Electronic Companions

14. Supplementary Experiment

Reward Recovering on a Tabular Grid World. In order to validate the proposed algorithm as a method for IRL and show we recover correct rewards, we test our algorithm on a tabular Grid world setting, by using an open-source implementation[3]. The classical method Ziebart et al. (2008) requires repeated backward and forward passes, to calculate soft-values and action probabilities under a given reward and optimize the rewards respectively. By using a single-loop algorithm structure, our proposed algorithm could avoid the expensive backward pass in optimizing the policy under each given reward without compromising reward estimation accuracy. In Figure 1, we visualize our recovered rewards in the discrete GridWorld environment. According to Fig. 1, we show that ML-IRL converges much faster compared with MaxEnt-IRL while achieving similar accuracy on the recovered reward function.

[3] https://github.com/yrlu/irl-imitation

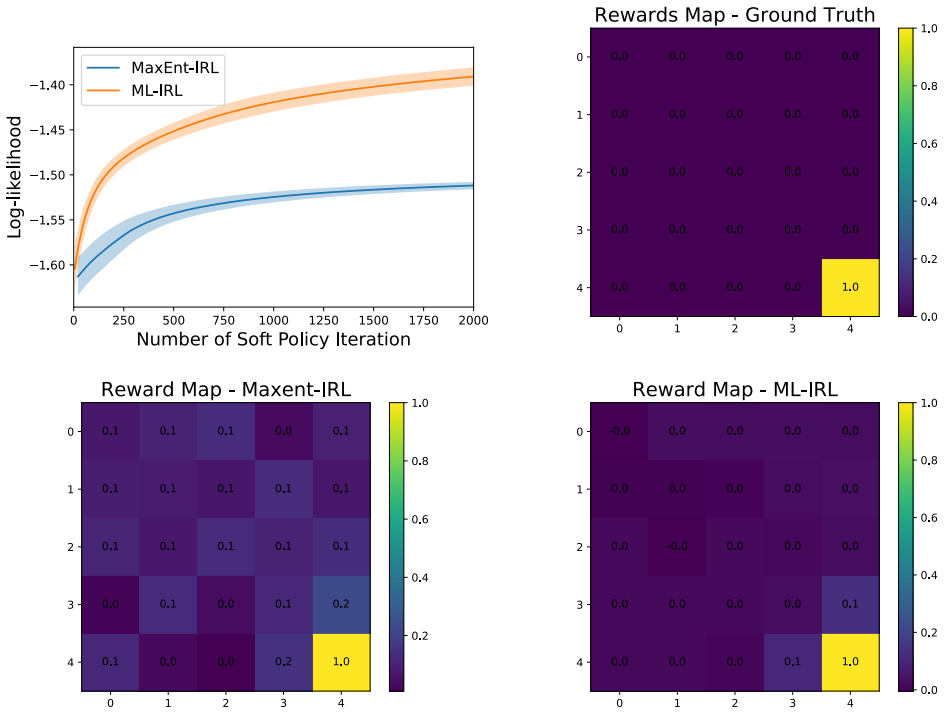

Figure 1 **Tabular Grid World.** We use a discrete GridWorld environment with 5 possible actions: up, down, left, right, stay. Agent starts in a random state. (With 30 expert demos).

