INDIAN INSTITUTE OF TECHNOLOGY HYDERABAD

ADDITIONAL SHEET

ROLL NO. :

1 (a)

States set = $\{S, 1, 3, 5, 6, 7, 8, W\}$

Transition matrix ⇒

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

(b)

$$R(s) = \begin{cases} -1 & s \in \{S, 1, 3, 5, 6, 7, 8\} \\ 0 & s = W \end{cases}$$

Discount factor $\gamma = 1$

we want to calculate the no. of throws required from states except W to reach W.

Therefore taking states $k \in \{S, 1, 3, 5, 6, 7, 8\}$

$$R = [-1, -1, -1, -1, -1, -1]^T$$

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

$$V = (I - \gamma P)^{-1} R = \begin{bmatrix} -7.118 \\ -7.050 \\ -6.711 \\ -6.77 \\ -5.35 \\ -5.35 \\ -5.35 \end{bmatrix}$$

ROLL NO. :

3 a

$$P^{\pi_1} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 0.9 & 0.1 & 0 \\ 0.1 & 0 & 0 & 0.9 \\ 0.9 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$P^{\pi_2} = \begin{array}{c} A \\ B \\ C \\ D \end{array} \begin{bmatrix} 0 & 0.1 & 0.9 & 0 \\ 0.9 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 0.9 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P^{\pi_3} = \begin{array}{c} A \\ B \\ C \\ D \end{array} \begin{bmatrix} 0 & 0.42 & 0.58 & 0 \\ 0.1 & 0 & 0 & 0.9 \\ 0.1 & 0 & 0 & 0.9 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{bmatrix} -10 & -10 \\ -10 & -10 \\ -10 & -10 \\ 100 & 100 \end{bmatrix} \end{array} \qquad \gamma = 1$$

$$V^{\pi_1} = \left(I - \gamma P^{\pi_1}\right)^{-1} R = \begin{bmatrix} -24.39 & -24.39 \\ -12.43 & -12.43 \\ -31.95 & -31.95 \end{bmatrix}$$

$$V^{\pi_2} = \left(I - \gamma P^{\pi_2}\right)^{-1} R = \begin{bmatrix} -24.39 & -24.39 \\ -31.95 & -31.95 \\ -12.43 & -12.43 \end{bmatrix}$$

$$V^{\pi_3} = \begin{bmatrix} -22.22 & -22.22 \\ -12.22 & -12.22 \\ -12.22 & -12.22 \end{bmatrix}$$

(b) π_3 is the best policy because all the individual elements / ~~reward~~ values in value matrix with respect to V^{π_1} and V^{π_2} ~~of~~ V^{π_3} is highest.

(c) No, because if in any policy π_1 and π_2 ~~bes any two values~~

$$V^{\pi_1} = \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix} \qquad V^{\pi_2} = \begin{bmatrix} 0.8 \\ 0.7 \end{bmatrix}$$

then $V^{\pi_1}[1] > V^{\pi_2}[1]$

$V^{\pi_1}[2] < V^{\pi_2}[2]$

then thy two policy are not comparable

(d) We will take values functions value corresponding to each policy π_1 and π_2 then we will pick maximum value actions corresponding to each state from policy π_1 and π_2 and create a new policy π with this actions (maximum values) for each state (can have mixture of actions as well) (This will be similar for the question ~~MD~~ MDP π_3 as we have different actions for different states)

INDIAN INSTITUTE OF TECHNOLOGY HYDERABAD

ADDITIONAL SHEET

ROLL NO. :

(4) (a) for policy π_1, the agent is preferring close exit and risk the cliff. Which means the agent is short sighted that means γ will be low (0.1) and not giving importance to future (distant).

