CS5500: Reinforcement Learning

Assignment No. 2

Name - Vijay Tadikamalla

Roll.no - CS17BTECH11040

Q1

a) We know $V(s) = E(r + \gamma V(s_{t+1}) \mid (s_t = s))$

Also $V(s_6) = 0$

$\therefore V(s_5) = 10 + 1(0) = 10$

Similarly $V(s_4) = 1 + 1(0) = 1$

$V(s_3) = (0.9)V(s_4) + (0.1)V(s_5) = 1.9$

$V(s_2) = 2 + V(s_3) = 2 + 1.9 = 3.9$

$V(s_1) = 1 + V(s_3) = 2.9$

b) Path 1 : $s_1 \rightarrow s_3 \rightarrow s_4 \rightarrow s_6$ (3 times) Reward = 2

Path 2 : $s_1 \rightarrow s_3 \rightarrow s_5 \rightarrow s_6$ (1 time) Reward = 11

Path 3 : $s_2 \rightarrow s_3 \rightarrow s_5 \rightarrow s_6$ (1 time) Reward = 12

$\therefore V(s_1) = \frac{2+2+2+11}{4} = 4.25$

$V(s_2) = \frac{12}{1} = 12$

c) We know $V(S_t) = V(S_t) + \alpha_t[r_{t+1} + \cancel{\gamma_s} \gamma V(S_{t+1}) - V(S_t)]$

where $\alpha_t = \frac{1}{t}$ for t^{th} episode

First, we initialize $V(S)=0 \quad \forall S$

At $t=1$ (episode 1) $S_1 \xrightarrow[+1]{} S_3 \xrightarrow[0]{} S_4 \xrightarrow[+1]{} S_6$ $\alpha_t = 1$

$V(S_1) = 0 + 1(0+1-0) = 1$
$V(S_3) = 0 + 1(0+1(0)-0) = 0$
$V(S_4) = 0 + 1(0+1-0) = 1$

At $t=2$ $\alpha_t = 1/2$ $S_1 \xrightarrow[+1]{} S_3 \xrightarrow[0]{} S_5 \xrightarrow[+10]{} S_6$

$V(S_1) = 1 + \frac{1}{2}(1+1(0)-1) = 1$
$V(S_3) = 0 + \frac{1}{2}(0+1(0)-0) = 0$
$V(S_5) = 0 + \frac{1}{2}(0+1(10)-0) = 5$

At $t=3$, $\alpha_t = 1/3$

$V(S_1) = 1 + \frac{1}{3}(0+1-1) = 0$
$V(S_3) = 0 + 1/3(1+0-0) = 1/3$
$V(S_4) = 1 + 1/3(0+1-1) = 1$

At $t=4$, $\alpha_t = 1/4$
$V(S_1) = 0 + 1/4(1+1/3-0) = 1/3$
$V(S_3) = 1/3 + 1/4(0+1-1/3) = 1/2$
$V(S_4) = 1 + 1/4(0+1-1) = 1$

At $t=5$, $\alpha_t = 1/5$
$V(S_2) = 0 + 1/5(2+1/2-0) = 1/2$
$V(S_3) = 1/2 + 1/5(0+5-1/2) = 14/10$
$V(S_5) = 5 + 1/5(10+0-5) = 6$

Q1d) In the given sample $S_3 \to S_4$ occurs three times while $S_3 \to S_5$ occurs two Times

Hence $P_{S_3 S_4} = 0.6$ & $P_{S_3 S_5} = 0.4$

$\therefore$ $V(S_6) = 0$

$V(S_5) = 10$, $V(S_4) = 1$

$V(S_3) = 0.6 \times V(S_4) + (0.4) V_{S5} = 4.6$

$V(S_2) = 6.6$

Q1 e) TD(0) estimate is closest to the true value as TD has low variance and employes bootstrapping technique.

