Subject: Date: / /

سوال ①:

Contraction mapping → $\|U(V_1) - U(V_2)\| \leq q \|V_1 - V_2\|$

$\|U(V_1) - U(V_2)\| = \|R + \gamma P V_1 - (R + \gamma P V_2)\| = \|\gamma P (V_1 - V_2)\| =$

$= \gamma \|P(V_1 - V_2)\| \leq \gamma \|P\| \|V_1 - V_2\|$ ~~$q = \gamma \|P\|$~~

↳ $U(V)$ یک نگاشت انقباضی است.

$$U(V_1(s)) - U(V_2(s)) = \sum_{s'} P(s'|s,a)\{R(s) + \gamma V_1(s')\} - \sum_{s'} P(s'|s,a)\{R(s) + \gamma V_2(s')\}$$

$$= \sum_{s'} P(s'|s,a)\gamma(V_1(s') - V_2(s')) \leq \gamma \max_{s'}\{V_1(s') - V_2(s')\} \underbrace{\sum P(s'|s,a)}_{1}$$

$\max_{s'}\{V_1(s') - V_2(s')\} = \|V_1 - V_2\|_\infty$

↳ می دانیم که $U(V)$ یک Contraction Mapping است.

~~d~~ → $U(V^\pi) = V^\pi$ در نتیجه داریم که

$d(U(x), U(y)) \leq q\, d(x,y) \quad \forall x, y \in X$

↳ $d(U(V^k), U(V^\pi)) \leq q\, d(V^k$

$d(V^k, V^\pi) = d(U(V^{k-1}), U(V^\pi)) \leq q\, d(V^{k-1}, V^\pi)$

$= q\, d(U(V^{k-2}), U(V^\pi))$

$\leq q^2 d(V^{k-2}, V^\pi) = q^2 d(U(V^{k-3}), U(V^\pi))$

$\Rightarrow d(V^k, V^\pi) \leq q^k d(V^0, V^\pi)$

↳ $0 \leq d(V^k, V^\pi) \leq q^k d(V^0, V^\pi)$

$\lim k \to \infty$ ↳ 0

s.a.m

Subject: Date: / /

$$\lim_{k\to\infty} d(V^k, V^\pi) \le \lim_{k\to\infty} \gamma^k d(V^0, V^\pi) = 0$$

در نتیجه $\lim_{n\to\infty} U^n(V) = V^\pi$

$$\|U(V_1) - U(V_2)\|_\infty \le \gamma \|V_1 - V_2\|_\infty$$

به صورت کلی برای norm داریم

$$\|x - y\|_\infty \le \|x - z\|_\infty + \|z - y\|_\infty \quad (I)$$

حال داریم

$$\|U^{(k+1)}(V) - U^k(V)\|_\infty \le \gamma \|U^k(V) - U^{k-1}(V)\|_\infty \le \gamma\epsilon$$

$$\|U^{(k+2)}(V) - U^{k+1}(V)\|_\infty \le \gamma \|U^{k+1}(V) - U^k(V)\|_\infty \le \gamma^2\epsilon$$

$$\vdots$$

$$\|U^{k+n}(V) - U^{k+n-1}(V)\|_\infty \le \gamma^n \epsilon$$

$$\Rightarrow \sum_{i=0}^{n} \|U^{k+i}(V) - U^{k+i-1}(V)\|_\infty \le \sum_{i=1}^{n} \gamma^i \epsilon$$

همچنین طبق نکته ۱ داریم

$$\|V^\pi - U^k(V)\|_\infty \le \lim_{n\to\infty} \sum_{i=1}^{n} \|U^{k+i}(V) - U^{k+i-1}(V)\|_\infty \le \lim_{n\to\infty} \sum_{i=1}^{n} \gamma^i \epsilon$$

$$\Rightarrow \|V^\pi - U^k(V)\|_\infty \le \frac{\gamma\epsilon}{1-\gamma} \le \frac{\epsilon}{1-\gamma}$$

s.a.m

Subject: Date: / /

Every visit Monte-Carlo: سوال 1 و 2

Sample Episode:

Define $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \cdots + \gamma^{T_i-1} r_{i,T_i}$

For each step t until T_i:

state s is the state visited at step t

$N(s) = N(s) + 1$

$G(s) = G(s) + G_{i,t}$

$V^{\pi}(s) = G(s)/N(s)$

تنها در اپیزود 21 $R=-10$

و در اپیزود 23 $R=10$

در بقیه اپیزود ها هیچ پاداشی نداریم.

