به نام خدا

یادگیری تقویتی

تمرین تئوری سوم

یلدا شعبان زاده

۹۸۱۰۱۸۲۲

نیمسال دوم ۰۲-۰۱

فهرست

سوال ۱- LQR

(آ)

LQR یک الگوریتم کنترل بهینه شناخته شده است که می‌تواند در مسائل یادگیری تقویتی (RL) استفاده شود. همگرایی LQR به دینامیک سیستم و تابع هزینه بستگی دارد. به طور کلی، LQR زمانی همگرا می شود که سیستم قابل کنترل و پایدار باشد. نرخ همگرایی LQR توسط مقادیر ویژه ماتریس سیستم A و ماتریس تابع هزینه Q تعیین می شود. اگر همه مقادیر ویژه A منفی و Q مثبت معین (positive definite) باشد، LQR به صورت نمایی سریع همگرا می‌شود.

شرایط ضروری برای همگرایی LQR در مسائل یادگیری تقویتی عبارت است از:

- محیط باید خطی و time-invariant باشد.
- Agent باید یک مدل کامل از محیط داشته باشد: LQR یک روش مبتنی بر مدل است، به این معنی که برای همگرایی به سیاست بهینه نیاز به یک مدل کامل از محیط دارد. Agent باید بتواند بر اساس مشاهدات خود از محیط، نتیجه اقدامات خود را به دقت پیش‌بینی کند.
- محیط باید fully observable باشد: Agent در هر مرحله زمانی به تمام اطلاعات مربوطه در مورد وضعیت فعلی محیط دسترسی داشته باشد.
- فضای state و action باید متناهی و پیوسته باشد.
- تابع هزینه باید quadratic باشد.

اگر موارد بالا صورت نگیرد LQR ممکن است به همگرایی نرسد یا به یک خطای زیر بهینه همگرا شود.

(منبع)

***

(ب)

رویکردهایی برای یادگیری تقویتی در محیط های partially observable وجود دارد. به عنوان مثال، یک رویکرد در مقاله‌ی

Reinforcement Learning under Partial Observability Guided by Learned Environment Models

آمده است که فرض می‌کند محیط مانند یک فرآیند تصمیم گیری مارکوف partially observable با اقدامات گسسته شناخته شده و بدون دانش در مورد ساختار یا احتمالات انتقال آن رفتار می‌کند.

در این رویکرد Q-learning را با IoAlergia، روشی برای یادگیری فرآیندهای تصمیم مارکوف (MDP) ترکیب می‌کند. با یادگیری مدل‌های MDP محیط از قسمت‌های RL agent، RL را در حوزه‌های partially observable بدون حافظه صریح و اضافی فعال می‌کند تا تعاملات قبلی را برای مقابله با ابهامات ناشی از partial observability ردیابی کند. در این روش در عوض با شبیه‌سازی تجربیات جدید در مدل‌های محیط آموخته شده برای ردیابی حالات کاوش شده، مشاهدات اضافی را در قالب حالت های محیطی انتزاعی به RL ارائه می‌کند.

(منبع)

رویکرد دیگر استفاده از روش Kalman filter است که یک تخمینگر حالت است که می تواند وضعیت فعلی یک سیستم خطی را بر اساس اندازه‌گیری های نویز تخمین بزند. در زمینه یادگیری تقویتی، از فیلتر کالمن می توان برای تخمین وضعیت فعلی محیط بر اساس مشاهدات عامل استفاده کرد که سپس می تواند به عنوان ورودی کنترل کننده LQR استفاده شود.

ترکیب روش LQR و فیلتر کالمن به عنوان Linear Quadratic Gaussian (LQG) شناخته می شود. کنترل LQG از دو بخش تشکیل شده است: state estimator و کنترل‌کننده. State estimator از فیلتر کالمن برای تخمین وضعیت سیستم بر اساس اندازه‌گیری‌های نویز استفاده می‌کند. کنترل‌کننده از حالت تخمین زده شده برای محاسبه ورودی کنترل بهینه با استفاده از روش LQR استفاده می‌کند.

