Markov Decision Processes-Homework 3

Name: Zhengwei Hu ID: 12443586

Problem 1

Solution:

(a) Let state *X* be the number of days since the last repair or maintenance. According to the statement of the problem if the system continues for 30 days without repair or maintenance, the system fails with probability 1 sometime during the 30th day, we could find that the states are 0, 1, 2, 3, …, 30, and there are 31 states in total. Note that, state 0 indicates the initial state or new. The two actions in the MDP are Produce and Maintain. Thus, the states and actions of this Markov chain could be shown below:

$$S = \{0, 1, 2, \cdots, 30\}, \quad A = \{\text{Produce, Maintain}\} \tag{1}$$

According to the statement that the system will fail on the *d*th day since the last repair or maintenance with probability of failure of $(1-\psi^d)$, the TPM of action “Produce” is given by

$$TPM_{\Pr oduce} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 1-\psi & 0 & \psi & 0 & 0 & \cdots & 0 & 0 \\ 1-\psi^2 & 0 & 0 & \psi^2 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & & & \\ 1-\psi^d & 0 & \cdots & \cdots & 0 & \psi^d & \cdots & 0 \\ \vdots & & & & & & & \\ 1-\psi^{29} & 0 & \cdots & 0 & \cdots & 0 & 0 & \psi^{29} \\ 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}_{31\times31} \tag{2}$$

where d=1, 2, …29.

Since the cost for repair is $Cr = -\$450$, the TRM of the action “Produce” is

$$TRM_{\Pr oduce} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ -450 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -450 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -450 & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -450 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ -450 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}_{31\times31} \tag{3}$$

After the system is repaired or maintained, it is assumed to be as good as new. It means that the machine will go to state 0 every time after maintenance. Then, the TPM of action “Maintain” could be shown as:

$$TPM_{Ma\,\mathrm{int}\,ain} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}_{31\times 31} \tag{4}$$

The cost for maintenance is $Cm = -\$175$, and the TRM of the action "Maintain" is

$$TRM_{Ma\,\mathrm{int}\,ain} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ -175 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -175 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -175 & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -175 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ -175 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}_{31\times 31} \tag{5}$$

(b) The MATLAB codes for this problem include ***main_HW3_Problem_1.m*** and ***func_HW3_Problem_1.m***. The ***main_HW3_Problem_1.m*** is the main function which shows the optimal policy as well as the iteration history. The ***func_HW3_Problem_1.m*** is a user defined general function for performing relative value iteration. The results of these codes are stored in the file named ***Results of HW3_Problem_1***.

(c) The optimal policy for $\psi = 0.9$ is listed in the Table below:

State	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Action	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2
State	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
Action	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	

where action "1" means "Produce", action "2" means "Maintain" .

The results indicate that the optimal policy for preventative maintenance of the machine is the machine should be maintained after 2 days since the last repair or maintenance.

Problem 2

Solution:

According to the statement of this problem, we have

$$P_1 = \begin{bmatrix} 0.1 & 0.9 \\ 0.8 & 0.2 \end{bmatrix}, \quad R_1 = \begin{bmatrix} 12 & 16 \\ -7 & 13 \end{bmatrix} \tag{6}$$

$$P_2 = \begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix}, \quad R_2 = \begin{bmatrix} 12 & -11 \\ 6 & 9 \end{bmatrix} \tag{7}$$

The Q-Learning algorithm in which the state (action) trajectory is: 1(1), 2(2), 1(2), 2(2), 2(X). Therefore, the system starts in state 1.

State 1. Set all the Q-factors to 0:

$$Q(1,1) = Q(1,2) = Q(2,1) = Q(2,2) = 0 \tag{8}$$

The set of actions allowed in state 1 is $A(1) = \{1,\ 2\}$ and that allowed in state 2 is $A(2) = \{1,\ 2\}$. Then we have $\left|A(i)\right| = 2$ for *i* =1, 2. Let the step-size α be defined by

$$\alpha = \frac{A}{B + k} \tag{9}$$

where $A = 5$, $B = 10$, and n denotes the number of state transitions.

According to the trajectory of this problem, the selected action is 1. Simulate action 1 and the next state is 2.