$V_{(1)} = -10$ $V_{(2)} = -10$ $V_{(3)} = -10$ $V_{(8)} = -10$

$V_{(7)} = -10$ $V_{(12)} = -10$ $V_{(16)} = -10$ $V_{(20)} = -10$

~~$V_{(21)} = -10/2 = -5$~~ $V_{(21)} = \frac{-10+0}{2} = -5$ $V_{(22)} = \frac{0+10}{2} = 5$

$V_{(17)} = 10$ $V_{(18)} = 10$ $V_{(23)} = 10$

و Value مربوط به بقیه اپیزود ها برابر با 0 باقی می ماند.

* برای هر دو قسمت سوال فرض گرفته شده است که $\gamma = 1$ باشد.

s.a.m

Subject: Date: / /

First-Visit Monte Carlo: (2

Sample Episode

$G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \dots + \gamma^{T_i - 1} r_{i,T_i}$

For each time step t until T_i

If this is the first viste of state s in this episode

$N(s) = N(s) + 1$

$G(s) = G(s) + G_{i,t}$

$V^{\pi}(s) = G(s) / N(s)$

$v(1) = -10$ $v(2) = -10$ $v(3) = -10$ $v(8) = -10$

$v(7) = -10$ $v(12) = -10$ $v(16) = -10$ $v(20) = -10$

$v(21) = -10$ $v(22) = 0$ $v(27) = 10$ $v(15) = 10$

$v(23) = 10$

و بقیه value ها همه برابر با 0 هستند.

s.a.m

سوال 3 -

الف)

$$v^{\pi}_{(s)} = \mathbb{E}^{\pi}\left\{\sum_{t=0}^{\infty} \gamma^t r_t \mid S_0 = s\right\} = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\{r_t \mid S_0 = s\},$$

$$= \sum_{a \in A} \pi_{(s,a)} \left\{R_{(s,a)} + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}\{r_t \mid S_0 = s\}\right\}$$

$$= \sum_{a \in A} \pi_{(s,a)} \left\{R_{(s,a)} + \sum_{s' \in S} P_{(s,a,s')} \sum_{a' \in A} \pi(s',a') \left\{\gamma R(s',a') + \sum_{t=2}^{\infty} \gamma^t \mathbb{E}\{r_t \mid S_0 = s\}\right\}\right\}$$

$$= \sum_{a \in A} \pi_{(s,a)} R_{(s,a)} + \sum_{a \in A} \pi_{(s,a)} \sum_{s' \in S} P_{(s,a,s')} \sum_{a' \in A} \gamma \pi(s',a') R_{(s',a')}$$

$$+ \sum_{a \in A} \pi_{(s,a)} \sum_{s' \in S} P_{(s,a,s')} \sum_{a' \in A} \pi(s',a') \sum_{s'' \in S} P_{(s',a',s'')} \sum_{a'' \in A} \gamma^2 \pi(s'',a'') R(s'',a'') + \cdots$$

با توجه به اینکه M و M_0 هر دو P و R یکسانی دارند، نتیجه می‌توان گفت که

$$v^{\pi}_{(s)} = v^{\pi}_{0}(s)$$

s.a.m

ب) این عبارت درست است.

$$v^{\pi}(s) = \mathbb{E}^{\pi}\left\{\sum_{t=0}^{\infty} \gamma^t r_t \mid S_0 = s\right\} \rightsquigarrow \pi^* = \underset{\pi}{\operatorname{argmax}}\left\{v^{\pi}_{(s)}\right\}$$

$$r_t \to c r_t \Rightarrow v'^{\pi}(s) = \mathbb{E}^{\pi}\left\{\sum_{t=0}^{\infty} \gamma^t c r_t \mid S_0 = s\right\} =$$

$$= c\,\mathbb{E}^{\pi}\left\{\sum_{t=0}^{\infty} \gamma^t r_t \mid S_0 = s\right\} = c v^{\pi}_{(s)}$$

$$\pi'^{*} = \underset{\pi}{\operatorname{argmax}}\left\{c v^{\pi}_{(s)}\right\} = \underset{\pi}{\operatorname{argmax}}\left\{v^{\pi}_{(s)}\right\} = \pi^*$$

ج) نادرست

$$v^{\pi}_{(s)} = \mathbb{E}^{\pi}\left\{\sum_{t=0}^{T} \gamma^t r_t \mid S_0 = s\right\} \rightsquigarrow \pi^* = \underset{\pi}{\operatorname{argmax}}\left\{v^{\pi}_{(s)}\right\}$$

$$r_t \to r_t + c \rightsquigarrow v'^{\pi}_{(s)} = \mathbb{E}\left\{\sum_{t=0}^{T} \gamma^t (r_t + c) \mid S_0 = s\right\} =$$

$$= \mathbb{E}\left\{\sum_{t=0}^{T} \gamma^t r_t \mid S_0 = s\right\} + \sum_{t=0}^{T} \gamma^t c$$

با توجه به عبارت بدست آمده مقدار $\sum_{t=0}^{T} \gamma^t c$ به $v^{\pi}_{(s)}$ اضافه می‌شود.

