

سوال ①

(الف) مقدار مورد انتظار بینه یا $v^*(s)$:

$v^*(s)$ بیان می کند که اگر در حالت s باقیماند و از آن حالت بعد را به صورت بینه optimal گل کنند و در واقع سیاست بینه را ادامه دهند و action هایمان را به صورت بینه انتخاب کنند، آنگاه $v^*(s)$ مقدار ماتریم Utility مورد انتظار که از حالت s دریافت خواهیم کرد، می باشد.

در واقع $v^*(s)$ بیش ترین مقدار مورد انتظار جمع discounted reward است که از استیت s دریافت می کنند با این فرق که عامل ماتریم γ را در بینه $v^*(s)$ می بینیم.

(ب) سیاست بینه ایک و منتهی یا $\pi^*(s)$:

رواقع سیاست بینه تعیین می کند که اگر در حالت s باقیماند، کدام کد از action را مقدار مورد انتظار utility را بیشینه کند. در واقع $\pi^*(s)$ بهترین action را منصفاً می کند که از در حالت s انجام دهیم. مقدار $\pi^*(s)$ را بیشینه می کند و هدف نایاب MDP شف سیاست بینه را بهتر می کند. $\pi^*(s)$ = optimal action from state (s)

(ج) می از v^* متوالی π^* را پیدا و بخواهیم v^* را بدید.

رواقع متوالی نفت این دو باید تبر عمال هستند و اگر هر کد را کشته باشیم، MDP را می توانیم حل کنیم. اینباره از policy extraction اینباره بدل آن لازم است بیار آنکه از v^* پیمایی متوالی علایت π^* را بدید. ① می توانیم v^* را با π^* را باز طبق رابطه زیر $v^*(s) = \pi^*(s)$ داراییم.

$$\pi^*(s) = \arg \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma v^*(s')]$$

رواقع باید نفت که کجا ایک state می شوند که با این action که می باز از حالت s به آن باید

② می توانیم از v^* پیمایی π^* را بدید (که نفت تراز ① می باشد)، لازم است از policy evaluation

$$v_\infty^\pi(s) = 0$$

$$v_{k+1}^\pi(s) \leftarrow \sum_{s'} T(s, \pi(s), s') [R(s, \pi(s), s') + \gamma v_k^\pi(s')]$$

کر با جایگزین ای $\pi^*(s) = \pi^*(s)$ مترادف به تعبیر زیر کنم:

$$V_{k+1}^{\pi^*}(s) \leftarrow \sum_{s'} T(s, \pi^*(s), s') [R(s, \pi^*(s), s') + \gamma V_k^{\pi^*}(s')]$$

از آنچه که سیاست ما بهینه باشد پس π^* را ب دست خواهیم آورد

(سال ۲)

(الف) (اصح) - زیرا حداکثر خاصیت مادکوف، استقلال از گذشته و بائمه،
Policy چون در واقع اتفاق نماید که قبل از کلام حالت ایدیم و آن در حالت مشخص باشیم،
جوا از تضمینات گذشته، مشخص شود با عمل و action ای که آن را انتخاب کنیم
بکلام حالت جدید می‌باشد. در واقع بدل آنکه تضمین گیری کنیم، state فعل، اطلاعات کافی
را در اختیار ما قرار نماید و نیاز به تأثیرپذیری از گذشته نداریم.

(ب) (غلط) - زیرا طبق معادل روبرو، خواهیم
کرد آن $v^*(s')$ پر اشیاء شود که اگر به حالت که بدم و در واقع داشته، خواهیم
شافت. معلوم است که هر چه مقدار لا بد از تأثیر پذیر آنده بیشتر
می‌شود و بنابراین تحقیق ما در این ایندیشی آنده طی بعده و نه در پیاده‌نمای این
هم‌جنین را این MDP کاری بسته نماید و با state فعل تضمین شود.

(ج) (غلط) - از آنها برای مطابقت با فرول بین در حالت که بد عمل در action بپردازیم
 $\max_a v^*(s) = \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma v^*(s')]$

حذف می‌شود و خواهیم داشت:
 $v^*(s) = \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma v^*(s')]$
بنابراین مقدار v^* نه تنها به R وابسته نباشد بلکه ممکن است توتنه از action
با اتفاقات مختلف به کار نمی‌افکار بدم و لذا به ازای کار اتفاق v^* تأثیر خود را در آنها
نمی‌داشت.

عبارت فوق تنها در صورت «است ایست که با اتفاق $a = (s,a)$ » $T(s,a)$ ، حالت s' terminal است
 $v^*(s) = \max_a Q^*(s,a)$ ، $v^*(s) = \max_a Q^*(s,a)$ با اینکه Q^* نمی‌تواند بعده کم تهم باشد

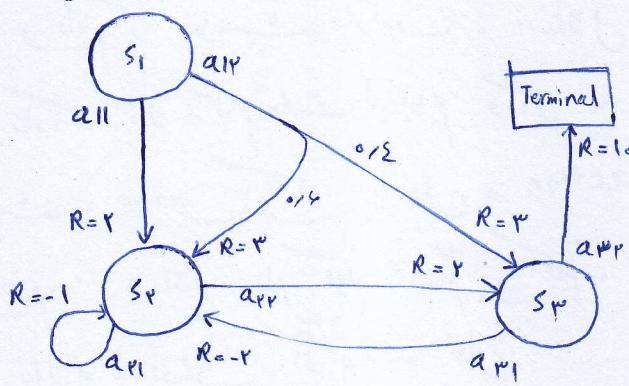
(د) (اصح) - چون در طبق معادل روبرو، $v^*(s) = \arg\max_a \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma v^*(s')]$
بله باید حالت را تبیین کرد و سپس به بعد Policy extraction که در فرول زیر آمده است
سیاست در برینه را یافت و می‌دانیم که یافتن سیاست بوسیله معادل حل MDP را می‌کند

Policy extraction: $\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma v^*(s')]$
و توجه باید کرد با این Q^* که ماقریم گفته شود v^* باید حالت بود آید.

(سؤال ۳)

(الف)

$$\gamma = 0.8$$



$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V_k(s')]$$

فرمول که دایم بصورت اوبرن لایع :

	s_1	s_r	s_3
V_p	v_1	v	10
V_r	$\delta/4$	v	10
V_1	γ	γ	10
V_0	0	0	0

پارچه بسؤال را پر ۳ام حل خواسته شده می توانیم
مقادیر را بصورت جدول درون آورده داشت
فرض ننمی :

حال بلکن $k=0$

$$① V_1(s_r) = \max_a \sum_{s'} T(s_r, a, s') [R(s_r, a, s') + \gamma V_0(s')]$$

- از حالت s_r به لکس باعث a_{rr} بصورت قطعی به حالت s_3 می روند و همچنان
و یا با a_{31} به طور قطعی به s_2 می روند . س :

$$\Rightarrow V_1(s_r) = \max \begin{cases} \text{if } (a=a_{rr}, s'=\text{terminal}) : 10 + 0 = 10 \\ \text{if } (a=a_{31}, s'=s_r) : -1 + 0 = -1 \end{cases} = 10$$

