یادگیری تقویتی عمیق

نیم‌سال دوم ۱۴۰۴-۱۴۰۵

دانشکده‌ی مهندسی کامپیوتر

مدرس: دکتر محمدحسین رهبان

میانترم

مدت: ۱۲۰ دقیقه

مسئله‌ی ۱. شکل دهی پاداش مبتنی بر تابع پتانسیل (۲۵ نمره)

شما در حال طراحی یک عامل ناوبری در یک محیط شبکه‌ای هستید. محیط یک جدول ۱۰ × ۱۰ است. عامل از خانهٔ پایین-چپ شروع می‌کند و هدف در خانهٔ بالا-راست قرار دارد. عامل می‌تواند در هر گام یکی از چهار عمل بالا، پایین، چپ یا راست را انتخاب کند. اگر حرکت انتخاب‌شده به خارج از جدول منتهی شود، عامل در همان خانه باقی می‌ماند.

عامل تنها زمانی پاداش مثبت دریافت می‌کند که به هدف برسد:

$$r(s,a,s') = \begin{cases} ۱, & \text{به هدف رسیده باشد,} \\ ۰, & \text{در غیر این صورت.} \end{cases}$$

اپیزود پس از رسیدن به هدف یا پس از ۲۰۰ گام پایان می‌یابد. عامل با الگوریتم Q-learning و سیاست اکتشافی ϵ-حریصانه آموزش داده می‌شود.

الف) مساله پاداش تنگ. توضیح دهید چرا یادگیری در این محیط ممکن است کند باشد. در پاسخ خود به تنک بودن پاداش، تأخیر پاداش، و نحوهٔ انتشار مقدارها در Q-learning اشاره کنید.

ب) شکل دهی پاداش مبتنی بر پتانسیل. برای کمک به یادگیری، می‌خواهیم از شکل‌دهی پاداش مبتنی بر پتانسیل استفاده کنیم. برای یک تابع پتانسیل کراندار

$$\Phi : \mathcal{S} \to \mathbb{R},$$

پاداش جدید به صورت زیر تعریف می‌شود:

$$r_\Phi(s,a,s') = r(s,a,s') + \gamma\Phi(s') - \Phi(s),$$

که در آن $۰ < \gamma < ۱$ ضریب تنزیل است.

ثابت کنید که برای هر سیاست π، مقدار کنشی در محیط شکل‌دهی‌شده برابر است با

$$Q^\pi_\Phi(s,a) = Q^\pi(s,a) - \Phi(s).$$

سپس نتیجه بگیرید که این نوع شکل‌دهی پاداش، مجموعهٔ اعمال بهینه را تغییر نمی‌دهد؛ یعنی

$$\arg\max_a Q^*_\Phi(s,a) = \arg\max_a Q^*(s,a).$$

ج) طراحی تابع پتانسیل. برای محیط شبکه‌ای بالا، یک تابع پتانسیل مناسب پیشنهاد دهید. تابع پیشنهادی شما باید باعث شود عامل در طول مسیر، سیگنال یادگیری مفیدتری دریافت کند، اما سیاست بهینه را تغییر ندهد.

مسئله‌ی ۲. ترفند علیت در گرادیان سیاست (۲۵ نمره)

فرض کنید $\tau \sim p_\theta(\tau)$ یک مسیر تولیدشده توسط سیاست π_θ باشد. تعریف کنید

$$r_t = r(s_t, a_t, s_{t+۱}).$$

ثابت کنید که

$$\mathbb{E}_{\tau\sim p_\theta}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'=1}^{t-1} r_{t'}\right] = 0.$$

نتیجه بگیرید که پاداش‌های گذشته را می‌توان از تخمین‌گر گرادیان سیاست حذف کرد؛ بنابراین بازگشت کامل را می‌توان با بازگشت از آن لحظه به بعد جایگزین کرد.

