تمرین دوم – یادگیری تقویتی

دانشجو: سایه جارالهی

شماره دانشجویی: ۹۸۱۰۱۳۳۹

استاد: دکتر رهبان- آقای حسنی

بهار ۱۴۰۲

سوال ۱

الف-

در این حالت baseline داریم:

$$\nabla_\theta \mathbb{E}[r] \approx \frac{1}{N}\sum_{i=1}^{N} [\nabla_\theta \log p_\theta(\tau_i)[\, r(\tau_i) - b]] \quad , b = \frac{1}{N}\sum_{i=1}^{N} r(\tau_i)$$

$$E[\nabla_\theta \log p_\theta(\tau) b] = \int p_\theta(\tau) \nabla_\theta log p_{\theta(\tau)b} d\tau =^{I} \int \nabla_\theta\, p_\theta(\tau) b d\tau = b \nabla_\theta \int p_\theta(\tau) d\tau = b\, \nabla_\theta 1 = 0$$

برابری بخش I به این دلیل برقرار است که داریم $p_\theta(\tau_i)\nabla_\theta log p_\theta(\tau_i) = \nabla_\theta p_\theta(\tau_i)$

درواقع در انتها به عبارتی میرسیم که انتگرال روی توزیع احتمال $\pi(a|s)$ است و چون انتگرال روی کل دامنه گرفته میشود مقدار آن عدد ثابت یک است و گرادیان عدد ثابت نیز صفر است. در این بخش ثابت کردیم که اضافه کردن مبنا باعث نمی شود که بایاس ایجاد شود واکسپکتد ترم اضافه شده صفر است.

ب-

می‌دانیم که $var(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2$ و در نتیجه :

$$var = \mathbb{E}_{\tau\sim p_\theta(\tau)}[(\nabla_\theta \log p_\theta(\tau)(r(\tau) - b))^2] - \mathbb{E}_{\tau\sim p_{\theta(\tau)}}\left[\nabla_{\theta \log}\, p_\theta(\tau)(r(\tau) - b)\right]^2$$

بخش دوم به دلیل الف برابر است با $\mathbb{E}_{\tau\sim p_{\theta(\tau)}}\left[\nabla_{\theta \log}\, p_\theta(\tau)\big(r(\tau)\big)\right]^2$ زیرا baseline بایاس اضافه نمیکند. همچنین

$$g(\tau) = \nabla_{\theta \log}\, p_\theta(\tau)$$

حال نسبت به بایس مشتق میگیریم تا اکسترمم پیدا شود. همچنین term دوم چون بایاس ندارد مشتقش صفر است.

$$\frac{dVar}{db} = \frac{d}{db}\mathbb{E}[g(\tau)^2(r(\tau) - b)^2] = \frac{d}{db}(\mathbb{E}[g(\tau)^2 r(\tau)^2] - 2\mathbb{E}[g(\tau)^2 r(\tau) b] + b^2\mathbb{E}[g(\tau)^2])$$
$$= 2b\mathbb{E}[g(\tau)^2] - 2\mathbb{E}[g(\tau)^2 r(\tau)] = 0 \rightarrow b = \frac{\mathbb{E}[g(\tau)^2 r(\tau)]}{\mathbb{E}[g(\tau)^2]}$$

ج-

د-

سوال ۲

آ)

این روش ها با نگه داری مقادیر در جدول محاسبات را انجام میدهند. در واقع به ازای هر استیت و اکشن باید یک مقدار داشته باشیم که در طول زمان اپدیت میشود. درواقع داده ها tabular هستند. اما در صورتی که فضای حالات پیوسته باشد امکان نگه داری مقادیر به دلیل بی شمار بودن آن ها وجود ندارد و در نتیجه محاسبه ممکن نیست. حتی اگر فضای اکشن ها را نیز گسسته سازی کنیم به دلیل زیاد بودن تعداد اکشن ها باز هم محاسبه ممکن نیست.

