Proximal Antagonistic Constrained Policy Search for Sample-Efficient Offline Actor-Critic (Supplementary Material)

Shuo Cao, Xuesong Wang, *Member, IEEE*, Jiazhi Zhang, and Yuhu Cheng, *Member, IEEE*

I. Proof for Theorem 1

Proof: By Lemma 1, it follows that

$$J(\pi_{k+1}) - J(\mu_{k+1}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\left[A^{\mu_{k+1}}\left(s,\pi_{k+1}(s)\right)\right]. \tag{30}$$

Considering the definition of the advantage function $A^{\mu_{k+1}}\left(s,\pi_{k+1}(s)\right) = Q^{\mu_{k+1}}\left(s,\pi_{k+1}(s)\right) - V^{\mu_{k+1}}(s)$, the properties of deterministic policy $V^{\mu_{k+1}}(s) = Q^{\mu_{k+1}}\left(s,\mu_{k+1}(s)\right)$, and combining with the shorthand $Q^{\mu_{k+1}}$ as Q_{k+1}, Eq. (30) can be reformulated as

$$\begin{aligned}J(\pi_{k+1}) - J(\mu_{k+1}) =& \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\Big[Q_{k+1}\left(s,\pi_{k+1}(s)\right)\\ &- Q_{k+1}\left(s,\mu_{k+1}(s)\right)\Big].\end{aligned} \tag{31}$$

By utilizing the grouping method for Eq. (31), we obtain

$$\begin{aligned}&J(\pi_{k+1}) - J(\mu_{k+1})\\ =&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\Bigg[\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big)\\ &-\Bigg(\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - Q_{k+1}\left(s,\pi_{k+1}(s)\right)\Big)\\ &+\Big(Q_{k+1}\left(s,\mu_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big)\Bigg)\Bigg].\end{aligned} \tag{32}$$

According to Lemma 2 and the trigonometric theorem, there are

$$\begin{aligned}&\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - Q_{k+1}\left(s,\pi_{k+1}(s)\right)\Big)\\ &+\Big(Q_{k+1}\left(s,\mu_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big)\\ \le&\left|\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - Q_{k+1}\left(s,\pi_{k+1}(s)\right)\right|\\ &+\left|\hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right) - Q_{k+1}\left(s,\mu_{k+1}(s)\right)\right|\\ \le&\frac{2\gamma C_{T,\delta}R_{\max}}{(1-\gamma)\sqrt{\mathcal{D}_c}}.\end{aligned} \tag{33}$$

Multiplying Eq. (33) by -1 and substituting it into Eq. (32) yields

$$\begin{aligned}&J(\pi_{k+1}) - J(\mu_{k+1})\\ \ge&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\Bigg[\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big)\\ &-\frac{2\gamma C_{T,\delta}R_{\max}}{(1-\gamma)\sqrt{\mathcal{D}_c}}\Bigg].\end{aligned} \tag{34}$$

Inspired by [23], here we define the first-order linear approximation error of the Q function as

$$\begin{aligned}&\delta_{PAC-PS}(s)\\ :=&\left|\bar{Q}_{k+1}(s,\tilde{a};\widehat{a}) - \hat{Q}_{k+1}(s,\tilde{a})\right|\\ :=&\Bigg|\hat{Q}_{k+1}(s,\widehat{a}) + \left[\nabla_{\tilde{a}}\hat{Q}_{k+1}(s,\tilde{a})\Big|_{\tilde{a}=\widehat{a}}\right]\times\left(\tilde{a}-\widehat{a}\right)\\ &-\hat{Q}_{k+1}(s,\tilde{a})\Bigg|,\end{aligned} \tag{35}$$

where δ_{PAC-PS} is zero when $\tilde{a}=\widehat{a}$ occurs.

Recalling Eqs. (10), (14), and (15), there are

$$\begin{aligned}\bar{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) :=&\lambda\bar{Q}_{k+1}\left(s,\pi_{k+1}(s);\mu_{k+1}(s)\right)\\ &+(1-\lambda)\bar{Q}_{k+1}\left(s,\pi_{k+1}(s);a\right)\end{aligned} \tag{36}$$

and

$$\begin{aligned}\bar{Q}_{k+1}\left(s,\mu_{k+1}(s)\right) :=&\lambda\bar{Q}_{k+1}\left(s,\mu_{k+1}(s);\mu_{k+1}(s)\right)\\ &+(1-\lambda)\bar{Q}_{k+1}\left(s,\mu_{k+1}(s);a\right)\\ :=&(1-\lambda)\bar{Q}^{\mu_{k+1}}\left(s,\mu_{k+1}(s);a\right).\end{aligned} \tag{37}$$