(منبع)

***

(ج)

روش LQR و روش‌های مدل رایگان مبتنی بر شبکه عصبی عمیق، رویکردهای متفاوتی برای کنترل و یادگیری تقویتی هستند. روش LQR یک رویکرد مبتنی بر مدل است که نیاز به یک مدل کامل و دقیق از محیط دارد، در حالی که روش‌های model-free مبتنی بر شبکه عصبی عمیق، رویکردهای بدون مدل هستند که مستقیماً یک پالیسی یا تابع ارزش را از تجربه یاد می‌گیرند.

برای ترکیب این دو داریم:

- Initialize کردن deep networks: یک رویکرد استفاده از روش LQR برای مقداردهی اولیه شبکه عصبی عمیق و سپس استفاده از روش های model-free برای تنظیم دقیق پالیسی است.
- یادگیری مدل از داده‌ها: شبکه‌های عصبی عمیق را می‌توان برای یادگیری مدلی از محیط از داده‌ها استفاده کرد، که سپس می‌توان از روش LQR برای محاسبه یک سیاست کنترل بهینه استفاده کرد. این رویکرد زمانی می‌تواند مفید باشد که محیط پیچیده یا غیرخطی باشد و یک مدل تحلیلی ساده در دسترس نباشد.
- ترکیب تابع هزینه LQR در الگوریتم های RL عمیق: این رویکرد زمانی می‌تواند مفید باشد که محیط partially observable باشد یا زمانی که روش LQR برای استفاده مستقیم از نظر محاسباتی گران است.

***

(د)

روش iLQR یک الگوریتم کنترل model-based است که می تواند برای یافتن یک سیاست کنترل محلی بهینه برای سیستم های دارای عدم قطعیت استفاده شود. با این حال، به طور مستقیم به مشکل exploration، که چالش کشف سیاست بهینه در یک محیط ناشناخته است، نمی‌پردازد.

برای رسیدگی به عدم قطعیت در محیط، الگوریتم iLQR را می توان به چند روش مختلف بهبود داد:

- Stochastic iLQR: یکی از راه‌های محاسبه عدم قطعیت در سیستم، وارد کردن تصادفی به الگوریتم iLQR است. این را می توان با ترکیب یک مدل احتمالی از دینامیک سیستم، مانند یک فرآیند **گاوسی** یا یک **شبکه عصبی بیزی** انجام داد. سپس می‌توان الگوریتم iLQRرا برای ایجاد توزیعی از سیاست‌های کنترلی تغییر داد که عدم قطعیت در سیستم را توضیح می‌دهد.
- Online iLQR: در برخی موارد، ممکن است با در دسترس قرار گرفتن اطلاعات جدید در مورد سیستم، لازم باشد سیاست کنترل را در زمان واقعی تطبیق دهید. این را می توان با استفاده از یک الگوریتم iLQR آنلاین انجام داد، که در آن پالیسی کنترل بر اساس اندازه گیری های جدید وضعیت سیستم به روز می شود.
- Adaptive iLQR: رویکرد دیگر استفاده از الگوریتم تطبیقی iLQR است که در آن مدل سیستم با در دسترس قرار گرفتن اطلاعات جدید به روز می شود. این را می توان با ترکیب یک الگوریتم تخمین مدل تطبیقی، مانند یک **فیلتر کالمن** یا یک الگوریتم **رگرسیون فرآیند گاوسی آنلاین** انجام داد.

برای پرداختن به مشکل exploration می‌توان گفت:

- Random exploration: یک استراتژی اکتشافی ساده اضافه کردن اغتشاشات تصادفی به ورودی های کنترلی تولید شده توسط الگوریتم iLQR است. این می تواند عامل را تشویق کند تا قسمت های مختلف محیط را کاوش کند و از گیر افتادن در بهینه محلی جلوگیری کند.
- Ensemble iLQR: Ensemble iLQR نوعی از الگوریتم iLQR است که از مجموعه ای از مدل‌ها برای نشان دادن عدم قطعیت محیط استفاده می‌کند. این رویکرد می‌تواند برای مدل‌سازی صریح عدم قطعیت و ایجاد توزیعی از سیاست‌های کنترلی استفاده شود که می‌تواند برای exploration و robustness استفاده شود.

