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

I. Proof for Theorem 1

Proof: By Lemma 1, it follows that

$$J(\pi_{k+1}) - J(\mu_{k+1}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\left[A^{\mu_{k+1}}\left(s,\pi_{k+1}(s)\right)\right]. \tag{24}$$

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. (24) can be reformulated as

$$J(\pi_{k+1}) - J(\mu_{k+1}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\Big[Q_{k+1}\left(s,\pi_{k+1}(s)\right) - Q_{k+1}\left(s,\mu_{k+1}(s)\right)\Big]. \tag{25}$$

By utilizing the grouping method for Eq. (25), we obtain

$$\begin{aligned} J(\pi_{k+1}) - J(\mu_{k+1}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\bigg[&\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big) \\ -\bigg(\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - Q_{k+1}\left(s,\pi_{k+1}(s)\right)\Big) &+ \Big(Q_{k+1}\left(s,\mu_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big)\bigg)\bigg]. \end{aligned} \tag{26}$$

According to Lemma 2 and the trigonometric theorem, there are

$$\begin{aligned} &\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - Q_{k+1}\left(s,\pi_{k+1}(s)\right)\Big) + \Big(Q_{k+1}\left(s,\mu_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big) \\ \le&\Big|\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - Q_{k+1}\left(s,\pi_{k+1}(s)\right)\Big| + \Big|\hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right) - Q_{k+1}\left(s,\mu_{k+1}(s)\right)\Big| \\ \le&\frac{2\gamma C_{T,\varepsilon}R_{\max}}{(1-\gamma)\sqrt{\mathcal{D}_c}}. \end{aligned} \tag{27}$$

Multiplying Eq. (27) by -1 and substituting it into Eq. (26) yields

$$J(\pi_{k+1}) - J(\mu_{k+1}) \ge \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\bigg[\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big) - \frac{2\gamma C_{T,\varepsilon}R_{\max}}{(1-\gamma)\sqrt{\mathcal{D}_c}}\bigg]. \tag{28}$$

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

$$\begin{aligned} \delta_{PAC-PS}(s) &:= \Big|\bar{Q}_{k+1}(s,\tilde{a};\widehat{a}) - \hat{Q}_{k+1}(s,\tilde{a})\Big| \\ &:= \Big|\hat{Q}_{k+1}(s,\widehat{a}) + \Big[\nabla_{\tilde{a}}\hat{Q}_{k+1}(s,\tilde{a})\Big|_{\tilde{a}=\widehat{a}}\Big] \times (\tilde{a}-\widehat{a}) - \hat{Q}_{k+1}(s,\tilde{a})\Big|, \end{aligned} \tag{29}$$

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

Recalling Eqs. (8), (11), and (12), there are

$$\bar{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) := \lambda\bar{Q}_{k+1}\left(s,\pi_{k+1}(s);\mu_{k+1}(s)\right) + (1-\lambda)\bar{Q}_{k+1}\left(s,\pi_{k+1}(s);a\right) \tag{30}$$

and

$$\bar{Q}_{k+1}\left(s,\mu_{k+1}(s)\right) := \lambda\bar{Q}_{k+1}\left(s,\mu_{k+1}(s);\mu_{k+1}(s)\right) + (1-\lambda)\bar{Q}_{k+1}\left(s,\mu_{k+1}(s);a\right). \tag{31}$$

Using the grouping method for $\bar{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)$, there is

$$\begin{aligned} &\bar{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right) \\ =&\lambda\Big(\bar{Q}_{k+1}\left(s,\pi_{k+1}(s);\mu_{k+1}(s)\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big) + (1-\lambda)\Big(\bar{Q}_{k+1}\left(s,\pi_{k+1}(s);a\right) - \hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big). \end{aligned} \tag{32}$$