State 2. The current state (*j*) is 2 and the old state (*i*) was 1. The action (*a*) selected in the old state was 1. So we now have to update *Q* (1,1). Now: $k = 0;\ \alpha = \frac{5}{10 + 0} = 0.5$.

$$r(i,\ a,\ j) = r(1,\ 1,\ 2) = 16 \tag{10}$$

$$\max_b\{Q(2,\ b)\} = \max_b\{Q(2,\ 1),\ Q(2,\ 2)\} = \max\{0,\ 0\} = 0 \tag{11}$$

$$\begin{aligned} Q(1,\ 1) &\leftarrow (1-\alpha)Q(1,\ 1) + \alpha[r(1,\ 1,\ 2) + \lambda \max_{b\in A(2)} Q(2,\ b)] \\ &= 0.5 \times 0 + 0.5 \times (16 + 0.7 \times 0) \\ &= 8 \end{aligned} \tag{12}$$

Current state is 2 and the selected action is 2. Simulate action 2 and the next state is 1.

State 1 (again). The current state (*j*) is 1 and the old state (*i*) was 2. The action (*a*) selected in the old state was 2. So we now have to update *Q* (2, 2). Now: $k = 1;\ \alpha = \frac{5}{10+1} = \frac{5}{11}$.

$$r(i,\ a,\ j) = r(2,\ 2,\ 1) = 6 \tag{13}$$

$$\max_{b}\{Q(1,\, b)\} = \max_{b}\{Q(1,\, 1),\, Q(1,\, 2)\} = \max\{8,\, 0\} = 8 \tag{14}$$

$$\begin{aligned} Q(2,\, 2) &\leftarrow (1-\alpha)Q(2,\, 2) + \alpha[r(2,\, 2,\, 1) + \lambda \max_{b\in A(1)} Q(1,\, b)] \\ &= \frac{6}{11}\times 0 + \frac{5}{11}[6 + 0.7\times 8] = \frac{58}{11} \end{aligned} \tag{15}$$

Current state is 1 and the selected action is 2. Simulate action 2 and the next state is 2.

State 2 (again). The current state (*j*) is 2 and the old state (*i*) was 1. The action (*a*) selected in the old state was 2. So we now have to update *Q* (1, 2). Now: $k = 2;\ \alpha = \frac{5}{10+2} = \frac{5}{12}$.

$$r(i,\, a,\, j) = r(1,\, 2,\, 2) = -11 \tag{16}$$

$$\max_{b}\{Q(2,\, b)\} = \max_{b}\{Q(2,\, 1),\, Q(2,\, 2)\} = \max\{0,\, \frac{58}{11}\} = \frac{58}{11} \tag{17}$$

$$\begin{aligned} Q(1,\, 2) &\leftarrow (1-\alpha)Q(1,\, 2) + \alpha[r(1,\, 2,\, 2) + \lambda \max_{b\in A(2)} Q(2,\, b) \\ &= \frac{7}{12}\times 0 + \frac{5}{12}[-11 + 0.7\times \frac{58}{11}] = -3.0455 \end{aligned} \tag{18}$$

Current state is 2 and the selected action is 2. Simulate action 2 and the next state is 2.

State 2 (a third time). The current state (*j*) is 2 and the old state (*i*) was 2. The action (a) selected in the old state was 2. So we now have to update Q (2, 2). Now: $k = 3;\ \alpha = \frac{5}{10+3} = \frac{5}{13}$.

$$r(i,\, a,\, j) = r(2,\, 2,\, 2) = 9 \tag{19}$$

$$\max_{b}\{Q(2,\, b)\} = \max_{b}\{Q(2,\, 1),\, Q(2,\, 2)\} = \max\{0,\, \frac{58}{11}\} = \frac{58}{11} \tag{20}$$

$$\begin{aligned} Q(2,\, 2) &\leftarrow (1-\alpha)Q(2,\, 2) + \alpha[r(2,\, 2,\, 2) + \lambda \max_{b\in A(2)} Q(2,\, b) \\ &= \frac{8}{13}\times \frac{58}{11} + \frac{5}{13}[9 + 0.7\times \frac{58}{11}] \\ &= 3.2448 + 0.3846(12.6909) \\ &= 8.1257 \end{aligned} \tag{21}$$

Current state is 2 and the selected action is X. Then, we stop here.