دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر
طراحان: امیدرضا معصومی، مبین باقری، محمد شفیع‌زاده، دنیا جعفری، سپهر ذوالفقاری، علیرضا ملک حسینی

هوش مصنوعی
بهار۱۴۰۴
استاد: احسان تن قطاری

تمرین پنجم | **یادگیری تقویتی** | مهلت ارسال: ۲۱ خرداد

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین سقف ۴ روز و در مجموع ۱۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر ساعت تأخیر غیر مجاز نیم درصد از نمره ی تمرین کم خواهد شد.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

سوالات نظری (۱۰۰ نمره)

۱. (۱۵ نمره) درستی یا نادرستی عبارت‌های زیر را با ذکر دلیل مشخص کنید.

(آ) اگر ضریب تخفیف γ شرط $0 < \gamma < 1$ را ارضا کند، می‌توان تضمین کرد که روش value iteration همگرا می‌شود

(ب) اگر برتر بودن یک سیاست را به معنی این در نظر بگیریم که موجب به دست آوردن reward بیشتری می‌شود، سیاست‌هایی که توسط روش value iteration به‌دست می‌آیند، از سیاست‌هایی که توسط روش policy iteration به‌دست می‌آیند، برتر هستند.

(ج) Q-learning می‌تواند تابع بهینه Q^* را بدون اینکه حتی یک بار سیاست بهینه را اجرا کند بیاموزد.

(د) اگر یک MDP مدل انتقال T ای داشته باشد که برای همهٔ سه‌تایی‌های $T(s, a, s')$ احتمالِ غیرصفر اختصاص دهد، آنگاهQ-learning شکست خواهد خورد.

(ه) فرض کنید عامل ۱ تابع مطلوبیت U_1 و عامل ۲ تابع مطلوبیت U_2 را دارد. اگر

$$U_1 = k_1 U_2 + k_2$$

که در آن $k_1 > 0$ و $k_2 > 0$ باشد، آنگاه عامل ۱ و عامل ۲ ترجیحات یکسانی دارند.

(و) به این سوال پاسخ کوتاه دهید. یک تغییر متداول در یادگیری Q، گنجاندن پاداش‌ها از گام‌های زمانی بیشتر در عبارت X است. بنابراین، عبارت معمول ما

$$r_t + \gamma \cdot \max_{a_{t+1}} q(s_{t+1}, a_{t+1}; w)$$

به این شکل تبدیل می‌شود:

$$r_t + \gamma \cdot r_{t+1} + \gamma^2 \cdot \max_{a_{t+2}} q(s_{t+2}, a_{t+2}; w)$$

به نظر شما مزایای استفاده از پاداش‌های بیشتر در این تخمین چیست؟

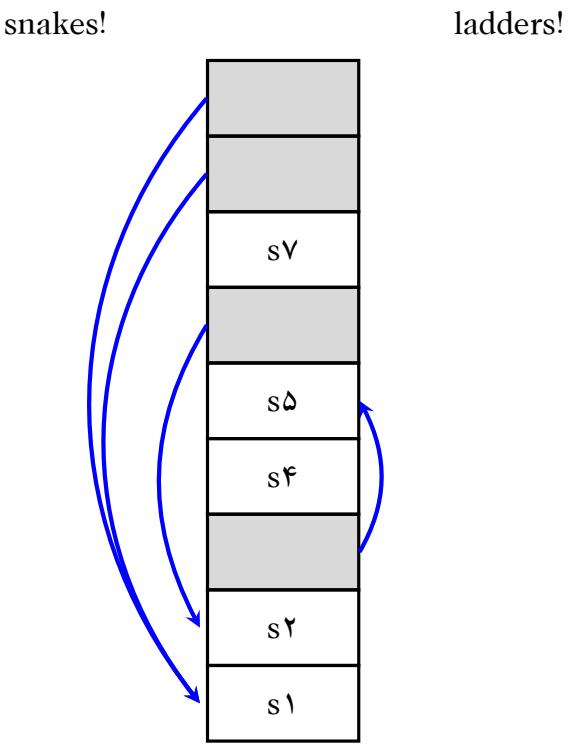
۲. (۱۴ نمره) یک بازی جدید در نظر بگیرید که هربار یک عدد رندوم با توزیع احتمال یکنواخت بین اعداد ۱ تا ۳ تولید می‌شود. شما در هر مرحله دو انتخاب دارید. با انتخاب اول شما می‌توانید به مقدار عدد تولید شده امتیاز بگیرید و بازی تمام شود و یا با انتخاب دوم امتیاز ۱ − بگیرید و ادامه دهید. ۰/۹ = γ در نظر بگیرید.