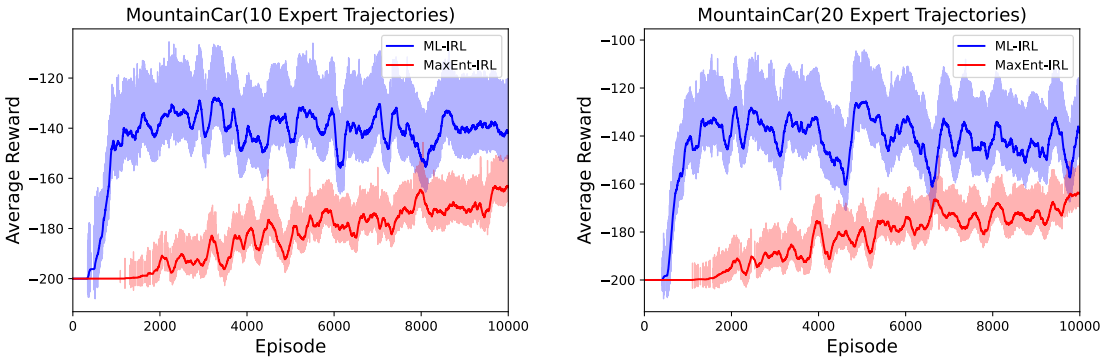

Figure 2 **Mountain Car.** We compare ML-IRL with MaxEnt-IRL under different numbers of expert trajectories.

Inverse Reinforcement Learning on Mountain Car. To further demonstrate the superiority of ML-IRL over MaxEnt-IRL, we test our algorithm on the classic reinforcement learning task – Mountain Car. According to Fig. 2, we show that ML-IRL is able to achieve faster convergence through leveraging the alternating updates between the policy and the reward estimator.

15. Proof of Lemma 1

Recall that in (8), the likelihood objective $L(\theta)$ can be expressed as:

$$L(\theta) = \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] - \mathbb{E}_{s_0\sim\rho}\Big[V_\theta(s_0)\Big].$$

Moreover, given a dataset of collected expert trajectories, we have defined the estimation problem $\widehat{L}(\theta;\mathcal{D})$ as below:

$$\widehat{L}(\theta;\mathcal{D}) = \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] - \mathbb{E}_{s_0\sim\rho}\Big[V_\theta(s_0)\Big].$$

Then we have the following result:

$$|L(\theta) - \widehat{L}(\theta;\mathcal{D})| = \Big|\mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] - \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big]\Big|.$$

According to Assumption 1, we obtain that the discounted cumulative reward of any trajectory follows $0 \le \sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta) \le \frac{C_r}{1-\gamma}$. Then by applying Hoeffdings inequality, for any $\epsilon > 0$, we have the following result:

$$P\bigg(\Big|\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big] - \mathbb{E}_{\tau\sim\mathcal{D}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big]\Big| \ge \epsilon\bigg) \le 2\exp\Big(-\frac{2|\mathcal{D}|\epsilon^2}{(\frac{C_r}{1-\gamma})^2}\Big).$$

Then by setting $\delta = 2\exp\Big(-\frac{2|\mathcal{D}|\epsilon^2}{(\frac{C_r}{1-\gamma})^2}\Big)$, with probability greater than $1-\delta$, we have

$$\left|\mathbb{E}_{\tau\sim\pi^{\mathrm{E}}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big]-\mathbb{E}_{\tau\sim\mathcal{D}}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\Big]\right|\leq\frac{C_r}{1-\gamma}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}},\tag{60}$$

where C_r is the constant defined in Assumption 1. According to (60), we obtain the concentration bound to quantify the approximation between $L(\theta)$ and $\widehat{L}(\theta;\mathcal{D})$ as below:

$$|L(\theta)-\widehat{L}(\theta;\mathcal{D})|\leq\frac{C_r}{1-\gamma}\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}|}},\quad\text{with probability greater than } 1-\delta.$$

This completes the proof of this lemma. ■

16. Proof of Lemma 2

Proof. Recall that we have defined V_θ, Q_θ as the soft value function, soft Q-function under reward parameter θ and the optimal policy π_θ in an entropy-regularized MDP. Hence, the following expressions hold for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$:

$$V_\theta(s):=\mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t,a_t;\theta)+\mathcal{H}(\pi_\theta(\cdot|s_t))\Big)\middle|s_0=s\right],\tag{61a}$$

$$Q_\theta(s,a):=r(s,a;\theta)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V_\theta(s')\big].\tag{61b}$$

