علی آقایاری ۴۰۰۱۲۷۱۵

سوال ۱) الف) درست است، در مسائل RL، agent به صورت online فعالیت می‌کند و با بهبود خطا policy را تقویت می‌کند

ب) نخیر، اگر داشته باشیم $Q(s,a) = Q(s,b)$، هردو اکشن a و b می‌توانند جزو policy بهینه باشند برای همین policy بهینه یکتا نیست

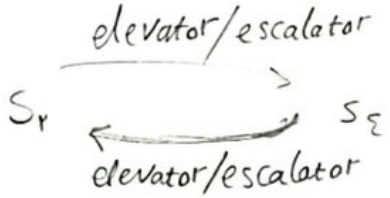
ج) نخیر، در این الگوریتم‌های خواهیم policy بهینه را مستقیم محاسبه کنیم برای همین این الگوریتم model free است.

د) درست، V^* بهینه است به همین دلیل V^π (π بهینه/غیربهینه) همیشه کمتر مساوی V^* است

سوال ۲) مانند policy extraction اکشنی که بهترین Q-value را دارد را انتخاب می‌کنیم پس:

$\pi^*(s_1) = elevator$ $\pi^*(s_2) = elevator$ $\pi^*(s_3) = escalator$

ب)



ج) مانند قسمت الف

$\pi^*(s_2) = escalator$

$\pi^*(s_4) = escalator$

د) فرض مارکوف این است که حرکت‌های گذشته و آینده از هم مستقل‌اند و هر state تمام اطلاعات مورد نیاز برای تصمیم‌گیری را دارد و نیازی به اطلاعات actionهای گذشته ندارد اما در اینجا ما فضا را کاهشی داده‌ایم و مقداری دنیا از دست داده‌ایم به همین دلیل ممکن است که ارتقالی بین اکشن‌های گذشته و آینده وجود نداشته باشد، به همین دلیل ممکن است به policy متفاوتی برسیم.

ع) الف) داریم:

$$U^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s, \pi(s), s') U^{\pi}(s')$$

$\longrightarrow$ $U^{\pi_0}(M) = 1 + \gamma\left(\frac{1}{2} U^{\pi_0}(R) + \frac{1}{2} U^{\pi_0}(D)\right)$

$\longrightarrow$ $U^{\pi_0}(R) = 2 + \gamma(1 \times U^{\pi_0}(R))$

$\longrightarrow$ $U^{\pi_0}(D) = -1 + \gamma(1 \times U^{\pi_0}(D))$

$$U^{\pi_0}(R) = \frac{2}{1-\gamma}$$

$$U^{\pi_0}(D) = \frac{-1}{1-\gamma}$$

$$U^{\pi_0}(M) = \frac{2-\gamma}{2-2\gamma}$$

ب)

$\pi_0 = P$ $\quad U^{\pi_0} = (\underbrace{5.5,\ 20,\ -10}_{M,\ R,\ D})$

$\pi_1(M) = \arg\max\left(P:\ 1 + 0.9\left(\frac{1}{2} \times 20 - \frac{1}{2} \times 10\right),\right.$

$\left. W:\ 1 + 0.9(0.1 \times 5.5 - 0.2 \times 10 + 0.7 \times 20)\right)$

$\longrightarrow$ $\pi_1(M) = W$

$\pi_1(R) = \arg\max\left(P:\ 2 + 0.9 \times 20,\ W:\ 2 + 0.9(0.2 \times 20 + 0.8 \times 5.5)\right)$

$\longrightarrow$ $\pi_1(R) = P$

$\pi_1(D) = \arg\max\left(P:\ -1 + 0.9 \times (-10),\ W:\ -1 + 0.9 \times 5.5\right)$

$\longrightarrow$ $\pi_1(D) = W$

$$\text{sample} = R(s,a,s') + \gamma \max_{a'} Q(s',a')$$

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

sample 1 = (D, W, D, -٢)
sample 2 = (D, W, R, ٣)
sample 3 = (R, P, M, 1)
sample 4 = (M, P, R, 1)

$Q(D,W) = 0 + \frac{1}{2}(-2 + 0 - 0) = -1$

$Q(D,W) = -1 + \frac{1}{2}(3 + 0 + 1) = +1$

$Q(R,P) = 0 + \frac{1}{2}(1 + 0 - 0) = \frac{1}{2}$

$Q(M,P) = 0 + \frac{1}{2}(1 + \frac{1}{2} - 0) = 1,75$

M-P	R-P	D-W
0	0	0
0	0	-1
0	0	+1
0	1/2	+1
3/4	1/2	+1