(آ) برای این بازی یک MDP ارائه دهید و اجزای مختلف آن را مشخص کنید.

(ب) سیاست بهینه‌ای که پس از ۳ مرحله اجرای value iteration بدست می‌آید را حساب کنید.

(ج) با داشتن سیاست اولیه پایان بازی برای اعداد ۱ و ۲ و ادامه برای عدد ۳، با اجرای ۳ مرحله از policy iteration سیاست بهینه را پیدا کنید.

۳. (۱۶ نمره)



شما در حال بازی نسخه‌ای ساده‌شده از یک بازی کلاسیک هستید که «مار و پله» نامیده می‌شود.

- شما در امتداد یک مسیر یک‌بعدی حرکت می‌کنید که خانه‌های $s_۱, s_۲, s_۴, s_۵, s_۷$ را شامل می‌شود.
- شما دو عمل دارید: climb و quit.
- اگر از حالت s_i عمل climb را انتخاب کنید، با احتمال ۰/۵ یک خانه بالا می‌روید و با احتمال ۰/۵ دو خانه بالا می‌روید.
- اما! اگر روی خانه‌ای فرود بیایید که پله‌ای از آن بالا می‌رود، بلافاصله و با احتمال ۱ به خانهٔ مقصد پله منتقل می‌شوید.
- و اگر روی خانه‌ای فرود بیایید که ماری از آن پایین می‌آید، بلافاصله و با احتمال ۱ به خانهٔ مقصد مار سقوط می‌کنید.
- بنابراین، برای مثال، اگر در $s_۵$ باشید و climb کنید:
  - با احتمال ۰/۵ یک خانه بالا می‌روید و به خانهٔ خاکستری $s_۶$ می‌رسید؛ چون $s_۶$ مار دارد، بلافاصله به $s_۲$ سقوط می‌کنید.
  - با احتمال ۰/۵ دو خانه بالا می‌روید و به $s_۷$ می‌رسید.

– اگر در $s_۷$ باشید و climb کنید، ابتدا به خانهٔ خاکستری $s_۸$ می‌روید و چون $s_۸$ مار دارد، به $s_۱$ منتقل می‌شوید.

- اگر عمل quit را انتخاب کنید، بازی تمام می‌شود و دیگر نمی‌توانید حرکتی انجام دهید.
- هر اپیزود جدید از $s_۱$ شروع می‌شود.
- پاداشِ انتخاب climb در هر حالت برابر با ۰ است.
- پاداشِ انتخاب quit در حالت s_i برابر با i است.

(آ) سیاستِ بهینه در افق زمانی یک چیست؟ برای هر یک از حالات $s_۱, s_۲, s_۴, s_۵, s_۷$ بنویسید که عملِ بهینه چیست.

(ب) اگر مقادیر Q تمام حالات را با صفر مقداردهی اولیه کنید و یک دور value iteration بدون تنزیل ($\gamma = ۱$) انجام دهید، تابع Q حاصل چه خواهد بود؟ مقدار $Q(s, \text{quit})$ و $Q(s, \text{climb})$ را برای هر یک از حالات $s_۱, s_۲, s_۴, s_۵, s_۷$ مشخص کنید.