سوال ۲- بازی اتاق فرار جایزه دار

آ)

$$r \sim Gamma(\alpha,\beta) \rightarrow p(r|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}$$

$$t \sim Exp(\lambda) \rightarrow p(t|\lambda) = \lambda e^{-\lambda t}$$

آلفا مشخص و بتا نامعلوم است.

$$\beta \sim Gamma(\epsilon,\omega)$$
$$\lambda \sim Gamma(\sigma,\eta)$$

$$\begin{aligned}
p(\beta|r_1,\alpha,\epsilon,\omega) &= \frac{p(\beta,r_1|\alpha,\epsilon,\omega)}{p(r_1|\alpha,\epsilon,\omega)} = \frac{p(\beta|\alpha,\epsilon,\omega)p(r_1|\beta,\alpha,\epsilon,\omega)}{p(r_1|\alpha,\epsilon,\omega)} = \frac{p(\beta|\alpha,\epsilon,\omega)p(r_1|\beta,\alpha,\epsilon,\omega)}{\int_\beta p(r_1,\beta|\alpha,\epsilon,\omega)} \\
&= \frac{p(\beta|\epsilon,\omega)p(r_1|\alpha,\beta)}{\int_\beta p(r_1,\beta|\alpha,\epsilon,\omega)} \\
&= \frac{\omega^\epsilon}{\Gamma(\epsilon)}\beta^{\epsilon-1}e^{-\omega\beta}\frac{\beta^\alpha}{\Gamma(\alpha)}r_1^{\alpha-1}e^{-\beta r_1} \times \frac{1}{\int_\beta \frac{\omega^\epsilon}{\Gamma(\epsilon)}\beta^{\epsilon-1}e^{-\omega\beta}\frac{\beta^\alpha}{\Gamma(\alpha)}r_1^{\alpha-1}e^{-\beta r_1}} \\
&= \frac{\omega^\epsilon}{\Gamma(\epsilon)}\beta^{\epsilon-1}e^{-\omega r}\frac{\beta^\alpha}{\Gamma(\alpha)}r_1^{\alpha-1}e^{-\beta r_1} \times \frac{1}{\frac{\omega^\epsilon}{\Gamma(\epsilon)\Gamma(\alpha)}r_1^{\alpha-1}\int_\beta \beta^{\alpha+\epsilon-1}e^{-\beta(r_1+\omega)}} \\
&= \frac{\beta^{\alpha+\epsilon-1}e^{-\beta(r_1+\omega)}}{\int_\beta \beta^{\alpha+\epsilon-1}e^{-\beta(r_1+\omega)}} = \frac{\frac{(r_1+\omega)^{\alpha+\epsilon}}{\Gamma(\alpha+\epsilon)}\beta^{\alpha+\epsilon-1}e^{-\beta(r_1+\omega)}}{\int_\beta \frac{(r_1+\omega)^{\alpha+\epsilon}}{\Gamma(\alpha+\epsilon)}\beta^{\alpha+\epsilon-1}e^{-\beta(r_1+\omega)}} \\
&= \frac{(r_1+\omega)^{\alpha+\epsilon}}{\Gamma(\alpha+\epsilon)}\beta^{\alpha+\epsilon-1}e^{-\beta(r_1+\omega)} = p(\beta|\alpha+\epsilon,r_1+\omega) \rightarrow \beta \sim Gamma(\alpha+\epsilon,r_1+\omega) \\
&\rightarrow \begin{cases} \epsilon' = \alpha+\epsilon \\ \omega' = r_1+\omega \end{cases}
\end{aligned}$$

