۴۰۳۱۰۶۶۱

Subject:

Date: محمدمهدی مرادی

۱- الف) درست

تابع ارزش به دلیل پیوسته بودن، ممکن است خیلی طول بکشد تا همگرا شود.

همچنین از یک جایی به بعد اکشنی که در هر استیت بهینه است را خیلی قبل‌تر

از همگرایی تابع ارزش بدست می‌آوریم پس همگرایی به π^* زودتر رخ می‌دهد

ب) نادرست

Q-learning یک الگوریتم off-policy است پس می‌تواند مستقل از سیاست بهینه

و اینکه اکشن بهینه چی است، مقادیر بهینه اکشن و درنهایت سیاست بهینه را

پیدا کند

ج) درست

در هر iteration الگوریتم value iteration $O(S^2A)$ پیچیدگی زمانی داریم

در هر iteration الگوریتم policy iteration دو بخش policy evaluation/improvement

داریم که در بخش اول به دلیل عدم وجود max روی اکشن‌ها $O(S^2)$ را داریم

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

$O(S^2A)$ پیچیدگی زمانی داریم پس: $O(S^2) + O(S^2A) = O(S^2A)$

DAT

پس اگر $|S| \approx |A|$ هر دو الگوریتم پیچیدگی زمانی $O(S^3)$ دارند

Subject:

Date:

د) درست

$$E\left(\sum_{t=0}^{\infty} \gamma^t (R_t + c)\right) = V^{\pi}(s) + \sum_{t=0}^{\infty} \gamma^t c = V^{\pi}(s) + \frac{c}{1-\gamma}$$

پس به تمامی حالات مقدار مثبت و ثابت $\frac{c}{1-\gamma}$ اضافه می‌شود و ترتیب و سیاست بهینه تغییر نمی‌کند (اگر $\gamma = 1$ بود و agent می‌توانست تا ابد بماند گزاره غلط میشد) در محیط

ه) نادرست

هرچه γ به صفر نزدیکتر باشد، سیاست بهینه پاداش زود هنگام را می‌خواهد اما اگر γ به یک نزدیک شود سیاست بهینه آینده نگرتر می‌شود و پاداش های بزرگ دیرهنگام را نیز می‌خواهد

و) درست

وقتی که یک finite horizon MDP است سیاست ها کاملا به تعداد حرکت باقی مانده بستگی دارد. مثلا اگر یک حرکت باقی مانده باشد بهترین سیاست انتخاب کشن با بیشترین reward است اما اگر ۵۰ حرکت مانده باشد بهترین سیاست سرمایه‌گذاری برای reward های بهتر در آینده است. پس چون سیاست بهینه به زمان وابسته شده non-stationary است

DAT

Subject: Date:

الف) تفاوت این دو در یک عبارت است که $Q(S_t, A_t)$ را آپدیت می‌کند و در

off-policy ← $\gamma \max_a Q(S_{t+1}, a)$ و در on-policy ← $\gamma Q(S_{t+1}, A_{t+1})$

این به این معنا است که در off-policy مستقل از سیاستی که داریم از بین استفاده

می‌کنیم، Q ها را آپدیت می‌کنیم (Q-learning) و در on-policy با استفاده از

همان سیاستی که داریم در محیط استفاده می‌کنیم، Q ها را آپدیت می‌کنیم (SARSA)

ب) در model-based ها سعی داریم تا R و T هر استیت رو محاسبه کنیم

تا با الگوریتم های MDP مسئله را حل کنیم اما در model-free ها T و R را

مستقیم در نظر نمی‌گیریم و با تعامل فیزیکی با محیط و مشاهده نتایج ارزش ها را یاد

می‌گیریم

DAT

Subject: Date:

۲-الف) در مرحله اول $V_{(1,2)}$ و $V_{(2,1)}$ آپدیت نمی‌شوند چون بیشتر از یک قدم با ترمینال‌ها فاصله دارند

$$V_{(1,2)} \Rightarrow a=R \Rightarrow \left.\begin{array}{l} 0/9(-10+0)=-9 \\ 0/05(0+0)=0 \\ 0/05(0+0)=0 \end{array}\right\} \Rightarrow -9$$

$$a=D/U \Rightarrow \left.\begin{array}{l} 0/9(0+0)=0 \\ 0/05(0+0)=0 \\ 0/05(-10+0)=-0/5 \end{array}\right\} \Rightarrow -0/5$$

$$a=L \Rightarrow \left.\begin{array}{l} 0/9(0+0)=0 \\ 0/05(0+0)=0 \\ 0/05(0+0)=0 \end{array}\right\} \Rightarrow 0$$

پس در مرحله اول: $V_1(1,2)=0$ $\quad \max_a \Rightarrow a=L$

برای بقیه نیز به همین صورت است:

$$V_{(2,2)} \Rightarrow \left.\begin{array}{l} a=R \Rightarrow 9+0+0=9 \\ a=D/U \Rightarrow 0+0+0/5=0/5 \\ a=L \Rightarrow 0+0+0=0 \end{array}\right\} \Rightarrow V_1(2,2)=9$$

DAT

CamScanner

Subject: Date:

مرحله دوم: باز هم $V_{(1,1)}$ صفر می‌شود

$$\left.\begin{array}{l} V_{(2,1)} \Rightarrow a=R \Rightarrow 1.29+0+0=1.29 \\ a=D/U \Rightarrow 0+0+0.450=0.450 \\ a=L \Rightarrow 0+0+0=0 \end{array}\right\} \Rightarrow V_2(2,1)=1.29$$

$$\left.\begin{array}{l} V_{(1,2)} \Rightarrow a=R \Rightarrow -9+0+0.450=-8.550 \\ a=D \Rightarrow 1.29+0-0.5=0.79 \\ a=U \Rightarrow 0+0-0.5=-0.5 \\ a=L \Rightarrow 0+0+0.450=0.450 \end{array}\right\} \Rightarrow V_2(1,2)=0.79$$

$$\left.\begin{array}{l} V_{(2,2)} \Rightarrow a=R \Rightarrow 9+0.450+0=9.450 \\ a=L \Rightarrow 0+0+0=0 \\ a=U \Rightarrow 0+0+0.5=0.5 \\ a=D \Rightarrow 0+0.5+1.29=1.79 \end{array}\right\} \Rightarrow V_2(2,2)=9.450$$

ب

$$TDL \Rightarrow \hat{V}(s)=(1-\alpha)\hat{V}(s)+\alpha\left(R(s,\pi(s),s')+\gamma\hat{V}(s')\right)$$

مسیر اول

$$V_{(1,1)}=0+0.1(0+0.9\times 0)=0$$

$$V_{(1,2)}=0+0.1(-10+0.9\times 0)=-1$$

(ادامه صفحه بعد)

DAT

Subject: Date:

مسیر دوم:

$V(1,1) = 0 + 0.1(0 + 0.9(-1)) = -0.09$

$V(1,2) = -1 \times 0.9 + 0.1(0 + 0.9(0)) = -0.9$

$V(2,2) = 0 + 0.1(10 + 0.9(0)) = 1$

ج) چون اکشن ها ثابت هستند و همه به سمت راست هستند مقادیر Q ها با روش SARSA برابر با مقادیر V در حالت ب می شود

مسیر اول:

$Q((1,1), R) = 0$

$Q((1,2), R) = -1$

مسیر دوم:

$Q((1,1), R) = -0.09$

$Q((1,2), R) = -0.9$

$Q((2,2), R) = 1$

DAT

CamScanner

Subject: Date:

د) مسیر اول:

$Q((1,1),R)=0$

$Q((1,2),R)=-1$

مسیر دوم:

$Q((1,1),R) \Rightarrow Q((1,2),R)=-1$ اینجا به جای انتخاب

از $Q((1,2),L/U/D)=0$ استفاده می‌شود چون $-1<0$

$\Rightarrow Q((1,1),R)=0+0/1(0)=0$

$Q((1,2),R)=-0/9$

$Q((2,2),R)=1$

مقادیر Q به جز $Q((1,1),R)$ با قسمت ج یکسان شدند چون در تمامی حالت‌ها

یا max در اکشن راست بود یا در ترمینال بودیم که $Q=0$ اما در حالت $Q((1,1),R)$

داشتیم: $Q((1,2),R)=-1$ و $Q((1,2),L/U/D)=0$

DAT

Subject: Date:

۵) برای تمامی حالت های دو مسیر $\Leftarrow$ $V_t(s) = \frac{\gamma^0 R_{t+1} + \gamma R_{t+1} + \dots + \gamma^{n-1} R_{t+n-1}}{n}$

$$V(1,1) = \frac{\left(0 + 0.9(-10)\right) + \left(0 + 0.9(10) + (0.9)^2(10)\right)}{2} = -0.45$$

$$V(1,2) = \frac{\left(1 \times (-10)\right) + \left(0 + 0.9(10)\right)}{2} = -0.5$$

$$V(2,2) = \frac{1 \times 10}{1} = 10$$

حالت (۱و۲) چون در مسیرها مشاهده نشده تخمین ندارد

در DE ارزش هر حالت مستقل و فقط بر اساس پاداش ته هر مسیر محاسبه می شود و درکی از ارتباط بین حالت ها ندارد و مثلا از ارزش (۱و۲) برای آپدیت (۱و۱) استفاده نمی کند

اما در TD تخمین یک حالت را با استفاده از تخمین حالت های بعدی آپدیت می کند که این باعث می شود در TD sample efficiency بالاتری داشته باشیم

DAT

Subject: Date:

۳- گزاره غلط است

اگر π_3 به صورت قطعی بود داشتیم:

$$V^{\pi_3}(s) = \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V^{\pi_3}(s')\right)$$

اما چون یک توزیع احتمال روی π_3 داریم پس:

$$V^{\pi_3}(s) = \sum_{a} \left(\frac{1}{2}\pi_1(a|s) + \frac{1}{2}\pi_2(a|s)\right) \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V^{\pi_3}(s')\right)$$

از فرمول اولیه یک E گرفتیم و به فرمول بالا رسیدیم

$$\Rightarrow V^{\pi_3}(s) = \frac{1}{2}\sum_{a} \pi_1(a|s) \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V^{\pi_3}(s')\right)$$

$$+ \frac{1}{2}\sum_{a} \pi_2(a|s) \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V^{\pi_3}(s')\right)$$

برای اینکه فرمول بالا درست باشد، عبارت $\gamma V^{\pi_3}(s')$ در حقیقت نقش‌ها باید به ترتیب $\gamma V^{\pi_1}(s')$ و $\gamma V^{\pi_2}(s')$ باشند که اینطور نیست

مثال نقض:

states = $\{S_0, S_{end}\}$ $\quad V(S_{end}) = 0$

action(A) = با احتمال یک در S_0 می‌ماند و ۱۰ امتیاز می‌گیرد

action(B) = با احتمال یک به S_{end} می‌رود و ۵ امتیاز می‌گیرد

$\gamma = \frac{1}{2}$

DAT

Subject: Date:

π_1 همواره اکشن A را انتخاب می‌کند

$$V_1(S) = 10 + \frac{1}{2} V_1(S) \Rightarrow V_1(S) = 20$$

π_2 همواره اکشن B را انتخاب می‌کند

$$V_2(S) = 0 + \frac{1}{2} V(S_{end}) \Rightarrow V_2(S) = 0$$

سوال می‌گوید:

$$V_3 = \frac{1}{2} \times 20 + 0 = 10$$

حالا همان فرمول صفحه قبل را می‌نویسیم

$$V_3(S) = \frac{1}{2}\left(10 + \frac{1}{2} V_3(S)\right) + \frac{1}{2} \times 0 \Rightarrow V_3(S) = \frac{20}{3} \neq$$

پس گزاره سوال غلط است

Subject: Date:

۴ـ الف) V ها را با این فرمول آپدیت می کنیم

$$V_{k+1}^{\pi_1}(s) = \sum_{s'} T(s,\pi_1(s),s')\left(R(s,\pi_1(s),s') + \gamma V_k^{\pi_1}(s')\right)$$

