به نام خدا

یادگیری تقویتی

پاییز ۱۴۰۲

مدرس: محمدحسین رهبان

زمان: ۱۰۰ دقیقه

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

امتحان میان‌ترم

تاریخ: ۱۳ اردیبهشت ۱۴۰۳

سوالات (۱۰۰ نمره)

۱. (۲۰ نمره) به سوالات زیر پاسخ دهید.

(آ) فرض کنید در الگوریتم Policy Gradient ساده مانند REINFORCE به جای دنبال کردن سیاست $\pi_\theta(.|s)$ سیاست دیگری را دنبال می‌کنیم که هدفش تقویت اکتشاف (exploration) است. به این منظور با احتمال ϵ یک عمل تصادفی و با احتمال $1-\epsilon$ یک عمل از توزیع سیاست π_θ نمونه‌برداری می‌کنیم. گرادیان مورد استفاده برای به روزرسانی سیاست π_θ به چه صورت خواهد بود؟ راه حل را به صورت کامل شرح دهید. توجه داشته باشید که در گرادیان سیاست داریم:

$$\nabla_\theta J(\theta) \approx \sum_i \left(\sum_t \nabla_\theta \log \pi_\theta(a_t^i|s_t^i)\right)\left(\sum_t r(s_t^i, a_t^i)\right) \qquad (1)$$

پاسخ: کافی است قرار دهیم: $\pi'_\theta(a|s) = \epsilon U(a) + (1-\epsilon)\pi_\theta(a|s)$. لذا $\nabla_\theta \log \pi'_\theta = \frac{(1-\epsilon)\nabla\pi_\theta}{\epsilon U(a)+(1-\epsilon)\pi_\theta}$.
اگر کسی پاسخی با این مضمون نوشته که الگوریتم را به صورت off-policy در نظر گرفته نیز نمره کامل در صورت صحیح بودن پاسخ داده شود. توجه کنید در این حالت بایستی از مفهوم importance sampling استفاده کرده باشد.

(ب) فرض کنید به یک مجموعه از trajectory های نمونه‌برداری شده از سیاست‌های قبلی دسترسی داریم و هدف این است که تغییری در الگوریتم گرادیان سیاست ساده مانند REINFORCE ایجاد کنیم که بتواند از این داده‌ها نیز استفاده نماید. می‌خواهیم در گرادیان تابع هدف به گونه‌ای تغییر ایجاد کنیم که این مسئله محقق شود. چگونه این کار را انجام دهیم؟ راه حل را شرح دهید.
پاسخ: در این حالت باید از روش وزن‌دهی نمونه استفاده کنیم تا وزن گرادیان‌هایی که از trajectoryهای قبلی بدست آمده تنظیم شود:

$$\nabla_\theta J = \mathbb{E}\left(\prod_{t=1}^{T} \frac{\pi_\theta(a_t|s_t)}{\pi_\theta^{\text{old}}(a_t|s_t)}\right)\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=1}^{T} r(s_t, a_t)\right)$$

.

(ج) در اجرای یک روش گرادیان سیاست مشاهده شده است که میزان بازده (return) نوسان‌های زیادی دارد. علت احتمالی این مسئله چیست؟ دو راه حل برای این مشکل به اختصار مطرح کنید.

پاسخ: روش گرادیان سیاست در حالت ایده‌آل باید در هر گام میزان بازده را بهبود دهد. با این حال، در صورتی که تغییرات سیاست در یک گام به روزرسانی بالا باشد، ممکن است به دلیل تفاوت $p_\theta(\tau)$ و $p_{\theta'}(\tau)$ نتوان بیشتر شدن $J(\theta') - J(\theta)$ را تضمین نمود. لذا برای حل این مشکل، از روش‌های trust region می‌توان استفاده نمود. همینطور، یک دلیل محتمل دیگر می‌تواند بالا بودن واریانس تخمین گرادیان سیاست دانست. برای حل این مشکل می‌توان از راهکارهایی مانند اضافه کردن baseline و همینطور استفاده از تخمین علی پاداش باقی‌مانده تا انتهای اپیزود استفاده کرد.

۲. (۲۰ نمره) فرض کنید در یک MDP پاداش، برخلاف مورد متداول بررسی شده در کلاس، یک متغیر تصادفی وابسته به حالت جاری است. (منظور از وابستگی این متغیر تصادفی به حالت جاری این است که در مدل گرافی توصیف کننده MDP، یک یال از متغیر حالت در هر زمان به متغیر پاداش در آن زمان متصل است.) لذا معادله بهینگی Bellman به صورت زیر در آمده است:

$$V^\star(s) = \max_a \mathbb{E}_{s',R}\left(R + \gamma V^\star(s')\right). \qquad (۲)$$

ثابت کنید الگوریتم Value Iteration در این حالت، به نقطه ثابت این معادله همگرا می‌شود.

