**Academy of Mathematics and Systems Science** May. 12, 2024
**Chinese Academy of Sciences** Assignment
Name: 李夏洋
Major: 运筹学与控制论
UID: 202328000206057
Personal Page: `https://xiayangli2301.github.io`

# Stochastic Operations Research-Homework 3

## Question 1

Derive transition probabilities and expected one-period rewards for the randomized Markov policy of Section 3.1.

**Solution.** Given that

$$q_{d_1^{\mathrm{MR}}(s_1)}(a_{1,1}) = 0.7, a_{d_1^{\mathrm{MR}}(s_1)}(a_{1,2}) \quad = 0.3, q_{d_1^{\mathrm{MR}}(s_2)}(a_{2,1}) = 1,$$

we have transition probabilities as

$$\begin{aligned} p_1(s_1 \mid s_1, d_1^{\mathrm{MR}}) &= 0.7 * 0.5 + 0.3 * 0 = 0.35, \\ p_1(s_2 \mid s_1, d_1^{\mathrm{MR}}) &= 0.7 * 0.5 + 0.3 * 1 = 0.65, \\ p_1(s_2 \mid s_1, d_1^{\mathrm{MR}}) &= 1. \end{aligned}$$

And, we have expected one-period rewards as

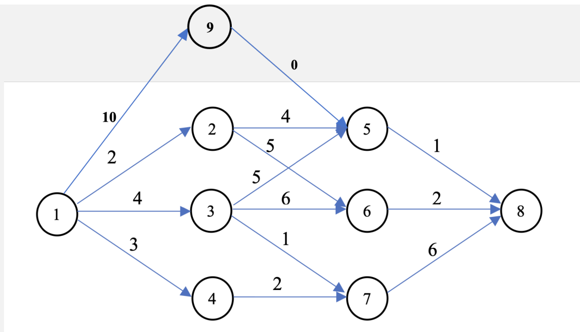
$$r_1(s_1, d_1^{\mathrm{MR}}) = 0.7 * 5 + 0.3 * 10 = 6.5, r_1(s_2, d_1^{\mathrm{MR}}) = -1$$

## Question 2

Consider the network in Figure 5, but suppose in addition that it contains an arc connecting node 1 to node 5 with length 10. given a modified network in which nodes are grouped by stages. Be sure to indicate all arc lengths in the modified network.

**Solution.**

The results follow.




**Question 3**

(A simple bandit model) Suppose there are two projects available for selection in each of three periods. Project 1 yields a reward of one unit and always occupies state s and the other, project 2, occupies either state $t$ or state $u$. When project 2 is selected, and it occupies state $u$, it yields a reward of 2 and moves to state $t$ at the next decision epoch with probability 0.5. When selected in state $t$, it yields a reward of 0 and moves to state $u$ at the next decision epoch with probability 1. Assume a terminal reward of 0, and that project 2 does not change state when it is not selected. Using backward induction determine a strategy that maximizes the expected total reward.

**Solution.** Let $a_1$ denote project 1, and $a_2$ denote project 2. Then the **backward induction** follows.

1. $u_4^*(s) = u_4^*(t) = u_4^*(u) = 0.$

2. $$\begin{aligned} u_3^*(s) &= \max\{r(s,a_1) + u_4^*(s), r(s,a_2) + 0.5 * u_4^*(t) + 0.5 * u_4^*(u)\} \\ &= r(s,a_1) + u_4^*(s) = 1 \\ u_3^*(u) &= \max\{r(u,a_1) + u_4^*(s), r(u,a_2) + 0.5 * u_4^*(u) + 0.5 * u_4^*(t)\} \\ &= r(u,a_2) + 0.5 * u_4^*(u) + 0.5 * u_4^*(t) = 2 \\ u_3^*(t) &= \max\{r(t,a_1) + u_4^*(s), r(t,a_2) + 0.5 * u_4^*(u) + 0.5 * u_4^*(t)\} \\ &= r(t,a_1) + u_4^*(s) = 1 \end{aligned}$$