$$\begin{aligned}
p(\lambda|t_1,\sigma,\eta) &= \frac{p(\lambda,t_1|\sigma,\eta)}{p(t_1|\sigma,\eta)} = \frac{p(\lambda|\sigma,\eta)p(t_1|\lambda,\sigma,\eta)}{p(t_1|\sigma,\eta)} = \frac{p(\lambda|\sigma,\eta)p(t_1|\lambda,\sigma,\eta)}{\int_\lambda p(\lambda,t_1|\sigma,\eta)} = \frac{p(\lambda|\sigma,\eta)p(t_1|\lambda)}{\int_\lambda p(\lambda,t_1|\sigma,\eta)} \\
&= \frac{\frac{\eta^\sigma}{\Gamma(\sigma)}\lambda^{\sigma-1}e^{-\eta\lambda}\lambda e^{-\lambda t_1}}{\int_\lambda \frac{\eta^\sigma}{\Gamma(\sigma)}\lambda^{\sigma-1}e^{-\eta\lambda}\lambda e^{-\lambda t_1}} = \frac{\frac{(\eta+t_1)^{\sigma+1}}{\Gamma(\sigma+1)}\lambda^\sigma e^{-\lambda(t_1+\eta)}}{\int_\lambda \frac{(\eta+t_1)^{\sigma+1}}{\Gamma(\sigma+1)}\lambda^\sigma e^{-\lambda(t_1+\eta)}} = \frac{(\eta+t_1)^{\sigma+1}}{\Gamma(\sigma+1)}\lambda^\sigma e^{-\lambda(t_1+\eta)} \\
&= p(\lambda|\sigma+1,t_1+\eta) \rightarrow \lambda \sim Gamma(\sigma+1,t_1+\eta) \rightarrow \begin{cases} \sigma' = \sigma+1 \\ \eta' = t_1+\eta \end{cases}
\end{aligned}$$

***

(ب)

$$p(t_2|t_1) = \int_\lambda p(t_2|\lambda, t_1)p(\lambda|t_1)d\lambda = \int_\lambda p(t_2|\lambda)p(\lambda|t_1)d\lambda = \int_\lambda p(t_2|\lambda)p(\lambda|t_1, \sigma, \eta)d\lambda$$
$$= \int_\lambda p(t_2|\lambda)p(\lambda|\sigma', \eta')d\lambda = \int_0^\infty \lambda e^{-\lambda t_2} \frac{(\eta')^{\sigma'}}{\Gamma(\sigma')} \lambda^{\sigma'-1} e^{-\lambda(\eta')} d\lambda$$
$$= \frac{(\eta')^{\sigma'}}{\Gamma(\sigma')} \underbrace{\int_0^\infty \lambda^{\sigma'} e^{-\lambda(\eta'+t_2)} d\lambda}_{I} = \frac{(\eta')^{\sigma'}}{\Gamma(\sigma')} \frac{\sigma'!}{(\eta'+t_2)^{\sigma'+1}} = \frac{(\eta')^{\sigma'}}{\sigma'!-1} \frac{\sigma'!}{(\eta'+t_2)^{\sigma'+1}}$$
$$= \frac{(\eta')^{\sigma'}\sigma'}{(\eta'+t_2)^{\sigma'+1}} = \frac{(\eta')^{\sigma'}\sigma'}{(\eta')^{1+\sigma'}\left(1+\frac{t_2}{\eta'}\right)^{\sigma'+1}} = \frac{\sigma'}{\eta'}\left(1+\frac{t_2}{\eta'}\right)^{-(\sigma'+1)} = p(t_2|\sigma', \eta')$$
$$\to t_2 \sim Lomax(\sigma', \eta')$$

$$I = \int_0^\infty \lambda^{\sigma'} e^{-\lambda(\eta'+t_2)} d\lambda = \frac{\lambda^{\sigma'} e^{-\lambda(\eta'+t_2)}}{-(\eta'+t_2)} |_0^\infty + \frac{\sigma'}{\eta'+t_2} \int_0^\infty \lambda^{\sigma'-1} e^{-\lambda(\eta'+t_2)} d\lambda$$
$$= 0 + \frac{\sigma'}{\eta'+t_2} \int_0^\infty \lambda^{\sigma'-1} e^{-\lambda(\eta'+t_2)} d\lambda = \frac{\sigma'!}{(\eta'+t_2)^{\sigma'+1}}$$