According to Cen et al. (2021), the soft value function V_θ and the policy π_θ in the entropy-regularized MDP satisfy the following relations for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$:

$$V_\theta(s)=\log\Big(\sum_{\tilde{a}\sim\mathcal{A}}\exp Q_\theta(s,\tilde{a})\Big),\tag{62a}$$

$$\pi_\theta(a|s)=\frac{\exp\big(Q_\theta(s,a)\big)}{\sum_{\tilde{a}\in\mathcal{A}}\exp\big(Q_\theta(s,\tilde{a})\big)}.\tag{62b}$$

According to (8), we are able to express the objective function $L(\theta)$ in (4a) as below:

$$L(\theta)=\mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t;\theta)\right]-\mathbb{E}_{s_0\sim\rho}\Big[V_\theta(s_0)\Big].\tag{63}$$

Based on (63), we calculate the exact gradient of the objective function $L(\theta)$ as below:

$$\nabla_\theta L(\theta):=\mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right]-\mathbb{E}_{s_0\sim\rho}\Big[\nabla_\theta V_\theta(s_0)\Big]$$

$$\begin{aligned}
&\overset{(i)}{=} \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}\left[\nabla_\theta \log\left(\sum_{\tilde{a}\in\mathcal{A}}\exp Q_\theta(s_0,\tilde{a})\right)\right] \\
&= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}\left[\sum_{a\in\mathcal{A}}\left(\frac{\exp Q_\theta(s_0,a)}{\sum_{\tilde{a}\in\mathcal{A}}\exp Q_\theta(s_0,\tilde{a})}\nabla_\theta Q_\theta(s_0,a)\right)\right] \\
&\overset{(ii)}{=} \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}\left[\sum_{a\in\mathcal{A}}\pi_\theta(a|s_0)\nabla_\theta Q_\theta(s_0,a)\right]. \qquad (64)
\end{aligned}$$

where (i) follows (62a) and (ii) is from (62b). Then we calculate $\nabla_\theta Q_\theta(s_0,a_0)$ as below:

$$\begin{aligned}
&\nabla_\theta Q_\theta(s_0,a_0) \\
&\overset{(i)}{=} \nabla_\theta\left(r(s_0,a_0;\theta) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[V_\theta(s_1)\right]\right) \\
&\overset{(ii)}{=} \nabla_\theta r(s_0,a_0;\theta) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[\nabla_\theta\log\left(\sum_{\tilde{a}\in\mathcal{A}}\exp Q_\theta(s_0,\tilde{a})\right)\right] \\
&= \nabla_\theta r(s_0,a_0;\theta) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[\sum_{a\in\mathcal{A}}\frac{\exp Q_\theta(s_1,a)}{\sum_{\tilde{a}\in\mathcal{A}}\exp Q_\theta(s_1,\tilde{a})}\nabla_\theta Q_\theta(s_1,a)\right] \\
&\overset{(iii)}{=} \nabla_\theta r(s_0,a_0;\theta) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[\sum_{a\in\mathcal{A}}\pi_\theta(a|s_1)\nabla_\theta Q_\theta(s_1,a)\right] \\
&\overset{(iv)}{=} \nabla_\theta r(s_0,a_0;\theta) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0),a_1\sim\pi_\theta(\cdot|s_1)}\left[\nabla_\theta\left(r(s_1,a_1;\theta) + \gamma\mathbb{E}_{s_2\sim P(\cdot|s_1,a_1)}\left[V_\theta(s_2)\right]\right)\right] \\
&\overset{(v)}{=} \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\mid s_0,a_0\right] \qquad (65)
\end{aligned}$$

where (i) and (iv) follows the definition of the soft Q-function, see (14b); (ii) follows (62a); (iii) is from (62b); (v) is shown by recursively applying (i) - (iv).

