization factor $\gamma(i, j) = g(j) - g(i)$.

Solution:

a. For any $X_{l_k} = j \in \mathcal{S}_0, k = 0, 1, \cdots$, we have

$$g(X_{l_k}) = E\{\sum_{n=0}^{\infty}[f(X_{l_k+n}) - \eta]|X_{l_k}\}.$$

Therefore,

$$g(X_{l_k}) = E\{\sum_{k=l_k}^{l_{k+1}-1}[f(X_k) - \eta] + g(X_{l_{k+1}})|X_{l_k}\}.$$

Then, the $TD(0)$ algorithm is

$$g(X_{l_k}) := g(X_{l_k}) + \kappa_k\{\sum_{l=l_k}^{l_{k+1}-1}[f(X_l) - \eta] + g(X_{l_{k+1}}) - g(X_{l_k})\}. \tag{6.8}$$

b. From (6.8), we have

$$\begin{aligned} & g(X_{l_k}) - g(X_{l_{k+1}}) \\ := \; & g(X_{l_k}) - g(X_{l_{k+1}}) + \kappa_k \{ \sum_{l=l_k}^{l_{k+1}-1} [f(X_l) - \eta] - [g(X_{l_k}) - g(X_{l_{k+1}})] \} \\ = \; & g(X_{l_k}) - g(X_{l_{k+1}}) - \kappa_k \{ [g(X_{l_k}) - g(X_{l_{k+1}})] - \sum_{l=l_k}^{l_{k+1}-1} [f(X_l) - \eta] \}. \end{aligned}$$

Thus, according to realization factor $\gamma(X_{l_{k+1}}, X_{l_k}) = E\{\sum_{l=l_k}^{l_{k+1}-1}[f(X_l) - \eta]\}$, the above $TD(0)$ algorithm in fact estimates the realization factor $\gamma(X_{l_{k+1}}, X_{l_k}) = g(X_{l_k}) - g(X_{l_{k+1}})$ by using the Robbins-Monro algorithm (6.6). Thus, the algorithm converges under some conditions that Robbins-Monro algorithm requires.

6.8 Consider a two-state Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

and reward function $f(1) = 1$ and $f(0) = 0$. We have $\eta = \frac{1}{2}$.

a. What are the potentials for the two states?

b. Write a computer program applying algorithm(6.15), (6.22) and (6.24), and observe the trends of the convergence of the sequences generated by these algorithms. (For Algorithm (6.22), observe the trend of convergence of $g(1) - g(0)$.)

Solution:

a. From the balance equations, we can obtain $\pi = (0.5, 0.5)$. From (2.13), the potentials defined as (6.13) is $g = (I - P + e\pi)^{-1} f - \eta = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$.

b. We apply algorithm (6.15), (6.22) and (6.24) to this problem with initial value $g(0) = g(1) = 0$, respectively. In applying algorithm (6.22), we set $G = 0.5$ and $i^* = 1$. In applying algorithm (6.25), we set $i^* = 1$. The simulation results are Figure 6.7, 6.8 and 6.9, respectively.

6.9 The $TD(0)$ algorithm (6.15) and (6.16) can only determine the potentials up to an additive constant. That is, starting from different initial values, the algorithm converges to

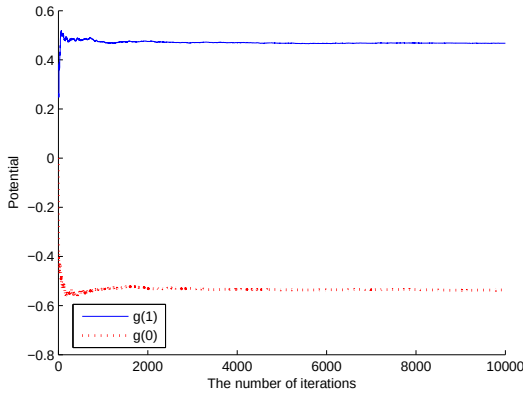

Figure 6.7: Algorithm (6.15)

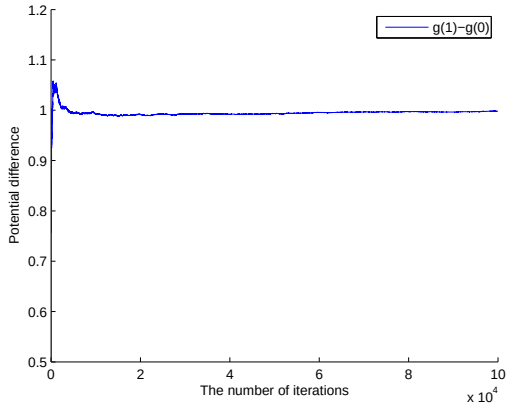

Figure 6.8: Algorithm (6.22)

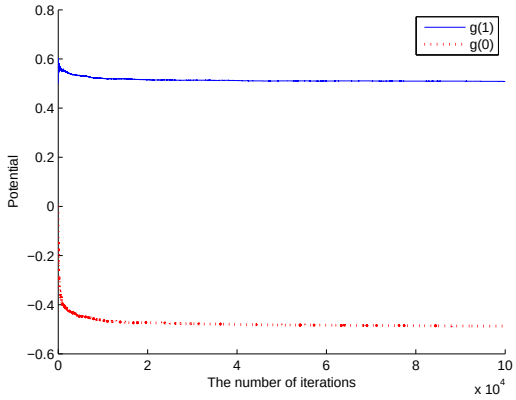

Figure 6.9: Algorithm (6.24)

different sets of potentials that have the same perturbation realization factor $\gamma(i,j), i,j \in \mathcal{S}$.

a. Can we fix a reference state i^* and set $g(i^*) = 0$ in the $TD(0)$ algorithm (6.15) and (6.16)?

b. If so, modify the algorithm.

c. Explain your algorithm using $g(i) = \gamma(i^*, i) = E\left\{\sum_{l=0}^{L(i^*|i)-1}[f(X_l) - \eta]\middle| X_0 = i\right\}$.

d. Apply this algorithm to the Markov chain in Example 6.5.

Solution:

a. Yes, we can fix a reference state and set $g(i^*) = 0$ in algorithm (6.15) and (6.16).

b. We can directly set $g(X_l) = 0$ when $X_l = i^*$. The updates at other states are still similar to that in algorithm (6.15).

c. Denote $X_l = i$, we have $g(i) = \gamma(i^*, i) = E\{\sum_{k=0}^{L(i^*|i)-1}[f(X_{l+k}) - \eta]|X_l = i\} = E\{[f(X_l) - \eta + g(X_{l+1})]|X_l = i\}$, From this, we have

$$g(X_l) = E\{[f(X_l) - \eta] + g(X_{l+1})|X_l\}.$$

Thus, we can use $[f(X_l) - \eta] + g(X_{l+1})$ as an estimate of $g(X_l)$. This results in the algorithm (6.15). The algorithm (6.15) can converges to different potentials if we use different initial values. We directly set $g(i^*) = 0$ when $X_l = i^*$, which limits the algorithm to keep $g(i^*) = 0$, thus, this algorithm converges to a special potential with $g(i^*) = 0$.

d. We apply the algorithm to Example 6.5. The simulation result is as Figure 6.10, where $i^* = 0$ and the initial potential is $(1, 1)^T$.

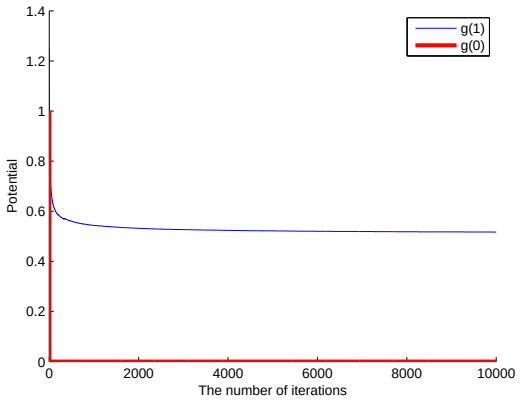

Figure 6.10: The simulation result in Problem 6.9 d)

6.10 Consider the modified algorithm (6.22).

a. Can we fix a reference state i^* and set $g(i^*) = 0$ in (6.22), as we considered in Problem 6.9? (to find the answer, apply it to the Markov chain in Example 6.5.)

b. If not, why?

Solution:

a. We cannot fix a reference state. Let's apply the modified algorithm to Example 6.5. Consider an approximate sample path $1, 0, 1, 0, 1, 0, \cdots$. We assume that state 0 is the reference state and initial values $g(1) = g(0) = 0$. We obtain

$$\begin{aligned}
&at\ l = 0: \quad g(1) = (1-1)g(1) + [f(1) + g(0)] = 1,\\
&at\ l = 1: \quad g(0) = 0,\\
&at\ l = 2: \quad g(1) = (1 - \frac{1}{3})g(1) + \frac{1}{3}[f(1) + g(0)] = 1\\
&at\ l = 3: \quad g(0) = 0;\\
&at\ l = 4: \quad g(1) = (1 - \frac{1}{5})g(1) + \frac{1}{5}[f(1) + g(0)] = 1\\
&\quad \vdots
\end{aligned}$$

Thus, we have $g(1) - g(0) = 1$, which does not converge to the true potential difference.

b. Since in algorithm (6.22), each update will result in a different potential, that is, the potentials at different times are up to a constant. If we fix a reference state, the potential at this state remains unchangeable, but the potentials at other states may be changed. Thus, the modified algorithm cannot converge to the true potential difference.

6.11 Derive an iterative numerical algorithm similar to the algorithm in (6.20) for potentials by using Equation (3.4).

Solution: From (3.4), we have

$$g = P_- g + f_-.$$

We can apply the stochastic approximation algorithm (6.6) to obtain an iterative numerical algorithm for the potential by using Equation (3.4):

$$g_{k+1}(i) := g_k(i) - \kappa_k \left\{ g_k(i) - \left[f(i) - f(S) + \sum_{j \in \mathcal{S}} [p(j|i) - p(j|S)] g_k(j) \right] \right\}.$$

6.12 Consider a finite state discrete-time birth-death process $\{X_l, l = 0, 1, \cdots\}$: The state space is $\mathcal{S} = \{0, 1, 2, \cdots, S\}$. The state is the population $n \in \mathcal{S}$. The transition probability from state n to $n+1$ (the birth rate) is $p(n+1|n) = a, n = 1, \cdots, S-1$, and the death rate is $p(n-1|n) = b, n = 1, 2, \cdots, S-1, a + b = 1$; and $p(1|0) = p(S-1|S) = 1$. Let the reward function be $f(n) = n$, the performance is defined as the average population $\eta = E_\pi[f(X_l)] = \sum_{n=0}^{S} \pi(n) f(n)$.

a. Derive a formula expressing the performance η as a function of the birth rate a.

b. Set $a = \frac{1}{2}$. Using the derivative formula (6.43) to derive the performance derivative $\frac{d\eta}{da}\Big|_{a=\frac{1}{2}}$.

c. Develop a $TD(0)$ algorithm for estimating $\frac{d\eta}{da}\Big|_{a=\frac{1}{2}}$

Solution:

a. When $a = 1$, the process will eventually cycle between $S-1$ and S. Thus, the average population is $\frac{2S-1}{2}$. When $a = 0$, the process will cycle between 0 and 1, thus the average population is $1/2$. Next, we assume $a \neq 1, 0$. By using the balance equation, we can obtain $\pi(1) = \frac{1}{1-a}\pi(0), \pi(2) = \frac{a}{(1-a)^2}\pi(0), \pi(3) = \frac{a^2}{(1-a)^3}\pi(0), \cdots, \pi(S-1) = \frac{a^{S-2}}{(1-a)^{S-1}}\pi(0), \pi(S) = \frac{a^{S-1}}{(1-a)^{S-1}}\pi(0)$. From, $\pi e = 1$, we have $[1 + \frac{1}{1-a} + \frac{a}{(1-a)^2} + \frac{a^2}{(1-a)^3} + \cdots + \frac{a^{S-2}}{(1-a)^{S-1}} + \frac{a^{S-1}}{(1-a)^{S-1}}]\pi(0) = 1$. Thus, if $a \neq \frac{1}{2}$, then $\pi(0) = \frac{(1-2a)(1-a)^{S-1}}{2[(1-a)^S - a^S]}$. The average population $\eta = \sum_{n=0}^{S}\pi(n)n = \frac{(1-2a)(1-a)^{S-1}}{2[(1-a)^S-a^S]}[\frac{1}{1-a} + \frac{2a}{(1-a)^2} + \frac{3a^2}{(1-a)^3} + \cdots + \frac{(S-1)a^{S-2}}{(1-a)^{S-1}} + \frac{Sa^{S-1}}{(1-a)^{S-1}}] = \frac{(1-a)^{S+1}+a^S(1-a)(4aS-2S-1)}{2[(1-a)^S-a^S](1-2a)(1-a)}$. If $a = \frac{1}{2}$, then $\pi(0) = \frac{1}{2S}$, the average population $\eta = \frac{S}{2}$.

b. From the derivative formula (6.43), we have

$$\frac{d\eta}{da} = E\Big\{[f(X_k) - \eta]\sum_{l=u_{m(k)}}^{k}\frac{dp(X_l|X_{l-1})/da}{p(X_l|X_{l-1})}\Big\},$$

where $u_m, m = 0, 1, \cdots$, are a sequence of regenerative points with $X_0 = 0 \in \mathcal{S}$, $u_0 = 0$, and $u_{m+1} = min\{n : n > u_m, X_n = 0\}$, and for any integer $k \geq 0$ we have $u_{m(k)} \leq k < u_{m(k)+1}$. Thus, we have

==========

$$\begin{aligned}
\frac{d\eta}{da} &= \sum_{i\in\mathcal{S}}\pi(i)[f(i)-\eta]E\left\{\sum_{l=u_{m(k)}}^{k}\frac{dp(X_l|X_{l-1})/da}{p(X_l|X_{l-1})}\Big| X_k = i\right\} \\
&= \sum_{i\in\mathcal{S}}\pi(i)[f(i)-\eta] \\
&\quad E\left\{\frac{dp(X_{u_{m(k)}}|X_{u_{m(k)}-1})/da}{p(X_{u_{m(k)}}|X_{u_{m(k)}-1})} + \frac{dp(X_{u_{m(k)+1}}|X_{u_{m(k)}})/da}{p(X_{u_{m(k)+1}}|X_{u_{m(k)}})} + \ldots\right. \\
&\quad \left. + \frac{dp(X_k|X_{k-1})/da}{p(X_k|X_{k-1})}\Big| X_k = i\right\}.
\end{aligned}$$

==========

c. The $TD(0)$ algorithm:

1. Set $\hat{r}_{-1} = 0$ and $\Delta_0 = 0$ and $k = 0$

2. For each state X_k visited, do

$$\hat{r}_k = \begin{cases} \hat{r}_{k-1} + \frac{dp(X_k|X_{k-1})/da}{p(X_k|X_{k-1})} & if\ X_k \neq i^*; \\ 0 & if\ X_k = i^* \end{cases}$$

where we assume $\frac{dp(X_0|X_{-1})/da}{p(X_0|X_{-1})} = 0$.

$$\Delta_{k+1} = \Delta_k + \kappa_k\{[f(X_k) - \eta]\hat{r}_k - \Delta_k\}.$$

6.13 Consider a randomized policy d_r. Denote $\mathcal{A}(i) := \{\alpha_{i,1}, \cdots, \alpha_{i,|\mathcal{A}(i)|}\}$, where $|\mathcal{A}(i)|$ is the number of actions in $\mathcal{A}(i), i \in \mathcal{S}$. At state i the system takes action $\alpha_{i,k} \in \mathcal{A}(i)$ with probability $p_{i,k}$, $k = 1, 2, \cdots, |\mathcal{A}(i)|$, and $\sum_{k=1}^{|\mathcal{A}(i)|} p_{i,k} = 1, i \in \mathcal{S}$. If action $\alpha \in \mathcal{A}(i)$ is taken at state i, then the transition probabilities are $p^{\alpha}(j|i), j \in \mathcal{S}$, and the performance function is $f(i, \alpha), i \in \mathcal{S}$. The Q-function are defined in (6.28) as follows

$$Q^{d_r}(i, \alpha) = \{\sum_{j=1}^{S} p^{\alpha}(j|i) g^{d_r}(j)\} + f(i, \alpha) - \eta^{d_r}, \quad \alpha \in \mathcal{A}(i), i \in \mathcal{S},$$

where $g^{d_r}(i), i \in \mathcal{S}$, are the performance potential of the system under this randomized policy d_r.

a. Determine the performance function and transition probabilities for the system under this randomized policy; Derive the Poisson equation for it.

b. Prove $g^{d_r}(i) = \sum_{k=1}^{|\mathcal{A}(i)|} p_{i,k} Q^{d_r}(i, \alpha_{i,k})$.

c. Given a deterministic policy $d(i) = \alpha_i^* \in \mathcal{A}(i), i \in \mathcal{S}$, we define an ϵ-randomized policy: with probability $1 - \epsilon$ the system takes action α_i^* and with probability $\frac{\epsilon}{(|\mathcal{A}(i)|-1)}$ it takes any other actions in $\mathcal{A}(i), i \in \mathcal{S}$. Let $g(i)$ be the potentials of the deterministic policy d, and $g_\epsilon(i)$, and $Q_\epsilon(i, \alpha), \alpha \in \mathcal{A}(i), i \in \mathcal{S}$, be the potentials and Q-function of the ϵ-randomized policy. Prove

$$\lim_{\epsilon \to 0} g_\epsilon(i) = g(i), \quad i \in \mathcal{S}$$
$$\lim_{\epsilon \to 0} Q_\epsilon(i, \alpha_i^*) = g(i), \quad i \in \mathcal{S},$$

and

$$\lim_{\epsilon\to 0} Q_\epsilon(i,\alpha) = \{\sum_{j=1}^{S} p^\alpha(j|i)g(j)\} + f(i,\alpha) - \eta, \alpha \neq \alpha_i^*, i \in \mathcal{S}.$$

Solution:

a. The performance function under the randomized policy is $f^{d_r}(i) := \sum_{k=1}^{|\mathcal{A}(i)|} p_{i,k} f(i,\alpha_{i,k})$ and the transition probability from state i to state j is $p^{d_r}(j|i) = \sum_{k=1}^{|\mathcal{A}(i)|} p_{i,k} p^{\alpha_{i,k}}(j|i)$. Then, we have the following Poisson equation

$$g^{d_r}(i) - \sum_{j\in\mathcal{S}} p^{d_r}(j|i)g^{d_r}(j) + \eta^{d_r} = f^{d_r}(i). \quad i \in \mathcal{S} \tag{6.9}$$

b. From the definition of performance potential, we have $g^{d_r}(i) = E^{d_r}\{\sum_{l=0}^{\infty}[f(X_l,A_l)-\eta]|X_0 = i\}$, where E^{d_r} denotes the expectation under the randomized policy d_r. We have

$$\begin{aligned}
& E^{d_r}\{\sum_{l=0}^{\infty}[f(X_l,A_l)-\eta]|X_0=i\} \\
= \;& E^{d_r}\Big\{E^{d_r}\{\sum_{l=0}^{\infty}[f(X_l,A_l)-\eta]|X_0=i,A_0\}|X_0=i\Big\} \\
= \;& \sum_{k=1}^{|\mathcal{A}(i)|} p_{i,k} E^{d_r}\{\sum_{l=0}^{\infty}[f(X_l,A_l)-\eta]|X_0=i,A_0=\alpha_{i,k}\} \\
= \;& \sum_{k=1}^{|\mathcal{A}(i)|} p_{i,k} Q^{d_r}(i,\alpha_{i,k}).
\end{aligned}$$

c. We firstly prove $g_\epsilon(i), i \in \mathcal{S}$ are continuous with respect to ϵ. We assume P_ϵ is the transition probability matrix under the ϵ-randomized policy, whose (i,j) component is $p_\epsilon(j|i) = (1-\epsilon)p^{\alpha_i^*}(j|i) + \frac{\epsilon}{|\mathcal{A}(i)|-1}\sum_{a\in\mathcal{A}(i)-\{\alpha_i^*\}} p^\alpha(j|i)$. From the balance equation $\pi_\epsilon P_\epsilon = \pi_\epsilon$ and $\pi_\epsilon e = 1$, where π_ϵ is the steady-state probability under the ϵ-randomized policy, we can prove π_ϵ is continuous with respect to ϵ since each component of P_ϵ is continuous with respect to ϵ. We consider a specified potential satisfying $\pi_\epsilon g_\epsilon = \eta_\epsilon$, where η_ϵ is the average performance. From Poisson equation (6.9), we have

$$g_\epsilon = (I - P_\epsilon + e\pi_\epsilon)^{-1} f_\epsilon.$$

From the continuity of inversion of matrix with respect to ε, we can easily prove $g_\epsilon(i), i \in \mathcal{S}$ are continuous with respect to ϵ. Thus, we have $\lim_{\epsilon\to 0} g_\epsilon(i) = g(i), i \in \mathcal{S}$.

Since $Q_\epsilon(i,\alpha_i^*) = \sum_{j=1}^{\mathcal{S}} p^{\alpha_i^*}(j|i)g_\epsilon(i) + f(i,\alpha_i^*) - \eta_\epsilon$, ,from the continuity of $g_\epsilon(i), i \in \mathcal{S}$ and η_ϵ with respect to ϵ, we have $\lim_{\epsilon\to 0} Q_\epsilon(i,\alpha_i^*) = \sum_{j=1}^{\mathcal{S}} p^{\alpha_i^*}(j|i)g(i) + f(i,\alpha_i^*) - \eta, \quad i \in \mathcal{S}$. Then, from the Poisson equation $g(i) = \sum_{j=1}^{\mathcal{S}} p^{\alpha_i^*}(j|i)g(i) + f(i,\alpha_i^*) - \eta$, we have $\lim_{\epsilon\to 0} Q_\epsilon(i,\alpha_i^*) = g(i), i \in \mathcal{S}$. Similarly, we have $\lim_{\epsilon\to 0} Q_\epsilon(i,\alpha) = \{\sum_{j=1}^{S} p^{\alpha}(j|i)g(j)\} + f(i,\alpha) - \eta, \alpha \neq \alpha_i^*, i \in \mathcal{S}$.

6.14 Suppose that we can only control the actions in the states in a subset of the state space $\mathcal{S}_0 \subset \mathcal{S}$ of a Markov chain, which is under a randomized policy that visits all the state-action pairs when the state is in $\mathcal{S}_0$. Denote the time sequence at which the Markov chain is in $\mathcal{S}_0$ as $l_0, l_1, \cdots, l_k, \cdots$; i.e. $X_{l_k} \in \mathcal{S}_0, k = 0, 1, \cdots$. Develop a $TD(0)$ algorithm for Q-factors $Q(i,\alpha), i \in \mathcal{S}_0$, with random steps K.

Solution: Denote $X_{l_k} = i \in \mathcal{S}_0$. From (6.33), we have

$$Q(i,\alpha) = E\{\sum_{l=l_k}^{\infty}[f(X_l,A_l) - \eta]|X_{l_k} = i, A_{l_k} = \alpha\}$$

$$= E\{\sum_{l=l_k}^{l_{k+1}-1}[f(X_l,A_l) - \eta] + Q(X_{l_{k+1}}, A_{l_{k+1}})|X_{l_k} = i, A_{k_k} = \alpha\}.$$

We can use $\sum_{l=l_k}^{l_{k+1}-1}[f(X_l,A_l) - \eta] + Q(X_{l_{k+1}}, A_{l_{k+1}})$ as an estimate of $Q(X_l,A_l)$. Thus, from (6.12), we can obtain the following $TD(0)$ algorithm:

$$Q(X_{l_k},A_{l_k}) := Q(X_{l_k},A_{l_k}) + \kappa_k\delta_{l_k},$$
$$\delta_{l_k} = \sum_{l=l_k}^{l_{k+1}-1}[f(X_l,A_l) - \eta] + Q(X_{l_{k+1}}, A_{l_{k+1}}) - Q(X_{l_k},A_{l_k}), k = 0, 1, \cdots.$$

6.15 Develop a K-step algorithms for estimating the Q-factors (c.f. (6.33) and (6.34)).

Solution:

$$Q(X_l,A_l) = E\{\sum_{k=l}^{l+K-1}[f(X_k,A_k) - \eta] + Q(X_{l+K}, A_{l+K})|X_l,A_l)\}.$$

Thus, we can use $\sum_{k=l}^{l+K-1}[f(X_k,A_k) - \eta] + Q(X_{l+K}, A_{l+K})$ as an estimate of $Q(X_l,A_l)$. From (6.12), we obtain the following K-step $TD(0)$ algorithm:

$$Q(X_l,A_l) := Q(X_l,A_l) + \kappa_l\Big\{\sum_{k=0}^{K-1}[f(X_{l+k},A_{l+k}) - \eta] + Q(X_{l+K},A_{l+K}) - Q(X_l,A_l)\Big\}$$

6.16 In (6.33) and (6.34), we may set the Q-factor of a pair of reference state-action to be zero; i.e. $Q(i^*, \alpha^*) = 0$. Develop a $TD(0)$-learning algorithm.

Solution: The $TD(0)$ algorithm with $Q(i^*, \alpha^*) = 0$:

$$Q(X_l, A_l) := Q(X_l, A_l) + \kappa_l \Big\{ [f(X_l, A_l) - \eta] + Q(X_{l+1}, A_{l+1}) - Q(X_l, A_l \Big\},$$
$$if\ X_l \neq i^* \text{ and } A_l \neq \alpha^*, l = 1, 2, \cdots$$
$$Q(i^*, a^*) = 0,\ if\ X_l = i^* \text{ and } A_l = \alpha^*$$

6.17 We partition the the state space $\mathcal{S}$ into $\mathcal{S}_0$ subsets: $\mathcal{S} = \cup_{k=1}^{S_0} \mathcal{S}_k, \mathcal{S}_k \cap \mathcal{S}_{k'} = \emptyset, k, k' = 1, 2, \cdots, S_0$. Let $\pi(i), i \in \mathcal{S}$, be the steady-state probability, and let $\pi(i|\mathcal{S}_k) = \frac{\pi(i)}{\sum_{j \in \mathcal{S}_k} \pi(j)}$ be the conditional steady-state probability of i given that $i \in \mathcal{S}_k$. The potential associated with the aggregation $\mathcal{S}_k$ is defined as (6.39):

$$g(\mathcal{S}_k) = \sum_{i \in \mathcal{S}_k} \pi(i|\mathcal{S}_k) g(i).$$

We wish to establish a Poisson equation for the aggregations:

$$g(\mathcal{S}_k) = \sum_{k'=1}^{S_0} p(\mathcal{S}_{k'}|\mathcal{S}_k) g(\mathcal{S}_{k'}) + f(\mathcal{S}_k) - \eta, \quad k = 1, 2, \cdots, S_0. \tag{6.10}$$

a. According to their physical meanings, determine the transition probabilities $p(\mathcal{S}_{k'}|\mathcal{S}_k)$ and the performance function $f(\mathcal{S}_k), k, k' = 1, 2, \cdots, S_0$.

b. Prove that the Poisson equation (6.10) holds for the aggregations if and only if for any $\mathcal{S}_{k'}, k' = 1, \cdots, \mathcal{S}_0$, and any $j \in \mathcal{S}_{k'}$, we have

$$\frac{\pi(j)}{\sum_{j' \in \mathcal{S}_{k'}} \pi(j')} = \frac{\sum_{i \in \mathcal{S}_k} \pi(i) p(j|i)}{\sum_{j' \in \mathcal{S}_{k'}} \sum_{i \in \mathcal{S}_k} \pi(i) p(j'|i)}, k = 1, \cdots, S_0. \tag{6.11}$$

c. Set

$$\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k) = \frac{\sum_{i \in \mathcal{S}_k} \pi(i) p(j|i)}{\sum_{j' \in \mathcal{S}_{k'}} \sum_{i \in \mathcal{S}_k} \pi(i) p(j'|i)}.$$

Then (6.11) becomes $\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k) = \pi(j|\mathcal{S}_{k'})$. Prove that (6.11) is equivalent to the following condition:

$$\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k) \textit{ is independent of } k. \tag{6.12}$$

d. Explain the meaning of $\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k)$ and condition (6.12).

e. Derive a $TD(0)$ algorithm for $g(\mathcal{S}_k), k = 1, \cdots, \mathcal{S}_0$.

f. Explain that the algorithm developed in e. may not work if the condition (6.12) does not hold.

Solution:

a.

$$p(\mathcal{S}_{k'}|\mathcal{S}_k) = \frac{p(i \in \mathcal{S}_k, j \in \mathcal{S}_{k'})}{p(i \in \mathcal{S}_k)} = \frac{\sum_{i\in\mathcal{S}_k} \pi(i) \sum_{j\in\mathcal{S}_{k'}} p(j|i)}{\sum_{i\in\mathcal{S}_k} \pi(i)}.$$

$$f(\mathcal{S}_k) = \sum_{i\in\mathcal{S}_k} \pi(i|\mathcal{S}_k) f(i).$$

b. For the original Markov chain, we have Poisson equation:

$$g(i) = \sum_{j\in\mathcal{S}} p(j|i)g(j) + f(i) - \eta.$$

Pre-multiplying $\pi(i|\mathcal{S}_k)$ and summing them over $\mathcal{S}_k$, we have

$$\begin{aligned}
&\sum_{i\in\mathcal{S}_k} \pi(i|\mathcal{S}_k) g(i) \\
= &\sum_{i\in\mathcal{S}_k} \pi(i|\mathcal{S}_k) \sum_{j\in\mathcal{S}} p(j|i)g(j) + \sum_{i\in\mathcal{S}_k} \pi(i|\mathcal{S}_k) f(i) - \eta \\
= &\sum_{j\in\mathcal{S}} \frac{\sum_{i\in\mathcal{S}_k} \pi(i)p(j|i)}{\sum_{i\in\mathcal{S}_k} \pi(i)} g(j) + f(\mathcal{S}_k) - \eta \\
= &\sum_{k'=1}^{S_0} \sum_{j\in\mathcal{S}_{k'}} \frac{\sum_{i\in\mathcal{S}_k} \pi(i)p(j|i)}{\sum_{i\in\mathcal{S}_k} \pi(i)} g(j) + f(\mathcal{S}_k) - \eta \\
= &\sum_{k'=1}^{S_0} \frac{\sum_{j'\in\mathcal{S}_{k'}} \sum_{i\in\mathcal{S}_k} \pi(i)p(j'|i)}{\sum_{i\in\mathcal{S}_k} \pi(i)} \sum_{j\in\mathcal{S}_{k'}} \frac{\sum_{i\in\mathcal{S}_k} \pi(i)p(j|i)}{\sum_{j'\in\mathcal{S}_{k'}} \sum_{i\in\mathcal{S}_k} \pi(i)p(j'|i)} g(j) \\
&+ f(\mathcal{S}_k) - \eta \\
= &\sum_{k'=1}^{S_0} p(\mathcal{S}_{k'}|\mathcal{S}_k) \sum_{j\in\mathcal{S}_{k'}} \frac{\sum_{i\in\mathcal{S}_k} \pi(i)p(j|i)}{\sum_{j'\in\mathcal{S}_{k'}} \sum_{i\in\mathcal{S}_k} \pi(i)p(j'|i)} g(j) + f(\mathcal{S}_k) - \eta.
\end{aligned}$$

If

$$\frac{\pi(j)}{\sum_{j'\in\mathcal{S}_{k'}} \pi(j')} = \frac{\sum_{i\in\mathcal{S}_k} \pi(i)p(j|i)}{\sum_{j'\in\mathcal{S}_{k'}} \sum_{i\in\mathcal{S}_k} \pi(i)p(j'|i)},$$

we have the Poisson equation (6.10).

c. From (6.11), it is obvious that $\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k)$ is independent of $\mathcal{S}_k$. That is, we have (6.12) from (6.11). If $\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k)$ is independent of $\mathcal{S}_k$, for any $\mathcal{S}_k$, we have

$$\begin{aligned}
& \pi(j|\mathcal{S}_{k'}, \mathcal{S}_k) \\
= & \frac{\sum_{i\in\mathcal{S}_k} \pi(i)p(j|i)}{\sum_{j'\in\mathcal{S}_{k'}} \sum_{i\in\mathcal{S}_k} \pi(i)p(j'|i)} \\
= & \frac{\sum_{k=1}^{\mathcal{S}_0} \sum_{i\in\mathcal{S}_k} \pi(i)p(j|i)}{\sum_{k=1}^{\mathcal{S}_0} \sum_{j'\in\mathcal{S}_{k'}} \sum_{i\in\mathcal{S}_k} \pi(i)p(j'|i)} \\
= & \frac{\sum_{i\in\mathcal{S}} \pi(i)p(j|i)}{\sum_{j'\in\mathcal{S}_{k'}} \sum_{i\in\mathcal{S}} \pi(i)p(j'|i)} \\
= & \frac{\pi(j)}{\sum_{j'\in\mathcal{S}_{k'}} \pi(j')} \\
= & \pi(j|\mathcal{S}_{k'}).
\end{aligned}$$

Thus, (6.11) is equivalent to "$\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k)$ is independent of $\mathcal{S}_k$".

d. $\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k)$ denotes the conditional steady-state probability that the system is in state j given that the system is in subset $\mathcal{S}_k$ in the previous time and in subset $\mathcal{S}'_k$ in the current time. Condition (6.12) indicates the conditional steady-state probability $\pi(j|\mathcal{S}_{k'}, \mathcal{S}_k)$ does not depend on the subset that the previous state belongs to. This means this conditional steady-state probability has a memoryless property.

e. Since

$$g(\mathcal{S}_k) = E\{f(X_l) - \eta + g(\mathcal{S}_{l+1})|X_l \in \mathcal{S}_k\},$$

we can use $f(X_l) - \eta + g(\mathcal{S}_{l+1})$ as an estimation of $g(\mathcal{S}_k)$. The $TD(0)$ algorithm can be developed as follows:

$$g(\mathcal{S}_l) := g(\mathcal{S}_l) + \kappa_l[f(X_l) - \eta + g(\mathcal{S}_{l+1}) - g(\mathcal{S}_l)], \quad if \quad X_l \in \mathcal{S}_l.$$

f. If the condition (6.12) does not hold, the Poisson equation will not hold. Thus, we cannot use $f(X_l) - \eta + g(\mathcal{S}_{l+1})$ as an estimation of $g(\mathcal{S}_k)$. The algorithm in e. cannot work.

6.18 In perturbation analysis of Markov chains, we have two Markov chains with transition probability matrices P and P', respectively. Let $\Delta P = P' - P$ and $P_\delta = P + \delta\Delta P$.

Let η_δ be the long-run average reward of the Markov chain with transition probability matrix P_δ. Assume that the reward function f_δ are the same as f for all $0 \le \delta \le 1$. Let π and g be the steady-state probability and performance potential of the Markov chain with transition probability P. Then the directional derivative of η_δ is (2.23):

$$\frac{d\eta_\delta}{d\delta} = \pi(\Delta P)g.$$

a. Write the performance derivative in the form of Q-factors.

b. Suppose that we do not know the values of P and P' and only know the corresponding actions. Develop a $TD(0)$ algorithm for the performance derivative.

