به نام خدا

# یادگیری تقویتی

**نیم‌سال بهار ۴۰۲-۱۴۰۱**

اساتید: دکتر رهبان، آقای حسنی

دانشکده‌ی مهندسی کامپیوتر

تمرین سری دوم | الگوریتم‌های گرادیان سیاست | زمان تحویل: ۱۸ فروردین

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. به اسکرین‌شات از منابع یا پاسخ افراد دیگر نمره‌ای تعلق نمی‌گیرد.
- 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 پنج روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید.

---

## سوال ۱: گرادیان سیاست (۲۵ نمره)

(آ) نشان دهید تخمین گرادیان با خط مبنا [1] وابسته به حالت $(\mathrm{b}(s_t))$ بدون بایاس است.

(ب) هدف از اضافه کردن خط مبنا به تخمین گرادیان، کاهش واریانس تخمین است. خط مبنا بهینه (که کمترین واریانس تخمین گرادیان را ایجاد می‌کند) را به دست آورید.

(ج) فرض کنید احتمال یک مسیر در MDP تحت سیاست $\pi$ و پاداش تخفیف داده شده به صورت زیر تعریف شده‌اند.

$$\mathrm{Pr}_{\mu}^{\pi}(\tau) = \mu\left(s_0\right)\pi\left(a_0 \mid s_0\right)P\left(s_1 \mid s_0, a_0\right)\pi\left(a_1 \mid s_1\right)\cdots \tag{۱}$$

$$R(\tau) := \sum_{t=0}^{\infty}\gamma^t r\left(s_t, a_t\right) \tag{۲}$$

هدف از روش گرادیان سیاست بیشینه‌سازی رابطه زیر است.

$$\max_{\theta\in\Theta} V^{\pi_\theta}(\rho) \tag{۳}$$

که در آن

$$V^{\pi}(\rho) := \mathbb{E}_{s_0\sim\rho}\left[V^{\pi}\left(s_0\right)\right] \tag{۴}$$

ثابت کنید:

$$\nabla V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau\sim\mathrm{Pr}_{\mu}^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t Q^{\pi_\theta}\left(s_t, a_t\right)\nabla\log\pi_\theta\left(a_t \mid s_t\right)\right] \tag{۵}$$

[1] Baseline

(د) برای سیاست $\pi$ می‌توان یک اندازه‌گیری احتمال [۲] برای بازدید حالت‌ها به صورت زیر تعریف کرد.

$$d_{s_0}^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^{\pi}\left(s_t = s \mid s_0\right) \tag{۶}$$

به جای $s_0$ می‌توان توزیع اولیه $\mu$ را در نظر گرفت که داریم:

$$d_{\mu}^{\pi}(s) = \mathbb{E}_{s_0\sim\mu}\left[d_{s_0}^{\pi}(s)\right] \tag{۷}$$

با استفاده از تعاریف فوق می‌توان نشان داد:

$$\mathbb{E}_{\tau\sim\Pr\pi}\left[\sum_{t=0}^{\infty}\gamma^t f\left(s_t, a_t\right)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[f(s,a)] \tag{۸}$$

با توجه به نتایج بخش‌های قبل روابط (۹) و (۱۰) را ثابت کنید.

$$\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a\mid s)\right] \tag{۹}$$

$$\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a\mid s)\right] \tag{۱۰}$$

## سوال ۲: الگوریتم‌های مبتنی بر ارزش برای مسائل با فعالیت‌های پیوسته (۲۰ نمره)

بسیاری از الگوریتم‌های یادگیری تقویتی از چهارچوب policy iteration برای یافتن سیاست مناسب استفاده می‌کنند. همانطور که در جلسات کلاس دیده‌ایم این الگوریتم شامل دو مرحله policy improvement و policy evaluation می‌باشد. در الگوریتم Q-learning نیز به صورت تکرار شونده اقدام به بهبود تخمین ارزش و یافتن سیاست با روابط ذیل می‌کنیم:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\left[r + \gamma\max_{a'}Q\left(s',a'\right)\right] \tag{۱۱}$$

$$\pi(s) = \underset{a}{\operatorname{argmax}}Q(s,a) \tag{۱۲}$$

(آ) عنوان کنید به چه دلیل این روابط در مسائلی با فضای فعالیت‌های پیوسته قابل اجرا نیستند؟

