تمرین سوم

۹۰۰۱۰۹۹۶۹ — آیلین رسته

۱) الف) درست. مسائل RL مسائل online هستند و عامل با انجام عمل به درستی یا غلطی آن پی می‌برد.

ب) نادرست. در صورتی که $Q(s,a) = Q(s,a')$ هر دو اکشن a و a' می‌توانند در policy بهینه باشند. در نتیجه پالیسی بهینه یکتا نیست.

ج) نادرست. در الگوریتم reinforce ما به دنبال به دست آوردن transition function یا reward function نیستیم و می‌خواهیم policy بهینه را به صورت مستقیم به دست آوریم. در نتیجه این الگوریتم model-free است.

د) درست. زیرا V^* همیشه بهینه است و value را در حالتی که agent بهینه عمل می‌کند اندازه می‌گیرد و مقدار آن از مقدار value هر پالیسی دیگری بیشتر است.

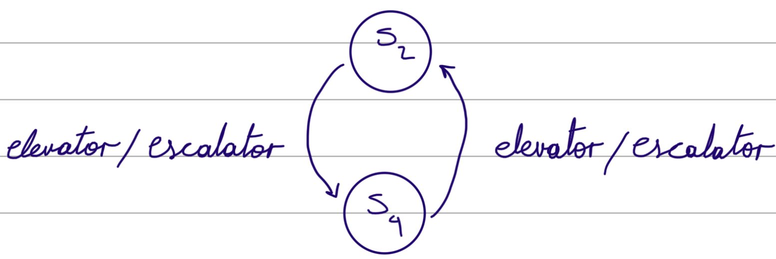
Q-values	Elevator	Escalator
$S_۱$	+۱/۵	−۰/۵
$S_۲$	+۲/۰	−۰/۳
$S_۳$	+۰/۸	+۰/۹

۲) الف) به کمک الگوریتم policy extraction می‌دانیم که باید اکشنی را انتخاب کنیم که ما را به Q-state ای با بیشترین Q_value ببرد. در نتیجه π^* به صورت روبه‌رو می‌شود:

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

$\pi^*(S_3) = escalator$

ب)



Q^*	Elevator	Escalator
S_2	$+0/08$	$+2/08$
S_4	$+1/18$	$+1/38$

ج) همانند قسمت الف عمل می‌کنیم و داریم:

$\pi^*(S_2) = \text{escalator}$ $\pi^*(S_4) = \text{escalator}$

د) فرض مارکوف بیان می‌کند که حرکت‌های گذشته و آینده از هم مستقل هستند و هر استیت اطلاعات کامل و کافی برای پیش‌بینی حرکت بعدی خود دارد و نیازی به اطلاعات استیت‌ها و اکشن‌های گذشته ندارد. در اینجا ما فضای حالت را ماهش دادیم و مقداری دیتا در مورد استیت‌ها از دست داده‌ایم در نتیجه استقلال حرکت‌های گذشته و آینده در اینجا معلوم است برقرار نباشد و با انجام الگوریتم بالا به policy متفاوتی برسیم.

3) الف) برای اثبات سازگاری تخمینگر باید ثابت کنیم:

$$\lim_{n\to\infty} p(|\hat{\theta}-\theta|>\varepsilon)=0$$

به عبارتی

$$E_{s\sim p(s),\, a\sim\pi_0(s,a)} \frac{\pi_1(s,a)}{\pi_0(s,a)} R(s,a) = \sum \frac{\pi_1(s,a)}{\pi_0(s,a)} \times R(s,a) \times p(s) \times \pi_0(s,a)$$

به برابر صورت سوال است که

$$= \sum R(s,a)\, p(s)\, \pi_1(s,a) = E_{s\sim p(s),\, a\sim\pi_1(s,a)} R(s,a)$$

ب)

$$E_{s\sim p(s),\, a\sim\pi_0(s,a)} \frac{\pi_1(s,a)}{\pi_0(s,a)} = \sum \frac{\pi_1(s,a)}{\pi_0(s,a)} \times p(s) \times \pi_0(s,a) = \sum \pi_1(s,a) \times p(s)$$

حرج هم برابر است صورت هم مورد الف است که

$$= \sum p(a|s)p(s) = 1$$

اثبات کردیم در نتیجه این تخمینگر سازگار است.

ج) وقتی فقط یک نمونه داشته باشیم تخمینگر ما برابر $R_{\pi_0}(s,a)$ می‌شود که از توزیع π_0 پیروی می‌کند. در نتیجه اگر $R_{\pi_0}(s,a)$ برابر $E[R_{\pi_1}(s,a)]$ نمی‌شود تخمینگر ما بایاس دارد.

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

(الف) 9

$$U^{\pi_0}(M) = 1 + \gamma(0.5 \times U^{\pi_0}(R) + 0.5 \times U^{\pi_0}(D))$$

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

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

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

π_0 = peace $\qquad U^{\pi_0} = (\overset{M}{5.5}, \overset{R}{20}, \overset{D}{-10})$

(ب)

$$\pi_1(M) = \text{argmax}(\text{peace}: 1 + 0.9(0.5 \times 20 - 0.5 \times 10),\ \text{war}: 1 + 0.9(0.1 \times 5.5 - 0.2 \times 10 + 0.7 \times 20))$$

$\pi_1(M)$ = war

$$\pi_1(R) = \text{argmax}(\text{peace}: 2 + 0.9 \times 20,\ \text{war}: 2 + 0.9(0.2 \times 20 + 0.8 \times 5.5))$$

$\pi_1(R)$ = peace

$$\pi_1(D) = \text{argmax}(\text{peace}: -1 + 0.9 \times -10,\ \text{war}: -1 + 0.9 \times 5.5)$$

$\pi_1(D)$ = war

$$\text{Sample} = R(s, a, s') + \gamma \max_{a'} Q(s', a') \qquad Q(s,a) \leftarrow (1-\alpha)Q(s',a') + \alpha \cdot \text{sample}$$

(ج)

(فرض کنید $\gamma = ۱, \alpha = ۰/۵$)

Start State	Action	State End	Reward
Desert	War	Desert	$-۲/۰$
Desert	War	Riverside	$۳/۰$
Riverside	Peace	Mountain	$۱/۰$
Mountain	Peace	Riverside	$۱/۰$

Mountain-peace	Riverside-peace	Desert-war
۰	۰	۰
0	0	-1
0	0	1
0	0.5	1
0.75	0.5	1