برای اثبات همگرایی این الگوریتم کافی است Lipschitz بودن اپراتور Bellman جدید τ را بررسی نماییم.

$$\tau V_۱(s) - \tau V_۲(s) = \max_a \sum_{s',r} p(s'|s,a)p(r|s)(r+\gamma V_۱(s')) - \max_a \sum_{s',r} p(s'|s,a)p(r|s)(r+\gamma V_۲(s'))$$

. فرض کنید پاسخ بهینه‌سازی اول بالا، عملی مانند $a_۱$ باشد. داریم:

$$\tau V_۱(s) - \tau V_۲(s) \le \sum_{s',r} p(s'|s,a_۱)p(r|s)(r+\gamma V_۱(s')) - \sum_{s',r} p(s'|s,a_۱)p(r|s)(r+\gamma V_۲(s')).$$

لذا بدست می‌آید:

$$\tau V_۱(s) - \tau V_۲(s) \le \sum_{s',r} \gamma p(s'|s,a_۱)p(r|s)(V_۱(s') - V_۲(s')) \le \gamma \max_{s'} V_۱(s') - V_۲(s') \sum_{s',r} p(s'|s,a)p(r|s).$$

لذا:

$$\tau V_۱(s) - \tau V_۲(s) \le \gamma \max_{s'} V_۱(s') - V_۲(s').$$

به این ترتیب اثبات می‌شود این اپراتور حالت contraction mapping دارد و همگرایی تضمین می‌شود.

۳. (۱۵ نمره) رضا هرروز سه حالت دارد، یا آسوده است، یا نگران است و یا خسته است. او هر روز دو کار می‌تواند انجام دهد، یا درس می‌خواند یا درس نمی‌خواند. اطلاعات مربوط به پاداش‌ها و احتمالات انتقال هر اکشن در شکل آمده است. $\gamma = ۰/۸$.

Reward	Probability	Next state	Action	First state
۴	۰/۸	Tired	Study	Worried
۶	۰/۲	Relaxed	Study	Worried
−۲	۱	Worried	study Don't	Worried
۴	۰/۷	Tired	Study	Tired
۳	۰/۳	Relaxed	Study	Tired
۰	۰/۲	Tired	Don't study	Tired
−۲	۰/۳	Worried	Don't study	Tired
۵	۰/۵	Relaxed	Don't study	Tired
۱۰	۰/۴	Relaxed	Study	Relaxed
۸	۰/۶	Tired	Study	Relaxed
۲	۰/۵	Relaxed	Don't study	Relaxed
−۸	۰/۵	Worried	Don't study	Relaxed

(آ) ارزش هر حالت را با استفاده از Value Iteration تا سه مرحله حساب کنید.

(ب) ارزش هر اکشن در هر حالت یا همان Q-Vlaue ها راحساب کنید.

(ج) رضا در کدام حالت‌های خود باید درس بخواند؟ در کدام حالت‌ها بهتر است درس نخواند؟ (نیازی به محاسبه تا همگرایی نیست و سه مرحله کافی‌است)

۴. (۱۵ نمره) به سوالات زیر در حوزه Deep Q-Learning پاسخ دهید.

(آ) دو دلیل استفاده از Replay Buffer در این الگوریتم را به اختصار توضیح دهید.
پاسخ: این بافر برای ذخیره‌سازی انتقال بین حالت‌های متوالی قبلی و استفاده از آنها در کمیته‌سازی تابع زیان در گام‌های بعدی استفاده می‌شود. در صورتی که این کار انجام نشود، مدل دچار فراموشی می‌شود. همینطور، این کار باعث data efficient شدن الگوریتم نیز می‌شود، چراکه جمع‌آوری داده‌های قبلی، هزینه محاسباتی قابل توجهی برای عامل داشته است و با ذخیره آنها این هزینه محاسباتی صرفه‌جویی می‌شود.

(ب) چرا در Replay Buffer مقدار target Q-value را نگهداری نمی‌کنیم؟
پاسخ: به این دلیل که target در طول زمان ثابت نیست و در هر لحظه به مقدار Q function که به صورت لگ‌دار در طول زمان Q function اصلی را دنبال می‌کند، وابسته است.