- از حالت s_r بصورت قطعی a_{rr} باشد $s_r \approx a_{rr}$ باشد s_3 با a_{rr} باشد قطعی

$$② V_1(s_r) = \max_a \sum_{s'} T(s_r, a, s') [R(s_r, a, s') + \gamma V_0(s')]$$

$$\Rightarrow V_1(s_r) = \max \begin{cases} \text{if } (a=a_{rr}, s=s_r) : \gamma + 0 = \gamma \\ \text{if } (a=a_{11}, s=s_r) : -1 + 0 = -1 \end{cases} = \gamma$$

- از حالت s_r بصورت قطعی a_{rr} باشد $s_r \approx a_{rr}$ باشد s_3 با a_{rr} باشد قطعی
حالت s_r با s_3 باشد s_r با a_{11} باشد s_3 با a_{11} باشد قطعی

$$③ V_1(s_1) = \max_a \sum_{s'} T(s_1, a, s') [R(s_1, a, s') + \gamma V_0(s')]$$

$$\Rightarrow V_1(s_1) = \max \begin{cases} \text{if } (a=a_{11}, s=s_r) : \gamma + 0 = \gamma \\ \text{if } (a=a_{12}, s=s_r) : 0.16(\gamma + 8V_0(s_r)) + 0.18(\gamma + 8V_0(s_3)) = 1.1\gamma + 1.1\gamma = \gamma \end{cases}$$

$\circ f^{1,1} \quad k=1$ حال بلو

$$\textcircled{1} \quad V_p(s_r) = \max \begin{cases} \text{if } (a=a_{rr}, s'= \text{terminal}) : R(s_r, a_{rr}, \text{terminal}) + \gamma(V_i(\text{terminal})) = 1_0 + 0 = 1_0 \\ \text{if } (a=a_{r1}, s' = s_r) : R(s_r, a_{r1}, s_r) + \gamma \times (V_i(s_r)) = -1 + 0 / \delta \times 1 \\ = -1 \\ = \max(1_0, -1) = 1_0 \end{cases}$$

$$\textcircled{2} \quad V_p(s_r) = \max \begin{cases} \text{if } (a=a_{rr}, s' = s_r) : R(s_r, a_{rr}, s_r) + \gamma(V_i(s_r)) = 1 + 0 / \delta \times 1 = V \\ \text{if } (a=a_{r1}, s' = s_r) : R(s_r, a_{r1}, s_r) + \gamma(V_i(s_r)) = -1 + 0 / \delta \times 1 = 0 \\ = \max(V, 0) = V \end{cases}$$

$$\textcircled{3} \quad V_p(s_1) = \max \begin{cases} \text{if } (a=a_{11}, s' = s_r) : R(s_1, a_{11}, s_r) + \gamma(V_i(s_r)) = 1 + 0 / \delta \times 1 = 1 \\ \text{if } (a=a_{1r}, s' = \begin{array}{c} \xrightarrow{\text{0/4}} s_r \\ \xrightarrow{\text{0/2}} s_r \end{array}) : T(s_1, a_{1r}, s_r) (R(s_1, a_{1r}, s_r) + \gamma V_i(s_r)) \\ + T(s_1, a_{1r}, s_r) (R(s_1, a_{1r}, s_r) + \gamma V_i(s_r)) \\ = 0/4 (1 + 0 / \delta \times 1) + 0/2 (1 + 0 / \delta \times 1) = 0/4 \\ = \max(1, 0/4) = 0/4 \end{cases}$$

$\circ f^{1,1} \quad k=1$ حال بلو

$$\textcircled{1} \quad V_p(s_r) = \max \begin{cases} \text{if } (a=a_{rr}, s'= \text{terminal}) : 1_0 + 0 / \delta \times 0 = 1_0 \\ \text{if } (a=a_{r1}, s' = s_r) : -1 + 0 / \delta \times V_p(s_r) = -1 + 0 / \delta \times V = 1 / \delta \end{cases} = 1_0$$

$$\textcircled{2} \quad V_p(s_r) = \max \begin{cases} \text{if } (a=a_{rr}, s' = s_r) : 1 + 0 / \delta \times V_p(s_r) = 1 + 0 / \delta \times 1 = V \\ \text{if } (a=a_{r1}, s' = s_r) : -1 + 0 / \delta \times V_p(s_r) = -1 + 0 / \delta \times V = 1 / \delta \end{cases} = V$$

$$\textcircled{3} \quad V_p(s_1) = \max \begin{cases} \text{if } (a=a_{11}, s' = s_r) : 1 + 0 / \delta \times V_p(s_r) = 1 + 0 / \delta \times V = 1 / \delta \\ \text{if } (a=a_{1r}, s' = \begin{array}{c} \xrightarrow{\text{0/4}} s_r \\ \xrightarrow{\text{0/2}} s_r \end{array}) : 0/4 (1 + 0 / \delta \times V) + 0/2 (1 + 0 / \delta \times V) = 1 / \delta \end{cases} = 1 / \delta$$

(تیسیت ب) بل - همانگونه در جدول زیر متنفس است، مقادیر مربوط به $V(S_1)$ و $V(S_2)$ دیگر تغییر نکرده و همگرا شده اند. بنابراین $V(S_1)$ هم قابل حمل کرد و پس از گام دیگر دوین انت

بنابراین $V(S_1) = V(S_2)$ را احتمل می کنیم.

	s_1	s_2	s_3	جدول
v_3	v_{11}	v	۱۰	
v_2	δ_{14}	v	۱۰	
v_1	۳	۲	۱۰	
v_0	۰	۰	۰	

حسب کار $k=3$ بازی $V(S_1)$ می باشد

$$V_3(S_1) = \max \left\{ \begin{array}{l} \text{if } (a=a_{11}, s'=s_2) : ۴ + ۰, \delta \times v = \delta_{14} \\ \text{if } (a=a_{12}, s' = \sum_{i=1}^{14} s_i) : ۰, \delta \times (۳ + ۰, \delta \times v) + ۰, \delta \times (۳ + ۰, \delta \times 1.) = v_{11} \end{array} \right\}$$

$$= \max(\delta_{14}, v_{11}) = \boxed{v_{11}}$$

	s_1	s_2	s_3	حال آنچه بازدیدکننده جدول را در نظر نمی نماید
v_4	v_{11}	v	۱۰	
v_3	v_{11}	v	۱۰	
v_2	δ_{14}	v	۱۰	
v_1	۳	۲	۱۰	
v_0	۰	۰	۰	

با نظر داشته باشند مقادیر v_3 و v_4 بنابراین کم شده و دیگر تغییر نکرده اند و در دفعه دیگر این انت

(value iteration)

قسمت ۸

	s_1	s_2	s_3
v_3	v_{11}	v	10
v_2	δ_{14}	v	10
v_1	3	2	10
v_0	0	0	0

جدول مارکوف هاست و در مرحله ۸

با توجه به اینکه مقدار s_3 مقادیر نهایی و آبیست ثده بود و همچنان اینتر مده بود از آن پاراملات
نتایج را نشان دهیم:

قول بحث زیر میگذرد:

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

حال پارامتراتی مختلف خواصی داشته باشد

$$s = s_4 \quad : \quad \begin{cases} \text{if } (a = a_{11}) \Rightarrow s' = s_1 \\ \text{if } (a = a_{14}) \Rightarrow s' = \text{terminal} \end{cases} \rightarrow \begin{array}{l} \text{عمل این ماقبل اند وهم این بایض} \\ \text{و نیازی نیست} \end{array}$$

$$\Rightarrow \pi(s_4) = \arg \max_a \begin{cases} \text{if } (a = a_{14}, s' = \text{terminal}) : 10 + 0 = 10 \\ \text{if } (a = a_{11}, s' = s_1) : -2 + 0 / \delta \times V = 1 / \delta \end{cases}$$

$$\Rightarrow \pi(s_4) = a_{11}$$

$$s = s_2 \quad : \quad \begin{cases} \text{if } (a = a_{11}) : s' = s_1 \\ \text{if } (a = a_{14}) : s' = s_4 \end{cases} \rightarrow \begin{array}{l} \text{عمل این ماقبل اند وهم این بایض} \\ \text{لیاز نیست} \end{array}$$

$$\Rightarrow \pi(s_2) = \arg \max_a \begin{cases} \text{if } (a = a_{11}, s' = s_1) : -1 + 0 / \delta \times V = 1 / \delta \\ \text{if } (a = a_{14}, s' = s_4) : 2 + 0 / \delta \times 10 = V \end{cases}$$

$$\Rightarrow \pi(s_2) = a_{11}$$

$$s = s_1 \quad : \quad \begin{cases} \text{if } (a = a_{11}) : s' = s_1 \\ \text{if } (a = a_{14}) : s' = \frac{0 / \delta}{1 / \delta} s_4 \end{cases} \rightarrow \begin{array}{l} \text{ابن عمل قطع است} \\ \text{ابن عمل غیرقطع است} \end{array}$$

$$\Rightarrow \pi(s_1) = \arg \max_a \begin{cases} \text{if } (a = a_{11}, s' = s_1) : 2 + 0 / \delta \times V = 2 / \delta \\ \text{if } (a = a_{14}, s' = \frac{0 / \delta}{1 / \delta} s_4) : 0 / \delta \times (2 + 0 / \delta \times V) + 1 / \delta (2 + 0 / \delta \times 1) = V / 1 \end{cases}$$

نهازی در state policy تعیین شد.

(قیمت)

فوجل بحثت زیر می باشد :

$$V_{k+1}^{\pi}(s) \leftarrow \sum_{s'} T(s, \pi(s), s') [R(s, \pi(s), s') + \gamma V_k^{\pi}(s')]$$

؛ \int_0^1 $k=0$ بار اول

① $\pi = a_{\text{up}}, s = s_u$

$$V_u^{\pi}(s_u) = 1_0 + \gamma V_0(\text{terminal}) = 1_0 + 0 = 1_0$$

② $\pi = a_{\text{up}}, s = s_r$

$$V_r^{\pi}(s_r) = r + \gamma \times V_0(s_r) = r + 0 = r$$

③ $\pi = a_{\text{up}}, s = s_l$

$$V_l^{\pi}(s_l) = 0.14 \times (r + 0) + 0.18(r + 0) = 0.32$$

؛ \int_0^1 $k=1$ بار

① $\pi = a_{\text{up}}, s = s_u$

$$V_u^{\pi}(s_u) = 1_0 + \gamma V_1(\text{terminal}) = 1_0 + 0 = 1_0$$

② $\pi = a_{\text{up}}, s = s_r$

$$V_r^{\pi}(s_r) = r + \gamma V_1(s_u) = r + 0.18 \times 1_0 = 0.18$$

③ $\pi = a_{\text{up}}, s = s_l$

$$V_l^{\pi}(s_l) = 0.14(r + 0.18 \times r) + 0.18(r + 0.18 \times 1_0) = 0.32$$

؛ \int_0^1 $k=2$ بار

① $\pi = a_{\text{up}}, s = s_u$

$$V_u^{\pi}(s_u) = 1_0 + \gamma V_2(\text{terminal}) = 1_0 + 0 = 1_0$$

② $\pi = a_{\text{up}}, s = s_r$

$$V_r^{\pi}(s_r) = r + \gamma V_2(s_u) = r + 0.18 \times 1_0 = 0.18$$

③ $\pi = a_{\text{up}}, s = s_l$

$$V_l^{\pi}(s_l) = 0.14(r + 0.18 \times r) + 0.18(r + 0.18 \times 1_0) = 0.32$$

	s_1	s_r	s_u
V_u	V_u	V	1_0
V_r	0.18	V	1_0
V_l	0.32	V	1_0
V_0	0	0	0

پس از اینکه مقدارهای شروع شده تابع V_k^{π} کامل شوند، این روش را کاملاً کامل نمایم.

(قیمت)

فرمول معرفت زیراست:

$$\pi_{i+1}(s) = \arg \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V^{\pi_i}(s')]$$

① $s = s_1$: $\pi_r(s_1) = \arg \max_a \sum_{s'} T(s_1, a, s') [R(s_1, a, s') + \gamma V^{\pi_i}(s')]$

$$= \arg \max_a \begin{cases} \text{if } (a=a_{11}, s'=s_r) : r + 0.1 \times v = 0.1 \\ \text{if } (a=a_{1r}, s' = \sum_{s_r} s_r) : 0.1(r + 0.1 \times v) + 0.1 \times (r + 0.1 \times 1.0) = v, 1 \end{cases}$$

$$\Rightarrow \boxed{\pi_r(s_1) = a_{1r}}$$

② $s = s_r$: $\pi_r(s_r) = \arg \max_a \sum_{s'} T(s_r, a, s') [R(s_r, a, s') + \gamma V^{\pi_i}(s')]$

$$= \arg \max_a \begin{cases} \text{if } (a=a_{rr}, s'=s_r) : -1 + 0.1 \times v = 0.1 \\ \text{if } (a=a_{rr}, s'=s_w) : r + 0.1 \times 1.0 = v \end{cases}$$

$$\Rightarrow \boxed{\pi_r(s_r) = a_{rr}}$$

③ $s = s_w$: $\pi_r(s_w) = \arg \max_a \sum_{s'} T(s_w, a, s') [R(s_w, a, s') + \gamma V^{\pi_i}(s')]$

$$= \arg \max_a \begin{cases} \text{if } (a=a_{ww}, s' = \text{terminal}) : 1.0 + 0 = 1.0 \\ \text{if } (a=a_{wr}, s' = s_r) : -r + 0.1 \times v = 0.1 \end{cases}$$

$$\Rightarrow \boxed{\pi_r(s_w) = a_{ww}}$$

جواب سوال پرسیده شده: خیر تغیری در سیاست نسبت به بخش ج متأثر نمی‌شود.

(سوال ۴)

الف) مرتکب مثال ای زیر ا نوشت :

۱) سیستم پیشنهاد دهنده اخبار پر کاربران \leftarrow عامل (agent) : محتوا بر زنگنه پیشنهاد دهنده اخبار

ترفیعه مثلاً به عنوان فونه news google را مرتکب نامید
کاربران \leftarrow محیط (environment) : جستجوی مختلف - اخبار
همین سیستم پیشنهاد دهنده محصولات شخصی باز شده بزرگ شده است

۲) ملین خودران \leftarrow عامل (agent) : ملین خودران

محیط (environment) : خیابان ها - موانع - عابرین پیاده - ملین های پر دید
چراغ راهنمای

ترفیعه ملین های در جایه جای بین لاین های مختلف خیابان مرتکب
از گفتگو Q-learning ...

