به نام خدا

یادگیری تقویتی

تمرین تئوری دوم

یلدا شعبان زاده

۹۸۱۰۱۸۲۲

نیمسال دوم ۰۲-۰۱

فهرست

سوال ۱-گرادیان سیاست

(آ) می‌دانیم پالیسی گرادیان مبتنی بر baseline به صورت زیر است:

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\big(a_{i,t}, s_{i,t}\big)\left[\Big(\sum_{t'=t}^{T}\gamma^{t'} r(s_{i,t'}, a_{i,t'}\Big) - b\right]$$

و در اسلایدهای درس نشان دادیم که:

$$E_{\tau\sim\pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau_i)\, b] = \int \pi_\theta(\tau_i)\nabla_\theta \log \pi_\theta(\tau_i)\, b d\tau_i \xRightarrow{\pi_\theta(\tau_i)\nabla_\theta \log \pi_\theta(\tau_i)=\nabla_\theta \pi_\theta(\tau_i)} \int \nabla_\theta \pi_\theta(\tau_i) b d\tau_i$$
$$= b\nabla_\theta \int \pi_\theta(\tau_i) d\tau_i = b\nabla_\theta 1 = 0$$

حال اگر بدانیم b تابعی از s باشد می‌توان گفت:

$$\begin{aligned}E_{\tau\sim\pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(a_t|s_t)\, b(s_t)] &= E_{s_{0:t},a_{0:t-1}}[E_{s_{t+1:T},a_{t:T-1}}[\nabla_\theta \log \pi_\theta(a_t|s_t)\, b(s_t)]\\ &= E_{s_{0:t},a_{0:t-1}}\left[b(s_t) E_{s_{t+1:T},a_{t:T-1}}[\nabla_\theta \log \pi_\theta(a_t|s_t)]\right]\\ &= E_{s_{0:t},a_{0:t-1}}\left[b(s_t) E_{a_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)]\right] = E_{s_{0:t},a_{0:t-1}}[b(s_t).0] = 0\end{aligned}$$

بنابراین در این حالت نیز unbiased است.

***

(ب) برای بدست آوردن baseline بهینه که کمترین واریانس تخمین گرادیان را ایجاد می‌کند داریم:

$$Var[x] = E[x^2] - E[x]^2$$
$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}\big[\nabla_\theta \log \pi_\theta(\tau)\, [r(\tau) - b]\big]$$
$$Var = E_{\tau\sim\pi_\theta(\tau)}\left[\Big(\underbrace{\nabla_\theta \log \pi_\theta(\tau)}_{g(\tau)}[r(\tau)-b]\Big)^2\right] - \underbrace{E_{\tau\sim p_\theta(\tau)}\big[\nabla_\theta \log \pi_\theta(\tau)\, [r(\tau) - b]\big]^2}_{\substack{This\ is\ E_{\tau\sim\pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau) r(\tau)]^2 = 0 \\ (b\ can\ be\ dependant\ to\ state)}}$$

$$\frac{\partial Var}{\partial b} = \frac{\partial}{\partial b} E_{\tau\sim p_\theta(\tau)}[g(\tau)^2 (r(\tau)-b)^2] = \frac{\partial}{\partial b}\big(E[g(\tau)^2 r(\tau)^2] - 2E[g(\tau)^2 r(\tau) b] + b^2 E[g(\tau)^2]\big)$$
$$= -2E[g(\tau)^2 r(\tau)] + 2bE[g(\tau)^2] = 0$$
$$\Rightarrow b = \frac{E[g(\tau)^2 r(\tau)]}{E[g(\tau)^2]} = \frac{E[\nabla_\theta \log \pi_\theta(\tau)^2\, r(\tau)]}{E[(\nabla_\theta \log \pi_\theta(\tau))^2]}$$

***

ج) هدف از روش گرادیان سیاست بهینه‌سازی عبارت زیر است:

$$\max_{\theta\in\Theta} V^{\pi_\theta}(\mu)$$
$$V^{\pi}(\mu) \coloneqq E_{s_0\sim\mu}[V^{\pi}(s_0)]$$

$$Pr_{\mu}^{\pi}(\tau) = \mu(s_0)\pi(a_0|s_0)P(s_1|s_0,a_0)\pi(a_1|s_1)\ \ldots$$
$$R(\tau) \coloneqq \sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)$$