Solution:

a. We can view P and P' as two transition probability matrix under two different policies v and v', respectively.

$$\frac{d\eta_\delta}{d\delta} = \pi[Q^{v'} - Q^v] = \sum_{i\in\mathcal{S}} \pi(i)[Q(i, v'(i)) - Q(i, v(i))]. \tag{6.13}$$

where $Q^{v'} = P'g + f$ and $Q^v = Pg + f$.

b. We consider the Markov chain with transition probability matrix P_δ. The policy corresponding to P_δ chooses action $v(i)$ with probability δ and chooses action $v'(i)$ with probability $1-\delta$ when the state is i. From (6.13), we have $\frac{d\eta_\delta}{d\delta} = E_\pi[Q(X_l, v'(X_l)) - Q(X_l, v(X_l))]$. Thus, we can firstly estimate Q-factor by using $TD(0)$ algorithm, then use the $TD(0)$ algorithm to estimate the expectation. We can develop the following algorithm:

$$Q(X_l, A_l) := Q(X_l, A_l) - \kappa_l\Big\{[f(X_l, A_l) - \eta] + Q(X_{l+1}, A_{l+1}) - Q(X_l, A_l)\Big\},$$
$$\frac{d\eta_\delta}{d\delta}|_{l+1} = \frac{d\eta_\delta}{d\delta}|_l + \kappa_l\Big\{Q(X_l, v'(X_l)) - Q(X_l, v(X_l)) - \frac{d\eta_\delta}{d\delta}|_l\Big\}.$$

6.19 Develop two $TD(0)$ algorithms, similar to (6.44) and (6.45), based on the performance derivative formula (3.44).

Solution: Since

$$\frac{d\eta_\delta}{d\delta} = \frac{E\{\sum_{k=u_m}^{u_{m+1}-1}\Big(\frac{\Delta p(X_{k+1}|X_k)}{p(X_{k+1}|X_k)}\hat{w}_{k+1}\Big)\}}{E[u_{m+1} - u_m]} \tag{6.14}$$

$$= E\Big(\frac{\Delta p(X_{k+1}|X_k)}{p(X_{k+1}|X_k)}\hat{w}_{k+1}\Big), \tag{6.15}$$

where

$$\hat{w}_{k+1} = \sum_{l=k+1}^{u_{m(k+1)+1}-1} [f(X_l) - \eta].$$

Based on (6.14), we can develop an algorithm to estimate the numerator of (6.14) similarly to (6.44),

$$\nu_{\delta,l+1} := \nu_{\delta,l} + \kappa_l \delta_l, \quad l = 0, 1, \cdots,$$
$$\delta_l = \sum_{k=u_l}^{u_{l+1}-1} \frac{\Delta p(X_{k+1}|X_k)}{p(X_{k+1}|X_k)} \hat{w}_{k+1} - \nu_{\delta,l}.$$

Similarly to (6.45), we can develop an algorithm based on (6.15):

$$\frac{d\eta_\delta}{d\delta}|_{l+1} := \frac{d\eta_\delta}{d\delta}|_l + \kappa_l \delta_l, \quad l = 0, 1, \cdots,$$
$$\delta_l = \frac{\Delta p(X_{l+1}|X_l)}{p(X_{l+1}|X_l)} \hat{w}_{l+1} - \frac{d\eta_\delta}{d\delta}|_l.$$

6.20 Suppose that algorithm (6.62) and (6.63) converge to optimal Q-factors. Are the following statements true? If so, please explain

a. With (6.62), when the algorithm converges we have $Q(i^*, \alpha^*) = \eta^*$, the optimal performance.

b. With (6.63), when algorithm converges we have $max_{\alpha \in \mathcal{A}(i^*)} Q(i^*, \alpha) = \eta^*$.

Solution:

a. Under some standard stochastic approximation conditions, the algorithm (6.62) converges to the solution of the following equation:

$$Q^{\hat{d}}(i, \alpha) = \{\sum_{j=1}^{S} p^{\alpha}(j|i)[max_{\beta \in \mathcal{A}(j)} Q^{\hat{d}}(j, \beta)]\} - Q(i^*, \alpha^*) + f(i, \alpha), \quad \alpha \in \mathcal{A}(i), i \in \mathcal{S} (6.16)$$

where $\hat{d}$ is the optimal policy. We have $\hat{d}(i) = argmax_{\alpha \in \mathcal{A}(i)} Q^{\hat{d}}(i, \alpha)$. Taking the maximum among the actions in $\mathcal{A}(i)$ on both sides of (6.16), we have

$$max_{\alpha \in \mathcal{A}(i)} Q^{\hat{d}}(i, \alpha) = Q^{\hat{d}}(i, \hat{d}(i)) = \{\sum_{j=1}^{S} p^{\hat{d}(i)}(j|i)[max_{\beta \in \mathcal{A}(j)} Q^{\hat{d}}(j, \beta)]\} - Q(i^*, \alpha^*) + f(i, \hat{d}(i)),$$

for all $i \in \mathcal{S}$. Pre-multiplying $\pi^{\hat{d}}(i)$ on both sides of the above equation and summing them over all $i \in \mathcal{S}$, we have $Q(i^*, \alpha^*) = \sum_{i \in \mathcal{S}} \pi^{\hat{d}}(i) f(i, \hat{d}(i)) = \eta^*$.

b. Under some standard stochastic approximation conditions, the algorithm (6.63) converges to the solution of the following equation:

$$Q^{\hat{d}}(i, \alpha) = \{\sum_{j=1}^{S} p^{\alpha}(j|i)[max_{\beta \in \mathcal{A}(j)} Q^{\hat{d}}(j, \beta)]\} - \max_{\alpha \in \mathcal{A}(i^*)} Q^(i^*, \alpha) + f(i, \alpha), \quad \alpha \in \mathcal{A}(i), i \in \mathcal{S}.$$

Similarly to the discussion in part a), we have $\max_{\alpha \in \mathcal{A}(i^*)} Q^(i^*, \alpha) = \sum_{i \in \mathcal{S}} \pi^{\hat{d}}(i) f(i, \hat{d}(i)) = \eta^*$

6.21 In this problem, we derive a performance derivative formula for closed Jackson networks in the form of sample path expectation. Consider a closed Jackson network consisting of M servers and N customers. The service times of server i are exponentially distributed with mean $\bar{s}_i = 1/\mu_i, i = 1, 2, \cdots, M$. The state of the system is $\boldsymbol{n} = (n_1, \cdots, n_M)$, n_i is the number of customers in server i, $\sum_{i=1}^{M} n_i = N$. Suppose that the system is in the steady state, and let $\pi(\boldsymbol{n})$ be the steady-state probability of state $\boldsymbol{n}$. Denote $\mu(\boldsymbol{n}) = \sum_{i=1}^{M} \epsilon(n_i)\mu_i$, with $\epsilon(n) = 1$ if $n > 0$ and 0 if $n = 0$. The system throughput is $\eta = \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n})\mu(\boldsymbol{n})$, its derivative with respect to $\bar{s}_v$, $v = 1, 2, \ldots, N$, is (2.109):

$$\frac{\bar{s}_v}{\eta} \frac{\partial \eta}{\partial \bar{s}_v} = - \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) c(\boldsymbol{n}, v),$$

where $c(\boldsymbol{n}, v)$ is the realization probability of a perturbation of server v when the system is in state $\boldsymbol{n}$, and "E_π" represents the steady-state mean.

a. Consider a sample path of the system. Denote the sequence of transition times as $T_0, T_1, \cdots, T_l, \cdots$. Suppose that the system is in state $\boldsymbol{n}$ in $[T_l, T_{l+1})$; i.e., $X_l = \boldsymbol{n}$. Assume that in this period server v obtain a (infinitesimal) perturbation. We define a *perturbation realization index* for this perturbation as follows:

$$RI(l, X_l, v) = \begin{cases} 1 & \textit{if the perturbation is realized on the sample path,} \\ 0 & \textit{otherwise;} \end{cases}$$

and set $\varsigma(t) = RI(l, X_l, v)$ for $t \in [T_l, T_{l+1})$. Then by definition, we have

$$E[RI(l, X_l, v)|X_l = \boldsymbol{n}] = c(\boldsymbol{n}, v),$$

where "E" denotes the expectation with respect to the probability space generated by all the sample paths. Explain the following equation:

$$\frac{\bar{s}_v}{\eta}\frac{\partial \eta}{\partial \bar{s}_v} = -E[RI(l, X_l, v)] = -E[\varsigma(t)] := -\lim_{L\to\infty}\frac{1}{T_L}\int_0^{T_L}\varsigma(t)dt, a.s. \qquad (6.17)$$

b. Can you determine the function $\varsigma(t)$ based on the sample path in Figure 2.18 of Chapter 2 (Note: $\varsigma(t)$ depends not only on the current state $\boldsymbol{n}$, but also on the future behavior of the system.)

c. Derive a sample-path based estimate of the performance derivative by using the above result.

d. Apply this equation (6.17) to a two-server closed Jackson network and verify the results. Can this be extended to networks with non-exponentially distributed service times?

e. Derive a recursive algorithm (c.f. Problem 6.5).

f. Discuss and compare your results with other algorithms.

Solution:

a.

$$\begin{aligned} & E[RI(l, X_l, v)] = E\{E[RI(l, X_l, v)|X_l]\} \\ = & \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) E[RI(l, X_l, v)|X_l = \boldsymbol{n}] \\ = & \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) c(\boldsymbol{n}, v). \end{aligned}$$

Thus, we have

$$\frac{\bar{s}_v}{\eta}\frac{\partial \eta}{\partial \bar{s}_v} = -E[RI(l, X_l, v)].$$

Next, we prove that

$$\lim_{L\to\infty}\frac{1}{T_L}\int_0^{T_L}\varsigma(t)dt = \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n}) c(\boldsymbol{n}, v), \quad a.s.$$

Since $\lim_{L\to\infty}\frac{1}{T_L}\int_0^{T_L}\varsigma(t)dt = \lim_{L\to\infty}\frac{1}{T_L}\sum_{l=0}^{L-1}S_lRI(l,X_l,v)$, where $S_l, l=0,1,\cdots$, denote the durations the system stays state X_l, thus we only need to prove

$$\lim_{L\to\infty}\frac{1}{T_L}\sum_{l=0}^{L-1}S_lRI(l,X_l,v) = \sum_{all\ \boldsymbol{n}}\pi(\boldsymbol{n})c(\boldsymbol{n},v). \tag{6.18}$$

We can find whether a perturbation is realized or lost depends on the initial state X_l and the customer transitions afterwards. It does not depend on the durations the system stays in all the states $S_l, S_{l+1},\cdots$. Therefore, $RI(l,X_l,v)$ is function of $X_l, X_{l+1},\cdots$.

Let $\chi_{\boldsymbol{n}}(X_l)=1$ if $X_l=\boldsymbol{n}$; $\chi_{\boldsymbol{n}}(X_l)=0$, otherwise. For all l with $\chi_{\boldsymbol{n}}(X_l)=1$, S_l are independent and identically distributed. By the law of large number, we have

$$\lim_{l\to\infty}\frac{\sum_{l=0}^{L-1}S_lRI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)}{\sum_{l=0}^{L-1}RI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)} = E\{S_l|X_l=\boldsymbol{n},R(l,X_l,v)=1\}.$$

From this, we have

$$\begin{aligned}
&\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}S_lRI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)\\
=\ & E\{S_l|X_l=\boldsymbol{n},RI(l,X_l,v)=1\}\times\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}RI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)\\
=\ & E\{S_l|X_l=\boldsymbol{n},RI(l,X_l,v)=1\}\times\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}Y_l,
\end{aligned}$$

where

$$Y_l = RI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l) = \phi(X_l,X_{l+1},\cdots)\chi_{\boldsymbol{n}}(X_l) := \psi(X_l,X_{l+1},\cdots).$$

From the fundamental ergodicity theorem in Chapter 3, we have

$$\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}Y_l = E(Y_l) = E[RI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)] = \pi(RI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)=1).$$

where $\pi(RI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)=1)$ is the steady-state probability that $RI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)=1$ for Markov chain $Y=\{Y_0,Y_1,\cdots\}$. Therefore,

$$\begin{aligned}
&\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}S_lRI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)\\
=\ & E\{S_l|X_l=\boldsymbol{n},RI(l,X_l,v)=1\}\times\pi(RI(l,X_l,v)\chi_{\boldsymbol{n}}(X_l)=1)\\
=\ & E\{S_l\chi_{\boldsymbol{n}}(X_l)RI(l,X_l,v)\}\\
=\ & E\{S_lRI(l,X_l,v)|X_l=\boldsymbol{n}\}\times\hat{\pi}(\boldsymbol{n}),
\end{aligned}$$

where $\hat{\pi}(\boldsymbol{n})$ is the stationary probability of state $\boldsymbol{n}$ of the embedded chain $X = \{X_0, X_1, \cdots$. By using the Markov property, given $X_l = \boldsymbol{n}$, the two random variables S_l and $RI(l, X_l, v)$ are independent. Thus, we have

$$\begin{aligned} & E\{S_l RI(l, X_l, v)|X_l = n\} \\ = \; & E(S_l|X_l = \boldsymbol{n}) \times E\{RI(l, X_l, v)|X_l = \boldsymbol{n}\} \\ = \; & \frac{c(\boldsymbol{n}, v)}{\mu(\boldsymbol{n})}. \end{aligned}$$

Finally, we have

$$\lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} S_l RI(l, X_l, v)\chi_{\boldsymbol{n}}(X_l) = c(\boldsymbol{n}, v)\frac{\hat{\pi}(\boldsymbol{n})}{\mu(\boldsymbol{n})}, \quad w.p.1.$$

According the relationship between the steady-state probabilities, $\hat{\pi}$ and π, of embedded chain and continuous time process, i.e.

$$\pi(\boldsymbol{n}) = \eta\frac{\hat{\pi}(\boldsymbol{n})}{\mu(\boldsymbol{n})},$$

we obtain

$$\lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} S_l RI(l, X_l, v)\chi_{\boldsymbol{n}}(X_l) = \frac{1}{\eta}\pi(\boldsymbol{n})c(\boldsymbol{n}, v), \quad w.p.1.$$

Summing up both sides of the above equation over all $\boldsymbol{n}$ and noting that $\sum_{all\ \boldsymbol{n}} \chi_{\boldsymbol{n}}(X_l) = 1$, we get

$$\lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} S_l RI(l, X_l, v) = \frac{1}{\eta} \sum_{all\ \boldsymbol{n}} \pi(\boldsymbol{n})c(\boldsymbol{n}, v), \quad w.p.1.$$

On the other hand, we can similarly prove

$$\begin{aligned} & \lim_{L\to\infty} \frac{T_L}{L} = \lim_{L\to\infty} \sum_{l=0}^{L-1} S_l \\ = \; & \lim_{L\to\infty} \frac{1}{L} \sum_{all\ \boldsymbol{n}} \sum_{l=0}^{L-1} S_l \chi_{\boldsymbol{n}}(X_l) \\ = \; & \sum_{all\ \boldsymbol{n}} \{\lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} S_l \chi_{\boldsymbol{n}}(X_l) \\ = \; & \sum_{all\ \boldsymbol{n}} E[\chi_{\boldsymbol{n}}(X_l)] \times E(S_l|X_l = \boldsymbol{n}) \quad w.p.1 \\ = \; & \sum_{all\ \boldsymbol{n}} \frac{\pi(\boldsymbol{n})}{\mu(\boldsymbol{n})} = \frac{1}{\eta} \quad w.p.1. \end{aligned}$$

Noting that

$$\lim_{L\to\infty}\frac{1}{T_L}\sum_{l=0}^{L-1}S_lRI(l,X_l,v)=\lim_{L\to\infty}\frac{L}{T_L}\times\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}S_lRI(l,X_l,v)=\sum_{all\ \boldsymbol{n}}\pi(\boldsymbol{n})c(\boldsymbol{n},v).$$

Thus, we have proved (6.18).

Reference:

Xi-Ren Cao, Realization Probability: The Dynamics of Queueing systems, Springer-Verlag, London, 1994.

b.

$$\varsigma(t)=\begin{cases}0, & T_0\le t<T_1\\ 0 & T_1\le t<T_2\\ 0 & T_2\le t<T_3\\ 1 & T_3\le t<T_4\\ ? & T_4\le t<T_5,\\ & \cdots\end{cases}$$

c. and e. Denote

$$\frac{\bar{s}_v}{\eta}\frac{\partial\eta}{\partial\bar{s}_v}|_{L+1}=-\frac{1}{T_{L+1}}\int_0^{T_{L+1}}\varsigma(t)dt,$$

Then, we have $\lim_{L\to\infty}\frac{\bar{s}_v}{\eta}\frac{\partial\eta}{\partial\bar{s}_v}|_{L+1}=\frac{\bar{s}_v}{\eta}\frac{\partial\eta}{\partial\bar{s}_v}$, $w.p.1$. For $\frac{\bar{s}_v}{\eta}\frac{\partial\eta}{\partial\bar{s}_v}|_{L+1}$, we have the following recursive algorithm.

$$\begin{aligned}\frac{\bar{s}_v}{\eta}\frac{\partial\eta}{\partial\bar{s}_v}|_{L+1} &= -\frac{1}{T_{L+1}}\{\int_0^{T_L}\varsigma(t)dt+(T_{L+1}-T_L)R(L,X_L,v)\}\\ &= -\frac{\bar{s}_v}{\eta}\frac{\partial\eta}{\partial\bar{s}_v}|_L-\frac{T_{L+1}-T_L}{T_{L+1}}\Big[R(L,X_L,v)-\frac{\bar{s}_v}{\eta}\frac{\partial\eta}{\partial\bar{s}_v}|_L\Big]\end{aligned}$$

f. Since this algorithm uses the function $\varsigma(t)$, which depends on the future information on the sample path, it is difficult to use this algorithm in an on-line way.

7

Solutions to Chapter 7

7.1 Repeat Example 7.1 with the data listed in Table 7.3.

state		Transition prob. $p^{\alpha}(j\|i)$			
i	action	$j=1$	2	3	Perf. func.
1	$\alpha_{1,1}$	0.3	0.6	0.1	10
	$\alpha_{1,2}$	0.4	0.2	0.4	
	$\alpha_{1,3}$	0.2	0.3	0.5	
2	$\alpha_{2,1}$	0.6	0	0.4	0
	$\alpha_{2,2}$	0.4	0.3	0.3	
3	$\alpha_{3,1}$	0.4	0.2	0.4	-5
	$\alpha_{3,2}$	0.3	0.5	0.2	
	$\alpha_{3,3}$	0.2	0.1	0.7	

Table 7.1: The Actions and Performance Function in Problem 7.1

Solution: We first solve the finite-step optimization problem by using (7.12). The values of $\eta^*_{\ell=L}(i), \hat{d}_{\ell=L}(i)$, and $g^*_{\ell=L}(i)$ in (7.11), (7.12), and (7.15) for $L = 1, 2, 3, 4$, are listed in Table 7.2. As shown in the table, the optimal decision-rule sequence $d_{\ell=L}$ converges to $\hat{d} = \{\hat{d}(1) = \alpha_{1,1}, \hat{d}(2) = \alpha_{2,1}, \hat{d}(3) = \alpha_{3,2}\}$.

L	1			2		
state i	$\eta^*_{\ell=1}(i)$	$g^*_{\ell=1}(i)$	$\hat{d}_{\ell=1}(i)$	$\eta^*_{\ell=2}(i)$	$g^*_{\ell=2}(i)$	$\hat{d}_{\ell=2}(i)$
1	10	15	$\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3}$	12.5	15.5	$\alpha_{1,1}$
2	0	5	$\alpha_{2,1}, \alpha_{2,2}$	4	7	$\alpha_{2,1}$
3	-5	0	$\alpha_{3,1}, \alpha_{3,2}, \alpha_{3,3}$	-3	0	$\alpha_{3,1}, \alpha_{3,2}$

L	3			4		
state i	$\eta^*_{\ell=3}(i)$	$g^*_{\ell=3}(i)$	$\hat{d}_{\ell=3}(i)$	$\eta^*_{\ell=4}(i)$	$g^*_{\ell=4}(i)$	$\hat{d}_{\ell=4}(i)$
1	15.85	15.7	$\alpha_{1,1}$	18.55	15.615	$\alpha_{1,1}$
2	6.3	6.15	$\alpha_{2,1}$	9.57	6.635	$\alpha_{2,1}$
3	0.15	0	$\alpha_{3,2}$	2.935	0	$\alpha_{3,2}$

Table 7.2: The Results for the L-step Optimization Problems

Now, let us solve the problem by policy iteration. Firstly, we choose an initial stationary policy d_0 and then determine the potentials under this policy. Suppose that we pick up $d_0 = \{\alpha_{1,1}, \alpha_{2,1}, \alpha_{3,1}\}$. The transition probability matrix is

$$P^{d_0} = \begin{bmatrix} 0.3 & 0.6 & 0.1 \\ 0.6 & 0 & 0.4 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}.$$

We use the approximation

$$\bar{g}_n = \sum_{k=0}^{n} P_-^k f_-,$$

where $P_- = P - ep_{S*}$, p_{S*} is the Sth row of P, and $f_-(i) = f(i) - f(S), i = 1, 2, \cdots, S$, to approximate the potentials. From P^{d_0}, we have

$$P_-^{d_0} = \begin{bmatrix} -0.1 & 0.4 & -0.3 \\ 0.2 & -0.2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The values for $\bar{g}_0$ to $\bar{g}_5$ are:

i	$\bar{g}_0(i)$	$\bar{g}_1(i)$	$\bar{g}_2(i)$	$\bar{g}_3(i)$	$\bar{g}_4(i)$	$\bar{g}_5(i)$
1	15	15.5	16.25	16.055	16.1585	16.1157
2	5	7	6.7	6.91	6.829	6.8659
3	0	0	0	0	0	0

By using the policy improvement

$$d_1(i) \in \arg\max_{\alpha}\{f(i,\alpha) + p^{\alpha}(j|i)\bar{g}_5(j)\}, \quad i \in \mathcal{S},$$

we can obtain $d_1 = \{\alpha_{1,1}, \alpha_{2,1}, \alpha_{3,2}\}$. We have

$$P^{d_1} = \begin{bmatrix} 0.3 & 0.6 & 0.1 \\ 0.6 & 0 & 0.4 \\ 0.3 & 0.5 & 0.2 \end{bmatrix},$$

and therefore

$$P_{-}^{d_1} = \begin{bmatrix} 0 & 0.1 & -0.1 \\ 0.3 & -0.5 & 0.2 \\ 0 & 0 & 0 \end{bmatrix}.$$

The values for $\bar{g}_0$ to $\bar{g}_5$ are:

i	$\bar{g}_0(i)$	$\bar{g}_1(i)$	$\bar{g}_2(i)$	$\bar{g}_3(i)$	$\bar{g}_4(i)$	$\bar{g}_5(i)$
1	15	15.5	15.7	15.615	15.6635	15.6367
2	5	7	6.15	6.635	6.367	6.5155
3	0	0	0	0	0	0

Applying the policy improvement, we obtain the same policy $d_2 = d_1 = \{\alpha_{1,1}, \alpha_{2,1}, \alpha_{3,2}\}$. Thus, this policy is optimal.

7.2 For any three operators P_1, P_2, and P_3, prove

a. For any function $h(x)$ on $\Re^n$, we have $(P_1P_2)h(x) = P_1[P_2h(x)]$(assuming the integrations exist); and

b. $(P_1P_2)P_3 = P_1(P_2P_3)$; and

c. $P^k = PP^{k-1} = P^{k-1}P$.

Solution:

a. For any $x \in \Re^n$, we have

$$(P_1P_2)h(x) = \int_{\Re^n} h(z)P_1P_2(dz|x) = \int_{\Re^n}\int_{\Re^n} h(z)P_2(dz|y)P_1(dy|x),$$

$$P_1[P_2h(x)] = \int_{\Re^n} P_2h(y)P_1(dy|x) = \int_{\Re^n}\int_{\Re^n} h(z)P_2(dz|y)P_1(dy|x).$$

Thus, we have $(P_1P_2)h(x) = P_1[P_2h(x)]$ for any function $h(x)$ on $\Re^n$.

b. For any $x \in \Re^n$ and $R \subset \mathcal{B}$, we have

$$\begin{aligned}(P_1P_2)P_3(R|x) &= \int_{\Re^n} P_3(R|y)(P_1P_2)(\mathrm{d}y|x) \\ &= \int_{\Re^n}\int_{\Re^n} P_3(R|y)P_2(\mathrm{d}y|z)P_1(\mathrm{d}z|x),\end{aligned}$$

and

$$\begin{aligned}P_1(P_2P_3)(R|x) &= \int_{\Re^n} (P_2P_3)(R|z)P_1(\mathrm{d}z|x) \\ &= \int_{\Re^n}\int_{\Re^n} P_3(R|y)P_2(\mathrm{d}y|z)P_1(\mathrm{d}z|x).\end{aligned}$$

Thus, $(P_1P_2)P_3 = P_1(P_2P_3)$.

c. From the definition of kth power of P, we have $P^k = PP^{k-1}$. Thus

$$P^{k-1}P = PP^{k-2}P = \cdots = \underbrace{PP\cdots P}_{k}$$

and

$$P^k = PP^{k-1} = PPP^{k-2} = \cdots = \underbrace{PP\cdots P}_{k}$$

So we have $P^k = P^{k-1}P$.

7.3 For any probability distribution ν, transition function P, and any function h, prove $\nu(Ph) = (\nu P)h$. Explain the meaning of both sides.

Solution:

$$\nu(Ph) = \int_{\mathcal{R}^n} \left\{ \int_{\mathcal{R}^n} h(y)P(\mathrm{d}y|x) \right\} \nu(\mathrm{d}x)$$

and

$$(\nu P)h = \int_{\mathcal{R}^n} h(y)(\nu P)(\mathrm{d}y) = \int_{\mathcal{R}^n} h(y) \int_{\mathcal{R}^n} P(\mathrm{d}y|x)\nu(\mathrm{d}x).$$

So we have $\nu(Ph) = (\nu P)h$.

$Ph(x)$ is the expected performance at the next time epoch when current state is x, which can be written as $\mathrm{E}\{h(x_1)|x_0 = x\}$. For any probability measure ν,

$$\nu(Ph) = \int_{\mathcal{R}^n} \nu(\mathrm{d}x)\mathrm{E}\{h(x_1)|x_0 = x\} =: \mathrm{E}\{h(x_1)|\nu\}.$$

It is the expected performance at the next time epoch when current state distribution is ν. It is the physical meaning of left side.

νP is state distribution at the next time epoch when current state distribution is ν. Let $\nu_1 = \nu P$. Therefore $\nu_1 h = \int_{\mathcal{R}^n} h(x)\nu_1(\mathrm{d}x)$ represent the expected performance at the next time epoch when current state distribution is ν. This is the physical meaning of right side. From their physical meaning, we can also obtain they are equal.

7.4 With the forward-time index used in (7.9), from (7.14) and (7.15), we can define the finite-step perturbation realization factor for any policy $\boldsymbol{d} = \{d_0, d_1, \cdots, d_{L-1}\}$ as follows:

$$g_{\ell=L}^{\boldsymbol{d}}(i) = \sum_{l=0}^{L-1} E\Big\{[f(X_l, d_l(X_l)) - f(X'_l, d_l(X'_l))]|X_0 = i, X'_0 = i^*\Big\},$$

where $\boldsymbol{X} = \{X_0, X_1, \cdots\}$ and $\boldsymbol{X'} = \{X'_0, X'_1, \cdots\}$ are two independent sample paths with initial state $X_0 = i$ and $X'_0 = i^*$, respectively. Note that the decision rules d_l may be different for different $l = 0, 1, \cdots$. Let L_{ii^*} be the time at which the two sample paths merge together, i.e. $X_{L_{ii^*}} = X'_{L_{ii^*}}$.

a. Prove that if $E(L_{ii^*}) < \infty$, then $\lim_{L\to\infty} g_{\ell=L}^{\boldsymbol{d}}(i)$ exists.

b. Find a condition under which $E(L_{ii^*}) < \infty$.

Solution:

a.

$$\begin{aligned} g_{\ell=L}^{\boldsymbol{d}}(i) &= E\Big\{\sum_{l=0}^{L-1}[f(X_l, d_l(X_l)) - f(X'_l, d_l(X'_l))]|X_0 = i, X'_0 = i^*\Big\} \\ &= E\Big\{\sum_{l=0}^{L_{ii^*}-1}[f(X_l, d_l(X_l)) - f(X'_l, d_l(X'_l))]\end{aligned}$$

$$+\sum_{l=L_{ii^*}}^{L-1}[f(X_l,d_l(X_l))-f(X'_l,d_l(X'_l))]|X_0=i,X'_0=i^*\Big\}.$$

By the strong Markov property, two Markov chains $\boldsymbol{X}$ and $\boldsymbol{X'}$ behave similarly statistically after L_{ii^*}. Thus, $\lim_{L\to\infty}E\Big\{\sum_{l=L_{ii^*}}^{L-1}[f(X_l,d_l(X_l))-f(X'_l,d_l(X'_l))]|X_0=i,X'_0=i^*\Big\}=0$. Therefore,

$$\lim_{L\to\infty}g^{\boldsymbol{d}}_{\ell=L}(i)=E\Big\{\sum_{l=0}^{L_{ii^*}-1}[f(X_l,d_l(X_l))-f(X'_l,d_l(X'_l))]|X_0=i,X'_0=i^*\Big\}.$$

Since performance function f is bounded and $E(L_{ii^*})<\infty$, $E\Big\{\sum_{l=0}^{L_{ii^*}-1}[f(X_l,d_l(X_l))-f(X'_l,d_l(X'_l))]|X_0=i,X'_0=i^*\Big\}$ is finite and $\lim_{L\to\infty}g^{\boldsymbol{d}}_{\ell=L}(i)$ exists.

b. If the Markov chain under policy d has a absorbing state, then $E(L_{ii^*})<\infty$.

7.5 Prove Lemma 7.1.

Solution:

Putting (7.32) into the left side of (7.29), we have

$$\begin{aligned}
&\{I+\sum_{k=1}^{\infty}(P^k-e\pi)\}f(x)-P\{I+\sum_{k=1}^{\infty}(P^k-e\pi)\}f(x)+\eta(x)\\
=\;& f(x)+(Pf)(x)-[(e\pi)f](x)+\{\sum_{k=2}^{\infty}(P^k-e\pi)\}f(x)-(Pf)(x)\\
&-P\{\sum_{k=1}^{\infty}(P^k-e\pi)\}f(x)+\eta(x)\\
=\;& f(x)-(\pi f)e(x)+\eta(x)\\
=\;& f(x).
\end{aligned}$$

7.6 For any bounded function $f(x), x\in\Re$, we define the e-norm $\|f(x)\|=\sup_x|f(x)|$. The e-norm of a linear operation $P(R|x)$ is defined as $\|P\|:=\sup\{\|Pu\|:\|u\|\le 1\}$. A transition probability matrix P is called e-ergodic, if

$$\lim_{k\to\infty}\|(P^k-e\pi)\|=0.$$

Prove if P is e-ergodic, then

$$\lim_{k\to\infty}g_k=g,\quad and\quad \lim_{k\to\infty}Pg_k=Pg,$$

where $g_k := \{I + \sum_{l=1}^{k}(P^l - e\pi)\}f$, for any bounded function f.

Solution:

From the assumption that $\lim_{k\to\infty} \|(P^k - e\pi)\| = 0$, there is an integer $K > 0$ and an $\epsilon, 0 < \epsilon < 1$ such that $\|(P^K - e\pi)f\| < \epsilon\|f\|$. Then $\|(P^{2K} - e\pi)f\| = \|(P^K - e\pi)[(P^K - e\pi)f]\| < \epsilon\|(P^K - e\pi)f\| < \epsilon^2\|f\|, \cdots$ and $\|P^{nK} - e\pi\| < \epsilon^n\|f\|$. Since $P^k - e\pi = (P - e\pi)^k$, we have

$$\begin{aligned} \|\sum_{l=nK}^{(n+1)K-1}(P^l - e\pi)f\| &= \|\{\sum_{l=0}^{K-1}(P - e\pi)^l\}(P - e\pi)^{nK}f\| \\ \leq \sum_{l=0}^{K-1}\|(P - e\pi)^l(P - e\pi)^{nK}f\| &\leq \sum_{l=0}^{K-1}\|(P - e\pi)^l\|\|(P - e\pi)^{nK}f\| \\ < \epsilon^n\sum_{l=0}^{K-1}\|(P - e\pi)^l\|\|f\| &= \epsilon^n G, \end{aligned}$$

where $G = \sum_{l=0}^{K-1}\|(P - e\pi)^l\|\|f\|$, which is bounded because f is bounded, and

$$\|\sum_{n=0}^{\infty}\sum_{l=nK}^{(n+1)K-1}(P - e\pi)^l f\| \leq \sum_{n=0}^{\infty}\|\sum_{l=nK}^{(n+1)K-1}(P - e\pi)^l f\| < \sum_{n=0}^{\infty}\epsilon^n G = \frac{G}{1-\epsilon}.$$

Therefore,

$$\begin{aligned} \lim_{k\to\infty} g_k &= I + \sum_{l=1}^{\infty}(P^l - e\pi)f \\ &= \sum_{n=0}^{\infty}\sum_{l=nK}^{(n+1)K-1}(P - e\pi)^l f =: g, \end{aligned}$$

exists. Since $\|P(g_k - g)\| \leq \|P\|\|g_k - g\|$ and $\|P\|$ is bounded, we have $\lim_{k\to\infty} Pg_k = Pg$.

7.7 Consider the two steady-state probability distributions π and π' defined as shown in Figure 7.7. The two distributions have discrete masses as follows: $\pi(-0.2) = \pi(-0.4) = \pi(-0.6) = \pi(-0.8) = \pi(-1) = 0.1$, and $\pi'(0.2) = \pi'(0.4) = \pi'(0.6) = \pi'(0.8) = \pi'(1) = 0.1$. The total probabilities on these discrete points are $\frac{1}{2}$ for both distributions. The other $1/2$ is evenly distributed on the interval $[-1, 1]$. Explain that these two distribution functions have the same state space, but they do not have the same support.

Solution: It is clear that these two distribution functions have the same state space. Since $\pi(-0.2) > 0, \pi(-0.4) > 0, \pi(-0.6) > 0, \pi(-0.8) > 0, \pi(-1) > 0$ and $\pi(R) > 0$ for

any $R \subseteq [0,1]$ with a positive volume and $\pi'(0.2) > 0, \pi'(0.4) > 0, \pi'(0.6) > 0, \pi'(0.8) > 0, \pi'(1) > 0$ and $\pi'(R') > 0$ for any $R' \subseteq [-1,0]$ with a positive volume, the supports of π and π' are different.

7.8 Consider a non-linear control system

$$X_{l+1} = uX_l + \xi_l, \qquad l = 0, 1, \cdots,$$

where u is a control variable. Let $p_\xi(\ldots)$ be the distribution density function of the independent and identically distributed random noises $\xi_l, l = 0, 1, \ldots$.

a. Derive the transition probability function $P^u(dy|x)$.

b. How do we estimate the discrete approximation $p(j|i), i, j = 1, 2, \cdots, S$? Can we reduce the number of the transition probabilities to be estimated?

