به نام خدا

یادگیری تقویتی

نیم‌سال بهار ۴۰۲ـ۱۴۰۱

اساتید: دکتر رهبان، آقای حسنی

دانشکده‌ی مهندسی کامپیوتر

مدت: ۱۲۰ دقیقه | امتحان میانترم | بارم: ۱۰۰ نمره

نام و نام خانوادگی : شماره دانشجویی:

سوال ۱: MDP and Value Iteration (۲۰ نمره)

محیط gridworld ساده‌ی یک‌بعدی زیر را در نظر بگیرید که در آن agent از خانه Start شروع کرده و در هر کدام از خانه‌های سفید توانایی انتخاب کنش‌های چپ و راست را دارد. در خانه‌های خاکستری رنگ، تنها انتخاب ممکن خروج از بازی است و به هنگام خروج از هر کدام، پاداشی دریافت می‌شود که در خانه مربوطه نوشته شده است. به هنگام ترک خانه‌های سفید هم پاداشی دریافت نمی‌شود.
ضریب تخفیف را $\gamma = 1$ در نظر گرفته و به سوالات زیر پاسخ دهید.

+1		Start			+5

(آ) میزان ارزش بهینه‌ی $V^*(Start)$ چقدر است؟

(ب) با فرض اجرای الگوریتم value iteration و مقدار اولیه $V_0(Start) = 0$، اولین گام k که در آن $V_k(Start)$ غیرصفر می‌شود، چند است؟

(ج) بعد از چند گام k، خواهیم داشت $V_k(Start) = V^*(Start)$؟ در صورتی که این دو مقدار هیچگاه مساوی نخواهند شد، بنویسید هیچوقت.

(د) حال فرض کنید $\gamma = 0.8$، مقدار ارزش بهینه $V^*(Start)$ چقدر است؟

(ه) به ازای چه مقادیری از γ، تصمیم بهینه در نقطه‌ی شروع، حرکت به سمت چپ خواهد بود؟

(و) حال فرض کنید $\gamma = 1$ باشد، ولی محیط به صورت تصادفی بوده و کنش‌های اتخاذ شده در خانه‌های سفید تنها با احتمال $p = 0.8$ منجر به گذار به خانه‌های مجاور می‌شوند. حال مقدار $V^*(Start)$ چیست؟

(ز) در این شرایط، اولین گام k از value iteration که در آن $V_k(Start)$ غیرصفر می‌شود، چند است؟

(ح) بعد از چند گام k، خواهیم داشت $V_k(Start) = V^*(Start)$؟ در صورتی که این دو مقدار هیچگاه مساوی نخواهند شد، بنویسید هیچوقت.

سوال ۲: Causality Trick (۲۰ نمره)

رابطه‌ی گرادیان سیاست را در نظر داشته و به سوالات زیر پاسخ دهید:

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

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

(ب) نشان دهید رابطه‌ی به دست 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.

سوال ۳: Generalized Advantage Estimation (۲۰ نمره)

تابع هدف actor-critic را در نظر گرفته و به سوالات زیر پاسخ دهید:

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})\hat{A}^\pi(s_{i,t},a_{i,t})$$

$$\hat{A}^\pi(s_{i,t},a_{i,t}) = r(s_{i,t},a_{i,t}) + \gamma\hat{V}_\phi^\pi(s_{i,t+1}) - \hat{V}_\phi^\pi(s_{i,t})$$

(آ) با استفاده از مفهوم n-step TD نسخه‌ای تعمیم یافته از advantage، یعنی $\hat{A}_n^\pi(s_t,a_t)$ را بنویسید.

(ب) مزایا و معایب استفاده از $\hat{A}_n^\pi(s_t,a_t)$ بجای $\hat{A}^\pi(s_t,a_t)$ در چیست؟

(ج) نشان دهید با جمع وزن دار $\hat{A}_n^\pi(s_t,a_t)$ ها به ازای تمامی n با وزن‌هایی به صورت $\omega_n = \lambda^{n-1}$ می‌توان به $\hat{A}_{GAE}^\pi(s_t,a_t)$ رسید. رابطه‌ی مربوط به آن را نوشته و نشان دهید آن را می‌توان به صورت جمع وزن‌دار خطاهای TD نوشت:

$$\hat{A}_{GAE}^\pi(s_t,a_t) = \sum_{t'=t}^{\inf}(\gamma\lambda)^{t'-t}\delta_{t'}, \quad where \quad \delta_{t'} = r(s_{t'},a_{t'}) + \gamma\hat{V}_\phi^\pi(s_{t'+1}) - \hat{V}_\phi^\pi(s_{t'})$$