MC is far from the true value because there are not sufficient samples for state (S_2). Moreover the only sample of S_2 was used to calculate the value of the state (S_2) which contained only $(S_3 \to S_5)$ transition which has a very low probability

Q2 i) $\alpha_t = 1/t$

We get $\sum_{t=1}^{\infty} \alpha_t = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}$ -----

$\geq 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right)$ ---

$\geq 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}$ -----

$\geq \infty$

$$\sum \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^2} \leq \int_1^{\infty} \frac{1}{t^2}\, dt < \infty$$

Hence $\alpha_t = \frac{1}{t}$ will converge

b) $\alpha_t = \frac{1}{t^2}$

We know $\sum \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty$

$\therefore$ It will not converge

c) $\alpha_t = \frac{1}{t^{2/3}}$

$$\sum \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t^{2/3}} > \sum_{t=1}^{\infty} 1/t = \infty$$

$$\sum \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^{4/3}} \leq \int_{t=1}^{\infty} \frac{1}{t^{4/3}}\, dt < \infty$$

$\therefore$ It will converge

d) $\alpha_t = 1/t^{1/2}$

$$\sum \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t^{1/2}} > \sum_{t=1}^{\infty} 1/t = \infty$$

$$\sum \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t} = \infty$$

$\therefore$ It will not converge.

$\therefore$ For any $p > 1$ $\sum_{t=1}^{\infty} \frac{1}{t^p} < \int_1^{\infty} \frac{1}{t^p}\, dt$

If $0 < p \leq 1$ $\sum_{t=1}^{\infty} \frac{1}{t^p} = \infty$

$\sum_{t=1}^{\infty} \frac{1}{t^{2p}} < \int_1^{\infty} \frac{1}{t^{2p}}\, dt$ iff $2p > 1$

$\Rightarrow p > \frac{1}{2}$

$\therefore$ $\alpha_t = \frac{1}{t^p}$ will lead to convergence if

$\frac{1}{2} < p \leq 1$

3. ϵ - greedy exploration for a policy π means that the action at state S which has the maximum $Q(S,a)$ is chosen with probability $(1-\epsilon)$ and any ~~a~~ random action is chosen uniformly with probability $\frac{\epsilon}{A}$.

$$\pi(a/s) = \begin{cases} (1-\varepsilon) + \varepsilon/A & \text{if } a = \underset{a}{\text{argmax}}\ Q(S,a) \\ \varepsilon/A & \text{, otherwise} \end{cases}$$

where A = total no. of actions at state S

For any policy π, ε-greedy policy π' wrt q_π is

$$q_\pi(s, \pi'(s)) = \sum_{a \in A} \pi'(a/s) q_\pi(s,a)$$

$$= \varepsilon/m \sum_{a \in A} q_\pi(s,a) + (1-\varepsilon) \max_{a \in A} q_\pi(s,a)$$

$$\geq \frac{\varepsilon}{m} \sum_{a \in A} q_\pi(s,a) + (1-\varepsilon) \sum_{a \in A} \frac{\pi(a/s) - \varepsilon/m}{1-\varepsilon} q_\pi(s,a)$$

$$= \sum_{a \in A} \pi(a/s) q_\pi(s,a) = V_\pi(s)$$

$\therefore$ From policy improvement theorem $V_{\pi'}(s) \geq V_\pi(s)$

Q4 Weight of n^{th} -term $= w_n = (1-\lambda)\lambda^{n-1}$

$\therefore w_1 = (1-\lambda)$

So when $w_n = \frac{w_1}{2}$

$\Rightarrow (1-\lambda)\lambda^{n-1} = \frac{(1-\lambda)}{2}$

$$2\lambda^{n-1} = 1$$

$$\log 2 + (n-1)\log\lambda = 0$$

$$n-1 = \frac{-\log 2}{\log\lambda}$$

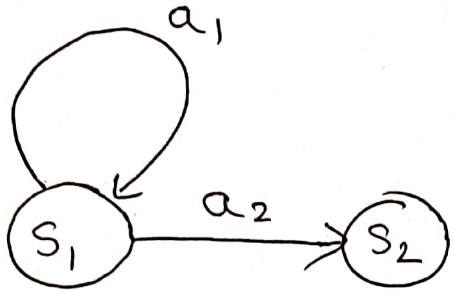
so $n(\lambda) = 1 - \frac{\log 2}{\log\lambda}$

Now $n(\lambda) = 3$

$1 - \frac{\log 2}{\log\lambda} = 3 \Rightarrow \log\lambda = -\frac{\log 2}{2}$

$\therefore \lambda = \frac{1}{\sqrt{2}}$

Q5 The trajectory $(S_1, a_1, 1, S_2, a_2, 2, S_2)$ can be definitely generated (if such transitions are allowed by environment) because we are following ϵ- greedy policy.



As there is no noise in the enviroment, we can make the above transition diagram. Therefore we can clearly ~~state~~ say that first ~~trans~~ transition is random because there exists a better choice of action(a_2) which will give us greater reward.

Nothing can be said about second transition because of lack of information. Therefore it might be greedy or random.