Solution: a. $P^u(\mathrm{d}y|x) = p_\xi(y - ux)dy$.

b. $p(j|i)$ can be estimated with equation (7.69). Indeed, we needn't estimate $S \times S$ transition probabilities. Assuming that $\Delta x_i = \Delta$, for any i, is very small, we have

$$p(j|i) \approx p_\xi[y - ux]\Delta, \quad y \in \Delta_j, x \in \Delta_i. \tag{7.1}$$

For simplicity, we consider a one-dimensional system. We divide $\Re$ with the points $k\Delta$, $k = -(S-1), \cdots, -1, 0, 1, \cdots, S-1$. There are $2S$ states corresponding to intervals $\Delta_1 = (-\infty, -(S-1)\Delta]$, $\Delta_2 = (-(S-1)\Delta, -(S-2)\Delta]$,$\cdots$, $\Delta_{2S-1} = ((S-2)\Delta, (S-1)\Delta]$ and $\Delta_{2S} = ((S-1)\Delta, \infty)$. We assume that the probability that the random noise ξ in Δ_1 and Δ_{2S} is very small. From (7.1), if we know $p_\xi(y)$, we can calculate the transition probability matrix $p(j|i)$. This means we can convert a problem of estimating a two-dimensional matrix P to a problem of estimating a one-dimensional vector $p_\xi(y), y \in \mathcal{S}$. As an example, if we take action $u = 0$, $P(j|i) \approx p_\xi(y)\Delta, y \in \Delta_j, x \in \Delta_i$. Thus, $P^0(1|i) = p_\xi[-(S-1)\Delta]\Delta, P^0(2|i) = p_\xi[-(S-2)\Delta]\Delta, \cdots, P^0(2S-1|i) = p_\xi[(S-2)\Delta]\Delta, P^0(2S|i) = p_\xi[(S-1)\Delta]\Delta, i \in \mathcal{S}$. Therefore, for any $i = 2, 3, \cdots, 2S-1$, $p^u(j|i)$, which can be estimated by using (7.69), correspond to the probabilities of the random noise ξ in the intervals $(-(S-1)\Delta, -(S-2)\Delta], \cdots, ((S-2)\Delta, (S-1)\Delta]$. That is, we need only to

estimate $2(S-1)$, rather than $2(S-1) \times 2(S-1)$, values. Thus, we can reduce the number of the transition probabilities to be estimated. For other control laws, we can use the similar method to reduce the number of the transition probabilities to be estimated.

7.9 Consider a JLQ problem.

a. Suppose that the modes changes slowly. That is, $p(i|i) \approx 1$ and $p(i|i) \approx 0$ for $j \neq i$. Show that the coupled Riccati equation is decoupled into M Riccati equations corresponding to M LQ problems.

b. We consider another extreme case: the mode changes rapidly. As an example, we consider a 2-mode system ($M = 2$). Suppose $p(2|1) = p(1|2) \approx 1$ and $p(1|1) = p(2|2) \approx 0$. What are the coupled Riccati equation in this case? Explain your results.

Solution:

a. If $p(i|i) \approx 1$ and $p(j|i) \approx 0$ for $j \neq i$, then $\hat{H}_i = \sum_{j\in\mathcal{M}} p(j|i)\hat{S}_j \approx \hat{S}_i$. Coupled Riccati equation becomes

$$\hat{S}_i = A_i^{\mathrm{T}}\hat{S}_i A_i + A_i^{\mathrm{T}}\hat{S}_i B_i(V_i + B_i^{\mathrm{T}}\hat{S}_i B_i)^{-1}B_i^{\mathrm{T}}\hat{S}_i A_i + Q_i \qquad i = 1, 2, \cdots, M$$

These are M Riccati equations.

b. Suppose $p(2|1) = p(1|2) \approx 1$ and $p(1|1) = p(2|2) \approx 0$, the Coupled Riccati equations are

$$\hat{S}_1 = A_1^{\mathrm{T}}\hat{S}_2 A_1 - A_1^{\mathrm{T}}\hat{S}_2 B_1(V_1 + B_1^{\mathrm{T}}\hat{S}_2 B_1)^{-1}B_1^{\mathrm{T}}\hat{S}_2 A_1 + Q_1, \tag{7.2}$$

$$\hat{S}_2 = A_2^{\mathrm{T}}\hat{S}_1 A_2 - A_1^{\mathrm{T}}\hat{S}_1 B_2(V_2 + B_2^{\mathrm{T}}\hat{S}_1 B_2)^{-1}B_2^{\mathrm{T}}\hat{S}_1 A_2 + Q_2. \tag{7.3}$$

Substituting (7.2) into (7.3), we can obtain the Riccati equation of the combined system

$$X_{k+1} = A_2 A_1 X_k + A_2 B_1 u_1 + B_2 u_2 + \varepsilon_k,$$

where $\varepsilon_k = A_2\xi_k + \zeta_k$, ξ_k and ζ_k are i.i.d.

7.10 Prove that in (7.48),

$$\sum_{k=1}^{\infty}(c_k - \eta) = -\int_{\Re^n}[z^T U z]p_\xi(z)dz,$$

with $U = \sum_{k=1}^{\infty} kW_k$. Prove

$$U - C^T UC = C^T SC.$$

Solution:

$$c_k = \int_{\mathcal{R}^n} z^{\mathrm{T}} \sum_{n=0}^{k-1} W_n z p_\xi(z) \mathrm{d}z$$

and

$$\eta = \lim_{k\to\infty} c_k = \int_{\mathcal{R}^n} z^{\mathrm{T}} \sum_{n=0}^{\infty} W_n z p_\xi(z) \mathrm{d}z$$

Therefore

$$c_k - \eta = -\int_{\mathcal{R}^n} z^{\mathrm{T}} \sum_{n=k}^{\infty} W_n z p_\xi(z) \mathrm{d}z$$

$$\begin{aligned}
\sum_{k=1}^{\infty}(c_k - \eta) &= -\sum_{k=1}^{\infty} \int_{\mathcal{R}^n} z^{\mathrm{T}} \sum_{n=k}^{\infty} W_n z p_\xi(z) \mathrm{d}z \\
&= -\int_{\mathcal{R}^n} z^{\mathrm{T}} \sum_{k=1}^{\infty} kW_k z p_\xi(z) \mathrm{d}z \\
&= -\int_{\mathcal{R}^n} z^{\mathrm{T}} U z p_\xi(z) \mathrm{d}z
\end{aligned}$$

The first equation is proved.

$$\begin{aligned}
U - C^{\mathrm{T}} UC &= \sum_{k=1}^{\infty} kW_k - C^{\mathrm{T}} \sum_{k=1}^{\infty} kW_k C \\
&= \sum_{k=1}^{\infty} kC^{\mathrm{T}} W_{k-1} C - \sum_{k=1}^{\infty} C^{\mathrm{T}} kW_k C \\
&= \sum_{k=1}^{\infty} (k-1) C^{\mathrm{T}} W_{k-1} C + \sum_{k=1}^{\infty} C^{\mathrm{T}} W_{k-1} C - \sum_{k=1}^{\infty} C^{\mathrm{T}} kW_k C \\
&= \sum_{k=0}^{\infty} C^{\mathrm{T}} kW_k C - \sum_{k=1}^{\infty} C^{\mathrm{T}} kW_k C + \sum_{k=0}^{\infty} C^{\mathrm{T}} W_k C \\
&= \sum_{k=0}^{\infty} C^{\mathrm{T}} W_k C \\
&= C^{\mathrm{T}} SC
\end{aligned}$$

So we have $U - C^{\mathrm{T}} UC = C^{\mathrm{T}} SC$. The second equation is proved.

7.11 Consider a linear system

$$X_{l+1} = CX_l + \xi_l, \quad l = 0, 1, \cdots,$$

with a discounted quadratic performance criterion

$$\eta(x) = \lim_{L\to\infty} E\{\sum_{l=0}^{L} \beta^l (X_l^T W X_l) | X_0 = x\}, \quad 0 < \beta < 1,$$

with W being a positive semi-definite matrix. Determine the performance potential of this LDQ (*Linear-discounted-quadratic*) problem.

Solution: Let performance potential $g_\beta = \{I + \sum_{k=1}^{\infty} \beta^k (P^k - e\pi)\} f$. Then we have discounted Poisson equation:

$$(I - \beta P + \beta e\pi) g_\beta = f.$$

$$\begin{aligned} g_\beta(x) &= f + \sum_{k=1}^{\infty} \beta^k (c_k e(x) + x^{\mathrm{T}} W_k x - \pi f e(x)) \\ &= \sum_{k=1}^{\infty} \beta^k (c_k - \pi f) e(x) + \sum_{k=0}^{\infty} x^{\mathrm{T}} \beta^k W_k x \end{aligned}$$

where $W_0 = W$. Let $S_\beta = \sum_{k=0}^{\infty} \beta^k W_k$, then

$$g_\beta(x) = \sum_{k=1}^{\infty} \beta^k (c_k - \pi f) e(x) + x^{\mathrm{T}} S_\beta x.$$

7.12 Consider a linear control problem

$$X_{l+1} = AX_l + Bu(X_l) + \xi_l, \quad l = 0, 1, \cdots,$$

with a discounted quadratic performance criterion

$$\eta(x) = \lim_{L\to\infty} E\{\sum_{l=0}^{L} \beta^l [X_l^T Q X_l + u_l^T V u_l | X_0 = x\}, \quad l = 0, 1, \cdots,$$

Applying policy iteration to this LDQ control problem to derive the (discounted) Riccati equation for the optimal policy.

Solution: Let $h(x) = g_\beta(x) = x^T S_\beta x$ in (7.43). From policy iteration approach for discounted MDP, we have

$$\begin{aligned} u' &= \arg\min_u \{\beta P^u g_\beta(x) + f^u(x)\} \\ &= \arg\min_u \{(Ax+Bu)^{\mathrm{T}} \beta S_\beta (Ax+Bu) + u^{\mathrm{T}} V u\} \\ &= -Dx \end{aligned}$$

where $D = (B^{\mathrm{T}} \beta S_\beta B + V)^{-1} B^{\mathrm{T}} \beta S_\beta A$. From the definition of S_β, we have

$$C^{\mathrm{T}} \beta S_\beta C = \sum_{k=0}^{\infty} \beta^{k+1} C^{\mathrm{T}} W_k C = \sum_{k=0}^{\infty} \beta^{k+1} W_{k+1} = S_\beta - W$$

Substituting $C = A - BD$, $W = Q + D^T V D$, and $D = (B^{\mathrm{T}} \beta S_\beta B + V)^{-1} B^{\mathrm{T}} \beta S_\beta A$ into the above equation, we obtain the discounted Riccati equation

$$S_\beta = A^{\mathrm{T}} \beta S_\beta A + A^{\mathrm{T}} \beta S_\beta B (V + B^{\mathrm{T}} \beta S_\beta B)^{-1} B^{\mathrm{T}} \beta S_\beta A + Q$$

7.13 Consider the JLQ problem

$$X_{l+1} = A_{M_l} X_l + B_{M_l} u_l + \xi_{M_l,l}, \tag{7.4}$$

in which the noises $\xi_{M_l,l}, M_l = 1, 2, \ldots, M$, have different probability distribution $P_{\xi_i}(y), y \in \Re^n$. Derive the solution to this problem.

Solution: We assume $P_{\xi_i}(y), y \in \Re^n$ has probability density $p_{\xi_i}(y)$, i.e. $P_{\xi_i}(dy) = p_{\xi_i}(y)dy$. For any quadratic function $h(i,x) = x^T W_i x$, where $W_i, i = 1, 2, \cdots, M$ are positive semi-definite matrices, and a control law $u(i,x)$, we have

$$\begin{aligned} &(P^u h)(i,x) = \sum_{j \in \mathcal{M}} \{p(j|i) h(j,y) P_i^u(dy|x)\} \\ = &\sum_{j \in \mathcal{M}} \Big\{ p(j|i) \int_{\Re^n} y^T W_j y p_{\xi_i}\{y - [A_i x + B_i u(i,x)]\} dy \Big\} \\ = &\sum_{j \in \mathcal{M}} \Big\{ p(j|i) \int_{\Re^n} \{z + [A_i x + B_i u(i,x)]\}^T W_j \{z + [A_i x + B_i u(i,x)]\} p_{\xi_i}(z) dz \Big\}. \end{aligned}$$

From $\int_{\Re^n} p_{\xi_i}(z)dz = 1$ and $\int_{\Re^n} z p_\xi(z)dz = 0$, we have

$$(P^u h)(i,x) = c_1(i) e(x) + \sum_{j \in \mathcal{M}} p(j|i) [A_i x + B_i u(i,x)]^T W_j [A_i x + B_i u(i,x)],$$

where $c_1(i) := \sum_{j\in\mathcal{M}} c_0(i,j)p(j|i)$ with

$$c_0(i,j) := \int_{\Re^n} [z^T W_j z] p_{\xi_i}(z) dz.$$

For a linear control $u_l = -D_{M_l} X_l$, the system (7.4) becomes

$$X_{l+1} = C_{M_l} X_l + \xi_{M_l,l},$$

where $C_i = A_i - B_i D_i, i = 1, 2, \cdots, M$. The performance function $f(i,x) = x^T W_i x$ with $W_i = Q_i + D_i^T V_i D_i$. Following the method in Section 7.3.2, we can obtain the similar results expect that $p_\xi(z)$ is replaced with $p_{\xi_i}(z)$.

8

Solutions to Chapter 8

8.1 In a discrete-time birth-death process, the system transits from state n to $n+1$ with a birth probability p_n, $0 < p_n < 1$, $n = 0, 1, \cdots$,and from state n to $n-1$ with a death probability q_n, $0 < q_n < 1$ and $p_n + q_n \leq 1$, or stays in the same state n with probability $1 - p_n - q_n$. When $n = 0$, the death probability is $q_0 = 0$. Define the events representing: a birth (denoted as event b), a death (denoted as event a), and no population change (denoted as event c), respectively.

Solution: A birth event b can be defined as

$$b = \{< n, n+1 >: n = 0, 1, 2, \cdots\}.$$

A death event a can be defined as

$$a = \{< n, n-1 >: n = 1, 2, \cdots\}.$$

No population change can be defined as

$$c = \{< n, n >: n = 0, 1, 2, \cdots\}.$$

8.2 In the discrete-time birth-death process considered in Problem 8.1, we set $p_n = p$ for all $n \geq 0$ and $q_n = q$ for all $n > 0$.

a. Find the steady-state probability $\pi(n)$, $n = 0, 1, \cdots$.

b. Suppose that we know a prior that at time l the system is at steady state, and we observed a birth event b at time $l - 1$, what is the conditional distribution $P(X_l|e_{l-1} = b)$?

c. What is the conditional probability of X_l if we have observed two consecutive birth events?

d. What if we observed a death event at steady state; i.e., what is $P(X_l|e_{l-1} = a)$?

Solution:

a. From the balance equation, we have

$$\pi(1) = \frac{p}{q}\pi(0), \pi(2) = \frac{p^2}{q^2}\pi(0), \pi(3) = \frac{p^3}{q^3}\pi(0), \cdots, \pi(n) = \frac{p^n}{q^n}\pi(0), \cdots.$$

Since $\sum_{i=0}^{\infty} \pi(i) = 1$, we have

$$(1 + \frac{p}{q} + \frac{p^2}{q^2} + \cdots)\pi(0) = 1$$

and

$$\pi(0) = 1 - \frac{p}{q}.$$

Thus, $\pi(n) = (1 - \frac{p}{q})\frac{p^n}{q^n}, n = 0, 1, 2, \cdots$.

b.

$$\begin{aligned} P(X_l = n|e_{l-1} = b) &= \frac{\pi(n-1)p(n|n-1)}{\sum_{n=1}^{\infty} \pi(n-1)p(n|n-1)} \\ &= \frac{(1 - \frac{p}{q})\frac{p^{n-1}}{q^{n-1}}p}{\sum_{n=1}^{\infty}(1 - \frac{p}{q})\frac{p^{n-1}}{q^{n-1}}p} \\ &= (1 - \frac{p}{q})\frac{p^{n-1}}{q^{n-1}}. \end{aligned}$$

c. The conditional probability of X_l if we have observed two consecutive birth events is

$$\begin{aligned} P(X_l = n | e_{l-2} = b, e_{l-1} = b) &= \frac{\pi(n-2)p(n-1|n-2)p(n|n-1)}{\sum_{n=2}^{\infty}\pi(n-2)p(n-1|n-2)p(n|n-1)} \\ &= (1-\frac{p}{q})\frac{p^{n-2}}{q^{n-2}}. \end{aligned}$$

d.

$$\begin{aligned} P(X_l = n | e_{l-1} = a) &= \frac{\pi(n+1)p(n|n+1)}{\sum_{n=0}^{\infty}\pi(n+1)p(n|n+1)} \\ &= (1-\frac{p}{q})\frac{p^{n}}{q^{n}}. \end{aligned}$$

8.3 Please define the following events in Problem 4.2 (the state of the system is the stock in every evening before the order):

a. the retailer ordered more than the next day's demand,

b. the retailer ordered less than or equal to the next day's demand, and

c. the retailer does not or may not have enough merchandise to sell.

Solution:

a. The event that the retailer ordered more than the next day's demand can be denoted as $\{< n, m >: n = 0, 1, \cdots, m = n+1, n+2, \cdots\}$.

b. The event that the retailer ordered less than or equal to the next day's demand can be denoted as $\{< n, m >: n = 0, 1, 2\cdots, m = 0, 1, \cdots, n\}$.

c. The event that the retailer does not or may not have enough merchandise to sell can be denoted as $\{< n, -1 >: n = 0, 1, 2, \cdots\}$, where -1 is a logical state to denote that the retailer does not or may not have enough merchandise to sell.

8.4 We modify and restate the retailer's problem (Problem 4.2 and Problem 8.2) as follows: The system state x is the stock left every evening. We only consider threshold types of policies. That is, the state space $\{0, 1, \cdots\}$ is divided into N intervals $I_1 := [0, n_1]$, $I_2 := [n_1 + 1, n_2]$, $\cdots$, $I_{N-1} := [n_{N-2}, n_{N-1}]$, $I_N := [n_{N-1}, \infty)$. The retailer is allowed to order M pieces of merchandise, or $2M$ pieces of merchandise, or not to order at all.

Assume that we can only observe that the state is in a particular interval and cannot observe the state itself. Based on the observation $x \in I_i$, $i = 1, 2, \cdots, N$, the retailer may choose different probabilities of ordering 0, M, or $2M$ pieces of merchandise. Every day's demand on merchandise can be described by an integer random variable with distribution p_n, $n = 0, 1, \cdots$. Describe the three types of events: the observable, the controllable, and the natural transition events.

Solution: The observable events include

$$\{< x, y >: x \in I_i, y \in \mathcal{S}\}, i = 1, 2, \cdots, N.$$

The controllable events include

$$\{< x, x - n >: x \in I_i, n = 0, 1, 2, \cdots\}, \{< x, x + M - n >: x \in I_i, n = 0, 1, \cdots\},$$
$$\{< x, x + 2M - n >: x \in I_i, n = 0, 1, 2, \cdots\}.$$

The natural transition events include

$$\{< x, x - n >\}, \{< x, x + M - n >\}, \{< x, x + 2M - n >\}, x \in \mathcal{S}, n = 0, 1, 2, \cdots.$$

8.5 Suppose that the derivative $\frac{df_\theta(i)}{d\theta}|_{\theta=0}$ is known. Derive a sample-path-based formula for the event-based average

$$\frac{df_\theta(k_1)}{d\theta}\Big|_{\theta=0} = \sum_{i=1}^{S} \left\{ \pi(i|e(k_1)) \frac{df_\theta(i)}{d\theta}\Big|_{\theta=0} \right\}.$$

Solution:

$$\frac{df_\theta(k_1)}{d\theta}\Big|_{\theta=0} = \lim_{L\to\infty} \frac{\sum_{l=0}^{L-1} I_{e(k_1)}(e_l) \frac{df_\theta(X_l)}{d\theta}\Big|_{\theta=0}}{\sum_{l=0}^{L-1} I_{e(k_1)}(e_l)}, \quad \text{w.p.1},$$

where $I_{e(k_1)}(e_l) = 1$ if $e_l = e(k_1)$, otherwise $I_{e(k_1)}(e_l) = 0$.

8.6* Derive equation (8.31), by using the arrival theorem and the steady state probabilities of the open Jackson networks.

Solution: From the arrival theorem, the conditional probability $\pi(\boldsymbol{n}|n)$ at the arriving times is equal to the time-average conditional probability. That is to say, $\pi(\boldsymbol{n}|n)$ is equal

to the steady-state conditional probability that the system stays at state $\boldsymbol{n}$ when the total number of customers in the system is n. Thus, we can use the results about the product-form solution in Section A.3.2 to conditional probability $\pi(\boldsymbol{n}|n)$ in (8.31).

Let λ_i be the overall arrival rate (including both external and internal arrivals) of the customers to server i. Then,

$$\lambda_i = \lambda_{0,i} + \sum_{j=1}^{M} \lambda_j q_{j,i}, i = 1, 2, \cdots, M. \tag{8.1}$$

where $\lambda_{0,i}$ is the external arrival rate to server i. Generally,

$$\lambda_{0,i} = \lambda q_{0,i}, \tag{8.2}$$

where λ is the total external arrival rate and $q_{0,i}$ is the probability that the external customer joins server i. The admission control policy can only affect the total external arrival rate λ. The total arrival rates λ^h and λ^d under different policies h and d have the following ratio relationship.

$$\lambda^h = k^{h,d}\lambda^d. \tag{8.3}$$

Putting (8.3) and (8.2) into (8.1), we can obtain $\lambda_i^h = k^{h,d}\lambda_i^d, i = 1, 2, \cdots, M$. Define $v_i = \frac{\lambda_i}{\mu_i}$. From the product-form solution of queueing networks (C.11), we have

$$\pi(\boldsymbol{n}|n) = \frac{1}{G_\Gamma(n)} \prod_{i\in\Gamma} \frac{v_i^{n_i}}{A_i(n_i)}$$

where $G_\Gamma(n) = \sum_{\sum_{i\in\Gamma} n_i = n} \prod_{i\in\Gamma} \frac{v_i^{n_i}}{A_i(n_i)}$. Since $v_i^h = k^{h,d} v_i^d$ and $A_i(n_i)$ does not change under different policies h and d, the ratio $\frac{1}{G_\Gamma(n)} \prod_{i\in\Gamma} \frac{v_i^{n_i}}{A_i(n_i)}$ remains the same under different policies. Therefore, (8.31) holds.

8.7 In Chapter 3, we derived a few sample-path-based direct-learning algorithms for the performance derivatives $\frac{d\eta_\delta}{d\delta}$, e.g., (3.30), (3.33), and (3.35). Derive similar direct-learning algorithms for the aggregated potentials (8.26) and the event-based performance derivatives by using formula (8.25).

Solution: Since

$$g_{\theta_1}(e_o = e_o(k_1), e_c = e_c(k_2)) = E_{\theta_1}\left\{ \sum_{l=0}^{\infty} f_{\theta_1}(X_l) | e_o^0 = e_o(k_1), e_c^0 = e_c(k_2) \right\},$$

we can view $\sum_{l=0}^{L-1} f_{\theta_1}(X_l)$ as an estimation of potential $g_{\theta_1}(k_1, k_2)$ when events $e_o(k_1)$ and $e_c(k_2)$ occur, where L is a truncation parameter. From (8.25), we have

$$\frac{d\eta(\theta)}{d\theta}\Big|_{\theta=\theta_1} = E_{\theta_1}\{\frac{df_\theta(X_l)}{d\theta}\Big|_{\theta=\theta_1} + \frac{\frac{d}{d\theta}p_\theta[e_c^l|e_o^l]}{p_\theta[e_c^l|e_o^l]}\Big|_{\theta=\theta_1} g_{\theta_1}(e_o^l, e_c^l)\}.$$

where e_o^l and e_c^l denote the observable event and the controllable event at time l, respectively. If the system is ergodic, we have

$$\begin{aligned}
\frac{d\eta(\theta)}{d\theta}\Big|_{\theta=\theta_1} &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\left\{\frac{df_\theta(X_n)}{d\theta}\Big|_{\theta=\theta_1} + \frac{\frac{d}{d\theta}p_\theta[e_c^n|e_o^n]}{p_\theta[e_c^n|e_o^n]}\Big|_{\theta=\theta_1} g_\theta(e_o^n, e_c^n)\right\} \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\left\{\frac{df_\theta(X_n)}{d\theta}\Big|_{\theta=\theta_1} + \frac{\frac{d}{d\theta}p_\theta[e_c^n|e_o^n]}{p_\theta[e_c^n|e_o^n]}\Big|_{\theta=\theta_1} \sum_{l=n}^{n+L-1} f_{\theta_1}(X_l)\right\} \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\left\{\frac{df_\theta(X_n)}{d\theta}\Big|_{\theta=\theta_1} + f_{\theta_1}(X_{n+L})\sum_{l=0}^{L-1}\frac{\frac{d}{d\theta}p_\theta[e_c^{n+l}|e_o^{n+l}]}{p_\theta[e_c^{n+l}|e_o^{n+l}]}\Big|_{\theta=\theta_1}\right\}.
\end{aligned}$$

8.8 Suppose that in an MDP problem, we can only apply control actions when the system is in a subset of state space, denoted as $\mathcal{I} \subset \mathcal{S}$. The observable events can be defined as when the system leaves any state $i \in \mathcal{I}$ or leaves the non-controllable set $\mathcal{S} - \mathcal{I}$.

a. Precisely define the observable events.

b. What are the controllable events?

c. Apply the event-based approach to this problem to derive the performance difference and derivative formulas for any two policies.

Solution:

a. The observable events can be defined as $e_o(i) := \{< i, j >: j \neq i \in \mathcal{S}\}$, $i \in \mathcal{I}$, $e_o(\bar{\mathcal{I}}) := \{< i, j >: i \in \mathcal{S} - \mathcal{I}, j \in \mathcal{I}\}$ and $\overline{\cup_{i\in\mathcal{I}} e_o(i) \cup e_o(\bar{\mathcal{I}})}$.

b. The controllable events are $\{< i, j >\}, i \in \mathcal{I}, j \neq i \in \mathcal{S}$.

c. Since the states in set $\mathcal{S} - \mathcal{I}$ cannot be controlled, the transition probability from observable event $e_o(\bar{\mathcal{I}})$ to any controllable event are 0. Then, according to the difference formula (8.19), we have

$$\begin{aligned}
&\eta^h - \eta^d \\
=\ &\sum_{i\in\mathcal{I}} \pi^h(i) \sum_{j\neq i\in\mathcal{S}} \{p^{h(i)}[j|i] - p^{d(i)}[j|i]\} g^d(j).
\end{aligned}$$

Defining $p^\delta(j|i) = p^{d(i)}(j|i) + \delta(p^{h(i)}(j|i) - p^{d(i)}(j|i)), i \in \mathcal{I}$, we have

$$\eta^\delta - \eta^d = \sum_{i\in\mathcal{I}} \pi^\delta(i) \sum_{j\neq i\in\mathcal{S}} \{p^\delta[j|i] - p^{d(i)}[j|i]\} g^d(j).$$

Let $\delta \to 0$, we have the following performance derivative

$$\frac{d\eta^\delta}{d\delta}\Big|_{\delta=0} = \sum_{i\in\mathcal{I}} \pi^d(i) \sum_{j\in\mathcal{S}} \{p^{h(i)}[j|i] - p^{d(i)}[j|i]\} g^d(j).$$

8.9 This problem is designed to further illustrate the ideas of natural transition events and potential aggregation. Compared with Example 8.1, there are two additional rooms 7 and 8, as shown in Figure 8.1. As in Example 8.1, after passing the green light on the right, the robot moves to the top; however, it will enter room 3 with probability u_1 and enter room 7 with probability u_2. Likewise, after passing the red light on the right, the robot will enter room 4 with probability v_1 and will enter room 8 with probability v_2.

a. Formulate this problem with the event-based approach, and define the observable, controllable, and natural transition events.

b. Derive the performance difference and derivative formulas.

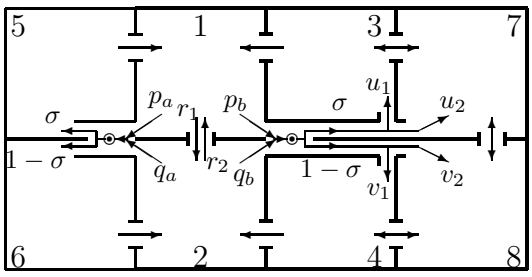

Figure 8.1: Extended Moving Robot Problem

Solution:

a. Similarly to formulation in Example 8.3, the process of the robot passing through a passage consists of three phases: First, the robot moves to the front of a light, either on the left or the right. (The robot moving to the front of the left light is called event a, and the robot moving to the front of the right light is called event b.) Second, an action is taken (turning on the red or the green light). We can control the probabilities of

the actions (red or green), by using the information obtained in the first phase (i.e., the robot moves to the front of the left, or the right, light). Third, the robot moves on to its destination following the instruction of the light, where the robot chooses the destination room with a natural probability distribution. These three phases can be modelled as three types of events: the observable events, the controllable events and the natural transition events. The observable events are:

the event that the robot moves to the front of the right light:

$$a = \{< 1,3 >, < 1,4 >, < 2,3 >, < 2,4 >, < 1,7 >, < 1,8 >, < 2,7 >, < 2,8 >\},$$

the event that the robot moves to the front of the left light:

$$b = \{< 1,5 >, < 1,6 >, < 2,5 >, < 2,6 >\},$$

and the event of "not an arrival":

$$\overline{a \cup b}.$$

The controllable events are:

the event that the robot moves to room 5:

$$c_1 = \{< 1,5 >, < 2,5 >\},$$

the event that the robot moves to room 6:

$$c_2 = \{< 1,6 >, < 2,6 >\},$$

the event that the robot moves to room 3 or room 7:

$$c_3 = \{< 1,3 >, < 2,3 >, < 1,7 >, < 2,7 >\},$$

the event that the robot moves to room 4 or room 8:

$$c_4 = \{< 1,4 >, < 2,4 >, < 1,8 >, < 2,8 >\}.$$

We can choose different σ to control the transition probabilities when an observable event occurs. The transition probabilities are:

$$p(c_1|b) = \sigma, p(c_2|b) = 1 - \sigma, p(c_3|a) = \sigma, P(c_4|a) = 1 - \sigma.$$

The natural transition events are:

$$n_1 = \{< 1,3 >, < 2,3 >\}, n_2 = \{< 1,7 >, < 2,7 >\}, n_3 = \{< 1,4 >, < 2,4 >\},$$
$$n_4 = \{1,8 >, < 2,8 >\}, n_5 = \{< 1,5 >, < 2,5 >\}, n_6 = \{< 1,6 >, < 2,6 >\},$$
$$n_7 = \{< 5,1 >, < 6,2 >, < 3,1 >, < 4,2 >, < 7,3 >, < 8,4 >, < 7,8 >, < 8,7 >\}.$$

b. Using the performance difference formula (8.19), we have

$$\begin{aligned} \eta' - \eta &= \sum_{e_o} \pi'(e_o) \sum_{e_c} [p'(e_c|e_o) - p(e_c|e_o)] g(e_o, e_c) \\ &= \pi'(a)(\sigma' - \sigma)[g(a, c_3) - g(a, c_4)] + \pi'(b)(\sigma' - \sigma)[g(b, c_1) - g(b, c_2)], \quad (8.4) \end{aligned}$$

where

$$g(e_o, e_c) = \sum_{i \in I[e_o]} \sum_{e_t} \pi'(i|e_o) p(e_t|e_o, e_c) g(j), \quad (8.5)$$

with $j = O_i[e_o \cap e_c \cap e_t]$. From (8.5), we have

$$g(b, c_1) = \sum_{i \in \mathcal{S}} \pi'(i|b) g(5) = g(5), g(b, c_2) = \sum_{i \in \mathcal{S}} \pi'(i|b) g(6) = g(6),$$
$$g(a, c_3) = \sum_{i \in \mathcal{S}} \pi'(i|a)[u_1 g(3) + u_2 g(7)] = u_1 g(3) + u_2 g(7),$$
$$g(a, c_4) = \sum_{i \in \mathcal{S}} \pi'(i|a)[v_1 g(4) + v_2 g(8)] = v_1 g(4) + v_2 g(8).$$

Thus, the performance difference formula is

$$\eta' - \eta = \pi'(a)(\sigma' - \sigma)\{u_1 g(3) + u_2 g(7) - [v_1 g(4) + v_2 g(8)]\} + \pi'(b)(\sigma' - \sigma)[g(5) - g(6)].$$

From (8.4), we can obtain the following performance derivative formula:

$$\frac{d\eta}{d\sigma} = \pi(a)\{u_1 g(3) + u_2 g(7) - [v_1 g(4) + v_2 g(8)]\} + \pi(b)[g(5) - g(6)].$$

8.10* A robot takes a random walk among four rooms, denoted as 1, 2, 3, and 4, as shown in Figure 8.2. When the robot is in room 3, in the next step, it moves to room 1. When it is in room 4, in the next step, it moves to rooms 2. There is a special passage that connects the four rooms as shown in the middle of Figure 8.2. When the robot is in room

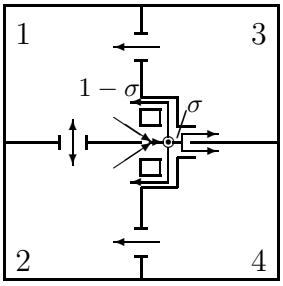

Figure 8.2: The Moving Robot System in Problem 8.10

1, in the next step, it moves to room 2 with probability $1-p_1$, or it tries to go through the passage with probability p_1. There is a traffic light, denoted as $\odot$ in the figure, in the passage. If it is red, the try fails and the robot moves back to room 1 in the next step; if the light is green, the robot passes the light and moves to room 3. The robot behaves in a similar way when it is in room 2: In the next step, it moves to room 1 with probability $1-p_2$, or it tries to go through the passage with probability p_2; and the robot moves back to room 2 in the next step if the light is red, and it passes the light and moves to room 4 in the next step, if the light is green. Denote the reward function as f.

Denote the probabilities of the light being green and red as σ and $1-\sigma$, respectively. We may control σ when we observe that the robot is in front of the light; we, however, do not know which room does the robot come from. Our goal is to determine the probability σ so that the long-run average reward is the maximum.

a. Formulate this problem with the event-based approach.

b. Derive the performance difference and derivative formulas.

c. Derive a policy iteration algorithm.

d. Show that one of the boundary points, $\sigma_{\max}$ or $\sigma_{\min}$, must be an optimal policy.

