DATE / / 400109843 SUBJECT : پیاده‌سازی

سوال ① :

الف) درست | در مسائل RL، با توجه به اینکه ابتدا باید تعدادی sample از محیط جمع‌آوری تعامل با محیط به صورت online انجام می‌شود.

ب) نادرست. در مسائل MDP، در نهایت به دنبال سیاستی هستیم که بیشترین reward را داشته باشد. در نهایت اگر در حالتی $Q(s,a) = Q(s,a')$ باشد، با دو سیاست متفاوت به یک مقدار رسیدیم، در نتیجه سیاست بهینه لزوماً یکتا نیست.

ج) نادرست. الگوریتم REINFORCE در واقع یک الگوریتم Model-free است. زیرا به دنبال پیدا کردن یک policy بهینه است و $\hat{T}$ و $\hat{R}$ را پیدا نمی‌کند.

د) درست. به صورت کلی $V^*(s)$، در واقع جواب بهینه برای $V(s)$ است در نتیجه باید حالتی $V^\pi(s)$ باشد و در نتیجه $V^*(s) \geq V^\pi(s)$

DATE / / SUBJECT :

سؤال (2):

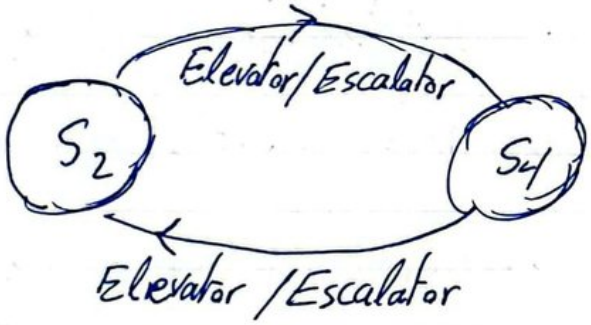
آ) با توجه به اینکه محدود Q-Value را داریم Infinite Horizon بودن، می‌توانیم با استفاده از الگوریتم Policy Extraction، سیاست بهینه را پیدا کنیم

$$\forall s \in S : \pi^*(s) = \underset{a}{\operatorname{argmax}}\, Q^*(s,a)$$

$\pi^*(S_1)$ = Elevator $\pi^*(S_2)$ = Elevator

$\pi^*(S_3)$ = Escalator

ب)



ج) مانند قسمت الف برای به دست آوردن سیاست بهینه از Policy Extraction استفاده می‌کنیم:

$\pi^*(S_2)$: Escalator $\pi^*(S_4)$: Escalator

د) فرض مارکف به این صورت است که احتمال رسیدن به 's' در زمان $t+1$ تنها به ایالت s و اکشن a در زمان t بستگی دارد. (بی حافظگی)

در اینجا با توجه به اینکه از Q-learning استفاده کردیم ممکن است که داده‌هایی که جمع کردیم به اندازه کافی نباشد. در نتیجه بی حافظگی در فرض مارکو ممکن است که رعایت نشده باشد. در نتیجه سیاست بهینه تحویل می‌شود.

DATE / / SUBJECT :

سوال (3):

(آ

$$E_{s\sim P(s),\, a\sim\pi_0(s,a)} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a) =$$

~~$= \sum R(s,a) = \sum \frac{\pi_1(s,a)}{\hat{\pi}}$~~

$$= \sum \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a) P(s) P(a|s) \xrightarrow{\pi_0(s,a)} \longrightarrow$$

$$\xrightarrow{\hat{\pi}_0(s,a) = \pi_0(s,a)} = \sum \pi_1(s,a) R(s,a) P(s) \Rightarrow$$

$$\Rightarrow E_{s\sim P(s),\, a\sim\pi_0(s,a)} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} R(s,a) = E_{s\sim P(s),\, a\sim\pi_1(s,a)} R(s,a)$$

تخمین سازگاری: $\forall \varepsilon > 0 : P(|\hat{\theta} - \theta| > \varepsilon) = 0$

با توجه به اینکه تخمین گر داده شده تابع R برای π_1 را دقیق تخمین می‌زند، یک تخمین سازگار است.