We will be putting noise η to be zero in the environment, so that there is no danger of tripping of the cliff.

for π_3, the agent is preferring close exit and not risk the cliff. Which means the agent is short-sighted taking γ to be low (0.1) and the noise η to be ~~..~~ more (0.5) in the environment so that there is danger of tripping in the cliff.

for π_2, the agent is preferring ~~close~~ distant exit and risk the cliff which means the agent is far sighted taking γ to be high (0.9) and the noise η to be low (0) in the environment so that the danger of the tripping is none.

for π_4, the agent is preferring distant exit and not risk the cliff. which means the agent is far sighted taking γ to be high loss and the noise to be more loss in the environment so that the dangerous tripping is exists.

(5) (a)

$$Q_1^\pi(s,a) = \mathbb{E}_\pi\left[r'_{t+1} + \gamma r'_{t+2} \dots \mid s_t = s, a_t = a\right]$$

$$Q_2^\pi(s,a) = \mathbb{E}_\pi\left[r^2_{t+1} + \gamma r^2_{t+2} \dots \mid s_t = s, a_t = a\right]$$

$$\Rightarrow Q_1^\pi(s,a) + Q_2^\pi(s,a)$$

$$\Rightarrow \mathbb{E}_\pi\left[r'_{t+1} + \gamma r'_{t+2} \dots \mid s_t = s, a_t = a\right] + \mathbb{E}_\pi\left[r^2_{t+1} + \gamma r^2_{t+2} \dots \mid s_t = s, a_t = a\right]$$

$$\Rightarrow \mathbb{E}_\pi\left[(r'_{t+1} + r^2_{t+2}) + \gamma(r'_{t+1} + r^2_{t+2}) \dots \mid s_t = s, a_t = a\right]$$

$$\Rightarrow \cancel{\mathbb{E}_\pi[} \; Q_3^\pi(s,a)$$

$$Q_3^\pi(s,a) = Q_2^\pi(s,a) + Q_1^\pi(s,a)$$

~~(b)~~

(c) from (a)

$$V_3^\pi(s,a) = V_1^\pi(s,a) + V_2^\pi(s,a)$$

when π^* is the optimal policy for ~~$Q_1(s,a)$ and Q~~ MDP M_1 and M_2 then we can say ~~that~~ M_3 will have π^* as the optimal policy

(d)

$$V_1^\pi(s) = \mathbb{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r'_{t+k+1}\right)$$

$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \left(r^2_{t+k+1} + \varepsilon\right)\right]$$

$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r^2_{t+k+1}\right] + \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \varepsilon\right]$$

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

(b) It's not possible to do so by computing π_1^* and π_2^* because the rewards for the optimal policies may be totally different and the optimal policy to be obtained ~~wit~~ for π_2^* will have totally different reward sums.

2 (a)

State space = $\{0, 1, \ldots \ldots N\}$

= No. of machines working

Action = [Repair , No repair]

Repair → repairing will be done

No repair → Repairing will not be done

$$\text{Rewards} = \begin{cases} K - \frac{N}{2} & \text{(Repairing was done with K working machines)} \\ K & \text{(working machines count)} \end{cases}$$

Transition

Transition probability matrix with no repair action

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

Transition matrix with repair action

$$\Rightarrow \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \\ N-1 \\ N \end{array} \begin{bmatrix} 0 & \cdots & 1 \\ 0 & \cdots & 1 \\ 0 & \cdots & 1 \\ \vdots & & \vdots \\ - & \cdots & 1 \\ - & \cdots & - \end{bmatrix}$$

(b) We are going to use discounted setting with $\gamma < 1$ as we don't have any absorbing states (can be thought as indefinite horizon problem)

(c)

$$
\begin{array}{c|cccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 \\
\hline
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\
2 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\
3 & 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\
4 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 \\
5 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6
\end{array}
$$

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

$$R = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 \end{bmatrix}^T$$

$\pi =$ no repair policy is chosen

$$V^{\pi} = \begin{bmatrix} 0 & 1.904 & 3.783 & 5.706 & 7.61 & 9.302 \end{bmatrix}^T$$