VAISHNAVI
AI20BTECH11025

Problems

a. States: $S = \{S, 1, 3, 5, 6, 7, 8, W\}$

Transition matrix:

$$
\begin{array}{c|cccccccc}
 & S & 1 & 3 & 5 & 6 & 7 & 8 & W \\
\hline
S & 0 & 1/4 & 1/4 & 0 & 0 & 1/4 & 1/4 & 0 \\
1 & 0 & 0 & 1/4 & 1/4 & 0 & 1/4 & 1/4 & 0 \\
3 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 \\
5 & 0 & 0 & 1/4 & 0 & 1/4 & 1/4 & 1/4 & 0 \\
6 & 0 & 0 & 1/4 & 0 & 0 & 1/4 & 1/4 & 1/4 \\
7 & 0 & 0 & 1/4 & 0 & 0 & 1/4 & 1/4 & 1/4 \\
8 & 0 & 0 & 1/4 & 0 & 0 & 0 & 1/2 & 1/4 \\
W & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

b) Reward function: $R(S) = -1$ for $s = \{S, 1, 3, 5, 6, 7, 8\}$
$R(W) = 0$

Discount factor: $\gamma = 1$

$$V = (I - P)^{-1} R \qquad \text{where } R = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$

$$\begin{array}{c} V(S) \\ V(1) \\ V(3) \\ V(5) \\ V(6) \\ V(7) \\ V(8) \\ V(W) \end{array} = \begin{bmatrix} -7.0833 \\ -6.9999 \\ -6.6666 \\ -6.666 \\ -5.33 \\ -5.33 \\ -5.33 \\ 0 \end{bmatrix}$$

We don't consider the states 2, 4 and 9 as ~~they are equivalent~~ the agent does not step over them and directly moves to 7, 8 and 3 respectively. So they are equivalent

The negative value functions of the states represent the expected no. of die throws / steps on average required to reach the goal state starting from them.

2a) State space : Number of presently working machines

$S = \{0, 1, 2, \dots N\}$

Action space : Call / Not call a repair man

$A = \{ a_0 :$ Call repair man,
$a_1 :$ not call repair man $\}$

Rewards : 'm' machines working $= \$ m$

Call a repair man $= -\frac{N}{2} \$$

$$\boxed{\begin{aligned} R^{a_0}_{ss'} &= S - N/2 \\ R^{a_1}_{ss'} &= S \end{aligned}}$$

Transition probabilities :

1. For action a_0 :

$$\begin{array}{c|c} & \begin{matrix} 0 & 1 & 2 & \dots & N-1 & N \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & \dots & & 1 \\ 0 & 0 & 0 & \dots & 1/2 & 1/2 \\ 0 & 0 & 0 & \dots 1/3 & 1/3 & 1/3 \\ & & & & & \\ & 1/N & 1/N & \dots 1/N & 1/N & 1/N \\ 1/(N+1) & 1/(N+1) & 1/(N+1) & \dots & 1/(N+1) & 1/(N+1) \end{bmatrix} \end{array}$$

$$\cancel{P^{a_0}_{ss'} = \begin{cases} 1 & s' = N \\ 0 & \text{otherwise} \end{cases}}$$

$$P^{a_0}_{ss'} = \begin{cases} \frac{1}{S+1} & s' \geq N-S \\ 0 & \text{otherwise} \end{cases}$$

2. For action a_1 :

$$\begin{array}{c|c} & \begin{matrix} 0 & 1 & 2 & 3 & \dots & N-1 & N \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ N-1 \\ N \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 & & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & & 0 & 0 \\ & & & & & & \\ 1/N & 1/N & 1/N & 1/N & & 1/N & 0 \\ 1/(N+1) & 1/(N+1) & 1/(N+1) & 1/(N+1) & & 1/(N+1) & 1/(N+1) \end{bmatrix} \end{array}$$

$$P^{a_1}_{ss'} = \begin{cases} \frac{1}{S+1} & s' \leq S \\ 0 & \text{otherwise} \end{cases}$$

b) Since it is an infinite horizon setting, to avoid infinite returns we use discounted setting for the above MDP.