(ب صورت گزاره را در قسمت الف) بدست آوردیم برای مخرج نیز به صورت زیر عمل می‌کنیم:

$$E_{s\sim P(s),\, a\sim\pi_0(s,a)} \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} = \sum \frac{\pi_1(s,a)}{\hat{\pi}_0(s,a)} P(s) \pi_0(s,a) =$$

$$= \sum P(s) \pi_1(s,a) = \sum P(s) P(a|s) = 1 \Rightarrow$$

DATE / / SUBJECT :

$$\frac{E_{S\sim P(S),\, a\sim\pi_0(S,a)} \frac{\pi_1(S,a)}{\pi_0(S,a)} R(S,a)}{E_{S\sim P(S),\, a\sim\pi_0(S,a)} \frac{\pi_1(S,a)}{\hat{\pi}_0(S,a)}} = E_{S\sim P(S),\, a\sim\pi_1(S,a)} R(S,a)$$

در نتیجه این تخمین گر، تخمین گری سازگار است.

ج) فرض می کنیم تنها یک داده داشته باشیم، در این صورت مقدار تخمینگر ها برابر با $R_{\pi_0}(S,a)$ که از توزیع احتمالی π_0 پیروی می کند در نتیجه $E(R_{\pi_1}(S,a)) \neq E(R_{\pi_0}(S,a))$ که بیانگر این است که این تخمینگر می تواند بایاس باشد.

DATE / / SUBJECT :

سوال ④ :

$$U^{\pi}_{(s)} = R_{(s)} + \gamma \sum_{s'} T_{(s,\pi(s),s')} U^{\pi}_{(s')}$$ (۱

$U^{\pi}_{(s)}$ با توجه به اینکه مقدار Reward تنها بستگی به s دارد. می توان

M: Mountain R: Riverside D: Desert ، π: Peace را به صورت بالا نوشت.

$$U^{\pi}_{(M)} = 1 + \gamma\{\tfrac{1}{2} \times U^{\pi}_{(D)} + \tfrac{1}{2} \times U^{\pi}_{(R)}\}$$

$$U^{\pi}_{R} = 2 + \gamma\{1 \times U^{\pi}_{(R)}\} \Rightarrow U^{\pi}_{(R)} = \frac{2}{1-\gamma}$$

$$U^{\pi}_{(D)} = -1 + \gamma\{1 \times U^{\pi}_{(D)}\} \Rightarrow U^{\pi}_{(D)} = -\frac{1}{1-\gamma}$$

$$U^{\pi}_{(M)} = 1 + \frac{\gamma}{2} \times \left\{\frac{2}{1-\gamma} + \left(-\frac{1}{1-\gamma}\right)\right\} = 1 + \frac{\gamma}{2(1-\gamma)} = \frac{2-\gamma}{2(1-\gamma)}$$

(←

π_0 : Peace $\rightarrow U^{\pi_0}_{(M)} = 0.5 \quad U^{\pi_0}_{(R)} = 20 \quad U^{\pi_0}_{(D)} = -10$

$$\pi_{1(s)} = \underset{a}{\operatorname{argmax}} \sum_{s'} T_{(s,a,s')} \{R_{(s,a,s')} + \gamma U^{\pi_0}_{(s')}\}$$

↳ Policy Iteration

DATE / / SUBJECT :

$\pi_{1(M)}$: ~~Max~~ argmax (peace: $1 + 0,9 \times (\frac{1}{2} \times 20 - \frac{1}{2} \times 10)$,

, war: $1 + 0,9 \times (0,1 \times 5,5 + 0,7 \times 20 - 0,2 \times 10)$)

$\Rightarrow \pi_{1(M)}$: War

$\pi_{1(D)}$: argmax (peace: $-1 + 0,9 \times (1 \times -1)$, war: $-1 + 0,9 \times (1 \times 5,5)$)

$\pi_1(D)$: war

$\pi_{1(R)}$: argmax (peace: $2 + 0,9 \times (1 \times 20)$,

war: $2 + 0,9 \times (0,2 \times 20 - 0,8 \times 10)$)

$\Rightarrow \pi_1(R)$: Peace

Sample : $R(s,a,s') + \gamma \max_{a'} (Q(s';a'))$ (2.

$Q(s,a) \leftarrow (1-\alpha) Q(s,a) + \alpha \cdot \text{Sample}$

Mountain - peace	Riverside - peace	Desert - war
0	0	0
0	0	-1
0	0	1
0	0,5	1
0,75	0,5	1