3. $$\begin{aligned} u_2^*(s) &= \max\{r(s,a_1) + u_3^*(s), r(s,a_2) + 0.5 * u_3^*(t) + 0.5 * u_3^*(u)\} \\ &= r(s,a_1) + u_4^*(s) = 2 \\ u_2^*(u) &= \max\{r(u,a_1) + u_3^*(s), r(u,a_2) + 0.5 * u_3^*(u) + 0.5 * u_3^*(t)\} \\ &= r(u,a_2) + 0.5 * u_3^*(u) + 0.5 * u_3^*(t) = 3.5 \\ u_2^*(t) &= \max\{r(t,a_1) + u_3^*(s), r(t,a_2) + 0.5 * u_3^*(u) + 0.5 * u_3^*(t)\} \\ &= r(t,a_1) + u_3^*(s) = 2 \end{aligned}$$

4.
$$
\begin{aligned}
u_1^*(s) &= \max\{r(s,a_1) + u_2^*(s), r(s,a_2) + 0.5 * u_3^*(t) + 0.5 * u_3^*(u)\} \\
&= r(s,a_1) + u_4^*(s) = 3 \\
u_2^*(u) &= \max\{r(u,a_1) + u_3^*(s), r(u,a_2) + 0.5 * u_3^*(u) + 0.5 * u_3^*(t)\} \\
&= r(u,a_2) + 0.5 * u_3^*(u) + 0.5 * u_3^*(t) = 4.75 \\
u_2^*(t) &= \max\{r(t,a_1) + u_3^*(s), r(t,a_2) + 0.5 * u_3^*(u) + 0.5 * u_3^*(t)\} \\
&= r(t,a_1) + u_3^*(s) = 3
\end{aligned}
$$

The optimal policy is

$$
\begin{aligned}
d_1^*(s) = a_1, d_1^*(u) = a_2, d_1^*(t) = a_1 \\
d_2^*(s) = a_1, d_2^*(u) = a_2, d_2^*(t) = a_1 \\
d_3^*(s) = a_1, d_3^*(u) = a_2, d_3^*(t) = a_1
\end{aligned}
$$

**Question 4**

Consider a two-state Markov decision process (MDP) with state $s_1$ and state $s_2$. In state $s_1$, the decision maker chooses either action $a_1$ or action $a_2$; In state $s_2$, only action $a_3$ is available. The immediate returns and transition probabilities are as follows.

$$
\begin{aligned}
r(s_1,a_1) = 4, r(s_1,a_2) = 10, r(s_2,a_3) = 2, \\
p(s_1 \mid s_1,a_1) = p(s_2 \mid s_1,a_1) = 0.5, p(s_2 \mid s_1,a_2) = 1 \\
p(s_1 \mid s_2,a_3) = 0.2, p(s_2 \mid s_2,a_3) = 0.8.
\end{aligned}
$$

1. Solve the three-period problem with terminal reward $r_4(s_1) = r_4(s_2) = 0$ to maximize the expected total rewards and find the optimal decision rule in each period.
2. Consider the infinite-horizon discounted MDP with discounted factor $\lambda = 0.5$. Calculate the expected total discounted reward of a stationary policy $\delta^\infty$ with $\delta(s_1) = a_1$ and $\delta(s_2) = a_3$. Also, use the optimality equations to check if it is the optimal policy,

**Solution.**

1. (a) $u_4^*(s_1) = 0, u_4^*(s_2) = 0.$

   (b)

$$
\begin{aligned}
u_3^*(s_1) &= \max \begin{cases} r(s_1,a_1) + p(s_1 \mid s_1,a_1) \times u_4^*(s_1) + p(s_2 \mid s_1,a_1) \times u_4^*(s_2) = 4, \\ r(s_1,a_2) + p(s_2 \mid s_1,a_2) \times u_4^*(s_2) = 10. \end{cases} \\
&= 10 \\
u_3^*(s_2) &= r(s_2,a_3) + p(s_1 \mid s_2,a_3) \times u_4^*(s_1) + p(s_2 \mid s_2,a_3) \times u_4^*(s_2) = 2
\end{aligned}
$$