(ج) چرا در محاسبه مقدار target Q-value از میانگین متحرک نمایی (exponential moving average) وزن‌های شبکه Q به صورت زیر استفاده می‌کنیم؟ در صورتی که بخواهیم مقدار τ را بر اساس یک زمان‌بندی در طول آموزش تغییر دهیم، چه زمان‌بندی برای این کار مناسب‌تر است؟
پاسخ: دلیل این مسئله آن است که نمی‌خواهیم اهداف یادگیری بین گام‌های متوالی تغییرات زیادی داشته باشد. در غیر اینصورت بهینه‌سازی تابع زیان همگرا نخواهد شد. برای زمان‌بندی τ بهتر است این مقدار از یک شروع شده و به تدریج به صفر میل داده شود.

$$\theta' = \tau\theta + (1-\tau)\theta'. \tag{3}$$

۵. (۱۵ نمره) رابطه گرادیان سیاست را در نظر داشته و به سوالات زیر پاسخ دهید:

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})\right)\left(\sum_{t=1}^{T} r(s_{i,t},a_{i,t})\right) \tag{4}$$

(آ) با حذف جملات غیر علی از رابطه بالا به عبارت ساده‌تری برسید.

$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})\sum_{t'=t}^{T} r(s_{i,t},a_{i,t})\right).$$

(ب) نشان دهید رابطه بدست آمده از قسمت قبل، چگونه باعث کاهش واریانس تخمین گرادیان سیاست می‌شود.

(ج) نشان دهید عبارات به دست آمده در قسمت اول، به تخمین گرادیان سیاست بایاس اضافه نمی‌کند.

۶. (۱۵ نمره) در رابطه با روش‌های Temporal Difference و Monte Carlo به سولات زیر پاسخ دهید.

(آ) در چه شرایطی نمی‌توان از روش‌های Monte Carlo استفاده نمود؟
در شرایطی که محیط episodic نباشد.

(ب) دو روش را از حیث Bias/Variance Trade-off مقایسه نمایید.
MC تخمین بدون سوگیری از $v_\pi(S_t)$ به دست می‌دهد در حالیکه تخمین TD دارای سوگیریست مگر آنکه $\mathbf{E}[v_t(S_{t+1})|S_{t+1}] = v_\pi(S_{t+1})$. ؛ اما تخمین TD واریانس کمتری دارد زیرا بر مبنای مشاهدات بیشتریست.

(ج) یک محیط با دو حالت A و B را در نظر بگیرید. دو دنباله‌ی نمونه از کنش‌ها و پاداش‌ها به صورت زیر داده شده است.

$$A \xrightarrow{۳} A \xrightarrow{۲} B \xrightarrow{۱} \text{Terminate}$$

$$A \xrightarrow{۳} A \xrightarrow{۳} A \xrightarrow{۲} B \xrightarrow{۱} \text{Terminate}$$

مقدار تابع Value برای حالات A و B را با روش TD دو قدمه و Monte Carlo به طور جداگانه به دست آورید. مقدار اولیه‌ی تابع را ۰ و γ را برابر ۱ در نظر بگیرید.

- MC:

$$G_{۱,۳} = ۱, G_{۱,۲} = ۳, G_{۱,۱} = ۶$$

$$G_{۲,۴} = ۳, G_{۲,۳} = ۳, G_{۲,۲} = ۶, G_{۲,۱} = ۹$$

$$G(A) = ۳ + ۶ + ۳ + ۶ + ۹ = ۲۷, N(A) = ۵ \rightarrow V(A) = ۵/۴$$

$$G(B) = ۱ + ۱, N(B) = ۲ \rightarrow V(B) = ۱$$

- TD-2: مقدار α به دانشجویان گفته نشده و بسته به مقداری که خودشان در نظر گرفته‌اند پاسخ بررسی می‌شود.

$$V(A) = \alpha(۳ + ۲)$$

$$V(A) = V(A) + \alpha(۲ + ۱ - V(A)) = \alpha(۳ + ۵(۱ - \alpha))$$

$$V(B) = \alpha$$

$$V(A) = \alpha(۶ + ۳(۱ - \alpha) + ۵(۱ - \alpha)^{۲})$$

$$V(A) = \alpha(۵ + ۶(۱ - \alpha) + ۳(۱ - \alpha)^{۲} + ۵(۱ - \alpha)^{۳})$$

$$V(A) = \alpha(۳ + ۵(۱ - \alpha) + ۶(۱ - \alpha)^{۲} + ۳(۱ - a\alpha)^{۳} + ۵(۱ - \alpha)^{۴})$$

$$V(B) = \alpha + \alpha(۱ - \alpha)$$

برای $\alpha = \frac{۱}{۲}$، $V(A) = ۳/۸۴۳۷۵ = \frac{۱۲۳}{۳۲}$ و $V(B) = ۰/۷۵ = \frac{۳}{۴}$ و برای $\alpha = ۱$، $V(A) = ۳$ و $V(B) = ۱$.