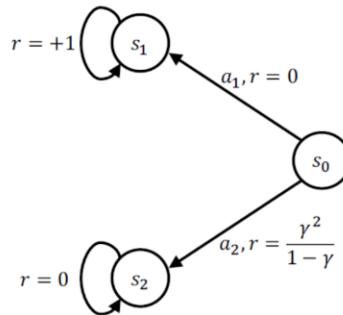


مسئله ۱. (۱۰ نمره) حد همگرایی در Value Iteration

زنگیره مارکوف زیر را در نظر بگیرید. ارزش اولیه تمام حالت‌ها را صفر فرض کنید. برای $1 < \gamma < 0$ به سوالات زیر پاسخ دهید.



(آ) (۱ نمره) عمل بهینه در زمان $t = 0$ در s . کدام است؟ توضیح دهید.

(ب) (۶ نمره) نشان دهید که الگوریتم value iteration پس از مرحله n^* برای ارزش s همگرا می‌شود؛ به طوری که n^* در رابطه زیر صدق می‌کند:

$$n^* \geq \frac{\log(1 - \gamma)}{\log \gamma}$$

(پ) (۳ نمره) با این فرض که اگر تغییرات در ارزش‌ها کمتر از ترشولد θ باشد الگوریتم همگرا می‌شود، حد بالایی برای n^* بر حسب θ پیدا کنید. با این حساب، برای یک γ خاص، کمترین مقدار θ چقدر باشد تا در سریعترین زمان ممکن همگرایی صورت بگیرد؟

(د) در این الگوریتم معمولاً به این صورت عمل می‌شود که در زمان $t=0$ ارزش تمام حالت‌ها با صفر در نظر گرفته می‌شود و احتمالاً عمل بهینه‌ای وجود نداشته باشد ولی پس از یک مرحله آپدیت ارزش حالت‌ها، ارزش حالت‌های s_1 و s_2 طبق محاسبات زیر برابر با ۱ و صفر خواهند شد و بنابراین عمل بهینه در زمان $t=1$ در حالت s_0 برابر با عمل a_1 خواهد بود.

$$\begin{aligned} V(s_1) &= P(s_1, a_1, s_1) [1 + \gamma V(s_1)] = [1 + 0] = 1 \\ V(s_2) &= P(s_2, a_2, s_2) [0 + \gamma V(s_2)] = 0 \end{aligned}$$

(ب) با توجه به اینکه احتمالات عملگرهای ۱ و ۲ داده نشده است، احتمال آنها را برابر با $\frac{1}{2}$ در نظر می‌گیریم و محاسبات به صورت زیر است.

t	0	1	2	3
s_0	0			
s_1	0			
s_2	0			

$$V_1(s_0) = P(s_0, \alpha_1, s_1) [0 + \gamma V(s_1)] + P(s_0, \alpha_2, s_2) \left[\frac{\gamma^2}{1-\gamma} + \gamma V(s_2) \right]$$

$$= \frac{1}{2}[0] + \frac{1}{2} \left[\frac{\gamma^2}{1-\gamma} + 0 \right] = \frac{1}{2} \left(0 + \frac{\gamma^2}{1-\gamma} \right) = \frac{1}{2} \cdot \frac{\gamma^2}{1-\gamma}$$

$$V_1(s_1) = P(s_1, \alpha, s_1) [1 + \gamma V(s_1)] = 1$$

$$V_1(s_2) = P(s_2, \alpha, s_2) [0 + \gamma V(s_2)] = 0$$

$$V_2(s_0) = \frac{1}{2}[0 + \gamma(1)] + \frac{1}{2} \left[\frac{\gamma^2}{1-\gamma} + \gamma(0) \right] = \frac{1}{2}\gamma + \frac{1}{2} \left[\frac{\gamma^2}{1-\gamma} \right] = \frac{1}{2} \left[\gamma + \frac{\gamma^2}{1-\gamma} \right]$$

