به نام خدا

# یادگیری تقویتی

زمستان ۱۴۰۲

مدرس: محمدحسین رهبان

زمان: ۱۵۰ دقیقه

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

امتحان پایان‌ترم

تاریخ: ۳۱ خرداد ۱۴۰۳

## سوالات (۱۰۰ نمره)

۱. (۲۰ نمره) به سوالات زیر در مورد Inverse Reinforcement Learning (IRL) و Imitation Learning (IL) پاسخ دهید.

(آ) در Feature Matching IRL قصد حل بهینه‌سازی زیر را داریم:

$$\max_{\psi,m} \quad m \quad \text{s.t} \quad \psi^T E_{\pi^\star}[\mathbf{f}(s,a)] \geq \max_{\pi\in\Pi} \psi^T E_{\pi}[\mathbf{f}(s,a)] + m \qquad (۱)$$

مشکل این کار چیست؟ آن را اصلاح کنید.

(ب) در یادگیری متغیر بهینگی (optimality) به منظور یادگیری تقویتی معکوس، با تابع هدف زیر روبرو هستیم:

$$\max_{\psi} \frac{۱}{N}\sum_{i=۱}^{N} r_\psi(\tau_i) - \log Z.$$

توضیح دهید این تابع هدف با چه منطقی بدست آمده است. همینطور، چگونه می‌توان گرادیان جمله $\log Z$ نسبت به $\psi$ را صرفا با استفاده از نمونه‌برداری تخمین زد؟

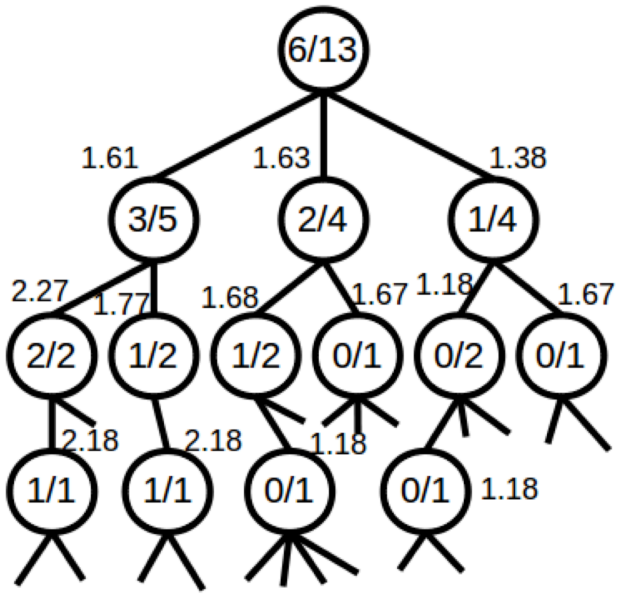
(ج) چرا در روش DAgger مقدار $\beta_i$ که احتمال استفاده از سیاست خبره (expert) در جمع‌آوری داده در تکرار $i$ است، با بزرگ شدن $i$ به صفر میل داده می‌شود؟

پاسخ: در ابتدای کار $\beta_i = ۱$ است. به این ترتیب در همان ابتدا فقط سیاست خبره برای جمع‌آوری داده و آموزش سیاست جدید $\hat{\pi}_i$ استفاده می‌شود. به تدریج که سیاست $\hat{\pi}_i$ آموزش می‌بیند، اولا بهتر می‌شود و ثانیا ممکن است مجموعه حالت‌هایی که با پیروی از آن مشاهده می‌شود متفاوت از سیاست خبره باشد. لذا داده‌هایی که جمع‌آوری می‌شود بایستی روی همین حالات متمرکز شود. به همین دلیل، در طول زمان $\beta_i \to ۰$.

۲. (۲۰ نمره) به سوالات زیر در مورد Model Based Reinforcement Learning (RL) پاسخ دهید.