۳) ربات - pick and place \leftarrow عامل (agent) : ربات

محیط (environment) : کارخانه - اثبات و لازم که باید پردازد و جایجا کند
محیط کارخانه

۴) ربات فرتابیست \leftarrow عامل (agent) : ربات فرتابیست
محیط (environment) : زمین باز تور - هیف - دوازه

(ب)

۱) سیستم پیشنهاد دهنده اخبار

در آنکه سیستم کار خودشند که حالات مختلف در طول زمان در محتوا آن را خود دهد، استفاده از مدل کار و مناسب نیست. بلکه مثال "اینها ممکن است علاقت فرد بعد از موقع تفسیر الله در آن میباشد" در قابل تمايل است. اخبار اتفاقی دیر علاقه نداشته باشد یا شرایط فرد بر دالیل تفسیر الله مبادر بلعد اخبار بگزید را دنبال کند. آن از قبل مدل داشته باشیم با توجه باین تفسیرات نیاز است مدل جویی را به بعد که هنوز برای بلکه و آخر مدل آپسیت نموده و این کار آنکه زمانی بود که باید مدل همچو این را بر حاکمیت یادگیری بود

مدل مرتکب این فونه را افزایش داد و همچنان سیستم را ببعود بگشید. میتوان شفعت یادگیری بدون مدل (model-free) کار آمد قدر بلکه

۱) مکانی خود ران :

به دلیل وسیع بودن پریده‌ها و همچنین وجود حالات مختلف که عامل نسبت به محیط نداد علاوه فقط به تعداد محدود سهل کرد هر چند زیاد هم باشد، نزدیک بسندگان و آن بنوای این مدل اراته دعیم باید آن مدل بسیار پیچیده باشد که شاید علاوه بر عکس باشد و از تلف حجم اطلاعات مورد نیاز هم سنتی و مستقیم مدل (model-tree) استفاده ننمی‌شود.

پس باید در مکانی خود ران از یادگیری بودن مدل (model-tree) استفاده ننمی‌شود.

باز مثال مکانی است معمولی تغییر لذت یا شرایط حیر مثل وجود سه، طرفه، عوض شکل فریض اصطلاح و راهنمایی ملکه و با این روش یادگیری بودن مدل دیگر نیافر با آنست که کل مدل نزدیک و بلکه یادگیری را با نمونه بیشتر از معمولی خوبی داشته باشد.

۲) ربات pick & place در کارخانه :

اگر این ربات بتواند در جایگاه که شرایط داشته باشد این تغییر لذت مدل خط تولید کند کارخانه، آنگاه شاید به دلیل محدود بودن هاست و امکان پذیر بودن مسافت مدل که همیشه درست کار کند، ایده به کارگیری model-based ایده برتری باشد.

اما مشابه حالات قبل در صورت میتواند دلیل کاری که تغییر پذیر نباشد، یادگیری پذیر مدل کار آمدترین باشد.

۳) ربات فوتbalیست :

در اینجا مکمل است ربات که پیدا نمی‌کند نزدیک جدید با اینکه شرایط مختلف متغیر شود و همچنین در محل بازی ممکن است موقعیت دشمن را بازتر تغییر کند باز مثال جایگاه قوه و بازیگران تغییر کند و یا سایر توابع نسبت به نیاز موضع شود و... «این» حالت به کارگیری یادگیری بودن مدل مناسب تر است «این» حالت نزدیک اتفاقات دقیقاً به جو صورت است و باخوض عده شرایط نمیدانیم دقیقاً اتفاق اتفاق نهان مگر منبع بخود رخ بازدزگل می‌شود چه قدر است و بقول آن دفعه نایم رس یادگیری پذیر مدل بترات

(نحوه ۲)

۱) سیتم پیشنهاد دهنده اخباره:

در اینجا باید به آنکه از ابتدا امر اطلاعات خاص از سلابق کارجو و صادر صور علاقه‌دار شایم، اینها باید explore کنیم تا اطلاعات کافی به دست آوریم و سپس رانش کسب شده را با exploit کنیم تا reward های را بشناسیم.

البته لازمات تا هر چند وقت بگذر بازهم explore انجام دهیم تا رانش سُبٰت به میتواند و علاقه‌دار و سه آژینت شود در واقع explore و exploit به طور هم‌زمان باشد بلکه دو آینه explore داریم.

۲) ملئین خود را:

در اینجا در آغاز کار، agent ما باید رانش کافی نسبت به محیط کسب کند وسیله exploit کرد. زمانکه رانش کافی راجح آوری کرد آنگاه من قوانه رانش خود را exploit کنیم تا reward را بست آوردم. در مطلب خود را

البته باید باین موضوع توجه داشت که به دلیل ماهیت متغیر محیط و reward از action باشد هر دفعه وسیله از کریست کلار explore کنیم تا رانش ما آژینت شود و تراینیم exploit را در آینده بشناسیم. در واقع در ابتدا explore کریم و در آینده exploit کریم.

ریاست - pick & place در کارخانه:

مشابه قبل «آغاز امر فقط explore کنیم و در نهایت رانش خود را exploit کنیم و بم»

رانش متناسب به محیط بیشتر شود دیگر نیازی به explore نخواهیم داشت زیرا تغییرات محیط کارخانه بار ریاست محدود و ناپذیر است و به مرور زمان explore بایانی باشد و در واقع ریاست میتواند رانش بسیار بیشتر شود تا exploit کنیم و از تبدیل شدن استفاده نماییم و از کوچه‌جایی بیشتر نیاز به explore کشیده باشیم.

۳) ریاست فوتbalیست:

«اینها معلم است ریاست درست زمین جیب یکی گرد وسیله ابتدا لازمات explore انجام دهد و بعد در نهایت رانش خود را exploit کنید reward گیرد. مثلاً این جا معلم است بار بازیگران تغییر شود و به نظر ثالثی و استراتژی بازی تغییر کند و مرقصیت توپ و بازیگران نیز به نور تغییر کند آن حالت تا حالا دیده نشده باشد، نیاز به آژینت بگزید وسیله اینها در ادامه exploit و explore به مرور هم‌زمان نیاز است.

مثال ۵
(الحل)

فرمول کل ره برای مطالعه تابع ارزش در هر حالت داریم به صورت زیر است:

$$V_{k+1}^{\pi}(s) \leftarrow \sum_{s'} T(s, \pi(s), s') [R(s, \pi(s), s') + \gamma V_k^{\pi}(s')]$$

پاترجه به ادلا بطور راهه متول نکت برانگشت از حالات، باید اینا داخل Episode باشند لذت و بینیم آن دادام بید از آن که هات مورد تطبیق باه بنوان حالت شروع samples میگذرد سپس با پید بینیم آن هر episode از آن

حالت بعد حی مقدار یادگار شونده ایم و مجموع آن را به دست آوریم. سپس بین مقادیر

ب درست آمده درجه episode مودتلق باید میگذیریم (که راه دیگر این بود انش اما + و - ۱۱ کسر) متوالی از اینات را باز نویسی کنیم (آنات را باز نویسی کنیم) که در آن جانب اندکه value خانه با انتخاب

Episode ۱: $(A_1, U, B_1, -0.1), (B_1, U, C_1, -0.1), (C_1, W, C_1, -0.1), (C_1, U, C_1, -0.1), (C_1, E, C_1, -0.1)$
 $(C_1, E, C_2, -0.1), (C_2, E, C_3, -0.1), (C_3, E, C_2, -0.1), (C_2, exit, X, +1)$