c) $\pi(S) = a_1$; $N = 5$

$$V = R + \gamma P V \qquad \gamma = 1$$

$$V = R + PV$$

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

$$\begin{bmatrix} V(0) \\ V(1) \\ V(2) \\ V(3) \\ V(4) \\ V(5) \end{bmatrix} = \left(I - \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix} \right)^{-1} \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$$

On solving, we get

$$= \begin{pmatrix} 0 \\ 2 \\ 4 \\ 6 \\ 8 \\ 10 \end{pmatrix}$$

$3/2 \times$

d) Policy iteration

$\pi_0(s) = a_1 \quad \forall s$

$\rightarrow s = 0$

$$Q^{\pi_0}(0, a_0) = \sum_{s' \in S} P_{ss'}^{a_0} \left[R_{ss'}^{a_0} + V^{\pi_0}(s') \right]$$

$$= 1\left[-\frac{5}{2} + V^{\pi_0}(5)\right] = 10 - \frac{5}{2} = \frac{15}{2}$$

$$Q^{\pi_0}(0, a_1) = V^{\pi_0}(0) = 0$$

$$Q^{\pi_0}(0, a_0) > V^{\pi_0} \quad \Rightarrow \quad \pi'(0) = a_0$$

$\rightarrow s = 1$

$$Q^{\pi_0}(1, a_0) = \frac{1}{2}\left[1 - \frac{5}{2} + V^{\pi_0}(5)\right] + \frac{1}{2}\left[1 - \frac{5}{2} + V^{\pi_0}(4)\right]$$

$$= \frac{1}{2}\left(1 - \frac{5}{2} + 10\right) + \frac{1}{2}\left(1 - \frac{5}{2} + 8\right)$$

$$= \frac{1}{2}\left(2 - 5 + 18\right) = 15/2$$

$$Q^{\pi_0}(1, a_1) = V^{\pi_0}(1) = 2$$

$$Q^{\pi_0}(1, a_0) > Q^{\pi_0}(1, a_1) \quad \Rightarrow \quad \pi'(1) = a_0$$

$\rightarrow s = 2$

$$Q^{\pi_0}(2, a_0) = \frac{1}{3}\left[2 - \frac{5}{2} + V^{\pi_0}(5)\right] + \frac{1}{3}\left[2 - \frac{5}{2} + V^{\pi_0}(4)\right] + \frac{1}{3}\left[2 - \frac{5}{2} + V^{\pi_0}(3)\right]$$

$$= 15/2$$

$$Q^{\pi_0}(2, a_1) = V^{\pi_0}(2) = 4$$

$$Q^{\pi}(2, a_0) > Q^{\pi_0}(1, a_1) \quad \Rightarrow \quad \pi'(2) = a_0$$

→ $s = 3$

$$Q^{\pi_0}(3, a_0) = \frac{1}{4}\left[3 - \frac{5}{2} + V^{\pi_0}(5)\right] + \frac{1}{4}\left[3 - \frac{5}{2} + V^{\pi_0}(4)\right] + \frac{1}{4}\left[3 - \frac{5}{2} + V^{\pi_0}(3)\right] + \frac{1}{4}\left[3 - \frac{5}{2} + V^{\pi_0}(2)\right]$$

$$= 15/2$$

$$Q^{\pi_0}(3, a_1) = V^{\pi_0}(3) = 6$$

$$Q^{\pi}(3, a_0) > V^{\pi}(3) \qquad \Rightarrow \pi'(3) = a_0$$

→ $s = 4$

$$Q^{\pi_0}(4, a_0) = \frac{1}{5}\left(4 - \frac{5}{2} + V^{\pi}(5)\right] + \frac{1}{5}\left(4 - \frac{5}{2} + V^{\pi_0}(4)\right] + \frac{1}{5}\left(4 - \frac{5}{2} + V^{\pi}(3)\right) + \frac{1}{5}\left(4 - \frac{5}{2} + V^{\pi}(2)\right)$$

$$+ \frac{1}{5}\left(4 - \frac{5}{2} + V^{\pi}(1)\right) = 15/2$$

$$Q^{\pi_0}(4, a_1) = V^{\pi_0}(4) = 8$$

$$Q^{\pi}(4, a_0) < V^{\pi}(4) \qquad \Rightarrow \pi'(4) = a_1$$

→ $s = 5$

$$Q^{\pi_0}(5, a_0) = \frac{1}{6}\left(5 - \frac{5}{2}\right) \cdot 6 + \frac{1}{6}\left(V^{\pi_0}(5) + \cdots V^{\pi_0}(0)\right] = 15/2$$