(آ) در ابتدا صفحه ۲ یک Monte-Carlo tree search (MCTS) را مشاهده می‌کنید که در آن عدد سمت چپ خط مورب تعداد مشاهده، عدد سمت راست خط مورب، Q-Value و عدد بالای هر نود UCB Value است. نودهایی که هنوز مشاهده نشده‌اند به صورت شاخه‌های بدون حباب نشان داده شده‌اند. نشان دهید که random rollout بعدی در کدام نود رخ می‌دهد. اگر random rollout بعدی به یک شکست منتج شود، برای هرکدام از UCB Value ها درخت نشان دهید که آیا این مقدار افزایش، کاهش یا عدم تغییر مقدار را تجربه خواهد کرد. (پیروزی امتیاز ۱ و شکست امتیاز ۱− را دارد.)

(ب) در یک روش مبتنی بر مدل با استفاده از برنامه‌ریزی مبتی بر Cross-Entropy Maximization، تابع $p(s_{t+۱}|s_t,a_t)$ را به کمک یک شبکه ژرف مدل‌سازی کرده‌ایم. فرض کنید دینامیک تغییرات حالت بسیار پیچیده است و لذا از یک شبکه بسیار بزرگ برای این مدل‌سازی استفاده کرده‌ایم. چه چالشی پیش می‌آید و برای حل آن باید چه راهکاری استفاده کنیم؟




پاسخ: در این حالت به دلیل بزرگ بودن شبکه، و اینکه در ابتدای آموزش داده کمی برای آموزش وجود دارد، شبکه دچار بیش‌برازش می‌شود و ممکن است مسیرهایی در فضای حالت را به عنوان نتیجه برنامه‌ریزی پیدا کند که پاداش بالا در آنها خوشبینانه باشد. در این حالت برای واقع‌بین کردن شبکه نیاز است عدم قطعیت در مقدار پاداش را که ناشی از عدم قطعیت معرفتی epistemic uncertainty است را اندازه‌گیری و در برنامه‌ریزی دخیل کنیم. یک راه استاندارد برای این کار، استفاده از ensemble ای از مدل‌ها است. به گونه‌ای که چند شبکه عصبی که با وزن‌های متفاوت مقداردهی اولیه شده‌اند را استفاده می‌کنیم و دنباله اعمال مورد بهینه‌سازی در cross-entropy maximization را به کمک هر کدام از آنها مورد ارزیابی و محسابه پاداش قرار می‌دهیم. سپس از پاداش محاسبه شده متوسط می‌گیریم تا عدم قطعیت مدل شده باشد.

۳. (۲۰ نمره) در روش Upper Confidence Bound، در مسئله Multi-Armed Bandit، مقدار $\sqrt{\frac{۲ \log T}{N(a)}}$ را به میانگین تجربی پاداش مشاهده شده در یک عمل اضافه می‌کنیم و بعدی عملی که بیشترین مقدار را کسب کند، انتخاب می‌کنیم. در این عبارت $N(a)$ تعداد اجراهای عمل $a$ در مرحله $T$ام است. اولا به صورت شهودی دلیل استفاده از جمله $\log T$ را توضیح دهید. ثانیا این مسئله را به لحاظ تئوری توجیه کنید.

۴. (۲۰ نمره) به سوالات زیر پاسخ دهید:

- تفاوت offline-RL با روش‌های imitation learning در چیست؟

پاسخ: در یادگیری تقلیدی، هدف الزاما تقلید خبره است و این در حالی است که خبره ممکن است بهینه نباشد. لذا سیاستی بهتر از خبره یاد گرفته نخواهد شد. در یادگیری تقویتی غیربرخط، اولا ممکن است دادگان غیربرخط جمع‌آوری شده الزاما کیفیت مطلوبی نداشته باشند. ثانیا هدف این است که به سیاستی برسیم که از سیاست‌های مورد استفاده برای جمع‌آوری داده بهتر باشد.

- بهینه‌سازی min-max استفاده شده در روش conservative Q-Learning را بیان کنید و دلیل وجود هر کدام از اجزای آن را بنویسید.

