



لطفا نکات زیر را رعایت کنید:

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- پاسخ ارسالی واضح و خوانا باشد.
- در هر کدام از سوالات، اگر از منابع خاصی استفاده کرده‌اید باید آن را ذکر کنید.
- اگر با افرادی همفکری کرده‌اید، نام ایشان را ذکر کنید.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد. به اسکرین‌شات از منابع یا پاسخ افراد دیگر نمره‌ای تعلق نمی‌گیرد.
- تمام پاسخ‌های خود را در یک فایل با فرمت `RL_HW#[SID]_[Fullname].zip` روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. علاوه بر آن، در هر تمرین می‌توانید تا سقف پنج روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید.

سوال ۱: (نظری) (۲۵ نمره)

فرض کنید (X, d) یک فضای متریک باشد آنگاه نگاشت f از X به X که به صورت $f: X \rightarrow X$ نشان داده می‌شود، یک نگاشت انقباضی است اگر مقدار $q \in [0, 1)$ وجود داشته باشد طوری که

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

حال می‌خواهیم با استفاده از نگاشت انقباضی همگرایی الگوریتم policy iteration را اثبات کنیم. همانطور که در کلاس دیدیم الگوریتم از دو مرحله بهبود سیاست و ارزیابی سیاست تشکیل شده است. در مرحله ارزیابی سیاست مقادیر ارزش‌ها برای هر سیاست را با استفاده از رابطه زیر محاسبه کرده و این رابطه را به صورت بازگشتی تکرار می‌کنیم تا زمانی مقادیر ارزش‌ها به V^π همگرا شوند.

$$V^\pi(S) = R(S, \pi(s)) + \gamma \sum_{S'} P(S'|S, \pi(s)) V^\pi(S') \quad \forall s$$

حال می‌خواهیم رابطه بالا را به فرم ماتریسی بنویسیم. فرض کنید:

$$\mathbf{V} = \mathbf{R} + \gamma \mathbf{P} \mathbf{V}$$

- \mathbf{R} : یک بردار $1 \times |S|$ از مقادیر پاداش هر وضعیت بر اساس سیاست π است.
- \mathbf{V} : یک بردار $1 \times |S|$ از مقادیر ارزش هر وضعیت بر اساس سیاست π است.
- \mathbf{P} : یک بردار $|S| \times |S|$ از مقادیر احتمال‌های انتقال بر اساس سیاست π است.

حال فرض کنید تابع $U(V)$ را به صورت زیر تعریف می‌کنیم:

$$\mathbf{U}(V) = \mathbf{R} + \gamma \mathbf{P} \mathbf{V}$$

الف) ثابت کنید $U(V)$ یک نگاشت انقباضی می باشد.

ب) حال پس از اثبات قسمت الف، نشان دهید که در مرحله ارزیابی سیاست به V^π همگرا می شویم. به عبارت دیگر، ثابت کنید رابطه زیر برقرار است:

$$\lim_{n \rightarrow \infty} U^n(V) = V^\pi$$

دقت شود منظور از $U^n(V)$ اعمال تابع U در n مرحله روی V می باشد، به طور مثال:

$$U^2(V) = U(U(V))$$

پ) حال با توجه به اینکه نمی توانیم در مرحله ارزیابی سیاست برای محاسبه V^π به صورت نامحدود حلقه را تکرار کنیم، فرض کنید حلقه را k مرحله تکرار کرده و پس از k مرحله رابطه زیر برقرار است:

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

آنگاه ثابت کنید رابطه زیر برقرار است:

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

پاسخ:

الف) برای آن که نشان دهیم U یک نگاشت انقباضی می باشد کافی است ثابت کنیم رابطه زیر برقرار است:

$$\|U(V') - U(V)\|_\infty < \gamma \cdot \|V' - V\|_\infty$$

$$\|U(V') - U(V)\|_\infty = \|(R + \gamma \cdot P \cdot V' - R - \gamma \cdot P \cdot V)\|_\infty = \|\gamma P(V' - V)\|_\infty$$

حال از نامساوی زیر استفاده می کنیم:

$$|AB| < |A| \cdot |B|$$

$$\|\gamma P(V' - V)\|_\infty < \gamma \cdot \|P\|_\infty \cdot \|V' - V\|_\infty$$

حال از آنجا که $\max_{s'} \sum_s P(s', s) = 1$:

$$\gamma \cdot \|P\|_\infty \cdot \|V' - V\|_\infty = \gamma \cdot \|V' - V\|_\infty$$

در نتیجه بر اساس تعریف ارائه شده برای نگاشت انقباضی، تابع U یک نگاشت انقباضی می باشد.

