محمد پیام تابش ۴۰۰۱۰۴۸۶۷   تمرین سوم هوش

۱) الف) صحیح است در RL، T و R را نمی‌دانیم و در حالت

active RL مستقیما خودمان باید episode ها را شروع کرده و بررسی کنیم

و در حالت passive RL هم یک π به عنوان behavioral policy داریم

که طبق آن sample جمع می‌کنیم البته اگر سخت‌گیری کنیم می‌توان

گزاره را غلط گرفت چون حالتی هم هست که بدون اینکه ما کاری در

محیط بکنیم با داشتن تعدادی sample از π مشخص شده که شخص

دیگری قبلا جمع کرده است می‌توان π^* را پیدا کرد ولی در آن حالت هم

شخص دیگری قبلا به شکل آنلاین آن sample ها را به دست آورده

ب) غلط حالتی را فرض کنید که تمامی state ها و Reward ها

مثل هم باشد در این صورت هر پالیسی یک پالیسی بهینه است

ج) غلط، این الگوریتم فارغ از تخمین زدن T و R مستقیما

به دنبال تخمین زدن π^* است در کل RL به دو دسته‌ی

model-base و model-free تقسیم می‌شود که تقریبا تمام الگوریتم‌های کاربردی

آن model-free هستند

د) صحیح. طبعا π^* از تمامی π های دیگر بهتر است پس v^*

قطعا از v^2 بیشتر خواهد بود که به راحتی با برهان خلف قابل اثبات

است: فرض کنید به ازای s داریم: $v^*_{(s)} < v^2_{(s)}$

یعنی وجود دارد a به طوری که $v^2_{(s)} = Q^2_{(s,a)}$.

پس $v^*_{(s)} < Q^{\pi}_{(s,a)}$ که خب یعنی v^*_s اشتباه است.

شود و می‌توان حداقل یک گام policy improvement داشت که فرض

را نقض می‌کند. از طرفی در policy improvement هم صرفا v صعودی است

و بعد از converge کردن دیگر تغییر نمی‌کند و آن بیشینه مقدارهای v^*

است.

۳) الف

$$\pi(s) = \arg\max_a Q(s,a)$$

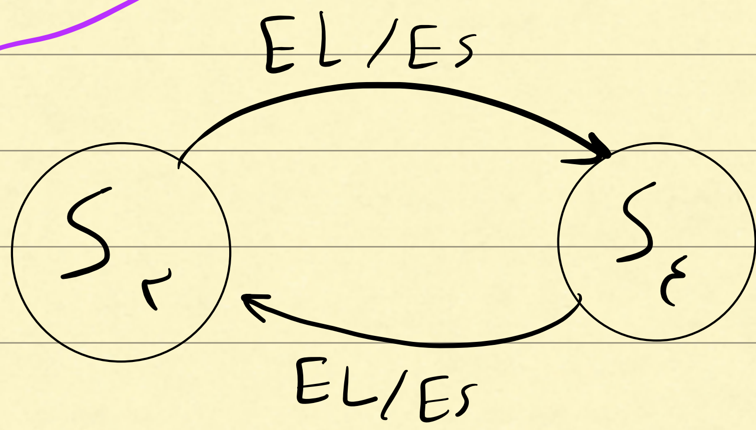
EL → Elevator

Es → Escalator

$\pi(s_1) = EL$ $\quad \pi(s_4) = Es$

$\pi(s_2) = EL$

ب)



ج)

$\pi(s_2) = Es \quad \pi(s_4) = Es$

د) مارکو فرض می‌کنه که هر حالت فقط تابع حالت قبلی خوداست و به همه حالات

قبلی بستگی نداره و اینجا چون تعداد state ها را کم کردیم طبیعتا

مقداری داده را از دست دادیم و مدلسازی اول کامل‌تر است از طرفی

۱ و ۲ تفاوت دارند مثلا اگر ۱ و ۲ و ۱ و ۲ را دارای ۲ حالت مقایسه کنیم این تفاوت

کاملا واضح است در واقع در حالت دوم استقلال حالت های گذشته و آینده که state لزوما برقرار نیست

۳) الف)

$$E_{s\sim P(s),\, a\sim \pi_0(s,a)}\left[\frac{\pi_1(s,a)}{\pi_0(s,a)} R_{(s,a)}\right] =$$

$$\sum \frac{\pi_1(s,a)}{\pi_0(s,a)} R_{(s,a)} \times P_{(s)} P_{(a|s)} =$$

$$\sum \frac{\pi_1(s,a)}{\cancel{\pi_0(s,a)}} R_{(s,a)} P_{(s)} \cancel{\pi_0(s,a)} =$$

$$\sum R_{(s,a)} P_{(s)} \pi_1(s,a) = E_{s\sim P(s),\, a\sim \pi_1(s,a)}\left[R(s,a)\right]$$