پاسخ: اسلاید ۳۰ مباحث offline-RL: دلیل جمله اول این است که اعمالی که تابع Q نسبت به آنها خوشبینی کاذب دارد را پنالتی بدهیم. جمله دوم به نوعی می‌خواهد اعمالی که در داخل خود دادگان وجود دارد و خوشبینی و مقدار بالای Q آنها بیجهت نیست را پنالتی ندهیم. جمله سوم برای منظم‌سازی توزیع کنش‌ها است که این توزیع روی یک عمل خاص متمرکز نشود. جمله چهارم هم در نهایت برای کمینه‌کردن خطای temporal difference است.

- مسئله بهینه‌سازی بخش قبل را به صورت یک بهینه‌سازی فقط مبتنی بر min بنویسید (با راه حل).

پاسخ: برای اینکار کافی است نسبت به $\mu(a|s)$ مشتق گرفته و مساوی صفر قرار دهیم:

$$\frac{\partial}{\partial \mu(a|s)} \sum_{a'} \mu(a'|s) Q(s,a') - \beta \sum_{a'} \mu(a'|s) \log \mu(a'|s) + \lambda(\sum_{a'} \mu(a'|s) - ۱) = ۰.$$

لذا خواهیم داشت:

$$Q(s,a) - \beta \log \mu(a|s) - \beta + \lambda = ۰.$$

در نتیجه:

$$\mu(a|s) = \exp \frac{\lambda - \beta}{\beta} \exp\left(\frac{۱}{\beta} Q(s,a)\right).$$

با اعمال قاعده مساوی یک بودن مجموع احتمالات، در نهایت بدست می‌آید:

$$\mu(a|s) = \frac{\exp Q(s,a)/\beta}{\sum_{a'} \exp Q(s,a')/\beta}.$$

۵. (۲۰ نمره) در الگوریتم Soft Actor-Critic، در هر گام به روزرسانی سیاست، در حالت ایده‌آل مسئله بهینه‌سازی زیر را حل می‌کنیم:

$$\pi_{\text{new}} = \arg\min_{\pi'} D_{\text{KL}}(\pi'(.|s) \parallel ۱/Z \exp Q^{\pi^{\text{old}}}(s,.)).$$

نشان دهید به ازاء هر حالت و عمل دلخواهی مقدار $Q^{\pi^{\text{new}}}$ سیاست نرم یا soft policy بهتر از $Q^{\pi^{\text{old}}}$ خواهد شد.

پاسخ: فرض کنید:

$$J_\pi(\pi') = \mathbb{E}_{a\sim\pi'} Q^\pi(s,a) + \mathcal{H}(\pi'(.|s)).$$

حال توجه کنید که با بهینه‌سازی تابع هدف مطرح شده در صورت سوال داریم:

$$D_{\text{KL}}(\pi'(.|s) \parallel ۱/Z \exp Q^{\pi^{\text{old}}}(s,.)) = \sum_a -\pi'(a|s) Q^{\pi^{\text{old}}}(s,a) + \pi'(a|s) \log \pi'(a|s) + \log Z.$$

که برابر با

$$-J_{\pi^{\text{old}}}(\pi')$$

است. حال توجه کنید که

$$J_{\pi^{\text{old}}}(\pi_{\text{new}}) \geq J_{\pi^{\text{old}}}(\pi^{\text{old}}).$$

چراکه در بهینه‌سازی مطرح شده در صورت سوال، حتما $\pi_{\text{new}}$ بهتر از هر سیاست دیگر از جمله $\pi^{\text{old}}$ است. حال بدست می‌آید:

$$\mathbb{E}_{a\sim\pi_{\text{new}}} Q^{\pi^{\text{old}}}(s,a) - \log \pi_{\text{new}}(a|s) \geq \mathbb{E}_{a\sim\pi^{\text{old}}} Q^{\pi^{\text{old}}}(s,a) - \log \pi^{\text{old}}(a|s) = V^{\pi^{\text{old}}}(s).$$

حال معادله soft Bellman را در نظر بگیرید:

$$\begin{aligned}
Q^{\pi^{\text{old}}}(s_t,a_t) &= r(s_t,a_t) + \gamma \mathbb{E}_{s_{t+۱}} V^{\pi^{\text{old}}}(s_{t+۱}) \\
&\leq r(s_t,a_t) + \gamma \mathbb{E}_{s_{t+۱}} \mathbb{E}_{a_{t+۱}\sim\pi_{\text{new}}} Q^{\pi^{\text{old}}}(s_{t+۱},a_{t+۱}) - \log \pi_{\text{new}}(a_{t+۱}|s_{t+۱}) \\
&\vdots \\
&\leq Q^{\pi_{\text{new}}}(s_t,a_t)
\end{aligned}$$