(c)

$$u_2^*(s_1) = \max \begin{cases} r(s_1,a_1) + p(s_1 \mid s_1,a_1) \times u_3^*(s_1) + p(s_2 \mid s_1,a_1) \times u_3^*(s_2) = 10 \\ r(s_1,a_2) + p(s_1 \mid s_1,a_2) \times u_3^*(s_1) + p(s_2 \mid s_1,a_2) \times u_3^*(s_2) = 12 \end{cases}$$
$$= 12$$
$$u_2^*(s_2) = r(s_2,a_3) + p(s_1 \mid s_2,a_3) \times u_3^*(s_1) + p(s_2 \mid s_2,a_3) \times u_3^*(s_2) = 5.6$$

(d)

$$u_1^*(s_1) = \max \begin{cases} r(s_1,a_1) + p(s_1 \mid s_1,a_1) \times u_2^*(s_1) + p(s_2 \mid s_1,a_1) \times u_2^*(s_2) = 12.8 \\ r(s_1,a_2) + p(s_1 \mid s_1,a_2) \times u_3^*(s_1) + p(s_2 \mid s_1,a_2) \times u_2^*(s_2) = 15.6 \end{cases}$$
$$= 15.6$$
$$u_1^*(s_2) = r(s_2,a_3) + p(s_1 \mid s_2,a_3) \times u_2^*(s_1) + p(s_2 \mid s_2,a_3) \times u_2^*(s_2) = 8.88$$

The optimal decision rule is

$$d_1^*(s_1) = a_2, d_1^*(s_2) = a_3,$$
$$d_2^*(s_1) = a_2, d_2^*(s_2) = a_3,$$
$$d_3^*(s_1) = a_2, d_3^*(s_2) = a_3,$$

2. We have

$$v_\lambda^{\delta^\infty}(s_1) = 4 + 0.5\lambda v_\lambda^{\delta^\infty}(s_1) + 0.5\lambda v_\lambda^{\delta^\infty}(s_2)$$
$$v_\lambda^{\delta^\infty}(s_2) = 2 + 0.2\lambda v_\lambda^{\delta^\infty}(s_1) + 0.8\lambda v_\lambda^{\delta^\infty}(s_2),$$

and use $\lambda = 0.5$, then we get

$$v_\lambda^{\delta^\infty}(s_1) = \frac{116}{17}$$
$$v_\lambda^{\delta^\infty}(s_2) = \frac{76}{17}.$$

The optimality equations are

$$v(s_1) = \max\{4 + 0.5\lambda v(s_1) + 0.5\lambda v(s_2), 10 + \lambda v(s_2)\}$$
$$v(s_2) = 2 + 0.2\lambda v(s_1) + 0.8\lambda v(s_2)$$

Substituting $v(s_1) = \frac{116}{17}, v(s_2) = \frac{76}{17}$ can not satisfy the optimality equations. So it is not the optimal policy.

**Question 5**

Each quarter the marketing manager of a retail store divides customers into two classes based on their purchase behavior in the previous quarter. Denote the classes as $L$ for low and $H$ for high. The manager wishes to determine to which classes of customers he should send quarterly catalogs. The cost of sending a catalog is \$15 per customer and the expected purchase depends on the customer's class and the manager's action. If a customer is in class $L$ and receives a catalog, then the expected purchase in the current quarter is \$20, and if a class $L$ customer does not receive a catalog his expected purchase is \$10. If a customer is in class H and receives a catalog, then his expected purchase is \$50, and if a class $H$ customer does not receive a catalog his expected purchase is \$25.

The decision whether or not to send a catalog to a customer also affects the customer’s classification in the subsequent quarter. If a customer is class $L$ at the start of the present quarter, then the probability he is in class $L$ at the subsequent quarter is 0.3 if he receives a catalog and 0.5 if he does not. If a customer is class $H$ in the current period, then the probability that he remains in class H in the subsequent period is 0.8 if he receives a catalog and 0.4 if he does not. Assume a discount rate of 0.9 and an objective of maximizing expected total discounted reward.

