



یادگیری تقویتی

نیم سال بهار ۱۴۰۱-۴۰۲

اساتید: دکتر رهبان، آقای حسنی

تمرین سری سوم

الگوریتم‌های مبتنی بر مدل و روش‌های بیزی

زمان تحویل: ۲۲ اردیبهشت

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 `[Fullname].[SID]_RL_HW#.zip` 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 پنج روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید.

سوال ۱: (نظری) LQR (۱۵ نمره)

- (آ) روش LQR در تمامی مسائل مربوط به یادگیری تقویتی همگرا نمی‌شود. شرایط لازم یک محیط و ایجنت برای همگرا شدن این الگوریتم چیست؟
- (ب) همان‌طور که در بخش قبلی پاسخ دادید، یکی از شرایط لازم $fully\ observable$ بودن محیط است تا بتوان از LQR استفاده کرد. چگونه می‌توان از این روش برای محیط‌های $partially\ observable$ استفاده کرد؟
- (ج) چگونه می‌توان از روش LQR در کنار روش‌های $model\ free$ که از شبکه‌های عصبی عمیق استفاده می‌کنند، بهره برد؟
- (د) با الهام گرفتن از آنچه که در کلاس در ارتباط با اعمال LQR در محیط‌های تصادفی دیدید، چگونه می‌توان از روش iLQR برای برطرف کردن عدم قطعیت محیط و یا $exploration$ استفاده کرد؟

پاسخ:

- (آ) شرایط همگرایی روش LQR در یادگیری تقویتی به عوامل متعددی از جمله ماهیت سیستم تحت کنترل، انتخاب تابع هزینه و طراحی خود کنترل کننده بستگی دارد. با این حال، به طور کلی، روش LQR زمانی موثرتر است که برای سیستم‌هایی استفاده شود که خطی و تغییرناپذیر زمان هستند و در آن‌ها تابع هزینه ماهیت درجه دوم دارد. علاوه بر این، روش LQR معمولاً نیاز به دسترسی به یک مدل دقیق از دینامیک سیستم ($Fully\ observable$) دارد، که می‌تواند یک عامل محدودکننده در برخی از برنامه‌های یادگیری تقویتی باشد که در آن مدل سیستم ناشناخته یا نامشخص است. با این وجود، تحت شرایط مناسب، روش LQR می‌تواند به یک کنترل کننده بهینه همگرا شود که عملکرد بهینه را در مسئله کنترل داده شده ارائه می‌دهد.
- (ب) برای غلبه بر این محدودیت، محیط‌های کاملاً قابل مشاهده در محیط‌های قابل مشاهده جزئی، یک رویکرد ممکن استفاده از فیلتر Kalman یا سایر تکنیک‌های تخمین برای تخمین $partial\ state$ محیط است. رویکرد دیگر استفاده از روش‌های یادگیری تقویتی $model-based$ است، مانند کنترل پیش‌بینی مدل، که در آن یک مدل دینامیک از سیستم یاد گرفته می‌شود و برای تولید $policy$ استفاده می‌شود. علاوه بر این، رویکرد دیگر استفاده از LQR با بازخورد $partial\ state$ است، که در آن تنها زیر مجموعه‌ای از متغیرهای حالت برای بازخورد استفاده می‌شود، به جای استفاده از حالت کامل. این رویکرد زمانی می‌تواند موثر باشد که حالت جزئی حاوی اطلاعات کافی برای کنترل سیستم باشد. به طور کلی، چندین رویکرد برای غلبه بر محدودیت محیط‌های کاملاً قابل مشاهده در محیط‌های قابل مشاهده جزئی وجود دارد و رویکرد مناسب با توجه به محیط، تغییر می‌کند.
- (ج) ترکیب روش‌های یادگیری تقویتی $model-free$ با LQR می‌تواند یک رویکرد مؤثر برای کنترل سیستم‌های $partially\ observable$ ارائه کند. یکی از راه‌های انجام این کار، استفاده از شبکه‌های عصبی عمیق برای تقریب تابع $value$ یا $policy$ در الگوریتم یادگیری تقویتی $model-free$ ، و سپس استفاده از $policy$ حاصل برای تولید دستورات مرجع برای کنترل کننده LQR است. رویکرد دیگر استفاده از LQR با بازخورد

