به نام خدا

دانشکده‌ی مهندسی کامپیوتر

یادگیری تقویتی

نیم‌سال بهار ۴۰۲-۱۴۰۱

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 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 می توانید 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 برطرف‌کردن عدم‌قطعیت محیط و یا exploration استفاده کرد؟

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

سروش یکی از دانشجویان فعال درس ۴۰۹۵۷ است. اخیرا یکی از دوستان او به نام روزبه، بازی خطرناک ولی وسوسه‌انگیزی را به او معرفی کرده است. این بازی به این صورت است که با پرداخت ۱۰ دلار، وارد یک اتاق بزرگ می‌شوید 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 تصادفی‌ای دریافت خواهید کرد که کران بالای آن بینهایت است! اما نکته غم‌انگیز 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 ویژگی‌های خود اتاق بستگی دارد و تجربه‌های قبلی شخص از جست‌وجو در اتاق باعث تقویت مهارت او در جست‌وجو نخواهد شد!
سروش که دانشجوی باهوشی است اقدام به مدلسازی مسئله می‌کند. او مسئله‌ی شرکت کردن متوالی در بازی را به صورت یک مسئله‌ی تصمیم‌گیری دنباله‌ای در نظر می‌گیرید 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 نقطه‌ی تصمیم گیری متوالی توزیع نمایی در نظر می‌گیرد:

$$\begin{aligned} r \sim Gamma(\alpha, \beta) \rightarrow p(r \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r} \\ t \sim Exp(\lambda) \rightarrow p(t \mid \lambda) = \lambda e^{-\lambda t} \end{aligned} \quad (۱)$$

سروش که دانشجوی محتاطی هم هست، سعی در جمع‌آوری اطلاعات موجود کرده و ابتدا از تجربه‌ی خود روزبه می‌پرسد. روزبه در پاسخ می‌گوید من دو بار متوالی در بازی شرکت کردم و به ترتیب ۵ ساعت و ۱۵ ساعت در اتاق گیر کردم ولی پاداش‌هایی به اندازه ۲۰۰ و ۱۰۰ دلار دریافت کرده‌ام. سروش با توجه به این اطلاعات و با استفاده از تکنیک بیشینه درست‌نمایی، یک توزیع پیشین برای متغیر مجهول مدل پاداش (برای سادگی α را مشخص و فقط β را مجهول در نظر می‌گیریم) و مدل زمان به صورت زیر به دست می‌آورد:

$$\beta \sim Gamma(\epsilon, \omega)$$
$$\lambda \sim Gamma(\sigma, \eta) \quad (۲)$$

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

(آ) با توجه به این مشاهدات، باور سروش نسبت به محیط را بروزرسانی 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 توزیع‌های پسین یعنی $\epsilon', \omega', \sigma', \eta'$ را محاسبه کنید.
(راهنمایی: توزیع‌های پیشین انتخاب شده از نوع conjugate prior بوده و جنس توزیع پسین آن‌ها هم مانند توزیع پیشین خواهد بود.)

(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 $p(t_2 \mid t_1)$ است. به او در محاسبه‌ی این احتمال کمک کرده و نشان دهید این مقدار از توزیع زیر پیروی می‌کند.

$$t_2 \sim Lomax(\sigma', \eta') \rightarrow p(t_2 \mid \sigma', \eta') = \frac{\sigma'}{\eta'} \left(1 + \frac{t_2}{\eta'}\right)^{-(\sigma'+1)} \quad (۳)$$

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

(ج) بعد از چند مرتبه بازی کردن که باور سروش از مدل محیط دقیق‌تر 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 می‌تواند 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 ریسک‌گریز (در نظر گرفتن احتمال بدترین رخدادها)، ریسک‌خنثی (در نظر گرفتن میانگین) و ریسک‌پذیر (در نظر گرفتن احتمال بهترین رخدادها) ارائه دهید.

سوال ۳: (نظری) بررسی روش گرادیان سیاست در رویکرد soft optimality (۳۰ 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 درستنمایی مشاهدات $O_{1:T}$ کران پایین احتمالاتی به صورت زیر به دست آمد:

$$\log p(O_{1:T}) \geq \sum_t \mathbb{E}_{(s_t,a_t)\sim q}[r(s_t, a_t)] + \mathbb{E}_{s_t \sim q(s_t)}[q(a_t, s_t)] \quad (۴)$$

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:

$$q(a_t|s_t) = exp(Q(s_t, a_t) - V(s_t)) \quad (۵)$$

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

$$\pi_\theta(s_t, a_t) = \pi_\theta(a_t|s_t)\pi_\theta(s_t) \quad (۶)$$

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:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \sum_t \nabla_\theta \log \pi(a_t|s_t) \big(\sum_{t'=t}^{T} [r(s(t'), a(t')) - \log \pi(a_{t'}, s_{t'})] - 1 \big) \quad (۷)$$

(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:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \sum_{t=1}^{T} \big(\nabla_\theta Q(a_t|s_t) - \nabla_\theta V(s_t)\big)\big(r(s_t, a_t) + Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) + V(s_t)\big) \quad (۸)$$

(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.