(ب) برای حل این مشکل و حل مسائل بهینه‌سازی $\max_{a'}Q(s,a')$ و $\underset{a}{\operatorname{argmax}}Q(s,a)$ روش‌های متنوعی ارائه شده‌اند. با افزایش ابعاد فضای فعالیت‌ها و پیچیده شدن مسئله بهینه‌سازی روش‌های مبتنی بر نمونه برداری کارایی خود را از دست می‌دهند. با این حال در صورت آگاهی از تابع هدف (روش‌های white-box) می‌توانیم از اطلاعاتی که گرادیان تابع هدف در اختیار ما قرار می‌دهد برای تسریع بهینه‌سازی استفاده کنیم. در این قسمت به بررسی برخی روش‌های در این دسته خواهیم پرداخت.

۱. یکی از روش‌های برای یافتن بهینه سراسری تابع Q فرض یک تابع محدب با بهینه به صورت فرم بسته برای این تابع بر حسب پارامتر a است. برای مثال تابع ارزش به فرم زیر را در نظر بگیرید.

$$Q_\phi(\mathbf{s},\mathbf{a}) = -\frac{1}{2}\left(\mathbf{a}-\mu_\phi(\mathbf{s})\right)^T P_\phi(\mathbf{s})\left(\mathbf{a}-\mu_\phi(\mathbf{s})\right) + V_\phi(\mathbf{s})$$

در این حالت مقادیر $\underset{a}{\operatorname{argmax}}Q(s,a)$ و $\max_{a}Q(s,a)$ را به دست بیاورید و بیان کنید در نظر گرفتن چنین فرم‌های ساده‌ای برای تابع ارزش چه مزایا و معایبی همراه خواهد داشت.

[2] Probability Measure

۲. رویکرد دیگر برای حل مسئله بهینه‌سازی یادگیری بهینه‌سازی است. به این ترتیب که مدل پارامتری در طول آموزش اقدام به یادگیری حل مسئله بهینه‌سازی می‌کند. این رویکرد در روش deep deterministic policy gradient مورد استفاده قرار گرفته شده است و مدل پارامتری برای حل مسئله بهینه‌سازی یک شبکه عصبی در نظر گرفته شده است.

آ. در ابتدا الگوریتم، تابع زیان و معماری الگوریتم DDPG را بیان کنید.

ب. نحوه استفاده از شبکه‌های actor و critic در این الگوریتم را با الگوریتم REINFROCE مقایسه کنید. عبور گرادیان از شبکه critic برای آموزش شبکه actor چه مزیت‌هایی به همراه دارد؟

ج. تابع هدف در بیشتر الگوریتم‌های یادگیری تقویتی بیشینه کردن امید مجموع پاداش دریافتی است که در ادامه نمایش داده شده است. در این رابطه شبکه actor با $\mu$ نمایش داده شده که با مجموعه پارامترهای $\theta$ مدل شده‌است.

$$\begin{aligned} J\left(\mu_\theta\right) &= \int_{\mathcal{S}} \rho^\mu(s) r\left(s, \mu_\theta(s)\right) \mathrm{d}s \\ &= \mathbb{E}_{s\sim\rho^\mu}\left[r\left(s, \mu_\theta(s)\right)\right] \end{aligned}$$

در جلسات پیشین درس دیدیم که بیشینه کردن این تابع هدف معادل با بیشینه کردن تابع ارزش $Q(s,a)$ یا $V(s)$ است. با توجه به این اطلاعات و پاسخ قسمت قبلی ثابت کنید که گرادیان تابع هدف برای پارامترهای شبکه actor به شکل زیر به دست خواهد آمد.

$$\begin{aligned} \nabla_\theta J\left(\mu_\theta\right) &= \int_{\mathcal{S}} \rho^\mu(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)\Big|_{a=\mu_\theta(s)} \mathrm{d}s \\ &= \mathbb{E}_{s\sim\rho^\mu}\left[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}\right] \end{aligned}$$

برای این منظور می‌توانید گام‌های زیر را دنبال کنید.