که با توجه به اینکه طول هر trajectory به اندازه‌ی T است و این مقادیر با هم

فرق داشته باعث می‌شود که argmax مقدار دیگری را به عنوان π

به ما دهد.

s.a.m

Subject: Date: / /

$$v^{\pi}_{(s)} = \mathbb{E}^{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t r_t \mid S_0 = s\right\} \rightsquigarrow \pi^* = \underset{\pi}{\operatorname{argmax}}\left\{v^{\pi}_{(s)}\right\}$$ (ب

$r_t \to r_t + c$ درست است.

$$v'^{\pi}_{(s)} = \mathbb{E}^{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t\{r_t + c\} \mid S_0 = s\right\} =$$

$$= \mathbb{E}^{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t r_t \mid S_0 = s\right\} + \sum_{t=0}^{\infty}\gamma^t c = v^{\pi}_{(s)} + \frac{c}{1-\gamma} \rightsquigarrow$$

$$\rightsquigarrow \pi'^* = \underset{\pi}{\operatorname{argmax}}\left\{v'^{\pi}\right\} = \underset{\pi}{\operatorname{argmax}}\left\{v^{\pi} + \frac{c}{1-\gamma}\right\} =$$

$$= \underset{\pi}{\operatorname{argmax}}\left\{v^{\pi}\right\} = \pi^* \Rightarrow \pi'^* = \pi^*$$

ج) این عبارت درست است.

در نظر بگیرید که $a = \underset{\pi}{\operatorname{argmax}}\left\{v^{\pi}_{(s)}\right\} \Rightarrow \pi^* = \underset{\pi}{\operatorname{argmax}}\left\{v^{\pi}\right\}$

$v^{\pi}_{(s)} = \mathbb{E}^{\pi}\left\{\sum_{t=0}^{\infty}\gamma^t r_t \mid S_0 = s\right\}$ باشد. حال

ابتدا در MDP بدیهی ~~Action~~

$$v^{\pi}_{(s)} = \sum_{a\in A}\pi(s,a)R(s,a) + \sum_{a\in A}\pi(s,a)\sum_{s'\in S}P(s,a,s')\sum_{a'\in A}\gamma\pi(s',a')R(s',a') +$$

$$+ \sum_{a\in A}\pi(s,a)\sum_{s'\in S}P(s,a,s')\sum_{a'\in A}\pi(s',a')\sum_{s''\in S}P(s',a',a'')\sum_{a''\in A}\gamma^2\pi(s'',a'')R(s'',a'') +$$

$$+ \dots$$

s.a.m

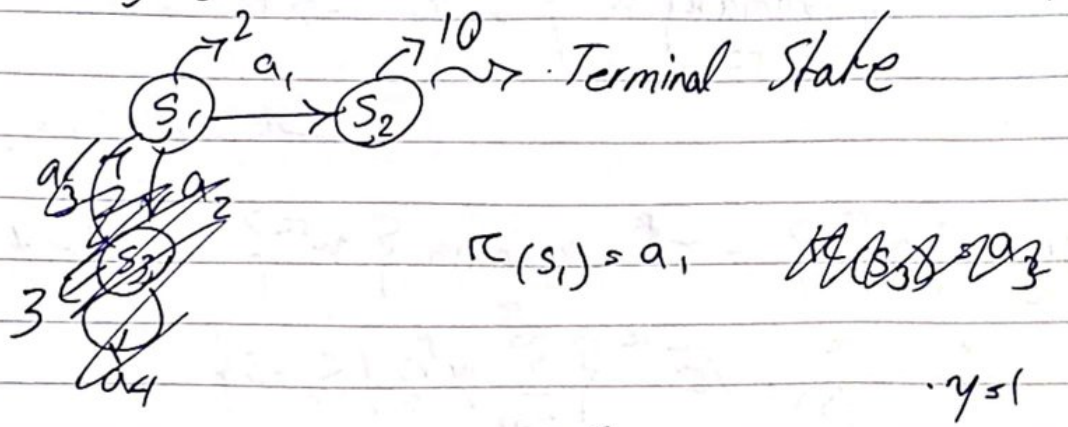
سوال ۶ـ

الف) نادرست است.