partial state به عنوان معماری baseline است و سپس از الگوریتم یادگیری تقویتی بدون مدل برای یادگیری یک خط مشی اغتشاش استفاده می کند که می تواند عملکرد کنترل کننده LQR را در محیط های نامشخص یا در حال تغییر بهبود بخشد. این رویکرد می تواند از نقاط قوت استراتژی های کنترل model-free و model-based برای دستیابی به کنترل مؤثر در طیف وسیعی از محیط های partially observable استفاده کند.

(د) iLQR می تواند برای غلبه بر عدم قطعیت در محیط یا برای اکتشاف (exploration) در زمینه یادگیری تقویتی استفاده شود. یک رویکرد گنجانیدن عدم قطعیت در مدل دینامیک سیستم و استفاده از الگوریتم iLQR برای بهبود کنترل کننده و در عین حال در نظر گرفتن عدم قطعیت است. رویکرد دیگر استفاده از iLQR در یک محیط یادگیری تقویتی model-free است که در آن از الگوریتم iLQR برای بهینه سازی سیاست های کنترل بر اساس داده های نمونه گیری شده استفاده می شود. این سیاست ها را می توان برای کشف محیط و یادگیری در مورد پویایی سیستم، بهبود عملکرد کنترل کننده در طول زمان استفاده کرد. علاوه بر این، iLQR را می توان همراه با سایر استراتژی های اکسپلور، مانند اکتشاف مبتنی بر کنجکاوی یا model-based planning، برای افزایش اکتشاف و بهبود عملکرد کنترل در محیط های نامشخص استفاده کرد.

سوال ۲: (نظری) بازی اتاق فرار جایزه دار (۳۰ نمره)

سروش یکی از دانشجویان فعال درس ۴۰۹۵۷ است. اخیراً یکی از دوستان او به نام روزبه، بازی خطرناک ولی وسوسه انگیزی را به او معرفی کرده است. این بازی به این صورت است که با پرداخت ۱۰ دلار، وارد یک اتاق بزرگ می شوید که در آن قفل است و باید راه حل برون رفت را درون اتاق بیابید. در صورت یافتن راه حل، علاوه بر بیرون آمدن از اتاق پاداش دلاری با مقدار تصادفی ای دریافت خواهید کرد که کران بالای آن بینهایت است! اما نکته غم انگیز و ترسناک ماجرا هم در این است که مدت زمان گیر کردن در اتاق هم کران بالا ندارد و ممکن است تا پایان عمر طول بکشد! پیدا کردن راه حل خروج از اتاق به ویژگی های خود اتاق بستگی دارد و تجربه های قبلی شخص از جست و جو در اتاق باعث تقویت مهارت او در جست و جو نخواهد شد!

سروش که دانشجوی باهوشی است اقدام به مدلسازی مسئله می کند. او مسئله ی شرکت کردن متوالی در بازی را به صورت یک مسئله ی تصمیم گیری دنباله ای در نظر می گیرد که در آن متغیر حالت، زمان بوده و فضای تصمیم به صورت تصمیم باینری شروع بازی است. او برای پاداش یک مدلسازی به صورت توزیع گاما و برای مدت زمان بین دو نقطه ی تصمیم گیری متوالی توزیع نمایی در نظر می گیرد:

$$\begin{aligned} r \sim \text{Gamma}(\alpha, \beta) &\rightarrow p(r | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r} \\ t \sim \text{Exp}(\lambda) &\rightarrow p(t | \lambda) = \lambda e^{-\lambda t} \end{aligned} \quad (1)$$