Episode 2: $(A_1, E, A_2, -0.1), (A_2, W, A_2, -0.1), (A_2, S, A_2, -0.1), (A_2, E, A_2, -0.1), (A_2, U, B_2, -0.1)$ $, (B_2, exit, X, +1)$

Episode 3: $(A_1, S, A_1, -0.1), (A_1, E, A_2, -0.1), (A_2, W, A_1, -0.1), (A_1, E, A_2, -0.1), (A_2, U, A_2, -0.1)$
 $(A_2, E, A_2, -0.1), (A_2, U, B_2, -0.1), (B_2, E, B_2, -0.1), (B_2, exit, X, +1)$

Episode 4: $(A_1, E, A_2, -0.1), (A_2, E, A_2, -0.1), (A_2, E, A_2, -0.1), (A_2, U, B_2, -0.1), (B_2, S, A_2, -0.1)$
 $(A_2, U, B_2, -0.1), (B_2, S, A_2, -0.1), (A_2, U, B_2, -0.1), (B_2, W, B_2, -0.1), (B_2, E, B_2, -0.1)$ $, (B_2, exit, X, +1)$

Episode 5: $(A_1, E, A_2, -0.1), (A_2, S, A_2, -0.1), (A_2, U, A_2, -0.1), (A_2, E, A_2, -0.1), (A_2, E, A_2, -0.1)$
 $(A_2, U, B_2, -0.1), (B_2, exit, X, +1)$

Episode 6: $(A_1, E, A_2, -0.1), (A_2, S, A_2, -0.1), (A_2, W, A_1, -0.1), (A_1, E, A_2, -0.1), (A_2, W, A_2, -0.1)$
 $(A_2, U, B_2, -0.1), (B_2, W, B_2, -0.1), (B_2, S, C_2, -0.1), ((C_2, S, C_2, -0.1), (C_2, E, C_2, -0.1))$ $, (C_2, exit, X, +1)$

$$V(A_1) = \frac{(-1/4) - (1/8) + (-1/\lambda - 1/V - 1/8) - (1/4) - (1/8) + (0.0 + 0.1/8)}{q} = -\frac{9/4}{q} = -1/0.4$$

$$V(A_2) = \frac{(-1/4 - 1/4) + (-1/8 - 1/4 - 1/4) + (-1/9 - 1/A) + (-1/8 - 1/4 - 1/4) + (0/1 + 0/1 + 0/1)}{q} = -1/0.94$$

$$V(A_w) = \frac{E_V(-1/2) + (-1/4) + (-1/V - 1/8 - 1/W) + (-1/P) + 0/8}{q} = -\frac{V/4}{q} = -1/0.18$$

$$V(A_\Sigma) = \frac{-1/1 - 1/1}{q} = -1/1$$

$$V(B_1) = \frac{0/4}{q} = 0/4$$

$$V(B_2) = \frac{(-1/1) + (-1/8 - 1/4 - 1/P - 1/1) + (0/4 + 0/V)}{q} = -\frac{8/4}{q} = -0/1.8$$

حریت از B_2 برآوردنی است.

$$V(B_2) = -1$$

$$V(C_1) = \frac{(0.1E + 0.1\Delta + 0.4 + 0.1N)}{E_1} = \frac{2.1}{2} = 1.05$$

$$V(C_r) = \frac{0.1A}{1} = 0.1A$$

$$V(C_w) = \frac{0.19 + (0.1A + 0.19)}{w} = \frac{0.19}{w}$$

$$V(C_E) = 1$$

$\alpha = 0.1$ روش با پیشگیری از انتقال transition را به این طبق سوال $\alpha = 0.1$ است. مقدار متاداده اولیه Q های نیز ۰ است.

نحوه بحث در نظر گرفته شد:

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

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

Episode 1:

$$\text{sample1} = R(A_1, U, B_1) + 1 \times \max_{a'} Q(B_1, a') = -0.1 + 1 \times 0 = -0.1$$

$$Q(A_1, U) = (1 - 0.1) \times Q(A_1, U) + 0.1 \times (-0.1) = -0.10$$

$$\text{sample2} = R(B_1, U, C_1) + 1 \times \max_{a'} Q(C_1, a') = -0.1 + 1 \times 0 = -0.1$$

$$Q(B_1, U) = (1 - 0.1) \times Q(B_1, U) + 0.1 \times (-0.1) = -0.10$$

$$\text{sample3} = R(C_1, W, C_1) + 1 \times \max_{a'} Q(C_1, a') = -0.1 + 1 \times 0 = -0.1$$

$$Q(C_1, W) = (1 - 0.1) \times Q(C_1, W) + 0.1 \times (-0.1) = -0.10$$

$$\text{sample4} = R(C_1, U, C_1) + 1 \times \max_{a'} Q(C_1, a') = -0.1 + 1 \times \max(0, -0.10) = -0.1$$

$$Q(C_1, U) = (1 - 0.1) \times Q(C_1, U) + 0.1 \times (-0.1) = -0.10$$

$$\text{sample5} = R(C_1, E, C_1) + 1 \times \max_{a'} Q(C_1, a') = -0.1 + 1 \times \max(0, -0.10, -0.10) = -0.1$$

$$Q(C_1, E) = (1 - 0.1) \times Q(C_1, E) + 0.1 \times (-0.1) = -0.10$$

$$\text{sample6} = R(C_1, E, C_r) + 1 \times \max_{a'} Q(C_r, a') = -0.1 + 1 \times 0 = -0.1$$

$$Q(C_1, E) = (1 - 0.1) \times Q(C_1, E) + 0.1 \times (-0.1) = -0.10$$

$$\text{samplev} = R(C_r, E, C_w) + 1 \times \max_{a'} Q(C_w, a') = -0.1 + 1 \times 0 = -0.1$$

$$Q(C_r, E) = (1 - 0.1) \times Q(C_r, E) + 0.1 \times (-0.1) = -0.10$$

$$\text{Sample 1} = R(C_r, E, C_E) + 1 \times \max_{a'} Q(C_E, a') = -0.1 + 0 = -0.1$$

$$Q(C_r, E) = (1 - 0.1\delta) \times Q(C_r, E) + 0.1\delta \times -0.1 = 0.1\delta$$

Episode 2 :

$$\text{Sample 1} = R(A_1, E, A_r) + 1 \times \max_{a'} Q(A_r, a') = -0.1 + 0 = -0.1$$

$$Q(A_1, E) = (1 - 0.1\delta) \times Q(A_1, E) + 0.1\delta \times (-0.1) = -0.1\delta$$

$$\text{Sample 2} = R(A_r, W, A_r) + 1 \times \max_{a'} Q(A_r, a') = -0.1 + 0 = -0.1$$

$$Q(A_r, W) = (1 - 0.1\delta) \times Q(A_r, W) + 0.1\delta \times (-0.1) = -0.1\delta$$

$$\text{Sample 3} = R(A_r, S, A_W) + 1 \times \max_{a'} Q(A_W, a') = -0.1 + 0 = -0.1$$

$$Q(A_r, S) = (1 - 0.1\delta) \times Q(A_r, S) + 0.1\delta \times (-0.1) = -0.1\delta$$

$$\text{Sample 4} = R(A_W, E, A_S) + 1 \times \max_{a'} Q(A_S, a') = -0.1 + 0 = -0.1$$

$$Q(A_W, E) = (1 - 0.1\delta) \times Q(A_W, E) + 0.1\delta \times (-0.1) = -0.1\delta$$

$$\text{sample 5} = R(A_S, U, B_E) + 1 \times \max_{a'} Q(B_E, a') = -1 + (-0.1) + 0 = -1.1$$

$$Q(A_S, U) = (1 - 0.1\delta) \times Q(A_S, U) + 0.1\delta \times (-1.1) = -0.1\delta$$