***

(ج)

ریسک‌گریز: درحالتی که ریسک‌گریز باشد، در بدترین حالت ممکن است زمان بازی infinity شود که در این صورت بهتر است به کار کارمندی خود ادامه دهد. زیرا در این حالت به نظر فرد احتمال رخداد حالت‌های بد بیشتر از حالت‌های خوب است و در اینصورت نسبت به زمان infinity ترسش باعث نمی‌شود ریسک دریافت جایزه‌ی بهتر را بپذیرد.

ریسک‌پذیر: در حالت ریسک‌پذیر ممکن است با یک زمان محدود و یا حتی کم پاداش infinity بگیرد که در آن صورت بهتر است به این بازی بپردازد و کار کارمندی خود را رها کند. زیرا میل به ریسک طلبی باعث می‌شود جایزه‌ی infinity بسیار برایش ارزش بیشتری نسبت به ماندن در بازی در زمان infinity را داشته باشد.

ریسک‌خنثی: در این حالت داریم:

- کارمند:

$$E[tK] = E[t] \times K = \frac{K}{\lambda}$$

- بازیکن:

$$E[r] = E[r] = \frac{\alpha}{E[\beta]} = \frac{\alpha\epsilon}{\omega}$$

حال اگر $\frac{K}{\lambda} > \frac{\alpha\epsilon}{\omega}$ بهتر است که بازی نکند در غیر اینصورت یعنی اکسپکتد سود کارکردن بیشتر خواهد بود.

سوال ۳- بررسی روش گرادیان سیاست در رویکرد soft optimality

(آ) برای بازنویسی عبارت زیر به فرم KL داریم:

$$\begin{aligned}
\log p(O_{1:T}) &\geq \sum_t E_{(s_t,a_t)\sim q}[r(s_t,a_t)] + \sum_t E_{s_t\sim q(s_t)}[\mathcal{H}(q(.|s_t))] \\
&= \sum_t E_{(s_t,a_t)\sim q}[r(s_t,a_t)] + \sum_t E_{s_t\sim q(s_t)}[\mathcal{H}(q(.|s_t))] \\
&= \sum_t E_{(s_t,a_t)\sim q}[r(s_t,a_t)] - \sum_t E_{s_t\sim q(s_t)}\left[E_{a_t\sim q(.|s_t)}[\log q(a_t|s_t)]\right] \\
&= \sum_t E_{(s_t,a_t)\sim q}[r(s_t,a_t)] - \sum_t E_{(s_t,a_t)\sim q}[\log q(a_t|s_t)] \\
&= \sum_t E_{(s_t,a_t)\sim q}[r(s_t,a_t) - \log q(a_t|s_t)] = E_{\tau\sim q(\tau)}[r(s_t,a_t) - \log q(a_t|s_t)] \\
&= E_{\tau\sim q(\tau)}\Bigg[\log \hat{q}(s_1) + \sum_t \log \hat{q}(s_{t+1}|s_t,a_t) + r(s_t,a_t) - \log \hat{q}(s_1) \\
&\quad - \sum_t \log \hat{q}(s_{t+1}|s_t,a_t)\Bigg] = -D_{KL}(q(\tau)||\hat{q}(\tau))
\end{aligned}$$

حال می‌توان گفت باید این backward messageها را از دیدگاه بهینه‌سازی به عنوان یک الگوریتم برنامه نویسی پویا استخراج کنیم. با حالت پایه بهینه سازی $q(a_t|s_t)$ شروع می کنیم که حداکثر می شود.

$$\begin{aligned}
&E_{(s_T,a_T)\sim q}[r(s_T,a_T) - \log q(a_T|s_T)] \\
&\qquad = E_{s_T\sim q(s_T)}[-D_{KL}(q(a_T|s_T)||\exp(r(s_T,a_T) - V(s_T))) + V(s_T)] \quad (1)
\end{aligned}$$