سروش که دانشجوی محتاطی هم هست، سعی در جمع آوری اطلاعات موجود کرده و ابتدا از تجربه ی خود روزبه می پرسد. روزبه در پاسخ می گوید من دو بار متوالی در بازی شرکت کردم و به ترتیب ۵ ساعت و ۱۵ ساعت در اتاق گیر کردم ولی پاداش هایی به اندازه ۲۰۰ و ۱۰۰ دلار دریافت کرده ام. سروش با توجه به این اطلاعات و با استفاده از تکنیک بیشینه درست نمایی، یک توزیع پیشین برای متغیر مجهول مدل پاداش (برای سادگی α را مشخص و فقط β را مجهول در نظر می گیریم) و مدل زمان به صورت زیر به دست می آورد:

$$\begin{aligned} \beta &\sim \text{Gamma}(\epsilon, \omega) \\ \lambda &\sim \text{Gamma}(\sigma, \eta) \end{aligned} \quad (2)$$

در نهایت سروش تصمیم به شروع بازی گرفته و در اولین تلاش به اندازه ی t_1 ساعت در اتاق مانده و به هنگام خروج پاداش r_1 دریافت می کند. حال به سوالات زیر پاسخ دهید:

(آ) با توجه به این مشاهدات، باور سروش نسبت به محیط را برورسانی کرده و توزیع پسین روی متغیرهای مسئله مدلسازی یعنی $p(\beta | r_1, \alpha, \epsilon, \omega)$ و $p(\lambda | t_1, \sigma, \eta)$ را به دست آورید. مقادیر پارامترهای توزیع های پسین یعنی $\epsilon', \omega', \sigma', \eta'$ را محاسبه کنید. (راهنمایی: توزیع های پیشین انتخاب شده از نوع conjugate prior بوده و جنس توزیع پسین آن ها هم مانند توزیع پیشین خواهد بود.)

پاسخ:

$$\begin{aligned} p(\beta | r_1, \alpha, \epsilon, \omega) &\propto p(r_1 | \beta, \alpha, \epsilon, \omega) p(\beta | \alpha, \epsilon, \omega) \\ &= p(r_1 | \beta, \alpha) p(\beta | \epsilon, \omega) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} r_1^{\alpha-1} e^{-\beta r_1} \frac{\omega^\epsilon}{\Gamma(\epsilon)} \beta^{\epsilon-1} e^{-\omega \beta} \\ &\propto \beta^{\alpha+\epsilon-1} e^{-\beta(r_1+\omega)} \end{aligned} \quad (3)$$

$$\rightarrow \beta' \sim \text{Gamma}(\epsilon', \omega') = \text{Gamma}(\epsilon + \alpha, \omega + r_1)$$

$$\begin{aligned} p(\lambda | t_1, \sigma, \eta) &\propto p(t_1 | \lambda, \sigma, \eta) p(\lambda | \sigma, \eta) \\ &= p(t_1 | \lambda) p(\lambda | \sigma, \eta) \\ &= \lambda e^{-\lambda t_1} \frac{\eta^\sigma}{\Gamma(\sigma)} \lambda^{\sigma-1} e^{-\eta \lambda} \\ &\propto \lambda^{\sigma+1-1} e^{-\lambda(t_1+\eta)} \end{aligned} \quad (4)$$

$$\rightarrow \lambda' \sim \text{Gamma}(\sigma', \eta') = \text{Gamma}(\sigma + 1, \eta + t_1)$$

دلیل استفاده از نماد \propto در روابط بالا، وجودی خاصیت conjugate prior بود که روابط را ساده کرده و صرفاً با در نظر گرفتن متغیر اصلی توزیع پسین، می‌توان پارامترهای آن را با انطباق محاسبه کرد.

(ب) با توجه به باور جدیدی که سروش نسبت به مدت زمان بازی به دست آورده است، می‌خواهد برای ادامه یا توقف بازی تصمیم بگیرد. او به دنبال محاسبه‌ی توزیع پسین predictive $p(t_2 | t_1)$ است. به او در محاسبه‌ی این احتمال کمک کرده و نشان دهید این مقدار از توزیع زیر پیروی می‌کند.

$$t_2 \sim Lomax(\sigma', \eta') \rightarrow p(t_2 | \sigma', \eta') = \frac{\sigma'}{\eta'} \left(1 + \frac{t_2}{\eta'}\right)^{-(\sigma'+1)} \quad (5)$$

(راهنمایی: فرض کنید σ' عددی طبیعی است، سپس برای محاسبه‌ی انتگرال، از تکنیک جزیج به صورت بازگشتی استفاده نمایید.)