$$V_2(s_1) = 1(1 + \gamma(1)) = 1 + \gamma$$

$$V_2(s_2) = 1(0 + \gamma(0)) = 0$$

$$V_3(s_0) = \frac{1}{2}(0 + \gamma(1 + \gamma)) + \frac{1}{2} \left[\frac{\gamma^2}{1-\gamma} + \gamma(0) \right] = \frac{1}{2} (\gamma + \gamma^2 + \frac{\gamma^2}{1-\gamma})$$

$$V_3(s_1) = 1(1 + \gamma(1 + \gamma)) = 1 + \gamma + \gamma^2$$

$$V_3(s_2) = 1(0 + \gamma(0)) = 0$$

$$V_4(s_0) = \frac{1}{2}(0 + \gamma(1 + \gamma + \gamma^2)) + \frac{1}{2} \left[\frac{\gamma^2}{1-\gamma} + \gamma(0) \right] = \frac{1}{2} (\gamma + \gamma^2 + \gamma^3 + \frac{\gamma^2}{1-\gamma})$$

$$V_4(s_1) = 1(1 + \gamma(1 + \gamma + \gamma^2)) = 1 + \gamma + \gamma^2 + \gamma^3$$

$$V_4(s_2) = 1(0 + \gamma(0)) = 0$$

استنادی معمولی خواهد بود که این مقدار s_0 پایه درجه n باشد:

$$V_n(s_0) = \frac{1}{2}(\gamma + \gamma^2 + \dots + \gamma^{n-1} + \frac{\gamma^n}{1-\gamma})$$

$$|V_{n+1} - V_n| < \varepsilon \rightarrow$$

که

$$\begin{aligned}
 |V_{n+1}(s_0) - V_n(s_0)| &< \varepsilon \rightarrow \left| \frac{1}{2} (\gamma + \gamma^2 + \dots + \gamma^n + \frac{\gamma^2}{1-\gamma}) - \frac{1}{2} (\gamma + \gamma^2 + \dots + \gamma^{n-1} + \frac{\gamma^2}{1-\gamma}) \right| < \varepsilon \\
 &= \left| \gamma + \gamma^2 + \dots + \gamma^n + \frac{\gamma^2}{1-\gamma} - \gamma - \gamma^2 - \dots - \gamma^{n-1} - \frac{\gamma^2}{1-\gamma} \right| < \varepsilon' \\
 &= \gamma^n < \varepsilon \Rightarrow n \log \gamma < \log \varepsilon' \Rightarrow n \log \gamma < \varepsilon'' \Rightarrow
 \end{aligned}$$

مسئله ۲. (۱۰ نمره امتیازی) گرادیان در Multi-armed Bandit

برای حل مسئله Bandit با k بازو می‌توان به طور مستقیم و بدون واسطه‌گریتابع ارزش نیز احتمال انتخاب کنش‌ها را مدل‌سازی کرد. اگر $H_t(a)$ میزان تمايل به انتخاب کنش a در زمان t را نشان دهد، می‌توان سیاست را به صورت زیر محاسبه کرد:

$$\pi_t(a) := \frac{e^{H_t(a)}}{\sum_{b=1}^k e^{H_t(b)}}$$

می‌توان توزیع مذکور را با بیشینه‌سازی $\tilde{\mathbb{E}}[R_t] = \sum_x \pi_t(x)q^*(x)$ آموخت داد. R_t پاداش لحظه‌ای حاصل از انجام A_t است و داریم: $q^*(a) := \mathbb{E}[R_t | A_t = a]$

(۱) (۳ نمره) نشان دهید که

$$\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x)(\mathbb{I}[a = x] - \pi_t(a))$$

ب) (۲ نمره) را برحسب $\frac{\partial \pi_t(x)}{\partial H_t(a)}$ بنویسید.