در حالت A و B ، T=1 یعنی اگر اکشن را فیکس کنیم حتما به s' می رویم

در C اینطوری نیست. و در هر ۳ استیت یک توزیع احتمال برای π_1

دارد و چون توزیعش یکنواخت است عمل میانگین گیری انجام می دهیم

$$V_2^{\pi_1}(A) = \frac{1}{2}\times\left(4+\frac{1}{2}\times 2\right)+\frac{1}{2}\left(2+\frac{1}{2}\times 2\right) = \boxed{4}$$

توزیع احتمال π_1

$$V_2^{\pi_1}(B) = \frac{1}{2}\times\left(2+\frac{1}{2}\times 2\right)+\frac{1}{2}\left(0+\frac{1}{2}\times 2\right) = \boxed{2}$$

$$V_2^{\pi_1}(C) = \frac{1}{2}\left(\frac{3}{4}\times\left(8+\frac{1}{2}\times 2\right)+\frac{1}{4}\left(8+\frac{1}{2}\times 2\right)\right)+\frac{1}{2}\left(1+\frac{1}{2}\times 2\right)$$

این یک $T(C,\pi_1(C),s')$ است

A یا C

$$\Longrightarrow V_2^{\pi_1}(C) = \boxed{5.5}$$

DAT

Subject: Date:

ب) π_2 را با فرمول زیر محاسبه می کنیم

$$\pi_2(s) = \arg\max_a \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V^{\pi_1}(s')\right)$$

$\pi_2(A) =$ $x \Rightarrow ۴ + \frac{۱}{۲} \times ۲ = ۵$

$y \Rightarrow ۲ + \frac{۱}{۲} \times ۵,۵ = ۴,۷۵$

$۵ > ۴,۷۵$

$\Rightarrow \boxed{\pi_2(A) = x}$

$\pi_2(B) =$ $x \Rightarrow ۲ + \frac{۱}{۲} \times ۵,۵ = ۴,۷۵$

$y \Rightarrow ۰ + \frac{۱}{۲} \times ۴ = ۲$

$۴,۷۵ > ۲$

$\Rightarrow \boxed{\pi_2(B) = x}$

$\pi_2(C) \Rightarrow$ $x \Rightarrow \frac{۱}{۴}(۸ + \frac{۱}{۲} \times ۴) + \frac{۳}{۴}(۸ + \frac{۱}{۲} \times ۵,۵) = \frac{۴۲,۲۵}{۴}$

$y \Rightarrow ۱ + \frac{۱}{۲} \times ۲ = ۲$

$\frac{۴۲,۲۵}{۴} > ۲$

$\Rightarrow \boxed{\pi_2(C) = x}$

DAT

CamScanner

Subject:

Date:

ج)

$$\pi'(s) = \arg\max_a Q^{\pi}(s, a)$$

پس ارزش کنش $\pi'(s)$ بزرگتر مساوی ارزش کنش $\pi(s)$ $(V^{\pi}(s))$ است

$$\Rightarrow Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s)$$

Q را باز می‌کنیم

$$\Rightarrow V^{\pi}(s) \leq E_{\pi'}\left[R_{t+1} + \gamma V^{\pi}(S_{t+1}) \mid S_t = s\right]$$

$V^{\pi}(S_{t+1})$ را باز می‌کنیم

$$V^{\pi}(s) \leq E_{\pi'}\left[R_{t+1} + \gamma Q^{\pi}(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s\right]$$

$$\Rightarrow V^{\pi}(s) \leq E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V^{\pi}(S_{t+2}) \mid S_t = s\right]$$

اگر ۳ ها را تا آخر باز کنیم به فرمول $V^{\pi'}(s)$ می‌رسیم

$$\boxed{\Rightarrow V^{\pi}(s) \leq V^{\pi'}(s)}$$

DAT

Subject: Date:

بخش دوم:

$$\forall s \in S \quad V^{\pi}(s) = V^{\pi'}(s)$$

$$V^{\pi'}(s) = \max_a Q^{\pi}(s,a)$$

$$\Rightarrow V^{\pi'}(s) = \max_a \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V^{\pi}(s')\right)$$