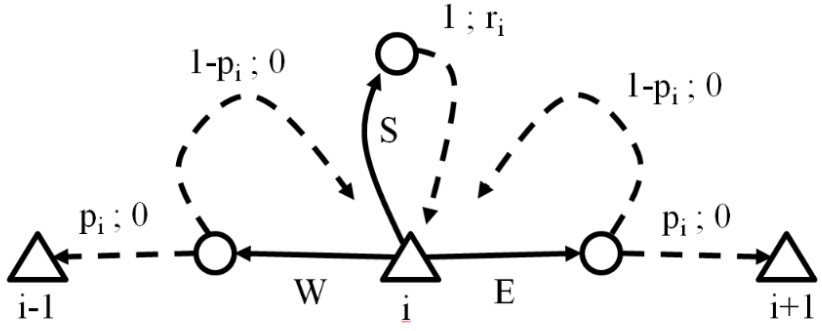
(ج) اگر $\gamma = ۱$ (یعنی هیچ تنزیلی نداریم)، سیاستِ بهینه در افق نامتناهی چیست؟ برای هر یک از حالات $s_۱, s_۲, s_۴, s_۵, s_۷$ بنویسید که عملِ بهینه چیست.

(د) اکنون تنزیل ($\gamma < ۱$) را در نظر بگیرید. نابرابری‌ای بنویسید که شامل مقادیر عددی، γ، $Q(s_۲, \text{climb})$ و $Q(s_۷, \text{climb})$ باشد و شرط را مشخص کند که تحت آن عملِ بهینه در حالت $s_۵$، انتخابِ quit باشد.

۴. (۱۸ نمره)

در امتداد یک بزرگراه اصلی، N شهر وجود دارد که با شماره‌های ۱ تا N شماره‌گذاری شده‌اند. شما تاجری از شهر ۱ هستید (شروع شما از آنجاست). هر روز، می‌توانید یکی از کارهای زیر را انجام دهید: به شهر مجاور بروید (حرکت به سمت شرق یا غرب)، یا در شهر فعلی بمانید و به تجارت بپردازید (عمل Stay). اگر تصمیم بگیرید از شهر i سفر کنید، با احتمال p_i با موفقیت به شهر بعدی می‌رسید، اما با احتمال $۱ - p_i$ گرفتار طوفان می‌شوید و در این صورت روزتان هدر می‌رود و هیچ‌جا نمی‌روید. اگر تصمیم بگیرید در شهر i بمانید و تجارت کنید، پاداشی برابر $r_i > ۰$ دریافت می‌کنید؛ روزهای سفر پاداشی برابر با صفر دارند، چه موفق به رفتن به شهر دیگر شوید و چه نشوید.

نمودار زیر، اقدامات و انتقال‌ها از شهر i را نشان می‌دهد. پیکان‌های پررنگ نشان‌دهنده اقدامات هستند؛ پیکان‌های خط‌چین انتقال‌های حاصل را با برچسبی شامل احتمال و پاداش (به همین ترتیب) نمایش می‌دهند.



(آ) اگر برای همه i، داشته باشیم $r_i = ۱$، $p_i = ۱$ و $\gamma = ۰/۵$ باشد، مقدار $V_{\text{stay}}(۱)$ در حالتی که همیشه stay انتخاب می‌شود چقدر است؟

(ب) اگر برای همه i، داشته باشیم $r_i = ۱$، $p_i = ۱$ و $\gamma = ۰/۵$ باشد، مقدار بهینه $V^*(۱)$ بودن در شهر ۱ چقدر است؟

(ج) اگر مقادیر r_i و p_i اعداد مثبت معلوم باشند و discount factor تقریباً برابر با یک باشد ($\gamma \approx ۱$)، سیاست بهینه را توصیف کنید.

(د) فرض کنید از الگوریتم Value Iteration استفاده می‌کنیم. $V_k(s)$ مقدار حالت s پس از k مرحله مقدارگذاری است و همه مقادیر اولیه صفر هستند.
اگر مقدار بهینه برای بودن در شهر ۱ مثبت باشد یعنی $V^*(۱) > ۰$، بیشترین مقدار k که در آن ممکن است $V_k(۱) = ۰$ باشد چیست؟

(ه) اگر همه‌ی r_i و p_i مثبت باشند، بیشترین مقدار k که در آن ممکن است $V_k(s) = ۰$ برای برخی حالت‌ها باشد چیست؟

(و) فرض کنید r_iها و p_iها را نمی‌دانیم، بنابراین تصمیم می‌گیریم از Q-Learning استفاده کنیم.
فرض کنید دنباله زیر از حالت‌ها، کنش‌ها و پاداش‌ها تجربه شده است:
(s=1, a=stay, r=4)،
(s=1, a=east, r=0)،
(s=2, a=stay, r=6)،
(s=2, a=west, r=0)،
(s=1, a=stay, r=4)