پ) (۵ نمره) نشان دهید که $H_t(a)$ با رابطه‌ی زیر بروزرسانی می‌شود (منظور از \bar{R}_t میانگین پاداش‌ها از لحظه‌ی اول تا t و α نرخ یادگیری صعود در امتداد گرادیان است).

$$H_{t+1}(a) \leftarrow H_t(a) + \alpha(R_t - \bar{R}_t)(\mathbb{I}[a = A_t] - \pi_t(a))$$

(۱)

$$\frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x) = \frac{\partial}{\partial H_t(a)} \left[\frac{e^{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}} \right]$$

$$= \frac{\frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_{y=1}^k e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_{y=1}^k e^{H_t(y)}}{\partial H_t(a)}}{\left(\sum_{y=1}^k e^{H_t(y)} \right)^2}$$

$$= \frac{\mathbb{I}_{a=x} e^{H_t(x)} \sum_{y=1}^k e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{\left(\sum_{y=1}^k e^{H_t(y)} \right)^2}$$

$$= \frac{\mathbb{I}_{a=x} e^{H_t(x)}}{\sum_{y=1}^k e^{H_t(y)}} - \frac{e^{H_t(x)} e^{H_t(a)}}{\left(\sum_{y=1}^k e^{H_t(y)} \right)^2}$$

$$= \mathbb{I}_{a=x} \pi_t(x) - \pi_t(x) \pi_t(a) = \pi_t(x) (\mathbb{I}_{a=x} - \pi_t(a))$$

(.)

$$\begin{aligned}\frac{\partial E[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x) q_*(x) \right] = \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\ &= \sum_x (q_*(x) - R_t) \frac{\partial \pi_t(x)}{\partial H_t(a)}\end{aligned}$$

(٤)

از خش ب محاسبه

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \sum_x (q_a(x) - R_t) \frac{\partial R_t(x)}{\partial H_t(a)}$$

دست دفعه را می خواهیم داشت

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \sum_x n_t(x) (q_a(x) - R_t) \frac{\partial R_t(x)}{\partial H_t(a)} / R_t(x)$$

$$= E[(q_a(A_t) - R_t) \frac{\partial R_t(A_t)}{\partial H_t(a)} / R_t(A_t)]$$

$$= E[(R_t - \bar{R}_t) \frac{\partial R_t(A_t)}{\partial H_t(a)} / R_t(A_t)]$$

$$= E[(R_t - \bar{R}_t) R_t(A_t) (\mathbb{I}_{a=A_t} - R_t(a)) / R_t(A_t)]$$

$$= E[(R_t - \bar{R}_t) (\mathbb{I}_{a=A_t} - R_t(a))]$$

که در اینجا می خواهیم داشت

$$H_{t+1}(a) = H_t(a) + d(R_t - \bar{R}_t) (\mathbb{I}_{a=A_t} - R_t(a))$$

مسئله‌ی ۳. (۱۲ نمره) الگوریتم‌های یادگیری ارزش حالات

در این سوال به دنبال بررسی عملکرد روش‌های تخمین ارزش حالات با دو الگوریتم temporal difference و monte carlo هستیم. همچنین در نهایت همگرایی دو الگوریتم Q-learning و monte carlo را بررسی می‌کنیم.

(آ) (۲ نمره) روش MC برای تخمین ارزش حالات را به صورت مختصر توضیح دهید و نشان دهید تخمین از ارزش حالات تخمینی unbiased است.

(ب) (۲ نمره) یکی از مشکلات روش MC الزام به پایان رساندن هر episode برای بهروزرسانی ارزش حالات است. موضوعی که به خصوص در مسائی long horizon چالش برانگیز است. روش TD چگونه این مشکل را بر طرف می‌کند؟ روابط بهروزرسانی ارزش حالات در روش TD را ذکر کنید.

(پ) (۳ نمره) برای درک بهتر تفاوت این دو روش، ارزش حالات مربوط به markov reward process زیر را با توجه به episode های بیان شده با هر دو روش محاسبه کنید. آیا تفاوتی در مقدار محاسبه شده وجود دارد؟ نتیجه را تفسیر کنید.