این همان معادله بلمن است که در آن π' همان π^* یا سیاست بهینه است

DAT

CamScanner

Subject:

Date:

States $\Rightarrow S = \{(0,C), (2,C), (3,C), (4,C), (5,C), Done\}$ ۵- الف)

یا $n>5$ یا اکشن stop انتخاب شود

ب)

$R((n,C), stop, Done) = n$

$R((n,C), draw, (n',C)) = 0 \quad n' \leq 5$

$R((n,C), draw, Done) = 0$

ج)

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

$V_0(s') = 0 \quad \forall s' \in S$, $V_k(Done) = 0 \quad \forall k \in [0,\infty)$

مرحله اول:

$V_1((5,C)) = \begin{cases} a = stop \Rightarrow 5 \\ a = draw \Rightarrow \frac{1}{3}(0) + \frac{1}{3}(0) + \frac{1}{3}(0) = 0 \end{cases}$

$\Rightarrow V_1((5,C)) = 5 = Q_1((5,C), stop)$, $Q_1((5,C), draw) = 0$

$V_1((4,C)) = \begin{cases} a = stop \Rightarrow 4 \\ a = draw \Rightarrow 0 \end{cases}$

$\Rightarrow V_1((4,C)) = 4 = Q_1((4,C), stop)$, $Q_1((4,C), draw) = 0$

DAT

Subject: Date:

$V_1((۳,C)) =$ $a = stop \Rightarrow ۳$
$a = draw \Rightarrow ۰$

$\Rightarrow V_1((۳,C)) = ۳ = Q_1((۳,C), stop)$ و $Q_1((۳,C), draw) = ۰$

$V_1((۲,C)) =$ $a = stop \Rightarrow ۲$
$a = draw \Rightarrow ۰$

$\Rightarrow V_1((۲,C)) = ۲ = Q_1((۲,C), stop)$ و $Q_1((۲,C), draw) = ۰$

$V_1((۱,C)) =$ $a = stop \Rightarrow ۰$
$a = draw \Rightarrow ۰$

$V_1((۰,C)) = ۰ = Q_1((۰,C), draw) = Q_1((۰,C), draw)$

مرحله دوم: (استیت های $(۵,C)$ و $(۴,C)$ همواره ارزش ۵ و ۴ را خواهند داشت

و اکشن بهینه stop خواهد بود پس برای $(۳,C)$ و $(۲,C)$ و $(۰,C)$

ادامه می دهیم)

$V_2((۳,C)) =$ $a = stop \Rightarrow ۳$
$a = draw \Rightarrow \frac{1}{۳}(۰ + ۵) + \frac{1}{۳}(۰) + \frac{1}{۳}(۰) = \frac{۵}{۳}$

$\Rightarrow V_2((۳,C)) = ۳ = Q_2((۳,C), stop)$ و $Q_2((۳,C), draw) = \frac{۵}{۳}$

DAT

CamScanner

Subject:

Date:

$V_2((2,C)) = \quad a=stop \Rightarrow 2$

$a=draw \Rightarrow \frac{1}{3}(0+4)+\frac{1}{3}(0+5)+\frac{1}{3}(0)=3$

$\Rightarrow V_2((2,C)) = 3 = Q_2((2,C),draw)$ و $Q_2((2,C),stop)=2$

$V_2((0,C)) = \quad a=stop \Rightarrow 0$

$a=draw \Rightarrow \frac{1}{3}(0+2)+\frac{1}{3}(0+3)+\frac{1}{3}(0+4)=3$

$\Rightarrow V_2((0,C)) = 3 = Q_2((0,C),draw)$ و $Q_2((0,C),stop)=0$

مرحله سوم: (از (2,C) تا (5,C) تغییری نمی‌کنند):

$V_3((0,C)) = \quad a=stop \Rightarrow 0$

$a=draw \Rightarrow \frac{1}{3}(0+3)+\frac{1}{3}(0+3)+\frac{1}{3}(0+4)=\frac{10}{3}$

$V_3((0,C)) = \frac{10}{3} = Q_3((0,C),draw)$ و $Q_3((0,C),stop)=0$

مرحله چهارم: هیچکدام از استیت ها V و π نشان تغییر نمی‌کنند پس همگرا شد.