Solution:

a. The process of the robot passing through a passage also consists of three phases. These three phases correspond to three types of events. The observable events are

$$a = \{<1,1>,<1,3>,<2,2>,<2,4>\}, \quad \text{and} \quad \overline{a}.$$

The controllable events are

$$c_1 = \{< 1,3 >, < 2,4 >\}, c_2 = \{< 1,1 >, < 2,2 >\}.$$

The natural transition events are

$$n_1 = \{< 1,3 >\}, n_2 = \{< 1,1 >\}, n_3 = \{< 2,4 >\}, n_4 = \{< 2,2 >\},$$
$$n_5 = \{< 3,1 >\}, n_6 = \{< 4,2 >\}, n_7 = \{< 1,2 >, < 2,1 >\}.$$

b. Using the performance difference formula (8.19), we have

$$\begin{aligned} \eta' - \eta &= \sum_{e_o} \pi'(e_o) \sum_{e_c} [p'(e_c|e_o) - p(e_c|e_o)] g(e_o, e_c) \\ &= \pi'(a)(\sigma' - \sigma)[g(a, c_1) - g(a, c_2)], \end{aligned}$$

where

$$g(e_o, e_c) = \sum_{i \in I[e_o]} \pi'(i|e_o) \sum_{e_t} p(e_t|e_o, e_c) g(j), \tag{8.6}$$

with $j = O_i(e_o \cap e_c \cap e_t)$. From (8.6),

$$g(a, c_1) = \pi'(1|a)g(3) + \pi'(2|a)g(4), \quad g(a, c_2) = \pi'(1|a)g(1) + \pi'(2|a)g(2).$$

Thus, the performance difference formula is

$$\eta' - \eta = \pi'(a)(\sigma' - \sigma)\{[\pi'(1|a)g(3) + \pi'(2|a)g(4)] - [\pi'(1|a)g(1) + \pi'(2|a)g(2)]\}. \tag{8.7}$$

For this problem, the transition probability matrix is

$$P(\sigma) = \begin{bmatrix} p_1(1-\sigma) & 1-p_1 & p_1\sigma & 0 \\ 1-p_2 & p_2(1-\sigma) & 0 & p_2\sigma \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

From the balance equation $\pi P(\sigma) = \pi$, we have

$$\pi(1)(1-p_1) = \pi(2)(1-p_2),$$

for any σ. Thus, the conditional steady-state probability

$$\pi'(1|a) = \frac{\pi'(1)p(a|1)}{\pi'(1)p(a|1) + \pi'(2)p(a|2)} = \frac{\pi'(1)}{\pi'(1) + \pi'(2)} = \frac{\pi'(1)}{\pi'(1) + \pi'(1)\frac{1-p_1}{1-p_2}} = \frac{1-p_2}{2-p_1-p_2} \tag{8.8}$$

and

$$\pi'(2|a) = \frac{\pi'(2)p(a|2)}{\pi'(1)p(a|1) + \pi'(2)p(a|2)} = \frac{\pi'(1)}{\pi'(1) + \pi'(2)} = \frac{\pi'(1)\frac{1-p_1}{1-p_2}}{\pi'(1) + \pi'(1)\frac{1-p_1}{1-p_2}} = \frac{1-p_1}{2-p_1-p_2} \quad (8.9)$$

The conditional steady-state probability $\pi'(1|a)$ and $\pi'(2|a)$ do not depend on σ'. We can design the policy iteration algorithm for this problem. The difference formula (8.7) can be written as

$$\eta' - \eta = \pi'(a)(\sigma' - \sigma)\{[\pi(1|a)g(3) + \pi(2|a)g(4)] - [\pi(1|a)g(1) + \pi(2|a)g(2)]\}.$$

c. Policy Iteration Algorithm:

1. Select an initial policy $d_0 = \sigma^{(0)}$ and set $k = 0$;

2. Compute the potentials $g^{d_k}(1), g^{d_k}(2), g^{d_k}(3)$ and $g^{d_k}(4)$ by using the Poisson equation $(I - P^{d_k})g^{d_k} + \eta^{d_k}e = f^{d_k}$ and compute $\pi(1|a)$ and $\pi(2|a)$ by (8.8) and (8.9).

3. If $\pi(1|a)g^{d_k}(3) + \pi(2|a)g^{d_k}(4) > \pi(1|a)g^{d_k}(1) + \pi(2|a)g^{d_k}(2)$, set $\sigma^{(k+1)} = \sigma_{\max}$, otherwise, set $\sigma^{(k+1)} = \sigma_{\min}$.

4. If $d_{k+1} = d_k$, stops; otherwise go to step 2.

d. From the process of the above policy iteration algorithm, the improved policy must be one of the boundary points, $\sigma_{\max}$ or $\sigma_{\min}$. Since the number of such policies is finite, thus the policy iteration must stop at such a policy. If we assume the algorithm stops at policy d^*, for any policy $d = \sigma$,

$$(\sigma^d - \sigma^{d^*})\{[\pi(1|a)g^{d^*}(3) + \pi(2|a)g^{d^*}(4)] - [\pi(1|a)g^{d^*}(1) + \pi(2|a)g^{d^*}(2)]\} \leq 0$$

Thus, from the difference formula, we have $\eta^d \leq \eta^{d^*}$. d^* is the optimal policy.

8.11* Derive equation (8.59), by using the arrival theorem and the product-form solution to the steady state probabilities of the closed Jackson networks.

Solution: For a closed network, equation (C.5) holds, that is

$$v_i = \sum_{j=1}^{M} \tilde{q}_{j,i} v_j, \qquad i = 1, 2, \cdots, M. \quad (8.10)$$

where $\tilde{q}_{j,i} = p_{j,n_j} q_{i,j}$, if $j \neq i$, and $\tilde{q}_{j,j} = 1 - p_{j,n_j}$. Writing the above equations in a matrix form, we have

$$v = vQ,$$

where $v = (v_1, v_2, \ldots, v_M)$ and $Q = [\tilde{q}_{j,i}]$. If we only change the server rates of server k to $p^h_{k,n_k} \neq p^d_{k,n_k}$ and set $p^h_{i,n_i} = p^d_{i,n_i}$ for all n_i and $i \neq k$, then only the kth row in Q will change. That is, $\tilde{q}^d_{k,i} = p^d_{k,n_k} q_{k,i}, k \neq i$ and $\tilde{q}^d_{k,i} = 1 - p^d_{k,n_k}$ are changed to $\tilde{q}^h_{k,i} = p^h_{k,n_k} q_{k,i}, k \neq i$ and $\tilde{q}^h_{k,i} = 1 - p^h_{k,n_k}$. Let $c^{h,d} = \frac{p^d_{k,n_k}}{p^h_{k,n_k}}$. We can prove if $(v^d_1, v^d_2, \cdots, v^d_k, \cdots, v^d_S)$ is a solution of (8.10) under policy d, then $(v^d_1, v^d_2, \cdots, c^{h,d} v^d_k, \cdots, v^d_S)$ is a solution of (8.10) under policy h. For the closed network, the conditional probability $\pi(\boldsymbol{n}|a_{-k}(n_k))$ at the departure times is equal to the time-average conditional probability, that is,

$$\pi(\boldsymbol{n}|a_{-k}(n_k)) = \pi(\boldsymbol{n}_k|N - n_k).$$

where $\boldsymbol{n}_k$ denotes the state of other $M-1$ servers except server k and $\pi(\boldsymbol{n}_k|N-n_k)$ denotes the conditional steady-state probability that the state of other $M-1$ servers except server k is $\boldsymbol{n}_k$ when the total number of customers at these servers is $N - n_k$. From the product-form solution (C.11) and the fact that the solution of (8.10) under policy h is $(v^d_1, v^d_2, \cdots, c^{h,d} v^d_k, \cdots, v^d_S)$, we have

$$\pi^h(\boldsymbol{n}_k|N - n_k) = \pi^d(\boldsymbol{n}_k|N - n_k).$$

Therefore, (8.69) holds.

8.12 Develop a sample-path-based estimation algorithm for $g[a_{-k}(n_k), fb]$ in (8.61) and $g[a_{-k}(n_k), dp]$ in (8.62).

Solution: Consider a sample path of the closed Jackson network with L transitions under policy d. Denote the sequence of the time instants at which events $a_{-k}(n_k)$ and fb happen (A customer at server k moves back to server k after the completion of its service) on the sample path as $\mathcal{T}_{a_{-k}(n_k)} := \{l_1, l_2, \cdots, l_{L_{-n_k}}\}$. Choose a large N and set

$$\tilde{g}_{l_n} = \sum_{l=0}^{N-1} f(X_{l_n+l}).$$

Next, we group the set $\mathcal{T}_{a_{-k}(n_k)}$ into sub-groups $\mathcal{T}_{a_{-k}(n_k)} = \cup_{\boldsymbol{n} \in \mathcal{S} \text{ with } n_k} \mathcal{T}_{a_{-k}(\boldsymbol{n})}$, where $\mathcal{T}_{a_{-k}(\boldsymbol{n})}$ denotes the time instants at which events $a_{-k}(n_k)$ and fb happen and the state is $\boldsymbol{n}$ and the subscript "$\boldsymbol{n} \in \mathcal{S}$ *with* n_k" denotes all the states $\boldsymbol{n}$ with the number of customers at server k equal n_k. Let $L_{-\boldsymbol{n}}$ be the number of instants in $\mathcal{T}_{a_{-k}(\boldsymbol{n})}$. We have $L_{-n_k} = \sum_{\boldsymbol{n} \in \mathcal{S} \text{ with } n_k} L_{-\boldsymbol{n}}$.

Then

$$\begin{aligned}\frac{1}{L_{-n_k}}\sum_{n=1}^{L_{-n_k}}\tilde{g}_{l_n} &= \frac{1}{L_{-n_k}}\sum_{l_n\in\mathcal{T}_{a_{-k}(n_k)}}\tilde{g}_{l_n}\\ &= \frac{1}{L_{-n_i}}\sum_{\boldsymbol{n}\in\mathcal{S}\ with\ n_k}\ \sum_{l_n\in\mathcal{T}_{a_{-k}(\boldsymbol{n})}}\tilde{g}_{l_n}\\ &= \sum_{\boldsymbol{n}\in\mathcal{S}\ with\ n_k}\frac{L_{-\boldsymbol{n}}}{L_{-n_k}}\frac{1}{L_{-\boldsymbol{n}}}\sum_{l_n\in\mathcal{T}_{a_{-k}(\boldsymbol{n})}}\tilde{g}_{l_n}.\end{aligned}$$

Similarly to the argument in Section 8.4.2, when L is large enough, we have

$$\frac{1}{L_{-n_k}}\sum_{n=1}^{L_{-n_k}}\tilde{g}_{l_n}\approx\sum_{all\ \boldsymbol{n}\ with\ n_k}\pi^d[\boldsymbol{n}|a_{-k}(n_k)]g^d(\boldsymbol{n})=g^d[a_{-k}(n_k),fb].$$

Thus, we can estimate the potential $g^d[a_{-k}(n_k),fb]$ by using $\frac{1}{L_{-n_i}}\sum_{k=1}^{L_{-n_i}}\tilde{g}_{l_k}$.

Similarly, denote the sequence of the time instants at which events $a_{-k}(n_k)$ and dp happen (A customer at server k leaves the server after the completion of its service) on the sample path as $\{l_1,l_2,\cdots,l_{L_{-n_k}}\}$, we have

$$\frac{1}{L_{-n_k}}\sum_{n=1}^{L_{-n_k}}\bar{g}_{l_n}\approx\sum_{all\ \boldsymbol{n}\ with\ n_k}\pi^d[\boldsymbol{n}|a_{-k}(n_k)]q_{k,j}g^d(\boldsymbol{n}_{-k,+j})=g[a_{-k}(n_k),dp],$$

where $\bar{g}_{l_k}=\sum_{l=1}^{N}f(X_{l_k+l})$. Thus, we can use $\frac{1}{L_{-n_i}}\sum_{n=1}^{L_{-n_k}}\bar{g}_{l_n}$ to estimate the potential $g[a_{-k}(n_k),dp]$.

8.13 Consider the policy iteration Algorithm 8.1 in the service rate control problem in Section 8.5.2.

a. Prove that the algorithm reaches a local optimal policy in a finite number of iterations. Why is this policy not a "global" optimal policy?

b. If we change the policy improvement step to

> 3. (Policy improvement) For $i=1,\cdots,M$, do for $n_i=1,\cdots,N$, do
>
> i. if $g^{d_k}[a_{-i}(n_i),fb]\geq g^{d_k}[a_{-i}(n_i),dp]$ then set $p_{i,n_i}^{d_{k+1}}=\max_{1\leq l\leq K_{i,n_i}}p_{i,n_i}(l)$;
>
> ii. if $g^{d_k}[a_{-i}(n_i),fb]<g^{d_k}[a_{-i}(n_i),dp]$ then set $p_{i,n_i}^{d_{k+1}}=\min_{1\leq l\leq K_{i,n_i}}p_{i,n_i}(l)$;
>
> If $p_{i,n_i}^{d_{k+1}}\neq p_{i,n_i}^{d_k}$ for any i and n_i, then set $k:=k+1$ and go to step 2; If $p_{i,n_i}^{d_{k+1}}=p_{i,n_i}^{d_k}$ for all $i=1,\ldots,M$, and $n_i=1,\ldots,N$, stop.

What is the difference that such a change makes to the algorithm? Will this algorithm stop? Will it reach a local optimal policy?

Solution:

a. From the algorithm, we know if $i \neq M$, we have $\eta^{d_{k+1}} > \eta^{d_k}$. Thus, the average reward increases at each iteration before it stops. Because the number of policies is finite, the iteration procedure has to stop after a finite number of iterations. When the algorithm stops at a policy $\hat{d}$, since for any policy h,

$$[p^h_{k,n_k} - p^{\hat{d}}_{k,n_k}]\{g^{\hat{d}}[a_{-k}(n_k), fb] - g^{\hat{d}}[a_{-k}(n_k), dp]\} \leq 0, k = 1, 2, \cdots, M, n_k = 1, 2, \cdots, N.$$

Thus, for any policy h, the directional derivative

$$\frac{d\eta}{d\delta} = \sum_{i=1}^{M} \sum_{n_i=1}^{N} \left\{ \pi^{\hat{d}}[a_{-i}(n_i)](p^h_{i,n_i} - p^{\hat{d}}_{i,n_i})\{g^{\hat{d}}[a_{-i}(n_i), fb] - g^{\hat{d}}[a_{-i}(n_i), dp]\} \right\} \leq 0.$$

But we cannot determine $\eta^h - \eta^{\hat{d}} \leq 0$ because condition (8.69) may not hold under policy h and $\hat{d}$. Therefore, $\hat{d}$ is only a local optimal policy.

b. The algorithm does not only change the service rate of one server at each iteration, but change the service rates of all the servers. Under this change, condition (8.69) cannot hold. Thus, this algorithm can not increase the average reward at each iteration and cannot stop. It will also not reach a local optimal policy.

8.14 In the policy iteration algorithm in the service rate control problem in Section 8.5.2, at every iteration we always start from server 1, in the order of server 1, server 2, and so on, to update the service rates of the servers. We may try to update the service rates of the servers in a round-robin way: e.g., if server 1's service rates are updated at an iteration, then in the next iteration, we start from server 2 to update the service rates, etc. Develop such an algorithm and discuss its advantages, if any.

Solution:

1. Guess an initial policy d_0, set $k := 0$, $i := 1$ and $c := 0$.

2. (Policy evaluation) Estimate the aggregated potentials $g^{d_k}[a_{-j}(n_j), fb]$ and $g^{d_k}[a_{-j}(n_j), dp]$ for $j = 1, \cdots, N$ defined in (8.61) and (8.62) on a sample path of the system under policy d_k.

3. (Policy improvement) for $n_i = 1, 2, \cdots, N$, do

   (a) if $g^{d_k}[a_{-i}(n_i), fb] \geq g^{d_k}[a_{-i}(n_i), dp]$ then set $p_{i,n_i}^{d_{k+1}} = \max_{1 \leq l \leq K_{i,n_i}} p_{i,n_i}(l)$;

   (b) if $g^{d_k}[a_{-i}(n_i), fb] < g^{d_k}[a_{-i}(n_i), dp]$ then set $p_{i,n_i}^{d_{k+1}} = \min_{1 \leq l \leq K_{i,n_i}} p_{i,n_i}(l)$.

4. If $p_{i,n_i}^{d_{k+1}} = p_{i,n_i}^{d_k}$ for all $n_i = 1, 2, \cdots, N$, set $c := c + 1$, otherwise, set $c := 0$.

   If $c = M$, stop;

   otherwise, set $k := k + 1$ and $i := i + 1$, if $i > M$, set $i := 1$, go to step 2.

The advantage: This algorithm does not need to start from server 1 at every iteration. It can update the service rates of the servers in a round-robin way.

8.15* (*Options* [15]) This problem is closely related to the time aggregation formulation. Consider a Markov process $\boldsymbol{X}$ with state space $\mathcal{S}$, and let $\mathcal{I} \subset \mathcal{S}$ be a subset of $\mathcal{S}$. As in Problem 8.8, we may define an observable event as when the system leaves a state in $\mathcal{I}$. Let us call the period between two consecutive events (i.e., two consecutive visits to $\mathcal{I}$) as an *option* period. The control problem is described as follows. There is a space, denoted as Π, of a finite number of options. An option corresponds to a state transition probability matrix in $\mathcal{S}$ (i.e., equivalent to a policy); however, it is only applied to an option period. After the system visits a state $i \in \mathcal{I}$, the system may evolve with any option in the available option set $\Pi_i \subseteq \Pi$ until it reaches the next state $j \in \mathcal{I}$. We assume that under any option in Π, the set $\mathcal{I}$ is reachable.

We consider randomized policies. Thus, in this problem for any given $i \in \mathcal{I}$ a policy specifies a probability distribution on Π_i. Precisely, let $o_{i,1}, o_{i,2}, \cdots, o_{i,n_i}$ be the options in Π_i. A policy d specifies a probability distribution $d(i) := (p_{i,1}, \ldots, p_{i,n_i})$. With policy d, the system operates under option $o_{i,k}$ with probability $p_{i,k}$, $\sum_{k=1}^{n_i} p_{i,k} = 1$. Our goal is to determine the policy that achieves the maximum long-run average reward. For simplicity, we assume that the reward function f is the same for all policies.

The standard event-based optimization approach discussed in this chapter does not directly apply to this problem. However, the basic principles and concepts can be easily modified and extended to this problem. In the standard formulation, a control action taken at a time instant only affects the transition to the next state and therefore the

controllable event can be defined. In the option problem, however, a control action affects the transitions in the entire option period.

Please formulate this problem in the framework of event-based optimization.

a. What are the observable events?

b. What are the aggregated potentials? (Hint: it can be denoted as $g(i, o_i)$.)

c. Derive the performance difference and derivative formulas for the two policies in the problem.

d. Comment on this event-option based optimization approach.

Solution:

a. The observable events are $e_o(i) := \{< i, j >: j \neq i \in \mathcal{S}\}, i \in \mathcal{I}$ and $\overline{\cup_{i\in\mathcal{I}} e_o(i)}$.

b. From (8.37), the aggregated potential is $g(i, o_{i,l}) = \sum_{j\in\mathcal{I}} p[j|i, o_{i,l}]g(j), i \in \mathcal{I}, l = 1, 2, \cdots, n_i$, where $p[j|i, o_{i,l}]$ denotes the probability that the process $\boldsymbol{X}$ transits from state i to state j in an option period.

c.

$$\eta^h - \eta^d = \sum_{i\in\mathcal{I}} \pi^h(i) \sum_{l=1}^{n_i} [p^h_{i,l} - p^d_{i,l}] g^d(i, o_{i,l}).$$

If we assume the policy depends on a parameter θ, then the performance derivative is

$$\frac{d\eta(\theta)}{d\theta} = \sum_{i\in\mathcal{I}} \pi^h(i) \sum_{l=1}^{n_i} \frac{dp_{i,l}(\theta)}{d\theta} g^d(i, o_{i,l}).$$

d. For this problem, we can find that condition (8.36) holds naturally. This is because for any observable event $e_o := \{< i, j >: j \neq i \in \mathcal{S}\}$, the conditional steady-state probability $\pi^h(i|e_o) \equiv 1$ for any policy h. However, for the option problem, it is difficult to define the controller events and natural transition events by using state transitions. This is because there exists many state transitions in an option period. If we only consider the time instants that the observable event occurs, the process is still Markovian and the event-based method can be directly utilized. This idea is the same as the time aggregation formulation.

8.16* Consider a partially observable Markov chain with the structure shown in Figure 8.3. The 15 states are grouped into three functionally similar groups. Group 1 consists

of 5 states denoted as 1, 111, 112, 121, and 122; Group 2 consists of 5 states denoted as 2, 211, 212, 221, and 222; and Group 3 consists of 5 states denoted as 3, 311, 312, 321, and 322. States 1, 2, and 3 are completely observable. The other 12 states are grouped in to 6 super-states, denoted as 11, 12, 21, 22, 31, and 32; each consisting of two states as shown in the figure; e.g, the super-state 11 consists of two states 111 and 112. Only the super-states are observable; for example, after the system transits out from state 1, we only know that the system is in super-state 11 or 12 and cannot know which exact state the system is in. The state transition probabilities are indicated in the figure. The transition probabilities from the observable states 1, 2, and 3, e.g., $p_{1,11}$, $p_{1,12}$, $p_{1,21}$, and $p_{1,22}$, are fixed and known. The transition probabilities from the non-observable states are controllable by actions and are denoted as $p^{\alpha}_{111;2}$, $p^{\alpha}_{111;3}$, $p^{\alpha}_{121;2}$, and $p^{\alpha}_{121;3}$, etc. The superscript α denotes any feasible action for the corresponding state. Because we cannot determine the exact state in a super-state, we need to assume that the sets of the feasible actions for the two states in a super-state are the same. For example, if we know that the system is in super-state 12 and decides to take action α, then this action must be feasible to both 121 and 122.

A sample path of the Markov chain may look like: $\boldsymbol{X} = \{2, 221, 1, 112, 2, 211, 3, 322, 1, 111, \cdots\}$, with an observable state followed by a non-observable state and followed by another observable state, etc. The corresponding observed random sequence is $\mathbf{Y} = \{2, 22, 1, 11, 2, 21, 3, 32, 1, 11, \cdots\}$.

Suppose that when the system is at state x, a random reward is received with $f(x)$ being its average. In addition, we assume that the function f is unknown but the reward at any time instant is observable. We consider the optimization of the long-run average reward. Please formulate this problem in the event-based formulation.

a. Explain that in this POMDP problem, a memoryless policy is a mapping from the space $\{11, 12, 21, 22, 31, 32\}$ to the action space.

b. What are the observable events?

c. What are the aggregated potentials?

d. Derive the performance difference and derivative formulas for the two policies in the

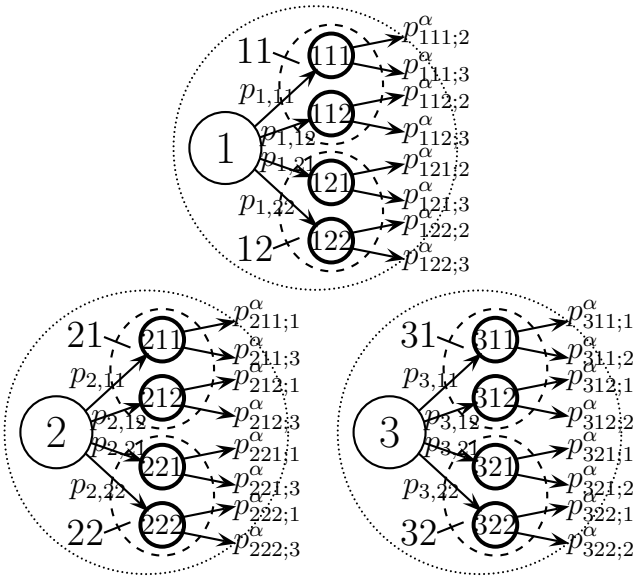

Figure 8.3: The POMDP in Problem 8.16

problem.

e. Can we develop a policy iteration algorithm for the performance optimization of this problem? If so, please describe the algorithm in detail.

Solution:

a. In POMDP, a memoryless policy is to choose action according to the current observation. Thus, this policy is a mapping from the space $\{11, 12, 21, 22, 31, 32\}$ to the action space. For example, when the system is in supper state 12, we choose an action from the action space according to the current supper state 12. A memoryless policy may also be a stochastic policy. For example, when the system is in supper state 12, we can choose an action from the action space with a probability distribution determined by super state 12. At that time, the memoryless policy is a mapping from the space $\{11, 12, 21, 22, 31, 32\}$ to the probability distribution set on the action space. The stochastic memoryless policy may be better than the deterministic memoryless policy. We consider the stochastic memoryless policy in this problem.

b. The observable events are $e_{11} = \{< i, j >: i \in \{111, 112\}, j \in \{2, 3\}\}$, $e_{12} = \{<$ $i, j >: i \in \{121, 122\}, j \in \{2, 3\}\}$, $e_{21} = \{< i, j >: i \in \{211, 212\}, j \in \{1, 3\}\}$,$e_{22} = \{<$ $i, j >: i \in \{221, 222\}, j \in \{1, 3\}\}$, $e_{31} = \{< i, j >: i \in \{311, 312\}, j \in \{1, 2\}\}$,$e_{32} = \{<$

$i,j>: i \in \{321, 322\}, j \in \{1,2\}\}$ and $\overline{e_{11} \cup e_{12} \cdots e_{32}}$.

c. For this problem, we have $\pi(ijk|e_{i,j}) = \frac{\pi(i)p(ijk|i)p(ij|ijk)}{\pi(i)\sum_{j=1,2} p(ijk|i)p(ij|ijk)} = \frac{p_{i,jk}}{\sum_{k=1,2} p_{i,jk}}, i = 1,2,3, j,k = 1,2$, which do not depend on the policy. Therefore, the aggregated potentials are $g^d(e_{i,j},\alpha) = \sum_k \pi(ijk|e_{i,j}) \sum_{m\neq i, m=1,2,3} p^{\alpha}_{ijk,m} g^d(m), i = 1,2,3, j = 1,2$.

d. The performance difference formula is

$$\eta^h - \eta^d = \sum_{i=1,2,3,j=1,2} \pi^h(e_{i,j}) \sum_{\alpha \in \mathcal{A}(e_{i,j})} [p^h(\alpha|e_{i,j}) - p^d(\alpha|e_{i,j})] g^d(e_{i,j},\alpha),$$

where $p^h(\alpha|e_{i,j})$ and $p^d(\alpha|e_{i,j})$ denotes the probabilities that the stochastic memoryless policies h and d choose action α when the observable event $e_{i,j}$ occurs and $\mathcal{A}(e_{i,j})$ denotes the available action space when event $e_{i,j}$ occurs. The performance derivative is

$$\frac{d\eta(\theta)}{d\theta} = \sum_{i=1,2,3,j=1,2} \pi^d(e_{i,j}) \sum_{\alpha \in \mathcal{A}(e_{i,j})} \frac{dp_\theta(\alpha|e_{i,j})}{d\theta} g^d(e_{i,j},\alpha),$$

when the probability $p(\alpha|e_{i,j})$ depends on parameter θ.

e. Since $\pi(ijk|e_{i,j})$ does not depend on the policy, we can develop the policy iteration algorithm for the performance optimization of this problem.

Algorithm:

1. Select an initial policy d_0, and set $k = 0$

2. Estimate the potential $g^{d_k}(e_{i,j},\alpha)$ based on a sample path under policy d_k (The estimation is similar to Problem 8.12).

3. Choose a policy d_{k+1} such that

$$d_{k+1} \in arg\max_{d\in\mathcal{D}} \sum_{\alpha \in \mathcal{A}(e_{i,j})} [p^d(\alpha|e_{i,j}) - p^{d_k}(\alpha|e_{i,j})] g^{d_k}(e_{i,j},\alpha),$$

   for all $e_{i,j}$.

4. If $d_{k+1} = d_k$, stop; otherwise, set $k := k+1$ and go to step 2.

8.17* We consider a POMDP problem with the structure shown in Figure 8.4. The 4 states 1, 2, 3, and 4 are grouped into 2 super-states a and b, with $a = \{1,2\}$ and $b = \{3,4\}$. The super-states are observable, but the states are not. A sample path

may look like $\boldsymbol{X} = \{1, 2, 4, 2, 3, 4, 1, 2, 4, 1, 4, 1, 2, 3, 2, 3, 4, 1, \cdots\}$, and the corresponding observed random sequence is

$$\mathbf{Y} = \{a, a, b, a, b, b, a, a, b, a, b, a, a, b, a, b, b, a, \cdots\}. \tag{8.11}$$

Unlike in Problem 8.16 where a super-state completely determines the probability distribution of the system state, here the state distribution may depend on the history of the observed super-states. For example, if we observe two a's in a row, from Figure 8.4 we know that the system must be in state 2. Similarly, two consecutive observations of "b" lead to a state 4. Therefore, after two consecutive a's or b's, denoted as (a, a) or (b, b), the system "regenerates" from state 2 or 4.

The regenerative property simplifies the analysis as well as the notation. Let x, or x', denote any sequence of super-states. Then an observation history (x', a, a, x) can be denoted as (a, a, x), and (x', b, b, x) can be denoted as (b, b, x), because the past history x' does not contain any extra information. Furthermore, if x is non-null, we may further omit the prefix (a, a) or (b, b) and simply denote them as x (if x starts with a, the prefix cannot be (a, a), and vise versa). Therefore, the observation histories correspond to the following cases: (a, a), (b, b), (a), (b), (a, b), (b, a), (a, b, a), (b, a, b), and (a, b, a, b), and so on. In general, the sequence alternates between a and b.

If at a time instant the observation history is $\mathbf{Y} = \{x', a, a, x\}$ or $\mathbf{Y} = \{x', b, b, x\}$, then x (or (a, a) and (b, b) if x is null) completely determines the probability distribution of the states at that time instant. For example, $x = (a)$ implies that the system just transits from state 4 to state 1 or 2. Thus, the state probability distribution is $p(3) = p(4) = 0$ and $p(1) = \frac{p^{\alpha}_{4,1}}{p^{\alpha}_{4,1}+p^{\alpha}_{4,2}}$ and $p(2) = \frac{p^{\alpha}_{4,2}}{p^{\alpha}_{4,1}+p^{\alpha}_{4,2}}$.

Therefore, in terms of the state probability distribution, the history $\mathbf{Y}$ in (8.11) is equivalent to

$$\{\bullet, 2, (b), (b, a), (b, a, b), 4, (a), 2, (b), (b, a), (b, a, b), (b, a, b, a), 2, (b), (b, a), (b, a, b), 4, (a), \cdots, \}$$

where "$\bullet$" represents the initial probability.

Suppose that when the system is at state i, a random reward is received with $f(i)$ being its average. In addition, we assume that the function f is unknown but the reward at any time instant is observable. We consider the long-run average reward, its existence is guaranteed by the regenerative property.

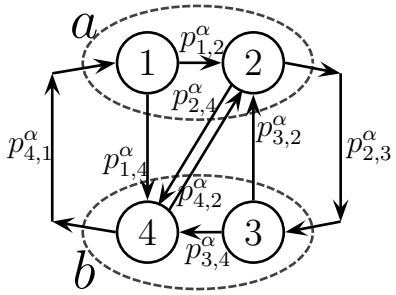

Figure 8.4: The POMDP in Problem **??**

a. Derive the state probability distributions corresponding to (b), (b,a) (a,b) and so on.

b. What are the observable events?

c. What are the aggregated potentials?

d. Derive the performance difference and derivative formulas for the two policies in the problem.

e. Can we develop a policy iteration algorithm for the performance optimization of this problem? If so, please describe the algorithm in details.

Solution:

a.

$$\pi(i|b) = \frac{\pi(i)p(b|i)}{\sum_{i\in\mathcal{S}}\pi(i)p(b|i)}.$$

Since $p(b|i) = 0$ when $i = 1, 2$ and $p(b|i) = 1$ when $i = 3, 4$, then $\pi(1|b) = \pi(2|b) = 0$ and $P(3|b) = \frac{\pi(3)}{\pi(3)+\pi(4)}$, $P(4|b) = \frac{\pi(4)}{\pi(3)+\pi(4)}$. For the observation history (b,a), we have

$$\pi(j|(b,a)) = \frac{\sum_{i\in\mathcal{S}}\pi(i)p(b|i)p(j|i)p(a|j)}{\sum_{i\in\mathcal{S}}\pi(i)p(b|i)\sum_{j\in\mathcal{S}}p(j|i)p(a|j)}.$$

Then, $p(1|(b,a)) = \frac{\pi(4)p^\alpha(1|4)}{\pi(3)p^\alpha(2|3)+\pi(4)}$, $p(2|(b,a)) = \frac{\pi(4)p^\alpha(2|4)+\pi(3)p^\alpha(2|3)}{\pi(4)+\pi(3)p(2|3)}$ and $p(3|(b,a)) = p(4|(b,a)) = 0$. For the observation history (a,b), we have

$$\pi(j|(a,b)) = \frac{\sum_{i\in\mathcal{S}}\pi(i)p(a|i)p(j|i)p(b|j)}{\sum_{i\in\mathcal{S}}\pi(i)p(a|i)\sum_{j\in\mathcal{S}}p(j|i)p(b|j)}.$$

Then, $\pi(3|(a,b)) = \frac{\pi(2)p^\alpha(3|2)}{\pi(1)p^\alpha(4|1)+\pi(2)}$, $\pi(4|(a,b)) = \frac{\pi(2)p^\alpha(4|2)+\pi(1)p^\alpha(4|1)}{\pi(1)p^\alpha(4|1)+\pi(2)}$. and $\pi(1|(a,b)) = \pi(2|(a,b)) = 0$.

b. The observable events are $a = \{< i, j >: i \in \{1, 2\}, j \in \mathcal{S}\}$,$b = \{< i, j >: i \in \{3, 4\}, j \in \mathcal{S}\}$.

c. Let e denote a sequence of observable event, then, from the regenerative property, $e \in \{(a, a), (b, b), (a), (b), (a, b), (b, a), (a, b, a), (b, a, b), (a, b, a, b)\}$. we have the following aggregated potential

$$g^{h,d}(e, \alpha) = \sum_{i \in \mathcal{S}} \pi^h(i|e) p^\alpha(j|i) g^d(j)$$

d. The performance difference formula is

$$\eta^h - \eta^d = \sum_e \pi^h(e) \sum_\alpha [p^h(\alpha|e) - p^d(\alpha|e)] \sum_{i \in \mathcal{S}} \pi^h(i|e) p^\alpha(j|i) g^d(j),$$

where $p^h(\alpha|e)$ and $p^d(\alpha|e)$ denote the probabilities that the policies h and d choose action α when the observable event sequence e occurs. The performance derivative formula is

$$\frac{d\eta}{d\theta} = \sum_e \pi^d(e) \sum_\alpha \frac{dp^d_\theta(\alpha|e)}{d\theta} \sum_{i \in \mathcal{S}} \pi^d(i|e) p^\alpha(j|i) g^d(j).$$

e. Since $\pi^h(i|e) \neq \pi^d(i|e)$ when $h \neq d$ in general, we cannot design the policy iteration algorithm for the performance optimization of this problem.