ب) اولاً براساس تعریف تابع U می دانیم رابطه زیر برقرار است:

$$V^\pi = U^\infty(0)$$

ولی نکته ای که وجود دارد، در مرحله ارزیابی سیاست، مقدار $U^\infty(V)$ به ازای هر مقدار اولیه V محاسبه می شود. حال در قسمت الف ثابت کردیم که $U(V)$ یک نگاشت انقباضی می باشد، لذا رابطه زیر برقرار است:

$$\|U(V') - U(V)\|_\infty < \gamma \cdot \|V' - V\|_\infty$$

حال هنگامی که n به سمت بی نهایت میل می کند، می توان رابطه زیر را نوشت:

$$\lim_{n \rightarrow \infty} \|U^n(V) - U^n(0)\|_\infty \rightarrow 0$$

حال بر اساس نتایج بالا می توان نوشت

$$V^\pi = U^\infty(V)$$

پ) برای پاسخ به این بخش از قسمت ب استفاده می کنیم:
اولاً در قسمت ب ثابت کردیم $U^\infty(V) = V^\pi$ ، در نتیجه رابطه بالا را می توان به صورت زیر نوشت:

$$\|V^\pi - U^k(V)\|_\infty = \|U^\infty(V) - U^k(V)\|_\infty$$

حال رابطه را به صورت زیر ساده می کنیم

$$\left\| \sum_{t=1}^{\infty} \left(U^{(t+k)}(V) - U^{(t+k-1)}(V) \right) \right\|_\infty$$

همچنین می دانیم به صورت کلی رابطه زیر برقرار است

$$|A + B|_\infty < |A|_\infty + |B|_\infty$$

در نتیجه می توان نوشت

$$\left\| \sum_{t=1}^{\infty} \left(U^{(t+k)}(V) - U^{(t+k-1)}(V) \right) \right\|_\infty < \sum_{t=1}^{\infty} \|U^{(t+k)}(V) - U^{(t+k-1)}(V)\|_\infty$$

که این رابطه را به صورت زیر می توانیم ساده کنیم

$$\sum_{t=1}^{\infty} \gamma^{(t)} \varepsilon = \frac{\varepsilon}{1 - \gamma}$$

در نتیجه عبارت نوشته شده در قسمت پ ثابت شد.

سوال ۲: (نظری) (۱۵ نمره)

یک gridworld به صورت زیر را در نظر بگیرید:

کنش های ممکن به صورت حرکت به سمت بالا، حرکت به سمت پایین، حرکت به سمت چپ و حرکت به سمت راست می باشد. همچنین فرض کنید هنگامی که عامل به یک سمت حرکت می کند:


(آ) با احتمال 0.8 در جهتی که می خواهد حرکت می کند

(ب) با احتمال 0.05، ۹۰ درجه به سمت راست منحرف می شود

(ج) با احتمال 0.05، ۹۰ درجه به سمت چپ منحرف می شود

(د) با احتمال 0.1 نمی تواند حرکتی کند و در جای خود می ماند

همچنین در صورتی که عامل با دیوار برخورد کند در جای خود می ماند. پاداش تمامی وضعیت هایی که ذکر نشده است را معادل صفر در نظر بگیرید. تمامی پاداش های مشخص برای هر وضعیت برای ورود به آن وضعیت می باشد.

Start $s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
$s = 6$	$s = 7$	$s = 8$	$s = 9$	$s = 10$
$s = 11$	$s = 12$		$s = 13$	$s = 14$
$s = 15$	$s = 16$		$s = 17$	$s = 18$
$s = 19$	$s = 20$	$R = -10$ $s = 21$ 	$s = 22$	$R = +10$ $s = 23$ Goal

شکل ۱: Gridworld

۱) فرض کنید عامل در اپیزود اول از اجرا خود به ترتیب وضعیت های ۱، ۲، ۳، ۸، ۷، ۱۲، ۱۶، ۲۰، ۲۱، ۲۲، ۲۱، ۲۲، ۱۷، ۱۸، ۲۳ را به ترتیب از چپ به راست ملاقات می کند. اگر عامل از every visit Mont Carlo برای تخمین مقادیر ارزش وضعیت ها استفاده نماید، مقدار ارزش هریک از وضعیت ها پس از اپیزود اول چه خواهد بود؟ (می توانید فرض کنید مقادیر اولیه ارزش وضعیت ها صفر می باشد)

۲) در صورتی که از الگوریتم اولین بازدید مونت کارلو استفاده شود، مقادیر ارزش ها در حالت قبلی به چه صورت خواهد بود؟

پاسخ:

الف) ارزش وضعیت ها به صورت زیر می شود

-۱۰	-۱۰	-۱۰	۰	۰
۰	-۱۰	-۱۰	۰	۰
۰	-۱۰	N/A	۰	۰
۰	-۱۰	N/A	۱۰	۱۰
۰	-۱۰	۵	۵	۰

جدول ۱: carlo mont visit every

ب) در این حالت ارزش وضعیت ها به صورت زیر می شود

-۱۰	-۱۰	-۱۰	۰	۰
۰	-۱۰	-۱۰	۰	۰
۰	-۱۰	N/A	۰	۰
۰	-۱۰	N/A	۱۰	۱۰
۰	-۱۰	۰	۰	۰

جدول ۲: carlo mont visit first

صورت کلی در یک MDP v^π به صورت زیر تعریف می‌شود:

$$v^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t | S_0 = s \right] \quad (۱)$$

در این سوال می‌خواهیم بررسی کنید در صورت ایجاد چه تغییراتی در یک MDP ممکن است سیاست بهینه تغییر کند

الف) فرض کنید M, M_0 دو MDP کاملاً مشابه می‌باشند که صرفاً در توزیع اولیه وضعیت‌ها با یکدیگر متفاوت می‌باشند. ثابت کنید v^π در هر دو MDP یکسان می‌باشد.