- در ابتدا با استفاده از تعریف تابع ارزش به شکل زیر

$$r\left(s, \mu_\theta(s)\right) + \int_{\mathcal{S}} \gamma p\left(s' \mid s, \mu_\theta(s)\right) V^{\mu_\theta}\left(s'\right) \mathrm{d}s'$$

و فرض پیوستگی توابع $p\left(s' \mid s, a\right), \mu_\theta(s), V^{\mu_\theta}(s)$ و مشتق آنها نسبت به $\theta$ که به دنبال آن امکان جابه‌جایی عملگرهای گرادیان و انتگرال را خواهیم داشت، به یک رابطه بازگشتی برای گرادیان تابع $V^{\mu_\theta}(s)$ بر حسب $V^{\mu_\theta}(s')$ برسید.

- در گام بعدی با جایگذاری متوالی رابطه بازگشتی به دست آماده به صورت حدی و با فرض محدود بودن نرم تابع ارزش و finite horizon بودن مسئله، به یک فرم بسته برای گرادیان تابع ارزش بر حسب پرامتر $\theta$ می‌رسیم.

- در گام آخر از این گرادیان برحسب $S$ امیدریاضی می‌گیریم.

## سوال ۳: روش بهینه‌سازی جدید با تغییر Trust Region (۳۰ نمره)

در این سوال سعی می‌کنیم تا با تغییر Trust Region الگوریتم بهینه‌سازی TRPO یک روش جدید و البته قابل اتکاتر برای به دست‌آوردن سیاست بهینه به دست آوریم. سعی می‌کنیم تا گام به گام به سمت حل مسأله پیش برویم. در این مسأله فضای $\mathcal{X}$ را برای سادگی یک مجموعه بسته و کراندار اندازه‌گیری‌پذیر در نظر می‌گیریم به صورتی که $\mathcal{P}(\mathcal{X})$ مجموعه‌ی تمام توزیع‌های احتمال موجود روی $\mathcal{X}$ است. همچنین با داشتن یک سیاست $\pi$ روی یک MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma)$ ارزش هر وضعیت یا حرکت_وضعیت را با $V^\pi(s)$ و $Q^\pi(s,a)$ نشان دهیم تابع مزیت را به صورت $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ می‌توان تعریف کرد.
در ابتدا با چند تعریف شروع می‌کنیم:

(آ) **هزینه انتقال:** تابع هزینه‌ی انتقال را به صورت یک تابع $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ نشان می‌دهیم.

(ب) همچنین برای هر دو توزیع $\mu, \nu \in \mathcal{P}(\mathcal{X})$ مجموعه توزیع‌های توأمی که حاشیه‌های $\mu$ و $\nu$ دارند را به صورت زیر تعریف می‌کنیم:

$$\Gamma(\mu,\nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \gamma(A \times \mathcal{X}) = \mu(A), \gamma(\mathcal{X} \times B) = \nu(B)\} \tag{۱۳}$$

(ج) **تابع اختلاف انتقال بهینه:** برای هر دو توزیع $\mu, \nu \in \mathcal{P}(\mathcal{X})$ تابع بهینه‌ی هزینه‌ی انتقال را به صورت زیر تعریف می‌کنیم:

$$C(\mu,\nu) = \min_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,x') d\gamma(x,x') \tag{۱۴}$$

(د) **تابع هدف:** تابع هدف که پاداش کاهش‌یابنده را امید ریاضی محاسبه می‌کند به صورت زیر تعریف می‌کنیم:

$$J(\pi) = \mathbb{E}_{\rho,\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right] \tag{۱۵}$$

که در آن $\rho$ توزیع اولیه روی وضعیت شروع است.

حال می‌توانیم گام به گام به سمت حل مسأله حرکت کنیم:

(آ) ابتدا نشان دهید در صورتی که $\pi, \tilde{\pi} \in \Pi$ دو سیاست دلخواه باشند، خواهیم داشت:

$$J(\tilde{\pi}) = J(\pi) + \int_{\mathcal{S}}\int_{\mathcal{A}} A^{\pi}(s,a)d\tilde{\pi}(a|s)d\rho_{\tilde{\pi}}(s) \tag{۱۶}$$

که در آن $\rho_{\tilde{\pi}}(s)$، *توزیع وضعیت آینده کاهشی*، به صورت زیر است:

$$\rho_{\tilde{\pi}}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}[s_t = s|\tilde{\pi}, \rho] \tag{۱۷}$$

