# یادگیری تقویتی

**نیم‌سال بهار ۱۴۰۲-۱۴۰۳**

استاد: محمدحسین رهبان

دانشکده‌ی مهندسی کامپیوتر

تمرین سری سوم | Policy Gradient | زمان تحویل: ۲۸ 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.
- 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 ۲ روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید و در مجموع ۵ روز تاخیر مجاز برای تمارین در اختیار دارید.

---

## سوال ۱: هدف یادگیری تقویتی (نظری) (۱۰ نمره)

هدف یادگیری تقویتی بیشینه‌سازی مجموع پاداش در طول زمان است

$$J(\theta) = \mathbb{E}[\sum_t \gamma^t r(s_t, a_t)]. \qquad (۱)$$

الف) منشأ تصادفی بودن میزان جمع پاداش که منجر می‌شود بخواهیم به جای مقدار آن در یک اپیزود، به امید ریاضی آن توجه کنیم چه چیزهایی است؟

ب) فرض کنید الگوریتم مورد استفاده، یادگیری یک سیاست $\pi_\theta : \mathcal{S} \to \mathcal{A}$ باشد. چرا نمی‌توان $\theta$ را مستقیم با استفاده از گرادیان کاهشی بر روی تابع هزینه تعریف شده در ۱ بهینه کرد؟ پیشنهادی دهید که چگونه می‌توان تنظیمات یادگیری تقویتی (شیوه‌ی اعمال کنش و دریافت پاداش) را تغییر داد به طوری که امکان بهینه‌سازی با اعمال گرادیان کاهشی به طور مستقیم بر روی رابطه‌ی ۱ وجود داشته باشد (راهنمایی: دلیلی که سبب می‌شود نتوان از رابطه‌ی ۱ به صورت مستقیم مشتق گرفت آن است که ما در حالت معمول در یادگیری تقویتی کنش را نمونه‌گیری کرده و انتخاب می‌کنیم و این عملیات قابلیت مشتق‌گیری ندارد).

## سوال ۲: بایاس و واریانس گرادیان سیاست (نظری) (۲۰ نمره)

الف) ثابت کنید تخمین رابطه‌ی ۲ از $\nabla_\theta J(\theta)$ ناریب است. چه فرض‌هایی را در اثباتتان در نظر گرفته‌اید؟

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}(\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}))(\sum_{t=0}^{T}\gamma^t r(s_{i,t}, a_{i,t})) \qquad (۲)$$

ب) تعریف می‌کنیم

$$\rho_\pi(s) = \sum_{t=0}^{\infty}\gamma^t p(s_t = s) \qquad (۳)$$

. نشان دهید

$$\mathbb{E}_\tau[\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t))] = \mathbb{E}_{\rho_\pi,\pi}[\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t, a_t)]$$

که در آن $\hat{Q}(s_t,a_t)=\sum_{t'=t}^{\infty}\gamma^{t'-t}r(s_{t'},a_{t'})$.

ب) تابع $b(s):\mathcal{S}\rightarrow\mathbb{R}$ را بیابید که واریانس تخمینگر نااریب

$$\nabla_\theta J(\theta)\approx\mathbb{E}_{\rho_\pi,\pi}[\sum_{t=0}^{\infty}\nabla_\theta\log\pi_\theta(a_t|s_t)(\hat{Q}(s_t,a_t)-b(s_t))] \tag{۴}$$

کمینه شود.

ج) حال با توجه به رابطه‌ای که در قسمت قبل به دست آورده‌اید، نشان دهید که اگر به جای $\hat{Q}(s_t,a_t)$ از مقدار واقعی $Q(s_t,a_t)$ استفاده کنیم، خواهیم داشت $b(s_t)=V^\pi(s_t)$.

## سوال ۳: TRPO (نظری) (۳۵ نمره)

الف) دو سیاست $\pi_\theta$ و $\pi_{\theta'}$ را در نظر بگیرید. فرض کنید

$$\forall s_t:|\pi_{\theta'}(a_t|s_t)-\pi_\theta(a_t|s_t)|\leq\epsilon. \tag{۵}$$

ثابت کنید

$$|p_{\theta'}(s_t)-p_\theta(s_t)|\leq 2t\epsilon.$$

اگر از قضایای کمکی استفاده می‌کنید، آن‌ها را اثبات نمایید.