برای اثبات عبارت خواسته شده از policy gradient theorem استفاده می‌کنیم:

$$\begin{aligned}
\nabla_\theta V^{\pi}(s) &= \nabla_\theta \sum_{a\in A}\pi_\theta(a\,|s)Q^{\pi}(s,a) = \sum_{a\in A}\nabla_\theta\pi_\theta(a\,|s)Q^{\pi}(s,a) + \pi_\theta(a\,|s)\nabla_\theta Q^{\pi}(s,a)\\
&= \sum_{a\in A}\nabla_\theta\pi_\theta(a\,|s)Q^{\pi}(s,a) + \pi_\theta(a\,|s)\left(\nabla_\theta\sum_{s',r}P(s',r\,|s,a)\big(r + V^{\pi}(s')\big)\right)\\
&= \sum_{a\in A}\nabla_\theta\pi_\theta(a\,|s)Q^{\pi}(s,a)\\
&+ \pi_\theta(a\,|s)\left(\sum_{s',r}P(s',r\,|s,a)\nabla_\theta V^{\pi}(s')\right) \qquad P(s',r|s,a) \text{ or } r \text{ is not a func of } \theta\\
&= \sum_{a\in A}\nabla_\theta\pi_\theta(a\,|s)Q^{\pi}(s,a) + \pi_\theta(a\,|s)\left(\sum_{s'}P(s'|s,a)\nabla_\theta V^{\pi}(s')\right)
\end{aligned}$$

دنباله بازدید زیر را در نظر بگیرید و احتمال انتقال از حالت s به حالت x را با پالیسی π_θ بعد از k قدم به فرم زیر نشان می‌دهیم:

$$\rho^{\pi}(s\to x,k)$$

$$s \xrightarrow{a\sim\pi_\theta(.|s)} s' \xrightarrow{a\sim\pi_\theta(.|s')} s'' \xrightarrow{a\sim\pi_\theta(.|s'')} \ldots$$

- در k=0 داریم: $\rho^{\pi}(s\to s,k=0)=0$
- در k=1 ا تمام اقدامات ممکن را بررسی می‌کنیم و احتمالات انتقال به حالت هدف را جمع می‌زنیم:

$$\rho^{\pi}(s\to s',k=1) = \sum_a \pi_\theta(a|s)P(s'|s,a)$$

- پس از k قدم داریم:

$$\rho^{\pi}(s\to x,k+1) = \sum_a \rho^{\pi}(s\to s',k)\rho^{\pi}(s'\to x,1)$$

برای راحتی محاسبات $\phi(s)=\sum_{a\in A}\nabla_\theta\pi_\theta(a\,|s)Q^{\pi}(s,a)$ درنظر می‌گیریم.

$$\nabla_\theta V^\pi(s) = \phi(s) + \sum_a \pi_\theta(a\,|s) \sum_{s'} P(s'|s,a)\nabla_\theta V^\pi(s')$$
$$= \phi(s) + \sum_{s'}\sum_a \pi_\theta(a\,|s)P(s'|s,a)\nabla_\theta V^\pi(s') = \phi(s) + \sum_{s'} \rho^\pi(s\ \to s',1)\nabla_\theta V^\pi(s')$$
$$= \phi(s) + \sum_{s'} \rho^\pi(s\ \to s',1)\left[\phi(s') + \sum_{s''} \rho^\pi(s' \to s'',1)\nabla_\theta V^\pi(s'')\right]$$
$$= \phi(s) + \sum_{s'} \rho^\pi(s\ \to s',1)\,\phi(s') + \sum_{s''} \rho^\pi(s \to s'',2)\nabla_\theta V^\pi(s'') = \cdots$$
$$= \sum_{x\in \mathrm{S}}\sum_{k=0}^{\infty} \rho^\pi(s\ \to x,k)\,\phi(x)$$

حال اگر داشته باشیم $\eta(s) = \sum_{x\in \mathrm{S}}\sum_{k=0}^{\infty} \rho^\pi(s\ \to x,k)$

$$\nabla_\theta V^\pi(s) = \sum_{\mathrm{S}}\sum_{k=0}^{\infty} \rho^\pi(s\ \to x,k)\,\phi(x) = \sum_{\mathrm{S}} \eta(s)\phi(x) = \left(\sum_{\mathrm{S}} \eta(s)\right)\sum_{\mathrm{S}} \frac{\eta(s)}{\sum_{\mathrm{S}} \eta(s)}\phi(x)$$
$$\propto \sum_{\mathrm{S}} \frac{\eta(s)}{\sum_{\mathrm{S}} \eta(s)}\phi(x) = \sum_{\mathrm{S}} d^\pi(s) \sum\nolimits_{a\in A} \nabla_\theta \pi_\theta(a\,|s)Q^\pi(s,a)$$

