



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.
- تمام پاسخ‌های خود را در یک فایل با فرمت zip [Fullname]_[SID]_RL_HW# روی کوثر قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. علاوه بر آن، در هر تمرین می‌توانید تا سقف ۲ روز از تأخیر مجاز باقیمانده خود استفاده کنید و در مجموع ۵ روز تأخیر مجاز برای تمرین در اختیار دارید.

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

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

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

ب) حال پس از اثبات قسمت الف، نشان دهید که در مرحله ارزیابی سیاست به V^π همگرا می‌شویم. 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 می‌باشد، به طور مثال:

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

پ) حال با توجه به اینکه نمی‌توانیم 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:

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

سوال ۲: (نظری) (۱۵ نمره)

یک gridworld به صورت زیر را در نظر بگیرید:

| | | | | |
|------------------|----------|--|----------|-------------------------------|
| 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

کنش‌های ممکن به صورت حرکت به سمت بالا، حرکت به سمت پایین، حرکت به سمت چپ و حرکت به سمت راست می‌باشد. همچنین فرض کنید هنگامی که عامل به یک سمت حرکت می‌کند:

(آ) با احتمال 0.8 در جهتی که می‌خواهد حرکت می‌کند

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

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

(د) با احتمال 0.1 نمی‌تواند حرکتی کند و در جای خود می‌ماند

همچنین در صورتی که عامل با دیوار برخورد کند در جای خود می‌ماند. پاداش تمامی وضعیت‌هایی که ذکر نشده است را معادل صفر در نظر بگیرید. تمامی پاداش‌های مشخص برای هر وضعیت برای ورود به آن وضعیت می‌باشد.

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

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

سوال ۳: (نظری) (۳۰ نمره)

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 و ضریب $\gamma < 1$ 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 درستی یا نادرستی جمله زیر را نشان دهید
در یک MDP با حالات محدود و پاداش با باند مشخص و $\gamma < 1$ در صورتی همه پاداش‌ها با عدد ثابت c جمع شود، سیاست بهینه تغییر نمی‌کند

(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 و $\gamma < 1$ 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 و مثبت c را 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.”

سوال ۴: (نظری) (۳۰ نمره)

در این سوال می‌خواهیم 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 می‌کند معادله بلمن را معکوس کند و مقدار یک حالت را بر اساس مقادیر قبلی بنویسد. این شخص به رابطه زیر می‌رسد. 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 می‌شود

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

(ب) نشان دهید دو عبارت زیر با یکدیگر معادل می‌باشند:

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

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

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

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

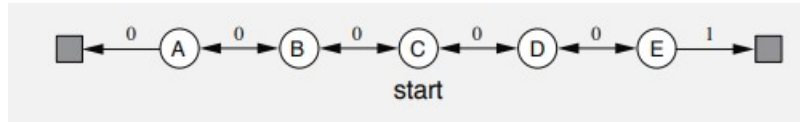
توسط یک سیاست π تولید شده‌است.
ثابت کنید دنباله زیر صعودی اکید می‌باشد

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

سوال ۵: (نظری) (۱۰ نمره)

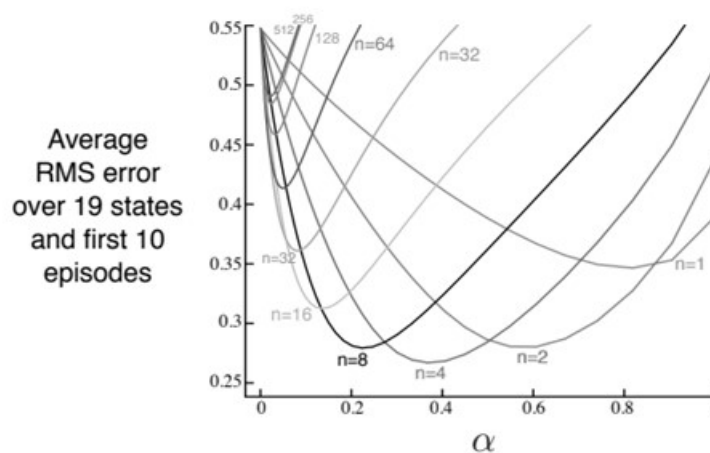
یک فرآیند پاداش مارکوف به صورت زیر را در نظر بگیرید. فرآیند پاداش مارکوف یک فرآیند تصمیم‌گیری مارکوف است که در آن کنش تعریف نمی‌شود. عملاً یک کنش داریم.

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 E، D، C، B، A ملاقات می‌شوند. توضیح دهید برای $n = 1$ تا $n = 5$ مقدار ارزش کدام حالت‌ها در این اپیزود آپدیت می‌شود؟



شکل ۲: mdp

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

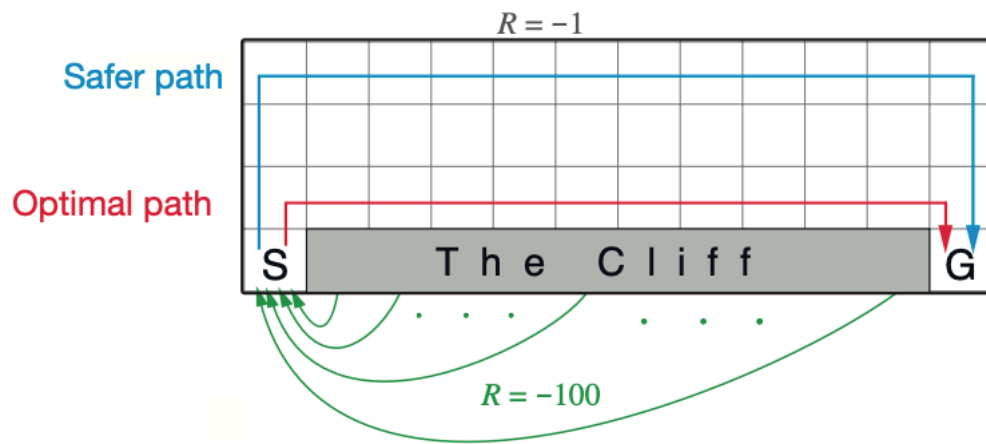
(ب) افزایش تعداد اپیزودها

(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 $\gamma = 0.25$, $\lambda = 0.8$ استفاده کنیم، چقدر است؟

سوال ۶: (عملی) (۱۰ نمره)

در این بخش شما موظف هستید که الگوریتم‌های MC و TD را بر روی محیط **Cliff Walking** پیاده‌سازی و نتایج داده شده را تکمیل کنید.



شکل ۳: Walking Cliff

در بخش MC شما باید الگوریتم Monte Carlo Online Control / On Policy Improvement را، که در آخرین صفحه‌ی اسلایدهای جلسه‌ی ششم آمده است، پیاده‌سازی کنید. با توجه به محدودیت‌های برآمده از روش‌های MC که با آن‌ها در جلسات درس آشنا شدید، به روش‌های TD روی می‌آوریم و تلاش می‌کنیم تا این محیط را با استفاده از الگوریتم‌های Q-learning و SARSA حل کنیم. پس شما باید این دو الگوریتم را بر روی این محیط اجرا کنید و تفاوت‌ها و شباهت‌های آن‌ها را تشخیص دهید. همچنین قرار است با استفاده از eligibility traces الگوریتم‌های $Q(\lambda)$ و $SARSA(\lambda)$ را پیاده‌سازی کنید و تفاوت‌های این دو الگوریتم را با نسخه‌ی بدون eligibility traces آن‌ها مقایسه کنید.

سوال ۷: (عملی) (۱۰۰ نمره)

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