Finally, plugging equation (65) into (64), the gradient expression of $L(\theta)$ follows:

$$\begin{aligned}
\nabla_\theta L(\theta) &= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}\left[\sum_{a\in\mathcal{A}}\pi_\theta(a|s_0)\nabla_\theta Q_\theta(s_0,a)\right] \\
&= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{s_0\sim\rho}\left[\sum_{a\in\mathcal{A}}\pi_\theta(a|s_0)\cdot\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\mid s_0,a\right]\right] \\
&= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^{\mathrm{E}}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right]. \qquad (66)
\end{aligned}$$

Following the same proof steps, we can also show the gradient expression of the surrogate objective $\widehat{L}(\theta;\mathcal{D})$ as below:

$$\nabla_\theta\widehat{L}(\theta;\mathcal{D}) = \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r(s_t,a_t;\theta)\right]. \qquad (67)$$

This completes the proof of this lemma. ■

17. Proof of Lemma 3

To prove Lemma 3, we show the inequalities (20a) and (20b) respectively. The constants L_q and L_c in Lemma 3 has the expression:

$$L_q := \frac{L_r}{1-\gamma}, \quad L_c := \frac{2L_q L_r C_d \sqrt{|\mathcal{S}| \cdot |\mathcal{A}|}}{1-\gamma} + \frac{2L_g}{1-\gamma},$$

where L_r, L_g are the constants defined in Assumption 3 and C_d is the constant in Lemma 5.

17.1. Proof of Inequality (20a)

Proof. The proof of (20a) consists of two steps: 1) Q_θ has bounded gradient with respect to any reward parameter θ, 2) the inequality (20a) holds due to the mean value theorem.

Recall that in (65), we have shown the explicit expression of $\nabla_\theta Q_\theta(s,a)$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Using this expression, we have the following relations:

$$\begin{aligned}
\|\nabla_\theta Q_\theta(s,a)\| &\overset{(i)}{=} \left\| \mathbb{E}_{\tau^{\mathrm{A}} \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta r(s_t, a_t; \theta) \;\middle|\; (s_0, a_0) = (s,a) \right] \right\| \\
&\overset{(ii)}{\leq} \mathbb{E}_{\tau^{\mathrm{A}} \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \left\| \nabla_\theta r(s_t, a_t; \theta) \right\| \;\middle|\; (s_0, a_0) = (s,a) \right] \\
&\overset{(iii)}{\leq} \mathbb{E}_{\tau^{\mathrm{A}} \sim \pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t L_r \;\middle|\; (s_0, a_0) = (s,a) \right] \\
&= \frac{L_r}{1-\gamma}
\end{aligned} \tag{68}$$

where (i) is from the equality (65) in the proof of Lemma 2, (ii) follows Jensen's inequality and (iii) follows the inequality (19) in Assumption 3. To complete this proof, we use the Mean Value Theorem to show that

$$|Q_{\theta_1}(s,a) - Q_{\theta_2}(s,a)| \leq \| \max_\theta \nabla_\theta Q_\theta(s,a) \| \cdot \|\theta_1 - \theta_2\| \leq L_q \|\theta_1 - \theta_2\|$$

where the last inequality follows (68) and we denote $L_q := \frac{L_r}{1-\gamma}$. Therefore, we have proved the Lipschitz continuous inequality in (20a). ■

17.2. Proof of Inequality (20b)

Proof. In this section, we prove the inequality (20b) in Lemma 3.

According to Lemma 2, the gradient $\nabla_\theta \widehat{L}(\theta;\mathcal{D})$ has the expression as follows:

$$\nabla_\theta \widehat{L}(\theta;\mathcal{D}) = \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta)\right]. \tag{69}$$

Using the above relation, we have

$$\begin{aligned}
&\|\nabla_\theta \widehat{L}(\theta_1;\mathcal{D}) - \nabla_\theta \widehat{L}(\theta_2;\mathcal{D})\| \\
&\overset{(i)}{=} \Bigg\| \left(\mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_1}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right]\right) \\
&\qquad - \left(\mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_2}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\right]\right)\Bigg\| \\
&\le \underbrace{\left\|\mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right] - \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\right]\right\|}_{:=\text{term A}} + \\
&\qquad \underbrace{\left\|\mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_1}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_2}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\right]\right\|}_{:=\text{term B}}
\end{aligned} \tag{70}$$

where (i) follows the exact gradient expression in equation (69). Then we separately analyze term A and term B in (70).