پاسخ:

$$p(t_2 | t_1) = \int p(t_2, \lambda | t_1) d\lambda = \int p(t_2 | \lambda, t_1) p(\lambda | t_1) d\lambda = \int p(t_2 | \lambda) p(\lambda | t_1) d\lambda \quad (6)$$

در عبارت بالا منظور از $p(\lambda | t_1)$ توزیع پسین روی متغیر λ بعد از مشاهده t_1 است که در قسمت قبل محاسبه شد (در واقع صحیح‌تر است بجای نمادگذاری $p(t_2 | t_1)$ از نماد گذاری کامل‌تر $p(t_2 | t_1, \sigma, \eta)$ استفاده می‌شد).

$$\begin{aligned} p(t_2 | t_1) &= \int p(t_2 | \lambda) p(\lambda | t_1) d\lambda \\ &= \int p(t_2 | \lambda) p(\lambda | t_1, \sigma, \eta) d\lambda \\ &= \int p(t_2 | \lambda) p(\lambda | \sigma', \eta') d\lambda \\ &= \int \lambda e^{-\lambda t_2} \frac{\eta'^{\sigma'}}{\Gamma(\sigma')} \lambda^{\sigma'-1} e^{-\eta' \lambda} d\lambda \\ &= \frac{\eta'^{\sigma'}}{\Gamma(\sigma')} \int \lambda^{\sigma'} e^{-\lambda(\eta' + t_2)} d\lambda \\ &= \frac{\eta'^{\sigma'}}{\Gamma(\sigma')} \mathbf{I}_\lambda(\sigma', \eta') \end{aligned} \quad (7)$$

حال با استفاده از راهنمایی سوال، به محاسبه انتگرال با تکنیک جزیج می‌پردازیم:

$$\begin{aligned} \mathbf{I}_\lambda(\sigma', \eta') &= \int_0^\infty \lambda^{\sigma'} e^{-\lambda(\eta' + t_2)} d\lambda \\ &= \frac{\lambda^{\sigma'} e^{-\lambda(\eta' + t_2)}}{-(\eta' + t_2)} \Big|_0^\infty + \frac{\sigma'}{\eta' + t_2} \int_0^\infty \lambda^{\sigma'-1} e^{-\lambda(\eta' + t_2)} d\lambda \\ &= 0 + \frac{\sigma'}{\eta' + t_2} \int_0^\infty \lambda^{\sigma'-1} e^{-\lambda(\eta' + t_2)} d\lambda \end{aligned} \quad (8)$$

همانطور که مشاهده می‌شود، در آخرین انتگرال توان λ یک عدد کاهش یافته است. با ادامه استفاده از تکنیک جزیج، در آخرین انتگرال به

$$\int_0^\infty \lambda^{\sigma'-1} e^{-\lambda(\eta' + t_2)} d\lambda = \frac{1}{\eta' + t_2} \quad (9)$$

می‌رسیم که با جایگذاری در رابطه بازگشتی بالا خواهیم داشت:

$$\mathbf{I}_\lambda(\sigma', \eta') = \frac{\sigma'}{\eta' + t_2} \cdot \frac{\sigma' - 1}{\eta' + t_2} \cdots \frac{1}{\eta' + t_2} = \frac{\sigma'!}{(\eta' + t_2)^{\sigma'+1}} \quad (10)$$

در نهایت با جایگذاری بالا در توزیع predictive برای متغیر t_2 خواهیم داشت:

$$\begin{aligned}
 p(t_2|\sigma', \eta') &= \frac{\eta'^{\sigma'} \sigma'!}{\Gamma(\sigma') (\eta' + t_2)^{\sigma'+1}} \\
 &= \frac{\eta'^{\sigma'} \sigma'!}{(\sigma' - 1)! (\eta' + t_2)^{\sigma'+1}} \\
 &= \frac{\sigma'}{\eta'} \frac{\eta'^{\sigma'+1}}{(\eta' + t_2)^{\sigma'+1}} \\
 &= \frac{\sigma'}{\eta'} \left(1 + \frac{t_2}{\eta'}\right)^{-(\sigma'+1)} \\
 &\rightarrow t_2 \sim Lomax(\sigma', \eta')
 \end{aligned} \tag{۱۱}$$