A 0 B 0 C 0

A 0 B 0

B 0

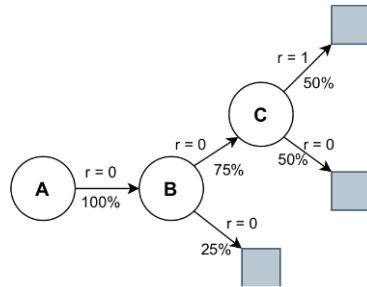
B 0 C 0

C 1

C 1

C 0

C 1



(ت) (۵ نمره) روش TD برای یادگیری ارزش حالات از حدس ارزش حالات بعدی استفاده می‌کند. آیا این موضوع همگرایی این الگوریتم را با مشکل مواجه می‌کند؟ اگر جواب مثبت است تحت چه شرایطی همگرایی قابل تضمین نیست؟ توضیح دهید. در مورد الگوریتم Q-learning که در آن کنش‌ها به صورت تصادفی انتخاب می‌شوند چطور؟ آیا همگرایی برای آن الگوریتم تضمین می‌شود؟

(ا) روش Monte Carlo روشهای احتمالی ندارد. در بسیاری از موارد تولید نمونه از توزیع احتمال ساده است ولی به دست آوردن توزیع احتمال غیر ممکن است. در ادامه روش برای روشنگری از حالت را توضیح می‌دهیم. می‌دانیم که ارزش هر حالت امید ریاضی reward به دست آمده با شروع از آن حالت است. با داشتن نمونه می‌توانیم این مقدار را با میانگین گیری از تمام مقادیر reward با شروع از این حالت، تخمین بزنیم. هر چه مقدار reward یا مقدار بازگشتی از حالت‌ها بیشتر داشته باشیم، تخمین دقیق‌تری خواهیم داشت. به طور مثال اگر بخواهیم مقدار $(s)_\pi$ را تخمین بزنیم، با داشتن مجموعه‌ای از episode‌ها که توسط π به دست آمده‌اند و از حالت s گشته‌اند. به طور مثال می‌توانیم مقدار برگشتی از هر باری که از حالت s شروع می‌کنیم را میانگین بگیریم که به این روش every-visit گفته می‌شود. روش دیگر این است که فقط مقادیر برگشتی را هنگامی که از حالت s شروع شده است میانگین بگیریم که به آن first-visit گفته می‌شود. هر دو این روش‌ها در صورتیکه تعداد مشاهده‌ها به سمت بی‌نهایت میل کند، تخمین درستی می‌دهند. در روش first-visit، می‌دانیم که مقادیر برگشتی، تخمین‌های i.i.d با واریانس متناهی هستند. با توجه به قانون اعداد بزرگ می‌دانیم که میانگین این تخمین‌ها به امید ریاضی آنها همگرا می‌شود. هر میانگین یک تخمین unbiased است و انحراف معیار خطای آن نیز در بازه $\frac{1}{\sqrt{n}}$ که n تعداد مقادیر برگشتی است، قرار می‌گیرد.

(ب) روش Temporal Difference مانند MC بدون داشتن مدل از محیط و بر اساس تجربه عمل می‌کند. همچنین شبیه به روش DP تخمین‌ها را با کمک سایر تخمین‌ها و بدون خروجی نهایی آپدیت می‌کنند. روش MC نیاز دارد که منتظر بماند تا مقدار برگشتی از حالتی که از آن شروع کردایم محاسبه شود و سپس از آن مقدار به عنوان target برای ارزش آن حالت استفاده می‌کند. روش every-visit MC به صورت زیر مقدار ارزش حالت را آپدیت می‌کند.

$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

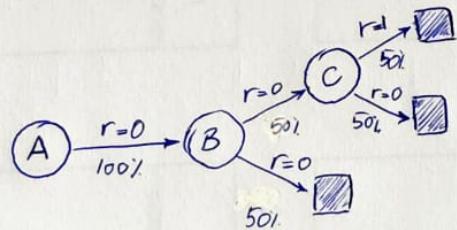
که در آن G_t مقدار برگشتی در زمان t است.

روش TD اما نیازی به داشتن مقدار G_t ندارد و فقط باید تا مرحله زمانی بعدی منتظر بماند. در زمان $t+1$ این روش براساس R_{t+1} و $V(S_{t+1})$ به صورت زیر ارزش حالت S را آپدیت می‌کند.