ب)

۱- باید از رابطه صورت سوال بر حسب a مشتق بگیریم تا argmax مشخص شود. همچنین با جایگزین کردن این مقدار، ماکزیمم تابع را به دست آوریم.

$$\frac{\partial Q}{\partial a} = -\frac{1}{2}\left[2P_\phi(s)\left(a-\mu_\phi(s)\right)\right] + 0 = -P_\phi(s)(a-\mu_\phi(s) = 0 \rightarrow a = \mu_\phi(s)$$

$$\rightarrow \operatorname*{argmax}_a Q(s,a) = \mu_\phi(s)$$

$$\rightarrow \max_a Q(s,a) = V_\phi(s)$$

عبارات بالا از نظر شهودی نیز درست هستند زیرا در این حالت ماکزیمم مقداری که میتوان برای Q متصور شد هم معادل با ارزش آن است (value)

در نظر فرم ساده مشابه این فرم میتواند مزایایی داشته باشد. برای مثال محاسبه آن ساده است و از نظر محاسباتی هزینه کمی دارد. همچنین محاسبه ماکزیمم آن و اکشن بهینه نیز هزینه زیادی ندارد زیرا پیدا کردن مشتق آن آسان است. همچنین برای هردو حالت پیوسته و گسسته بودن فضای حالت میتواند استفاده شود .

از معایب آن میتوان گفت که پیچیدگی‌های فضای اکشن و استیت را در نظر نمیگیرد و در میانگین و ماتریس کواریانس اعمال نمیکند. همچنین از نظر اندازه برای نگه داری P و a میتوانند ماتریس های بزرگی باشند که محاسبه آن‌ها به مشکل مواجه میشود.

۲-

الف) الگوریتم DDPG یک الگوریتم model-free, actor-critic است که برای مسائل با اکشن اسپیس پیوسته در نظر گرفته میشود. در این الگوریتم actor وظیفه مپ کردن استیت به اکشن را دارد. درواقع actor مقدار ریوارد را با استفاده از q-value های تخمین زده شده بیشینه میکند. critic وظیفه تخمین Q-value ها را دارد به طوری که اختلاف آن‌ها با مقادیر واقعی return کمینه شود.

این الگوریتم برای stable کردن خود از دو راهکار در معماری خود استفاده می‌کند. یکی از آن‌ها استفاده از replay buffer است که هر بار یک mini-batch از آن برای ترین استفاده می‌شود. و دیگری آن است که یک target network قرار می‌دهد که وزن‌های هردوی critic و actor را دارد اما سرعت تغییر آن کمتر از این دو نتورک است و اینکار باعث می‌شود که یادگیری stable شود.

حال به بیان الگوریتم و تابع هزینه می‌پردازیم، برای اینکار نوتیشن ها را مشخص میکنیم.

برای شبکه actor داریم: $\mu(s|\theta^\mu)$ که θ^μ همان وزن‌های شبکه است.

برای شبکه critic داریم: $Q(s,a|\theta^Q)$ که θ^Q همان وزن‌های شبکه است.

همچنین یک target network هم داریم که وزن‌های آن به دلیل استیبل شدن با سرعت کمتری آپدیت می‌شود. برای آن نیز داریم:

برای شبکه actor داریم: $\mu'(s|\theta^\mu)$ که $\theta^{\mu'}$ همان وزن‌های شبکه است و با مقدار θ^μ مقداردهی میشود.

برای شبکه critic داریم: $Q'(s,a|\theta^Q)$ که $\theta^{Q'}$ همان وزن‌های شبکه است و با مقدار θ^Q مقداردهی اولیه میشود.