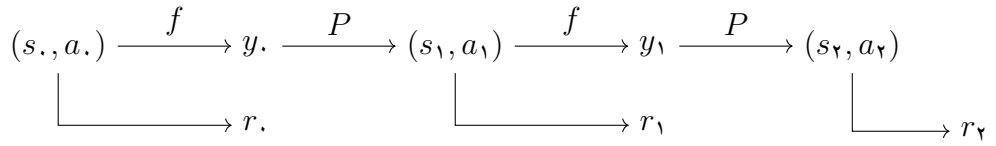
اگر نرخ یادگیری برابر ۰/۵، discount factor برابر ۱ و همه مقادیر اولیه $Q(s,a) = ۰$ باشند، مقدارهای نهایی $Q(s,a)$ را در جدول زیر وارد کنید. هر سطر باید مقادیر q-value پس از انتقال مشخص‌شده در ستون اول را نشان دهد. مقادیر بدون تغییر را می‌توان خالی گذاشت.

(s,a,r,s')	Q(1,S)	Q(1,E)	Q(2,W)	Q(2,S)
initial	0	0	0	0
(1,S,4,1)				
(1,E,0,2)				
(2,S,6,2)				
(2,W,0,1)				
(1,S,4,1)				

۵. (۱۸ نمره) یک مدل MDP با افق نامتناهی و ضریب تخفیف γ با مشخصات زیر در نظر بگیرید:

$$T(s,a,s') = P(s'|f(s,a)), \quad R(s,a,s') = R(s,a)$$

که در آن $f : S \times A \to Y$ یک تابع قطعی به مجموعه حالت‌های پس‌ازتصمیم Y است. دنباله حالت‌ها s_t، اقدامات a_t، حالت‌های پس‌ازتصمیم y_t و پاداش‌ها r_t به صورت زیر است:



مقدار مورد انتظار با تخفیف پاداش‌ها تحت سیاست π به صورت زیر تعریف می‌شود:

$$V^\pi(s_۰) = \mathbb{E}\left[\sum_{t=۰}^{\infty} \gamma^t R(s_t, a_t) \;\middle|\; a_t = \pi(s_t)\right]$$

و تابع ارزش بهینه به صورت $V^*(s) = \max_\pi V^\pi(s)$. برای حالت‌های پس‌ازتصمیم:

$$W^\pi(y.) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t) \mid y.\right], \quad W^*(y) = \max_\pi W^\pi(y)$$

(آ) W^* را بر حسب V^* به دست آورید.

(ب) V^* را بر حسب W^* بنویسید. راهنمایی: ابتدا V^* را بر حسب Q بنویسید و سپس آن را بر اساس W^* بنویسید.

(ج) می‌دانیم معادله بلمن فورد برای تابع ارزش به شکل زیر تعریف می‌شود.

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

معادل این معادله برای W^* را ارائه دهید.

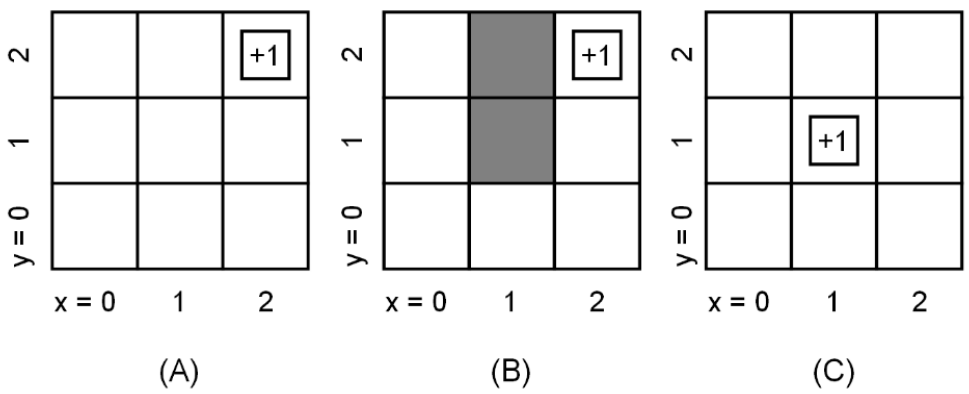
(د) جاهای خالی را در الگوریتم زیر طوری پر کنید که یک الگوریتم policy iteration به دست آید. این الگوریتم باید تضمین کند که π^* را پیدا می‌کند. همچنین هر مورد را توضیح دهید:

- سیاست $\pi^{(1)}$ را به صورت دلخواه مقداردهی اولیه کنید
- برای $i = 1, 2, 3, \ldots$:
  - $W^{\pi^{(i)}}(y)$ را برای تمام $y \in Y$ محاسبه کن
  - سیاست جدید $\pi^{(i+1)}$ را طوری محاسبه کن که _____ $\pi^{(i+1)}(s) = \arg\max_a$
  - اگر _____ برای تمام $s \in S$، $\pi^{(i)}$ را برگردان

(ه) با توجه به دنباله مشاهده شده $s_t, a_t, y_t, s_{t+1}, a_{t+1}, y_{t+1}$ و نرخ یادگیری $\alpha \in (0, 1)$، قاعده به‌روزرسانی برای تخمین W^* را مشابه قاعده به‌روزرسانی الگوریتم Q-learning کامل کنید.

$$W(y_t) \leftarrow (1-\alpha) W(y_t) + \alpha \_\_\_\_\_$$

۶. (۱۹ نمره) در مسئله gridworld زیر، عامل می‌تواند اعمال N، S، E، W را انجام دهد که عامل را به ترتیب به سمت شمال، جنوب، شرق و غرب یک خانه حرکت می‌دهند. هیچ نویزی وجود ندارد، بنابراین این اعمال همیشه عامل را در جهت مورد نظر حرکت می‌دهند، مگر آنکه آن جهت به بیرون از شبکه یا به خانه‌ای مسدود (خاکستری) منتهی شود، که در آن صورت هیچ حرکتی انجام نمی‌شود. خانه‌هایی که دارای +۱ هستند همچنین اجازهٔ اجرای عمل X را می‌دهند که عامل را از شبکه خارج کرده و وارد حالت نهایی (terminal) می‌کند. پاداش برای تمام انتقال‌ها صفر است، به جز انتقال خروج که پاداش آن +۱ است. discount factor برابر 0.5 فرض می‌شود.



(آ) مقدارهای بهینه برای شبکه (A) را کامل کنید.

(ب) سیاست بهینه برای شبکه (B) را مشخص کنید.

(ج) فرض کنید برای هر حالت غیرنهایی $s = (x, y)$ مجموعه‌ای از ویژگی‌های حقیقی $f_i(s)$ داریم، و می‌خواهیم مقدار بهینه $V^*(s)$ را با رابطهٔ خطی زیر تقریب بزنیم:

$$V(s) = \sum_i w_i \cdot f_i(s)$$

اگر ویژگی‌ها $f_۱(x, y) = x$ و $f_۲(x, y) = y$ باشند، مقادیر $w_۱$ و $w_۲$ را طوری بدهید که سیاست به‌دست‌آمده از (one-step look-ahead) در شبکه (A) بهینه باشد.

(د) آیا می‌توان مقادیر واقعی و بهینه V^* را برای شبکه (A) تنها با استفاده از این دو ویژگی نمایش داد؟ چرا یا چرا نه؟

(ه) برای هر یک از مجموعه ویژگی‌های زیر مشخص کنید کدام یک (در صورت وجود) از هایMDP شبکه‌ای بالا می‌توانند «حل شوند»؛ به این معنا که بتوان مقادیری (احتمالاً غیربهینه) برای $V(s)$ پیدا کرد که سیاست حاصل از نگاه یک‌مرحله‌ای به آینده، بهینه باشد.

i. $f_۱(x, y) = x$ ، $f_۲(x, y) = y$

ii. برای هر (i, j)، ویژگی $f_{i,j}(x, y) = (x, y) = (i, j)$ اگر ۱، و صفر در غیر این صورت.

iii. $f_۱(x, y) = (x - ۱)^۲$ ، $f_۲(x, y) = (y - ۱)^۲$ ، و $f_۳(x, y) = ۱$