منبع: اثبات standard policy gradient

حال می‌توان گفت:

$$\nabla_\theta V^\pi(s) = \sum_{\mathrm{S}} d^\pi(s) \sum\nolimits_{a\in A} \nabla_\theta \pi_\theta(a\,|s)Q^\pi(s,a) = \sum_{\mathrm{S}} d^\pi(s) \sum\nolimits_{a\in A} \frac{\pi_\theta(a\,|s)\nabla_\theta\pi_\theta(a\,|s)}{\pi_\theta(a\,|s)}Q^\pi(s,a)$$
$$= \sum_{\mathrm{S}} d^\pi(s) \sum\nolimits_{a\in A} \pi_\theta(a\,|s)\nabla_\theta \log \pi_\theta(a\,|s)\, Q^\pi(s,a)$$
$$= \sum_{\mathrm{S}}\sum_{t=0}^{\infty} \gamma^t E_{s_0\sim\mu}\left[\Pr^{\pi}(s_t = s|s_0)\right] \sum\nolimits_{a\in A} \pi_\theta(a\,|s)\nabla_\theta \log \pi_\theta(a\,|s)\, Q^\pi(s,a)$$
$$= E_{\tau\sim Pr_\mu^{\pi_\theta}}[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta\,(a_t|s_t)Q^{\pi_\theta}(s_t,a_t)]$$

یعنی درواقع می‌توانیم expectation را بر روی مسیرهای $\tau\sim Pr_\mu^{\pi_\theta}$ تحت سیاست رفتار μ بگیریم تا گرادیان value function را با توجه به θ بدست آوریم و همچنین از آنجا که discount factor داریم می‌توان گفت:

$$\nabla_\theta V^{\pi_\theta}(\mu) = E_{\tau\sim Pr_\mu^{\pi_\theta}}[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta\,(a_t|s_t)Q^{\pi_\theta}(s_t,a_t)]$$

بنابراین عبارت خواسته شده ثابت شد.

***

(د) می‌دانیم:

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

با توجه به قسمت‌های قبل داریم:

$$\begin{aligned}\nabla V^{\pi_\theta}(\mu)&=E_{\tau\sim Pr_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t Q^{\pi_\theta}(s_t,a_t)\nabla\log\pi_\theta(a_t|s_t)\right]\\&=E_{s_0\sim\mu}\left[E_{\tau\sim Pr^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t Q^{\pi_\theta}(s_t,a_t)\nabla\log\pi_\theta(a_t|s_t)\right]\right]\\&=E_{s_0\sim\mu}\left[\frac{1}{1-\gamma}E_{s\sim d_{s_0}^{\pi_\theta}}E_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]\right]\\&=\frac{1}{1-\gamma}E_{s_0\sim\mu}\left[E_{s\sim d_{s_0}^{\pi_\theta}}E_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]\right]\\&=\frac{1}{1-\gamma}E_{s\sim d_\mu^{\pi_\theta}}[E_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]\end{aligned}$$

برای عبارت دوم نیز می‌توان گفت:

$$A^{\pi_\theta}(s_t,a_t)=Q^{\pi_\theta}(s_t,a_t)-V^{\pi_\theta}(s_t)$$

$$\begin{aligned}E_{a\sim\pi_\theta(\cdot|s)}[A^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]&=E_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)-V^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]\\&=E_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]-V^{\pi_\theta}(s)\underbrace{E_{a\sim\pi_\theta(\cdot|s)}[\nabla\log\pi_\theta(a|s)]}_{=\nabla E_{a\sim\pi_\theta(\cdot|s)}[1]=\nabla 1=0}\\&=E_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]\end{aligned}$$

$$\begin{aligned}\Rightarrow\nabla V^{\pi_\theta}(\mu)&=\frac{1}{1-\gamma}E_{s\sim d^{\pi_\theta}}E_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]\\&=\frac{1}{1-\gamma}E_{s\sim d_\mu^{\pi_\theta}}E_{a\sim\pi_\theta(\cdot|s)}[A^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)]\end{aligned}$$

بنابراین هردو عبارت خواسته شده اثبات می‌شوند.

سوال ۲-الگوریتم های مبتنی بر ارزش برای مسائل با فعالیت های پیوسته

آ) در فضای پیوسته الگوریتم‌هایی مانند Policy iteration و Q-learning به دو مشکل اساسی برخورد می‌کنند.

- **بعد زیاد فضای state**: در فضاهای پیوسته، تعداد حالت های ممکن معمولاً بسیار زیاد و حتی بی‌نهایت است. این امر نشان دادن پالیسی فانکشن و ولیو فانکشن را به طور صریح، همانطور که توسط پالیسی ایتریشن و Q لرنینگ ضروری است، دشوار می کند.
- **exploration problem**: در فضاهای پیوسته، کاوش در فضای حالت به اندازه کافی برای یادگیری یک تابع دقیق value function یا policy function، اغلب دشوار است. برخلاف فضاهای گسسته که امکان برشمردن همه حالت های ممکن به طور کامل وجود دارد، در فضاهای پیوسته، امکان بازدید از هر حالت ممکن وجود ندارد. این امر یادگیری پالیسی یا value function بهینه را بدون گیر کردن در یک راه حل غیربهینه چالش برانگیز می کند.