(ب) اکنون به جای استفاده از عبارت بالا به صورت مستقیم (که محاسبه‌ی آن هزینه‌ی زیادی در بر دارد) از شکل تغییر یافته‌ی زیر استفاده می‌کنیم:

$$L_{\pi}(\tilde{\pi}) = J(\pi) + \int_{\mathcal{S}}\int_{\mathcal{A}} A^{\pi}(s,a)d\tilde{\pi}(a|s)d\rho_{\pi}(s) \tag{۱۸}$$

حال استفاده از این تابع هدف برای بیشینه سازی معقول‌تر به نظر می‌رسد، در صورتی که تغییر سیاست به گونه‌ای نباشد که توزیع وضعیت آینده کاهشی سیاست آتی با سیاست فعلی تفاوت قابل توجهی داشته باشد. که این فرض بنا بر شرایط مسأله فرض معقولی است و برای سادگی می‌توان از آن در ادامه‌ی کار استفاده کرد.

(ج) حال در ادامه با توجه به معادله ۱۸ می‌توانیم مسأله‌ی بهینه‌سازی را با توجه به بیشینه کردن عبارت دوم در آن، به شکل زیر تعریف کرد:

$$\begin{aligned} \sup_{\tilde{\pi}\in\Pi} \quad & \int_{\mathcal{S}}\int_{\mathcal{A}} A^{\pi}(s,a)d\tilde{\pi}(a|s)d\rho_{\pi}(s) \\ \text{s.t.} \quad & \tilde{\pi} \in \mathcal{T}_{\epsilon} := \left\{\tilde{\pi} \in \Pi : \int_{\mathcal{S}} C(\pi(.|s), \tilde{\pi}(.|s))d\rho_{\pi}(s) \leq \epsilon\right\} \end{aligned} \tag{۱۹}$$

همانطور که در مسأله‌ی بهینه‌سازی بالا مشاهده می‌شود $\mathcal{T}_{\epsilon}$ در واقع یک تعریف جدید از Trust Region مسأله است.

در ادامه باید گفت که در مسأله‌ی بهینه‌سازی جدید معرفی‌شده محاسبه‌ی تابع اختلاف انتقال بهینه‌ی مورد نظر باز هم از لحاظ محاسباتی دشواری‌های قابل توجهی دارد. اما برای حل آن از می‌توان از تکنیک تبدیل مسأله به دوگان آن استفاده کرد که در ادامه خواهیم دید تا چه حدی می‌تواند حل مسأله را بهبود بخشد.
اما پیش از آغاز به صورتی گذرا تکنیک دوگان‌گیری از یک مسأله‌ی بهینه‌سازی را بیان می‌کنیم:

(آ) فرض کنید یک مسأله‌ی بهینه‌سازی به صورت زیر داده شده‌است:

$$\begin{aligned} \min_{x\in X} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \quad \forall i \in \{1, 2, \ldots, k\} \end{aligned} \tag{۲۰}$$

حال تابع لاگرانژیان آن را به صورت زیر تعریف می‌کنیم:

$$L(x, \lambda_1, \lambda_2, \ldots, \lambda_k) = f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x) + \ldots + \lambda_k g_k(x) \tag{۲۱}$$

همچنین تابع دوگان آن را به صورت زیر تعریف می‌کنیم:

$$g(\lambda_1, \ldots, \lambda_k) = \begin{cases} \inf_{x\in X} L(x, \lambda_1, \ldots, \lambda_k) & \text{if} \lambda_i \geq 0, \forall i \\ -\infty & \text{o.w.} \end{cases} \tag{۲۲}$$

حال مسأله‌ی دوگان را به صورت زیر تعریف می‌کنیم:

$$\begin{aligned} \max \quad & g(\lambda_1, \ldots, \lambda_k) \\ \text{s.t.} \quad & \lambda_i \geq 0 \end{aligned} \tag{۲۳}$$

(ب) به سادگی قابل مشاهده است که اگر پاسخ مسأله‌ی ۲۰ را با $f^*$ و پاسخ مسأله‌ی ۲۳ را با $g^*$ نمایش دهیم، آنگاه با توجه به تعریف بالا خواهیم داشت: $g^* \leq f^*$. که این نتیجه به دوگانی ضعیف معروف است.