For term A, it follows that

$$\begin{aligned}
&\left\|\mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right] - \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\right]\right\| \\
&\overset{(i)}{\le} \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t \left\|\nabla_\theta r(s_t,a_t;\theta_1) - \nabla_\theta r(s_t,a_t;\theta_2)\right\|\right] \\
&\overset{(ii)}{\le} \mathbb{E}_{\tau^{\mathrm{E}}\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\gamma^t L_g\|\theta_1-\theta_2\|\right] \\
&= \frac{L_g}{1-\gamma}\|\theta_1-\theta_2\|
\end{aligned} \tag{71}$$

where (i) follows Jensen's inequality and (ii) is from (19) in Assumption 3.

For the term B, it holds that

$$\left\|\mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_1}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_2}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\right]\right\|$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \left\| \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_1}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_2}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right]\right\| \\
&\quad + \left\| \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_2}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_1)\right] - \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_2}}\left[\sum_{t=0}^{\infty}\gamma^t \nabla_\theta r(s_t,a_t;\theta_2)\right]\right\| \\
&\overset{(ii)}{\leq} \frac{1}{1-\gamma}\left\| \mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{\theta_1})}\left[\nabla_\theta r(s_t,a_t;\theta_1)\right] - \mathbb{E}_{(s,a)\sim d(\cdot,\cdot;\pi_{\theta_2})}\left[\nabla_\theta r(s_t,a_t;\theta_1)\right]\right\| \\
&\quad + \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_2}}\left[\sum_{t=0}^{\infty}\gamma^t \left\|\nabla_\theta r(s_t,a_t;\theta_1) - \nabla_\theta r(s_t,a_t;\theta_2)\right\|\right] \\
&\overset{(iii)}{\leq} \frac{1}{1-\gamma}\left\| \sum_{s\in\mathcal{S},a\in\mathcal{A}} \nabla_\theta r(s_t,a_t;\theta_1)\Big(d(s,a;\pi_{\theta_1}) - d(s,a;\pi_{\theta_2})\Big)\right\| + \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{\theta_2}}\left[\sum_{t=0}^{\infty}\gamma^t L_g\|\theta_1-\theta_2\|\right] \\
&\overset{(iv)}{\leq} \frac{2L_r}{1-\gamma}\|d(\cdot,\cdot;\pi_{\theta_1}) - d(\cdot,\cdot;\pi_{\theta_1})\|_{TV} + \frac{L_g}{1-\gamma}\|\theta_1-\theta_2\| \qquad (72)
\end{aligned}
$$

where (i) follows the triangle inequality, (ii) is from Jensen's inequality and the definition of the discounted state-action visitation measure $d(s,a;\pi) := (1-\gamma)\pi(a|s)\sum_{t=0}^{\infty}\gamma^t P^{\pi}(s_t = s|s_0\sim\rho)$; (iii) is from (19) in Assumption 3; (iv) is from (19) and the definition of the total variation norm.

