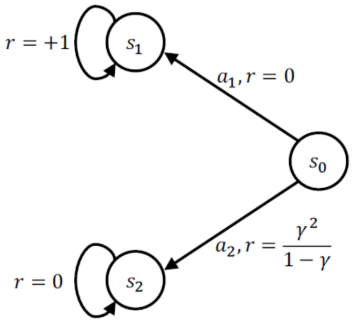
یادگیری ژرف
نیم‌سال اول ۰۱-۰۰
مدرس: حمید بیگی
دانشکده‌ی مهندسی کامپیوتر

تمرین سری ششم | یادگیری ژرف | زمان تحویل: ۸ بهمن ماه

نکات زیر را رعایت کنید:
فایل گزارش را به همراه تمامی کدها در یک فایل فشرده و با عنوان #HW6_STD در سایت Quera بارگذاری نمایید.
سوالات خود را از طریق Quera مطرح کنید.

مسئله‌ی ۱. (۱۰ نمره) حد همگرایی در Value Iteration

زنجیره مارکوف زیر را در نظر بگیرید. ارزش اولیه تمام حالت‌ها را صفر فرض کنید. برای $0 < \gamma < 1$ به سوالات زیر پاسخ دهید.



آ) (۱ نمره) عمل بهینه در زمان $t = 0$ در s_0 کدام است؟ توضیح دهید.

ب) (۶ نمره) نشان دهید که الگوریتم value iteration پس از مرحله n^* برای ارزش s_0 همگرا می‌شود؛ به طوری که n^* در رابطه زیر صدق می‌کند:

$$n^* \geqslant \frac{log(1-\gamma)}{log\gamma}$$

پ) (۳ نمره) با این فرض که اگر تغییرات در ارزش‌ها کمتر از ترشولد θ باشد الگوریتم همگرا می‌شود، حد بالایی برای n^* برحسب θ پیدا کنید. با این حساب، برای یک γ خاص، کمترین مقدار θ چقدر باشد تا در سریعترین زمان ممکن همگرایی صورت بگیرد؟

مسئله‌ی ۲. (۱۰ نمره امتیازی) گرادیان در Multi-armed Bandit

برای حل مسئله Bandit با k بازو می‌توان به طور مستقیم و بدون واسطه‌گری تابع ارزش نیز احتمال انتخاب کنش‌ها را مدل‌سازی کرد. اگر $H_t(a)$ میزان تمایل به انتخاب کنش a در زمان t را نشان دهد، می‌توان سیاست را به صورت زیر محاسبه کرد:

$$\pi_t(a) := \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}}$$

می‌توان توزیع مذکور را با بیشینه‌سازی $\mathbb{E}[R_t] = \sum_x \pi_t(x) q^*(x)$ آموزش داد. R_t پاداش لحظه ای حاصل از انجام A_t است و داریم: $q^*(a) := \mathbb{E}[R_t | A_t = a]$.

آ) (۳ نمره) نشان دهید که

$$\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x)(\mathbb{I}[a = x] - \pi_t(a))$$

ب) (۲ نمره) $\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$ را برحسب $\frac{\partial \pi_t(x)}{\partial H_t(a)}$ بنویسید.

پ) (۵ نمره) نشان دهید که $H_t(a)$ با رابطه‌ی زیر بروزرسانی می‌شود (منظور از $\bar{R}_t$ میانگین پاداش‌ها از لحظه‌ی اول تا t و α نرخ یادگیری صعود در امتداد گرادیان است).

$$H_{t+1}(a) \leftarrow H_t(a) + \alpha(R_t - \bar{R}_t)(\mathbb{I}[a = A_t] - \pi_t(a))$$

مسئله‌ی ۳. (۱۲ نمره) الگوریتم‌های یادگیری ارزش حالات

در این سوال به دنبال بررسی عملکرد روش‌های تخمین ارزش حالات با دو الگوریتم temporal difference و monte carlo هستیم. همچنین در نهایت همگرایی دو الگوریتم Q-learning و temporal difference را بررسی می‌کنیم.

آ) (۲ نمره) روش MC برای تخمین ارزش حالات را به صورت مختصر توضیح دهید و نشان دهید تخمین MC از ارزش حالات تخمینی unbiased است.