Episode 3 :

$$\text{sample 1} = R(A_1, S, A_1) + 1 \times \max_{a'} Q(A_1, a') = -0.1 + \max(0, -0.1\delta) = -0.1$$

$$Q(A_1, S) = (1 - 0.1\delta) \times Q(A_1, S) + 0.1\delta \times (-0.1) = -0.1\delta$$

$$\text{sample 2} = R(A_1, E, A_r) + 1 \times \max_{a'} Q(A_r, a') = -0.1 + \max(0, -0.1\delta) = -0.1$$

$$Q(A_1, E) = (1 - 0.1\delta) \times Q(A_1, E) + 0.1\delta \times (-0.1) = -0.1\delta - 0.1\delta = -0.1\delta$$

$$\text{sample 3} = R(A_r, W, A_1) + 1 \times \max_{a'} Q(A_1, a') = -0.1 + \max(0, -0.1\delta, -0.1\delta) = -0.1$$

$$Q(A_r, W) = (1 - 0.1\delta) \times Q(A_r, W) + 0.1\delta \times (-0.1) = -0.1\delta + (-0.1\delta) = -0.1\delta$$

$$\text{sample 4} = R(A_1, E, A_r) + 1 \times \max_{a'} Q(A_r, a') = -0.1 + \max(0, -0.1\delta, -0.1\delta) = -0.1$$

$$Q(A_1, E) = (1 - 0.1\delta) \times Q(A_1, E) + 0.1\delta \times (-0.1) = -0.1\delta$$

$$\text{sample 5} = R(A_r, U, A_r) + 1 \times \max_{a'} Q(A_r, a') = -0.1 + \max(0, -0.1\delta, -0.1\delta) = -0.1$$

$$Q(A_r, U) = (1 - 0.1\delta) \times Q(A_r, U) + 0.1\delta \times (-0.1) = -0.1\delta$$

$$\text{sample 4} = R(A_r, E, A_w) + 1 \times \max_{a'} Q(A_w, a') = -0.1 + \max(0, -0.1 \delta) = -0.1$$

$$Q(A_r, E) = (1 - 0.1 \delta) \times Q(A_r, E) + 0.1 \delta \times (-0.1) = [-0.1 \delta]$$

$$\text{sample 5} = R(A_r, U, B_w) + 1 \times \max_{a'} Q(B_w, a') = -0.1 + 0 = -0.1$$

$$Q(A_w, U) = (1 - 0.1 \delta) \times Q(A_w, U) + 0.1 \delta \times (-0.1) = [-0.1 \delta]$$

$$\text{sample 6} = R(B_r, E, B_s) + 1 \times \max_{a'} Q(B_s, a') = -1 - 0.1 + 0 = -1.1$$

$$Q(B_r, E) = (1 - 0.1 \delta) \times Q(B_r, E) + 0.1 \delta \times (-1.1) = [-0.1 \delta]$$

Episode 4 :

$$\text{sample 1} = R(A_1, E, A_r) + \max_{a'} Q(A_r, a') = -0.1 + -0.1 \delta = -0.1 \delta$$

$$Q(A_1, E) = (1 - 0.1 \delta) \times Q(A_1, E) + 0.1 \delta \times (-0.1 \delta) = [-0.1 \delta \text{ AND}]$$

$$\text{sample 2} = R(A_r, E, A_r) + \max_{a'} Q(A_r, a') = -0.1 + (-0.1 \delta) = -0.1 \delta$$

$$Q(A_r, E) = (1 - 0.1 \delta) \times Q(A_r, E) + 0.1 \delta \times (-0.1 \delta) = [-0.1]$$

$$\text{sample 3} = R(A_r, E, A_w) + \max_{a'} Q(A_w, a') = -0.1 + \max(0, -0.1 \delta) = -0.1$$

$$Q(A_w, E) = (1 - 0.1 \delta) \times Q(A_w, E) + 0.1 \delta \times (-0.1) = [-0.1]$$

$$\text{sample 4} = R(A_w, U, B_w) + \max_{a'} Q(B_w, a') = -0.1 + \max(0, -0.1 \delta) = -0.1$$

$$Q(A_w, U) = (1 - 0.1 \delta) \times Q(A_w, U) + 0.1 \delta \times (-0.1) = [-0.1 \delta]$$

$$\text{sample 5} = R(B_r, S, A_w) + \max_{a'} Q(A_w, a') = -0.1 + \max(0, -0.1 \delta) = -0.1$$

$$Q(B_r, S) = (1 - 0.1 \delta) \times Q(B_r, S) + 0.1 \delta \times (-0.1) = [-0.1 \delta]$$

$$\text{sample 6} = R(A_r, U, B_w) + \max_{a'} Q(B_w, a') = -0.1 + \max(0, -0.1 \delta) = -0.1$$

$$Q(A_w, U) = (1 - 0.1 \delta) \times Q(A_w, U) + 0.1 \delta \times (-0.1) = [-0.1 \delta \text{ AND}]$$

$$\text{sample 5} = R(B_r, S, A_r) + \max_{a'} Q(A_r, a') = -0.1 + \max(0, -0.1 \delta) = -0.1$$

$$Q(B_r, S) = (1 - 0.1 \delta) \times Q(B_r, S) + 0.1 \delta \times (-0.1) = [-0.1 \delta]$$

$$\text{sample 6} = R(A_w, U, B_w) + \max_{a'} Q(B_w, a') = -0.1 + \max(0, -0.1 \delta, -0.1 \delta) = -0.1$$

$$Q(A_w, U) = (1 - 0.1 \delta) \times Q(A_w, U) + 0.1 \delta \times (-0.1) = [-0.1 \delta \text{ AND}]$$

$$\text{Sample 9} = R(B_w, w, B_w) + \max_{a'} Q(B_w, a') = -0.1 + \max(0, -0.1 \times 0.1, -0.1 \times 0) = -0.1$$

$$Q(B_w, w) = (1 - 0.1) \underbrace{Q(B_w, w)}_{-0.1} + 0.1 \times (-0.1) = \boxed{-0.1 \times 0}$$

$$\text{Sample 10} = R(B_w, E, B_E) + \max_{a'} Q(B_E, a') = -0.1 - 1 = -1.1$$

$$Q(B_w, E) = (1 - 0.1) \underbrace{Q(B_w, E)}_{-0.1 \times 0.1} + 0.1 \times (-1.1) = \boxed{-0.1 \times 0.1}$$