ب) در یک MDP محدود با پاداش‌های دارای باند مشخص اگر همه پاداش‌ها را در عدد مثبت ضرب کنیم سیاست بهینه تغییر نمی‌کند. درستی یا نادرستی جمله بالا را بررسی کنید و در صورت درست بودن جمله گزاره اثبات شده و در غیر این صورت مثال نقض آورده شود.

ج) به صورت کلی ثابت کنید در یک MDP با حالات محدود و پاداش با باند مشخص در صورتی همه پاداش‌ها با عدد ثابت c جمع شود سیاست بهینه ممکن است تغییر کند

د) حال گزاره قسمت ج در حالتی بررسی کنید که Terminating state نداشته باشیم. به نظر شما در این این گزاره درست است یا خیر. در صورت درست بودن جمله گزاره را اثبات و در غیر این صورت مثال نقض آورده شود.

ه) یک MDP محدود با پاداش‌های دارای باند مشخص را در نظر بگیرید و فرض کنید این MDP یک سیاست بهینه قطعی دارد. حال از روی این MDP یک MDP جدید می‌سازیم به این صورت که اگر کنش a در یک وضعیت s بهینه نباشد، $r(s, a)$ را از مقدار ثابت c کم می‌کنیم (منظور از $r(s, a)$ پاداش گرفته در وضعیت s به ازای کنش a می‌باشد) و در صورتی که a کنش بهینه باشد، مقدار پاداش آن تغییری نمی‌کند. حال درستی یا نادرستی ادعای زیر را بررسی کنید:

”سیاست بهینه در MDP جدید با سیاست بهینه در MDP اولیه برابر است.“

پاسخ:

الف) فرض کنید تابع انتقال و تابع پاداش در MDP اول را برابر P, R در نظر گرفته و فرض کنید تابع انتقال و تابع پاداش در MDP دوم را برابر P', R' در نظر بگیریم

آنگاه می‌توان گفت رابطه زیر برقرار است

$$\forall s \in S, \forall a \in A, \forall s' \in S \quad P(s, a, s') = P'(s, a, s') \quad \text{and} \quad R(s, a) = R'(s, a)$$

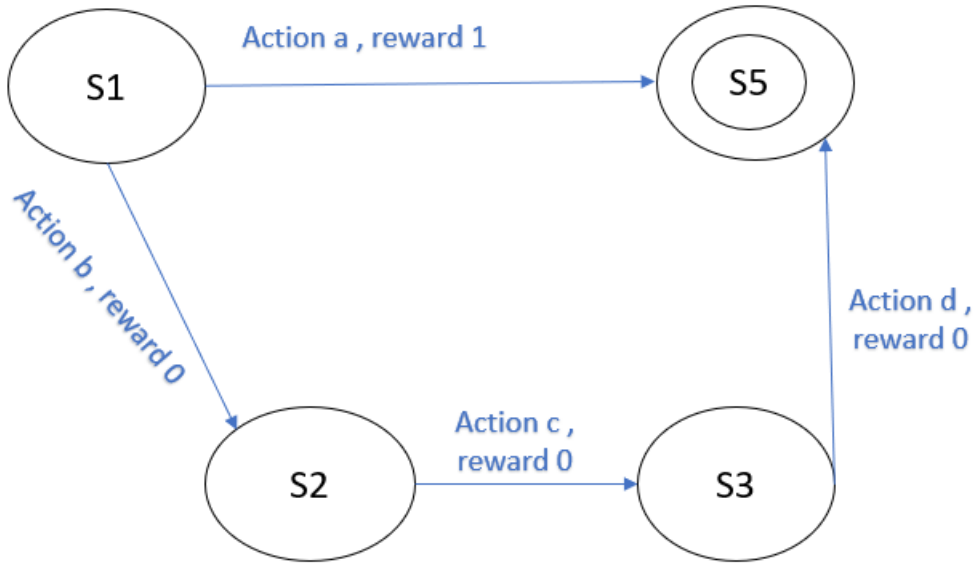
حال می‌توانیم بر اساس روابط زیر ثابت کرد که ارزش وضعیت‌ها در هر دو فرآیند تصمیم‌گیری مارکوف با یکدیگر برابرند