❄❄❄

ب)

۱)

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

$$\Rightarrow \frac{\partial Q_\phi(s,a)}{\partial a} = -\frac{1}{2}\left(a-\mu_\phi(s)\right)^T P_\phi(s)^T - \frac{1}{2}\left(a-\mu_\phi(s)\right)^T P_\phi(s) = 0$$

$$\Rightarrow -\frac{1}{2}\left(a-\mu_\phi(s)\right)^T \left(P_\phi(s)+P_\phi(s)^T\right) = 0 \Rightarrow a^* = \mu_\phi(s)$$

بنابراین، حداکثر مقدار $Q_\phi(s,a)$ زمانی به دست می آید که a اکشنی باشد که تابع مقدار میانگین $\mu_\phi(s)$ را به حداکثر می‌رساند، که عمل بهینه برای state داده شده‌ی s است.

$$\arg\max_a Q_\phi(s,a) = \mu_\phi(s)$$
$$\max_a Q_\phi(s,a) = Q_\phi\left(s,\mu_\phi(s)\right) = V_\phi(s)$$

معادله ای که ارائه شده نمونه ای از تقریب درجه دوم تابع Q است که در آن تابع Q با مجموعه پارامترها ϕ پارامتر می شود.

مزایا:

- تقریب درجه دوم می تواند در تنظیمات خاصی مفید باشد که در آن تابع Q واقعی برای مدل سازی دقیق بسیار پیچیده است، اما تقریب درجه دوم تقریب کافی برای تصمیم گیری فراهم می کند.
- این پارامتر می تواند منجر به همگرایی سریعتر و تعمیم بهتر نسبت به سایر انواع تقریبگرهای تابع شود.

معایب:

- تقریب درجه دوم فقط در یک منطقه محلی کوچک در اطراف state و action فعلی معتبر است، و ممکن است به خوبی به سایر بخش‌های فضای state-action تعمیم ندهد.
- انتخاب فرم مناسب برای تقریب درجه دوم می تواند دشوار باشد و ممکن است به دانش خاص دامنه نیاز داشته باشد.

- مسئله بهینه سازی برای یافتن پارامترهای بهینه ϕ می تواند از نظر محاسباتی گران باشد و ممکن است به روش های عددی برای حل نیاز داشته باشد.

به طور کلی، اینکه آیا این پارامترسازی تابع Q مناسب است یا خیر، بستگی به مشکل خاص در دست و مبادله بین دقت، تعمیم و کارایی محاسباتی دارد.

۲)

(آ) **الگوریتم:** الگوریتم DDPG یک نسخه اکشن پیوسته از الگوریتم DQN (شبکه Q عمیق) است. از replay experience استفاده می‌کند، جایی که عامل مجموعه‌ای از انتقال‌ها (state, action, reward, next state) و نمونه‌هایی را به‌طور تصادفی از این حافظه برای به‌روزرسانی شبکه‌های بازیگر و منتقد ذخیره می‌کند. شبکه بازیگر با استفاده از پالیسی گرادیان با شبکه منتقد به عنوان baseline به روز می‌شود. شبکه منتقد با استفاده از معادله بلمن و Q-value هدف به روز می شود که با استفاده از شبکه های هدف برای هر دو شبکه بازیگر و منتقد به دست می‌آید.

Algorithm 1 DDPG algorithm

Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer R
for episode = 1, M **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for t = 1, T **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in R
    Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
    Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$
    Update the actor policy using the sampled policy gradient:

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

    Update the target networks:

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

  end for
end for

Taken from *"Continuous Control With Deep Reinforcement Learning"* (Lillicrap et al, 2015)

تابع زیان: تابع هزینه استفاده شده در الگوریتم DDPG از دو بخش تشکیل شده است. بخش اول خطا بین مقدار Q پیش بینی شده و مقدار Q هدف است. بخش دوم، میانگین منفی مقدار Q پیش بینی شده توسط شبکه منتقد نسبت به عملکرد پیش بینی شده توسط شبکه بازیگر است. تابع هزینه کل از میانگین مربعات جمع این دو بخش تشکیل شده است.