در این معادله V^{π} را به این صورت محاسبه می‌کنیم که از چه اکشن‌هایی می‌توان به اکشن 's' رسید و V^{π} را به این اساس محاسبه می‌کنیم.

معادله بلمن اصلی معادله‌ای است که در پایین داریم:

$$V^{\pi}(s) = \sum_{s'} \sum_{a} \pi(s,a) P(s',a,s) \{R(s,a) + \gamma V^{\pi}(s')\}$$



$\pi(S_1) = a_1$

$\gamma = 1$.

معادله بلمن عادی: $V^{\pi}(S_2) = 10 \quad V^{\pi}(S_1) = 12$

معادله داده شده: $V^{\pi}(S_2) = 2 \quad V^{\pi}(S_2) = -8$

که به این است اگر ریوارد‌ها همه مثبت باشند، مقدار Value نمی‌تواند منفی باشد.

همچنین به این مسئله می‌توان اشاره کرد که اگر γ نزدیک به ۰ باشد، این معادله بلمن، مقدار زیاد برای Value به دست نمی‌آورد.

s.a.m

Subject: Date: / /

$$v^{\pi}_{(s)} = \mathbb{E}\{G_t | S_t = s, \pi\} = \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,|\, S_t = s, \pi\right\} = \quad (ب$$

$$= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}\{R_{t+k} | S_t = s, \pi\} = \sum_{a \in A} \pi_{(s,a)} \left\{R_{(s,a)} + \sum_{k=1}^{\infty} \gamma^k \mathbb{E}\{R_{t+k} | S_t = s, \pi\}\right\}$$

$$= \sum_{a \in A} \pi_{(s,a)} \left\{R_{(s,a)} + \sum_{s' \in S} P_{(s,a,s')} \sum_{a'} \pi_{(s',a')} \{\gamma R_{(s',a')} + \sum_{k=2}^{\infty} \gamma^k \mathbb{E}\{R_{t+k} | S_t = s, \pi\}\}\right\}$$

$$= \sum_{a \in A} \pi_{(s,a)} R_{(s,a)} + \sum_{a} \pi_{(s,a)} \sum_{s'} P_{(s,a,s')} \sum_{a'} \gamma \pi_{(s',a')} R_{(s',a')}$$

$$+ \sum_{a} \pi_{(s,a)} \sum_{s'} P_{(s,a,s')} \sum_{a'} \pi_{(s',a')} \sum_{s''} P_{(s',a',s'')} \sum_{a''} \gamma^2 \pi_{(s'',a'')} R_{(s'',a'')} + \cdots$$

$$v^{\pi}_{(s)} = \mathbb{E}\{G_t | S_0 = s, \pi\} = \mathbb{E}\left\{\sum_{t=0}^{\infty} \gamma^t R_t \,|\, S_0 = s, \pi\right\}$$

$$= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\{R_t | S_0 = s, \pi\} = \sum_{a} \pi_{(s,a)} \left\{R_{(s,a)} + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}\{R_t | S_0 = s, \pi\}\right\}$$

$$= \sum_{a} \pi_{(s,a)} \left\{R_{(s,a)} + \sum_{s'} P_{(s,a,s')} \sum_{a'} \pi_{(s',a')} \{\gamma R_{(s',a')} + \sum_{t=2}^{\infty} \gamma^t \mathbb{E}\{R_t | S_0 = s, \pi\}\}\right\}$$

$$= \sum_{a} \pi_{(s,a)} R_{(s,a)} + \sum_{a} \pi_{(s,a)} \sum_{s'} P_{(s,a,s')} \sum_{a'} \gamma \pi_{(s',a')} R_{(s',a')} +$$

$$+ \sum_{a} \pi_{(s,a)} \sum_{s'} P_{(s,a,s')} \sum_{a'} \pi_{(s',a')} \sum_{s''} P_{(s',a',s'')} \sum_{a''} \gamma^2 \pi_{(s'',a'')} R_{(s'',a'')} + \cdots$$

$$\mathbb{E}\{G_t | S_t = s, \pi\} \neq \mathbb{E}\{G | S_0 = s, \pi\}$$ نیست

s.a.m

$$v^{\pi}(s) = \mathbb{E}\{G_t / S_0 = s, \pi\} = \mathbb{E}\left\{\sum_{t=0}^{\infty} R_t / S_0 = s, \pi\right\} \quad (2$$