$$\begin{aligned}
v_M^\pi(s) &= \mathbb{E}[G_t | S_t = s, \pi] \\
&= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}[R_{t+k} | S_t = s, \pi] \\
&= \sum_{a \in A} \pi(s, a) \left[R(s, a) + \sum_{s' \in S} \gamma P(s, a, s') \sum_{a' \in A} \pi(s', a') \left[R(s', a') + \sum_{k=2}^{\infty} \gamma^{k-1} \mathbb{E}[R_{t+k} | S_t = s, \pi] \right] \right] \\
&= \sum_{a \in A} \pi(s, a) \left[R(s, a) + \sum_{s' \in S} \gamma P(s, a, s') \sum_{a' \in A} \pi(s', a') R(s', a') \right] \\
&\quad + \gamma^2 \sum_{a \in A} \pi(s, a) \left[P(s, a, s') \sum_{a' \in A} \pi(s', a') P(s', a', s'') \sum_{a'' \in A} \pi(s'', a'') R(s'', a'') \right] + \dots \\
&= \gamma^0 \sum_{a \in A} \pi(s, a) R(s, a) + \gamma^1 \sum_{a \in A} \pi(s, a) \sum_{s' \in S} P(s, a, s') \sum_{a' \in A} \pi(s', a') R(s', a') \\
&\quad + \gamma^2 \sum_{a \in A} \pi(s, a) \sum_{s' \in S} P(s, a, s') \sum_{a' \in A} \pi(s', a') P(s', a', s'') \sum_{a'' \in A} \pi(s'', a'') R(s'', a'') + \dots \\
&\text{(a) } v_M^\pi(s).
\end{aligned}$$

ب) عبارت گفته شده صحیح است. یک فرآیند تصمیم مارکوف با پاداش‌های دارای باند مانند M را در نظر بگیرید. اگر π^* یک سیاست بهینه در این زنجیره باشد، می‌خواهیم نشان دهیم که این سیاست بهینه، در تمامی فرآیند تصمیم مارکوف مشابه M مانند M' که در تمامی خواص به غیر از پاداش‌ها با M مشترک می‌باشند و $R'_t = \alpha R_t$ باشد، تغییر نمی‌کند. فرض کنید $J_\alpha(\pi) = E[\sum_{t=0}^{\infty} \gamma^t \cdot \alpha \cdot R_t]$ تابع هدف فرآیند تصمیم مارکوف M' باشد. اگر فرض کنید J تابع هدف فرآیند تصمیم مارکوف اصلی یعنی M باشد، واضح است که رابطه زیر به صورت کلی برقرار است:

$$\begin{aligned}
J(\pi^*) &\geq J(\pi) \\
E[\sum_{t=0}^{\infty} \gamma^t R_t | \pi^*] &\geq E[\sum_{t=0}^{\infty} \gamma^t R_t | \pi] \\
\alpha \cdot E[\sum_{t=0}^{\infty} \gamma^t R_t | \pi^*] &\geq \alpha \cdot E[\sum_{t=0}^{\infty} \gamma^t R_t | \pi] \\
E[\sum_{t=0}^{\infty} \alpha \cdot \gamma^t R_t | \pi^*] &\geq E[\sum_{t=0}^{\infty} \alpha \cdot \gamma^t R_t | \pi] \\
J_a(\pi^*) &> J_a(\pi)
\end{aligned}$$

ج) این عبارت در حالت کلی نادرست است زیرا برای آن مثال نقض وجود دارد. شکل زیر را در نظر بگیرید: در این شکل، اقدام بهینه در $S1$ عمل a می‌باشد. حال با اضافه کردن پاداش‌ها به مقدار $+4$ ، اکشن بهینه در $S1$ تغییر کرده و اکشن b در حالت بهینه انتخاب می‌شود.



(د) عبارت گفته شده در این حالت صحیح است. برای اثبات، فرض کنید c $R'(s, a) = R(s, a) + c$.
اثبات:

$$\begin{aligned}
 V'^{k+1}(s) &= \max_a \sum_{s'} P(s'|s, a) [R'(s', a, s) + \gamma V'^k(s')] \\
 &= \max_a \sum_{s'} P(s'|s, a) [R(s', a, s) + \gamma \frac{1 - \gamma^k}{1 - \gamma} r_0 + V^k(s)] \\
 &= \max_a \sum_{s'} P(s'|s, a) [R(s', a, s) + \frac{1 - \gamma^{k+1}}{1 - \gamma} r_0 + \gamma V^k(s)] \\
 &= \frac{1 - \gamma^{k+1}}{1 - \gamma} r_0 + \max_a \sum_{s'} P(s'|s, a) [R(s', a, s) + \gamma V^k(s)]
 \end{aligned}$$

همانطور که مشخص است، عبارت بدست آمده نسبت به حالت اصلی تغییری نکرده و در این حالت نیز کنشی انتخاب می شود که در MDP قبلی انتخاب می شد.

