RL - Assgnt 2

2017164

Ans 5)

$$\sum_{a \in A(S_t)} \pi(a|S_t)\, q_\pi(S_t, a) = V_\pi(S_t)$$

Now when $\pi = \pi^*$ optimal

then $V_\pi = V_*$, $q_\pi = q_*$.

and π^* assigns non-zero probabilities only to those actions ($a' \in A(S_t)$) which assign the maximum value to $q_*(S_t, a)$

∴ we can write this as :-

$$V_*(S_t) = \sum_{a \in A(S_t)} \pi^*(a|S_t)\, q_*(S_t, a)$$

$$= \sum_{a = \arg\max_{a \in A(S_t)}} \pi^*(a|S_t) \max(q_*(S_t, a))$$

here 'a' is that value of 'a' that maximizes $q_*(S_t, a)$ and that max value of $q_*(S_t, a)$ is assigned to $V_*(S_t)$

const. so take out of summation

$$= \max_{a = \arg\max_{a \in A(S_t)}} (q_*(S_t, a)) \sum_{a = \arg\max_{a \in A(S_t)}} \pi^*(a|S_t)$$

$$\sum_{a = \arg\max_{a \in A(S_t)}} \pi^*(a|S_t) = 1$$

$$\Rightarrow \boxed{V_*(S_t) = \max_{a' \text{ such that } a' = \arg\max_{a \in A(S_t)}} (q_*(S_t, a'))}$$

Ans 1)

current state (s)	actions $a \in A(s)$	next state (s')	rewards (r)	$p(s', r \mid s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1-\alpha$
low	search	high	-3	$1-\beta$
low	search	low	r_{search}	β
high	wait	high	r_{wait}	1
low	wait	low	r_{wait}	1
low	recharge	high	0	1

The table was obtained by using the figure of "transition graph" given in the book for that problem.

Also $\sum_{s',r} p(s', r \mid s, a) = 1 \quad \forall s, a; s \in S$ and $a \in A(s)$. So it is right

Ans 3) (3.15) and 3.16))

Signs of the rewards are irrelevant because they can be changed by adding appropriate const. 'c' to all rewards and rewards are something that we decide. Now the intervals of the values are important as they show the relative preferences of states. And those intervals can get affected in different ways by adding a constant 'c', based on the task is episodic or continuous.

for Episodic :-

$$V_\pi'(s) = E_\pi[G_t' \mid S_t = s], \quad G_t = R_{t+1} + R_{t+2} \cdots + R_T$$

$$G_t' = \sum_{k=1}^{M} \gamma^k (R_{t+k+1} + c) \rightarrow \text{terminates after some M steps.}$$

$$G_t' = \underbrace{\sum_{k=1}^{M} \gamma^k R_{t+k+1}}_{= G_t} + c\left(\frac{1-\gamma^M}{1-\gamma}\right) \rightarrow \text{GP formula}$$

$$\Rightarrow V_\pi'(s) = E_\pi\left[G_t + c\left(\frac{1-\gamma^M}{1-\gamma}\right) \mid S_t = s\right]$$

$$= E_\pi[G_t \mid S_t = s] + E\left[c\left(\frac{1-\gamma^M}{1-\gamma}\right) \mid S_t = s\right]$$

$$V_\pi'(s) = \boxed{V_\pi(s) + c\left(\frac{1-\gamma^M}{1-\gamma}\right)}$$

$\hookrightarrow$ const.

$M < \infty$ with probability = 1 and $0 \le \gamma < 1$

So, if M is smaller, then $c\left(\frac{1-\gamma^M}{1-\gamma}\right)$ will be smaller, so, states closer to termination will see smaller increase in their preference.

for continuous task :- $G'_t = G_t + \sum_{k=0}^{\infty} \gamma^k C = G_t + \frac{C}{1-\gamma}$

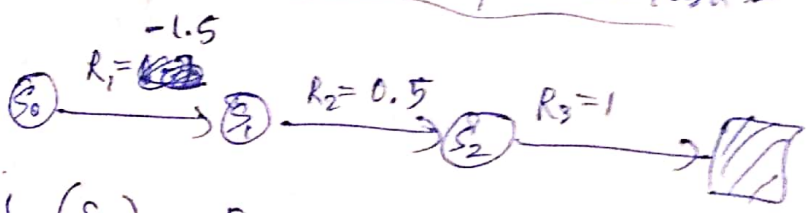
$$V'_\pi(s) = E_\pi \left\{ G_t + \frac{C}{1-\gamma} \mid S_t = s \right\}$$

$$\boxed{V'_\pi(s) = V_\pi(s) + \left(\frac{C}{1-\gamma}\right)}$$

So, In this case, each values in offset by the same constant $\frac{C}{1-\gamma}$, so the relative preferences of states doesn't change.

Example of the effect in Episodic task :-

Suppose given



So here $V_\pi(S_0) = 0$, $V_\pi(S_1) = -1.5$, $V_\pi(S_2) = 1$

now let $C = 100$ and $\gamma = 0.1$ so, $V'_\pi = V_\pi + C\left(\frac{1-\gamma^M}{1-\gamma}\right)$

so, for $V'_\pi(S_0) = 100\left(\frac{1-(0.1)^3}{0.9}\right) + 0 = 111$

$V'_\pi(S_1) = 100\left(\frac{1-(0.1)^2}{0.9}\right) + 1.5 = 111.5$

$V'_\pi(S_2) = 100\left(\frac{1-(0.1)}{0.9}\right) + 1 = 101$

$$\boxed{\text{So, } V_\pi(S_2) > V_\pi(S_0) \text{ but } V'_\pi(S_2) < V'_\pi(S_0)}$$