According to Eq. (36), it can be further obtained that

$$\begin{aligned}&\left|\bar{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)\right|\\ =&\Bigg|\lambda\Bigg(\hat{Q}_{k+1}(s,\mu_{k+1}(s))\\ &+\left[\nabla_{\pi_{k+1}(s)}\hat{Q}_{k+1}(s,\pi_{k+1}(s))\Big|_{\pi_{k+1}(s)=\mu_{k+1}(s)}\right]\\ &\times\left(\pi(s)-\mu_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)\Bigg)\\ &+(1-\lambda)\Bigg(\hat{Q}_{k+1}(s,a)\\ &+\left[\nabla_{\pi_{k+1}(s)}\hat{Q}_{k+1}(s,\pi_{k+1}(s))\Big|_{\pi_{k+1}(s)=a}\right]\\ &\times\left(\pi(s)-a\right) - \hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)\Bigg)\Bigg|\\ =&\lambda\delta_{PAC-PS}(s) + (1-\lambda)\delta_{PAC-PS}(s)\\ =&\delta_{PAC-PS}(s).\end{aligned} \tag{38}$$

Similar to the derivation of Eq. (38), it further follows from Eq. (37) that

$$\begin{aligned}&\left|\bar{Q}_{k+1}\left(s,\mu_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\right|\\ =&(1-\lambda)\delta_{PAC-PS}(s).\end{aligned} \tag{39}$$

Using Eqs. (38) and (39), it can be easily obtained that

$$\begin{aligned}&\left(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)-\hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\right)\\ \geq&\left(\bar{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)-\bar{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\right)\\ &-(2-\lambda)\delta_{PAC-PS}(s).\end{aligned} \tag{40}$$

Substitute Eq. (40) into Eq. (34) and ultimately obtain the following inequality w.h.p. than $1-\varepsilon$:

$$\begin{aligned}&J(\pi_{k+1})-J(\mu_{k+1})\\ \geq&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\left[\left(\bar{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)-\bar{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\right)\right.\\ &\left.-(2-\lambda)\delta_{PAC-PS}(s)-\frac{2\gamma C_{T,\delta}R_{\max}}{(1-\gamma)\sqrt{\mathcal{D}_c}}\right]:=\zeta.\end{aligned} \tag{41}$$

Thus, Theorem 1 is proved. ■

II. Proof for Theorem 2

Proof: In the context of HBP $\mu(s)$, according to the grouping method and the trigonometric inequality theorem, it can be obtained that

$$\begin{aligned}&\lim_{k\to\infty}|J(\pi^*)-J(\pi_{k+1})|\\ \leq&\lim_{k\to\infty}|J(\pi^*)-J(\mu_{k+1})|+\lim_{k\to\infty}|J(\pi_{k+1})-J(\mu_{k+1})|.\end{aligned} \tag{42}$$

First, considering $\lim_{k\to\infty}|J(\pi^*)-J(\mu_{k+1})|$ and recalling the definitions of $J(\pi^*)$ and $J(\mu_{k+1})$ in Section II-A, we can get

$$\begin{aligned}&\lim_{k\to\infty}|J(\pi^*)-J(\mu_{k+1})|\\ =&\lim_{k\to\infty}\left|\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^*}(s)}\left[r(s)\right]-\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\mu_{k+1}}(s)}\left[r(s)\right]\right|\\ \leq&\frac{1}{1-\gamma}\lim_{k\to\infty}\mathbb{E}_s\left[\left|d^{\pi^*}(s)-d^{\mu_{k+1}}(s)\right||r(s)|\right]\\ \leq&\frac{R_{\max}}{1-\gamma}\lim_{k\to\infty}\mathbb{E}_s\left|d^{\pi^*}(s)-d^{\mu_{k+1}}(s)\right|.\end{aligned} \tag{43}$$

According to Lemma 3, $\lim_{k\to\infty}\mathbb{E}_s\left|d^{\pi^*}(s)-d^{\mu_{k+1}}(s)\right|\leq C_dK_T\lim_{k\to\infty}\left[\max_{s\in S}\|\pi^*(s)-\mu_{k+1}(s)\|\right]$ holds, and substituting it into Eq. (43) yields

$$\begin{aligned}&\lim_{k\to\infty}|J(\pi^*)-J(\mu_{k+1})|\\ \leq&\frac{R_{\max}C_dK_T}{1-\gamma}\lim_{k\to\infty}\left[\max_{s\in S}\|\pi^*(s)-\mu_{k+1}(s)\|\right]\\ \leq&\frac{R_{\max}C_dK_T}{1-\gamma}\tilde{\epsilon}^*.\end{aligned} \tag{44}$$

where $\lim_{k\to\infty}\left[\max_{s\in S}\|\pi^*(s)-\mu_{k+1}(s)\|\right]\leq\tilde{\epsilon}^*$.