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

(ب)

بررسی مدل MDP با محدوده زمانی episode و TD برای این مسأله از حالت A و حالت B که هم به ایستار 0 و هم به ایستار 1 دارند، فتحاً ایستار 0 و هم به ایستار 1 دارند. از حالت C که هم به ایستار 0 و هم به ایستار 1 دارند، فتحاً ایستار 0 و هم به ایستار 1 دارند.



* بررسی این ایستار C، $\frac{1}{2}$ موقعیت ایستار صفر و $\frac{1}{2}$ موقعیت ایستار 1 داریم، از این حالت C بدلبا $\frac{1}{2}$ خواهد بود.

$$V(C) = \frac{1}{2} \times 0 + \frac{1}{2} \times 1 = \frac{1}{2}$$

از ایستار B نیز با توجه به این احتمالات ایستار صفر و ایستار 1 داریم، از این حالت B بدلبا $\frac{1}{2}$ خواهد بود.

$$V(B) = \frac{1}{2} \times 0 + \frac{1}{2} \times \left(\frac{1}{2}\right) = \frac{1}{4}$$

از ایستار A نیز با توجه به این احتمالات ایستار صفر و ایستار 1 داریم، از این حالت A بدلبا $\frac{1}{2}$ خواهد بود.

$$V(A) = 1 \times \frac{1}{4} = \frac{1}{4}$$

روش MC نیز ارزش حالت C برابر با $\frac{1}{2}$ است زیرا بهتر است از های حاصل نزدیک حالت را می‌گذین شویم /

$$V(C) = \frac{3}{6} \times 1 + \frac{3}{6} \times 0 = \frac{1}{2}$$

با این $\frac{1}{2}$ خوبیده.

از ارزش حالت B ضر خوبیده زیرا این حالت هم رطایت مرند است خوبیده هم در حالت C حالت C

$$V(B) = \frac{2}{4} \times 0 + \frac{2}{4} \times 0 = 0$$

می‌بود.

از ارزش حالت A ترتیب اسوبه به اینه خوبیده است، صفر را، بلطف باصف خوبیده.

$$V(A) = \frac{2}{2} \times 0 = 0$$

با توجه به اینکه روش TD سعی دارد ارزش حالت بعدی را نیز در نظر بگیرد و MRP را تولید کند، این روش ارزش‌های بهتری برای حالات تولید می‌کند و مقدار خطای آن نیز نسبت به روش MC سریعتر کاهش می‌آید.

ت) روش TD در صورتی تضمین می‌کند که تمامی جفت حالت و عملکرد در آن حالت را به تعداد نامتناهی بار تجربه کند و در غیر این صورت تضمینی بر همگرایی نمی‌دهد. در الگوریتم Q-learning که تابع حالت-ارزش، تابع بینه حالت-ارزش را تخمین می‌زند و الگوریتم را بسیار ساده می‌کند و همچنین فرایند همگرایی را بسیار سریع می‌کند. در این روش تنها کافی است که تمام جفت حالت و عملکردها مرتب آپدیت شوند و در این صورت این الگوریتم همگرا خواهد شد.

مسئله‌ی ۴. (۱۳ نمره) معماری Actor Critic و روش‌های Policy Gradient

گرادیان تابع هدف ساده شده روش‌های policy based در ۱ نشان داده شده است. یکی از ویژگی‌های گرادیان این تابع هدف واریانس بالای آن به دلیل ذات تصادفی تولید یک trajectory و دریافت پاداش است، موضوعی که فرآیند آموزش شبکه عصبی را با چالش همراه می‌کند. یکی از رویکردها برای کاهش واریانس این گرادیان استفاده از مقداری تحت عنوان baseline است. در این سوال ابتدا به بررسی تاثیر یک مقدار ثابت به عنوان baseline پرداخته و سپس با مطالعه دسته مهمی از معماری‌های شبکه‌های یادگیری تقویتی با نام actor critic که به دنبال یادگیری این baseline هستند، با دو روش ارائه شده در این شاخه آشنا می‌شویم.

$$\mathbb{E}_{\tau \sim p(\tau; \theta)} [\nabla_{\theta} \log p(\tau; \theta) r(\tau)] \quad (1)$$