8.18* Suppose that in Problem 8.17, for simplicity we only take (a, a), (b, b), (a), (b), (a, b), and (b, a) as the possible events; i.e., we aggregate the histories according to the latest two super-states. For example, history (a, b, a, b, a) is aggregated into (b, a) and so on. In this formulation, the action taken at a time instant depends only on the last two super-states in the observation history.

a. Derive the performance difference formula.

b. Explain that in general, policy iteration cannot be developed from such a performance difference formula.

c. Do this problem and Problem 8.15 help you understand the POMDP problems?

Solution:

a. Define $\mathcal{E} = \{(a, a), (b, b), (a), (b), (a, b), (b, a)\}$. The performance difference formula is

$$\eta^h - \eta^d = \sum_{e \in \mathcal{E}} \pi^h(e) \sum_\alpha [p^h(\alpha|e) - p^d(\alpha|e)] \sum_{i \in \mathcal{S}} \pi^h(i|e) p^\alpha(j|i) g^d(j).$$

b. In general, the conditional steady-state probability $\pi^h(i|e)$ cannot be equal to $\pi^d(i|e)$. For example, from part a) in Problem 8.17, we can find $\pi(3|(a,b))$ depends on $\pi(3)$ and $\pi(2)$, which are different under different policies in general. Thus, policy iteration cannot be developed from such a performance difference formula.

c. From this problem, we can find the policy of POMDP generally depends on the history. To obtain the optimal policy, e may depend on the whole history. If we only want to obtain a suboptimal policy, we can consider the finite history. For example, in this problem, histories (a,b,b,a) and (b,a,b,a) can be truncated into history (b,a).

8.19* In Problem 8.17, if we can trace back from the observation history, we can estimate the earlier system state better. For example, as shown in (8.11), the observations from $l=0$ to $l=5$ are $\{a,a,b,a,b,b\}$. We know that at $l=1$, the system is at $X_1=2$, and the state probability distributions at times $l=3$, $l=4$, and $l=5$ can be calculated, see Problem 8.17. However, at $l=5$ we have observed (b,b) and therefore we know that the system state is $X_5=4$. Knowing so, from the structure shown in Figure 8.4, we may trace back to $l=4$ and assert that $X_4=3$. Similarly, we can know for sure that $X_3=2$.

a. Update the state probability distribution at $l=2$ after observing $\{a,a,b,a,b,b\}$ at $l=5$.

b. Does this posterior information help in determining the optimal policy?

Solution:

a. From observed histories (a,a) and (b,b), we can completely determine $X_1=2$ and $X_5=4$. From the structure shown in Figure 8.4 or Figure 8.17 in the textbook, we can trace back to $l=4$ and assert $X_4=3$. Similarly, we can know for sure that $X_3=2$. Since the observation is b at $l=2$, we can assert $X_l=3$ or $X_l=4$. From $X_1=2$ and $X_3=2$, the state probability distribution at $l=2$ is $p(3|(a,a,b,a,b,b))=\frac{p^\alpha(3|2)p^\alpha(2|3)}{p^\alpha(3|2)p(2|3)+p^\alpha(4|2)p^\alpha(2|4)}$ and $p^\alpha(4|(a,a,b,a,b,b))=\frac{p^\alpha(4|2)p(2|4)}{p^\alpha(3|2)p^\alpha(2|3)+p^\alpha(4|2)p^\alpha(2|4)}$.

b.This posterior information does not help in determining the optimal policy because the policy depends only on the history and cannot depend on the future information.

8.20* In the analytical approach for MDPs, the reward function $f(i)$ is assumed to be known; and in the reinforcement learning approach, the reward at every time instant is

assumed to be observable. In MDPs, the state i is assumed to be completely observable, therefore, both assumptions are equivalent. In POMDPs, however, the state is not observable; therefore, knowing the form of the function $f(i)$ does not allow us to know the actual reward at every instant. As such, we may have four different situations regarding the rewards:

i. The function f is known, and the reward at every instant is observable;

ii. The function f is known, but the reward at every instant is not observable;

iii. The function f is not known, but the reward at every instant is observable; and

vi. The function f is not known, and the reward at every instant is also not observable, but the final reward at the completion of each sample path is known.

In Problems 8.16 and 8.17, we take the learning approach and therefore we were dealing with the third situation. In addition, we assumed that the reward is random with a unknown mean $f(i)$.

Now, let us further assume that the reward at any state i is a fixed deterministic number $f(i)$, which is an unknown function but the reward received at every time instant is observable. In this case, we may determine the state i by the reward received. For instance, in Problem 8.16, when super-state 11 is observed, the system may be in either state 111 or 112 with probabilities $\sigma_{111} := \frac{p_{1,11}}{p_{1,11}+p_{1,12}}$ or $\sigma_{112} := \frac{p_{1,12}}{p_{1,11}+p_{1,12}}$, respectively. Thus, the reward received is either $f(111)$ or $f(112)$ with probabilities σ_{111} or σ_{112}, respectively. To be more precise, suppose $\sigma_{111} = 0.4$ and $\sigma_{112} = 0.6$. Let us observe the sample path for a while. We may find that when 11 is observed, we have 0.4 chance of obtaining a reward of 0 and 0.6 chance of obtaining a reward of 1. Then we can easily know that $f(111) = 0$ and $f(112) = 1$, and later on when 11 is observed, if we receive 0 we know that the state is 111 and if we receive 1, we know it is in 112. The following questions are for your further investigation:

a. Can we develop an algorithm from this reasoning?

b. Can we apply the same reasoning to Problem 8.17?

c. Can we apply the same reasoning to the general POMDPs?

Solution:

a. If have known that $f(111) = 0$ and $f(112) = 1$, then we have $p(f = 0|111) = 1$, which denotes we can obtain the reward 0 with probability 1 when the state is 111. Similarly, we have $p(f = 0|112) = 0$ and $p(f = 1)|111) = 0, p(f = 1|112) = 1$. After that, we can use the reward information to estimate the state information. In problem 8.16, when super-state 11 is observed, we only use the observation information from the super-state to estimate the state and obtain the system may be in either state 111 or 112 with probabilities $\sigma_{111} := \frac{p_{1,11}}{p_{1,11}+p_{1,12}}$ or $\sigma_{112} := \frac{p_{1,12}}{p_{1,11}+p_{1,12}}$, respectively. If the reward information is added, there are two types of observations: super-states and rewards. Thus, if we receive a reward $f = 0$, the system may be in state 111 with probability $\frac{p_{1,11}p(f=0|111)}{p_{1,11}p(f=0|111)+p_{1,12}p(f=0|112)} = \frac{p_{1,11}}{p_{1,11}} = 1$ or in state 112 with probability $\frac{p_{1,12}p(f=0|112)}{p_{1,11}p(f=0|111)+p_{1,12}p(f=0|112)} = \frac{0}{p_{1,11}} = 0$. Similarly, if we receive a reward $f = 1$, then the system is in state 111 with probability $\frac{p_{1,11}p(f=1|111)}{p_{1,11}p(f=1|111)+p_{1,12}p(f=1|112)} = 0$ or in state 112 with probability $\frac{p_{1,12}p(f=1|112)}{p_{1,11}p(f=1|111)+p_{1,12}p(f=1|112)} = 1$. The general algorithm will be given for the POMDP problem in part c).

b. We consider a simple case in Problem 8.17 . If at a time the observation history is $\mathbf{Y} = \{x', b, b, x\}$, $x = (a)$ implies that the system just transits from state 4 to state 1 or 2. Thus, the system may be in either state 1 or 2 with probabilities $p(1) = \frac{p^{\alpha}_{4,1}}{p^{\alpha}_{4,1}+p^{\alpha}_{4,2}}$ and $p(2) = \frac{p^{\alpha}_{4,2}}{p^{\alpha}_{4,1}+p^{\alpha}_{4,2}}$, respectively. We may find that when (b, b, a) is observed, we have $p(1)$ chance of obtaining a reward of $f(1)$ and $p(2)$ chance of obtaining a reward of $f(2)$, then we can easily know that $f(1)$ and $f(2)$, and later on when (b, b, a) and $f(1)$ is observed, we know the state is 1 and if $f(2)$ is observed we know that the state is 2.

c. We can apply the same reasoning to the general POMDPs. For the general POMDPs, we can firstly compute the steady-state probability $\pi(i)$, then we have $\pi(i)$ chance of obtaining a reward $f(i), i \in \mathcal{S}$. Then, if the rewards are different for different states, we can know $p(f = f(i)|i) = 1$. Then, we can update the state probability distribution that the system is in state $i \in \mathcal{S}$ as follows:

$$b_0(i) = \frac{\pi_0(i)p(o_0|i)p(f_0|i)}{\sum_{i\in\mathcal{S}} \pi_0(i)p(o_0|i)p(f_0|i)},$$
$$b_l(i) = \frac{b_{l-1}(i)p(o_l|i)p(f_l|i)}{\sum_{i\in\mathcal{S}} b_{l-1}(i)p(o_l|i)p(f_l|i)}, l = 1, 2, \cdots,$$

where o_l and f_l denote the observation and the reward at time l, respectively.

9

Solutions to Chapter 9

9.1 As explained in Section 9.2, in the performance difference construction approach shown in Figure 9.1, the construction is done in the following way:

i. On the perturbed sample path $A - B - E - D$, we use the same random variable ξ_l to determine whether or not there is a jump at each transition l; and

ii. when a jump is identified, we use another independent sequence of random variables to generate an auxiliary path, e.g., $W - C$.

While the above construction is convenient, it is not necessary. Convince yourself that we can derive the same results as those in Section 9.2 if we construct the sample paths in the following way:

i. On the perturbed sample path $A - B - E - D$, we use two independent random variable ξ_l and ξ_l' to determine whether or not there is a jump at each transition l;

i.e., a jump from j to j' occurs if after visiting state i, the system moves to state j according to ξ_l and P, but it moves to state j' according to ξ'_l and P'; and

ii. we generate the auxiliary paths by using the same sequence of random variables as the perturbed path, e.g., we generate $W - C$ by using the same sequence as that used for generating the perturbed path $G - D$.

Solution: On the sample path $A - B - E - D$, we may use two independent random variables ξ_l and ξ'_l to determine whether or not there is a jump at each transition. If after visiting state i, the system moves to state j according to ξ_l and P, but it moves to state j' according to ξ'_l and P', a jump from j to j' occurs.

When we use the same random variable ξ_l to determine a jump, the probability of a jump from u to v after visiting i is $p(u, v|i)$, with $\sum_{u,v\in\mathcal{S}} p(u, v|i) = 1$. When we use two independent random variables ξ_l and ξ'_l, the probability of a jump from u to v after visiting i is $p(u|i)p(v|i)$, with $\sum_{u,v\in\mathcal{S}} p(u|i)p(v|i) = 1$. In fact, in the latter case, we have $p(u, v|i) = p(u|i)p(v|i)$. On this basis, even if we use two independent random variables ξ_l and ξ'_l to determine the jump, this does not affect the results in Section 9.2.

For the generation of the auxiliary paths, we use the same sequence of random variables as the perturbed path. In fact, this is a coupling in realization factors, which can reduce the variance of the estimation of realization factors and does not affect the value of realization factors. We can refer to Section 3.1.3.

9.2 For two ergodic transition probability matrices P and P', set $P(\delta) := P + \delta(P' - P)$. Assume that δ is very small. Apply the construction approach described in Section 9.2 by following a sample path of the Markov chains with $P(\delta)$. Show that this is equivalent to the performance derivative construction described in Section 2.1.3 . (In Section 9.2, we follow the perturbed sample path, while in Section 2.1.3 , we follow the original path.)

Solution: When δ is very small, the transition probability matrices P and $P(\delta)$ are very close. Thus, the transition according to $P(\delta)$ is the same as the transition according to P in most cases. Following the construction described in Section 9.2 , we start from a sample path, X_δ, of the Markov chains with $P(\delta)$. When the transitions according to $P(\delta)$ and P are different, a jump is generated, for example, states u_1 and v_1 at time 4

in Figure 9.1 . After the jump, two sample path are generated according to P_δ and P, respectively. Since P_δ and P are very close, the two sample path will generally merge before a new jump is generated. For example, in Figure 9.1, X_δ and X merge at time 7. Then at time 9, a new jump is generated. We can also follow the sample path X and generate similar perturbations in Figure 9.1 by using the construction method in section 2.1.3. Therefore, the two methods are equivalent.

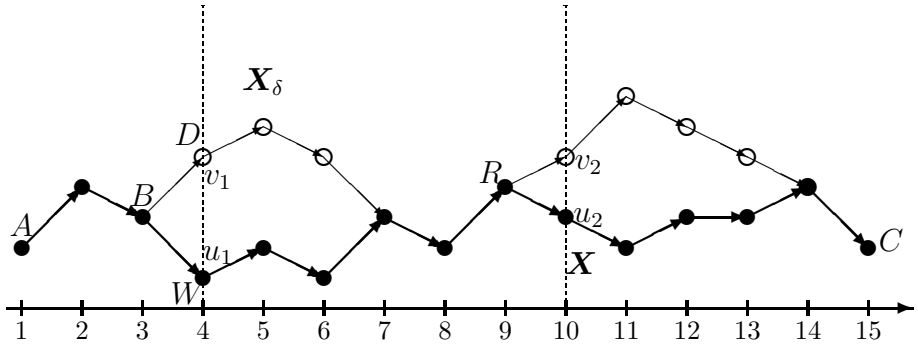

Figure 9.1: The Effect of Two Perturbations

9.3 Suppose that the transition probability matrices of all the policies in an MDP problem are uni-chains on the same finite state space $\mathcal{S}$. (A uni-chain is a special case of a multi-chain defined in (B.1) with $m = 1$.)

a. Apply the construction approach shown in Section 9.2 to any two uni-chain policies and derive the performance difference formula. Show that it is a special case of the performance difference formula (4.36) in Chapter 4 for the multi-chain case.

b. Derive the Poisson equation for a uni-chain policy, prove that its solution exists, and express the potentials of the transient states in terms of those of the recurrent states.

c. Develop the policy iteration algorithm for uni-chain MDPs, and show that it is the same as that for ergodic chains.

d. Explain point (c) using the policy iteration algorithm for the general case of multi-chain MDPs.

Solution:

a. We consider two uni-chains with transition probability matrices P^h and P^d and the same state space $\mathcal{S} = \{1, 2, \cdots, S'\}$. We assume $\{1, 2, \cdots, S\}$ are the recurrent states under policy h. For a uni-chain, we know the long run average performance $\eta^h(i)$ is independent of the initial state, i.e. $\eta^h(i) = \eta^h$. Applying the construction approach shown in Section 9.2, we can obtain the performance difference formula similarly to (9.3)-(9.5).

$$\eta^h - \eta^d = \sum_{i=1}^{S} \pi^h(i) \Big\{ \sum_{j=1}^{S'} [p^h(j|i) - p^d(j|i)] g^d(j) + f^h(i) - f^d(i) \Big\}. \tag{9.1}$$

For a uni-chain,

$$(P^h)^* = \begin{bmatrix} e_S \pi^h & 0 \\ e_{S'-S} \pi^h & 0 \end{bmatrix},$$

where e_S denotes a S-dimensional row vector in which all components are 1, $\pi^h = (\pi^h(1), \cdots, \pi^h(S))$ and $\pi^h(i)$ is the steady-state probability of state i under policy h. Putting $(P^h)^*$ and $\eta^d = \eta^d e$ into (4.36), we have

$$\eta^h - \eta^d = \sum_{i=1}^{S} \pi^h(i) \Big\{ \sum_{j=1}^{S'} [p^h(j|i) - p^d(j|i)] g^d(j) + f^h(i) - f^d(i) \Big\}.$$

Thus, (9.1) is a special case of (4.36) in Chapter 4 for the multi-chain case.

b. For a uni-chain, we assume $\{1, 2, \cdots, S\}$ are the recurrent states. We have the Poisson equation

$$\Big(I_{S'} - \begin{bmatrix} P & 0 \\ R_1 & R_2 \end{bmatrix} + \begin{bmatrix} e_S \pi & 0 \\ e_{S'-S} \pi & 0 \end{bmatrix} \Big) \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix},$$

where I'_S is a S'-dimensional unit matrix and $\pi = (\pi(1), \cdots, \pi(S))$. Thus, we have

$$(I_S + P + e_S \pi) g_1 = f_1$$

$$(e_{S'-S} \pi - R_1) g_1 + (I_{S'-S} - R_2) g_2 = f_2.$$

Since $I_S + P + e_S \pi$ and $I_{S'-S} - R_2$ are invertible, we have

$$g_1 = (I_S + P + e_S \pi)^{-1} f_1, \tag{9.2}$$

$$g_2 = (I_{S'-S} - R_2)^{-1} [f_2 - (e_{S'-S} \pi - R_1) g_1]. \tag{9.3}$$

c. Policy Iteration Algorithm:

1. Guess an initial policy d_0, set $k = 0$.

2. Obtain the potential g^{d_k} by using (9.2) and (9.3).

3. Choose

$$d_{k+1} \in arg \max_{d \in \mathcal{D}} \{f^d + P^d g^{d_k}\}.$$

component-wisely (i.e., to determine an action for each state). If at a state i, action $d_k(i)$ attains the maximum, then set $d_{k+1}(i) = d_k(i)$.

4. If $d_{k+1} = d_k$, stop; otherwise set $k := k + 1$ and go to step 2.

This algorithm is the same as that for ergodic chains.

d. From the point view of policy iteration for the general multi-chain MDPs, since $\eta^d = \eta e$ is independent of the initial state, then we have $P^h \eta^d = \eta^d$. Therefore, from comparison lemma (4.41), we should choose actions for all state as that in the step 3.

9.4 Prove that the policy iteration algorithm developed in Example 9.2 converges to an optimal policy.

Solution: From the step 3 of the algorithm, if $(\alpha_{i_{k+1}}, \beta_{i_{k+1}}) \neq (\alpha_{i_k}, \beta_{i_k})$, then

$$\begin{aligned} & p_0 \frac{\alpha_{i_{k+1}} - \alpha_{i_k}}{1 - \alpha_{i_{k+1}}} [g(0) - g(1)] + p_N \frac{\beta_{i_{k+1}} - \beta_{i_k}}{1 - \beta_{i_{k+1}}} [g(N) - g(N-1)] \\ > \; & p_0 \frac{\alpha_{i_k} - \alpha_{i_k}}{1 - \alpha_{i_k}} [g(0) - g(1)] + p_N \frac{\beta_{i_k} - \beta_{i_k}}{1 - \beta_{i_k}} [g(N) - g(N-1)] = 0. \end{aligned}$$

Using (9.7), we have $\eta_{k+1} > \eta_k$, where η_{k+1} and η_k are the performances under control pairs $(\alpha_{i_{k+1}}, \beta_{i_{k+1}})$ and $(\alpha_{i_k}, \beta_{i_k})$, respectively. That is, the average reward increases at each iteration before it stops. Because the number of policies is finite, the iteration procedure has to stop after a finite number of iterations. When it stops at step k, we set $(\widehat{\alpha}, \widehat{\beta}) := (\alpha_{i_{k+1}}, \beta_{i_{k+1}}) = (\alpha_{i_k}, \beta_{i_k})$. From the step 3 of the algorithm, we have

$$(\widehat{\alpha}, \widehat{\beta}) = arg \max_{(\alpha,\beta) \in \{(\alpha_i, \beta_i), i=1,2,\cdots,M\}} \left\{ p_0 \frac{\alpha - \widehat{\alpha}}{1 - \alpha} [g(0) - g(1)] + p_N \frac{\beta - \widehat{\beta}}{1 - \beta} [g(N) - g(N-1)] \right\}.$$

Thus, for any control pairs (α, β), from (9.7), we have $\eta - \widehat{\eta} \leq 0$, where η and $\widehat{\eta}$ are the performances under control pairs (α, β) and $(\widehat{\alpha}, \widehat{\beta})$, respectively. That is to say, for any

policy (α, β), it performance η is less than $\widehat{\eta}$. Therefore, the policy $(\widehat{\alpha}, \widehat{\beta})$ is the optimal policy.

9.5 In this exercise, we modify the random walk problem studied in Examples 9.1 and 9.2 as follows. First we simplify the problem by assuming that the random walker can take only $N + 1 = 5$ positions denoted as $0, 1, 2, 3$, and 4. When the walker hits the wall 0 or 4, s/he stays there with probability α_0, or α_4, respectively, and jumps to position 1, or 3, with probability $1 - \alpha_0$, or $1 - \alpha_4$, respectively. Second, we assume that when the walker is at position 1, 2, or 3, s/he will also stays there with probability α_1, α_2, and α_3, respectively, and leaves the position with probability $1 - \alpha_1$, $1 - \alpha_2$, and $1 - \alpha_3$, respectively. If s/he leaves position i, $i = 1, 2, 3$, s/he will have an equal probability of 0.5 to jump to one of its neighboring position $i - 1$ or $i + 1$, $i = 1, 2, 3$.

Now suppose that at each position i we may choose α_i from a finite set denoted as $\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,M}$, $i = 0, 1, \cdots, 4$.

a. Derive the performance difference formula (similar to (9.6)) and the policy iteration algorithm for this problem.

b. Furthermore, we assume that $\alpha_{0,i}$ and $\alpha_{4,i}$ (with the same i), $i = 1, 2, \cdots, M$, have to be chosen together, and $\alpha_{1,i}, \alpha_{2,i}$, and $\alpha_{3,i}$ (with the same i), $i = 1, 2, \cdots, M$, have to be chosen together. Derive a performance difference formula (similar to (9.7)) for this problem.

c. Based on the performance difference formula derived in (b), develop a policy iteration algorithm for the optimization problem in which actions at different states cannot be chosen independently.

Solution:

a. We consider the Markov chain X' under policy $(\alpha'_0, \alpha'_1, \alpha'_2, \alpha'_3, \alpha'_4)$ (without loss of generality, we assume $\alpha'_i > \alpha_i, i = 0, 1, 2, 3, 4$). At state 0, X' may jump from state 1 to 0 with probability $\alpha'_0 - \alpha_0$, and at state 4, it may jump from state 3 to 4 with probability $\alpha'_4 - \alpha_4$. Moreover, at state $i, i = 1, 2, 3$, it may jump from state Δ_i to i with probability $\alpha'_i - \alpha_i$, Δ_i is a stochastic state, which is $i - 1$ with probability $1/2$ and $i + 1$ with

probability 1/2. Thus, the potential $g(\Delta_i) = \frac{1}{2}[g(i-1)+g(i+1)]$. Then, by construction, we have

$$\begin{aligned}\eta' - \eta &= \pi'(0)\Big\{[\alpha_0' - \alpha_0][g(0) - g(1)]\Big\} + \pi'(4)\Big\{[\alpha_4' - \alpha_4][g(4) - g(3)]\Big\} + \\ &\quad \sum_{i=1,2,3} \pi'(i)\Big\{[\alpha_i' - \alpha_i][g(i) - g(\Delta_i)]\Big\}.\end{aligned}$$

Based on the above difference formula, similar to the method in Chapter 4, we obtain the following policy iteration algorithm.

1. Guess an initial policy $d_0 = (\alpha_0^0, \alpha_1^0, \alpha_2^0, \alpha_3^0, \alpha_4^0)$, set $k = 0$;

2. Obtain the potential g^{d_k} by solving the Poisson equation $(I - P^{d_k})g^{d_k} + \eta^{d_k} e = f^{d_k}$, or by estimation on a sample path of the system under policy d_k.

3. Choose d_{k+1} such that

$$\begin{aligned}\alpha_0^{k+1} &\in arg \max_{\alpha_0 \in \{\alpha_{0j}, j=1,2,\cdots,M\}} [\alpha_0 - \alpha_0^k][g^{d_k}(0) - g^{d_k}(1)], \\ \alpha_4^{k+1} &\in arg \max_{\alpha_4 \in \{\alpha_{4j}, j=1,2,\cdots,M\}} [\alpha_4 - \alpha_4^k][g^{d_k}(4) - g^{d_k}(3)], \\ \alpha_i^{k+1} &\in arg \max_{\alpha_i \in \{\alpha_{ij}, j=1,2,\cdots,M\}} [\alpha_i - \alpha_i^k][g^{d_k}(i) - g^{d_k}(\Delta_i)], i = 1,2,3,\end{aligned}$$

4. If $d_{k+1} = d_k$, stop; otherwise set $k := k+1$ and go to step 2.

b. Similar to Example 9.2, we have

$$\begin{aligned}\eta' - \eta &= \pi'(0,4)\kappa_1 \left\{ p_0 \frac{\alpha_0' - \alpha_0}{1 - \alpha_0'}[g(0) - g(1)] + p_4 \frac{\alpha_4' - \alpha_4}{1 - \alpha_4'}[g(4) - g(3)] \right\} + \\ &\quad + \pi'(1,2,3)\kappa_2 \left\{ p_1 \frac{\alpha_1' - \alpha_1}{1 - \alpha_1'}[g(1) - g(\Delta_1)] + p_2 \frac{\alpha_2' - \alpha_2}{1 - \alpha_2'}[g(2) - g(\Delta_2)] + \right. \\ &\quad \left. p_3 \frac{\alpha_3' - \alpha_3}{1 - \alpha_3'}[g(1) - g(\Delta_3)] \right\},\end{aligned}$$

where $\pi'(0,4) = \pi'(0) + \pi'(4), \pi'(1,2,3) = \pi'(1) + \pi'(2) + \pi'(3), \kappa_1 > 0, \kappa_2 > 0$ and p_0, p_1, p_2, p_3, p_4 is similar to p_0, p_N in Example 9.2.

c. The policy iteration can be designed similarly to Example 9.2.

9.6 Study the random walk problem in Example 9.3 by using the system with $N+2$ positions as the original system and the system with $N+1$ positions as the perturbed one. Derive the performance difference formula similar to (9.32).

Solution: Suppose that in Example 9.3 the number of positions of the random walker decreases from N + 2 to N + 1. According to (9.32), we need to determine $(\Delta P)'_-$. Comparing P and P', we can find that $[P, 0]$ and $[P'_1, P'_{1,2}]$ differ only on the last rows. Thus, $(\Delta P)'_-$ is zero everywhere expect its last row is

$$(0, \cdots, 0, 1 - \beta - \sigma_N, \beta, \sigma_N - 1).$$

From the difference formula (9.32), we have the following difference formula:

$$\begin{aligned} & \eta - \eta' \\ = \; & \pi(N)[(1 - \beta - \sigma_N)g'(N-1) + \beta g'(N) - (1 - \sigma_N)g'(N+1)]. \end{aligned}$$

9.7 Extend the performance derivative formulas (9.29) and (9.33) to the case with $f(i) \neq f'(i)$, $i = 1, \cdots, S$.

Solution: From the difference formula (9.28), the performance difference formula with $f(i) \neq f'(i), i = 1, \cdots, S$ is

$$\eta' - \eta = \pi'_-\{(P'_- - [P, 0])\tilde{g} + (f'_- - f)\}. \tag{9.4}$$

where $P'_- = [P'_1, P'_{12}]$ and $f'_- = (f'(1), \cdots, f'(S))^T$. For performance derivatives, we define

$$P_\delta = \tilde{P} + \delta[P' - \tilde{P}], f_\delta = \tilde{f} + \delta[f' - \tilde{f}]$$

where $\tilde{P}$ was defined as (9.14) and $\tilde{f} = (f(1), \cdots, f(S), f'(S+1), \cdots, f'(S'))^T$. Applying (9.4) to P_δ and P, we obtain $\eta_\delta - \eta = \pi_{\delta-}\delta[\Delta P_- \tilde{g} + h_-]$, where $\pi_{\delta-} = (\pi_\delta(1), \cdots, \pi_\delta(S)), \Delta P_- = [P'_1, P'_{12}] - [P, 0]$ and $h_- = f'_- - f$. Letting $\delta \to 0$, we get

$$\frac{d\eta_\delta}{d\delta} = \pi[\Delta P_- \tilde{g} + h_-].$$

Similarly, we can obtain the performance derivative formulas with $f(i) \neq f'(i)$, $i = 1, \cdots, S$ for (9.33),

$$\frac{d\eta_\delta}{d\delta} = \pi'_-[(\Delta P)'_- g' + h'_-],$$

where $(\Delta P)'_- = -\Delta P_-$ and $h'_- = -h_-$.

9.8 In Section 9.4.2, suppose $f'(i) \neq f(i)$, for $i = 1, \cdots, M$. Modify the performance difference formula (9.37) (i.e., derive the formula similar to (9.5) and (9.28).

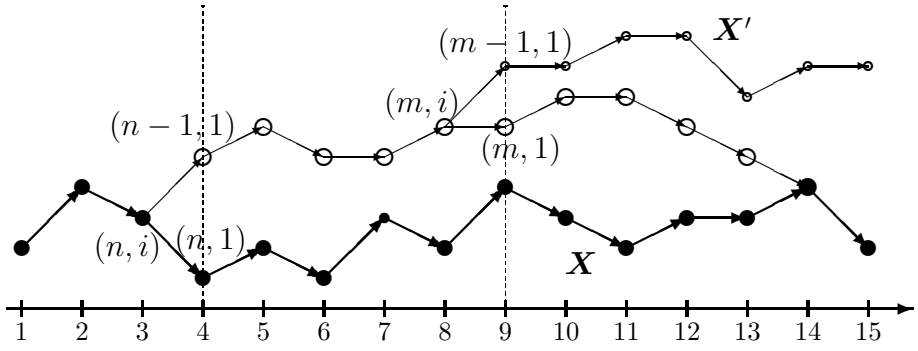

Figure 9.2: The jumps of the parameterized system in Section 9.3.2

Solution: If $f'(i) \neq f(i)$, for $i = 1, \cdots, M$, we should consider the effect of one-step performance for the performance difference $\eta' - \eta$. Similar to (9.37), we follow the sample path X' with $L, L >> 1$ transitions, which has transition probability matrix P'. Considering the effect of one-step performance, we have

$$E(F'_L - F_L) \approx \sum_{i \in \mathcal{S}_0} \left\{ \sum_{u \in \mathcal{S}} \sum_{v \in \mathcal{S}'} L\pi'(i)[p(u,v|i)\tilde{\gamma}(u,v) + f'(i) - f(i)] \right\},$$

where $\gamma(\tilde{u}, v) = \tilde{g}(v) - \tilde{g}(u)$. Following the same argument as (9.37), we can obtain the following difference formula:

$$\eta' - \eta = \pi'_-[\Delta P_- \tilde{g} + (f'_- - f_-)],$$

where $\Delta P_- = [0, P'_0, P'_{0-1}] - [P_{01}, P_0, 0]$, $f'_- = (f'(S_0), \cdots, f(1))^T$, $f_- = (f(S_0), \cdots, f(1))^T$.

9.9 Draw a sample-path diagram to illustrate the effect of one jump in the example of the parameterized system in Section 9.3.2.

Solution:

A sample path of the parameterized system in Section 9.3.2 is as Figure 9.2. At state (n, i), the customer will prepare to leave M_1 with probability $p_i, i = 1, 2, 3$ with $p_3 = 1$. After that, it will move back to M_1 with probability $1 - \theta$, which means the state transits to $(n, 1)$, and go to M_2 with probability θ, which means the state transits to state $(n-1, 1)$. If $\theta(n)$ change to $\theta(n) + \delta_n$, then this change in the system parameter may cause "jumps" of the system state on the sample path from $(n, 1)$ to $(n-1, 1)$ (the original sample path transits to state $(n, 1)$ but the perturbed path transits to state $(n-1, 1)$), for example,

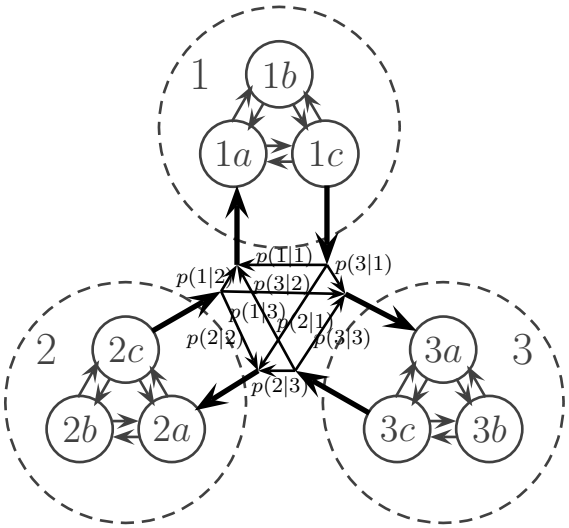

Figure 9.3: The Transition Probabilities in Problem **??**

at time 4 and time 9, the jumps occur.

9.10 Consider a discrete-time Markov chain consisting of three super states denoted as 1, 2, and 3, respectively; each of them is further composed of three phases a, b, and c, as shown in Figure 9.3. Each phase represents a state of the Markov chain and thus it has altogether 9 states denoted as $1a$, $1b$, $1c$; $2a$, $2b$, $2c$; and $3a$, $3b$, and $3c$. The transition probabilities between any two phases in the same super state are denoted by $p(1b|1a)$, $p(3a|3c)$ etc. When the system leaves a phase, it does not feed back immediately, i.e., $p(1a|1a) = 0$, etc. At each super state, phase a is an input phase, i.e., the system enters phase a to start its journey in the corresponding super state. Phase c is an exit phase, i.e, the system leaves a super state from phase c. At super state 1, for example, we have $p(1b|1a) + p(1c|1a) = 1$ and $p(1a|1b) + p(1c|1b) = 1$. At phase $1c$, there is a positive probability $p(0|1c)$ to leave the super state 1. Thus, $p(1a|1c) + p(1b|1c) + p(0|1c) = 1$. When a system leaves a super state i, $i = 1, 2, 3$, it transits to super state j, or enters phase ja, $j = 1, 2, 3$, with probability $p(j|i)$, $\sum_{j=1}^{3} p(j|i) = 1$. The reward function is denoted as $f(1a)$, $f(1b)$, etc.

Suppose that the transition probabilities $p(j|i)$ depend on a parameter θ and are denoted as $p_\theta(j|i)$, $i, j = 1, 2, 3$. Construct the performance derivative and difference formulas for this system, similar to (9.12) and (9.13).