Plugging the inequalities (71), (72) to (70), it holds that

$$
\begin{aligned}
\|\nabla_\theta \widehat{L}(\theta_1;\mathcal{D}) - \nabla_\theta \widehat{L}(\theta_2;\mathcal{D})\| &\leq \frac{2L_r}{1-\gamma}\|d(\cdot,\cdot;\pi_{\theta_1}) - d(\cdot,\cdot;\pi_{\theta_2})\|_{TV} + \frac{2L_g}{1-\gamma}\|\theta_1-\theta_2\| \\
&\overset{(i)}{\leq} \frac{2L_r C_d}{1-\gamma}\|Q_{\theta_1} - Q_{\theta_2}\| + \frac{2L_g}{1-\gamma}\|\theta_1-\theta_2\| \\
&\overset{(ii)}{\leq} \frac{2L_r C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma}\|Q_{\theta_1} - Q_{\theta_2}\|_\infty + \frac{2L_g}{1-\gamma}\|\theta_1-\theta_2\| \\
&\overset{(iii)}{\leq} \left(\frac{2L_q L_r C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma} + \frac{2L_g}{1-\gamma}\right)\|\theta_1-\theta_2\|. \qquad (73)
\end{aligned}
$$

Given the fact that π_θ is a softmax policy parameterized by Q_θ where $\pi_\theta(a|s)\propto\exp(Q_\theta(s,a))$, we show the inequality (i) from the inequality (30) in Lemma 5. Moreover, the inequality (ii) follows the conversion between the Frobenius norm and the infinty norm, where the inequality $|Q_{\theta_1}(s,a) - Q_{\theta_2}(s,a)| \leq \|Q_{\theta_1} - Q_{\theta_2}\|_\infty$ holds for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$ so that $\|Q_{\theta_1} - Q_{\theta_2}\| \leq \sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}\ \|Q_{\theta_1} - Q_{\theta_2}\|_\infty$. Last, (iii) is from the inequality (20a) in Lemma 3.

Define the constant $L_c := \frac{2L_q L_r C_d\sqrt{|\mathcal{S}|\cdot|\mathcal{A}|}}{1-\gamma} + \frac{2L_g}{1-\gamma}$, we have the following inequality:

$$
\|\nabla_\theta \widehat{L}(\theta_1;\mathcal{D}) - \nabla_\theta \widehat{L}(\theta_2;\mathcal{D})\| \leq L_c\|\theta_1-\theta_2\|.
$$

Therefore, we complete the proof of the inequality (20b) in Lemma 3. ■

18. Proof of Lemma 4

Proof. Suppose the expert trajectories τ in (4a) is sampled from an expert policy π^E. Moreover, we parameterize the state-only reward as $r(s;\theta)$. Then the objective function $L(\theta)$ can be rewritten as follows:

$$\begin{aligned} L(\theta) &:= \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^E}\left[\sum_{t=0}^{\infty}\gamma^t \log \pi_\theta(a_t|s_t)\right] \\ &\overset{(i)}{=} \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t;\theta)\right] - \mathbb{E}_{s_0\sim\eta(\cdot)}\Big[V_\theta(s_0)\Big] \\ &\overset{(ii)}{=} \mathbb{E}_{s_0\sim\eta(\cdot)}\big[V^{\mathrm{E}}(s_0)\big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\big[V_\theta(s_0)\big] - \mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^E}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{H}(\cdot|s_t)\right] \end{aligned} \tag{74}$$

where (i) follows (63) and the fact that the reward is a state-only function $r(s;\theta)$; (ii) follows the definitions of the soft value function.

Ignoring the constant term $\mathbb{E}_{\tau^{\mathrm{E}}\sim\pi^E}\big[\sum_{t=0}^{\infty}\gamma^t\mathcal{H}(\cdot|s_t)\big]$ in (74), the maximum likelihood formulation (4a) is equivalent to the following bi-level problem:

$$\begin{aligned} \min_{\theta} \quad & \mathbb{E}_{s_0\sim\eta(\cdot)}\big[V_\theta(s_0)\big] - \mathbb{E}_{s_0\sim\eta(\cdot)}\big[V^{\mathrm{E}}_\theta(s_0)\big] \\ \text{s.t.} \quad & \pi_\theta := \arg\max_{\pi}\ \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t;\theta)+\mathcal{H}(\pi(\cdot|s_t))\Big)\right]. \end{aligned}$$