ب) (۲ نمره) یکی از مشکلات روش MC الزام به پایان رساندن هر episode برای به‌روزرسانی ارزش حالات است. موضوعی که به خصوص در مسائل long horizon چالش برانگیز است. روش TD چگونه این مشکل را بر طرف می‌کند؟ روابط به‌روزرسانی ارزش حالات در روش TD را ذکر کنید.

پ) (۳ نمره) برای درک بهتر تفاوت این دو روش، ارزش حالات مربوط به markov reward process زیر را با توجه به episode های بیان شده با هر دو روش محاسبه کنید. آیا تفاوتی در مقدار محاسبه شده وجود دارد؟ نتیجه را تفسیر کنید.

A 0 B 0 C 0

A 0 B 0

B 0

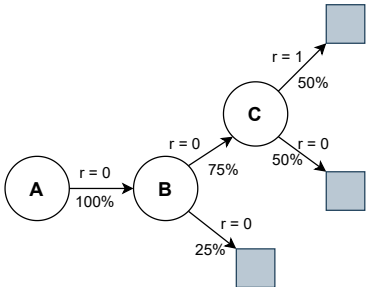
B 0 C 0

C 1

C 1

C 0

C 1



ت) (۵ نمره) روش TD برای یادگیری ارزش حالات از حدس ارزش حالات بعدی استفاده می‌کند. آیا این موضوع همگرایی این الگوریتم را با مشکل مواجه می‌کند؟ اگر جواب مثبت است تحت چه شرایطی همگرایی قابل تضمین نیست؟ توضیح دهید. در مورد الگوریتم Q-learning که در آن کنش‌ها به صورت تصادفی انتخاب می‌شوند چطور؟ آیا همگرایی برای آن الگوریتم تضمین می‌شود؟

مسئله‌ی ۴. (۱۳ نمره) معماری Actor Critic و روش‌های Policy Gradient

گرادیان تابع هدف ساده شده روش‌های policy based در ۱ نشان داده شده است. یکی از ویژگی‌های گرادیان این تابع هدف واریانس بالای آن به دلیل ذات تصادفی تولید یک trajectory و دریافت پاداش است، موضوعی که فرآیند آموزش شبکه عصبی را با چالش همراه می‌کند. یکی از رویکردها برای کاهش واریانس این گرادیان استفاده از مقداری تحت عنوان baseline است. در این سوال ابتدا به بررسی تاثیر یک مقدار ثابت به عنوان baseline پرداخته و سپس با مطالعه دسته مهمی از معماری‌های شبکه‌های یادگیری تقویتی با نام actor critic که به دنبال یادگیری این baseline هستند، با دو روش ارائه شده در این شاخه آشنا می‌شویم.

$$\mathbb{E}_{\tau\sim p(\tau;\theta)}\left[\nabla_\theta \log p(\tau;\theta) r(\tau)\right] \tag{۱}$$

این گرادیان با استفاده از یک مقدار ثابت c تحت عنوان baseline به شکل زیر تغییر می‌کند:

$$\mathbb{E}_{\tau\sim p(\tau;\theta)}\left[\nabla_\theta \log p(\tau;\theta)(r(\tau)-c)\right] \tag{۲}$$

آ) (۲ نمره) گرادیان تابع هدف ۲ را به شکل $\mathbb{E}[f(x)-\phi(x)]+\mathbb{E}[\phi(x)]$ باز نویسی کنید که در آن $\mathbb{E}[f(x)]$ همان عبارت ۱ است. مقدار $\mathbb{E}[\phi(\tau)]$ را نیز محاسبه کنید.

ب) (۳ نمره) ثابت کنید c بهینه که سبب کمینه شدن $\mathrm{Var}[f(\tau)-\phi(\tau)]$ می‌گردد برابر است با:

$$c=\frac{\mathbb{E}\left[\left(\nabla_\theta \log p(\tau;\theta)\right)^2 r(\tau)\right]}{\mathbb{E}\left[\left(\nabla_\theta \log p(\tau;\theta)\right)^2\right]}$$

پ) (۴ نمره) یکی از روش‌های موفق یادگیری تقویتی off policy الگوریتم SAC می‌باشد. در این مقاله تابع هدف یادگیری تقویتی

$$J(\pi)=\sum_{t=0}^{T}\mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right)\right] \tag{۳}$$

به

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \alpha \mathcal{H}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right) \right] \tag{۴}$$

تغییر پیدا کرده است.