(ه) ادعای گفته شده صحیح می باشد. اولاً، به صورت کلی می توان گفت برای همه $s \in S$ و $a \in A$ داریم: $R(s, a) > R'(s, a)$ که منظور از R' مقدار پاداش در MDP جدید می باشد. زیرا بر اساس تعریف مسئله، در صورتی که کنش در سیاست بهینه باشد $R(s, a) = R'(s, a)$ و در غیر این صورت $R'(s, a) = R(s, a) - c$. حال روابط را به صورت زیر مینویسیم:

$$\begin{aligned}
 v_M^{\pi^*}(s) &= E \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s, \pi^*, M \right] \\
 &= E \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s, \pi^*, M' \right] = v_{M'}^{\pi^*}(s) \quad s \in S \text{ همه برای}
 \end{aligned}$$

به صورت کلی برای تمام سیاست ها نیز می توان رابطه زیر را نوشت:

$$E \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s, \pi^*, M \right] \geq E \left[\sum_{k=0}^{\infty} \gamma^k R'_{t+k} | S_t = s, \pi^*, M' \right]$$

در نتیجه می توان گفت برای هر سیاستی:

$$v_M^{\pi^*}(s) > v_{M'}^{\pi^*}(s)$$

$$v_M^{\pi^*}(s) = v_{M'}^{\pi^*}(s)$$

در نتیجه ثابت کردیم که در M' نیز سیاست π^* سیاست بهینه می باشد.

سوال ۴: (نظری) (۳۰ نمره)

در این سوال می خواهیم رابطه را بلمن را کمی دقیق تر بررسی کنیم به صورت کلی تعاریف گفته شده را به صورت زیر در نظر بگیرید

$$G = \sum_{t=0}^{\infty} \gamma^{(t)} R_t$$

$$G_t = \sum_{k=0}^{\infty} \gamma^{(k)} R_{t+k}$$

الف) فرض کنید شخصی پیشنهاد می کند معادله بلمن را معکوس کند و مقدار یک حالت را بر اساس مقادیر قبلی بنویسد. این شخص به رابطه زیر می رسد. به نظر شما این رابطه درست است؟ در صورت درستی اثبات آن و در صورت نادرستی مثال نقض آن را بیان کنید رابطه به صورت زیر بیان می شود

$$v^{\pi}(s') = \sum_s \sum_a P(s, a, s') \pi(s, a) \left[\frac{v^{\pi}(s) - R(s, a)}{\gamma} \right]$$

ب) نشان دهید دو عبارت زیر با یکدیگر معادل می باشند:

$$v^{(\pi)}(s) = E[G_t | S_t = s, \pi]$$

$$v^{(\pi)}(s) = E[G | S_0 = s, \pi]$$

ج) فرض کنید یک MDP محدود با پاداش های دارای باند مشخص داریم، که تمامی پاداش ها در این MDP منفی هستند. همچنین فرض کنید ضریب کاهش یا discount factor برابر یک می باشد. MDP به صورت finit-horizon می باشد و تابع انتقال و توزیع اولیه حالت ها به صورت قطعی می باشد. (دقت کنید پاداش ها لزوماً صورت قطعی نیستند.)
حال فرض کنید

$$H_{\infty} = (S_0, A_0, R_0, S_1, A_1, R_1, \dots, S_{L-1}, A_{L-1}, R_{L-1})$$

توسط یک سیاست π تولید شده است.

ثابت کنید دنباله زیر صعودی اکید می باشد

$$v^{\pi}(S_0), v^{\pi}(S_1), v^{\pi}(S_2), \dots, v^{\pi}(S_{L-1})$$

پاسخ:

الف) می توان با یک مثال نقض اثبات کرد که این رابطه درست نیست. یک MDP به نام m را فرض کنید. در این MDP یک وضعیت به نام s' وجود دارد که از هیچ وضعیت دیگری قابل دسترس نیست ولی $v^{\pi}(s') \neq 0$. در این حالت بر اساس رابطه بالا، چون $P(s, a, s') = 0$ برابر صفر می باشد، همیشه $v^{\pi}(s') = 0$ می شود که می دانیم حالتی وجود دارد که در آن $P(s, a, s') = 0$ و $v^{\pi}(s') \neq 0$ مخالف صفر باشد. در نتیجه عبارت بالا درست نیست.

