* "Assignment - 1" *

Nelakuditi Rahul Naga
AI20BTECH11029

1) a) The states in the given Markov Reward process are S, 1, 3, 5, 6, 7, 8, W. This is because the states 2, 4 and 9 are equivalent to states 7, 8 and 3 respectively. The transition matrix P is given by :-

$$P = \begin{array}{c} \\ S \\ 1 \\ 3 \\ 5 \\ 6 \\ 7 \\ 8 \\ W \end{array} \begin{array}{c} \begin{array}{cccccccc} S & 1 & 3 & 5 & 6 & 7 & 8 & W \end{array} \\ \begin{bmatrix} 0 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0.25 & 0 & 0 & 0 & 0.5 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

b) The goal of this question is to calculate the expected number of die throws required to ~~the~~ reach the goal state W. from any other state. for this problem, the suitable reward function and discount factor γ are as follows:-

$$R(s) = \begin{cases} -1, & s \neq W \\ 0, & s = W \end{cases}$$

$$\gamma = 1$$

→ The Bellman evaluation equation for a Markov Reward process is :-

$$V = (I - \gamma P)^{-1} R$$

⇒ The vector "V" would give the required average number of moves to reach goal state W from a state S.

→ As the number of moves are required from states other than W, we will consider only the submatrix that constitutes the first 7 rows & columns of matrix P as the transition matrix in the Bellman equation. Subsequently, R becomes a column vector with length 7 consisting of all -1's. Upon substituting the matrices in the Bellman equation we obtain :-

$$v(s) : \begin{bmatrix} -7.083 \\ -7 \\ -6.67 \\ -6.67 \\ -5.33 \\ -5.33 \\ -5.33 \end{bmatrix}$$

→ But as we require the expected no. of moves, we take modulus of above values and the expected moves are thus given by :-

$$v(s) : \begin{bmatrix} 7.083 \\ 7 \\ 6.67 \\ 6.67 \\ 5.33 \\ 5.33 \\ 5.33 \end{bmatrix}$$

2) a) We are required to formulate the given problem as an MDP. We know that an MDP is given by the tuple $\langle S, A, P, R, \gamma \rangle$. For the given problem, they are given by :-

i) State space 'S' is the enumeration of all possible number of working machines on a particular day. Therefore it is given by $\{0, 1, 2, \dots N-1, N\}$.

ii) Action space 'A' is given by {Call the Repair Man, Don't call the Repair Man}.

iii) Our goal in this problem is to maximize the profits earned. Therefore the

Reward function can be formulated as follows :-

$$R(s_t, a_t) = \begin{cases} s_t - \frac{N}{2}, & \text{if } a_t = \text{Call the Repair Man} \\ s_t, & \text{if } a_t = \text{Don't call the Repair Man} \end{cases}$$

iv.) As there are 2 possible actions - Call the Repair Man (CR) and Don't call the Repair Man (DCR) we will have 2 state transition matrices, one for each action. They are given by :-

$$P^{CR} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & N-1 & N \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \end{array}$$

$$P^{DCR} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & N-1 & N \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & \cdots & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} & 0 \\ \frac{1}{N+1} & \frac{1}{N+1} & \frac{1}{N+1} & \cdots & \frac{1}{N+1} & \frac{1}{N+1} \end{bmatrix} \end{array}$$

b) The problem deals with maximizing long-time profit, the horizon is infinite. Hence, for the expected rewards/profit to converge it is better to use discounted setting for the above MDP formulation.

c) The policy given is to never call the repair man. We are also given that $N=5$. The Bellman Evaluation equation in matrix form is given by :-

$$V^{\pi} = (I - \gamma P^{\pi})^{-1} R^{\pi}$$

→ We will take $\gamma = 0.5$.

→ Also, we know that :-

$$P^{\pi}(s'|s) = \sum_{a \in A} \pi(a|s) P^{a}_{ss'} \quad \text{and} \quad R^{\pi}(s) = \sum_{a \in A} \pi(a|s) \sum_{s'} P^{a}_{ss'} R^{a}_{ss'}$$

Upon calculating we have :- (taking $N=5$)

$$P^{\pi} = \begin{array}{c|cccccc} & 0 & 1 & 2 & 3 & 4 & 5 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 2 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 3 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 4 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 \\ 5 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{array}$$

$$R^{\pi} = [0, 1, 2, 3, 4, 5]^T$$

$\Rightarrow$ Upon substituting in the Bellman Evaluation equation we obtain :-

$$V^{\pi} = \left[0, \frac{4}{3}, \frac{8}{3}, 4, \frac{16}{3}, \frac{20}{3}\right]^T$$

d) The policy iteration algorithm involves initialization of policy and then repeated evaluation and improvement of policy steps.