(ج) بعد از چند مرتبه بازی کردن که باور سروش از مدل محیط دقیق تر شد، اکنون فکر دیگری ذهن سروش را درگیر کرده است. او که درآمد ناشی از بازی کردن را معقول دریافته و به نوعی معتاد بازی شده است، حالا در یک دوراهی جدیدی قرار گرفته است. او می تواند برای کسب درآمد، به ادامه این بازی تا زمان دلخواه ادامه دهد و یا به کار کارمندی خود با درآمد ثابت ساعتی K دلار بازگردد. به سروش در اتخاذ تصمیم بهینه کمک کنید و تحلیل خود را در سه سناریو ریسک گریز (در نظر گرفتن احتمال بدترین رخدادها)، ریسک خنثی (در نظر گرفتن میانگین) و ریسک پذیر (در نظر گرفتن احتمال بهترین رخدادها) ارائه دهید.

پاسخ:

سناریو ریسک پذیر: با توجه به توزیع های مربوط به پاداش و مدت زمان دریافت آن، در بهترین حالت (با احتمال بزرگتر از صفر) می توان پاداش بزرگی را در مدت زمان کم دریافت کرد و در واقع نسبت این دو می تواند به بینهایت میل کند. فلذا در بهترین حالت، پاداش دریافتی بسیار بیشتر از K خواهد بود و اگر سروش فرد ریسک پذیری باشد می تواند به بازی ادامه دهد.

سناریو ریسک گریز: با توجه به توزیع های مربوط به پاداش و مدت زمان دریافت آن، در بدترین حالت (با احتمال بزرگتر از صفر) ممکن است پاداش کمی در مدت زمان طولانی دریافت شود و در واقع نسبت این دو می تواند به صفر میل کند. فلذا در بدترین حالت، پاداش دریافتی بسیار کمتر از K خواهد بود و اگر سروش فرد ریسک گریزی باشد به بازی ادامه نمی دهد.

سناریو ریسک خنثی: پاسخ این قسمت، با مفروضات متفاوت، می تواند تغییر کند (به راه حل های متفاوت، بسته به مفروضات در نظر گرفته شده نمره تخصیص داده خواهد شد). برای محاسبه نرخ درآمد با تقریب مرتبه دوم تیلور خواهیم داشت:

$$\mathbb{E}\left[\frac{r}{t}\right] = \frac{\mathbb{E}[r]}{\mathbb{E}[t]} \left(1 - \frac{Cov(r, t)}{\mathbb{E}[r]\mathbb{E}[t]} + \frac{Var(t)}{(\mathbb{E}[t])^2}\right) \tag{۱۲}$$

در ابتدای بازی، r و t به ترتیب از توزیع های پسین گاما و نمایی پیروی می کنند، با فرض استقلال این دو متغیر خواهیم داشت:

$$\mathbb{E}\left[\frac{r}{t}\right] = \frac{\frac{\alpha}{\beta}}{\frac{1}{\lambda}} \left(1 - 0 + \frac{\frac{1}{\lambda^2}}{(\frac{1}{\lambda})^2}\right) = 2 \frac{\lambda \alpha}{\beta} \tag{۱۳}$$

پس اگر K از $2 \frac{\lambda \alpha}{\beta}$ کوچکتر باشد، بهتر است بازی کند و در غیر این صورت بهتر است به بازی ادامه ندهد. توجه شود که با تقریب تیلور مرتبه اول به عبارت زیر خواهیم رسید (که مورد پذیرش است):

$$\mathbb{E}\left[\frac{r}{t}\right] = \frac{\mathbb{E}[r]}{\mathbb{E}[t]} = \frac{\lambda \alpha}{\beta}. \tag{۱۴}$$