Solution: Following the same procedure as in Section 9.2, we consider a perturbed sample path $\boldsymbol{X}'$ with super-state transition probabilities $p_{\theta'}(j|i)$ for $L >> 1$ transitions. Let $\pi'(ic), i = 1, 2, 3$ be the steady-state probability that state is in state ic under super-state transition probabilities $p_{\theta'}(j|i)$. A jump can occur only when the system stays at states $ic, i = 1, 2, 3$. Suppose that after visiting state ic, $\boldsymbol{X}'$ has a jump from ua to va, $u, v = 1, 2, 3$. Denote the probability of a jump from ua to va after visiting ic as $p(u, v|i)$. Then, $\sum_{u=1}^{3} p(u, v|i) = p(0|ic)p_{\theta'}(v|i)$ and $\sum_{v=1}^{3} p(u, v|i) = p(0|ic)p_{\theta}(u|i)$. On the average, on the sample path there are $L\pi'(ic)p(u, v|i)$ jumps from ua to va that happen after visiting ic. Since each jump has on the average an effect of $\gamma(ua, va)$ on F_L, on the average the total effect on F_L due to the change from $p_\theta(j|i)$ to $p_{\theta'}(j|i)$ is

$$\begin{aligned} E(F'_L - F_L) &\approx \sum_{i=1}^{M}\{\sum_{u,v=1}^{M} L\pi'(ic)p(u, v|i)\gamma(ua, va)\} \\ &= \sum_{i=1}^{M}\{\sum_{u,v=1}^{M} L\pi'(ic)p(u, v|i)[g(va) - g(ua)]\} \\ &= \sum_{i=1}^{3}\Big\{L\pi'(ic)p(0|ic)\{\sum_{j=1}^{3}[p_{\theta'}(j|i) - p_\theta(j|i)]g(ja)\}\Big\}. \end{aligned}$$

Finally, we have

$$\eta' - \eta = \lim_{L\to\infty} \frac{1}{L} E(F'_L - F_L) = \sum_{i=1}^{3}\Big\{\pi'(ic)p(0|ic)\{\sum_{j=1}^{3}[p_{\theta'}(j|i) - p_\theta(j|i)]g(ja)\}\Big\}$$

Letting $\theta' \to \theta$, we have the performance derivative

$$\frac{d\eta}{d\theta} = \sum_{i=1}^{3} \pi'(ic)p(0|ic)\{\sum_{j=1}^{3} \frac{dp_\theta(j|i)}{d\theta} g(ja)\}.$$

9.11 Consider a discrete-time M/M/1/N queue with capacity N. The system state is the number of customers in the system (in the queue plus in the server), denoted as n. The transition probabilities are $p(1|0) = p$, $p(0|0) = q$, $p(N - 1|N) = q$, $p(N|N) = p$, and $p(n + 1|n) = p$, $p(n - 1|n) = q$, $p > 0$, $q > 0$, $p + q = 1$. Suppose that

a. the capacity changes to $N - 1$, or

b. the capacity changes to $N + 1$.

Construct the difference formula for the mean response time.

Solution: For the discrete-time $M/M/1/N$ system, the mean response time is

$$\tilde{\eta} = \lim_{L\to\infty} \frac{\sum_{l=0}^{L-1} n_l}{K},$$

where n_l denotes the number of customers in the system and K denotes the number of customers that have been served until time L. It is a customer-average performance. From the long-run point of view, we have $K = L(1-\pi(0))q = L\left(1 - \frac{1-\frac{p}{q}}{1-(\frac{p}{q})^{N+1}}\right) q = \frac{p\left(1-\left(\frac{p}{q}\right)^N\right)}{1-\left(\frac{p}{q}\right)^{N+1}} L$. Therefore,

$$\tilde{\eta} = \lim_{L\to\infty} \frac{L}{K} \frac{\sum_{l=0}^{L-1} n_l}{L} = \frac{1-\left(\frac{p}{q}\right)^{N+1}}{p\left(1-\left(\frac{p}{q}\right)^{N}\right)} \eta = h(N)\eta.$$

where $h(N) = \frac{1-\left(\frac{p}{q}\right)^{N+1}}{p\left(1-\left(\frac{p}{q}\right)^{N}\right)}$ and $\eta = \lim_{L\to\infty} \frac{\sum_{l=0}^{L-1} n_l}{L}$ is a time-average performance. For time-average performance η, we can apply the construction approach in Section 9.4 to obtain the difference formula.

a. We consider the case that the capacity changes to $N-1$. We assume the potentials of M/M/1/N system with capacity N are $g(i), i = 0, 1, \ldots, N$. Moreover, the transition probability matrix of discrete-time M/M/1/N system with capacity $N-1$ is

$$P' = \begin{bmatrix} q & p & 0 & 0 & \cdots & 0 & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & q & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & q & p \end{bmatrix} \tag{9.5}$$

and its steady-state probability and time-average performance are $\pi' = (\pi'(0), \cdots, \pi'(N-1))$ and η', respectively. The transition probability matrix P for the M/M/1/N queue has the same form as (9.5) except that its size is larger by one. Let η' be the average performance of P'. From the performance difference formula (9.32), we have

$$\eta' - \eta = \pi'(N-1)[pg(N-1) - pg(N)].$$

Thus, the performance difference formula of the mean response time is

$$\begin{aligned}\tilde{\eta}' - \tilde{\eta} &= h(N-1)\eta' - h(N)\eta \\ &= h(N-1)\eta' - h(N-1)\eta + h(N-1)\eta - h(N)\eta \\ &= h(N-1)(\eta' - \eta) + (h(N-1) - h(N))\eta \\ &= h(N-1)\pi'(N-1)[pg(N-1) - pg(N)] + (h(N-1) - h(N))\eta.\end{aligned}$$

b. We consider the case that the capacity changes to $N+1$. We assume its time-average performance and steady-state probability are η' and π', respectively. Comparing P and P', we can construct $\tilde{P}$ in (9.14). Indeed, we have

$$P'_{21} = [0, 0, \cdots, q]$$

and $P'_{22} = p$. Therefore, from (9.25), we have

$$\tilde{g}(N+1) = \frac{1}{1-p}[N + 1 - \eta + qg(N)].$$

From the performance difference formula (9.27), we have

$$\eta' - \eta = \pi'(N)[-pg(N) + p\tilde{g}(N+1)].$$

Thus, the performance difference formula of the mean response time is

$$\begin{aligned}\tilde{\eta}' - \tilde{\eta} &= h(N+1)\eta' - h(N)\eta \\ &= h(N+1)\eta' - h(N+1)\eta + h(N+1)\eta - h(N)\eta \\ &= h(N+1)\pi'(N)[-pg(N) + p\tilde{g}(N+1)] + [h(N+1) - h(N)]\eta.\end{aligned}$$

9.12 Suppose that we have two independent M/M/1/N queues with parameters p_1, q_1, N_1 and p_2, q_2, and N_2 respectively, as explained in Problem 9.11. If we have one more buffer space, to which queue should we allocate this extra buffer space to maximally reduce the customers' mean response time? Please develop an on-line approach.

Remark: Because the mean response time is

$$\tilde{\tau} = \frac{\bar{n}}{p(1 - \pi(N))},$$

where $\bar{n}$ denotes the average queue length and $\pi(N)$ denotes the steady-state probability that the system in state N, it will increase if the buffer space

becomes larger. Thus, for this problem, we should consider the increment of the mean response time. Here, we only describe the idea to solve this problem.

Solution: If we have another buffer space N_3, we can allocate this extra buffer space to every queue and obtain two independent M/M/1/N queues with parameters p_1, q_1, $N_1 + N_3$ and p_2, q_2, and $N_2 + N_3$. We assume the mean response times of the M/M/1/N queues with parameters p_1, q_1, N_1 and p_1, q_1, $N_1 + N_3$ are $\tilde{\eta}_1'$ and $\tilde{\eta}_1$, respectively, and the mean response times of the M/M/1/N queues with parameters p_2, q_2, N_2 and p_2, q_2, $N_2 + N_3$ are $\tilde{\eta}_2'$ and $\tilde{\eta}_2$, respectively. Then, we need to compare $\tilde{\eta}_1' - \tilde{\eta}_1$ and $\tilde{\eta}_2' - \tilde{\eta}_2$ to determine which queue we allocate this extra buffer space to. Similarly to part b) in Problem 9.11, we have the difference formulas:

$$\tilde{\eta}_1' - \tilde{\eta}_1 = h(N_1 + N_3)\pi'(N_1)[-p_1 g(N_1) + p_1\tilde{g}(N_1 + 1)] + [h(N_1 + N_3) - h(N_1)]\eta_1,$$
$$\tilde{\eta}_2' - \tilde{\eta}_2 = h(N_2 + N_3)\pi'(N_2)[-p_2 g(N_2) + p_2\tilde{g}(N_2 + 1)] + [h(N_2 + N_3) - h(N_2)]\eta_2,$$

where $\tilde{g}(N_1 + 1) = \frac{1}{1-p_1}[N_1 + 1 - \eta_1 + q_1 g(N_1)]$ and $\tilde{g}(N_2 + 1) = \frac{1}{1-p_2}[N_2 + 1 - \eta_2 + q_2 g(N_2)]$. We can estimate $\eta_1, g(N_1)$ and $\eta_2, g(N_2)$ based on the sample paths of the queues with parameters p_1, q_1 and N_1 and p_2, q_2 and N_2, respectively. Computing $h(N_1 + N_3), h(N_1), \pi'(N_1), \tilde{g}(N_1 + 1)$ and $h(N_2 + N_3), h(N_2), \pi'(N_2), \tilde{g}(N_2 + 1)$, where $\pi'(N_1) = \frac{(1-p_1/q_1)\left(\frac{p_1}{q_1}\right)^{N_1}}{1-\left(\frac{p_1}{q_1}\right)^{N_1+N_3}}$ and $\pi'(N_2) = \frac{(1-p_2/q_2)\left(\frac{p_2}{q_2}\right)^{N_2}}{1-\left(\frac{p_2}{q_2}\right)^{N_2+N_3}}$, we can obtain the values of $\tilde{\eta}_1' - \tilde{\eta}_1$ and $\tilde{\eta}_2' - \tilde{\eta}_2$ and compare them.

9.13* Extend the construction approach in Section 9.2 to (continuous-time) Markov processes. *(Hint: This extension is not as straightforward as what it may appear. To develop a construction approach to the changes in transition probabilities of the embedded Markov chains $p(j|i)$ in (A.12) may be easy; the extension to the changes in transition rate $\lambda(i)$ may be more involved.)*

Solution: Consider an ergodic Markov process $\boldsymbol{X} = \{X_t, t \geq 0\}$ with a finite state space $\mathcal{S} = \{1, 2, \cdots, S\}$ and an infinitesimal generator $B = [b(i, j)]$, where

$$b(i,j) = \begin{cases} -\lambda(i) & if\ i = j \\ \lambda(i)p(j|i) & if\ i \neq j \end{cases}$$

for all $i, j \in \mathcal{S}$. As we know, Markov process X stays at state i for an exponentially distributed period with distribution $F(t) = 1 - exp(-\lambda(i)t)$ and then transits to state j

with probability $p(j|i)$. Let $X_l, l = 0, 1, \cdots$, be the embedded Markov chain and $T_l(i)$ be the holding time at state $X_l = i$. The holding time $T_l(i)$ can be simulated by using the inverse transform method. That is,

$$T_l(i) = -\frac{1}{\lambda(i)} \ln(1 - \xi_l), i \in \mathcal{S}, l = 0, 1, 2, \cdots, \tag{9.6}$$

where ξ_l is a uniformly distributed random variable on $[0, 1)$. The transitions of states can be simulated as (2.3).

We consider another perturbed Markov chain $\boldsymbol{X}'$ with transition rate $\lambda'(i)$ and transition probability $p'(j|i)$. Thus, its infinitesimal generator is $B' = [b'(i, j)]$, where

$$b'(i,j) = \begin{cases} -\lambda'(i) & if\ i = j \\ \lambda'(i)p'(j|i) & if\ i \neq j \end{cases}$$

We firstly consider the effect of a perturbation of transition rate from $\lambda(i)$ to $\lambda'(i)$ at one stage. We follow the perturbed sample path of Markov process $\boldsymbol{X}'$. At state $X_0 = i$, by using the transition rate $\lambda'(i)$ and $\lambda(i)$, respectively, and the same ξ_0 in (9.6), we have different holding time T_0' and T_0. We assume $\lambda'(i) > \lambda(i)$, then $T_0'(i) < T_0(i)$. The perturbation from $\lambda(i)$ to $\lambda'(i)$ results in the change of holding time at state i, $\Delta T_0(i) := T_0(i) - T_0'(i)$. From (9.6), we have

$$\Delta T_0(i) = \frac{\lambda'(i) - \lambda(i)}{\lambda'(i)} T_0(i). \tag{9.7}$$

The effect on $F_T = E\{\int_{t_0}^{T} f(X_t)dt\}$ due to this perturbation in the holding time is

$$\begin{aligned} \Delta_i \ := \ & E\Big\{ \int_{t_0}^{T_0'(i)} f(X_t')dt + \int_{T_0'(i)}^{T-\Delta T_0(i)} f(X_t')dt + \int_{T-\Delta T_0(i)}^{T} f(X_t')dt \Big\} \\ & -E\Big\{ \int_{t_0}^{T_0'(i)} f(X_t)dt + \int_{T_0'(i)}^{T_0(i)} f(X_t)dt + \int_{T_0(i)}^{T} f(X_t)dt \Big\}. \end{aligned}$$

In the right side of the above equation, the first and second items in the first bracket are equal to the first and third items in the second bracket, respectively, thus,

$$\Delta_i = E\{ \int_{T-\Delta T_0(i)}^{T} f(X_t')dt - \int_{T_0'(i)}^{T_0(i)} f(X_t)dt.$$

When T is large enough, for $t \in [T - \Delta T_0(i), T]$, we have $E[f(X_t')] \approx \pi f = \eta$, thus,

$$\Delta_i \approx E[\Delta T_0(i)](\eta - f(i)).$$

From (9.7), we have

$$\Delta_i \approx \frac{\lambda^{'}(i) - \lambda(i)}{\lambda^{'}(i)} E(T_0(i))(\eta - f(i)).$$

From Poisson equation $Bg = -f + \eta e$, we have $\eta - f(i) = \lambda(i)[\sum_{j \in \mathcal{S}} p(j|i)g(j) - g(i)]$. Since $E[T_0(i)] = \frac{1}{\lambda(i)}$, we have

$$\begin{aligned} \Delta_i &\approx \frac{\lambda'(i) - \lambda(i)}{\lambda'(i)} \sum_{j \in \mathcal{S}} p(j|i)g(j) - g(i) \\ &= \frac{\lambda'(i) - \lambda(i)}{\lambda'(i)} \sum_{j \in \mathcal{S}} p(j|i)\gamma(i,j). \end{aligned} \tag{9.8}$$

Now, we consider the effect of all these perturbations at different states and all stages. Let $\pi'(i)$ denote the steady-state probability that $\boldsymbol{X}'$ is at state $i \in \mathcal{S}$. During the time interval $[0, T]$, the time that the perturbed process $\boldsymbol{X}'$ stays at state i is $T\pi'(i)$ on the average. Since the mean holding time is $\frac{1}{\lambda'(i)}$, then, there are on the average $T\pi'(i)\lambda'(i)$ transitions from state i. Each of them has an effect as (9.8) on F_T on the average. Then the total effect on F_T due to all the perturbations in the holding times is

$$\begin{aligned} & E[F'_T - F_T] \\ = & \sum_{i=1}^{S} T\pi'(i)\lambda'(i) \frac{\lambda'(i) - \lambda(i)}{\lambda'(i)} \sum_{j \in \mathcal{S}} p(j|i)\gamma(i,j) \\ = & T\pi'(\Lambda' - \Lambda)[P - I]g, \end{aligned}$$

where $\Lambda' = diag\{\lambda'(1), \cdots, \lambda'(S)\}$ and $\Lambda = diag\{\lambda(1), \cdots, \lambda(S)\}$. Dividing by T on both sides of the above equation and letting $T \to \infty$, we have the following performance difference formula

$$\eta' - \eta = \pi'(\Lambda' - \Lambda)[P - I]g. \tag{9.9}$$

From (9.9), let $\lambda'(i) \to \lambda(i)$ and $\lambda'(j) = \lambda(j), j \neq i$, we can obtain performance derivative formula

$$\frac{d\eta}{d\lambda(i)} = \pi(i)\{\sum_{j \in \mathcal{S}} p(j|i)g(j) - g(i)\}.$$

Next, we consider the effect of perturbation in the transition probabilities. After visiting state i, $\boldsymbol{X}$ transits to state u based on the transition probabilities $p(j|i), i, j \in \mathcal{S}$,

while $\boldsymbol{X}'$ transits to v according to $p'(j|i), i, j \in \mathcal{S}$. Define the probability that following visiting state i such jumps happen from state u to state v as $p(u, v|i)$, $i, u, v \in \mathcal{S}$. Since there are on the average $T\pi'(i)\lambda'(i)$ transitions from state i, then on the average, on the sample path there are $T\pi'(i)\lambda'(i)p(u, v|i)$ jumps from u to v that happen after visiting i. Since each jump has on the average an effect of $\gamma(i, j)$ on F_L, on the average the total effect on F_L due to the change from $p(j|i)$ to $p'(j|i), i, j \in \mathcal{S}$ is

$$\begin{aligned} & E(F'_T - F_T) \\ \approx \; & T \sum_{i=1}^{S} \pi'(i)\lambda'(i) \sum_{u,v \in \mathcal{S}} p(u, v|i)\gamma(u, v) \\ = \; & T \sum_{i=1}^{S} \pi'(i)\lambda'(i) \sum_{j=1}^{S} [p'(j|i) - p(j|i)]g(j). \end{aligned}$$

Dividing by T on both sides of the above equation and letting $T \to \infty$, we have the performance difference formula under two different transition probabilities,

$$\eta' - \eta = \pi'\Lambda'(P' - P)g.$$

If we consider the perturbations of transition rates and transition probabilities simultaneously, we can decompose these perturbations into perturbations of transition rates and perturbations of transition probabilities, then we have

$$\begin{aligned} \eta' - \eta &= \pi'(\Lambda' - \Lambda)[P - I]g + \pi'\Lambda'(P' - P)g \\ &= \pi'[\Lambda'(P' - I) - \Lambda(P - I)]g \\ &= \pi'[B' - B]g. \end{aligned}$$

9.14* Propose a construction approach for the performance differences and derivatives for a (continuous-time) closed Jackson (Gordon-Newell) network (Section C.2) with respect to the changes in routing probabilities. *(Hint: Use the results in Problem 9.13 for the transition probability matrix of the embedded chain.)*

Solution: We consider a closed Jackson (or Gordon-Newell) network. There are N customers circulating among M servers according to routing probabilities $q_{i,j}$, with $\sum_{j=1}^{M} q_{i,j} = 1, i = 1, 2, \cdots, M$. Let n_k denote the number of customers at server $k, k = 1, 2, \cdots, M$. The state of the network can be denoted by $\boldsymbol{n} = (n_1, n_2, \cdots, n_M)$. Viewing the network

as a continuous time Markov process, the effective service rate is $\mu(\boldsymbol{n}) = \sum_{i=1}^{M} \epsilon(n_i)\mu_{i,n_i}$, where $\epsilon(n_i) = 1$, if $n_i > 0$, otherwise $\epsilon(n_i) = 0$. The probability that a customer completing the service at server i transits to server $j \neq i$ with probability $\frac{\epsilon(n_i)\mu_{i,n_i}q_{ij}}{\sum_{i=1}^{M}\epsilon(n_i)\mu_{i,n_i}}$ and transits to itself with probability q_{ii}.

We consider a sample path of Jackson network with perturbed routing probabilities $q'_{i,j}$ on time interval $[0, T]$. On the sample path, the time that the system stays at state $\boldsymbol{n}$ is $T\pi'(\boldsymbol{n})$ on the average. Since the average time that the system stays at state $\boldsymbol{n}$ is $\frac{1}{\mu(\boldsymbol{n})}$, there are $T\pi'(\boldsymbol{n})\mu(\boldsymbol{n})$ times that the system transits from state $\boldsymbol{n}$. After visiting state $\boldsymbol{n}$, the system transits to state $\boldsymbol{u}$ based on the original routing probabilities $q_{i,j}$, while it transits to state $\boldsymbol{v}$ based on the routing probabilities $q'_{i,j}$. Let $b(\boldsymbol{n}, \boldsymbol{u}, \boldsymbol{v})$ be the probability that such jump will happen from $\boldsymbol{u}$ to state $\boldsymbol{v}$ at state $\boldsymbol{n}$. Similarly to the argument in Problem 9.14, we have

$$\begin{aligned} E\{F'_T - F_T\} &= \sum_{\boldsymbol{n}} T\pi'(\boldsymbol{n})\mu(\boldsymbol{n}) \sum_{\boldsymbol{u},\boldsymbol{v}} b(\boldsymbol{n}, \boldsymbol{u}, \boldsymbol{v})\gamma(\boldsymbol{u}, \boldsymbol{v}) \\ &= \sum_{\boldsymbol{n}} T\pi'(\boldsymbol{n})\mu(\boldsymbol{n}) \sum_{\boldsymbol{u}} [p'(\boldsymbol{u}|\boldsymbol{n}) - p(\boldsymbol{u}|\boldsymbol{n})]g(\boldsymbol{u}) \\ &= \sum_{\boldsymbol{n}} T\pi'(\boldsymbol{n}) \Big\{ \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \epsilon(n_i)\mu_{i,n_i}[q'_{i,j} - q_{i,j}]g(\boldsymbol{n}_{i,j}) + \sum_{i=1}^{M} \epsilon(n_i)\mu_{i,n_i}[q'_{i,i} - q_{i,i}]g(\boldsymbol{n}) \Big\}. \end{aligned}$$

Dividing by T on both sides of the above equation and letting $T \to \infty$, we have

$$\eta' - \eta = \sum_{\boldsymbol{n}} \pi'(\boldsymbol{n}) \Big\{ \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \epsilon(n_i)\mu_{i,n_i}[q'_{i,j} - q_{i,j}]g(\boldsymbol{n}_{i,j}) + \sum_{i=1}^{M} \epsilon(n_i)\mu_{i,n_i}[q'_{i,i} - q_{i,i}]g(\boldsymbol{n}) \Big\}.$$

If let $Q^\delta = Q + \delta(Q' - Q)$, we have

$$\eta^\delta - \eta = \delta \sum_{\boldsymbol{n}} \pi^\delta(\boldsymbol{n}) \Big\{ \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \epsilon(n_i)\mu_{i,n_i}[q'_{i,j} - q_{i,j}]g(\boldsymbol{n}_{i,j}) + \sum_{i=1}^{M} \epsilon(n_i)\mu_{i,n_i}[q'_{i,i} - q_{i,i}]g(\boldsymbol{n}) \Big\}.$$

Thus, dividing by δ on both sides and letting $\delta \to 0$, we have the performance derivative formula:

$$\frac{d\eta^\delta}{d\delta} = \sum_{\boldsymbol{n}} \pi(\boldsymbol{n}) \Big\{ \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \epsilon(n_i)\mu_{i,n_i}[q'_{i,j} - q_{i,j}]g(\boldsymbol{n}_{i,j}) + \sum_{i=1}^{M} \epsilon(n_i)\mu_{i,n_i}[q'_{i,i} - q_{i,i}]g(\boldsymbol{n}) \Big\}.$$

Solutions to the Appendix

Appendix A

A.1 Consider the Coxian distribution shown in Figure A.1.

a. Derive the probability distribution density function for the Coxian distribution.

b. Derive the Laplace transform of the density function.

c. Construct a Coxian distribution such that the Laplace transform of its density function is the rational function given below:

$$F(s) = \frac{2 + 1.08s + 0.2s^2}{2 + 5s + 4s^2 + s^3}. \tag{9.10}$$

[**Solution**] a. Suppose that the Coxian distribution has k stages. See Figure A.1. And the service rate in stage i is $\lambda_i = \frac{1}{s_i}$. Denote the probability distribution density function for the Coxian distribution as $g(x)$ and denote the probability distribution density function for the exponential distribution with parameter λ_i as $f_i(x) = \lambda_i e^{-\lambda_i x}$. Then,

$$\begin{aligned} g(x) &= q_1 f_1(x) + p_1 q_2 f_1(x) * f_2(x) + p_1 p_2 q_3 f_1(x) * f_2(x) * f_3(x) \\ &\quad + \cdots + p_1 p_2 \cdots p_{k-2} q_{k-1} f_1(x) * f_2(x) * \cdots * f_{k-1}(x) \\ &\quad + p_1 p_2 \cdots p_{k-1} f_1(x) * f_2(x) * \cdots * f_k(x) \\ &= \sum_{j=1}^{k} \prod_{l=1}^{j-1} p_l q_j f_1(x) * f_2(x) * \cdots * f_j(x), \end{aligned}$$

where "$*$" denotes the convolution.

b. First, we consider the Laplace transform of $f_i(x)$. Denote $F_i(s)$ the Laplace transform of $f_i(x)$.

$$\begin{aligned} F_i(s) &= \int_0^\infty f_i(x)e^{-sx}dx = \int_0^\infty \lambda_i e^{-\lambda_i x}e^{-sx}dx \\ &= \int_0^\infty \lambda_i e^{-(\lambda_i+s)x}dx = \frac{\lambda_i}{\lambda_i+s} \end{aligned}$$

Denote $G(s)$ the Laplace transform of $g(x)$.

$$\begin{aligned} G(s) &= q_1F_1(s) + p_1q_2F_1(s)F_2(s) + p_1p_2q_3F_1(s)F_2(s)F_3(s) \\ &\quad + \cdots + p_1p_2\cdots p_{k-2}q_{k-1}F_1(s)F_2(s)\cdots F_{k-1}(s) \\ &\quad + p_1p_2\cdots p_{k-1}F_1(s)F_2(s)\cdots F_k(s) \\ &= q_1\frac{\lambda_1}{\lambda_1+s} + p_1q_2\frac{\lambda_1\lambda_2}{(\lambda_1+s)(\lambda_2+s)} + p_1p_2q_3\frac{\lambda_1\lambda_2\lambda_3}{(\lambda_1+s)(\lambda_2+s)(\lambda_3+s)} \\ &\quad + \cdots + p_1p_2\cdots p_{k-2}q_{k-1}\frac{\prod_{i=1}^{k-1}\lambda_i}{\prod_{i=1}^{k-1}(\lambda_i+s)} \\ &\quad + p_1p_2\cdots p_{k-1}\frac{\prod_{i=1}^{k}\lambda_i}{\prod_{i=1}^{k}(\lambda_i+s)} \\ &= \sum_{j=1}^{k}\prod_{l=1}^{j-1}p_lq_j\frac{\prod_{i=1}^{j}\lambda_i}{\prod_{i=1}^{j}(\lambda_i+s)} \\ &= \frac{\sum_{j=1}^{k}\prod_{l=1}^{j-1}p_lq_j\prod_{i=1}^{j}\lambda_i\prod_{m=j+1}^{k}(\lambda_m+s)}{\prod_{i=1}^{k}(\lambda_i+s)} \\ &= \frac{\sum_{j=1}^{k}\prod_{l=1}^{j-1}p_l(1-p_j)\prod_{i=1}^{j}\lambda_i\prod_{m=j+1}^{k}(\lambda_m+s)}{\prod_{i=1}^{k}(\lambda_i+s)} \end{aligned} \tag{9.11}$$

c. From part b), we may firstly find λ_i, $i=1,2,3$ such that $\prod_{i=1}^{3}(\lambda_i+s) = 2+5s+4s^2+s^3 =: f(s)$. It is easy to find $f(-1)=0$. So, we know that $s+1$ is a factor of $f(s)$. By using the division with residue, we have $f(s) = (s+1)(s^2+3s+2) = (s+1)(s+1)(s+2)$. Here, we can make different choices about the values of λ_i, $i=1,2,3$ and different choices may correspond to different Coxian distributions. For example, we may choose $\lambda_1=1, \lambda_2=2$ and $\lambda_3=1$. Putting them into the numerator in (9.11) and letting the numerator be equal to $2+1.08s+0.2s^2$, $p_1=0.8, p_2=0.7$ can be obtained and $p_3=1$ is obvious. It is noted that the Coxian distribution corresponding to (9.10) is not unique since the

different choices of $\lambda_i, i = 1, 2, 3$.

A.2 Consider an independent random sequence X_n with $\mathcal{P}(X_n = 1) = \frac{1}{n}$ and $\mathcal{P}(X_n = 0) = 1 - \frac{1}{n}$. Does the sequence converge in probability, w.p.1, in mean, or in mean square? **[Solution]**

The random sequence $\{X_n\}$ converges in probability to a random variable X, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathcal{P}[|X_n - X| \geq \epsilon] = 0.$$

Obviously, we can see that $X = 0$.

$$\lim_{n\to\infty} \mathcal{P}[|X_n| \geq \epsilon] = \lim_{n\to\infty} \mathcal{P}(X_n = 1) = \lim_{n\to\infty} \frac{1}{n} = 0.$$

Therefore, the sequence converges in probability.

Next, we show that this sequence does not converges to zero with probability 1. To establish that fact, we assume that the convergence with probability 1 holds true and then obtain the contradiction. If the convergence with probability 1 holds true, then

$$\lim_{n\to\infty} \mathcal{P}(\sup_{k\geq n} X_k = 1) = 0.$$

Notice that $\{\sup_{k\geq n} X_k = 1\} = \bigcup_{k\geq n}\{X_k = 1\}$. Hence, taking into consideration the fact that $\{X_n\}$ is the sequence of independent random variables and

$$\begin{aligned}\lim_{n\to\infty} \mathcal{P}(\bigcup_{k\geq n}\{X_k = 1\}) &= \lim_{n\to\infty}\left\{1 - \mathcal{P}(\bigcap_{k\geq n}\{X_k = 0\})\right\} \\ &= 1 - \lim_{n\to\infty}\prod_{k\geq n}\mathcal{P}(X_k = 0) = 1 - \lim_{n\to\infty}\prod_{k\geq n}(1 - \frac{1}{k}) \equiv 1,\end{aligned}$$

we arrive at announced contradiction. Therefore, the sequence $\{X_n\}$ may not converge with probability one.

$$\lim_{n\to\infty} E[|X_n - X|] = \lim_{n\to\infty} E[X_n] = \lim_{n\to\infty} 1 \times \mathcal{P}(X_n = 1) = \lim_{n\to\infty} \frac{1}{n} = 0$$

and

$$\lim_{n\to\infty} E[|X_n - X|^2] = \lim_{n\to\infty} E[|X_n|^2] = \lim_{n\to\infty} 1^2 \times \mathcal{P}(X_n = 1) = \lim_{n\to\infty} \frac{1}{n} = 0,$$

so, the sequence converges in mean and in mean square.

A.3 Consider a random sequence X_n with $\mathcal{P}(X_n = 1) = \frac{1}{n^2}$ and $\mathcal{P}(X_n = 0) = 1 - \frac{1}{n^2}$. Does the sequence converge in probability, w.p.1, in mean, or in mean square?

[**Solution**]

It is obvious that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathcal{P}[|X_n| \geq \epsilon] = \lim_{n\to\infty} \mathcal{P}(X_n = 1) = \lim_{n\to\infty} \frac{1}{n^2} = 0.$$

Therefore, the sequence converges in probability to zero.

$\{X_n\}$ converges with probability 1 to a random variable X, if

$$\mathcal{P}(\omega : \lim_{n\to\infty} X_n = X) = 1,$$

or equivalently, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathcal{P}(\sup_{k\geq n} |X_k - X| \geq \epsilon) = 0,$$

where $\{\sup_{k\geq n} |X_k - X| \geq \epsilon\} = \{\bigcup_{k\geq n} |X_k - X| \geq \epsilon\} = \{|X_k - X| \geq \epsilon \quad \text{for some} \quad k \geq n\}$.

For every $\epsilon > 0$, we have

$$\begin{aligned}
& \lim_{n\to\infty} \mathcal{P}(\sup_{k\geq n} |X_k| \geq \epsilon) \\
\leq & \lim_{n\to\infty} \left[\sum_{k=n}^{\infty} \mathcal{P}(|X_k| \geq \epsilon) \right] \\
= & \lim_{n\to\infty} \left[\sum_{k=n}^{\infty} \mathcal{P}(X_k = 1) \right] \\
= & \lim_{n\to\infty} \left[\sum_{k=n}^{\infty} \frac{1}{k^2} \right] = 0.
\end{aligned}$$

Therefore, the sequence $\{X_n\}$ converges with probability 1 to zero, of course converge in probability.

$$\lim_{n\to\infty} E[|X_n - X|] = \lim_{n\to\infty} E[X_n] = \lim_{n\to\infty} 1 \times \mathcal{P}(X_n = 1) = \lim_{n\to\infty} \frac{1}{n^2} = 0$$

and

$$\lim_{n\to\infty} E[|X_n - X|^2] = \lim_{n\to\infty} E[|X_n|^2] = \lim_{n\to\infty} 1^2 \times \mathcal{P}(X_n = 1) = \lim_{n\to\infty} \frac{1}{n^2} = 0,$$

so, the sequence converges in mean and in mean square to zero.

A.4 Let X and Y be two random variables with probability distributions $\Phi(x)$ and $\Psi(y)$, respectively. Their means are denoted as $\bar{x} = E(X)$ and $\bar{y} = E(Y)$. We wish to estimate $\bar{x} - \bar{y} = E(X - Y)$ by simulation. We generate random variables X and Y using the inverse transformation method. Thus, we have $X = \Phi^{-1}(\xi_1)$ and $Y = \Psi^{-1}(\xi_2)$, where ξ_1 and ξ_2 are two uniformly distributed random variables in $[0, 1)$. Prove that if we choose $\xi_1 = \xi_2$, then the variance of $X - Y$, $Var[X - Y]$, is the smallest among all possible pairs of ξ_1 and ξ_2.

[Solution]

$$\begin{aligned} Var[X - Y] &= E[((X - Y) - E(X - Y))^2] = E[(X - Y)^2] - (E[X - Y])^2 \\ &= E[X^2] + E[Y^2] - 2E[XY] - (E[X - Y])^2 \end{aligned}$$

For given distribution $\Phi(x)$ and $\Psi(y)$, $E[X^2]$, $E[Y^2]$ and $E[X-Y]$ are determined. Thus, to minimize $Var[X - Y]$ is equivalent to maximize $E[XY]$.

Denote $H(x, y) = P(X \leq x, Y \leq y)$. We have

$$H(x, y) = P(X \leq x, Y \leq y) \leq P(X \leq x) = \Phi(x)$$

and similarly, $H(x, y) \leq \Psi(y)$. Therefore, $H(x, y) \leq \Phi(x) \wedge \Psi(y)$. We know $X = \Phi^{-1}(\xi_1)$ and $Y = \Psi^{-1}(\xi_2)$, and Φ, Ψ both are non-decreasing functions. We have

$$H(x, y) = P(\Phi^{-1}(\xi_1) \leq x, \Psi^{-1}(\xi_2) \leq y) = P(\xi_1 \leq \Phi(x), \xi_2 \leq \Psi(y))$$

If $\xi_1 = \xi_2$, we have $H(x, y) = P(\xi_1 \leq \Phi(x) \wedge \Psi(y))$. Because ξ_1 is uniformly distributed on $[0, 1)$, then $H(x, y) = \Phi(x) \wedge \Psi(y)$. That means, if $\xi_1 = \xi_2$, then $H(x, y)$ reaches its maximum. From Hoeffding's Lemma in the reference,

$$E(XY) - E(X)E(Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [H(x, y) - \Phi(x)\Psi(y)]\mathrm{d}x\mathrm{d}y$$

$E(X), E(Y), \Phi(x), \Psi(y)$ are all determined. Thus if $H(x, y)$ reaches its maximum, then $E(XY)$ reaches its maximum. Therefore, if we choose $\xi_1 = \xi_2$, the variance $Var[X - Y]$ is the smallest among all possible pairs of ξ_2 and ξ_1.