According to Eq. (30), it can be further obtained that

$$\begin{aligned}
&\left|\bar{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)-\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)\right| \\
=&\Bigg|\lambda\Bigg(\hat{Q}_{k+1}(s,\mu_{k+1}(s))+\left[\nabla_{\pi_{k+1}(s)}\hat{Q}_{k+1}(s,\pi_{k+1}(s))\Big|_{\pi_{k+1}(s)=\mu_{k+1}(s)}\right]\times(\pi(s)-\mu_{k+1}(s))-\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)\Bigg) \\
&+(1-\lambda)\Bigg(\hat{Q}_{k+1}(s,a)+\left[\nabla_{\pi_{k+1}(s)}\hat{Q}_{k+1}(s,\pi_{k+1}(s))\Big|_{\pi_{k+1}(s)=a}\right]\times(\pi(s)-a)-\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)\Bigg)\Bigg| \\
=&\lambda\delta_{PAC-PS}(s)+(1-\lambda)\delta_{PAC-PS}(s) \\
=&\delta_{PAC-PS}(s).
\end{aligned} \tag{33}$$

Similar to the derivation of Eqs. (32)-(33), it further follows from Eq. (31) that

$$\left|\bar{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)-\hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\right|=(1-\lambda)\delta_{PAC-PS}(s). \tag{34}$$

Using Eqs. (33) and (34), it can be easily obtained that

$$\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)-\hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big)\geq\Big(\bar{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)-\bar{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big)-(2-\lambda)\delta_{PAC-PS}(s). \tag{35}$$

According to $\pi_{k+1}=\arg\max_{\mu_{k+1}}\bar{Q}_{k+1}$, Eq. (35) can be converted into

$$\Big(\hat{Q}_{k+1}\left(s,\pi_{k+1}(s)\right)-\hat{Q}_{k+1}\left(s,\mu_{k+1}(s)\right)\Big)\geq(\lambda-2)\delta_{PAC-PS}(s). \tag{36}$$

Substitute Eq. (36) into Eq. (28) and ultimately obtain the following inequality w.h.p. than $1-\varepsilon$:

$$J(\pi_{k+1})-J(\mu_{k+1})\geq\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{k+1}}}\left[(\lambda-2)\delta_{PAC-PS}(s)-\frac{2\gamma C_{T,\varepsilon}R_{\max}}{(1-\gamma)\sqrt{\mathcal{D}_c}}\right]:=\zeta. \tag{37}$$

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

$$\lim_{k\to\infty}\left|J(\pi^*)-J(\pi_{k+1})\right|\leq\lim_{k\to\infty}\left|J(\pi^*)-J(\mu_{k+1})\right|+\lim_{k\to\infty}\left|J(\pi_{k+1})-J(\mu_{k+1})\right|. \tag{38}$$

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

$$\begin{aligned}
\lim_{k\to\infty}\left|J(\pi^*)-J(\mu_{k+1})\right|=&\lim_{k\to\infty}\left|\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^*}(s)}\left[r(s)\right]-\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\mu_{k+1}}(s)}\left[r(s)\right]\right| \\
\leq&\frac{1}{1-\gamma}\lim_{k\to\infty}\mathbb{E}_s\left[\left|d^{\pi^*}(s)-d^{\mu_{k+1}}(s)\right|\left|r(s)\right|\right] \\
\leq&\frac{R_{\max}}{1-\gamma}\lim_{k\to\infty}\mathbb{E}_s\left|d^{\pi^*}(s)-d^{\mu_{k+1}}(s)\right|.
\end{aligned} \tag{39}$$

According to Lemma 3, $\lim\limits_{k\to\infty}\mathbb{E}_s\left|d^{\pi^*}(s)-d^{\mu_{k+1}}(s)\right|\leq C_dK_T\lim\limits_{k\to\infty}\left[\max\limits_{s\in S}\left\|\pi^*(s)-\mu_{k+1}(s)\right\|\right]$ holds, and substituting it into Eq. (39) yields

$$\begin{aligned}
\lim_{k\to\infty}\left|J(\pi^*)-J(\mu_{k+1})\right|\leq&\frac{R_{\max}C_dK_T}{1-\gamma}\lim_{k\to\infty}\left[\max_{s\in S}\left\|\pi^*(s)-\mu_{k+1}(s)\right\|\right] \\
\leq&\frac{R_{\max}C_dK_T}{1-\gamma}\tilde{\epsilon}^*.
\end{aligned} \tag{40}$$