حال اگر بعد از چند مرتبه بازی کردن برآورد ریسک انجام شود، بهتر است از اطلاعات جمع آوری شده استفاده شده و از توزیع های predictive r و t استفاده شود. در قسمت قبل نشان دادیم t از $Lomax(\sigma', \eta')$ پیروی می کند که میانگین آن برابر است با $\frac{\eta'}{\sigma'-1}$. برای متغیر r توزیع predictive محاسبه نشد، ولی می توان نشان داد (این قسمت ها برای مطالعه بیشتر است و اگر کسی در پاسخ خود اشاره کرده باشد نمره اضافه دریافت خواهد کرد) از توزیع generalized beta prime پیروی می کند:

$$\beta'(r; \alpha, \epsilon', 1, \omega') = \int_0^\infty Gamma(\beta|\epsilon', \omega') Gamma(r|\alpha, \beta) d\beta \tag{۱۵}$$

که میانگین آن $\frac{\alpha \omega'}{\epsilon'-1}$ است. در نتیجه (برای تقریب مرتبه اول تیلور) خواهیم داشت:

$$\mathbb{E}\left[\frac{r}{t}\right] = \frac{\frac{\alpha \omega'}{\epsilon'-1}}{\frac{\eta'}{\sigma'-1}} = \frac{\alpha \omega' (\sigma' - 1)}{\eta' (\epsilon' - 1)}. \tag{۱۶}$$

سوال ۳: (نظری) بررسی روش گرادیان سیاست در رویکرد soft optimality (۲۰ نمره)

در این مساله میخواهیم به بررسی روش گرادیان سیاست تحت استنتاج تقریبی، برای مساله‌ی کنترل با رویکرد soft optimality پرداخته و با روش soft Q-learning مقایسه کنیم. به این منظور به سوالات زیر پاسخ دهید:

(آ) همانطور که در کلاس بررسی شد، برای استنتاج مساله‌ی soft optimality با رویکرد variational، برای درستی‌مندی مشاهدات $O_{1:T}$ کران پایین احتمالاتی به صورت زیر به دست آمد:

$$\log p(O_{1:T}) \geq \sum_t \mathbb{E}_{(s_t, a_t) \sim q} [r(s_t, a_t)] + \mathbb{E}_{s_t \sim q(s_t)} [H(q(a_t|s_t))] \quad (۱۷)$$

این کران پایین را به صورت یک رابطه بر اساس D_{kl} بازنویسی کنید. سپس با استفاده از خواص D_{kl} نشان دهید که برای بهینه‌سازی این کران پایین، $q(a_t|s_t)$ باید به فرم زیر باشد:

$$q(a_t|s_t) = \exp(Q(s_t, a_t) - V(s_t)) \quad (۱۸)$$

پاسخ:

$$\begin{aligned} \log p(O_{1:T}) &\geq \sum_t \mathbb{E}_{(s_t, a_t) \sim q} [r(s_t, a_t)] + \mathbb{E}_{s_t \sim q(s_t)} [H(q(a_t|s_t))] \\ &= \sum_t \mathbb{E}_{(s_t, a_t) \sim q} [r(s_t, a_t)] - \mathbb{E}_{s_t \sim q(s_t)} \mathbb{E}_{a_t \sim q(a_t|s_t)} [\log(q(a_t|s_t))] \\ &= \sum_t \mathbb{E}_{s_t \sim q(s_t)} \mathbb{E}_{a_t \sim q(a_t|s_t)} [r(s_t, a_t) - \log(q(a_t|s_t))] \\ &= \sum_t \mathbb{E}_{s_t \sim q(s_t)} \mathbb{E}_{a_t \sim q(a_t|s_t)} [\log(\exp(r(s_t, a_t))) - \log(q(a_t|s_t))] \\ &= \sum_t \mathbb{E}_{s_t \sim q(s_t)} \mathbb{E}_{a_t \sim q(a_t|s_t)} [\log(\exp(r(s_t, a_t))) - \log \int \exp(r(s_t, a_t)) da_t \\ &\quad + \log \int \exp(r(s_t, a_t)) da_t - \log(q(a_t|s_t))] \\ &= \sum_t \mathbb{E}_{s_t \sim q(s_t)} \mathbb{E}_{a_t \sim q(a_t|s_t)} [\log(\exp(Q(s_t, a_t) - V(s_t))) - \log(q(a_t|s_t))] \\ &\quad + \mathbb{E}_{s_t \sim q(s_t)} [V(s_t)] \\ &= - \sum_t \mathbb{E}_{s_t \sim q(s_t)} [D_{kl}(q(a_t|s_t) || \exp(Q(s_t, a_t) - V(s_t)))] + \mathbb{E}_{s_t \sim q(s_t)} [V(s_t)] \end{aligned} \quad (۱۹)$$