یعنی این ۲ امید ریاضی با هم برابر هستند

تخمین گر معرفی شده importance sampling estimator است که به شکل :

$$\hat{J} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_1(s,a)}{\pi_0(s,a)} R_{(s,a)}$$

هست که طبق قانون اعداد بزرگ در $n \to \infty$

$$\hat{J} = E_{s\sim P(s),\, a\sim \pi_0(s,a)}\left[\frac{\pi_1(s,a)}{\pi_0(s,a)} R_{(s,a)}\right]$$

برای نشان دادن سازگاری باید نشان دهیم که

$$\lim_{n\to\infty} \left(\left| \hat{J} - \mathop{E}_{s\sim p(s),\, a\sim \pi_1(s,a)} \left[R(s,a) \right] \right| > \varepsilon \right) = 0$$

که می‌توان به جای $\hat{J}$ ، $\mathop{E}_{s\sim p(s),\, a\sim \pi_0(s,a)} \left[\frac{\pi_1(s,a)}{\pi_0(s,a)} R(s,a) \right]$

را گذاشت و این یعنی کافی است نشان دهیم امید ریاضی این

۲ یکی است یعنی همین چیزی که اثبات کردیم

$$\mathop{E}_{s\sim p(s),\, a\sim \pi_0(s,a)} \left[\frac{\pi_1(s,a)}{\pi_0(s,a)} R(s,a) \right] = \mathop{E}_{s\sim p(s),\, a\sim \pi_1(s,a)} \left[R(s,a) \right]$$

پس این تخمین‌گر سازگار است

ب) این تخمین‌گر به شکل

$$\hat{J} = \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{\pi_1(s,a)}{\pi_0(s,a)} R(s,a)}{\frac{1}{n} \sum_{i=1}^{n} \frac{\pi_1(s,a)}{\pi_0(s,a)}}$$

مثل بخش الف باید نشان داد که

$$\lim_{n\to\infty} \left(\left| \hat{J} - \mathop{E}_{s\sim p(s),\, a\sim \pi_1(s,a)} \left[R(s,a) \right] \right| > \varepsilon \right) = 0$$

پس داریم

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_1(s,a)}{\pi_0(s,a)} R(s,a) =$$

طبق الف

$$\mathbb{E}_{s\sim P(s),\, a\sim \pi_0(s,a)} \left[\frac{\pi_1(s,a)}{\pi_0(s,a)} R(s,a) \right] = \mathbb{E}_{s\sim P(s),\, a\sim \pi_1(s,a)} \left[R(s,a) \right]$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_1(s,a)}{\pi_0(s,a)} = \mathbb{E}\left[\frac{\pi_1(s,a)}{\pi_0(s,a)} \right]$$

$$= \sum \frac{\pi_1(s,a)}{\pi_0(s,a)} P(s) P(a|s) = \sum \frac{\pi_1(s,a)}{\cancel{\pi_0(s,a)}} P(s) \cancel{\pi_0(s,a)}$$

$$= \sum \pi_1(s,a) P(s) = 1$$ $\rightarrow$ جمع همه احتمالات

$$\Rightarrow \lim_{n\to\infty} \hat{J} = \frac{\mathbb{E}_{s\sim P(s),\, a\sim \pi_1(s,a)} \left[R(s,a) \right]}{1} = \mathbb{E}_{s\sim P(s),\, a\sim \pi_1(s,a)} \left[R(s,a) \right]$$

پس این مورد نیز نشان داده شد برابر همان امید ریاضی صورت سوال است و این تخمینگر نیز سازگار است.

ج) کافی است که $n=1$ باشد (فقط یک نمونه داشته باشیم)

a_1 و s_1 از همان یک نمونه هستند

$$\hat{J} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_1(s,a)}{\pi_0(s,a)}R(s,a)}{\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_1(s,a)}{\pi_0(s,a)}} = R(s_1,a_1)$$

در این حالت s_1 و a_1 از توزیع π_0 آمده اند و با E_{π_0} برابر

است و با E_{π_1} برابر نیست که یعنی bias داریم

د) ابتدا فرض می‌کنیم این R با $R(s)$ باشد یعنی هر

وقت وارد یک state می‌شویم این ریوارد

را دریافت می‌کنیم:

$\forall_{(s)} : \pi_{(s)} = \text{peace}$ الف

$M \rightarrow$ mountain $\quad R \rightarrow$ Riverside $\quad D \rightarrow$ Desert

$$V^{\pi}_{(s)} = \sum_{s'} T_{(s,\, \pi_{(s)},\, s')} \left[R_{(s,\, \pi_{(s)},\, s')} + \gamma V^{\pi}_{(s')} \right]$$

$$\rightarrow V^{\pi}_{(M)} = \frac{1}{2}\left[2 + \gamma V^{\pi}_{(R)}\right] + \frac{1}{2}\left[-1 + \gamma V^{\pi}_{(D)}\right]$$

$$V^{\pi}_{(R)} = \left[2 + \gamma V^{\pi}_{(R)}\right] \rightarrow V^{\pi}_{(R)} = \frac{2}{1-\gamma}$$

$$V^{\pi}_{(D)} = \left[-1 + \gamma V^{\pi}_{(D)}\right] \rightarrow V^{\pi}_{(D)} = -\frac{1}{1-\gamma}$$

$$\rightarrow V^{\pi}_{(M)} = \frac{1}{2}\left[2 + \frac{2\gamma}{1-\gamma}\right] - \frac{1}{2}\left[+1 + \frac{\gamma}{1-\gamma}\right]$$

$$= \frac{1}{1-\gamma} - \frac{1}{2}\,\frac{1}{1-\gamma} \rightarrow V^{\pi}_{(M)} = \frac{1}{2(1-\gamma)}$$

ب) در بخش الف π را evaluate کردیم که اگر $\gamma = 0.9$ باشد:

$$\boxed{V^{\pi}_{(M)} = +5, \quad V^{\pi}_{(D)} = -10 \quad V^{\pi}_{(R)} = 20}$$

در این بخش policy improvement می‌زنیم یعنی:

$$\pi_{i+1}(s) = \underset{a}{\arg\max}\ Q^{\pi_i}(s,a)$$

Peace → p

war → w

$$, \quad Q^{\pi_i}(s,a) = \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma v^{\pi_i}(s')\right]$$

$$\rightarrow \pi_{i+1}(s) = \underset{a}{\arg\max} \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma v^{\pi_i}(s')\right]$$

برای ساده سازی

$$\forall_{(s,a)}:\ R(s,a,M) + \gamma v^{\pi}(M) = ۱ + \frac{۹}{۱۰} \times ۵ = ۵/۵$$

$$\forall_{(s,a)}:\ R(s,a,D) + \gamma v^{\pi}(D) = -۱ + \frac{۹}{۱۰} \times -۱۰ = -۱۰$$

$$\forall_{(s,a)}:\ R(s,a,R) + \gamma v^{\pi}(R) = ۴ + \frac{۹}{۱۰} \times ۴۰ = ۴۰$$

$$Q^{\pi}(M,p) = \frac{۱}{۲} \times ۴۰ + \frac{۱}{۲} \times -۱۰ = +۵$$

$$Q^{\pi}(M,w) = \frac{۱}{۱۰} \times ۵/۵ + \frac{۲}{۱۰} \times -۱۰ + \frac{۷}{۱۰} \times ۴۰ = ۱۲/۵۵$$

$$\rightarrow \boxed{\pi^{i+1}(M) = \text{war}}$$

$$Q^{\pi}(R,p) = ۴۰ \quad , \quad Q^{\pi}(R,w) = \frac{۲}{۱۰} \times ۴۰ + \frac{۸}{۱۰} \times ۵/۵ = ۸/۴$$

$$\rightarrow \boxed{\pi^{i+1}(R) = \text{peace}}$$

$Q^2_{(D,P)} = -10$, $Q^2_{(D,W)} = 5.5$

$\rightarrow \boxed{\pi^{i+1}_{(D)} = \text{war}}$

ج) الگوریتم Q-learning را پیاده سازی می کنیم:

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

$$Q_{(s,a)} = Q_{(s,a)} + \alpha\left[\text{sample} - Q_{(s,a)}\right]$$

طبق این اپیزود ورودی، داریم:

ابتدا همه $Q_{(s,a)}$ ها صفر هست

$\text{Sample}_1 = (D, W, D, -2)$ $\quad$ $\text{sample}2 = (D, W, R, 3)$

$\text{Sample}_3 = (R, P, M, 1)$ $\quad$ $\text{sample}4 = (M, P, R, 1)$

sample 1:

$$Q_{(D,W)} = 0 + \frac{1}{2}\left[-2 + 0 - 0\right] = -1$$

sample 2:

$$Q_{(D,W)} = -1 + \frac{1}{2}\left[3 + 0 + 1\right] = +1$$

sample 3 :

$$Q_{(R,P)} = ۰ + \frac{۱}{۲}\left[۱ + ۰ - ۰\right] = ۰/۵$$

Sample 4 :

$$Q_{(M,P)} = ۰ + \frac{۱}{۲}\left[۱ + \frac{۱}{۲} - ۰\right] = ۰/۷۵$$

→ Q_table :