Proof of Hoeffding's Lemma:

Let (X_1, Y_1), (X_2, Y_2) be independent, each distributed according to $H(x, y)$. Then

$$\begin{aligned} & 2[\mathrm{E}(X_1Y_1) - \mathrm{E}(X_1)\mathrm{E}(Y_1)] = \mathrm{E}[(X_1 - X_2)(Y_1 - Y_2)] \\ = \; & \mathrm{E}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[I(u, X_1) - I(u, X_2)][I(v, Y_1) - I(v, Y_2)]\mathrm{d}u\mathrm{d}v \end{aligned}$$

where $I(u, x) = 1$ if $u \leq x$ and $= 0$ otherwise. Since $E(XY), E(X)$ and $E(Y)$ are finite, we can take expectation under the integral sign, then above equation becomes

$$\mathrm{E}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[I(u, X_1) - I(u, X_2)][I(v, Y_1) - I(v, Y_2)]\mathrm{d}u\mathrm{d}v = 2\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[H(x, y) - \Phi(x)\Psi(y)]\mathrm{d}x\mathrm{d}y$$

This completes the proof.

Reference: Lehmann E.L., "Some concepts of dependence," *Ann. Math. Statist.*, vol. 37, pp. 1137–1153, 1966.

A.5 Consider a sequence of independent and identically distributed random variables $\{X_n, n = 1, 2, \cdots\}$ with mean $E(X_n) = E(X)$. Define another sequence of $0 - 1$ valued independent and identically distributed random variables $\{\chi_n, n = 1, 2, \cdots\}$ where $\chi_n = 1$ with probability $1 > p > 0$ and $\chi_n = 0$ with probability $1 - p$. Let

$$N_n = \sum_{k=1}^{n} \chi_k$$

be the number of 1's in the first n samples. Define

$$M_n := \frac{1}{N_n}\sum_{k=1}^{n}(\chi_k X_k).$$

Prove M_n converges to $E(X)$ with probability 1 as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} M_n = E(X), \qquad w.p.1,$$

and M_n converges to $E(X)$ in probability as $n \to \infty$, i.e., for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathcal{P}[|M_n - E(X)| \geq \epsilon] = 0.$$

[Solution]

$$M_n = \frac{1}{N_n}\sum_{k=1}^{n}(\chi_k X_k) = \frac{n}{\sum_{k=1}^{n}\chi_k} * \frac{1}{n}\sum_{k=1}^{n}(\chi_k X_k).$$

According to the strong law of large numbers, $\frac{\sum_{k=1}^{n}\chi_k}{n}$ converges to p with probability 1 and $\frac{1}{n}\sum_{k=1}^{n}(\chi_k X_k)$ converges to $E[\chi_k X_k] = pE(X)$ with probability 1. So, M_n converges to $E(X)$ with probability 1.

From the property of convergence with probability 1, the convergence in probability can be easily obtained.

A.6 Consider a sequence of independent random variables $\{X_n, n = 1, 2, \cdots\}$. The mean value of X_n, $E(X_n)$, converges to a constant $\bar{X}$, $\lim_{n\to\infty} E(X_n) = \bar{X}$, and $Var(X_n) < \infty$. Prove that the mean sample $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ converges to $\bar{X}$ both with probability 1 and in probability .

[Solution]

Denote $a_n = E(X_n)$. We have $\lim_{n\to\infty} a_n = \lim_{n\to\infty} E(X_n) = \bar{X}$. Let $Y_n = X_n - a_n$. Since a_n, $\bar{X}$ are constant and $\{X_n, n = 1, 2, \cdots\}$ are independent random variables, we know that $\{Y_n, n = 1, 2, \cdots\}$ are also independent random variables and $E(Y_n) = 0$. From the strong law of large numbers, we know that $\frac{\sum_{k=1}^{n} Y_k}{n} \longrightarrow 0$ with probability 1 and in probability. Since $\frac{\sum_{k=1}^{n} Y_k}{n} = \frac{\sum_{k=1}^{n} X_k}{n} - \frac{\sum_{k=1}^{n} a_k}{n}$, $\lim_{n\to\infty} a_n = \bar{X}$ and $\lim_{n\to\infty} \frac{\sum_{k=1}^{n} a_k}{n} = \bar{X}$, $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ converges both with probability 1 and in probability to $\bar{X}$.

A.7 Let $\boldsymbol{X}$ be an irreducible but periodic Markov chain with transition probability matrix P. The asymptotic stationarity (A.8) does not hold. However, we may define $\pi(i)$ as the time average

$$\pi(i) = \lim_{L\to\infty} \frac{1}{L} E\{\sum_{l=0}^{L-1} \chi_i(X_l) | X_0 = j\}, \quad i, j \in \mathcal{S}, \tag{A.17}$$

with $\chi_i(x) = 1$, if $x = i$, and $\chi(i) = 0$, otherwise. Prove

a. Prove that the $\pi(i)$ in (A.17) indeed does not depend on j.

b. Let $\pi = (\pi(1), \cdots, \pi(S))$, then

$$P^* := \lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} P^l = e\pi.$$

c. $\pi P = \pi$, and $\pi e = 1$. That is, the time average π plays the same role as the steady-state probability.

d. Prove $\lim_{L\to\infty} \frac{1}{L}\{\sum_{l=0}^{L-1} \chi_i(X_l)\}$, $i \in \mathcal{S}$, converges with probability 1 to $\pi(i)$. Therefore, $\pi(i)$ can also be defined as the limit of the sample-path average of $\chi_i(X_l)$, $l = 0, 1, \cdots$.

[Solution] a. Let $f^k(j|i)$ be the probability that the Markov chain transits firstly to state j from initial state i at time k. Since the Markov chain is irreducible and periodic, we have

$$\sum_{k=0}^{\infty} f^k(j|i) = P(\text{Markov chain transits to state } j \text{ early or late from initial state } i) = 1.$$

Moreover,

$$p^l(j|i) = \sum_{k=0}^{l} f^k(j|i)p^{l-k}(j|j).$$

Then,

$$\begin{aligned} \frac{1}{L}\sum_{l=0}^{L-1} p^l(j|i) &= \frac{1}{L}\sum_{l=0}^{L-1}\sum_{k=0}^{l} f^k(j|i)p^{l-k}(j|j) \\ &= \sum_{k=0}^{L-1} f^k(j|i)\frac{1}{L}\sum_{l=k}^{L-1} p^{l-k}(j|j). \end{aligned}$$

Let $L \to \infty$, we have

$$\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1} p^l(j|i) = \sum_{k=0}^{\infty} f^k(j|i)\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1} p^l(j|j) = \lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1} p^l(j|j).$$

Thus,

$$\pi(i) := \lim_{L\to\infty}\frac{1}{L}E\{\sum_{l=0}^{L-1}\chi_i(X_l)|X_0 = j\} = \lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1} p^l(i|j) \tag{9.12}$$

is independent of the initial state j.

b. From (9.12), we naturally have the following matrix form

$$P^* := \lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1} P^l = e\pi.$$

c.

$$\begin{aligned}
\sum_{i=1}^{S}\pi(i)P(k|i) &= \sum_{i=1}^{S}\lim_{L\to\infty}\frac{1}{L}E\{\sum_{l=0}^{L-1}\chi_i(X_l)\}P(k|i) \\
&= \lim_{L\to\infty}\frac{1}{L}\sum_{i=1}^{S}E\{\sum_{l=0}^{L-1}\chi_i(X_l)\}P(k|i) \\
&= \lim_{L\to\infty}\frac{1}{L}\sum_{i=1}^{S}E\{\sum_{l=0}^{L-1}\chi_i(X_l)P(k|i)\} \\
&= \lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}E\{\sum_{i=1}^{S}\chi_i(X_l)P(k|i)\} \\
&= \lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}E\{\chi_k(X_{l+1})\} = \pi(k).
\end{aligned}$$

$$\begin{aligned}
\pi e &= \sum_{i=1}^{S}\pi(i) \\
&= \sum_{i=1}^{S}\lim_{L\to\infty}\frac{1}{L}E\{\sum_{l=0}^{L-1}\chi_i(X_l)\} \\
&= \lim_{L\to\infty}\frac{1}{L}E\{\sum_{l=0}^{L-1}\sum_{i=1}^{S}\chi_i(X_l)\} = \lim_{L\to\infty}\frac{1}{L}E\{\sum_{l=0}^{L-1}1\} = 1.
\end{aligned}$$

d. Proof: Set

$$N_i(L) = \sum_{l=0}^{L-1}\chi_i(X_l).$$

Let us fix a reference state i and define the r-th passage time to state i as

$$T_i(r) = \inf\{l \geq T_i(r-1)+1 : X_l = i\},$$

where $T_i(0) = 0$. Suppose the period is d for P, then,

$$Y_i(r) := T_i(r) - T_i(r-1) = nd < \infty,$$

and $Y_i(2), Y_i(3), \cdots$ are independently and identically distributed with mean μ_i. Now note that

$$Y_i(1) + Y_i(2) + \cdots + Y_i(N_i(L)-1) \leq L-1,$$

the left side being the time of last visit to i before time n. Also,

$$Y_i(1) + Y_i(2) + \cdots + Y_i(N_i(L)) \geq L.$$

Then we have

$$\frac{\sum_{r=1}^{N_i(L)-1} Y_i(r)}{N_i(L)} < \frac{L}{N_i(L)} \leq \frac{\sum_{r=1}^{N_i(L)} Y_i(r)}{N_i(L)}. \tag{9.13}$$

By using the strong law of large number, we have

$$\frac{1}{L}\sum_{r=1}^{L} Y_i(r) \to \mu_i, \quad w.p.1.$$

and also, since P is recurrent, we have

$$N_i(L) \to \infty, \quad \text{as} \quad L \to \infty \quad \text{with probability 1.}$$

So, letting $L \to \infty$ in (9.13), we can prove $\frac{N_i(L)}{L}$ converges to $\frac{1}{\mu_i} = \pi(i)$ with probability 1. Therefore, $\pi(i)$ can be also defined as the limit of the sample-path average of $\chi_i(X_l), l = 0, 1, \cdots$.

A.8 *(Uniformization)* Consider a Markov process $\boldsymbol{X}$ with transition rates $\lambda(i)$, $i \in \mathcal{S} = \{1, 2, \cdots, S\}$. Let $P = [p(j|i)]$ be the transition probability matrix of the embedded Markov chain, with $p(i|i) = 0$. Define another Markov process $\boldsymbol{X}'$ as follows: the transition rate at state i changes to $\lambda'(i) = \frac{\lambda(i)}{1-c_i}$, where $c_i \in (0, 1)$ is a fixed number, $i \in \mathcal{S}$; the transition probabilities change to $p'(i|i) = c_i$ and $p'(j|i) = p(j|i)[1 - c_i]$, $i \neq j$.

1. Prove the steady-state probabilities of the both processes are equal; i.e., $\pi'(i) = \pi(i)$, $i \in \mathcal{S}$.

2. Explain the relation between the sample paths of both processes.

3. Find the values for c_i, $i \in \mathcal{S}$, such that the embedded Markov chain of $\boldsymbol{X}'$, $\boldsymbol{X}'^{\dagger}$, has the same steady-state probabilities as those of $\boldsymbol{X}'$ and $\boldsymbol{X}$; i.e., $\pi'^{\dagger}(i) = \pi'(i) = \pi(i)$, $i \in \mathcal{S}$.

[Solution]

1. Denote B as the infinitesimal generator of the Markov process $\boldsymbol{X}$. We have $B = \text{diag}(\lambda(1), \ldots, \lambda(S))(P - I)$.

From $p'(i|i) = c_i$, $p'(j|i) = (1 - c_i)p(j|i)$ for all $j \in \mathcal{S} - \{i\}$ and $\lambda'(i) = \frac{\lambda(i)}{1-c_i}$, we have

$$\begin{aligned} B' &= \text{diag}(\lambda'(1), \ldots, \lambda'(S))(P' - I) \\ &= \text{diag}(\lambda(1), \ldots, \lambda(S))\text{diag}(\frac{1}{1-c_1}, \ldots, \frac{1}{1-c_S})\text{diag}(1 - c_1, \ldots, 1 - c_S)(P - I) = B. \end{aligned}$$

Since the Markov process $\boldsymbol{X}$ has the same infinitesimal generator with the Markov process $\boldsymbol{X}'$, they must have the same steady-state probabilities, i.e., $\pi'(i) = \pi(i)$, $i \in \mathcal{S}$.

2. Since the Markov process $\boldsymbol{X}$ has the same infinitesimal generator with the Markov process $\boldsymbol{X}'$, they must have the same statistical behaviors. Compared with $\boldsymbol{X}$, since the transition probability is $p'(i|i) = c_i$, Markov process $\boldsymbol{X}'$ can transit back to state $i \in \mathcal{S}$ with probability c_i after it stays at state i for a time with exponential distribution. Thus, the times that the process stays at state i follows a geometric distribution with c_i. Moreover, the sojourn time at state i is exponential distribution with rate $\frac{\lambda(i)}{1-c_i}$. From the sample path, we cannot observe the state transits to itself, thus, the total sojourn time at state i follows the exponential distribution with mean $\frac{1}{1-c_i}\frac{1}{\frac{\lambda(i)}{1-c_i}} = \frac{1}{\lambda(i)}$. Therefore, these two processes have the same sample path statistically. The only difference is on the sample path of $\boldsymbol{X}'$ there are some points the state of Markov process transits to itself, which cannot be observed by the observer.

3. Since we know $P' = \text{diag}(1 - c_1, \ldots, 1 - c_S)P + \text{diag}(c_1, \ldots, c_S)$ and $\pi'^{\dagger}P' = \pi'^{\dagger}$,

$$\pi'^{\dagger}\text{diag}(1 - c_1, \ldots, 1 - c_S)P = \pi'^{\dagger}\text{diag}(1 - c_1, \ldots, 1 - c_S).$$

That is,

$$\pi'^{\dagger}\text{diag}(1 - c_1, \ldots, 1 - c_S)(P - I) = 0. \tag{9.14}$$

By $\pi B = 0$, we have

$$\pi\text{diag}(\lambda(1), \ldots, \lambda(S))(P - I) = 0. \tag{9.15}$$

Comparing (9.14) and (9.15), we get

$$\pi'^{\dagger}\text{diag}(1 - c_1, \ldots, 1 - c_S) = K\pi\text{diag}(\lambda(1), \ldots, \lambda(S)),$$

where K is a constant. That is, $\pi'^{\dagger}(i) = K\pi(i)\frac{\lambda(i)}{1-c_i}$, for all $i \in \mathcal{S}$. If $\pi'^{\dagger}(i) = \pi(i)$, for all $i \in \mathcal{S}$, we obtain $K\frac{\lambda(i)}{1-c_i} = 1$ for all i. Then $c_i = 1 - K\lambda(i)$, which also need to satisfy $0 < K < \frac{1}{\max \lambda(i)}$ since $0 < c_i < 1$ for all i. We also can get $\lambda'(i) = \frac{\lambda(i)}{1-c_i} = \frac{1}{K}$.

A.9 Let $\boldsymbol{X}^{\dagger}$ be the embedded Markov chain of Markov process $\boldsymbol{X}$. Assume $\boldsymbol{X}^{\dagger}$ is ergodic. Let $\lambda(i)$, $i \in \mathcal{S} = \{1, 2, \cdots, S\}$ be the transition rates of $\boldsymbol{X}$; and $\pi^{\dagger}(i)$, $\pi(i)$, $i \in \mathcal{S}$, be the steady-state probabilities of $\boldsymbol{X}^{\dagger}$ and $\boldsymbol{X}$, respectively. Prove

$$\pi(i) = c\frac{\pi^{\dagger}(i)}{\lambda(i)}$$

where

$$c = \sum_{i\in\mathcal{S}} \pi(i)\lambda(i) = \frac{1}{\sum_{i\in\mathcal{S}} \frac{\pi^{\dagger}(i)}{\lambda(i)}}.$$

[Solution]

We assume $P^{\dagger}$ is the transition probability matrix of the embedded Markov Chain X†. From the definition of infinitesimal generator, we know:

$$B = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_S)(P^{\dagger} - I).$$

Since $\pi^{\dagger} = \pi^{\dagger}P^{\dagger}$, we have

$$\pi^{\dagger}(P^{\dagger} - I) = 0. \tag{9.16}$$

Moreover, for Markov process $\boldsymbol{X}$ we have $\pi B = 0$ and $\pi e = 1$, which have the unique solution. Then, we get

$$\pi\text{diag}(\lambda_1, \lambda_2, ..., \lambda_S)(P^{\dagger} - I) = 0. \tag{9.17}$$

Comparing the aforementioned two equations (9.16) and (9.17), we obtain

$$\pi\text{diag}(\lambda_1, \lambda_2, ..., \lambda_S) = c\pi^{\dagger}.$$

Therefore, $\pi(i) = c\frac{\pi^{\dagger}(i)}{\lambda(i)}$.

Since $c\sum\limits_{i=1}^{S} \frac{\pi^{\dagger}(i)}{\lambda(i)} = \sum\limits_{i=1}^{S} \pi(i) = 1$ and $\sum\limits_{i=1}^{S} \pi(i)\lambda(i) = \sum\limits_{i=1}^{S} c\pi^{\dagger}(i) = c$,

$$c = \sum_{i=1}^{S} \pi(i)\lambda(i) = \frac{1}{\sum\limits_{i=1}^{S} \frac{\pi^{\dagger}(i)}{\lambda(i)}}.$$

A.10 Consider an ergodic Markov chain $\boldsymbol{X} = \{X_0, X_1, \cdots\}$ with transition probability matrix $P = [p(j|i)]$. Let π be the steady-state probability vector. Define a performance function that depends on two consecutive states: $f(i,j)$, $i, j \in \mathcal{S}$. Prove that the following ergodicity equation holds:

$$\lim_{n\to\infty}\{\frac{1}{L}\sum_{l=0}^{L-1} f(X_l, X_{l+1})\} = E_{\pi,P}[f(X_l, X_{l+1})]$$
$$:= \sum_{i=1}^{S}\sum_{j=1}^{S}\{f(i,j)\pi(i)p(j|i)\} = \sum_{i=1}^{S}[\bar{f}(i)\pi(i)], \qquad \text{w.p.1}, \tag{9.18}$$

where $\bar{f}(i) = \sum_{j=1}^{S}[f(i,j)p(j|i)]$. Extend this results to function $f(X_l, X_{l+1}, \cdots, X_{l+N})$ for a finite integer N.

[Solution]

We define $Z_l = (X_l, X_{l+1})$, then we can easily prove $\mathbf{Z} = \{Z_l, l = 0, 1, 2, \ldots\}$ is Markov chain. Since $\boldsymbol{X}$ is ergodic, $\mathbf{Z}$ is also ergodic. By using the ergodicity theorem for ergodic Markov chain, we have

$$\lim_{n\to\infty}\{\frac{1}{L}\sum_{l=0}^{L-1} f(X_l, X_{l+1})\} = E_{\pi,P}[f(X_l, X_{l+1})],$$

where $E_{\pi,P}$ is the steady-state expectation of Markov chain $\mathbf{Z}$. Because the steady state probability of $\mathbf{Z}$ is $\pi(i,j) = \pi(i)p(j|i), i, j \in \mathcal{S}$. Thus we have

$$\lim_{n\to\infty}\{\frac{1}{L}\sum_{l=0}^{L-1} f(X_l, X_{l+1})\} = E_{\pi,P}[f(X_l, X_{l+1})]$$
$$:= \sum_{i=1}^{S}\sum_{j=1}^{S}\{f(i,j)\pi(i)p(j|i)\} = \sum_{i=1}^{S}[\bar{f}(i)\pi(i)], \qquad \text{w.p.1}.$$

For the function $f(X_l, X_{l+1}, \cdots, X_{l+N})$, we can define $Z_l = \{X_l, X_{l+1}, \ldots, X_{l+N})$. Similarly, applying the ergodicity theorem, we can obtain

$$\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1} E\left[f(X_l, X_{l+1}, \ldots, X_{l+N})\right]$$
$$= E_{\pi,P}[f(X_l, X_{l+1}, \ldots, X_{l+N})]$$
$$= \sum_{i\in\mathcal{S}}\sum_{j_1\in\mathcal{S}}\cdots\sum_{j_N\in\mathcal{S}} f(i, j_1, \ldots, j_N)\pi(i)p(j_1|i)\Pi_{k=1}^{N-1}p(j_{k+1}|j_k)$$

$$= \sum_{i \in \mathcal{S}} \pi(i) \bar{f}(i),$$

where $\bar{f}(i) = \sum_{j_1 \in \mathcal{S}} \cdots \sum_{j_N \in \mathcal{S}} f(i, j_1, \ldots, j_N) p(j_1|i) \Pi_{k=1}^{N-1} p(j_{k+1}|j_k)$.

A.11 Prove that the sojourn time that a Markov process stays in a state i is exponentially distributed, using the Markov property (A.10).

[Solution]

Let $T_0 = 0$, T_1, T_2, $\cdots$, be the instants of transitions for the Markov process $\boldsymbol{X} = \{X_t\}$ and X_0, X_1, X_2, $\cdots$ be the successive states visited by $\boldsymbol{X}$. $T_{l+1} - T_l$ is called the *sojourn time* in state X_l. We assume that the sample paths are right-continuous, i.e., $X_l = X_{T_l+0}$, and $X_l = i$, then $T_{l+1} - T_l$ is the sojourn time in state i.

Next, to prove the result, we prove

$$\mathcal{P}\{T_{l+1} - T_l \geq t | X_l = i\} = exp(-\lambda(i)t),$$

where $\lambda(i)$ is the *transition rate* of Markov process at state i. Because of the right-continuous property, we have

$$\mathcal{P}\{T_{l+1} - T_l \geq t | X_l = i\} = \mathcal{P}\{X_u = i, T_l \leq u \leq T_l + t | X_l = i\}.$$

Firstly, set $B := \{X_u = i, T_l \leq u \leq T_l + t\} = \bigcap_{T_l \leq u \leq T_l + t} \{X_u = i\}$. Dividing $[0, t]$ into 2^n equal parts, set

$$A_n := \{X_{T_l + \frac{kt}{2^n}} = i, k = 0, 1, \ldots, 2^n\} = \bigcap_{k=0}^{2^n} \{X_{T_l + \frac{kt}{2^n}} = i\}.$$

Since $A_{n+1} \subset A_n$, set $A := lim_{n \to \infty} A_n$. Obviously, $B \subset A$. On the other hand, from the right-continuous property, we have $\mathcal{P}(A - B) = 0$, so,

$$\begin{aligned}
\mathcal{P}\{T_{l+1} - T_l \geq t | X_l = i\} &= \mathcal{P}\{X_u = i, T_l \leq u \leq T_l + t | X_l = i\} \\
&= \mathcal{P}\{B | X_l = i\} = \mathcal{P}\{A | X_l = i\} \\
&= \lim_{n \to \infty} \mathcal{P}\{A_n | X_l = i\} \\
&= \lim_{n \to \infty} \mathcal{P}\{X_{T_l + \frac{kt}{2^n}} = i, k = 0, 1, \ldots, 2^n | X_l = i\} \\
&= \lim_{n \to \infty} P_{ii}(t/2^n)^{2^n} (\text{Markov property})
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} exp\{2^n ln[P_{ii}(t/2^n)]\} \\
&= \lim_{n\to\infty} exp\Big\{\frac{ln[1-\lambda(i)t/2^n + o(t/2^n)]}{-\lambda(i)t/2^n}(-\lambda(i)t)\Big\} \\
&= exp(-\lambda(i)t).
\end{aligned}
$$

A.12 Is the following statement true?

If the inter-transition times of a semi-Markov process are exponentially distributed, i.e., if $\mathcal{P}[T_{l+1} - T_l \le t | X_l = i] = 1 - e^{-\lambda(i)t}$, $i \in \mathcal{S}$, *then the semi-Markov process is a Markov process.*

If your answer is "yes", prove it; if the answer is "no", explain why and give a counter example.

[Solution]

This argument is wrong.

From the definition of Semi-Markov process, we have

$$
\begin{aligned}
& \mathcal{P}[X_{l+1} = j, T_{l+1} - T_l \le t | X_0, \ldots, X_l = i; T_0, \ldots, T_l] \\
= \; & \mathcal{P}[X_{l+1} = j, T_{l+1} - T_l \le t | X_l = i] \\
= \; & \mathcal{P}[T_{l+1} - T_l \le t | X_l = i] \mathcal{P}[X_{l+1} = j | X_l = i, T_{l+1} - T_l \le t] \\
= \; & [1 - e^{-\lambda(i)t}] \mathcal{P}[X_{l+1} = j | X_l = i, T_{l+1} - T_l \le t].
\end{aligned}
$$

As we know, by the Markov process definition, we have

$$
\mathcal{P}[X_{l+1} = j, T_{l+1} - T_l \le t | X_0, \ldots, X_l = i; T_0, \ldots, T_l] = p(j|i)[1 - e^{-\lambda(i)t}].
$$

As we can see from these two definitions, the state transition probability in the Markov process $p(j|i)$ will have no relation with state sojourn time $T_{l+1} - T_l$. Therefore, although the state sojourn time $T_{l+1} - T_l$ of Semi-Markov process has memory-less property, we cannot assert it is Markov process. Here is a counter-example. Suppose the state transition probability $P(X_{l+1}|X_l)$ of Semi-Markov process is related to $T_{l+1} - T_l$. In this situation, although $T_{l+1} - T_l$ is memory-less, the Semi-Markov process is still not a Markov process. If we have further condition that $\mathcal{P}[X_{l+1} = j | X_l = i, T_{l+1} - T_l \le t] = p(j|i)$, then this

semi-Markov process is a Markov process.

Appendix B

B.1 In the canonical form (B.1), what if R_{m+1} may be further irreducible.

a. Write R_{m+1} in a canonical form, and

b. Explain the meaning of this canonical form in terms of the transitions of the transient states.

[Solution]

R_{m+1} may be further reduced to

$$R_{m+1} = \begin{bmatrix} Q_1 & 0 & 0 & \cdots & \cdot & 0 \\ 0 & Q_2 & 0 & \cdots & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & Q_c & 0 \\ T_1 & T_2 & T_3 & \cdots & T_c & T_{c+1} \end{bmatrix},$$

b. This canonical form means the transient states can be further divided into $c+1$ parts. The part corresponding to $Q_i, i = 1, 2, \ldots, c$, can only transit to itself. The $c+1$th part can transit to any part.

B.2 Derive a general form for the solution to (B.6) and (B.7).

[Solution]

Denote P as it's canonical form:

$$P = \begin{bmatrix} P_1 & 0 & 0 & \cdots & \cdot & 0 \\ 0 & P_2 & 0 & \cdots & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & P_m & 0 \\ R_1 & R_2 & R_3 & \cdots & R_m & R_{m+1} \end{bmatrix}. \tag{9.19}$$

Recall that (B.6) and (B.7) are respectively

$$P^*e = e, \tag{9.20}$$

and

$$P^*P = PP^* = P^*P^* = P^*. \tag{9.21}$$

Denote the solution to (9.20) and (9.21) as

$$P^* = \begin{bmatrix} P^*_{11} & P^*_{12} & P^*_{13} & \cdots & \cdot & P^*_{1(m+1)} \\ P^*_{21} & P^*_{22} & P^*_{23} & \cdots & \cdot & P^*_{2(m+1)} \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ P^*_{m1} & P^*_{m2} & P^*_{m3} & \cdots & P^*_{mm} & P^*_{m(m+1)} \\ P^*_{(m+1)1} & P^*_{(m+1)2} & P^*_{(m+1)3} & \cdots & P^*_{(m+1)m} & P^*_{(m+1)(m+1)} \end{bmatrix}, \tag{9.22}$$

Then from $PP^* = P^*$, we get

$$P_jP^*_{ji} = P^*_{ji}, \quad j = 1, 2, \cdots, m, i = 1, 2, \cdots, m+1, \tag{9.23}$$

$$\sum_{k=1}^{m+1} R_kP^*_{kl} = P^*_{(m+1)l}, \quad l = 1, 2, \cdots, m+1. \tag{9.24}$$

Since P_i, $i = 1, 2, \cdots$ are irreducible non-negative matrix, then it is well-known that 1 is the simple eigenvalue of P_i. Combining with $P_ie_i = e_i$ and (9.23), we know that

$$P^*_{ji} = [c_1(j,i)e_j, c_2(j,i)e_j, \cdots, c_{n_i}(j,i)e_j], \quad j = 1, 2, \cdots, m, i = 1, 2, \cdots, m+1, \tag{9.25}$$

and

$$P^*_{(m+1)l} = (I - R_{m+1})^{-1} \sum_{k=1}^{m} R_kP^*_{kl}, \quad l = 1, 2, \cdots m+1, \tag{9.26}$$

where $e_i = [1, 1, \cdots, 1]^T$ and it's dimension is the same as P_i, denoted by n_i and $c_k(j,i)$ is a constant scalar.

Noting that $P^*e = e$, we know that

$$\sum_{i=1}^{m+1} \sum_{l=1}^{n_i} c_l(j,i) = 1 \text{ for } j = 1, 2, \cdots, m.$$

Then from $P^*P = P^*$, we get

$$P^*_{ji}P_i + P^*_{j(m+1)}R_i = P^*_{ji}, \quad i = 1, 2, \cdots, m, j = 1, 2, \cdots, m+1, \tag{9.27}$$

$$P^*_{j(m+1)}R_{m+1} = P^*_{j(m+1)}, \quad j = 1, 2, \cdots, m+1, \tag{9.28}$$

By (9.28), we know $P^*_{j(m+1)} = 0$ noting $R_{m+1}e \preceq e$, $j = 1, 2, \cdots, m+1$. Then (9.27) becomes

$$P^*_{ji}P_i = P^*_{ji}, \quad i = 1, 2, \cdots, m, j = 1, 2, \cdots, m+1.$$

Combining with (9.25), we get that

$$P^*_{ji} = c(j, i)e_j\pi_i, \quad i = 1, 2, \cdots, m, j = 1, 2, \cdots, m+1, \tag{9.29}$$

where $0 \leq c(j, i) \leq 1$ is any constant and π_i is the steady state probability of P_i. That is, $c_l(j, i) = c(j, i)\pi_i(l)$. Noting that $P^*e = e$ and $c(j, m+1) = 0$, we know that

$$\sum_{i=1}^{m+1}\sum_{l=1}^{n_i} c(j, i)\pi_i(l) = \sum_{i=1}^{m+1} c(j, i)\sum_{l=1}^{n_i}\pi_i(l) = \sum_{i=1}^{m+1} c(j, i) = \sum_{i=1}^{m} c(j, i) = 1 \text{ for } j = 1, 2, \cdots, m.$$

Finally from $P^*P^* = P^*$, we obtain

$$\sum_{k=1}^{m+1} P^*_{jk}P^*_{ki} = P^*_{ji}, \quad i, j = 1, 2, \cdots, m+1. \tag{9.30}$$

Since we have proved that $P^*_{j(m+1)} = 0$ and $P^*_{ji} = c(j, i)e_j\pi_i$, we get

$$\sum_{k=1}^{m} c(j, k)c(k, i) = c(j, i), \quad i, j = 1, 2, \cdots, m,$$

and

$$\sum_{k=1}^{m} c(m+1, k)c(k, i) = c(m+1, i), \quad i = 1, 2, \cdots, m.$$

Combining (9.26) and (9.29), we know that

$$\sum_{k=1}^{m} R_k c(k, l)e_k\pi_l = (I - R_{m+1})c(m+1, l)e_{m+1}\pi_l, l = 1, 2, \cdots, m.$$

$$P^* = \begin{bmatrix} c(1,1)e_1\pi_1 & c(1,2)e_1\pi_2 & \cdot & \cdots & \cdot & c(1,m)e_1\pi_m & 0 \\ c(2,1)e_2\pi_1 & c(2,2)e_2\pi_2 & \cdot & \cdots & \cdot & c(2,m)e_2\pi_m & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \\ c(m,1)e_m\pi_1 & c(m,2)e_m\pi_2 & \cdot & \cdots & \cdot & c(m,m)e_m\pi_m & 0 \\ c(m+1,1)e_{m+1}\pi_1 & c(m+1,2)e_{m+1}\pi_2 & \cdot & \cdots & \cdot & c(m+1,m)e_{m+1}\pi_m & 0 \end{bmatrix}, \tag{9.31}$$

where $c(j,i)$ satisfy

$$0 \le c(j,i) \le 1, \quad i = 1,2,\cdots,m, j = 1,2,\cdots,m+1$$
$$\sum_{k=1}^{m} c(j,k)c(k,i) = c(j,i), \quad i = 1,2,\cdots,m, j = 1,2,\cdots,m+1$$
$$\sum_{i=1}^{m} c(j,i) = 1, \quad j = 1,2,\cdots,m+1$$
$$\sum_{k=1}^{m} R_k c(k,l) e_k \pi_l = (I - R_{m+1}) c(m+1,l) e_{m+1} \pi_l, l = 1,2,\cdots,m.$$

The following is a group of solutions

$$P^* = \begin{bmatrix} c(1)e_1\pi_1 & c(2)e_1\pi_2 & c(3)e_1\pi_3 & \cdots & \cdot & c(m)e_1\pi_m & 0 \\ c(1)e_2\pi_1 & c(2)e_2\pi_2 & c(3)e_2\pi_3 & \cdots & \cdot & c(m)e_2\pi_m & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \\ c(1)e_m\pi_1 & c(2)e_m\pi_2 & c(3)e_m\pi_3 & \cdots & \cdot & c(m)e_m\pi_m & 0 \\ c(1)e_{m+1}\pi_1 & c(2)e_{m+1}\pi_2 & c(3)e_{m+1}\pi_3 & \cdots & \cdot & c(m)e_{m+1}\pi_m & 0 \end{bmatrix}. \tag{9.32}$$

where $0 \le c(i) \le 1$ and $\sum_{i=1}^{m} c(i) = 1$.

B.3 Many results for a series of real numbers have their counterparts in matrix form. For example, for real number series we have $\frac{1}{1-x} = 1 + x + x^2 + \cdots$ if $|x| < 1$; and for matrix series we have $(I-P)^{-1} = I + P + P^2 + \cdots$ if $\rho(P) < 1$. In real analysis we have the following Stolz theorem: for two series of real numbers x_n and y_n, $n = 1,2,\cdots$, if $y_{n+1} > y_n$, $n = 1,2,\cdots$, $\lim_{n\to\infty} y_n = \infty$, and $\lim_{n\to\infty} \frac{x_{n+1}-x_n}{y_{n+1}-y_n}$ exists, then

$$\lim_{n\to\infty} \frac{x_n}{y_n} = \lim_{n\to\infty} \frac{x_{n+1} - x_n}{y_{n+1} - y_n}.$$

a. Prove the Stolz theorem.

b. Prove if $\lim_{n\to\infty} x_n$ exists, then $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_n = \lim_{n\to\infty} x_n$.

c. Prove the matrix formula (B.8).

[**Solution**] a. Denote $\lim_{n\to\infty} \frac{x_{n+1}-x_n}{y_{n+1}-y_n} = a$.