$$Q^{\pi_0}(5, a_1) = V^{\pi_0}(5) = 10$$

$$Q^{\pi_0}(5, a_0) < V^{\pi_0}(5) \qquad \Rightarrow \pi'(5) = a_1$$

<u>Improved policy</u> π':

$$\pi'(s) = \begin{cases} a_0 & s = \{0, 1, 2, 3\} \\ a_1 & s = \{4, 5\} \end{cases}$$

$[\text{greedy}(V^{\pi_0}(s))]$

Problem 3

a) P_{π_1}

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

$$R_{\pi_1} = \begin{bmatrix} -10 \\ -10 \\ -10 \\ 100 \end{bmatrix}$$

$\gamma = 1$

$$V_{\pi_1} = R_{\pi_1} + \gamma \cdot P_{\pi_1} V_{\pi_1}$$

$$\Rightarrow V_{\pi_1} = (I - P_{\pi_1})^{-1} \cdot R$$

$$= \begin{bmatrix} 1.21 & 1.09 & 0.12 & 1 \\ 0.12 & 1.109 & 0.012 & 1 \\ 1.09 & 0.987 & 1.109 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -10 \\ -10 \\ -10 \\ 100 \end{bmatrix}$$

$$= \begin{bmatrix} 75.61 \\ 87.6 \\ 67.92 \\ 100 \end{bmatrix}$$

b) P_{π_2}

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

$$R_{\pi_2} = \begin{bmatrix} -10 \\ -10 \\ -10 \\ 100 \end{bmatrix}$$

$\gamma = 1$

$$V_{\pi_2} = R_{\pi_2} + \gamma P_{\pi_2} V_{\pi_2}$$

$$\Rightarrow V_{\pi_2} = (I - P_{\pi_1})^{-1} R$$

$$= \begin{bmatrix} 75.61 \\ 67.92 \\ 87.6 \\ 100 \end{bmatrix}$$

c) P_{π_3}

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 0.4\times0.9 + 0.6\times0.1 = 0.42 & 0.4\times0.1 + 0.6\times0.9 = 0.58 & 0 \\ B & 0.1 & 0 & 0 & 0.9 \\ C & 0.1 & 0 & 0 & 0.9 \\ D & 0 & 0 & 0 & 0 \end{array}$$

$$R_{\pi_3} = \begin{pmatrix} -10 \\ -10 \\ -10 \\ 100 \end{pmatrix} \qquad V_{\pi_3} = \begin{pmatrix} 77.78 \\ 87.78 \\ 87.78 \\ 100 \end{pmatrix}$$

b) Since $V^{\pi_3}(S) \geq V^{\pi_1}(S)$ & $V^{\pi_3}(S) \geq V^{\pi_2}(S)$ $\forall s \in S$

By partial ordering over policies, π_3 is the best policy among all policies suggested.

$\pi_3 > \pi_2$

$\pi_3 > \pi_1$

c) Not all policies are comparable.

A policy is better than another only if its value function for all states is greater (or equal) than to that of the other policy.

Policies π_1 & π_2 are not comparable as

$V^{\pi_1}(B) > V^{\pi_2}(B)$

but $V^{\pi_2}(C) < V^{\pi_2}(C)$

d) We can construct a policy π_3 better than two given policies π_1 and π_2 for an MDP, if they are not already optimal.

WLOG - let us start policy iteration from the policy π_1

And iterate as long as we do not reach a policy $\pi_3 > \pi_2$

~~We can~~ As in policy iteration, policy at next iteration is at least as good as the current policy, we can say than $\pi_3 > \pi_1$

Also as there always exists an optimal policy for an MDP, we can be sure of finding a $\pi_3 > \pi_2$ (or at least as good as if π_2 is itself optimal)

Problem 4

Case I : γ - low
η - low (or) zero

As γ is low, the agent is myopic - concerned only with immediate rewards. As the noise is also low, environment is predictable, The agent thus takes the close exit and risks the cliff

Case II : γ - low
η - high

As γ is low, the agent is myopic - concerned only with imm. rewards - thus prefers the close exit and settles with low reward
As the environment is noisy, highly unpredictable, it does not prefer going near the cliff.
Thus RL agent takes - close exit but avoids the cliff.