مسئله‌ی ۳. تحلیل گرادیان ساختار های کنشگر_منتقد (۳۰ نمره)

در صورت لزوم در این سوال فرضیات زیر را برای محاسبه کران ها و تحلیل های خود درنظر بگیرید:

- کران‌های مطلق پاداش در هر گام: $|r_t| \leqslant R_{\max}$ یا $|r(s,a,s')| \leqslant R_{\max}$.
- نُرم تابع امتیاز سیاست کران‌دار است: $\|\nabla_\theta \log \pi_\theta(a|s)\| \leqslant L_\pi$.
- تصادفی‌بودن انتخاب کنش کران‌دار است: $\max_s \mathrm{Var}_{a\sim\pi_\theta(\cdot|s)}\left(A^{\pi_\theta}(s,a)\right) \leqslant \sigma_{\text{adv}}^2$.
- بیشینهٔ خطای تقریب تابع مزیت: $\max_{s,a} |A_\phi(s,a) - A^\pi(s,a)| \leqslant \epsilon_A$.

ساختارهای کنشگر_منتقد، فضای ردیابی در روش‌های مبتنی بر سیاست را تغییر می‌دهند. این روش‌ها به‌جای استفاده از بردار ردیابی بازگشت تجربی چندگامی مونت‌کارلو، یعنی $G_t - V^\pi(s_t)$، از یک تقریب‌گر تابعی پارامتری، محلی و یادگرفته‌شده، یعنی $A_\phi(s_t,a_t)$ با پارامترهای ϕ، استفاده می‌کنند تا تابع مزیت حقیقی $A^\pi(s_t,a_t)$ را ردیابی کند. هدف گرادیان پارامتری g و برآوردگر معماریِ عملی آن، یعنی $\hat{g}_{\text{AC}}$، به‌صورت زیر نگاشت می‌شوند:

$$g \coloneqq \mathbb{E}_\tau\left[\sum_{t=0}^{H-1}\gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t) A^\pi(s_t,a_t)\right]$$
$$\hat{g}_{\text{AC}} \coloneqq \sum_{t=0}^{H-1}\gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t) A_\phi(s_t,a_t)$$

*یادآوری: بسط میانگین مربعات خطا: $\mathbb{E}\left[\|\hat{g}-g\|^2\right] = \mathrm{Bias}^2(\hat{g}) + \mathrm{Var}(\hat{g})$ که در آن $\mathrm{Bias}^2(\hat{g}) = \|\mathbb{E}[\hat{g}] - g\|^2$.

الف) تحلیل بایاس. بایاسی را که به‌طور صریح از طریق این جایگزینی معماری وارد جهت سیاست می‌شود تحلیل کنید. ثابت کنید که: $\mathrm{Bias}^2(\hat{g}_{\text{AC}}) \in \mathcal{O}\left(H^2 L_\pi^2 \epsilon_A^2\right)$.

ب) کران بندی ممان دوم منتقد. برای نامساوی میانی زیر اثباتی ارائه دهید که ممان دوم منتقد یادگرفته‌شده را کران‌بندی می‌کند: $\mathbb{E}\left[(A_\phi(s_t,a_t))^2\right] \leqslant 2\sigma_{\text{adv}}^2 + 2\epsilon_A^2$.

ج) تحلیل واریانس. واریانس برآوردگر کنشگر_منتقد را تحلیل کنید. کران بالای آن را استخراج کنید تا وابستگی دقیق آن به افق زمانی و کران‌های مزیت نشان داده شود: $\mathrm{Var}(\hat{g}_{\text{AC}}) \in \mathcal{O}\left(H^2 L_\pi^2(\sigma_{\text{adv}}^2 + \epsilon_A^2)\right)$.

مسئله‌ی ۴. یادگیری تقویتی مدل محور در محیط های نامانا (۲۰ نمره)

شما در حال طراحی یک عامل ناوبری هوشمند هستید که باید سریع‌ترین مسیر رفت‌وآمد روزانهٔ یک راننده را پیدا کند. عامل از الگوریتم استاندارد Dyna-Q استفاده می‌کند تا دینامیک گذار و «هزینه»ی مسیرهای مختلف را یاد بگیرد. منظور از هزینه، پاداش منفی است که تعداد دقیقه‌های سپری‌شده در ترافیک را نشان می‌دهد. سپس عامل از مدل داخلی خود برای شبیه‌سازی تجربه‌ها و برنامه‌ریزی مسیر بهینه استفاده می‌کند.