→ The initial policy in this problem is the no-repair policy.

→ We have already evaluated this policy in part (c) of this question.

→ Now to get an improved policy we will find the greedy policy $\pi' =$ greedy(V^{π}) where V^{π} is the value function calculated above for Policy π (no repair policy). We know that :-

$$\pi'(s) = \text{Greedy}(V^{\pi}) = \begin{cases} 1, & \text{if } a = \arg\max_{a \in A} \left[\sum_{s' \in S} P^a_{ss'} (R^a_{ss'} + \gamma V^{\pi}(s'))\right] \\ 0, & \text{otherwise} \end{cases}$$

→ Let us denote the actions Don't call Repair Main and Call the Repair Man as DCR and CR respectively. Let us also denote $\sum_{s' \in S} P^a_{ss'} (R^a_{ss'} + \gamma V^{\pi}(s'))$ by 'V'.

* Upon calculation, the results are as follows :-

i) for $s=0$; $V^{CR} = \frac{5}{6}$, $V^{DCR} = 0$ $\Rightarrow$ optimal action is CR

ii) for $s=1$; $V^{CR} = \frac{11}{6}$, $V^{DCR} = \frac{4}{3}$ $\Rightarrow$ optimal action is CR

iii) for $s=2$; $V^{CR} = \frac{17}{6}$, $V^{DCR} = \frac{8}{3}$ $\Rightarrow$ optimal action is CR iv.) for $S = 3$; $V^{CR} = \frac{23}{6}$, $V^{DCR} = 4$ $\Rightarrow$ Optimal action is DCR

v.) for $S = 4$; $V^{CR} = \frac{29}{6}$, $V^{DCR} = \frac{16}{3}$ $\Rightarrow$ Optimal action is DCR

vi.) for $S = 5$; $V^{CR} = \frac{35}{6}$, $V^{DCR} = \frac{20}{3}$ $\Rightarrow$ Optimal action is DCR

$\therefore$ The improved policy π' is as follows:-

$$\pi'(s) = \begin{cases} \text{Call the Repair Man, if } s = 0,1,2 \\ \text{Don't call the Repair Man, if } s = 3,4,5 \end{cases}$$

3) a) The Bellman evaluation equation is given by :-

$$V^{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s'} P^{a}_{ss'} [R^{a}_{ss'} + \gamma . V^{\pi}(s')]$$

We will choose $\gamma = 1$ for the given problem. The state transition matrix corres-ponding to policy π_1 is given by :-

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 0.9 & 0.1 & 0 \\ B & 0.1 & 0 & 0 & 0.9 \\ C & 0.9 & 0 & 0 & 0.1 \\ D & 0 & 0 & 0 & 1 \end{array}$$

Clearly, as D is the terminal state with reward 100, we have $V^{\pi_1}(D) = 100$. Now using Bellman evaluation equation we have :-

$$V^{\pi_1}(A) = 0.9[V^{\pi_1}(B) - 10] + 0.1[V^{\pi_1}(C) - 10]$$
$$V^{\pi_1}(B) = 0.9[V^{\pi_1}(D) - 10] + 0.1[V^{\pi_1}(A) - 10]$$
$$V^{\pi_1}(C) = 0.9[V^{\pi_1}(A) - 10] + 0.1[V^{\pi_1}(D) - 10]$$

Upon solving we obtain :- $V^{\pi_1} = [75.609, 87.56, 68.048, 100]^T$

→ The state transition matrix for policy π_2 is :-

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 0.1 & 0.9 & 0 \\ B & 0.9 & 0 & 0 & 0.1 \\ C & 0.1 & 0 & 0 & 0.9 \\ D & 0 & 0 & 0 & 1 \end{array}$$

We now have :-

$$V^{\pi_2}(D) = 100$$
$$V^{\pi_2}(A) = 0.9[V^{\pi_2}(C) - 10] + 0.1[V^{\pi_2}(B) - 10]$$
$$V^{\pi_2}(B) = 0.9[V^{\pi_2}(A) - 10] + 0.1[V^{\pi_2}(D) - 10]$$
$$V^{\pi_2}(C) = 0.9[V^{\pi_2}(D) - 10] + 0.1[V^{\pi_2}(A) - 10]$$

Upon solving we obtain:- $V^{\pi_2} = [75.609, 68.048, 87.56, 100]^T$

→ The state transition matrix for policy π_3 is :-

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 0.42 & 0.58 & 0 \\ 0.1 & 0 & 0 & 0.9 \\ 0.1 & 0 & 0 & 0.9 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