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

Then, considering $\lim\limits_{k\to\infty}\left|J(\pi_{k+1})-J(\mu_{k+1})\right|$ and recalling the definitions of $J(\pi_{k+1})$ and $J(\mu_{k+1})$ in Section II-A, similar to Eq. 39, we can get

$$\lim_{k\to\infty}\left|J(\pi_{k+1})-J(\mu_{k+1})\right|\leq\frac{R_{\max}}{1-\gamma}\lim_{k\to\infty}\mathbb{E}_s\left|d^{\pi_{k+1}}(s)-d^{\mu_{k+1}}(s)\right|, \tag{41}$$

According to Lemma 3, $\lim_{k\to\infty} \mathbb{E}_s \left|d^{\pi_{k+1}}(s) - d^{\mu_{k+1}}(s)\right| \leq C_d K_T \lim_{k\to\infty} \left[\max_{s\in S} \|\pi_{k+1}(s) - \mu_{k+1}(s)\|\right]$ holds, and substituting it into Eq. (41) yields

$$\begin{aligned}\lim_{k\to\infty} |J(\pi_{k+1}) - J(\mu_{k+1})| \leq& \frac{R_{\max} C_d K_T}{1-\gamma} \lim_{k\to\infty} \left[\max_{s\in S} \|\pi_{k+1}(s) - \mu_{k+1}(s)\|\right] \\ =& \frac{R_{\max} C_d K_T}{1-\gamma} \lim_{k\to\infty} \left[\max_{s\in S} \|\pi_{k+1}(s) - a + a - \mu_{k+1}(s)\|\right] \\ \leq& \frac{R_{\max} C_d K_T}{1-\gamma} \left(\lim_{k\to\infty} \left[\max_{s\in S} \|\pi_{k+1}(s) - a\|\right] + \lim_{k\to\infty} \left[\max_{s\in S} \|a - \mu_{k+1}(s)\|\right]\right) \\ \leq& \frac{R_{\max} C_d K_T}{1-\gamma} \left(\tilde{\epsilon}^{\pi} + \tilde{\epsilon}^{\mu}\right),\end{aligned} \tag{42}$$

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. (40) and (42) into Eq. (38) yields

$$\lim_{k\to\infty} |J(\pi^*) - J(\pi_{k+1})| \leq \frac{R_{\max} C_d K_T}{1-\gamma} \left(\tilde{\epsilon}^* + \tilde{\epsilon}^{\pi} + \tilde{\epsilon}^{\mu}\right). \tag{43}$$

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

$$\lim_{k\to\infty} |J(\pi^*) - J(\pi_{k+1})| \leq \lim_{k\to\infty} |J(\pi^*) - J(\pi_\beta)| + \lim_{k\to\infty} |J(\pi_{k+1}) - J(\pi_\beta)|. \tag{44}$$

For $\lim_{k\to\infty} |J(\pi^*) - J(\pi_\beta)|$, a similar derivation of Eq. (39) - Eq. (40) leads to

$$\lim_{k\to\infty} |J(\pi^*) - J(\pi_\beta)| \leq \frac{R_{\max} C_d K_T}{1-\gamma} \epsilon^*, \tag{45}$$

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. (41) - Eq. (42) leads to

$$\begin{aligned}\lim_{k\to\infty} |J(\pi_{k+1}) - J(\pi_\beta)| \leq& \lim_{k\to\infty} |J(\pi_{k+1}) - J(\mu_{k+1})| + \lim_{k\to\infty} |J(\mu_{k+1}) - J(\pi_\beta)| \\ \leq& \frac{R_{\max} C_d K_T}{1-\gamma} \left(\tilde{\epsilon}^{\pi} + 2\tilde{\epsilon}^{\mu}\right).\end{aligned} \tag{46}$$

Substituting Eqs. (45) and (46) into Eq. (44) yields

$$\lim_{k\to\infty} |J(\pi^*) - J(\pi_{k+1})| \leq \frac{R_{\max} C_d K_T}{1-\gamma} \left(\epsilon^* + \tilde{\epsilon}^{\pi} + 2\tilde{\epsilon}^{\mu}\right). \tag{47}$$

By comparing Eq. (43) and Eq. (47), 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. ■