OJUS SINGHAL
2020094

Page No.
Date

Answer 1 :

From the description, we can form the policy π as :

$$\pi(a) = \begin{cases} \frac{2}{15}, & a \in \{2,4,6,8,10\} \\ \frac{1}{15}, & a \in \{1,3,5,7,9\} \end{cases}$$

Now,

$$E[R_1 + R_2 + \dots + R_{10}] = \sum_{i=1}^{10} E[R_i]$$

$$= 10\, E[R_1]$$

(This is true since the policy is constant and $E[R_i]$ is only dependent on policy in this case)

$$= 10 \sum_{a} \pi(a)\, q(a)$$

(where $q(a)$ is the expected reward for action a)

$$= 10\left(\frac{2}{15}(2+4+6+8+10) + \frac{1}{15}(1+3+5+7+9)\right)$$

$$= \boxed{56.67}$$

Page No. ______
Date

Answer 2

From the description we have

$$q(a) = \begin{cases} 0.5 & , a \in \{1,2,4,5,7,9,10\} \\ 0.46 & , a \in \{3,6,8\} \end{cases}$$

A simple stochastic policy maximizing the expected reward is:

$$\pi(a) = \begin{cases} 1 & , a = 1 \\ 0 & , a \in \{2,3,4 \ldots, 9, 10\} \end{cases}$$

Similarly, we can have 5 other optimal stochastic policies where $\pi(a) = 1$ where $a = 2$ or 4 or 5 or 7 or 9 and 0 otherwise.

Answer 3

t \ a	a=1: Q(a), π(a)	a=2: Q(a), π(a)	a=3: Q(a), π(a)	A_t	R_t	
0	1 , 0	2 , 0	3 , 1	3	0	exploit
1	1 , 0.5	2 , 0	0 , 0.5	3	1	explore
2	1 , 0	2 , 1	0.5 , 0	2	1	exploit
3	1 , 0	1 , 0	0.5 , 1	3	1	explore
4	1 , 0.5	1 , 0.5	0.67 , 0	1	0	exploit
5	0 , 0.5	1 , 0	0.67 , 0.5	1	1	explore
6	0.5 , 0	1 , 1	0.67 , 0	2	0	exploit

Convention is that R_t gives after A_t, $Q_0(a)$ are initial assumptions.

($\overline{X}$ denotes the new reward system)

Answer 6

Let $\overline{G_t} = R_t + C + \gamma \overline{G_{t+1}}$

In terms of G_t, $\overline{G_t}$ is :

$$\overline{G_t} = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + C)$$

$$\overline{G_t} = c \sum_{k=0}^{\infty} \gamma^k + \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$$\overline{G_t} = \frac{c}{1-\gamma} + G_t \quad \text{— (I)}$$

Putting this in the Bellman Eq^n,

$$\overline{V_\pi(s)} = E[\overline{G_t} \mid S_t = s]$$

$$= E\left[R_t + \frac{c}{1-\gamma} + \gamma G_{t+1} \mid S_t = s\right]$$

$$\bar{V}_\pi(s) = \frac{c}{1-\gamma} + E[R_t + \gamma G_{t+1} \mid S_t = s]$$

$$\bar{V}_\pi(s) = \frac{c}{1-\gamma} + V_\pi(s)$$

Hence $\boxed{V_c = \dfrac{c}{1-\gamma}}$

Hence the signs of the rewards are not important.

However, for episodic tasks,

$$\overline{G_t} = \sum_{k=0}^{T-t-1} \gamma^k \left(R_{t+k+1} + C\right)$$

$$\overline{G_t} = \frac{C\left(1-\gamma^{T-t}\right)}{1-\gamma} + G_t$$

Here the additional factor is dependent on the current time-step, hence the constant can change the dynamics of the problem.

~~Bec~~ In the maze running example, the agent will try to increase the number of time steps just to gain this constant reward, but we want the time-steps to be less.

Hence adding constants is not equivalent in episodic tasks.

<u>Answer 8</u> $\quad v_*(s) = \max_a q_*(s,a)$