We trivially have $v^{\pi_3}(D) = 100$. We also have :-

$$v^{\pi_3}(A) = 0.42[v^{\pi_3}(B) - 10] + 0.58[v^{\pi_3}(C) - 10]$$

$$v^{\pi_3}(B) = 0.9[v^{\pi_3}(D) - 10] + 0.1[v^{\pi_3}(A) - 10]$$

$$v^{\pi_3}(C) = 0.1[v^{\pi_3}(A) - 10] + 0.9[v^{\pi_3}(D) - 10]$$

upon solving we obtain :- $v^{\pi_3} = [77.77, 87.77, 87.77, 100]^T$

b) It is easy to observe that $v^{\pi_3}(s) \geq v^{\pi_1}(s)$ and $v^{\pi_3}(s) \geq v^{\pi_2}(s)$ for all $s \in \{A, B, C, D\}$. Hence, we conclude that π_3 is the best policy among the given policies.

c) It is easy to see that the policies π_1 and π_2 are not comparable. This is because $v^{\pi_1}(B) > v^{\pi_2}(B)$ but $v^{\pi_1}(C) < v^{\pi_2}(C)$. Hence, we conclude that all policies are not comparable.

d) We are given two deterministic policies π_1 & π_2. of an MDP M. A new policy π that is better than policies π_1 and π_2 can be constructed as follows:-

→ Suppose the states in MDP are $S_1, S_2, \ldots S_n$ where S_n is the terminal state without loss of generality. Now policy π is as follows:-

$$\pi(S_i) = \begin{cases} \pi_1(S_i), \text{ if } v^{\pi_1}(S_i) > v^{\pi_2}(S_i) \\ \pi_2(S_i), \text{ if } v^{\pi_2}(S_i) > v^{\pi_1}(S_i) \\ \pi_1(S_i) \text{ with probability } p,\ \pi_2(S_i) \text{ with probability } 1-p, \text{ o/w} \end{cases}$$

for $i = 1, 2, \ldots n$ and $p \in [0,1]$. If $p = 0$ (or) 1, π becomes a deterministic policy else if $p \in (0,1)$, the policy π is stochastic. We are constructing the new policy by greedily choosing an action suggested by the determinstic policy which is ~~suggests~~ an action that maximizes the value function.

→ When both policies suggest an action that is equally optimal, we choose from either of them with appropriate probabilities as mentioned above.

→ Hence, the new policy π is better than both π_1 and π_2.

4) a) We will 4 possible pairs of γ and η :- $(0.9, 0)$, $(0.9, 0.5)$, $(0.1, 0)$ and $(0.1, 0.5)$. We will analyze what is the optimal chosen in each of the four scenarios.

i) $\gamma = 0.9$ and $\eta = 0$ i.e; high discount factor and zero noise environment.

→ As γ is high, the agent focuses on rewards than can be obtained in faraway future. Also, as there is no noise in the environment, the agent can go along the dashed path i.e; by risking the cliff. This is because in zero noise environment, the agent will successfully avoid the negative payoff terminal states while going along the dashed path.

∴ for high γ and low η, agent prefers distant exit but risks the cliff.

ii) $\gamma = 0.9$ and $\eta = 0.5$ i.e; high discount factor and noisy environment.

→ As γ is high, the agent focuses on rewards that can be obtained in faraway future. Also, as the environment is noisy, the agent chooses to go along the solid path i.e; by avoiding the cliff. This is because in a noisy environment, there is high probability that the agent will fall off into the negative payoff terminal states if it chooses to go along the dashed path.

∴ for high γ and high η, agent prefers distant exit and avoids the cliff.

iii.) $\gamma = 0.1$ and $\eta = 0$ i.e; low discount factor and zero noise environ-
-ment.

→ As γ is low, the agent focuses on rewards than can be obtained in imme-
-diate future. Also, as there is no noise in the environment, the agent choo-
-ses to go along the dashed path i.e; by risking the cliff. This is because in zero noise environment, the agent will successfully avoid the negative pay-off terminal states while going along the dashed path.

∴ For low γ and low η, agent prefers close exit but risks the cliff.

iv.) $\gamma = 0.1$ and $\eta = 0.5$ i.e; low discount factor and noisy environment.

→ As γ is low, the agent focuses on rewards that can be obtained in immediate future. Also, as the environment is noisy, the agent chooses to go along the solid path i.e; by avoiding the cliff. This is because in a noisy environment, there is high probability that the agent will fall off into the negative payoff terminal states if it chooses to go along the dashed path.

∴ For low γ and high η, agent prefers close exit and avoids the cliff.