$$Loss = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$$

معماری: شبکه های عصبی عمیق معمولاً برای شبکه های بازیگر (actor) و منتقد (critic) استفاده می‌شوند تا بتوانند با فضای وضعیت و عمل بلند مدت سازگار شوند. شبکه بازیگر وظیفه مپ کردن وضعیت به عملکرد را دارد در حالی که شبکه منتقد، تابع ارزش را برای جفت وضعیت-عمل فعلی ارزیابی می کند.

$$\begin{aligned} \theta^Q &: \text{Q network} \\ \theta^\mu &: \text{Deterministic policy function} \\ \theta^{Q'} &: \text{target Q network} \\ \theta^{\mu'} &: \text{target policy network} \end{aligned}$$

شبکه Q و شبکه پالیسی بسیار شبیه Advantage Actor-Critic ساده است، اما در DDPG، Actor به جای خروجی دادن توزیع احتمال در یک فضای عمل گسسته، مستقیماً حالت‌ها را به اقدامات (خروجی شبکه مستقیماً خروجی) نگاشت می‌کند.

شبکه‌های target کپی‌هایی با تاخیر زمانی از شبکه‌های اصلی خود هستند که به آرامی شبکه های آموخته شده را ردیابی می‌کنند. استفاده از این شبکه های ارزش هدف تا حد زیادی ثبات در یادگیری را بهبود می‌بخشد. به این دلیل است: در روش‌هایی که از شبکه‌های هدف استفاده نمی‌کنند، معادلات به‌روزرسانی شبکه به مقادیر محاسبه‌شده توسط خود شبکه وابسته است، که آن را مستعد واگرایی می‌کند.

***

(ب) هر دو الگوریتم DDPG (Deep Deterministic Policy Gradient) و REINFORCE در یادگیری تقویتی استفاده می شوند که شامل تعامل یک عامل با محیط برای یادگیری رفتار بهینه از طریق آزمون و خطا است.

DDPG از دو شبکه عصبی استفاده می کند: یک شبکه actor که اقدام بهینه را در یک حالت مشخص پیش بینی می کند و یک شبکه critic که پاداش مورد انتظار انجام آن اقدام را تخمین می‌زند. شبکه actor برای به حداکثر رساندن پاداش مورد انتظار همانطور که توسط شبکه critic تخمین زده شده است، با استفاده از شکلی از gradient ascent آموزش دیده است.

از سوی دیگر، الگوریتم REINFORCE یک روش policy-based reinforcement learning است که مستقیماً تابع policy را می‌آموزد که حالت ها را به اقدامات نگاشت می‌کند، بدون اینکه یک value function را تخمین بزند. از یک شبکه عصبی منفرد، معروف به شبکه policy، برای خروجی توزیع احتمال بر روی اقداماتی که یک حالت داده شده است، استفاده می‌کند و سپس پارامترهای شبکه policy را به روز می‌کند تا بازده مورد انتظار سیاست را افزایش دهد.

به طور خلاصه، تفاوت اصلی بین دو الگوریتم این است که DDPG از هر دو شبکه actor و critic برای تخمین عمل بهینه و پاداش مورد انتظار استفاده می‌کند، در حالی که REINFORCE مستقیماً تابع policy را بدون استفاده از شبکه critic یاد می گیرد.

تقاطع گرادیان شبکه critic به استفاده از گرادیان خروجی شبکه critic با توجه به ورودی آن (یعنی اقدام انجام شده توسط شبکه actor) برای آموزش شبکه actor اشاره دارد. این تکنیک معمولاً در الگوریتم های یادگیری تقویتی مانند DDPG و Advantage Actor-Critic استفاده می شود، جایی که شبکه actor و شبکه critic برای یادگیری سیاستی که پاداش مورد انتظار را به حداکثر می‌رساند، با هم کار می‌کنند.

یکی از مزیت‌های تقاطع گرادیان این است که راه پایدارتر و کارآمدتری برای آموزش شبکه actor ارائه می‌دهد. در روش‌های پالیسی گرادیان سنتی، گرادیان مستقیماً از پاداش مورد انتظار محاسبه می‌شود که می‌تواند نویزدار باشد و منجر به واریانس بالا شود. با استفاده از گرادیان خروجی شبکه بحرانی، که نشان‌دهنده مقدار مورد انتظار جفت state-action فعلی است، واریانس گرادیان کاهش می یابد و منجر به به روز رسانی های پایدارتر می شود.

مزیت دیگر تقاطع گرادیان این است که به شبکه actor اجازه می دهد تا از تخمین‌های شبکه critic از value function بیاموزد، که می‌تواند اطلاعات ارزشمندی در مورد کیفیت اقدامات انتخاب شده ارائه دهد. این به شبکه actor کمک می‌کند تا سیاستی را بیاموزد که نه تنها پاداش مورد انتظار را به حداکثر می‌رساند، بلکه پاداش های مورد انتظار آینده را نیز در نظر می‌گیرد.