(د) مزایا و معایب استفاده از $\hat{A}_{GAE}^\pi(s_t,a_t)$ بجای $\hat{A}_n^\pi(s_t,a_t)$ چیست؟

سوال ۴: Policy Gradient as Policy Iteration (۲۰ نمره)

تابع هدف مورد استفاده در گرادیان سیاست را در نظر گرفته و به سوالات زیر پاسخ دهید.

$$J(\theta) = E_{\tau\sim p_\theta(\tau)}\left[\sum_t \gamma^t r(s_t,a_t)\right]$$

(آ) نشان دهید اگر با به‌روزرسانی پارامترهای θ و رسیدن به θ'، عبارت زیر برقرار است:

$$J(\theta') - J(\theta) = E_{\tau\sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t r(s_t,a_t)\right]$$

(ب) با توجه به رابطه‌ی قسمت قبل، به صورت مختصر توضیح دهید که چرا این رابطه می‌تواند نشان دهنده‌ی بهبود گرادیان در گام‌های متوالی باشد؟

(ج) آیا از رابطه‌ی نوشته شده در قسمت اول را می‌توان به صورت مستقیم هدف بهینه‌سازی قرار داد؟ اگر مشکلی دارد، ایراد آن را مرتفع نمایید.

سوال ۵: Trust Region and Natural Gradient (۲۰ نمره)

رابطه‌ی بروز رسانی گرادیان سیاست را در نظر گرفته و به سوالات زیر پاسخ دهید.

$$\theta \leftarrow \theta + \alpha\nabla_\theta J(\theta)$$

(آ) نشان دهید رابطه زیر معادل بهینه‌سازی بالاست:

$$\theta \leftarrow \arg\min_{\theta'}(\theta'-\theta)^T\nabla_\theta J(\theta) \quad s.t. \|\theta'-\theta\|^2 \le \epsilon$$

(ب) نشان دهید بعد از بهینه‌سازی رابطه قسمت قبل خواهیم داشت:

$$\theta' = \theta + \sqrt{\frac{\epsilon}{\|\nabla_\theta J(\theta)\|^2}}\nabla_\theta J(\theta)$$

(ج) به هنگام بهینه‌سازی سیاست، چرا استفاده از قید $D_{KL}(\pi'_\theta(a_t|s_t)\|\pi_\theta(a_t|s_t))$ بهتر از قید داده شده در قسمت اول است؟ بر این مبنا رابطه‌ی داده شده در قسمت اول را بازنویسی کنید.

(د) با استفاده از تقریب $D_{KL}(\pi'_\theta(a_t|s_t)\|\pi_\theta(a_t|s_t)) \approx \frac{1}{2}(\theta'-\theta)^T\mathbf{F}(\theta'-\theta)$ نشان دهید رابطه‌ی بروز رسانی تابع هدف قسمت قبل به صورت زیر خواهد بود:

$$\theta' = \theta + \sqrt{\frac{\epsilon}{\nabla_\theta J(\theta)^T\mathbf{F}^{-1}\nabla_\theta J(\theta)}}\mathbf{F}^{-1}\nabla_\theta J(\theta)$$

سوال ۶: Model-Based RL (۲۰ نمره)

در رابطه با روش‌های یادگیری تقویتی مبتنی بر مدل، به سوالات زیر پاسخ کوتاه دهید.

(آ) منظور از برنامه‌ریزی به صورت open-loop و closed-loop چیست و کدام یک ارجحیت دارد؟

(ب) دو مزیت و دو عیب از روش برنامه‌ریزی random shooting را بیان کرده و دو روش برای بهبود آن معرفی کنید.

(ج) روش bootstrap ensemble برای غلبه به کدام مشکل در روش‌های مبتنی بر مدل استفاده می‌شود؟

(د) به اختصار توضیح دهید چگونه می‌توان برای sample-efficient کردن روش‌های model-free، از رویکرد model-based بهره گرفت.