۶. (۲۰ نمره) در رابطه با الگوریتم Deep Deterministic Policy Gradient (DDPG) به سوالات زیر پاسخ دهید. (راهنمایی: پاسخ تمام بخش‌ها (جز قسمت (ج) که نیاز به کمی توضیح دارد) نصف خط و بعضا یکی دو کلمه است)

(آ) چرا نمی‌توان از Q-learning در حالتی که فضای کنش پیوسته باشد استفاده کرد؟

زیرا امکان محاسبه‌ی $max$ در فضای پیوسته وجود ندارد.

(ب) در الگوریتم DDPG برای دور زدن مشکل گفته شده، بیان می‌شود که به جای محاسبه $\max_a Q(s,a)$، از یک شبکه‌ی actor برای محاسبه‌ی $...argmax_a Q(s,a)...$ استفاده می‌شود. نوع خروجی actor در این الگوریتم با الگوریتم actor-critic چه تفاوتی دارد؟

خروجی actor در الگوریتم actor-critic توزیع احتمال روی کنش‌هاست ولی در DDPG یک کنش به صورت deterministic است.

(ج) ممکن است یکی از ورودی‌های critic در الگوریتم DDPG بعد پایین و دیگری بعد بالا داشته باشد. یک راه برای حل مشکل تفاوت اندازه‌ی ابعاد دو ورودی بیان نمایید.

یکی از موارد زیر:

i. به جای ورودی دادن کنش به صورت مستقیم به critic، کنش در لایه‌های مختلف به آن ورودی داده شود یا در آن‌ها ضرب شود.

ii. کنش ابتدا به بعد بالاتری (مثلا با استفاده از بیک شبکه‌ی دیگر) برده شود و سپس به critic ورودی داده شود.

(د) برای آموزش هر دو شبکه‌ی actor و critic در DDPG از الگوریتم‌های مبتنی بر گرادیان کاهشی استفاده می‌شود. تابع هزینه [۱] برای هر کدام را بنویسید. برای یکسان‌سازی نمادها، شبکه‌ی actor و شبکه‌ی مؤخر آن را به ترتیب $Q_\theta$ و $Q_{\theta'}$، شبکه‌ی critic و شبکه‌ی مؤخر آن را $\mu_\phi$ و $\mu_{\phi'}$، و mini-batch نمونه‌برداری شده را به صورت $\{s_j, a_j, s'_j, r_j\}$ و ضریب کاهش پاداش را $\gamma$ در نظر بگیرید. نیازی به نوشتن رابطه‌ی بهینه‌سازی پارامترها نیست.

$$\text{critic:}(Q_\theta(s_j,a_j) - (r_j + \gamma Q_{\theta'}(s'_j, \mu_{\phi'}(s'_j))))^2$$

$$\text{actor:} - Q_\theta(s_j, \mu_\phi(s_j))$$

(ه) به خروجی actor در DDPG یک نویز گاوسی اضافه می‌گردد.

- این کار برای حل چه مشکلیست؟ exploration
- آیا این روش قابلیت اعمال در زمانی که فضای کنش گسسته است را دارد؟
  خیر. اعمال نویز گاوسی بر روی خروجی پیوسته امکان‌پذیر نیست.
- ممکن است اعمال این روش باعث مشکل در همگرایی DDPG شود. در این صورت، چه تغییری در نوع اعمال روش در طول زمان بدهیم تا مشکل همگرایی مرتفع شود؟
  واریانس نویز را در طول زمان کاهشی در نظر می‌گیریم.
- temporal correlation چه مزیتی به روش گفته شده اضافه می‌کند؟
  کمک می‌کند که به exploration زمان کافی داشته باشد تا تاثیر خودش را نشان دهد.

[۱] loss function