برای انتخاب بهینه‌ی $q(a_t|s_t)$ در رابطه‌ی به دست آمده، عبارت دوم نقشی ندارد. از آن‌جا که D_{kl} یک عبارت نامنفی است، برای بیشینه کردن کران پایین لازم است D_{kl} صفر باشد که با توجه به خواص آن، این حالت زمانی اتفاق می‌افتد که دو توزیع برابر باشند، فلذا:

$$q(a_t|s_t) = \exp(Q(s_t, a_t) - V(s_t)) \quad (۲۰)$$

(ب) حال برای $q(s_t, a_t)$ فرم پارامتری زیر را در نظر بگیرید

$$\pi_\theta(s_t, a_t) = \pi_\theta(a_t|s_t) \pi_\theta(s_t) \quad (۲۱)$$

کران پایین درست‌نمایی را به عنوان تابع هدف در نظر بگیرید. مانند روش گرادیان سیاست از این تابع هدف نسبت به پارامتر θ مشتق بگیرید و نشان دهید این گرادیان را میتوان به صورت زیر تقریب زد:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \sum_t \nabla_\theta \log \pi(a_t|s_t) \left(\sum_{t'=t}^T [r(s(t'), a(t')) - \log \pi(a_{t'}, s_{t'})] - 1 \right) \quad (۲۲)$$

پاسخ:

به دلیل اطلاع دیر هنگام (بعد از ددلاین) از ناقص بودن مفروضات و وجود تایپو و اشکال در روابط سوال، این قسمت و دو قسمت بعدی حذف شده و نمره آن برای تمام دانشجویان لحاظ می‌شود.

(ج) با بازنویسی رابطه‌ی به دست آمده برای گرادیان سیاست در قسمت ب، و جایگذاری تابع سیاست داده‌شده در قسمت الف در آن، و با کمک خواص گرادیان سیاست نشان دهید رابطه‌ی زیر برقرار است:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_i \sum_{t=1}^T (\nabla_{\theta} Q(a_t|s_t) - \nabla_{\theta} V(s_t)) (r(s_t, a_t) + Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) + V(s_t)) \quad (23)$$

(د) با استفاده از خواص گرادیان سیاست، رابطه‌ی به دست آمده در قسمت قبل را یک مرحله ساده‌تر کنید. سپس گرادیان تابع هدف soft Q-learning را برای N نمونه‌ی داده و پنجره‌ی زمانی به طول T بازنویسی کنید و تا جای ممکن این عبارت را شبیه به عبارت به دست آمده برای گرادیان سیاست بازنویسی کنید. در نهایت شباهت‌ها و تفاوت‌های این دو عبارت و مزایای احتمالی هر کدام نسبت به دیگری را بنویسید.

سوال ۴: (عملی) پیاده‌سازی Monte Carlo Tree Search (۴۵ نمره)

هدف این تمرین پیاده‌سازی الگوریتم Monte Carlo Tree Search و اجرای این الگوریتم روی محیط CartPole از کتابخانه‌ی gym است. با کمک نوت‌بوک MCTS.ipynb این الگوریتم را پیاده‌سازی کرده و روی محیط Cartpole اجرا کنید.

سوال ۵: (عملی) پیاده‌سازی Multi-Armed Bandit (۳۰ نمره)

هدف این تمرین پیاده‌سازی الگوریتم Thompson Sampling و اجرای این روش با در نظر گرفتن یک توزیع پیشین گاوسی است. با اجرای مراحل بیان شده در نوت‌بوک ThompsonSampling.ipynb به سوالات گفته شده پاسخ دهید و نتایج را تحلیل کنید.