(ج) همچنین می‌توان نشان داد در صورتی که بر روی مجموعه‌ی feasible مسأله‌ی ۲۰ شرایطی برقرار باشد آنگاه دوگانی قوی است یعنی: $f^* = g^*$. از دوگانی قوی در حل مسأله‌ی بهینه‌سازی گذشته استفاده خواهد شد که البته با توجه به اینکه این موضوع از حوزه‌ی این تمرین خارج است برقراری آن را در این سوال فرض می‌کنیم.

اما اکنون به سراغ مسأله‌ی بهینه‌سازی ۱۹ می‌رویم:

(آ) پیش از ادامه برای به دست آوردن دوگان مناسب دو فرض که به صورت معمول برفرار است را در نظر می‌گیریم:

- فضای وضعیت $\mathcal{S}$ و کنش $\mathcal{A}$ مجموعه‌هایی بسته و کران‌دار هستند. به علاوه تابع پاداش $r$ یک تابع پیوسته و امید ریاضی هر تابع پیوسته $w$ روی $\mathcal{S}$ یک تابع پیوسته است.
- برای هر سیاست $\pi$ تابع مزیت $A^\pi$ یک تابع پیوسته است همچنین تابع هزینه انتقال $c$ یک تابع پیوسته است که هزینه‌ی هر کنش با خودش برابر صفر است: $c(a,a) = 0$.

(ب) حال فرم دوگان مسأله‌ی ۱۹ را به دست آورید. سپس یک کران بالا برای آن ارائه کنید به صورتی که به $\tilde{\pi}$وابستگی نداشته باشد. (راهنمایی: یک کران بالا برای فرم دوگان را به صورت زیر می‌توان نوشت:

$$\begin{aligned} \min \quad & \lambda\epsilon + \int_{\mathcal{S}} \int_{\mathcal{A}} \max_{a' \in \mathcal{A}} \{A^\pi(s,a') - \lambda c(a,a')\} d\pi(a|s) d\rho_\pi(s) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned} \tag{۲۴}$$

. همچنین برای به دست آوردن این کران می‌توانید از قضیه Kantorovich در مورد اختلاف انتقال بهینه استفاده کنید.)

(ج) **قضیه Kantorovich**: می‌توان نشان داد که پاسخ اختلاف انتقال بهینه ۱۴ را می‌توان به صورت زیر نوشت:

$$C(\mu,\nu) = \sup_{\substack{\phi,\psi \\ \phi(x)+\psi(x') \leq c(x,x')}} \left\{ \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{X}} \psi(x') d\nu(x') \right\} \tag{۲۵}$$

با عرض تبریک سوال تا اینجا به پایان رسید! :)
در ادامه به بیان روش به دست‌آوردن سیاست بهینه مبتنی بر این کران بالای به دست آمده می‌پردازیم.
می‌توان نشان داد که مسأله‌ی ۲۴ یک مسأله‌ی بهینه‌سازی محدب است که با توجه به تک پارامتره بودن می‌توان آن را با استفاده از حل‌کننده‌های مرسوم به سرانجام رساند. حال پس از پیدا کردن $\lambda^*$ بهینه برای این مسأله می‌توان با استفاده از آن سیاست (تقریباً) بهینه‌ی مربوط به مسأله‌ی اول ۱۹ را با تکنیک‌هایی به دست آورد که به جهت جلوگیری از اطناب از ذکر آن خودداری می‌کنیم.
حال تنها مسأله‌ای که باقی می‌ماند تخمین‌زدن تابع مزیت و محاسبه‌ی انتگرال‌هاست که می‌توان آن‌ها را با استفاده از یک شبکه‌ی عصبی یا روش های MC یا TD به دست آورد.

## سوال ۴: پیاده‌سازی (۲۵ نمره)

هدف این بخش از تمرین پیاده‌سازی دو الگوریتم PPO و DDPG و مقایسه نتایج این دو الگوریتم در محیط Pendulum-v1 از کتابخانه gym است. با استفاده از نوت‌بوک داده شده این دو الگوریتم را پیاده‌سازی کنید. برای مقایسه، نمودارهای هزینه شبکه‌های actor و critic و نمودار پاداش در طول اپیزودها را بر هر دو الگوریتم رسم کنید. (برای کاهش نوسانات شدید نمودار پاداش در طول اپیزود و مقایسه بهتر نمودارها می‌توانید با روش پنجره لغزان [3] میانگین بگیرید.)

[3] Sliding Window