ب) فرض کنید منظور از عبارت اول $v_t^\pi(s)$ و منظور از عبارت دوم $v_0^\pi(s)$ باشد. آنگاه داریم:

$$\begin{aligned}
v_t^\pi(s) &= \mathbb{E}[G_t | S_t = s, \pi] \\
&= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}[R_{t+k} | S_t = s, \pi] \\
&= \sum_{a \in A} \pi(s, a) \left(R(s, a) + \sum_{k=1}^{\infty} \gamma^k \mathbb{E}[R_{t+k} | S_t = s, \pi] \right) \\
&= \sum_{a \in A} \pi(s, a) \left(R(s, a) + \sum_{s' \in S} \gamma P(s, a, s') \sum_{a' \in A} \pi(s', a') \left(R(s', a') + \sum_{k=2}^{\infty} \gamma^{k-1} \mathbb{E}[R_{t+k} | S_t = s, \pi] \right) \right) \\
&= \gamma^0 \sum_{a \in A} \pi(s, a) R(s, a) + \gamma^1 \sum_{a \in A} \pi(s, a) \sum_{s' \in S} P(s, a, s') \sum_{a' \in A} \pi(s', a') R(s', a') \\
&\quad + \gamma^2 \sum_{a \in A} \pi(s, a) \sum_{s' \in S} P(s, a, s') \sum_{a' \in A} \pi(s', a') P(s', a', s'') \sum_{a'' \in A} \pi(s'', a'') R(s'', a'') + \dots \\
&= \gamma^0 \sum_{a \in A} \Pr(A_0 = a | S_0 = s) R(s, a) \\
&\quad + \gamma^1 \sum_{a \in A} \Pr(A_0 = a | S_0 = s) \sum_{s' \in S} \Pr(S_1 = s' | A_0 = a, S_0 = s) \sum_{a' \in A} \pi(s', a') R(s', a') + \dots \\
&= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s, \pi \right] \\
&= \mathbb{E}[G | S_0 = s, \pi] \\
&= v_0^\pi(s).
\end{aligned}$$

ج) برای حل این قسمت روابط زیر را می نویسیم

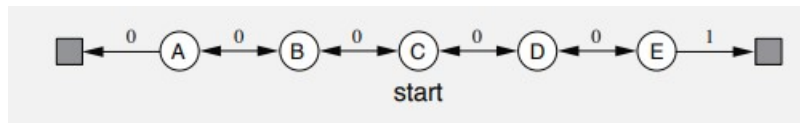
$$\begin{aligned}
v^\pi(S_t) &= v^\pi(s_t) \\
&= v^\pi(s_t) \\
&= \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \mid S_t = s_t, \pi \right] \\
&= \sum_{k=0}^{\infty} \mathbb{E}[R_{t+k} \mid S_t = s_t, \pi] \\
&= \sum_{k=0}^{\infty} \mathbb{E}[R_{t+k} \mid \pi] \\
&= \mathbb{E}[R_t \mid \pi^*] + \sum_{k=0}^{\infty} \mathbb{E}[R_{t+k+1} \mid \pi] \\
&= \mathbb{E}[R_t \mid \pi] + \sum_{k=0}^{\infty} \mathbb{E}[R_{t+k+1} \mid S_{t+1} = s_{t+1}, \pi] \\
&= \mathbb{E}[R_t \mid \pi] + v^\pi(s_{t+1}) \\
&\leq v^\pi(s_{t+1}),
\end{aligned}$$

سوال ۵: (نظری) (۱۰ نمره)

یک فرآیند پاداش مارکوف به صورت زیر را در نظر بگیرید. فرآیند پاداش مارکوف یک فرآیند تصمیم‌گیری مارکوف است که در آن کنش تعریف نمی‌شود. عملاً یک کنش داریم.

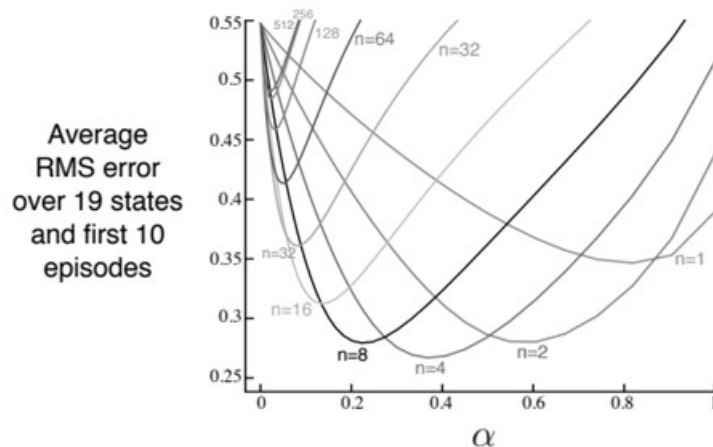
مقادیر اولیه هر وضعیت را برابر 0.5 در نظر بگیرید. همچنین مقدار α را برابر 0.1 در نظر بگیرید.