به طور کلی، تقاطع گرادیان از شبکه انتقادی یک راه قدرتمند و موثر برای آموزش شبکه بازیگر در الگوریتم‌های یادگیری تقویتی فراهم می‌کند که منجر به یادگیری پایدارتر و کارآمدتر سیاست‌های بهینه می‌شود.

***

(ج) علاوه بر پیوستگی توابع گفته شده فرض شده‌است که:

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

داریم:

$$\begin{aligned}
\nabla_\theta V^{\mu_\theta}(s) &= \nabla_\theta Q^{\mu_\theta}\big(s, \mu_\theta(s)\big) = \nabla_\theta \left(r\big(s, \mu_\theta(s)\big) + \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) V^{\mu_\theta}(s') ds' \right) \\
&= \nabla_\theta\, \mu_\theta(s) \nabla_a r(s,a)\big|_{a=\mu_\theta(s)} + \nabla_\theta \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) V^{\mu_\theta}(s') ds' \\
&= \nabla_\theta\, \mu_\theta(s) \nabla_a r(s,a)\big|_{a=\mu_\theta(s)} \\
&+ \int_S \gamma \big[p\big(s' \big| s, \mu_\theta(s)\big) \nabla_\theta V^{\mu_\theta}(s') ds' \\
&+ V^{\mu_\theta}(s') \nabla_\theta \mu_\theta(s) \nabla_a p(s'|s,a)\big|_{a=\mu_\theta(s)} \big] ds' \qquad \text{(Leibniz integral rule)} \\
&= \nabla_\theta\, \mu_\theta(s) \nabla_a r(s,a)\big|_{a=\mu_\theta(s)} + \nabla_\theta\, \mu_\theta(s) \nabla_a \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) V^{\mu_\theta}(s') ds' \,\big|_{a=\mu_\theta(s)} \\
&+ \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) \nabla_\theta V^{\mu_\theta}(s') ds' \\
&= \nabla_\theta\, \mu_\theta(s) \nabla_a \left(r(s,a) + \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) V^{\mu_\theta}(s') ds' \right) \Big|_{a=\mu_\theta(s)} \\
&+ \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) \nabla_\theta V^{\mu_\theta}(s') ds' \\
&= \nabla_\theta\, \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)\big|_{a=\mu_\theta(s)} + \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) \nabla_\theta V^{\mu_\theta}(s') ds'
\end{aligned}$$

حال می‌توان با جایگذاری همین عبارت به جای $\nabla_\theta V^{\mu_\theta}(s')$ گفت:

$$
\begin{aligned}
\nabla_\theta V^{\mu_\theta}(s) = {} & \nabla_\theta\, \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \\
& + \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) \bigg[\nabla_\theta\, \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} \\
& + \int_S \gamma p\big(s'' \big| s', \mu_\theta(s')\big) \nabla_\theta V^{\mu_\theta}(s'') ds'' \bigg] ds' \\
& = \nabla_\theta\, \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) \nabla_\theta\, \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' \\
& + \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) \left(\int_S \gamma p\big(s'' \big| s', \mu_\theta(s')\big) \nabla_\theta V^{\mu_\theta}(s'') ds'' \right) ds' \\
& = \nabla_\theta\, \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p\big(s' \big| s, \mu_\theta(s)\big) \nabla_\theta\, \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' \\
& + \int_S \gamma^2 p\big(s_{t+2} = s'' \big| s_t = s, \mu_\theta(s)\big) \nabla_\theta V^{\mu_\theta}(s') ds' \qquad (sum\ over\ probability)
\end{aligned}
$$

با جایگذاری $\nabla_\theta V^{\mu_\theta}(s')$ به همین شکل داریم:

$$
\nabla_\theta V^{\mu_\theta}(s) = \int_S \sum_{i=0}^{T} \gamma^i p\big(s_{t+i} = s' \big| s_t = s, \mu_\theta(s)\big) \nabla_\theta\, \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds'
$$

به همین ترتیب رابطه بازگشتی بدست آمد.

حال داریم:

$$
\begin{aligned}
\nabla_\theta J(\mu_\theta) = \nabla_\theta \int_S p_1(s) V^{\mu_\theta}(s) = {} & \int_S p_1(s) \nabla_\theta V^{\mu_\theta}(s) \\
& = \int_S p_1(s) \int_S \sum_{i=0}^{T} \gamma^i p\big(s_{t+i} = s' \big| s_t = s, \mu_\theta(s)\big) \nabla_\theta\, \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' \\
& = \int_S \rho^{\mu_\theta} \nabla_\theta\, \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\, ds = E_{s\sim\rho^{\mu_\theta}} \left[\nabla_\theta\, \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} \right]
\end{aligned}
$$

بنابراین عبارت خواسته شده اثبات می‌شود.