حال برای تابع هزینه داریم:

$$L(\theta^Q) = E[(Q(s_t,a_t|\theta^Q) - y_t)^2]$$

$$y_t = r(s_t,a_t) + \gamma Q(s_{t+1}, \mu(s_{t+1})|\theta^Q)$$

همچنین برای روند الگوریتم داریم:

(برای replay buffer که آن را به اختصار R میگوییم)

برای هر اپیزود مراحل زیر انجام می شود:

یک رندوم پراسس N در نظر گرفته میشود که برای نویز محیط است

به ازای هر t در بازه [1,T] داریم:

۱. انتخاب اکشن به صورت $a_t = \mu(s_t|\theta^\mu) + N_t$ که N برای در نظر گرفتن نویز exploration است.
۲. اکشن در محیط اجرا میشود و ریوارد r_t دریافت میشود و استیت بعدی s_{t+1} نیز مشخص می‌شود
۳. تاپل (s_t, a_t, s_{t+1}) به N اضافه میشود و طبق پیپر درصورتی که اندازه بافر پر باشد، قدیمی ترین داده حذف میشود.
۴. حال یک mini-batch با اندازه مشخص از replay buffer سمپل میشود.
۵. برای هر سمپل مقدار $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$ محاسبه میشود. توجه داریم که این مقدار با استفاده از target network محاسبه میشود.
۶. نتورک critic با استفاده از loss ای که در بخش قبل گفته شد آپدیت میشود. یعنی داریم:

$$L = \frac{1}{N}\sum_i (y_i - Q(s_i, a\_i|\theta^\mu))^2$$

۷. نتورک actorبا استفاده از sampled policy gradient آندیت میشود.

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s,a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i}$$

۸. نتورک target با استفاده از رابطه زیرو به ازای یک $\tau \in [0,1]$ آپدیت میشود. بدیهی است که هرچقد مقدار τ بزرگتر باشد تاثیر سمپل ها بیشتر است و برعکس.

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

لازم به توضیح است که این بخش دقیقا با خواندن متن پیپر اصلی نوشته شده است.

ب) در الگوریتم DDPG بخش actor وظیفه پیدا کردن پالیسی بهینه و critic وظیفه تخمین زدن تابع ارزش را دارد. همچنین داده ها از یک replay buffer در می آیند و هربار با استفاده پالیسی بهینه تا آنجا و با درنظر گرفتن نویز سمپل جدید اضافه میشود.

اما در الگوریتم reinforce مقدار ارزش با استفاده از return ها و به طورت مستقیم به دست می آید و پالیسی مستقیما با استفاده از گرادیان log-prob برای اکشن ها باتوجه به پالیسی به دست می آید. (توجه داریم که اینجا منظور از ارزش لزوما value نیست و Qنیز هست) درواقع تنها بخش اکتور را داریم و critic را نداریم.

در نتیجه در الگوریتمREINFORCE به دلیل نویزی بودن log-probability ها و ریواردهای تجمیعی، گرادیان ها نویزی میشوند و ممکن است نویز محیط باعث شود که یادگیری درست انجام نشود و گرادیان ها هم نویزی باشند. اما در DDPG به دلیل وجود critic و تخمین ارزش با استفاده از آن واریانس گرادیان کاهش یافته و در نتیجه یادگیری سرعت میگیرد و بهبود می یابد. همچنین critic میتواند یک بیس لاین برای policy gradient باشد و واریانس را کاهش دهد.

همانطور که در روابط بالا نیز دیده شد، گرادیان critic در گرادیان actor عبور میکند و در آن تاثیر دارد. وجود این گرادیان باعث می شود که: ۱. از آنجا که critic وظیفه به دست آوردن بهبود برای ارزش‌ها را دارد، باعث می‌شود پالیسی در جهتی تغییر کند که ارزش‌ها نیز بیشتر شوند و در نتیجه پالیسی بهتری یاد گرفته شود. در نتیجه کمک می‌کند تا exploration بهتری انجام شود و مسیر بهبود پالیسی هموارتر شود. ۲. همچنین میتوان گفت مسیر رسیدن به پالیسی بهینه را optimize میکند و رسیدن به آن نقطه با تعداد گام کمتری امکان پذیر است. ۳. استفاده از critic و گرادیان آن میتواند به عنوان یک baseline واریانس را کاهش دهد و یادگیری را stable کند. ۳. از آنجا که critic با تابع ارزش‌ها سر و کار دارد میتوان از experience آن و وقایع گذشته استفاده کرد تا آموزش actor بهینه‌تر باشد.