M - P	R_ P	D- W
۰	۰	۰
۰	۰	-۱
۰	۰	+۱
۰	۰/۵	+۱
۰/۷۵	۰/۵	+۱

حالا همین سوال را به شرطی حل می‌کنیم که R ، $R_{(s)}$ باشد

یعنی وقتی از state خارج می‌شویم این ریوارد

را دریافت کنیم :

الف

$$\forall_{(s)} : \pi_{(s)} = \text{peace}$$

$M \rightarrow$ mountain $\quad R \rightarrow$ Riverside $\quad D \rightarrow$ Desert

$$V^{\pi}_{(s)} = \sum_{s'} T_{(s,\pi_{(s)},s')}\left[R_{(s,\pi_{(s)},s')} + \gamma V^{\pi}_{(s')}\right]$$

$$\rightarrow V^{\pi}_{(M)} = \frac{۱}{۲}\left[۱ + \gamma V^{\pi}_{(R)}\right] + \frac{۱}{۲}\left[۱ + \gamma V^{\pi}_{(D)}\right]$$

$$V^{\pi}_{(R)} = \left[۲ + \gamma V^{\pi}_{(R)}\right] \rightarrow V^{\pi}_{(R)} = \frac{۲}{۱-\gamma}$$

$$V^{\pi}_{(D)} = \left[-۱ + \gamma V^{\pi}_{(D)}\right] \rightarrow V^{\pi}_{(D)} = -\frac{۱}{۱-\gamma}$$

$$\rightarrow V^{\pi}_{(M)} = ۱ + \frac{\gamma}{۲} \times \frac{۱}{۱-\gamma} = \frac{۲-۲\gamma+\gamma}{۲-۲\gamma}$$

$$= \frac{۲-\gamma}{۲-۲\gamma} \longrightarrow V^{\pi}_{(M)} = \frac{۲-\gamma}{۲-۲\gamma}$$

ب) در بخش الف π را evaluate کردیم که اگر $\gamma = ۰/۹$ باشد:

$$\boxed{V^{\pi}_{(M)} = ۵/۵,\quad V^{\pi}_{(D)} = -۱۰ \quad V^{\pi}_{(R)} = ۲۰}$$

در این بخش policy improvement می زنیم یعنی:

$\pi_{i+1}(s) = \underset{a}{\operatorname{argmax}}\, Q^{\pi_i}(s,a)$ Peace → p

war → w

$, Q^{\pi_i}(s,a) = \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^{\pi_i}(s')\right]$

$\rightarrow \pi_{i+1}(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^{\pi_i}(s')\right]$

$Q^{\pi}(M,P) = \frac{۱}{۲}\left[۱+\frac{۹}{۱۰}\times ۲۰\right] + \frac{۱}{۲}\left[۱+\frac{۹}{۱۰}\times -۱۰\right] = ۵/۵$

$Q^{\pi}(M,W) = \frac{۱}{۱۰}\left[۱+\frac{۹}{۱۰}\times ۵/۵\right] + \frac{۲}{۱۰}\left[۱+\frac{۹}{۱۰}\times -۱۰\right] + \frac{۷}{۱۰}\left[۱+\frac{۹}{۱۰}\times ۲۰\right]$

$= ۰/۵۹۵ + -۱/۶ + ۱۳/۳ = ۱۲/۲۹۵$

$\rightarrow \boxed{\pi^{i+1}_{(M)} = \text{war}}$

$Q(R,P) = ۲ + \frac{۹}{۱۰}\times ۲۰ = ۲۰$

$Q(R,W) = \frac{۲}{۱۰}\left[۲+\frac{۹}{۱۰}\times ۲۰\right] + \frac{۸}{۱۰}\left[۲+\frac{۹}{۱۰}\times ۵/۵\right] = ۹/۵۶$

$\rightarrow \boxed{\pi^{i+1}_{(R)} = \text{Peace}}$

$$Q_{(D,P)} = -1 + \frac{9}{10} \times -10 = -10$$

$$Q_{(D,W)} = -1 + \frac{9}{10} \times 5.5 = 3.95$$

$\rightarrow$ $\pi^{i+1}_{(D)} = war$

بخش ج هیچ تفاوتی نمی‌کند R روی وارد شدن یا

خارج شدن state باشد.