***

سوال ۳-روش بهینه سازی جدید با تغییر Trust Region

(آ) اگر π و $\tilde{\pi}$ دو سیاست دلخواه باشند طبق اسلاید lecture 10-10 برای محاسبه کردن $J(\tilde{\pi}) - J(\pi)$ داریم:

$$
\begin{aligned}
J(\tilde{\pi}) - J(\pi) &= J(\tilde{\pi}) - E_\rho[V^\pi(s_0)] = J(\tilde{\pi}) - E_{\rho,\tilde{\pi}}[V^\pi(s_0)] = J(\tilde{\pi}) - E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t V^\pi(s_t) - \sum_{t=1}^{\infty} \gamma^t V^\pi(s_t)\right] \\
&= J(\tilde{\pi}) + E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \big(\gamma V^\pi(s_{t+1}) - V^\pi(s_t)\big)\right] \\
&= E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] + E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \big(\gamma V^\pi(s_{t+1}) - V^\pi(s_t)\big)\right] \\
&= E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t, a_t) + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)\big)\right] = E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A^\pi(s_t, a_t)\right] \\
&= \frac{1}{1-\gamma} E_{\rho,\tilde{\pi}}[A^\pi(s_t, a_t)]
\end{aligned}
$$

حال برای رسیدن به خواسته سوال داریم:

$$A(s,a) = Q(s,a) - V(s)$$

$$J(\tilde{\pi}) - J(\pi) = E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A^\pi(s_t, a_t)\right] = E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t Q^\pi(s_t, a_t)\right] - E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t V^\pi(s_t)\right]$$

همچنین می‌دانیم (طبق state-action value function و state-value function):

$$
\begin{aligned}
E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t V(s_t)\right] &= \frac{1}{1-\gamma} \int_S V(s) d\rho_{\tilde{\pi}}(s) \\
&= \frac{1}{1-\gamma} \int_S \int_A V(s) d\rho_{\tilde{\pi}}(s)\, d\tilde{\pi}(a|s) \qquad (sum\ of\ probability)
\end{aligned}
$$

$$E_{\rho,\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t Q(s_t, a_t)\right] = \frac{1}{1-\gamma} \int_S \int_A Q(s,a) d\rho_\pi(s)\, d\tilde{\pi}(a|s)$$

حال می‌توان گفت:

$$
\begin{aligned}
J(\tilde{\pi}) - J(\pi) &= \frac{1}{1-\gamma} E_{\rho,\tilde{\pi}}\,[A(s,a)] = \frac{1}{1-\gamma} \int_S \int_A \big(Q^\pi(s,a) - V^\pi(s)\big) d\rho_{\tilde{\pi}}(s)\, d\tilde{\pi}(a|s) \\
&= \frac{1}{1-\gamma} \int_S \int_A A^\pi(s,a) d\rho_{\tilde{\pi}}(s)\, d\tilde{\pi}(a|s)
\end{aligned}
$$

***

ب) مسئله به فرم زیر است:

$$\max_{\tilde{\pi}\in\Pi}\left\{\int_S\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)d\rho_{\pi}(s):\int_S C\big(\pi(.|s),\tilde{\pi}(.|s)\big)d\rho_{\pi}(s)\le\varepsilon\right\}$$
$$=\min_{\tilde{\pi}\in\Pi}\left\{-\int_S\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)d\rho_{\pi}(s)+\lambda\left(\int_S C\big(\pi(.|s),\tilde{\pi}(.|s)\big)d\rho_{\pi}(s)-\varepsilon\right)\right\}$$

که در آن تابع لاگرانژیان به فرم زیر است:

$$L=-\int_S\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)d\rho_{\pi}(s)+\lambda\left(\int_S C\big(\pi(.|s),\tilde{\pi}(.|s)\big)d\rho_{\pi}(s)-\varepsilon\right)$$

فرم مسئله dual برابر است با:

$$\max_{\tilde{\pi}\in\Pi}\left\{\int_S\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)d\rho_{\pi}(s):\int_S C\big(\pi(.|s),\tilde{\pi}(.|s)\big)d\rho_{\pi}(s)\le\varepsilon\right\}$$
$$=\min_{\lambda>0}\left\{\lambda\varepsilon+\int_S\int_A\max_{a'\in A}\{A^{\pi}(s,a')-\lambda c(a,a')\}d\pi(a|s)d\rho_{\pi}(s)\right\}$$