در حالت کمتر از T:

$$E_{(s_t,a_t)\sim q}[r(s_t,a_t) - \log q(a_t|s_t)] + E_{(s_t,a_t)\sim q}\left[E_{s_{t+1}\sim q(s_{t+1}|s_t,a_t)}[V(s_{t+1})]\right] \quad (2)$$

عبارت اول مستقیماً از هدف objective پیروی می کند، در حالی که عبارت دوم سهم π(at|st) را در انتظارات تمام مراحل زمانی بعدی نشان می دهد. ابتدا حالت پایه را در نظر بگیرید: با توجه به معادله (π(aT|sT، می‌توانیم هدف سیاست را با جایگزین کردن مستقیم این معادله با معادله 1 ارزیابی کنیم. از آنجایی که واگرایی KL به صفر می رسد، ما فقط با عبارت V (sT) باقی می‌مانیم. در حالت بازگشتی، توجه می‌کنیم که می‌توانیم هدف را در رابطه (۱۴) به صورت بازنویسی کنیم.

$$V(s_T) = \log \int \exp(r(s_T,a_T))\, da_T$$

$$(2) = E_{s_t\sim q(s_t)}[-D_{KL}(q(a_t|s_t)||\exp(r(s_t,a_t) - V(s_t))) + V(s_t)]$$

حال این وقتی ماکزیمم می‌شود که عبارت D_{KL} صفر بشود.

$$D_{KL}(q(a_t|s_t)||\exp(Q(s_t,a_t) - V(s_t))) = 0 \rightarrow q(a_t|s_t) = \exp(Q(s_t,a_t) - V(s_t))$$

(ب)

$$\nabla_\theta J(\theta) = \sum_{t=1}^{T} \nabla_\theta E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q(\mathbf{s}_t,\mathbf{a}_t)} \left[r(\mathbf{s}_t,\mathbf{a}_t) + \mathcal{H}(q_\theta(\mathbf{a}_t|\mathbf{s}_t)) \right]$$
$$= \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q(\mathbf{s}_t,\mathbf{a}_t)} \left[\nabla_\theta \log q_\theta(\mathbf{a}_t|\mathbf{s}_t) \left(\sum_{t'=t}^{T} r(\mathbf{s}_{t'},\mathbf{a}_{t'}) - \log q_\theta(\mathbf{a}_{t'}|\mathbf{s}_{t'}) - 1 \right) \right]$$
$$= \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q(\mathbf{s}_t,\mathbf{a}_t)} \left[\nabla_\theta \log q_\theta(\mathbf{a}_t|\mathbf{s}_t) \left(\sum_{t'=t}^{T} r(\mathbf{s}_{t'},\mathbf{a}_{t'}) - \log q_\theta(\mathbf{a}_{t'}|\mathbf{s}_{t'}) - b(\mathbf{s}_{t'}) \right) \right]$$

در اینجا خط دوم از likelihood ratio trick و تعریف آنتروپی برای بدست آوردن عبارت $\log q_\theta(a_{t'}|s_{t'})$ پیروی می کند. 1- از مشتق عبارت آنتروپی می‌آید. خط آخر به این دلیل است که برآوردگر گرادیان نسبت به ثابت‌های وابسته به حالت افزایشی ثابت است و ۱- را با یک baseline وابسته به حالت جایگزین می‌کند. تخمین‌گر گرادیان پالیسی حاصل دقیقاً با برآوردگر گرادیان خط‌مشی استاندارد مطابقت دارد.