Episode 5 :

$$\text{Sample 1} = R(A_1, E, A_r) + \max_{a'} Q(A_r, a') = -0.1 + \max(-0.1 \times 0.1, -0.1) = -0.1$$

$$Q(A_1, E) = (1 - 0.1) \underbrace{Q(A_1, E)}_{-0.1 \times 0.1} + 0.1 \times (-0.1) = \boxed{-0.1 \times 0.1}$$

$$\text{Sample 2} = R(A_r, S, A_r) + \max_{a'} Q(A_r, a') = -0.1 + \max(-0.1 \times 0.1, -0.1) = -0.1$$

$$Q(A_r, S) = (1 - 0.1) \underbrace{Q(A_r, S)}_{-0.1 \times 0.1} + 0.1 \times (-0.1) = \boxed{-0.1}$$

$$\text{Sample 3} = R(A_r, U, A_r) + \max_{a'} Q(A_r, a') = -0.1 + \max(-0.1 \times 0.1, -0.1) = -0.1$$

$$Q(A_r, U) = (1 - 0.1) \underbrace{Q(A_r, U)}_{-0.1 \times 0.1} + 0.1 \times (-0.1) = \boxed{-0.1}$$

$$\text{Sample 4} = R(A_r, E, A_r) + \max_{a'} Q(A_r, a') = -0.1 + \max(0, -0.1 \times 0.1) = -0.1$$

$$Q(A_r, E) = (1 - 0.1) \underbrace{Q(A_r, E)}_{-0.1 \times 0.1} + 0.1 \times (-0.1) = \boxed{-0.1}$$

$$\text{Sample 5} = R(A_r, E, A_\Sigma) + \max_{a'} Q(A_\Sigma, a') = -0.1 + \max(0, -0.1 \times 0.1) = -0.1$$

$$Q(A_r, E) = (1 - 0.1) \underbrace{Q(A_r, E)}_{-0.1 \times 0.1} + 0.1 \times (-0.1) = \boxed{-0.1}$$

$$\text{Sample 6} = R(A_\Sigma, U, B_\Sigma) + \max_{a'} Q(B_\Sigma, a') = -1 - 0.1 = -1.1$$

$$Q(A_\Sigma, U) = (1 - 0.1) \underbrace{Q(A_\Sigma, U)}_{-0.1 \times 0.1} + 0.1 \times (-1.1) = \boxed{-0.1 \times 0.1}$$

Episode 6 :

$$\text{Sample 1} = R(A_1, E, A_r) + \max_{a'} (A_r, a') = -0.1 + \max(-0.1 \times 0.1, -0.1, -0.1, -0.1) = -0.1$$

$$Q(A_1, E) = (1 - 0.1) \underbrace{Q(A_1, E)}_{-0.1 \times 0.1} + 0.1 \times (-0.1) = \boxed{-0.1 \times 0.1}$$

$$\text{Sample 2} = R(A_r, S, A_r) + \max_{a'} (A_r, a') = -0.1 + \max(-0.1 \times 0.1, -0.1, -0.1, -0.1) = -0.1$$

$$Q(A_r, S) = (1 - 0.1) \underbrace{Q(A_r, S)}_{-0.1 \times 0.1} + 0.1 \times (-0.1) = \boxed{-0.1 \times 0.1}$$

$$\text{sample 3} = R(A_r, w, A_1) + \max_{a'} Q(A_1, a') = -0.1 + 0 = -0.1$$

$$Q(A_r, w) = (1 - 0.1\delta) Q(A_r, w) + 0.1\delta(-0.1) = \boxed{-0.10\delta}$$

$$\text{sample 5} = R(A_1, E, A_r) + \max_{a'} Q(A_r, a') = -0.1 + \max(-0.10\delta, -0.1, -0.11\delta) = -0.11\delta$$

$$Q(A_1, E) = (1 - 0.1\delta) Q(A_1, E) + 0.1\delta(-0.11\delta) = \boxed{-0.110\delta9\delta^2}$$

$$\text{sample 8} = R(A_r, w, A_w) + \max_{a'} Q(A_w, a') = -0.1 + \max(0, -0.1\delta, -0.19\delta) = -0.1$$

$$Q(A_r, w) = (1 - 0.1\delta) Q(A_r, w) + 0.1\delta \times (-0.1) = \boxed{-0.10\delta}$$

$$\text{sample 4} = R(A_w, V, B_w) + \max_{a'} Q(B_w, a') = -0.1 + 0 = -0.1$$

$$Q(A_w, V) = (1 - 0.1\delta) Q(A_w, V) + 0.1\delta(-0.1) = \boxed{-0.1094\delta}$$

$$\text{sample 5} = R(B_w, w, B_w) + \max_{a'} Q(B_w, a') = -0.1 + 0 = -0.1$$

$$Q(B_w, w) = (1 - 0.1\delta) Q(B_w, w) + 0.1\delta \times (-0.1) = \boxed{-0.10\delta}$$

$$\text{sample 1} = R(B_w, S, C_w) + \max_{a'} Q(C_w, a') = -0.1 + \max(0, 0.18\delta) = 0.18\delta$$

$$Q(B_w, S) = 0.1\delta \times Q(B_w, S) + 0.1\delta \times 0.18\delta = \boxed{0.113\delta}$$

$$\text{sample 9} = R(C_w, S, C_w) + \max_{a'} Q(C_w, a') = -0.1 + \max(0, 0.18\delta) = 0.18\delta$$

$$Q(C_w, S) = 0.1\delta \times 0 + 0.1\delta \times 0.18\delta = \boxed{0.11\delta}$$

$$\text{sample 10} = R(C_w, E, C_E) + \max_{a'} Q(C_E, a') = -0.1 + 1 = 0.9$$

$$Q(C_w, E) = 0.1\delta \times 0.18\delta + 0.1\delta \times 0.9 = \boxed{0.14\delta}$$

توجه کنید که اگر بهالت تینیال بازش $\rightarrow C_E$ رفیم مقدار دیوارد $0.9 + 0.05 = 0.95$ متر شده
به لای \rightarrow transition

اگر بهhalt تینیال بازش $\rightarrow (B_E)$ رفیم مقدار دیوارد $0.9 - 0.05 = 0.85$ متر

راه دیوار این بود که استیت جداگانه مثل \rightarrow تعریف کنیم و برای این استیت تینیال به آن رازش $\rightarrow +1$ داریم.

همچنین مقادیر آیینه ای شده Q ببارزی Episode ۱ مقدار اولیه است.

این پریده است

(بعد از)
Episode 1 :

	۱	۲	۳	۴	
C	-0.100	0	0	0.100	0
B	-0.100	-0.100	0	0	0
A	0	0	0	0	0
	0	0	0	0	0

(بعد از)
Episode 2 :

	۱	۲	۳	۴	
C	-0.100	0	0	0.100	0
B	-0.100	-0.100	0	0	0
A	0	0	0	0	-0.100
	0	-0.100	-0.100	0	0

(بعد از)
Episode 3 :

	۱	۲	۳	۴	
C	-0.100	0	0	0.100	0
B	-0.100	-0.100	0	0	0
A	0	-0.100	-0.100	0	0
	0	-0.100	-0.100	0	0

	۱	۲	۳	۴	
C	-0.100	0	0	0.100	0
B	-0.100	-0.100	0	-0.100	-0.100
A	0	-0.100	-0.100	0	0
	0	-0.100	-0.100	0	0

$\rightarrow -0.1111V0$

(بعد از)
Episode 5 :