اولاً ترم $\mathcal{H}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right)$ چه تاثیری دارد؟

دوماً طبق مقاله بیان کنید که سیاست جدید به دست آمده در گام policy improvement در الگوریتم policy iteration برای این تابع هدف جدید به چه شکل خواهد بود؟

ت) (۳ نمره) یکی از دسته روش‌های بهینه‌سازی روش‌های trust region هستند که در آن‌ها همانند سایر روش‌های بهینه‌سازی تکرار شونده، در هرگام از حدس گام قبل با مکانیزمی به حدس گام بعد می‌رسیم. نکته مهم در این روش‌ها کنترل نزدیکی حدس بعد به حدس قبلی و به اصطلاح باقی ماندن در فضای اطمینان است. TRPO با الهام‌گیری از همین موضوع سعی در کنترل میزان تغییرات سیاست در هرگام با استفاده از تابع هزینه ذیل دارد. این کنترل میزان تغییرات و جلوگیری از تغییرات ناگهانی در سیاست سبب ایجاد پایداری در یادگیری سیاست می‌گردد.

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t \right] \\ \text{subject to} \quad & \hat{\mathbb{E}}_t \left[\mathrm{KL}\left[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t) \right] \right] \leqslant \delta \end{aligned}$$

با این حال این مسئله بهینه‌سازی دارای hard constraint ای است که حل آن را چالش برانگیز می‌کند. به صورت مختصر و کلی بیان کنید روش پیشنهادی PPO چگونه این مشکل را حل می‌کند؟

مسئله‌ی ۵. (۲۵ نمره) REINFORCE

در این سوال الگوریتم REINFORCE را در دو حالت با baseline و بدون آن در محیط `CartPole-v0` پیاده‌سازی و با یک‌دیگر مقایسه می‌کنید. لطفا نوت‌بوک `REINFORCE.ipynb` را مطابق توضیحات و با رعایت فرمت کلی تکمیل کنید. مقادیر پیش‌فرض پارامترها را می‌توانید به تناسب کد خود تغییر دهید. سعی کنید از پردازنده Google Colab برای انجام این تمرین استفاده کنید. آموزش این دو عامل ممکن است چندین دقیقه روی GPU طول بکشد.

مسئله‌ی ۶. (۴۰ نمره) DQN و DRQN

یکی از موارد چالش برانگیز یادگیری تقویتی، فعالیت عامل در فضایی است که حالت محیط به صورت کامل قابل دستیابی نیست. به عبارت دیگر حالت به عنوان ورودیِ هر لحظه در فرمول‌بندی MDP ، به مشاهده در فرمول‌بندی POMDP[۱] تغییر می‌کند. مثالی از این تنظیمات مسئله را می‌توان در ماشین‌های خودران ملاحظه کرد. محیطی که در آن مشاهده تصویر در هر لحظه اطلاعاتی درباره سرعت و جهت حرکت موانع و سایر خودروها نمی‌دهد. یکی از رویکردها برای حل این مشکل استفاده از یک حافظه بازگشتی برای تجمیع اطلاعات در طول زمان و امکان تخمین حالت محیط بر اساس این مشاهدات انباشه است.

در این تمرین به دنبال مقایسه عملکرد روش DQN و روش DRQN هستیم تا تاثیر وجود حافظه بازگشتی را بررسی کنیم. برای این منظور عملکرد این دو روش را بر روی بسطی از بازی آرکید Pong بررسی می‌کنیم. در این

[۱] partially observable markov decision process

بسط در هر گام زمانی مشاهده عامل با احتمال ۰/۵ به طور کامل مخدوش شده و عامل برای اینکه تخمین درستی از جایگاه توپ و جهت حرکت آن در گام‌هایی با مشاهده مخدوش شده داشته باشد باید اطلاعات گام‌های پیشین را به درستی تجمیع کرده باشد. برای درک بهتر این دو روش می‌توانید مقاله‌های آن‌ها را مطالعه مختصری بفرمائید. قالب کلی تکمیل کد در نوت‌بوک مربوط به این سوال قرار داده شده است.