الف) فرض کنید از الگوریتم TD با n گام برای حل این مسئله استفاده می‌کنیم و در یک اپیزود به ترتیب C، D، E ملاقات می‌شوند. توضیح دهید برای $n = 1$ تا $n = 5$ مقدار ارزش کدام حالت‌ها در این اپیزود آپدیت می‌شود؟



شکل ۲: mrp

ب) در یک آزمایش برای حالتی که به جای ۵ وضعیت، ۱۹ وضعیت داشته باشیم، ۱۰ تکرار و در هر تکرار ۱۰ اپیزود را طی کرده‌ایم و مقادیر ارزش وضعیت‌ها را آپدیت نموده‌ایم. نتایج این آزمایش با مقادیر α و n های مختلف در شکل زیر نشان داده شده است



توضیح دهید مقادیر α چگونه بر مقدار خطا تاثیر می‌گذارد؟

ج) به نظر شما کدام یک از موارد زیر می‌تواند در کاهش خطای نشان داده در نمودار بالا موثر باشد؟ دلیل خود را ذکر کنید (فرض کنید پارامترهای دیگر ثابت هستند و صرفاً موارد ذکر شده تغییر کنند).

(آ) افزایش تعداد حالت‌ها در مسئله

(ب) افزایش تعداد اپیزودها

(ج) افزایش تعداد تکرارها

د) فرض کنید که در یک مسئله خاص، عامل به طور مداوم به همان حالت در یک حلقه برمی‌گردد. بیشترین مقداری که می‌توان توسط اثر پذیرش (eligibility trace) این حالت اخذ شود، در صورتی که ما از اثرات تجمعی با $\lambda = 0.8$, $\gamma = 0.25$ استفاده کنیم، چقدر است؟

پاسخ:

الف) اولاً، می‌دانیم مقدار اولیه ارزش وضعیت‌ها برابر 0.5 می‌باشد. در الگوریتم TD با یک قدم، صرفاً مقدار ارزش $V(E)$ آپدیت شده و برابر یک می‌گردد. در صورتی که از TD با دو قدم استفاده شود، صرفاً مقدار $V(D)$ و $V(E)$ آپدیت می‌شوند و در صورتی که از TD با بیشتر از دو قدم استفاده گردد، مقدار ارزش تمام وضعیت‌های دیده شده را برابر یک قرار می‌دهد

ب)

- پایین (نزدیک به 0): وقتی کوچک است، به‌روزرسانی‌های تابع مقدار عامل کوچک است، به این معنی که عامل بسیار آهسته یاد می‌گیرد. تخمین‌های عامل از تابع ارزش با هر قطعه اطلاعات جدید تغییر بسیار کمی خواهد کرد. اگر تخمین‌های اولیه ضعیف باشند، ممکن است به سوگیری زیادی منجر شود، اما می‌تواند از نوسان یا واگرایی تخمین‌های ارزش به دلیل واریانس زیاد در به‌روزرسانی‌ها جلوگیری کند.

- بالا (نزدیک به 1): بزرگتر به این معنی است که تابع مقدار عامل با شدت بیشتری به روز می شود. این می تواند خوب باشد اگر تخمین های اولیه عامل از مقادیر واقعی دور باشد زیرا امکان یادگیری سریعتر را فراهم می کند. با این حال، همچنین ممکن است به این معنی باشد که تابع مقدار دارای واریانس زیادی خواهد بود، که احتمالاً از مقدار بهینه فراتر می رود و باعث بی ثباتی در یادگیری می شود.
- متوسط: به طور معمول، بهترین نه خیلی زیاد است و نه خیلی پایین، اما تعادلی بین سوگیری و واریانس برقرار می کند و به عامل اجازه می دهد تا به طور موثر یاد بگیرد. جایی که خطای RMS به حداقل می رسد.

(ج)

(الف) افزایش تعداد حالت های MDP (فرایند تصمیم گیری مارکوف): افزایش تعداد حالت ها در یک MDP به طور کلی مشکل را پیچیده تر می کند، زیرا مقادیر حالت بیشتری برای تخمین وجود دارد، که اگر الگوریتم هنوز در مراحل اولیه یادگیری باشد، در ابتدا می تواند خطای RMS را افزایش دهد. حالت های بیشتر به معنای پتانسیل بیشتر برای تخمین های نادرست است، به خصوص اگر تعداد قسمت ها یا میزان یادگیری به نسبت افزایش نیابد. بنابراین، این مرحله لزوماً خطای RMS را کاهش نمی دهد و در بسیاری از موارد ممکن است آن را حداقل در کوتاه مدت افزایش دهد.

(ب) افزایش تعداد قسمت هایی که خطا در آنها محاسبه می شود: اگر الگوریتم یادگیری در طول زمان بهبود یابد، افزایش تعداد قسمت هایی که خطا در آنها میانگین می شود، می تواند منجر به کاهش خطای RMS شود. همانطور که عامل از قسمت های بیشتری یاد می گیرد، تابع خط مشی و ارزشی که استخراج می کند باید به بهینه نزدیک تر شود، با این فرض که نرخ های یادگیری به طور مناسب تنظیم شده اند. بنابراین، میانگین خطا در هر قسمت تمایل به کاهش دارد. با این حال، این فرض را بر این می گذارد که الگوریتم یادگیری به طور موثر خط مشی خود را با قسمت های بیشتر بهبود می بخشد. اگر عامل قبلاً به دلیل تنظیمات ضعیف پارامتر یا محدودیت ها در استراتژی اکتشاف خود به یک خط مشی کمتر از بهینه همگرا شده باشد، صرفاً افزایش تعداد قسمت ها ممکن است منجر به کاهش خطا نشود.