(a) Formulate this as an infinite-horizon discounted Markov decision problem, and write the optimality equations.

(b) Find an optimal policy using policy iteration starting with the stationary policy which has greatest one-step reward.

**Solution**. (a) The Markov decision problem can be formulated as follows.

1. Decision epochs:
$$T = \{1, 2, \cdots\}.$$
2. States:
$$S = \{L, H\}.$$
3. Action:
$$A_s = \{a_1, a_2\},$$
which $a_1$ means sending a catalog while $a_2$ not.
4. Expected rewards:
$$r(L, a_1) = 5, r(L, a_2) = 10, r(H, a_1) = 35, r(H, a_2) = 25.$$
5. Transition probability:
$$\begin{aligned} p(L \mid L, a_1) = 0.3, p(H \mid L, a_1) = 0.7, p(L \mid H, a_1) = 0.2, p(H \mid H, a_1) = 0.8 \\ p(L \mid L, a_2) = 0.5, p(H \mid L, a_2) = 0.5, p(L \mid H, a_2) = 0.6, p(H \mid H, a_2) = 0.4 \end{aligned}$$
.

Therefore, the optimality equations are as follows
$$\begin{aligned} v(L) &= \max\{0.3v(L) + 0.7v(H) + 5,\ 0.5v(L) + 0.5v(H) + 10\} \\ v(H) &= \max\{0.2v(L) + 0.8v(H) + 35,\ 0.6v(L) + 0.4v(H) + 25\} \end{aligned}$$

(b) The Policy Iteration Algorithm.

**The First Iteration**

1. Set $n = 0$, and select an arbitrary decision rule $d_0(s_1) = a_2, d_0(s_2) = a_1$.

2. Obtain $v_0$ by solving
$$(I - \lambda P_{d_0})v = r_{d_0}$$
where
$$I - \lambda P_{d_0} = \begin{pmatrix} 0.55 & -0.45 \\ -0.18 & 0.28 \end{pmatrix}$$
.
Then we get $v_0 = \begin{pmatrix} 254.11 \\ 288.36 \end{pmatrix}$.

3. Find $d_1 \in \arg\max_{d \in D^{MD}} \{r_d + \lambda P_d v^0\}$, and $d_1(s_1) = a_1, d_1(s_2) = a_1$.

4. n : = 1, return step 2.

**The Second Iteration**

1. Set $n = 1$, and select an arbitrary decision rule $d_1(s_1) = a_1, d_0(s_2) = a_1$.

2. Obtain $v_1$ by solving
$$(I - \lambda P_{d_1})v = r_{d_1}$$
where
$$I - \lambda P_{d_1} = \begin{pmatrix} 0.73 & -0.63 \\ -0.18 & 0.28 \end{pmatrix}$$
.
Then we get $v_1 = \begin{pmatrix} 257.69 \\ 290.66 \end{pmatrix}$.

3. Find $d_2 \in \arg\max_{d \in D^{MD}} \{r_d + \lambda P_d v^0\}$, and $d_2(s_1) = a_1, d_2(s_2) = a_1$.

4. $d_2 = d_1$, stop.

Therefore, the optimal policy is $d^*(s_1) = a_1, d^*(s_2) = a_1$.

**Question 6**

A decision maker observes a discrcte-time system which moves between states $\{s_1, s_2, s_3, s_4\}$, according to the following transition probability matrix:

$$\boldsymbol{P} = \begin{pmatrix} 0.3 & 0.4 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 & 0 \\ 0.1 & 0 & 0.8 & 0.1 \\ 0.4 & 0 & 0 & 0.6 \end{pmatrix}$$

At each point of time, the dicision maker may leave the system and receive a reward of $R = 20$ units, or alternatively remain in the system and receive a reward of $r(s_i)$ units if the system occupies state $s_i$. If the decision maker decides to remain in the system, its state at the next decision epoch is determined by $P$ . Assume a a discounted rate of 0.9 and that $r(s_i) = i$.