5) a) We are given action value functions $Q_1^{\pi}(s,a)$ and $Q_2^{\pi}(s,a)$ correspon-
-ding to MDP's M_1 and M_2. Now, M_3 is a composite MDP of M_1 & M_2 and $Q_3^{\pi}(s,a)$ is it's corresponding action-value function. The expressions for the action value functions are as follows:

$$Q_1^{\pi}(s,a) = E_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1} \mid s_t = s, a_t = a\right) ; \; r \sim R_1(s,a)$$

$$Q_2^{\pi}(s,a) = E_{\pi}\left(\sum_{k=0}^{\infty} (\gamma^k \cdot r_{t+k+1}) \mid s_t = s, a_t = a\right) ; \; r \sim R_2(s,a)$$

$$Q_3^{\pi}(s,a) = E_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1} \mid s_t = s, a_t = a\right) ; \; r \sim (R_1+R_2)(s,a)$$

→ From above equations, we can conclude that it is possible to combine the action-value functions $Q_1^{\pi}(s,a)$ and $Q_2^{\pi}(s,a)$ linearly (since the expectation operator is linear) to calculate $Q_3^{\pi}(s,a)$.

b) We are given optimal policies π_1^* and π_2^* corresponding to MDP's M_1 and M_2 respectively. The optimal policies for M_1 and M_2 can be obtained by solving :-

$$\pi_1^* = \underset{\pi}{\arg\max}\left[E_\pi\left(\sum_{t=0}^{\infty} \gamma^t \cdot r_{t+1} \right)\right] \quad ; \; r \sim R_1$$

$$\pi_2^* = \underset{\pi}{\arg\max}\left[E_\pi\left(\sum_{t=0}^{\infty} \gamma^t \cdot r_{t+1} \right)\right] \quad ; \; r \sim R_2$$

→ Now the optimal policy π_3^* for MDP M_3 is given by :-

$$\pi_3^* = \underset{\pi}{\arg\max}\left[E_\pi\left(\sum_{t=0}^{\infty} \gamma^t \cdot r_{t+1} \right)\right] \quad ; \; r \sim R_1 + R_2$$

→ Due to non-linearity of the "max" operator, we can conclude that it is not possible to simply combine optimal policies π_1^* and π_2^* in order to obtain π_3^*.

c) Consider an arbitrary policy π. We have :-

$$v_1^\pi(s) = E_\pi\left(\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1} \mid s_t = s \right) \quad ; \; r \sim R_1(s)$$

$$v_2^\pi(s) = E_\pi\left(\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1} \mid s_t = s \right) \quad ; \; r \sim R_2(s)$$

$$v_3^\pi(s) = E_\pi\left(\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1} \mid s_t = s \right) \quad ; \; r \sim (R_1 + R_2)(s)$$

It is easy to see that $v_3^\pi(s) = v_1^\pi(s) + v_2^\pi(s)$. Now it is given that π^* is an optimal policy for MDP's M_1 & M_2. We will prove that this is also an optimal policy of M_3 using proof by contradiction as follows :-

→ Assume that there exists a ~~better~~ policy π' such that π^* performs worser than π' on a state $s \in S$ i.e; $v_3^{\pi'}(s) > v_3^{\pi^*}(s)$. Hence we have :-

$$v_1^{\pi'}(s) + v_2^{\pi'}(s) > v_1^{\pi^*}(s) + v_2^{\pi^*}(s)$$

which implies that $v_1^{\pi'}(s) > v_1^{\pi^*}(s)$ (or) $v_2^{\pi'}(s) > v_2^{\pi^*}(s)$ which is a contradiction since π^* is an optimal policy for both M_1 and M_2. Hence,

π^* is an optimal policy for M_3.

d) We know that value functions under policy π for MDP's M_1 and M_2 are given as follows :-

$$V_1^{\pi}(s) = E_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k \cdot r^1_{t+k+1} \mid s_t = s\right) \; ; \; r^1 \sim R_1(s)$$

$$V_2^{\pi}(s) = E_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k \cdot r^2_{t+k+1} \mid s_t = s\right) \; ; \; r^2 \sim R_2(s)$$

$\Rightarrow$ As ist is given that $R_1(s,a,s') - R_2(s,a,s') = \epsilon$, we have :-

$$V_1^{\pi}(s) - V_2^{\pi}(s) = E_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k \cdot (r^1_{t+k+1} - r^2_{t+k+1}) \mid s_t = s\right)$$

$$V_1^{\pi}(s) - V_2^{\pi}(s) = E_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k \cdot \epsilon \mid s_t = s\right)$$

$$V_1^{\pi}(s) - V_2^{\pi}(s) = \epsilon \cdot \sum_{k=0}^{\infty} \gamma^k = \frac{\epsilon}{1-\gamma}$$

$$\therefore \; V_1^{\pi}(s) = V_2^{\pi}(s) + \frac{\epsilon}{1-\gamma}$$