(ج) افزایش تعداد تکرارهایی که خطا بر روی آنها محاسبه می شود: افزایش تعداد تکرارها (میانگین کردن خطا در اجرای بیشتر الگوریتم) می تواند منجر به تخمین دقیق تری از خطای RMS مورد انتظار شود. ممکن است به خودی خود خطای RMS را کاهش ندهد، اما می تواند واریانس در اندازه گیری خطا را کاهش دهد. با تکرارهای بیشتر، تصویر واضح تری از عملکرد میانگین الگوریتم به دست می آوریم، که ممکن است خطای میانگین RMS کمتری را نشان دهد اگر برخی از اجراها به دلیل تصادفی بودن در فرآیند یادگیری یا شرایط اولیه بالا یا پایین بودند.

به طور خلاصه، گزینه (ب) محتمل ترین گزینه برای کاهش خطای RMS به نظر می رسد، اما با این نکته مهم که بستگی به بهبود فرآیند یادگیری در طول زمان دارد. اگر یادگیری بالا رفته باشد، افزایش تعداد قسمت ها منجر به خطای کمتری نخواهد شد. گزینه (ج) خود خطای RMS را کاهش نمی دهد، اما به طور بالقوه تخمین دقیق تری از خطا را با کاهش واریانس در اندازه گیری ارائه می دهد. گزینه (الف) بعید است که خطای RMS را کاهش دهد و در واقع ممکن است آن را افزایش دهد.

(د) حداکثر (eligibility trace) در صورتی رخ می دهد که رابطه زیر برقرار باشد:

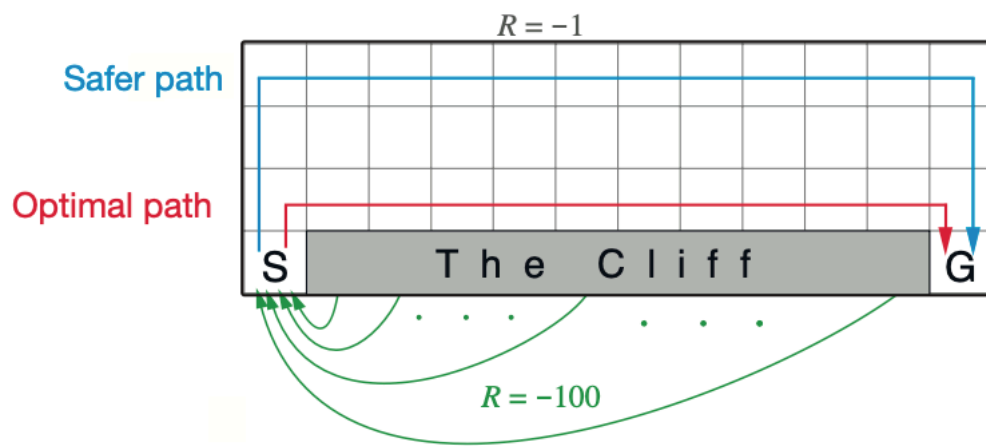
$$e_t(s) = e_{t-1}(s)$$

$$e_t(s) = \gamma \lambda e_{t-1}(s) + 1$$

$$e_t(s) = \frac{1}{1 - \gamma \lambda}$$

سوال ۶: (عملی) (۱۰۰ نمره)

در این بخش شما موظف هستید که الگوریتم های MC و TD را بر روی محیط **Cliff Walking** پیاده سازی و نوتبک داده شده را تکمیل کنید. در بخش MC شما باید الگوریتم Monte Carlo Online Control / On Policy Improvement را، که در آخرین صفحه ای اسلایدهای جلسه ی ششم آمده است، پیاده سازی کنید. با توجه به محدودیت های برآمده از روش های MC که با آن ها در جلسات درس آشنا شدید، به روش های TD روی می آوریم و تلاش می کنیم تا این محیط را با استفاده از الگوریتم های Q-learning و SARSA حل کنیم. پس شما باید این دو الگوریتم را بر روی این محیط اجرا کنید و تفاوت ها و شباهت های آن ها را تشخیص دهید. همچنین قرار است با استفاده از eligibility traces الگوریتم های $Q(\lambda)$ و $SARSA(\lambda)$ را پیاده سازی کنید و تفاوت های این دو الگوریتم را با نسخه ی بدون eligibility traces آن ها مقایسه کنید.



شکل ۳: Walking Cliff

سوال ۷: (عملی) (۱۰۰ نمره)

در این بخش شما باید الگوریتم DQN را بر روی محیط Lunar Lander آموزش دهید و نونتوک داده شده را تکمیل کنید.