(a) Formulate this model as a Markov decision process, and write the optimality equations.

(b) Use policy iteration to find a stationary policy which maximizes the expected total discounted reward.

(c) Find the smallest value of $R$ so that it is optimal to leave the system in state 2.

(a)

The Markov decision problem can be formulated as follows.

1. Decision epochs:
$$T = \{1, 2, \cdots\}.$$
2. States:
$$S = \{s_0, s_1, s_2, s_3, s_4\}$$
3. Action:
$$A_s = \{a_1, a_2\}, s \in \{s_1, s_2, s_3, s_4\}; A_s = \{a_2\}, s \in \{s_0\},$$
where $a_1$ means remaining in the system, and $a_2$ means leaving the system.
4. Expected rewards:
$$r(s_i, a_1) = \sum_{j=1}^{4} p_{ij} \cdot j, i = 1, 2, 3, 4$$
$$r(s_i, a_2) = 20, i = 1, 2, 3, 4$$
$$r(s_0, a_2) = 0$$
5. Transition probability:
$$p(s_j \mid s_i, a_1) = P_{ij}, p(s_0 \mid s_i, a_2) = 1, i, j = 1, 2, 3, 4$$

Therefore, the optimality equations are as follows

$$v(s_0) = 0$$
$$v(s_1) = \max\{0.3\lambda v(s_1) + 0.4\lambda v(s_2) + 0.2\lambda v(s_3) + 0.1\lambda v(s_4) + 2.1, R\}$$
$$v(s_2) = \max\{0.2\lambda v(s_1) + 0.3\lambda v(s_2) + 0.5\lambda v(s_3) + 2.3, R\}$$
$$v(s_3) = \max\{0.1\lambda v(s_1) + 0.8\lambda v(s_3) + 0.1\lambda v(s_4) + 2.9, R\}$$
$$v(s_4) = \max\{0.4\lambda v(s_1) + 0.6\lambda v(s_4) + 2.8, R\}$$

(b)

1. Select an arbitrary decision rule $d_0 \in D$.

$$d_0(s_0) = a_2, d_0(s_1) = a_1, d_0(s_2) = a_1, d_0(s_3) = a_1, d_0(s_4) = a_1.$$

2. Obtain $v_0$ by solving

$$(I - \lambda P_{d_0})v = r_{d_0}$$

where

$$I - \lambda P_{d_1} = \begin{pmatrix} 0.73 & -0.36 & -0.18 & -0.09 & 0 \\ -0.18 & 0.73 & -0.45 & 0 & 0 \\ -0.09 & 0 & 0.28 & -0.09 & 0 \\ -0.36 & 0 & 0 & 0.46 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \end{pmatrix}$$

. Then we get $v_0 = \begin{pmatrix} 25.641651934 \\ 26.120727891 \\ 27.00585336 \\ 26.154336296 \\ 0. \end{pmatrix}$.

3. Find $d_1 \in \arg\max_{d \in D^{MD}} \{r_d + \lambda P_d v^0\}$, and $d_1(s_i) = a_1, d_1(s_0) = a_2, i = 1, 2, 3, 4.$

4. $d_1 = d_0$, stop. Thus, the optimal stationary policy we find is

$$d(s_i) = a_1, d(s_0) = a_2, i = 1, 2, 3, 4$$

(c)

Solve

$$(I_4 - \lambda P')v = r,$$

where $P' = \begin{pmatrix} 0.73 & -0.36 & -0.18 & -0.09 \\ -0.18 & 0.73 & -0.45 & 0 \\ -0.09 & 0 & 0.28 & -0.09 \\ -0.36 & 0 & 0 & 0.46 \end{pmatrix}$, we get

$$v(2) = 26.1207.$$

So if $R \geqslant 26.1207$, it is optimal to leave the system.