ج)

- محاسبه رابطه بازگشتی:

$$\nabla_\theta V^{\mu_\theta}(s) = \nabla_\theta \left[r(s, \mu_\theta(s)) + \int_S \gamma p(s' | s, \mu_\theta(s)) V^{\mu_\theta}(s') ds' \right]$$

از قوانین گرادیان و ضرب و قانون زنجیره ای استفاده میکنیم و چون میتوان گرادیان را داخل انتگرال برد عبارارت زیر به دست می آید. همچنین توجه داریم که همه انتگرال ها روی S گرفته شده و به دلیل سادگی بازه را دیگر نمینویسیم.

$$= \nabla_\theta r(s, \mu_\theta(s)) + \gamma \int \nabla_\theta (p(s'|s, \mu_\theta(s) V^{\mu_\theta}(s')) ds'$$

$$\nabla_\theta \mu_\theta(s) \nabla_a r(s,a)|_{a=\mu_\theta(s)} + \int \gamma \left[p(s' | s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') + \nabla_\theta p(s' | s, \mu_\theta(s)) V^{\mu_\theta}(s') \right] ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)|_{a=\mu_\theta(s)} + \int \gamma \left[p(s' | s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') + \nabla_\theta \mu_\theta(s) \nabla_a p(s'|s,a)|_{a=\mu_\theta(s)} V^{\mu_\theta}(s') \right] ds'$$

حال میتوان انتگرال را با جمع جدا کرد.

$$= \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)|_{a=\mu_\theta(s)} + \int \gamma p(s' | s, \mu_\theta(s) \nabla_\theta V^{\mu_\theta}(s') ds' + \int \gamma \nabla_\theta \mu_\theta(s) \nabla_a p(s'|s,a)\Big|_{a=\mu_\theta(s)} V^{\mu_\theta}(s') ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a \left(r(s,a) + \int \gamma p(s'|s,a) V^{\mu_\theta}(s') ds' \right)|_{a=\mu_\theta(s)} + \int \gamma p(s' | s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int \gamma p(s' | s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') ds'$$

- جایگذاری رابطه بازگشتی:

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int \gamma p(s' | s, \mu_\theta(s)) \left[\nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} + \int \gamma p(s'' | s', \mu_\theta(s')) \nabla_\theta V^{\mu_\theta}(s'') ds'' \right] ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int \gamma p(s' | s, \mu_\theta(s)) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' + \int \int \gamma^2 p(s'|s, \mu_\theta(s)) p(s'' | s', \mu_\theta(s')) \nabla_\theta V^{\mu_\theta}(s'') ds'' ds'$$

در این بخش از منبع برای ساده سازی بیشتر عبارت بالا استفاده میکنیم. یک توصیف احتمالاتی را اینطور بیان میکنیم که $p(s \to s', 1, \mu_\theta)$ یعنی با انجام یک اکشن از s به s' برسیم و پالیسی نیز μ_θ باشد.

حال با استفاده از این توصیف، تعویض انتگرال روی s’ و s’’ با یکدیگر و انتگرال گیری به عبارت زیر میرسیم.

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_{\theta(s')}} ds' + \int \gamma^2 p(s \to s', 2, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds' + \cdots.$$

$$= \int \sum_{i=0}^{\infty} \gamma^i p(s \to s', t' \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds'$$

•

$$\int \int p_1(s) \sum_{i=0}^{\infty} \gamma^i p(s \to s', t' \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds'\, ds$$

$$\int \int \sum_{i=0}^{\infty} \gamma^i p_1(s) p(s \to s', t' \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds'\, ds$$

پس از باز کردن انتگرال و انتگرال گیری برحسب s’ به عبارت زیر میرسیم.

$$= \int \rho^{\mu_\theta}(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}\,(s,a)|_{a=\mu_\theta(s)}$$

سوال ۳

الف)