$V^*((5,C))=5$ و $V^*((4,C))=4$ و $V^*((3,C))=3$

$V^*((2,C))=3$ و $V^*((0,C)=\frac{10}{3}$ و $\pi^*((5,C))=stop$

$\pi^*((4,C))=stop$ و $\pi^*((3,C))=stop$ و $\pi^*((2,C))=draw$

DAT

و $\pi^*((0,C))=draw$

Subject:

Date:

د) چون فقط کارت ۲ در ماه پس از حالت های (۳و۱) و (۵و۱) همواره

Q طولیه خود را دارند (از (۱و۰) شروع می شود بازی)

$Q_1((4,1), stop) = \frac{1}{2}(0) + \frac{1}{2}(0+4) = 2$

$Q_2((4,1), stop) = \frac{1}{2}(2) + \frac{1}{2}(0+4) = 3$

$Q_3((4,1), stop) = \frac{1}{2}(3) + \frac{1}{2}(0+4) = 3.5$

چون تعداد زیادی سمپل داریم حدنهایی به ۴ می رسد

$Q_1((2,1), stop) = \frac{1}{2}(0) + \frac{1}{2}(0+2) = 1$

$Q_2((2,1), stop) = \frac{1}{2}(1) + \frac{1}{2}(0+2) = 1.5$

حدنهایی این نیز ۲ است

حدنهایی $Q((0,1), stop)$ هم بدیهتا صفر است

$Q_1((4,1), draw) = \frac{1}{2}(0) + \frac{1}{2}(0+0) = 0$

همواره صفر خواهد بود

DAT

Subject:

Date:

$Q_1((۲,C), draw) = \frac{1}{2}(0) + \frac{1}{2}(0+4) = 2$

$Q_2((۲,C), draw) = \frac{1}{2}(2) + \frac{1}{2}(0+4) = 3$

حد نهایی این همه ۴ است

$Q_1((۱,C), draw) = \frac{1}{2}(0) + \frac{1}{2}(0+4) = 2$

$Q_2((۱,C), draw) = \frac{1}{2}(2) + \frac{1}{2}(0+4) = 3$

$Q_3((۱,C), draw) = \frac{1}{2}(3) + \frac{1}{2}(0+4) = 3.5$

حد نهایی این همه ۴ است

$(۱,C) - stop \Rightarrow 0$ $\quad$ $draw \Rightarrow 4$

$(۲,C) - stop \Rightarrow 2$ $\quad$ $draw \Rightarrow 4$

$(۳,C) = stop \Rightarrow 0$ $\quad$ $draw \Rightarrow 0$

$(۴,C) = stop \Rightarrow 4$ $\quad$ $draw \Rightarrow 0$

$(۵,C) = stop \Rightarrow 0$ $\quad$ $draw \Rightarrow 0$

DAT

Subject: Date:

ه)

$$\varepsilon\text{-greedy} \Rightarrow a_t = \begin{cases} \text{random action} & \text{به احتمال } \varepsilon \\ \arg\max_a Q(s_t, a) & 1-\varepsilon \text{ '' '' } \end{cases}$$

با احتمال ۱-۰/۴=۰/۶ اکشن stop را انتخاب می‌کنیم چون

$$Q((۴,C), stop) = ۴ > ۳ = Q((۴,C), draw)$$

و با احتمال ۰/۴ به صورت رندوم ⇐

$$stop \Rightarrow \frac{۰/۴}{۲} = ۰/۲ \qquad draw \Rightarrow \frac{۰/۴}{۲} = ۰/۲$$

پس در کل با احتمال $\boxed{\text{۰/۸ stop و ۰/۲ draw}}$

طبق بخش ج داریم:

$$Q((۴,C), stop) = ۴ \qquad Q((۴,C), draw) = ۰$$

$$\Rightarrow E[Q] = ۰/۸ \times ۴ + ۰/۲ \times ۰ = \boxed{۳/۲}$$

DAT

Subject:

Date:

و) با توجه به مقادیر بخش ج داریم:

$$regret_1 = V^*((0,c)) - Q^*((0,c), draw) = \frac{10}{3} - \frac{10}{3} = \boxed{0}$$

$$regret_2 = V^*((2,c)) - Q^*((2,c), stop) = 3 - 2 = \boxed{1}$$

$$regret_3 = V^*((0,c)) - Q^*((0,c), draw) = \frac{10}{3} - \frac{10}{3} = \boxed{0}$$

$$regret_4 = V^*((4,c)) - Q^*((4,c), draw) = 4 - 0 = \boxed{4}$$

$$\Rightarrow \text{مجموع } regret = \sum_{t=1}^{4} regret_t = 0 + 1 + 0 + 4 = \boxed{5}$$

DAT

Subject:

Date:

$Q^*((۵,C), stop) = ۵$ $\quad$ $Q^*((۵,C), draw) = ۰$

$Q^*((۴,C), stop) = ۴$ $\quad$ $Q^*((۴,C), draw) = ۰$

$Q^*((۳,C), stop) = ۳$ $\quad$ $Q^*((۳,C), draw) = \frac{۵}{۳}$

$Q^*((۲,C), stop) = ۲$ $\quad$ $Q^*((۲,C), draw) = ۳$

$Q^*((۰,C), stop) = ۰$ $\quad$ $Q^*((۰,C), draw) = \frac{۱۰}{۳}$

DAT

CamScanner

Subject: Date:

$T(X, a, Y) = \frac{۲}{۳}$ $T(X, a, Z) = \frac{۱}{۳}$ غ- الف)

$T(X, b, X) = 1$ برای سایر حالات $T=0$

$T(Y, a, Z) = \frac{۲}{۳}$ $T(Y, a, Y) = \frac{۱}{۳}$

$T(Y, b, X) = 1$

$T(Z, a, Z) = 1$ $T(Z, b, X) = 1$

ب)

$R(X, a, Y) = R(X, a, Z) = ۳$

برای سایر حالات $R=0$

$R(X, b, X) = 1$

$R(Y, a, Z) = R(Y, a, Y) = ۴$

$R(Y, b, X) = ۲$

$R(Z, a, Z) = 0$ $R(Z, b, X) = ۵$

DAT

Subject: Date:

ج)

$$V_{k+1}(s) = \max_a \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V_k(s')\right)$$

$V_1(X) = a \Rightarrow \frac{2}{3}(3+0) + \frac{1}{3}(3+0) = 3$

$b \Rightarrow 1(1+0) = 1$

$V_1(X) = 3 \quad \pi(X) = a \quad Q_1(X,a) = 3 \quad Q_1(X,b) = 1$

$V_1(Y) = a \Rightarrow \frac{2}{3}(4+0) + \frac{1}{3}(4+0) = 4$

$b \Rightarrow 1(2+0) = 2$

$V_1(Y) = 4 \quad \pi(Y) = a \quad Q_1(Y,a) = 4 \quad Q_1(Y,b) = 2$

$V_1(Z) = a \Rightarrow 1(0+0) = 0$

$b \Rightarrow 1(5+0) = 5$

$V_1(Z) = 5 \quad \pi(Z) = b \quad Q_1(Z,a) = 0 \quad Q_1(Z,b) = 5$

$\Rightarrow V_1(X) = 3,\ V_1(Y) = 4,\ V_1(Z) = 5$

$\pi(X) = a \quad \pi(Y) = a \quad \pi(Z) = b$

DAT

CamScanner

Subject: Date:

د) به همگرایی تضمین شده است چون که هر دفعه عدد ۰/۹ در $V_k(z)$ ضرب

میشه که در مرحله اول صفره بعد ۵ بعد $0/9 \times 5$ بعد $(0/9)^2 \times 5$ بعد $(0/9)^3 \times 5$

همینطور تا $(0/9)^n \times 5$ که برای n های بزرگ $V(z)$ به صفر میل می کند

$$V_{n+1}(z) = 1 \times (0 + (0/9)^n \times 5) \longrightarrow 0$$

DAT