Then, considering $\lim_{k\to\infty}|J(\pi_{k+1})-J(\mu_{k+1})|$ and recalling the definitions of $J(\pi_{k+1})$ and $J(\mu_{k+1})$ in Section II-A, similar to Eq. 43, we can get

$$\begin{aligned}&\lim_{k\to\infty}|J(\pi_{k+1})-J(\mu_{k+1})|\\ \leq&\frac{R_{\max}}{1-\gamma}\lim_{k\to\infty}\mathbb{E}_s\left|d^{\pi_{k+1}}(s)-d^{\mu_{k+1}}(s)\right|,\end{aligned} \tag{45}$$

According to Lemma 3, $\lim_{k\to\infty}\mathbb{E}_s\left|d^{\pi_{k+1}}(s)-d^{\mu_{k+1}}(s)\right|\leq C_dK_T\lim_{k\to\infty}\left[\max_{s\in S}\|\pi_{k+1}(s)-\mu_{k+1}(s)\|\right]$ holds, and substituting it into Eq. (45) yields

$$\begin{aligned}&\lim_{k\to\infty}|J(\pi_{k+1})-J(\mu_{k+1})|\\ \leq&\frac{R_{\max}C_dK_T}{1-\gamma}\lim_{k\to\infty}\left[\max_{s\in S}\|\pi_{k+1}(s)-\mu_{k+1}(s)\|\right]\\ =&\frac{R_{\max}C_dK_T}{1-\gamma}\lim_{k\to\infty}\left[\max_{s\in S}\|\pi_{k+1}(s)-a+a-\mu_{k+1}(s)\|\right]\\ \leq&\frac{R_{\max}C_dK_T}{1-\gamma}\left(\lim_{k\to\infty}\left[\max_{s\in S}\|\pi_{k+1}(s)-a\|\right]\right.\\ &\left.+\lim_{k\to\infty}\left[\max_{s\in S}\|a-\mu_{k+1}(s)\|\right]\right)\\ \leq&\frac{R_{\max}C_dK_T}{1-\gamma}\left(\tilde{\epsilon}^{\pi}+\tilde{\epsilon}^{\mu}\right),\end{aligned} \tag{46}$$

where $\lim_{k\to\infty}\left[\max_{s\in S}\|\pi_{k+1}(s)-a\|\right]\leq\tilde{\epsilon}^{\pi}$ and $\lim_{k\to\infty}\left[\max_{s\in S}\|a-\mu_{k+1}(s)\|\right]\leq\tilde{\epsilon}^{\mu}$. Finally, substituting Eqs. (44) and (46) into Eq. (42) yields

$$\lim_{k\to\infty}|J(\pi^*)-J(\pi_{k+1})|\leq\frac{R_{\max}C_dK_T}{1-\gamma}\left(\tilde{\epsilon}^*+\tilde{\epsilon}^{\pi}+\tilde{\epsilon}^{\mu}\right). \tag{47}$$

However, in the context of BP $\pi_\beta(s)$, similar results can be obtained

$$\begin{aligned}&\lim_{k\to\infty}|J(\pi^*)-J(\pi_{k+1})|\\ \leq&\lim_{k\to\infty}|J(\pi^*)-J(\pi_\beta)|+\lim_{k\to\infty}|J(\pi_{k+1})-J(\pi_\beta)|.\end{aligned} \tag{48}$$

For $\lim_{k\to\infty}|J(\pi^*)-J(\pi_\beta)|$, a similar derivation of Eq. (43) - Eq. (44) leads to

$$\lim_{k\to\infty}|J(\pi^*)-J(\pi_\beta)|\leq\frac{R_{\max}C_dK_T}{1-\gamma}\epsilon^*, \tag{49}$$

where $\max_{s\in S}\|\pi^*(s)-\pi_\beta(s)\|\leq\epsilon^*$. Since the HBP $\mu(s)$ with RL signal is closer to the optimal policy compared with BP $\pi_\beta(s)$, $\mu(s)$ can form a smaller error measure, i.e., $\tilde{\epsilon}^*\leq\epsilon^*$ holds.

For $\lim_{k\to\infty}|J(\pi_{k+1})-J(\pi_\beta)|$, a similar derivation of Eq. (45) - Eq. (46) leads to

$$\begin{aligned}&\lim_{k\to\infty}|J(\pi_{k+1})-J(\pi_\beta)|\\ \leq&\lim_{k\to\infty}|J(\pi_{k+1})-J(\mu_{k+1})|+\lim_{k\to\infty}|J(\mu_{k+1})-J(\pi_\beta)|\\ \leq&\frac{R_{\max}C_dK_T}{1-\gamma}\left(\tilde{\epsilon}^{\pi}+2\tilde{\epsilon}^{\mu}\right).\end{aligned} \tag{50}$$

Substituting Eqs. (49) and (50) into Eq. (48) yields

$$\lim_{k\to\infty}|J(\pi^*)-J(\pi_{k+1})|\leq\frac{R_{\max}C_dK_T}{1-\gamma}\left(\epsilon^*+\tilde{\epsilon}^{\pi}+2\tilde{\epsilon}^{\mu}\right). \tag{51}$$

By comparing Eq. (47) and Eq. (51), it can be seen that the performance bounds of PAC-PS using HBP $\mu(s)$ can further refine to $\tilde{\epsilon}^*+\tilde{\epsilon}^{\pi}+\tilde{\epsilon}^{\mu}\leq\epsilon^*+\tilde{\epsilon}^{\pi}+2\tilde{\epsilon}^{\mu}$. Thus, Theorem 2 is proved. ■