ب) در صورتی که رابطه‌ی ۵ برقرار باشد، مرتبه‌ی اختلاف

$$\mathbb{E}_{p_{\theta'}(s_t)}[E_{a_t\sim\pi_\theta(a_t|s_t)}[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)}\gamma^t A^{\pi_\theta}(s_t,a_t)]] \tag{۶}$$

و

$$\mathbb{E}_{p_\theta(s_t)}[E_{a_t\sim\pi_\theta(a_t|s_t)}[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)}\gamma^t A^{\pi_\theta}(s_t,a_t)]] \tag{۷}$$

را در حالت افق محدود و نامتناهی به دست آورید.

ج) در صورت برقراری رابطه‌ی ۵، کران بالایی بر روی $D_{KL}(\pi_{\theta'}(a_t|s_t)||\pi_\theta(a_t|s_t))$ بیان نموده و آن را ثابت کنید.

د) بسط تیلور مرتبه‌ی اول

$$\sum_t\mathbb{E}_{s_t\sim p_\theta(s_t)}[E_{a_t\sim\pi_\theta(a_t|s_t)}[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)}\gamma^t A^{\pi_\theta}(s_t,a_t)]] \tag{۸}$$

را حول $\theta'=\theta$ بنویسید.

## سوال ۴: DPG (نظری) (۲۰ نمره)

در این سوال می‌خواهیم گرادیان سیاست را برای یک سیاست قطعی[۱] $\mu_\theta:\mathcal{S}\rightarrow\mathcal{A}$ به دست آوریم که $A=\mathbb{R}^m$ و $S$ یک زیر فضای فشرده از $\mathbb{R}^d$ است. محیط، مارکوف با تابع انتقال[۲] $p$ و توزیع حالت شروع $p_1$ است. احتمال حضور در حالت $s'$ پس از انجام $t$ کنش مطابق سیاست $\mu$ با شروع از $s$ را با $p(s\rightarrow s',t,\mu)$ نشان می‌دهیم. تعریف می‌کنیم

$$\rho_\mu(s'):=\int_{\mathcal{S}}\sum_{t=}^{\infty}\gamma^{t-1}p_1(s)p(s\rightarrow s',t,\mu)ds. \tag{۹}$$

تابع هدف به صورت

$$J(\mu_\theta)=\mathbb{E}[\sum_{t=1}^{\infty}\gamma^{t-1}r(s_t,a_t)|\mu]=\int_{\mathcal{S}}\rho_\mu(s)r(s,\mu_\theta(s))ds=\mathbb{E}_{s\sim\rho_\mu}[r(s,\mu_\theta(s)]] \tag{۱۰}$$

[۱] Deterministic
[۲] Transition

است. فرض کنید $p(s'|s,a)$، $\nabla_a p(s'|s,a)$، $r(s,a)$، $\nabla_a r(s,a)$، و $p_1(s)$ همگی نسبت به تمام پارامترها پیوسته‌اند. همچنین فرض کنید $b$ وجود دارد به طوری که $\sup_{a,s} r(s,a) < b$، $\sup_{a,s,s'} p(s'|s,a) < b$، $\sup_s p_1(s) < b$. به علاوه، $L$ نیز وجود دارد به طوری که $\sup_{a,s} \|\nabla_a r(s,a)\| < L \sup_{a,s,s'} \|\nabla_a p(s'|s,a)\| < L$. نشان دهید

$$\nabla_\theta J(\mu_\theta) = \int_S \rho_\mu(s) \nabla_\theta \mu_\theta(S) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} ds = \mathbb{E}_{s\sim\rho_\mu}[\nabla_\theta \mu_\theta(S) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}]. \quad (۱۱)$$

راهنمایی ۱: اثبات، شباهت‌هایی به قضیه‌ی استاندارد گرادیان سیاست در کتاب درس دارد.
راهنمایی ۲: برای جا به جایی ترتیب مشتق و انتگرال و همچنین ترتیب انتگرال‌ها در اثبات، توجه نمایید که با استفاده از شرایط گفته شده می‌توان نشان داد که $V^{\mu_\theta}(s)$ و $\nabla_\theta V^{\mu_\theta}(s)$ توابع پیوسته‌ای از $\theta$ و $s$ هستند. به علاوه، می‌توان از فشرده بودن $S$ نتیجه گرفت که برای هر $\theta$، $\|\nabla_\theta V^{\mu_\theta}(s)\|$، $\|\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\|$، و $\|\nabla_\theta \mu_\theta(s)\|$ توابع کران‌داری از $s$ هستند.