Case III : γ - high
η - zero

γ is high - the agent is farsighted, considers future rewards strongly - thus prefers the distant exit with high ~~pay~~ reward
~~As environment is~~: As noise is low, env is predictable - it can risk going through the cliff.
Thus RL agent takes - Distant exit and risks the cliff.

Case IV γ - high
η - high

γ is high - agent is far sighted, considers future rewards strongly - prefers distant exit which has a high reward.
As env is noisy, highly unpredictable - it cannot risk going towards the cliff.
Thus RL agent takes - distant exit but avoids the cliff

5 a) g $R^{\pi_3} = R^{\pi_1} + R^{\pi_2}$

$$V^{\pi_1} = (I - \gamma P)^{-1} R^{\pi_1}$$

$$V^{\pi_2} = (I - \gamma P)^{-1} R^{\pi_2}$$

$$V^{\pi_3} = (I - \gamma P)^{-1} (R^{\pi_1} + R^{\pi_2}) = V^{\pi_1} + V^{\pi_2}$$

$$Q_3^{\pi}(s,a) = \sum_{s' \in S} P_{ss'}^{\pi_3 a} [R_{ss'}^{\pi_3 a} + \gamma \underbrace{V^{\pi_3}(s')}_{V^{\pi_1}(s') + V^{\pi_2}(s')}]$$

$$R_{ss'}^{\pi_1 a} + R_{ss'}^{\pi_2 a}$$

$$= \sum_{s' \in S} P_{ss'}^{\pi_1 a} (R_{ss'}^{\pi_1 a} + \gamma V^{\pi_1}(s')) + \sum_{s' \in S} P_{ss'}^{\pi_2 a} (R_{ss'}^{\pi_2 a} + \gamma V^{\pi_2}(s'))$$

$$= Q_1^{\pi}(s,a) + Q_2^{\pi}(s,a) \quad //$$

b) No. We cannot obtain an optimal policy for M_3 using the optimal policies for M_1 and M_2

The reward function being the sum of the other two doesn't guarantee an optimal action for M_3 which can be obtained from the optimal policies of M_1 & M_2

π^* is an optimal policy for M_1 and M_2

c)

$$V_1^{\pi^*}(s) = \max_a Q_1^{\pi^*}(s,a)$$
$$V_2^{\pi^*}(s) = \max_a Q_2^{\pi^*}(s,a)$$

Consider π^* to be optimal policy for M_3

$$V_3^{\pi^*}(s) = V_1^{\pi^*}(s) + V_2^{\pi^*}(s)$$
$$= \max_a Q_1^{\pi^*}(s,a) + \max_a Q_2^{\pi^*}(s,a)$$

~~Since policy is the same~~

$$\geq \max_a \left(Q_1^{\pi^*}(s,a) + Q_2^{\pi^*}(s,a) \right)$$
$$\geq \max_a Q_3^{\pi^*}(s,a) \Rightarrow V_3^{\pi^*}(s) = \max_a Q_3^{\pi^*}(s,a)$$

$\Rightarrow \pi_3^*$ is an optimal policy for M_3 as well.

d) ~~$V^{\pi}(s) \leftarrow R_1(s) + \epsilon \times P$~~

$$V_1 = R_1 + \gamma P V_1 \Rightarrow R_1 = V_1 - \gamma P V_1$$
$$V_2 = R_2 + \gamma P V_2 \Rightarrow R_2 = V_2 - \gamma P V_2$$

$$R_1 - R_2 = \left(\begin{array}{c} \sum \pi(a|s_1) \sum P_{ss'}^a (R_{1ss'}^a - R_{2ss'}) \\ \vdots \end{array} \right) = \epsilon \vec{k}.$$

$$V_1 - \gamma P V_1 - (V_2 - \gamma P V_2) = \epsilon \vec{k}$$
$$(V_1 - V_2)(1 - \gamma P) = \epsilon \vec{k}$$
$$V_1 = V_2 + \frac{\epsilon \vec{k}}{1 - \gamma P} \quad //.$$