$\hookrightarrow v^{\pi}(S_{L-1}) = \mathbb{E}\{R_{L-1}\}$

$$v^{\pi}(S_{L-2}) = \mathbb{E}\{R_{L-2} + R_{L-1}\} = \mathbb{E}\{R_{L-2}\} + \mathbb{E}\{R_{L-1}\} =$$
$$= \mathbb{E}\{R_{L-2}\} + v^{\pi}(S_{L-1})$$

$$\vdots$$

$$v^{\pi}(S_n) = \mathbb{E}\{R_n + R_{n+1} + \cdots + R_{L-1}\} = \mathbb{E}\{R_n\} + \mathbb{E}\{R_{n+1} + \cdots + R_{L-1}\}$$
$$= \mathbb{E}\{R_n\} + v^{\pi}\{S_{n+1}\}$$

با توجه به اینکه هر R_n یک عدد مثبت است، $\mathbb{E}\{R_n\}$ نیز مثبت است در نتیجه

$$v^{\pi}(S_n) \{ v^{\pi}(S_{n+1}) \Rightarrow$$

$$\Rightarrow v^{\pi}(S_0) \{ v^{\pi}(S_1) \{ \cdots \{ v^{\pi}(S_{L-1})$$

$S_0: C \quad S_1: D \quad S_2: E$

سوال 5.

الف)

n-Step Returns: $G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n}$

$V(S_t) \leftarrow V(S_t) + \alpha \left(G_t^{(n)} - V(S_t) \right) + \gamma^n V(S_{t+n})$

$n:1 \Rightarrow$ Temporal-Difference:

$G_0^{(1)} = 0 + 0,5 \rightarrow V(S_0) = V(S_0) + \alpha(0,5 - 0,5) = 0,5$

$G_1^{(1)} = 0 + 0,5 \rightarrow V(S_1) = V(S_1) + \alpha(0,5 - 0,5) = 0,5$

$G_2^{(1)} = 1 \rightarrow V(S_2) = 0,5 + \alpha(-0,5 + 1) = 0,55$

تنها مقدار Value مربوط به E آپدیت می‌شود.

$n=2 \rightsquigarrow G_0^{(2)} = 0 + 0 + 0,5 \rightarrow V(S_0) = 0,5$

$G_1^{(2)} = 0 + 0 + 1 + 0,5 \rightarrow V(S_1) = 0,55$

$G_2^{(2)} = 1 \rightsquigarrow V(S_2) = 0,55$

D و E هر دو آپدیت می‌شوند

برای $n \geq 3$ با توجه به اینکه کل sample ما تنها 3 اپیزود دارد، در نتیجه در هر 3 حالت هر 3 اپیزود کار، D و E آپدیت می‌شوند.

s.a.m

ب) مقدار α در واقع به bias-Variance Trade-off اشاره دارد. زمانی که α کم باشد نمونه‌های اولیه اهمیت بیشتری دارند و داده‌های جدید اهمیت کمتری دارند و زمانی که α زیاد باشد، نیز نمونه‌های جدید خیلی اهمیت دارند.

ج)

آ) با افزایش تعداد حالت‌ها نیاز است که تکرارها و اپیزودها را بیشتر کنیم که نمونه‌های بیشتری داشته باشیم. در غیر این صورت خطا افزایش پیدا خواهد کرد.

ج و ب) افزایش و تعداد اپیزودها و تعداد تکرارها باعث می‌شوند که نمونه‌های بیشتری داشته باشیم که همین امر منجر به کاهش ارور می‌شود.

د)

$$E_t(s) = \gamma\lambda E_{t-1}(s) + \mathbb{1}(S_t = s)$$

$$E_t(s) = 1 + \gamma\lambda E_{t-1}(s) = 1 + \gamma\lambda + (\gamma\lambda)^2 E_{t-2}(s) + \dots$$

$$\Rightarrow E_t(s) = \sum_{n=0}^{t-1} (\gamma\lambda)^n$$

$$\lim_{t\to\infty} E_t(s) = \lim_{t\to\infty} \sum_{n=0}^{t-1} (\gamma\lambda)^n = \frac{1}{1-\gamma\lambda} = \frac{1}{1-0,2} = \frac{1}{0,8} = 1,25$$

s.a.m