	۱	۲	۳	۴	
C	-0.100	0	0	0.100	0
B	-0.100	-0.100	0	-0.100	-0.100
A	0	-0.100	-0.100	0	-0.100
	0	-0.100	-0.100	0	-0.100

$\rightarrow -0.11111V0$

	۱	۲	۳	۴	
C	-0.100	0	0	0.100	0
B	-0.100	-0.100	0	-0.100	-0.100
A	0	-0.100	-0.100	0	-0.100
	0	-0.100	-0.100	0	-0.100

$\downarrow \downarrow \rightarrow -0.111111V0$

(بعد از)
Episode 6 :

(قسمت ۲)

پاتریج ب آخرین جدول براسط آمده در مفهوم قبل که بعداز Episode ۶ بلزه است
آورده از ۷ هر حالت (state) ، ازین Q بر این state مالکیم شرکت
نایابی داریم

$$V(A_1) = \max(0, -0.1\delta, -0.1\delta, -0.1\delta) = 0.$$

$$V(A_2) = \max(-0.1, -0.1, -0.1\delta, -0.1\delta) = -0.1\delta$$

$$V(A_3) = \max(0, -0.1\delta, 0, -0.1\delta) = 0$$

$$V(A_4) = \max(0, 0, 0, -0.1\delta) = 0$$

$$V(B_1) = \max(0, 0, 0, -0.1\delta) = 0$$

$$V(B_2) = \max(0) = 0$$

$$V(B_3) = \max(0, -0.1\delta, 0.1\delta, -0.1\delta) = 0.1\delta$$

$$V(B_4) = -1$$

$$V(C_1) = \max(-0.1\delta, -0.1\delta, -0.1\delta, 0) = 0$$

$$V(C_2) = \max(0, 0, 0, -0.1\delta) = 0$$

$$V(C_3) = \max(0, 0, 0.1\delta, 0.1\delta) = 0.1\delta$$

$$V(C_4) = 1$$

همانطور که از مقادیر بدست آمده مشخص است با وجود اینه مطالبات لملان و پیویمه از
انجام دادن، همیناً مقدار ۷ مربوط به هر state است و میتوان ریکارڈ و مقادیر
بدست آمده با حالت قبل مقایسه است و هنوز خیلی که دارد تا به
۷ هر واقعه همچرا سووند

(نواں ۶)

Q-learning را از نوع بدون سیاست و نامحدود ذیا policy آبیت شده آن با policy دنگ متفاوت می‌باشد. در واقع بار پاکش action متر آنده تغییر مزد و بعد از آنکه این سیاست در پیشانی ارزش را به state حیثی افراحت کند.

حال بار توضیح دهن تا off-policy learning، on-policy learning مثال توضیح دهم:

on-policy: در واقع تپیات بایس آخرين سیاست learn جمع آوری خود را

و سپس از تپیات جدید بار بجود policy استفاده کنیم. یعنی سیاست π^k با اینکه جمع آوری شده تعلق خود π^k به اذم شوند سیاست ما بعد از

جمع آوری نتایج تعلق خوش evaluate خواهد شد.

(در واقع مژده است اینکه سیاست بار انتخاب action استفاده شده، در پایانی نیز استفاده شده)

value Iteration - policy Iteration (مثال ۶):

off-policy: «این دئی در واقع دس اجله که سیاست π^k را جایز دسته شود و بار بروز رسانی سیاست، تپیات قبل و پرسنلیتی در واقع اینکه استفاده از تپیات مربوط به سیاست که متفاوت و قدیمی تر مادام بیش سیاست. آن را بجود داشتم لزوماً سیاست مورد استفاده «مان بینت action

متناهی به بیان لفت: (Behaviour policy \neq target policy used for action selection)

Q-learning (مثال):

(متاید و تفاهه Q-learning و value iteration)

در Q-learning عامل b agent را بهار R (پاکش) و T (امصال افق) به حالت s اطلاع نداریم. بنابراین بعد از افق s به کد حالت (state) جدید و دریافت پاکش، در بارهای پاکش مربوط به آن حالت آگاه شود و از آن تاریخی transition اینکه عالیت به حالت افقی t استخراج کند. بسیاری از عامل فقط کشت و لش کریل افق که حالت ایجاد کردند پاکش موجود است، آن را هم زمانی پاکش کردند که بآن حالت بود و پاکش شکرید. در value iteration، عامل b agent R و T بار اینجا s به state بین را فراخواهد داد و دادن امصال ایمن از حالت s به t باعث می‌شوند است.

عالی s از اینکه action خاص از حالت s ، آنکه «با مرحلت s »

مآموزد

نمودنی و expected discount Cost حین که دستیار مسیر فرآوریم Q-learning
بعبارتی قاری اینکه لغت: حینکه یک black box داشته باشیم که فقط اجزاء
شیوه را به ما بدهد و مقدار (دقیق) reward، اتفاقات را به من ندهد باید از
آنکه داده شده باشیم که یک دوشی model free Q-learning باشد.

از ویکی‌پدیا، زیر متنفاوِه می‌باشد (الف)

- ۱- تراکم جمعیت ۲- دلایل تغیر ۳- داکتر مردم ملین (دستیط) ۴- موقعیت ملین
نشیت به افراد و فاصله از آن ۵- قوانین راهنمای انتگری مثل تایبلوو... ۶- جدول کش داین

از افزایش و تغییرات بابت رئود عرضه شده قابل استفاده نمایند. این رئود معمولاً با وزن مناسب (منفی یا مثبت) باعث تقویت یا خنثی کردن از میزان از دست دادن و بگذارن باوزن مناسب (منفی یا مثبت) باعث تقویت یا خنثی کردن میشوند.

بنابریں « اتبیان چونکہ میو لاً عابر نیست سرت ۱۲۰ تائیر مثبت فی نہاد»
 شامل شدید ماس دفعہ را یهیار (شدّہ دفعہ غرض) یا حدائق مرغت مجاز وید خیابان بین با توجه
 به وزنی و سرگرد ارزش state را تعین فی نہاد

دیگر نیز همراه باشد که مراتین بین خلط و پلنت نا اسیب حداقل شود و با جداول
و مرانع برخوران کنیم. هر دوین طبق تماره ۵ هر چه بیشتر قوانین رعایت شوند
باشه و شود جمیع نکومی هی اندی state در صورت دعایت قوانین بالا ترجیح دارد.
و دیگر ۲ که مبلغ راهنماییات نیز ضرورت هر دو دنباید لازم نظر مثلاً عبور کن

(ب) بار اسکو "شہر فر state ہی دیکھ لے کتے شدہ دبلا را دارد و فقط وزن آنے

$$v(s) = f_1(s) w_1 + f_r(s) w_r + \dots + f_n(s) w_n$$

می تواند متناسب باشد

نیازی ب حالت به ذهن ماردید این است که در ۲ استیت «حال داشتگ» شرکتمند و وزیر عوامل اد۳ و ۵ و ۶ بیان هم ۲ استیت یکسان نباشد . اما آن فاصله تابعی «اینجا دینه شماره ۴ بیان state اول کرن از state دوم بلند کرده آن میان فاصله زیادی تا پذیرفته دارد ، «این» حالت وزیر ۴ بیان state اول بسیار منف خواهد شد چون این

مطلب و امن نیست و لی فرض این ویکی پار بار state دوم مثبت است و state دوم با وجود آنکه ویکی پار کیسان با state اول هر دو، اما از نظر بالاتر دو در نتیجه و فرض ویکی پار تصور اینکه نایب state می باشد.