ابتدا فرم dual مسئله را می‌نویسیم:

$$\begin{aligned}\sup_{\tilde{\pi}\in\Pi}\inf_{\lambda>0}\int_S\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)d\rho_{\pi}(s)+\lambda\left(\varepsilon-\int_S C\big(\pi(.|s),\tilde{\pi}(.|s)\big)d\rho_{\pi}(s)\right)&\\
\le\inf_{\lambda>0}\left\{\lambda\varepsilon+\sup_{\tilde{\pi}\in\Pi}\left\{\int_S\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)-\lambda C\big(\pi(.|s),\tilde{\pi}(.|s)\big)d\rho_{\pi}(s)\right\}\right\}&\qquad(1)\\
\le\inf_{\lambda>0}\left\{\lambda\varepsilon+\int_S\sup_{\tilde{\pi}(.|s)\in\Pi}\left\{\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)-\lambda C\big(\pi(.|s),\tilde{\pi}(.|s)\big)\right\}d\rho_{\pi}(s)\right\}&\end{aligned}$$

(۱) اینفیمم سوپریمم یک مجموعه کمترمساوی سوپریمم اینفیمم آن است.

حال با استفاده از kantrovich duality داریم:

$$\sup_{\tilde{\pi}(.|s)\in\Pi}\left\{\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)-\lambda C\big(\pi(.|s),\tilde{\pi}(.|s)\big)\right\}$$
$$=\sup_{\tilde{\pi}(.|s)\in\Pi}\left\{\int_A A^{\pi}(s,a)d\tilde{\pi}(a|s)-\lambda\sup_{\phi,\psi\le c}\left\{\int_A\phi(a)d\pi(a|s)+\int_A\psi(a)d\tilde{\pi}(a|s)\right\}\right\}$$

فرض می‌کنیم λ مثبت است. در این صورت برای انتخاب ϕ,ψ داریم:

$$\psi(a)=\frac{A^{\pi}(s,a)}{\lambda},\qquad\phi(a)=\inf_{a'\in A}\{c(a,a')-\psi(a')\}$$

در این صورت

$$\phi(a_1) + \psi(a_2) = \inf_{a' \in A}\{c(a_1, a') - \psi(a')\} + \psi(a_2) \le c(a_1, a_2) - \psi(a_2) + \psi(a_2) \le c(a_1, a_2)$$

که شرط $\phi(a_1) + \psi(a_2) \le c(a_1, a_2)$ برقرار است.

پس می‌توان گفت:

$$\begin{aligned}
&\int_A \inf_{a' \in A}\left\{c(a, a') - \frac{A^{\pi}(s, a')}{\lambda}\right\} d\pi(a|s) + \int_A \frac{A^{\pi}(s, a)}{\lambda} d\tilde{\pi}(a|s) \le C\big(\pi(.\,|s), \tilde{\pi}(.\,|s)\big) \\
&\qquad \Rightarrow \sup_{\tilde{\pi}(.|\mathrm{S}) \in \Pi}\left\{\int_A A^{\pi}(s, a) d\tilde{\pi}(a|s) - \lambda C\big(\pi(.\,|s), \tilde{\pi}(.\,|s)\big)\right\} \\
&\qquad \le \sup_{\tilde{\pi}(.|\mathrm{S}) \in \Pi}\left\{\int_A - \inf_{a' \in A}\{\lambda c(a, a') - A^{\pi}(s, a')\}\, d\pi(a|s)\right\} \\
&\qquad = \sup_{\tilde{\pi}(.|\mathrm{S}) \in \Pi}\left\{\int_A \sup_{a' \in A}\{A^{\pi}(s, a') - \lambda c(a, a')\}\, d\pi(a|s)\right\} \\
&\qquad = \int_A \sup_{a' \in A}\{A^{\pi}(s, a') - \lambda c(a, a')\}\, d\pi(a|s)
\end{aligned}$$

پس در نهایت داریم:

$$\begin{aligned}
&\sup_{\tilde{\pi} \in \Pi} \inf_{\lambda > 0} \int_S \int_A A^{\pi}(s, a) d\tilde{\pi}(a|s) d\rho_{\pi}(s) + \lambda\left(\varepsilon - \int_S C\big(\pi(.\,|s), \tilde{\pi}(.\,|s)\big) d\rho_{\pi}(s)\right) \\
&\qquad \le \inf_{\lambda > 0}\left\{\lambda\varepsilon + \int_S \int_A \sup_{a' \in A}\{A^{\pi}(s, a') - \lambda c(a, a')\}\, d\pi(a|s)\, d\rho_{\pi}(s)\right\}
\end{aligned}$$