در ابتدا، عامل سه مسیر اصلی برای رسیدن به محل کار را بررسی و مدل‌سازی می‌کند:

- مسیر A یا بزرگراه: زمان سفر ۳۰ دقیقه است؛ بنابراین پاداش برابر است با $R_A = -۳۰$.
- مسیر B یا خیابان‌های شهری: زمان سفر ۴۵ دقیقه است؛ بنابراین پاداش برابر است با $R_B = -۴۵$.
- مسیر C یا جادهٔ فرعی خوش‌منظره: زمان سفر ۶۰ دقیقه است؛ بنابراین پاداش برابر است با $R_C = -۶۰$.

عامل به‌درستی مسیر A را به‌عنوان مسیر بهینه برنامه‌ریزی می‌کند و به‌طور مداوم از آن به‌عنوان مسیر روزانه استفاده می‌کند.

۱ عامل استاندارد Dyna-Q

الف) منطقه ساخت و ساز. پس از چند ماه، عملیات ساخت‌وساز گسترده‌ای در مسیر A شروع می‌شود و زمان سفر این مسیر به‌طور دائمی به ۹۰ دقیقه افزایش می‌یابد. بنابراین پاداش مسیر A از -۳۰ به مقدار زیر تغییر می‌کند: $R_A = -۹۰$. عامل استاندارد Dyna-Q نسبت به این تغییر چگونه واکنش نشان می‌دهد؟ آیا در نهایت مسیر جدید بهتر، یعنی مسیر B، را پیدا می‌کند؟ سازوکار به‌روزرسانی مدل را توضیح دهید.

ب) بزرگراه سریع السیر جدید. فرض کنید محیط دوباره به حالت اولیه بازگردانده شده است و عامل با رضایت از مسیر A استفاده می‌کند که پاداش آن برابر است با: $R_A = -۳۰$.

پس از چند ماه، شهر یک توسعهٔ بزرگ در مسیر C را تکمیل می‌کند و زمان سفر این مسیر از ۶۰ دقیقه به تنها ۱۵ دقیقه کاهش می‌یابد. بنابراین پاداش مسیر C به مقدار زیر تغییر می‌کند: $R_C = -۱۵$.

ترافیک و زمان سفر مسیر A کاملاً بدون تغییر باقی می‌ماند. آیا عامل استاندارد Dyna-Q کشف می‌کند که مسیر C اکنون سریع‌ترین مسیر است؟ پاسخ خود را با استفاده از تعارض میان اکتشاف و بهره‌برداری در زمینهٔ برنامه‌ریزی توضیح دهید.

۲ معرفی Dyna-Q+

وقتی محیط تغییر می‌کند، عامل‌های استاندارد یادگیری تقویتی گاهی نمی‌توانند خود را با برخی تغییرات سازگار کنند. برای رفع بخشی از این ضعف، نسخهٔ بهبودیافته‌ای به نام Dyna-Q+ معرفی شده است.

الگوریتم Dyna-Q+ یک ابتکار برای تشویق «کنجکاوی محاسباتی» اضافه می‌کند. این عامل برای هر زوج حالت—عمل یک متغیر τ نگه می‌دارد. این متغیر تعداد گام‌های زمانی سپری‌شده از آخرین باری را نشان می‌دهد که آن زوج حالت—عمل در محیط واقعی امتحان شده است. فرض اصلی این است که هرچه یک مسیر مدت طولانی‌تری امتحان نشده باشد، احتمال این‌که الگوی ترافیک آن تغییر کرده و مدل داخلی عامل نادرست شده باشد، بیشتر است.

در مرحلهٔ برنامه‌ریزی، اگر یک گذار مدل‌شده معمولاً پاداش r تولید کند، عامل به جای آن از پاداش افزودهٔ زیر استفاده می‌کند: $r + \kappa\sqrt{\tau}$, که در آن κ یک ثابت مقیاس‌دهندهٔ کوچک و مثبت است.
الف) از نظر ریاضی و مفهومی توضیح دهید این فرمول پاداش افزوده چگونه بر رفتار عامل در سناریوی ۲ اثر می‌گذارد. آیا این فرمول نتیجهٔ نهایی را تغییر می‌دهد؟
ب) چرا به جای استفاده از پاداش زمانی خطی مانند $r + \kappa\tau$, از ریشهٔ دوم τ استفاده می‌شود؟ استفاده از پاداش زمانی خطی چه مشکل رفتاری‌ای ممکن است برای عامل رفت‌وآمد ایجاد کند؟

موفق باشید :)