ابتدا در نظر میگیریم که

$$J(\pi) = \mathbb{E}_{s\sim D}[V_\pi(s)]$$

حال ابتدا اختلاف $V_{\pi'} - V_\pi$ را پیدا میکنیم. همچنین $\pi' = \tilde{\pi}$ است و به دلیل سادگی بیشتر اینطور تغییر کرده.

$$V_{\pi'}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(a_t, s_t)\sim \pi' P_t}[R(s_t, a_t)] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(a_t, s_t, s_{t+1})\sim \pi' P_t P(s_{t+1}; s_t; a_t)} [(1-\gamma) R(s_t, a_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t))] + V_\pi(s)$$

$$= V_{\pi}(s) + \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(a_t,s_t)\sim\pi' P_t}[A_{\pi}(s_t,a_t)] = V_{\pi}(s) + \frac{1}{1-\gamma}\mathbb{E}_{(a,s')\sim\pi',\rho}[A\_\pi(s',a)]$$

حال از دو طرف رابطه زیر expected میگیریم:

$$\mathbb{E}[V_{\pi'}(s) - V_{\pi}(s)] = \mathbb{E}[\frac{1}{1-\gamma}\mathbb{E}_{(a,s')\sim\pi',\rho}\left[A_{\pi(s',a)}\right]]$$

و به عبارت:

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma}\int\int A^{\pi}(s,a)d\pi'(a|s)d\rho_{\pi'}(s)$$

میرسیم.

منبع: پس از مقدار زیادی جستجو پیپر مربوط به این ایده را پیدا کردم و در منبع ۱۱ ام آن اثبات این بخش مشخص شده بود .

ب-

هدف پیدا کردن دوال برای مسئله زیر است:

$$\sup_{\pi'\in\Pi}\int\int A^{\pi}(s,a)d\pi'(a|s)d\rho_{\pi}(s): s.t. \pi' \in \mathrm{T}_{\epsilon} \coloneqq \{\pi' \in \Pi: \int C\big(\pi(.|s),\pi'(.|s)\big)d\rho_{\pi(s)} \le \epsilon\}$$

باتوجه به توضیحات سوال داریم:

$$B = \int\int A^{\pi}(s,a)d\pi'(a|s)d\rho_{\pi}(s)$$

$$\sup\inf B + \lambda(\epsilon - \int C\big(\pi(.|s),\pi'(.|s)\big)d\rho_{\pi(s)})$$

$$\le \inf\big\{\sup\big\{B - \int C\big(\pi(.|s),\pi'(.|s)\big)d\rho_{\pi(s)}\big\} + \lambda\epsilon\big\}$$

میتوان انتگرال ها را به صورت یکسان نوشت:

$$\le \inf\{\{\int \sup\{\int A^{\pi}(s,a)d\pi'(a|s) - \lambda C(\pi(.\mid s),\pi'.|s))d\rho_{\pi}(s)\}\} + \lambda\epsilon\}$$

$$= \inf\{\left\{\int sup\int A^{\pi}(s,a)d\pi'(a|s) - \lambda[\sup_{\phi,\psi:\,\phi(x)+\psi(x')\le c(x,x')}\{\int\phi(a)d\pi(a|s) + \int\psi(a)d\pi'(a|s)\}]d\rho_{\pi}(s)\right\} + \lambda\epsilon\}$$

همچنین فرض های زیر را میکنیم:

$$\Psi(.) = \frac{A^{\pi}(s,.)}{\lambda}\ and\ \phi(.) = \inf\{c(.,a) - \psi(a)\}$$

از روی فرض های سوال میتوان گفت که

$$\phi(x) + \psi(y) = \inf\{c(x,a\} - \psi(a)\} + \psi(y) \le c(x,y) - \psi(y) + \psi(y) \le c(x,y)$$

در نتیجه ϕ,ψ یک انتخاب suboptimal است و با جایگذاری به عبارت زیر برای upper boundمیرسیم.

$$\inf\{\int \int sup\, \{A^{\pi}(s,a') - \lambda c(a,a')\}\, d\pi(a|s) d\rho_{\pi}(s)\}$$