From the definition of convergence, for every $\epsilon > 0$ there is $N(\epsilon) \in \mathbf{N}$ such that $\forall n \ge N(\epsilon)$, we have :

$$a - \epsilon < \frac{x_{n+1} - x_n}{y_{n+1} - y_n} < a + \epsilon.$$

Because y_n is strictly increasing we can multiply the aforementioned equation with $y_{n+1} - y_n$ to get :

$$(a-\epsilon)(y_{n+1}-y_n) < x_{n+1} - x_n < (a+\epsilon)(y_{n+1}-y_n).$$

Let $k > N(\epsilon)$ be a natural number. Summing the last relation we get :

$$(a-\epsilon)\sum_{i=N(\epsilon)}^{k}(y_{i+1}-y_i) < \sum_{i=N(\epsilon)}^{k}(x_{i+1}-x_i) < (a+\epsilon)\sum_{i=N(\epsilon)}^{k}(y_{i+1}-y_i).$$

$$\Longrightarrow (a-\epsilon)(y_{k+1}-y_{N(\epsilon)}) < x_{k+1} - x_{N(\epsilon)} < (a+\epsilon)(y_{k+1}-y_{N(\epsilon)}).$$

Divide the last relation by $y_{k+1} \geq 0$ to get :

$$(a-\epsilon)(1-\frac{y_{N(\epsilon)}}{y_{k+1}}) < \frac{x_{k+1}}{y_{k+1}} - \frac{x_{N(\epsilon)}}{y_{k+1}} < (a+\epsilon)(1-\frac{y_{N(\epsilon)}}{y_{k+1}}).$$

$$\Longleftrightarrow (a-\epsilon)(1-\frac{y_{N(\epsilon)}}{y_{k+1}}) + \frac{x_{N(\epsilon)}}{y_{k+1}} < \frac{x_{k+1}}{y_{k+1}} < (a+\epsilon)(1-\frac{y_{N(\epsilon)}}{y_{k+1}}) + \frac{x_{N(\epsilon)}}{y_{k+1}}.$$

Since $\lim_{n\to\infty} y_n = \infty$, this means that there is some K such that for $k \geq K$ we have:

$$a - \epsilon < \frac{x_{k+1}}{y_{k+1}} < a + \epsilon.$$

Therefore,

$$\lim_{n\to\infty}\frac{x_n}{y_n} = a = \lim_{n\to\infty}\frac{x_{n+1}-x_n}{y_{n+1}-y_n}.$$

b. Let $y_n = n$ and $z_n = \sum_{k=1}^{n} x_k$. From part a, we know

$$\lim_{n\to\infty}\frac{z_n}{y_n} = \lim_{n\to\infty}\frac{z_{n+1}-z_n}{y_{n+1}-y_n}.$$

That is,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}x_k = \lim_{n\to\infty}x_{n+1} = \lim_{n\to\infty}x_n.$$

c. Denote real matrix $A(n) = (a_{i,j}(n))_{S\times S}$. If for any $i, j = 1, \ldots, S$ and for two series of real numbers $a_{i,j}(n)$ and y_n, $n = 1, 2, \cdots$, if $y_{n+1} > y_n$, $n = 1, 2, \cdots$, $\lim_{n\to\infty} y_n = \infty$, and $\lim_{n\to\infty}\frac{a_{i,j}(n+1)-a_{i,j}(n)}{y_{n+1}-y_n}$ exists, then from part a) we have

$$\lim_{n\to\infty}\frac{a_{i,j}(n)}{y_n} = \lim_{n\to\infty}\frac{a_{i,j}(n+1)-a_{i,j}(n)}{y_{n+1}-y_n}, \text{ for all } i, j = 1, \ldots S.$$

We write the aforementioned equation in matrix form,

$$\lim_{n\to\infty}\frac{A_n}{y_n}=\lim_{n\to\infty}\frac{A_{n+1}-A_n}{y_{n+1}-y_n}. \tag{9.33}$$

If let $A_n=\sum_{k=0}^{n-1}P^k$ and $y_n=n$, then by using (9.33), we can easily obtain

$$P^*=\lim_{n\to\infty}\frac{\sum_{k=0}^{n-1}P^k}{n}=\lim_{n\to\infty}P^n.$$

B.4 Let P be an irreducible periodic stochastic matrix. We have $p(i|i)=0$ for all $i\in\mathcal{S}$. To break the periodicity, it is enough to simply introduce a "feedback probability" $p(i|i)=\epsilon$ for only one state i, not all the states. Therefore, we define an aperiodic matrix by setting $p'(i|i)=\epsilon$, $p'(j|i)=(1-\epsilon)p(j|i)$, $j\neq i$ for one particular state i, and $p'(k|j)=p(k|j)$ for $k\in\mathcal{S}$, $j\neq i$.

1. Express the steady-state probabilities $\pi'(i)$ of P' in terms of ϵ and the steady-state probabilities $\pi(i)$ of P.

2. Let f denote the reward function and $\eta=\pi f$ be the long-run average reward for the Markov chain with transition probability matrix P. Define a reward function f' so that the long-run average performance of the Markov chain with transition probability matrix P', $\eta'=\pi' f'$, equals η.

[Solution]

1. We can get that

$$P'=\text{diag}(1,\ldots,1-\epsilon,\ldots,1)P+\text{diag}(0,\ldots,\epsilon_i,\ldots,0).$$

By $\pi'P'=\pi'$,

$$\pi'\text{diag}(1,\ldots,1-\epsilon,\ldots,1)P+\pi'\text{diag}(0,\ldots,\epsilon,\ldots,0)=\pi'.$$

That is,

$$\pi'\text{diag}(1,\ldots,1-\epsilon,\ldots,1)P=\pi'\text{diag}(1,\ldots,1-\epsilon,\ldots,1).$$

By this, we can know $\pi'\text{diag}(1,\ldots,1-\epsilon,\ldots,1)=c\pi$, where c is a constant. Finally, we get $\pi'=c\pi\text{diag}(1,\ldots,\frac{1}{1-\epsilon},\ldots,1)$. Noting that $\sum_{i=1}^{S}\pi'(i)=1$, thus $c=\frac{1}{\sum_{k=1,k\neq i}^{S}\pi(k)+\pi(i)\frac{1}{1-\epsilon}}$. Then, $\pi'(k)=c\pi(k)$ for $k\neq i$ and $\pi'(i)=c\pi(i)\frac{1}{1-\epsilon}$.

2. f' should be such that $\pi' f' = \pi f$. That is,

$$\sum_{k=1,k\neq i}^{S} c\pi(k)f'(k) + c\pi(i)\frac{1}{1-\epsilon}f'(i) = \sum_{k=1}^{S}\pi(i)f(i).$$

If we define $f'(k) = \frac{1}{c}f(k)$ for $k \neq i$ and $f'(i) = \frac{1-\epsilon}{c}f(i)$, then $\eta' = \pi' f' = \pi f = \eta$.

Appendix C

C.1 Write the steady-state probability flow-balance equation for M/M/1 queue.

[**Solution**]

Suppose λ and μ are the arrival rate and the service rate of $M/M/1$ queue respectively. Denote $p(n)$ as the steady-state probability of event that there are n customers in the system. We have the following flow balance equations,

$$\lambda p(n) = \mu p(n+1) \quad \text{for } n \geq 0.$$

Let $\rho = \frac{\lambda}{\mu}$. We know that $p(n+1) = \rho p(n) = \rho^{n+1}p(0)$ for $n \geq 0$. By $\sum_{n=0}^{\infty} p(n) = 1$, we get $p(0) = 1 - \rho$. Then $p(n) = \rho^n(1-\rho)$ for $n \geq 0$.

C.2 Consider an M/G/1 queue with arrival rate λ and mean service time $\bar{s}$. Prove that the average of the number of customers served in a busy period is $\frac{1}{1-\lambda\bar{s}}$.

[**Solution**] Let N_{bp} be the number of customers served in a busy period and $f_n = \mathcal{P}[N_{bp} = n]$. Next, we obtain a functional equation for f_n's z-transform defines as

$$F(z) = E[z^{N_{bp}}] = \sum_{n=1}^{\infty} f_n z^n.$$

The term for $n = 0$ is omitted from this definition since at least one customer must be served in a busy period. Let $\tilde{v}$ denote the number of arrivals during a service period.

We firstly consider $\tilde{v}$'s z-transform defined as

$$V(z) = E[z^{\tilde{v}}] = \sum_{k=0}^{\infty} \mathcal{P}[\tilde{v} = k]z^k.$$

Then we have

$$V(z) = \sum_{k=0}^{\infty} \int_0^{\infty} \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x)dx z^k$$

$$
\begin{aligned}
&= \int_0^\infty e^{-\lambda x} \left(\sum_{k=0}^{\infty} \frac{(\lambda x z)^k}{k!} \right) b(x) dx \\
&= \int_0^\infty e^{-\lambda x} e^{\lambda x z} b(x) dx \\
&= \int_0^\infty e^{-(\lambda - \lambda z)x} b(x) dx =: B^*(\lambda - \lambda z),
\end{aligned}
$$

where $B^*(s) = \int_0^\infty e^{-sx} b(x) dx$ and $b(x)$ denotes the service time probability density function. We assume that k customers arrive during the service period of the first customer. Moreover, since each of these arrivals will generate a sub-busy period and the number of customers served in each of these sub-busy periods will have a distribution given by f_n. Let M_i denote the number of customers served in the ith sub-busy period. We have

$$E[z^{N_{bp}} | \tilde{v} = k] = E[z^{1+M_1+M_2+\cdots+M_k}]$$

and since the M_i are independent and identically distribution we have

$$E[z^{N_{bp}} | \tilde{v} = k] = z \prod_{i=1}^{k} E[z^{M_i}].$$

But each of the M_i is distributed exactly the same as N_{bp} and, therefore

$$E[z^{N_{bp}} | \tilde{v} = k] = z[F(z)]^k.$$

Removing the condition on the number of arrivals we have

$$
\begin{aligned}
F(z) &= \sum_{k=0}^{\infty} E[z^{N_{bp}} | \tilde{v} = k] \mathcal{P}[\tilde{v} = k] \\
&= z \sum_{k=0}^{\infty} \mathcal{P}[\tilde{v} = k] [F(z)]^k \\
&= zV[F(z)].
\end{aligned}
$$

Thus, we have

$$F(z) = zB^*[\lambda - \lambda F(z)].$$

Then, we have

$$
\begin{aligned}
E(N_{bp}) &= F^{(1)}(1) = B^{*(1)}(0)[-\lambda F^{(1)}(1)] + B^*(0) \\
&= \lambda \bar{s} E(N_{bp}) + 1,
\end{aligned}
$$

thus,

$$E(N_{bp}) = \frac{1}{1-\lambda\bar{s}}.$$

Reference: L. Kleinrock, *Queueing Systems*, Volume I: Theory, John Wiley & Sons, New York, 1975.

C.3 An M/M/1 queue with arrival rate λ and departure rate μ can be constructed as follows. Choose an initial state n_0 at time 0 and a rate $\sigma > \lambda + \mu$. Generate a Poisson process with rate σ, denoted as $t_0, t_1, \ldots, t_l, \ldots$. An instant t_l, $l = 0, 1, \ldots$, is chosen as an arrival point with probability $\frac{\lambda}{\sigma}$ and as a departure point with probability $\frac{\mu}{\sigma}$. At an arrival point, we increase the population by one: $n := n + 1$, and at a departure point if $n > 0$ then we decrease the population by one: $n := n - 1$, and at other points we keep the population unchanged. Prove that the discrete-time Markov chain embedded at $t_l, l = 0, 1, \ldots$, is the discrete M/M/1 queue described on Page 526. Determine its parameters p_a and p_d [148].

[Solution] When $n > 0$, we have $p(n+1|n) = \frac{\lambda}{\sigma}$ and $p(n-1|n) = \frac{\mu}{\sigma}$. Thus , $p_a = \frac{\lambda}{\sigma}$ and $p_b = \frac{\mu}{\sigma}$. when $n = 0$, we have $p(1|0) = \frac{\lambda}{\sigma}$ and $p(0|0) = 1 - \frac{\lambda}{\sigma}$.

C.4 Many results in this book are stated only for discrete-time Markov models, but the queueing systems are usually modelled by continuous-time Markov models. Therefore, we need to use the embedded Markov chain.

a. Find the transition probabilities of the Markov chain embedded at the arrival and departure instants of an M/M/1 queue with arrival rate λ and service rate μ.

b. If we use the reward function $f(n) = n$, does the long-run average of the embedded chain equal to the mean length of the original M/M/1 queue?

c. If the answer to (b) is "No", what can we do? (cf. Problem C.9)

[Solution]

a. It is easy to know that the transition probabilities of the embedded Markov chain of M/M/1 queue is that, $p(n+1|n) = \frac{\lambda}{\lambda+\mu}, p(n-1|n) = \frac{\mu}{\lambda+\mu}$, for $n > 0$; $p(1|0) = 1$; and all the other probabilities are zero.

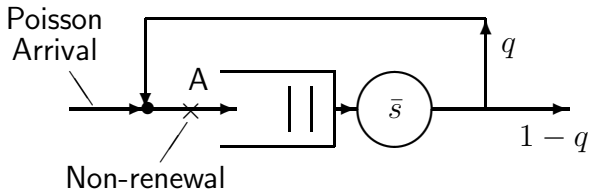

Figure 9.4: An $M/M/1$ Queue with Feedback

b. It is obvious that the steady-state probability of embedded Markov chain is not equal to that of the original queueing system. So the long-run average of the embedded chain is not equal to the mean length of the original M/M/1 queue.

c. This problem can solved by the idea of uniformization in Markov process. We should change the transition probability of embedded chain to $p(1|0) = \frac{\lambda}{\lambda+\mu}, p(0|0) = \frac{\mu}{\lambda+\mu}$ and keep the other probabilities unchanged. The steady-state probability of this embedded chain will be equal to the original queueing system and the corresponding system performance will also be equivalent.

C.5* Consider the queueing system with an $M/M/1$ queue and a feedback loop shown in Figure 9.4. This is the simplest non-acyclic open queueing network. The external arrival process to the system is a Poisson process. After the completion of its service at the server, a customer leaves the system with probability $1-q$ and returns back to the queue with probability q, $0 < q < 1$. The total arrival process to the queue at point A is a composition of both the external arrival process and the feedback process. Explain that this total arrival process at point A is not a renewal process. *(Hint: When the server is idle, the inter-arrival time is larger on average. Explain that the consecutive inter-arrival times at point A are not independent.)*

[Solution]

It is known that the renewal process requires the inter-arrival time sequence should be independent and identically distributed. In this problem, if we think the situation where the server is idle, it is easy to know that during the idle period the inter-arrival process is only contributed by the external arrival process. So the inter-arrival time is larger than the average. Thus, the inter-arrival time sequence does not have the same distribution in the combined arrival process. The total arrival process at point A is not a

renewal process is not a renewal process. Moreover, we can find the distributions of the inter-arrival times, when there are customers in the system, are different from those when there are no customers. That is to say, the inter-arrival time depends on the current state. Thus, according to the Markov property of state transitions, the internal-arrival time will also depend on the previous state. Therefore, the consecutive inter-arrival times at point A are not independent.

C.6 A nonblocking cross-bar switch can be modelled as a closed queueing network. Figure 9.5 illustrates the structure of a nonblocking packet switch consisting of N input links and M output links. Packets arriving at each input queue are put in a buffer waiting to be transmitted. Suppose that all packets belong to the same class in terms of the statistics of their destinations: A packet arriving at any input has probability $q_{i,j}$ of being destined for output j given that the previous packet at that input was destined for output i, $i, j = 1, \ldots, M$. Every packet destined to output j requires an exponentially distributed transmission time with mean $\bar{s}_j$. At a time, only the head of line (HOL) packet (the first packet) in an input queue can be transmitted and the switch can only transmit one packet to every output queue at a time. The HOL packet of an input queue contends with the HOL packets of other input queues that have the same destination in a FCFS manner. We wish to determine the maximum throughput of this $N \times M$ switch, i.e., how much packets that this switch can transmit to their destinations per second if there are always packets at every input waiting for transmission. Develop a queueing model for this problem [52, 68]. *(Hint: The HOL packet of an input queue makes a request to the switch asking for being transmitted to its destination at the time when it moves to the head position. All the requests to the same destination output queue form a logical queue called a request queue. The M request queues constitute a closed queueing network.)*

[Solution]

Although the packet served by the switch will leave from output link, we can consider the packet turn back to the input queue equivalently. This problem can just be modelled as an M-server N-customer closed Jackson network. The routing probability is $q_{i,j}, i, j = 1, \cdots, M$. The service time of each server is exponentially distributed with mean service

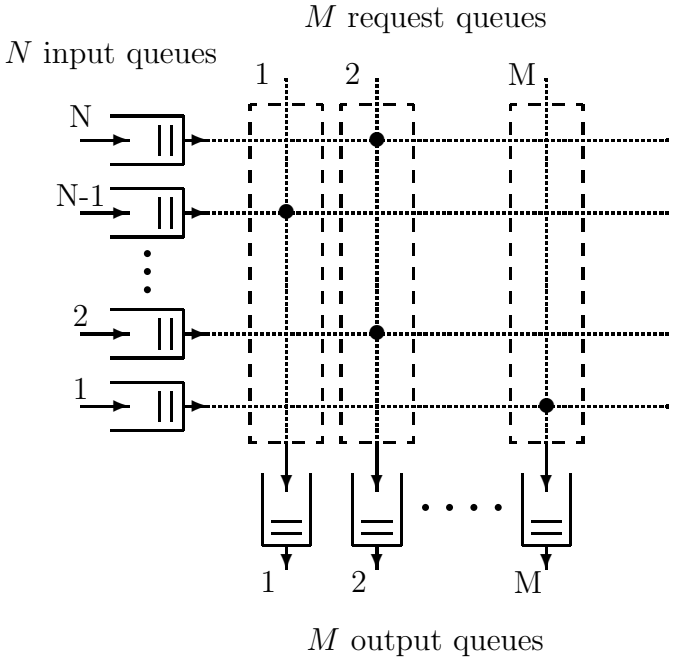

Figure 9.5: The Model of a Nonblocking Switch

time $\bar{s}_j, j = 1, \cdots, M$. Our objective is to maximize the throughput of the queueing network. It can be solved by the classical algorithm of closed Jackson network.

C.7 A cyclic queueing network of M servers is a closed network that contains M servers connecting as a circle. A two-server cyclic network is a network of two servers with routing probabilities $q_{1,2} = q_{2,1} = 1$ and $q_{1,1} = q_{2,2} = 0$. Consider a two-server cyclic network with service rates λ and μ, and a population K. Show that this closed network is equivalent to an M/M/1/K queue with arrival rate λ and service rate μ.

[Solution]

Denote the number of customers in server 2 is n. Since this is a closed network, then the number of customers in server 1 is $K - n$. From the memoryless property, the state of this closed network can be denoted as n. Since $q_{1,2} = q_{2,1} = 1$ and $q_{1,1} = q_{2,2} = 0$, there are no feedback loops. Then the arrival process to each server is a Poisson process. For server 2, the arrival process is a Poisson process with rate λ and it's service time is exponentially distributed with rate μ. Moreover, the customers in server 2 can not exceed K. In the physical meaning, this closed network is equivalent to an M/M/1/K queue with

arrival rate λ and service rate μ. We also get the flow-balance equations for this closed network,

$$\lambda p(n) = \mu p(n+1) \quad \text{for } 0 \le n \le K-1.$$

Solving it, we get $p(n) = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}$, $n \le K$, $\rho = \frac{\lambda}{\mu} \ne 1$, and if $\lambda = \mu$, $p(n) = \frac{1}{K+1}$, $0 \le n \le K$. This steady-state probabilities are the same as an M/M/1/K queue with arrival rate λ and service rate μ.

C.8 Consider an open Jackson network with M servers. The service times at server i are exponentially distributed with mean $\bar{s}_i$, $i = 1, 2, \cdots, M$; the routing probabilities are $q_{i,j}$, $i, j = 1, 2, \cdots, M$; the external arrival rate to server i is $\lambda_{0,i}$ and the leaving rate from server i is q_{i0}, $i = 1, 2, \cdots, M$. The state of the network is $\boldsymbol{n} = (n_1, \cdots, n_M)$. Let $N := \sum_{k=1}^{M} n_k$.

1. Find the conditional steady-state probability $p(\boldsymbol{n}|N)$.
2. Show that this conditional probability is the same as an equivalent closed Jackson network with a population N.
3. Find the routing probabilities of this equivalent closed Jackson network and give your explanation.

[Solution]

1. Let $p(\boldsymbol{n})$ be the steady-state probability of state $\boldsymbol{n}$, we have

$$p(\boldsymbol{n}) = p(n_1, n_2, \cdots, n_M) = \prod_{k=1}^{M} p(n_k) \tag{9.34}$$

with

$$p(n_k) = (1-\rho_k)\rho_k^{n_k}, \qquad \rho_k = \frac{\lambda_k}{\mu_k}, \qquad k = 1, 2, \cdots, M,$$

where

$$\lambda_k = \lambda_{0,k} + \sum_{j=1}^{M} \lambda_j q_{j,k}, \qquad k = 1, 2, \cdots, M \tag{9.35}$$

and $\mu_k = \frac{1}{\bar{s}_k}$.

This shows that in an open Jackson network, each server behaves as if an independent M/M/1 queue with arrival rate λ_k and service rate μ_k, $k = 1, 2, \cdots, M$, respectively.

Then,

$$\begin{aligned} p(\boldsymbol{n}|N) &= \frac{\prod_{k=1}^{M} p(n_k)}{\sum_{n_1+\cdots+n_M=N} \prod_{k=1}^{M} p(n_k)} \\ &= \frac{\prod_{k=1}^{M} (1-\rho_k)\rho_k^{n_k}}{\sum_{n_1+\cdots+n_M=N} \prod_{k=1}^{M} (1-\rho_k)\rho_k^{n_k}} \\ &= \frac{\prod_{k=1}^{M} \rho_k^{n_k}}{\sum_{n_1+\cdots+n_M=N} \prod_{k=1}^{M} \rho_k^{n_k}}. \end{aligned}$$

2. Let $G_M(N) = \sum_{n_1+\cdots+n_M=N} \prod_{k=1}^{M} \rho_k^{n_k}$, then $p(\boldsymbol{n}|N) = \frac{1}{G_M(N)} \prod_{k=1}^{M} \rho_k^{n_k}$. We can see that $p(\boldsymbol{n}|N)$ has the same formula as $p(\boldsymbol{n})$ in a equivalent closed Jackson network with a population N if $\rho_k = cx_k$, that is, $\lambda_k = cv_k$, $k = 1, \ldots, M$, where c is any non-zero constant.

3. We need to have $\lambda_k = cv_k$, $k = 1, \ldots, M$. Let $\lambda_0 = \sum_{k=1}^{M} \lambda_{0,k}$, and $q_{0,i} = \frac{\lambda_{0,i}}{\lambda_0}$, for $i = 1, \ldots, M$. Suppose the routing probabilities of this equivalent closed Jackson network is $q'_{i,j}$, $i, j = 1, \ldots, M$. Then

$$v_i = \sum_{j=1}^{M} q'_{j,i} v_j, \qquad j = 1, 2, \cdots, M. \tag{9.36}$$

Let

$$q'_{i,j} = q_{i,j} + q_{i,0} q_{0,j}, \quad i, j = 1, 2, \ldots, M. \tag{9.37}$$

(9.36) can be rewritten as

$$v_i = \sum_{j=1}^{M} q_{j,i} v_j + \sum_{j=1}^{M} q_{j,0} q_{0,i} v_j, \qquad j = 1, 2, \cdots, M.$$

Summing (9.35) from $i = 1$ to $i = M$, we get

$$\lambda_0 = \sum_{i=1}^{M} \lambda_{0,i} = \sum_{i=1}^{M} \lambda_i q_{i,0}.$$

By the aforementioned two equations and (9.35), we can prove that $\lambda_i = cv_i$ satisfies

$$\lambda_i = \sum_{j=1}^{M} q'_{j,i} \lambda_j, \qquad j = 1, 2, \cdots, M.$$

Therefore, if the routing probabilities of this equivalent closed Jackson network is defined as (9.37), then we show that this conditional probability is the same as an equivalent closed Jackson network with a population N.

Observing (9.37), we can see that the routing of the customers in this closed network is the same as that in the open network with the following modification: When a customer completes its service at server i, he/she will leave the network with probability $q_{i,0}$ and then be immediately routed to server j with probability $q_{0,j}$ or will be directly routed to server j with probability $q_{i,j}$. Thus the routing probability from server i to server j is $q'_{i,j} = q_{i,j} + q_{i,0}q_{0,j}$.

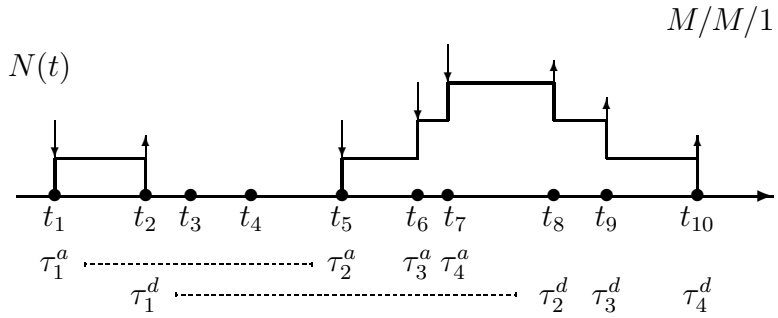

Figure 9.6: The Arrival Theorem and PASTA

C.9 Figure 9.6 illustrates a sample path $N(t)$ of an M/M/1 queue, in which the upward arrows indicate the departure instants and the downward arrows indicate the arrival instants. Let the arrival rate and service rate be λ and μ, respectively. We simulate the M/M/1 queue with the uniformization approach (cf. Problem 10.8):

i. Generate a Poisson process with rate $\lambda + \mu$, shown in Figure 9.6 as $\{t_1, t_2, t_3, \ldots\}$. Set $N(0) = n_0$ being the initial state ($n_0 = 0$ in Figure 9.6)

ii. At t_k, $k = 1, 2, \ldots$, generate an independent and uniformly distributed random variable $\xi_k \in [0, 1)$,

   (1) If $\xi_k < \frac{\lambda}{\lambda+\mu}$, then t_k is an arrival instant; set $N(t_k+) := N(t_k) + 1$.

   (2) If $\xi_k > \frac{\lambda}{\lambda+\mu}$ and $N(t_k) > 0$, then t_k is a departure instant; set $N(t_k+) := N(t_k) - 1$.

   (3) If $\xi_k > \frac{\lambda}{\lambda+\mu}$ and $N(t_k) = 0$, do nothing.

The process $N(t)$ thus generated is left-continuous. In Figure 9.6, τ_k^a, $k = 1, 2, \ldots$, indicate the arrival instants and τ_k^d, $k = 1, 2, \ldots$, indicate the departure instants; at t_3 and t_4 the server is idle and nothing changes, these instants are call "dummy instants".

a. Explain that the process $N(t)$ generated by the above algorithm is indeed an M/M/1 queue with arrival rate λ and service rate μ.

b. Define $X_k := N(t_k)$. Prove that the embedded chain $\boldsymbol{X} := \{X_1, X_2, \ldots\}$ is a Markov chain and its steady-state distribution is the same as that of the M/M/1 queue process $N(t)$(PASTA).

c. Prove that the average of the number of visits where the arriving customer or the departing customer sees n customers in the queueing system at the non-dummy instants $t_1, t_2, t_5, t_6, \ldots$, equals the steady-state probability of the state n, $n = 0, 1, \ldots$. Further, prove that the average of the number of visits where n customers are seen by the arriving customer in the system at the arrival instants τ_k^a, $k = 1, 2, \ldots$, (or the departure instants τ_k^d, $k = 1, 2, \ldots$) equals the steady-state probability of the state n, $n = 0, 1, \ldots$ (the arrival theorem).

d. Extend this explanation to (open or closed) Jackson networks.

[**Solution**]

a. It is known that these processes are generated independently. Since the total rate of the generated process is $\lambda + \mu$ and we adopt it as arrival process with probability $\frac{\lambda}{\lambda+\mu}$, with the memoryless property of Poisson process we know that the arrival process is a Poisson process with rate $(\lambda + \mu) \cdot \frac{\lambda}{\lambda+\mu} = \lambda$. We assume a customer begins to be served at time t_k. After that, the service will be completed at t_{k+1} with probability $\frac{\mu}{\lambda+\mu}$, at t_{k+2} with probability $\frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}$, $\cdots$, at t_{k+n} with probability $\left(\frac{\lambda}{\lambda+\mu}\right)^{n-1}\frac{\mu}{\lambda+\mu}$, $\cdots$, and so on. We assume ξ is a random variable which is exponential distributed with rate $\lambda + \mu$. Then from the construction of the process, we know the service time is $\frac{\mu}{\lambda+\mu}\xi + \frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}2\xi + \cdots + \left(\frac{\lambda}{\lambda+\mu}\right)^{n-1}\frac{\mu}{\lambda+\mu}n\xi + \cdots = \frac{\lambda+\mu}{\mu}\xi$. Since ξ is is a random variable which is exponential distributed with rate $\lambda + \mu$, we have

$$\mathcal{P}(\frac{\lambda+\mu}{\mu}\xi \leq x) = \mathcal{P}(\xi \leq \frac{\mu}{\lambda+\mu}x) = 1 - \exp(-(\lambda+\mu)\frac{\mu}{\lambda+\mu}x) = 1 - \exp(-\mu x).$$

The service time is also an exponential distribution with rate μ. So, the generated process $N(t)$ is indeed an M/M/1 queue with arrival rate λ and service rate μ. If a customer leave the system and there are no customers in the system at t_k, then we can similarly obtain

the arrival time is

$$\frac{\mu}{\lambda+\mu}\xi + \frac{\mu}{\lambda+\mu}\frac{\lambda}{\lambda+\mu}2\xi + \cdots + \left(\frac{\mu}{\lambda+\mu}\right)^{n-1}\frac{\lambda}{\lambda+\mu}n\xi + \cdots = \frac{\lambda+\mu}{\lambda}\xi.$$

Thus, the arrival process is Poisson process with rate λ when there are no customers in the system.

b. From the generation of process $N(t)$ we can see that the next state X_{n+1} is only dependent on the current state X_n, i.e., X has the Markovian property and X is a Markov chain. It is easy to know that the transition probability of X is $p(n+1|n) = \frac{\lambda}{\lambda+\mu}, p(n-1|n) = \frac{\mu}{\lambda+\mu}, n > 0; p(1|0) = \frac{\lambda}{\lambda+\mu}, p(0|0) = \frac{\mu}{\lambda+\mu}$; others probabilities are all zero. From the equation of steady-state probability $\pi P = \pi, \pi e = 1$, we can easily know that the steady-state probability of X is $\pi(n) = (1-\rho)\rho^n$, where $\rho = \lambda/\mu$. It is equivalent with the steady-state probability of the M/M/1 queue.

c. Let $ad(n)$ be the average of the number of arrivals or departures where the arriving customer or the departing customer sees n customers in the queueing system at the non-dummy instants $t_1, t_2, t_5, t_6, \ldots$ and $\pi(n)$ be steady-state probability of the state n. Viewing $ad(n)$ and $\pi(n)$ as the limiting probabilities, we have:

$$\pi(n) = \lim_{t\to\infty} P\{N(t) = n\},$$
$$ad(n) = \lim_{t\to\infty} P\{N(t) = n|\text{an arrival or a departure just after time } t\}.$$

This is right since time average probabilities are equal to limiting probabilities for ergodic systems

Let $A(t, t+\delta)$ be the event an arrival or a departure occurs in the time interval $[t, t+\delta)$. Then,

$$\begin{aligned}
ad(n) &= \lim_{t\to\infty}\lim_{\delta\to 0} P\{N(t) = n|A(t,t+\delta)\} \\
&= \lim_{t\to\infty}\lim_{\delta\to 0} P\{N(t) = n, A(t,t+\delta)|A(t,t+\delta)\} \\
&= \lim_{t\to\infty}\lim_{\delta\to 0} \frac{P\{A(t,t+\delta)|N(t)=n\}P\{N(t)=n\}}{P\{A(t,t+\delta)\}} \\
&= \lim_{t\to\infty}\lim_{\delta\to 0} \frac{P\{A(t,t+\delta)\}P\{N(t)=n\}}{P\{A(t,t+\delta)\}} \\
&= \lim_{t\to\infty} P\{N(t) = n\} \\
&= \pi(n).
\end{aligned}$$

Further, from the above result, we can proved that the average of the number of visits where n customers are seen by the arriving customer in the system at the arrival instants $\tau_k^a,\ k = 1, 2, \ldots$, (or the departure instants $\tau_k^d,\ k = 1, 2, \ldots$) equals the steady-state probability of the state n, $n = 0, 1, \ldots$

d. We may also simulate a (closed or open) Jackson network with uniformization approach. We only consider the open Jackson network. In the network, we have more than one arrival rates $\lambda_{0i}, i = 1, \ldots, M$ and service rates $\mu_{k,n_k}, k = 1, 2 \ldots, M$. We may generate a Poisson process with rate $R = \sum_{i=1}^{M} \lambda_{0i} + \sum_{k=1}^{M} \mu_{k,n_k}$ as $\{t_1, t_2, \ldots\}$. Set the network state at epoch t_k as $N(t_k) = (n_1, n_2, \ldots, n_M)$. At each epoch t_k, generate an independent and uniformly distributed random variable $\xi_k \in [0, 1)$,

1. If $\frac{\sum_{i=1}^{m-1} \lambda_{0i}}{R} \leq \xi_k < \frac{\sum_{i=1}^{m} \lambda_{0i}}{R}, m = 1, 2, \ldots, M$ with $\lambda_{00} = 0$, then t_k is an arrival instant and the customer arrives at server m; set $N(t_k+) = (n_1, n_2, \ldots, n_{m-1}, n_m + 1, n_{m+1}, \ldots, n_M)$.

2. If $\frac{\sum_{i=1}^{M} \lambda_{0i} + \sum_{j=1}^{m-1} \mu_{j,n_j}}{R} \leq \xi_k < \frac{\sum_{i=1}^{M} \lambda_{0i} + \sum_{j=1}^{m} \mu_{j,n_j}}{R}$ with $\mu_{0,n_0} = 0$ and $n_m > 0$, then t_k is an instant when the service of a customer at server m is finished. The customer transfers according to routing probability q_{ml}. If $l = 0$, the customer leaves the network; set $N(t_k+) = (n_1, n_2, \ldots, n_m - 1, \ldots, n_l, \ldots, n_M)$; otherwise, the customer enters server l; set $N(t_k+) = (n_1, n_2, \ldots, n_m - 1, \ldots, n_l + 1, \ldots, n_M)$.

3. If $\frac{\sum_{i=1}^{M} \lambda_{0i} + \sum_{j=1}^{m-1} \mu_{j,n_j}}{R} \leq \xi_k < \frac{\sum_{i=1}^{M} \lambda_{0i} + \sum_{j=1}^{m} \mu_{j,n_j}}{R}$ with $\mu_{0,n_0} = 0$ and $n_m = 0$, do nothing.