این گرادیان با استفاده از یک مقدار ثابت c تحت عنوان baseline به شکل زیر تغییر می‌کند:

$$\mathbb{E}_{\tau \sim p(\tau; \theta)} [\nabla_{\theta} \log p(\tau; \theta) (r(\tau) - c)] \quad (2)$$

(آ) (۲ نمره) گرادیان تابع هدف ۲ را به شکل $\mathbb{E}[f(x) - \phi(x)] + \mathbb{E}[\phi(x)]$ بازنویسی کنید که در آن $\mathbb{E}[f(x) - \phi(x)]$ همان عبارت ۱ است. مقدار $\mathbb{E}[\phi(\tau)]$ را نیز محاسبه کنید.

(ب) (۳ نمره) ثابت کنید c بهینه که سبب کمینه شدن $\text{Var}[f(\tau) - \phi(\tau)]$ می‌گردد برابر است با:

$$c = \frac{\mathbb{E} [(\nabla_{\theta} \log p(\tau; \theta))^T r(\tau)]}{\mathbb{E} [(\nabla_{\theta} \log p(\tau; \theta))^T]}$$

(پ) (۴ نمره) یکی از روش‌های موفق یادگیری تقویتی off policy الگوریتم SAC می‌باشد. در این مقاله تابع هدف یادگیری تقویتی

$$J(\pi) = \sum_{t=1}^T \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_{\pi}} [r(\mathbf{s}_t, \mathbf{a}_t)] \quad (3)$$

$$J(\pi) = \sum_{t=1}^T \mathbb{E}_{(s_t, a_t) \sim \rho_\pi} [r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot | s_t))] \quad (4)$$

تغییر پیدا کرده است.

اولاً ترم $\mathcal{H}(\pi(\cdot | s_t))$ چه تاثیری دارد؟

دوماً طبق مقاله بیان کنید که سیاست جدید به دست آمده در گام policy improvement در الگوریتم policy iteration برای اینتابع هدف جدید به چه شکل خواهد بود؟

ت) (۳ نمره) یکی از دسته روش‌های بهینه‌سازی روش‌های trust region هستند که در آن‌ها همانند سایر روش‌های بهینه‌سازی تکرار شونده، در هرگام از حدس گام قبل با مکانیزمی به حدس گام بعد می‌رسیم. نکته مهم در این روش‌ها کنترل نزدیکی حدس بعد به حدس قبلی و به اصطلاح باقی ماندن در فضای اطمینان است. **TRPO** با الهام‌گیری از همین موضوع سعی در کنترل میزان تغییرات سیاست در هرگام با استفاده از تابع هزینه ذیل دارد. این کنترل میزان تغییرات و جلوگیری از تغییرات ناگهانی در سیاست سبب ایجاد پایداری در پادگیری سیاست می‌گردد.

$$\begin{aligned} & \underset{\theta}{\text{maximize}} \quad \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \hat{A}_t \right] \\ & \text{subject to} \quad \hat{\mathbb{E}}_t [\text{KL}[\pi_{\theta_{\text{old}}}(\cdot | s_t), \pi_\theta(\cdot | s_t)]] \leq \delta \end{aligned}$$

با این حال این مسئله بهینه‌سازی دارای hard constraint ای است که حل آن را چالش برانگیز می‌کند. به صورت مختصر و کلی بیان کنید روش پیشنهادی **PPO** چگونه این مشکل را حل می‌کند؟

ت) الگوریتم **PPO** به جای حل مسئله بهینه‌سازی بالا مساله زیر را حل می‌کند که soft constraint دارد به جای hard constraint همچنین به ما اجازه می‌دهد که بتوانیم از Stochastic Gradient Descent به جای conjugate gradient constraint برای حل این مسئله استفاده کنیم.

$$\underset{\theta}{\text{maximize}} \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \hat{A}_t - \beta \text{KL}[\pi_{\theta_{\text{old}}}(\cdot | s_t), \pi_\theta(\cdot | s_t)] \right]$$