***

(ج)

$$J(\theta) = \sum_t E_{(s_t,a_t)\sim q}[r(s_t,a_t) - \log \pi(a_t|s_t)]$$

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_i \sum_t \nabla_\theta \log \pi(a_t|s_t) \left(\sum_{t'=t}^{T} [r(s_{t'},a_{t'}) - \log \pi(a_{t'}|s_{t'})] - 1 \right)$$
$$= \frac{1}{N}\sum_i \sum_t \nabla_\theta \log \pi(a_t|s_t) \left(\sum_{t'=t}^{T} [r(s_{t'},a_{t'}) - \log \pi(a_{t'}|s_{t'})] - 1 \right)$$
$$= \frac{1}{N}\sum_i \sum_t \nabla_\theta \log \pi(a_t|s_t) \left(r(s_t,a_t) - \log \pi(a_t|s_t) + \underbrace{\sum_{t'=t+1}^{T} [r(s_{t'},a_{t'}) - \log \pi(a_{t'}|s_{t'})]}_{\approx Q(s_{t+1},a_{t+1})} - 1 \right)$$
$$= \frac{1}{N}\sum_i \sum_t \left(\nabla_\theta Q(s_t,a_t) - \nabla_\theta V(s_t,a_t)\right)\left(r(s_t,a_t) - Q(s_t,a_t) + V(s_t) + Q(s_{t+1},a_{t+1}) - 1\right)$$
$$= \frac{1}{N}\sum_i \sum_t \left(\nabla_\theta Q(s_t,a_t) - \nabla_\theta V(s_t,a_t)\right)\left(r(s_t,a_t) - Q(s_t,a_t) + V(s_t) + Q(s_{t+1},a_{t+1})\right)$$

تساوی آخر به این دلیل است که می‌توان baseline را در نظر نگرفت.

(د)

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_i \sum_t \big(\nabla_\theta Q(s_t, a_t) - \nabla_\theta V(s_t, a_t)\big)\big(r(s_t, a_t) - Q(s_t, a_t) + V(s_t) + Q(s_{t+1}, a_{t+1})\big)$$
$$\approx \frac{1}{N}\sum_i \sum_t \big(\nabla_\theta Q(s_t, a_t) - \nabla_\theta V(s_t, a_t)\big)\big(r(s_t, a_t) - Q(s_t, a_t) + Q(s_{t+1}, a_{t+1})\big)$$

به دلیل اینکه در پالیسی گرادیان نشان دادیم تخمین گرادیان با baseline وابسته به s بدون بایاس است.

می‌دانیم:

$$V(\mathbf{s}_t) = \log \int_{\mathcal{A}} \exp(Q(\mathbf{s}_t, \mathbf{a}_t)) d\mathbf{a}_t$$
$$q(\mathbf{a}_t|\mathbf{s}_t) = \exp(Q(\mathbf{s}_t, \mathbf{a}_t) - V(\mathbf{s}_t))$$

بنابراین می‌توان آپدیت گرادیان در soft Q-learning را به فرم زیر نوشت:

$$\phi \leftarrow \phi - \alpha E\left[\frac{dQ_\phi}{d\phi}(\mathbf{s}_t, \mathbf{a}_t)\left(Q_\phi(\mathbf{s}_t, \mathbf{a}_t) - \left(r(\mathbf{s}_t, \mathbf{a}_t) + \log \int_{\mathcal{A}} \exp(Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})) d\mathbf{a}_{t+1}\right)\right)\right]$$

که بسیار شبیه به استانداردش است:

$$\phi \leftarrow \phi - \alpha E\left[\frac{dQ_\phi}{d\phi}(\mathbf{s}_t, \mathbf{a}_t)\left(Q_\phi(\mathbf{s}_t, \mathbf{a}_t) - \left(r(\mathbf{s}_t, \mathbf{a}_t) + \max_{\mathbf{a}_{t+1}} Q_\phi(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})\right)\right)\right]$$

در مورد actionهای گسسته، پیاده سازی این به روز رسانی ساده است، زیرا انتگرال با یک جمع جایگزین می شود و سیاست را می توان به سادگی با نرمالایز کردن تابع Q بدست آورد.

در مورد actionهای پیوسته، سطح بیشتری از تقریب برای ارزیابی انتگرال با استفاده از نمونه ها مورد نیاز است. نمونه‌برداری از پالیسی ضمنی نیز بی‌اهمیت است و به یک روش استنتاج تقریبی نیاز دارد.