Therefore, we complete the proof of Lemma 4. As an alternative interpretation to (4), the formulation above aims to minimize the gap between the soft value function of π_θ and π^E under the state-only IRL setting. ■

19. Proof of Lemma 6

Proof. Based on the definition of soft Q-functions $Q_{k+\frac{1}{2}}$ and Q_{k+1}, it follows

$$Q_{k+\frac{1}{2}}(s,a) := r(s,a;\theta_k) + \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=1}^{\infty}\gamma^t\Big(r(s_t,a_t;\theta_k)+\mathcal{H}(\pi_{k+1}(\cdot|s_t))\Big)\Big|(s_0,a_0)=(s,a)\right], \tag{75}$$

$$Q_{k+1}(s,a) := r(s,a;\theta_{k+1}) + \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=1}^{\infty}\gamma^t\Big(r(s_t,a_t;\theta_{k+1})+\mathcal{H}(\pi_{k+1}(\cdot|s_t))\Big)\Big|(s_0,a_0)=(s,a)\right]. \tag{76}$$

Then it holds that

$$|Q_{k+\frac{1}{2}}(s,a) - Q_{k+1}(s,a)| = \left|\mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t,a_t;\theta_k)-r(s_t,a_t;\theta_{k+1})\Big)\Big|(s_0,a_0)=(s,a)\right]\right|$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t\left|r(s_t,a_t;\theta_k)-r(s_t,a_t;\theta_{k+1})\right|\middle|(s_0,a_0)=(s,a)\right]\\
&\overset{(ii)}{=} \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t\left|\left\langle\nabla_\theta r(s_t,a_t;\theta^c),\theta_k-\theta_{k+1}\right\rangle\right|\middle|(s_0,a_0)=(s,a)\right]\\
&\overset{(iii)}{=} \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t\left\|\nabla_\theta r(s_t,a_t;\theta^c)\right\|\cdot\left\|\theta_k-\theta_{k+1}\right\|\middle|(s_0,a_0)=(s,a)\right]\\
&\leq \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t\left\|\max_\theta\nabla_\theta r(s_t,a_t;\theta)\right\|\cdot\left\|\theta_k-\theta_{k+1}\right\|\middle|(s_0,a_0)=(s,a)\right]\\
&\overset{(iv)}{\leq} \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t L_r\|\theta_k-\theta_{k+1}\|\right]\\
&= \frac{L_r}{1-\gamma}\|\theta_k-\theta_{k+1}\| \qquad (77)
\end{aligned}
$$

where (i) follows Jensen's inequality; (ii) follows the mean value theorem where θ^c is a convex combination of θ_k and θ_{k+1}; (iii) follows the CauchySchwarz inequality; (iv) follows inequality (19) in Assumption 3. ■

20. Proof of Lemma 7

In this section, we prove the inequalities (32) and (33) respectively.

20.1. Proof of Inequality (32)

Proof. Recall the definitions of the soft value function and the soft Q-function in (14a) - (14b). Moreover, we also defined the soft Q-function $Q_{k+\frac{1}{2}}$ in (75). Similarly, let us define the corresponding soft value function $V_{k+\frac{1}{2}}$ (under reward parameter θ_k and policy π_{k+1}) as:

$$
V_{k+\frac{1}{2}}(s) := \mathbb{E}_{\tau^{\mathrm{A}}\sim\pi_{k+1}}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t,a_t;\theta_k)+\mathcal{H}(\pi_{k+1}(\cdot|s_t))\Big)\middle|s_0=s\right],\quad \forall s\in\mathcal{S}. \qquad (78)
$$

According to the definitions of V_k in (14a) and $V_{k+\frac{1}{2}}$ in (78), we could rewrite Q_k and $Q_{k+\frac{1}{2}}$ as:

$$
\begin{aligned}
Q_k(s,a) &:= r(s,a;\theta_k)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V_k(s')\big],\\
Q_{k+\frac{1}{2}}(s,a) &:= r(s,a;\theta_k)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V_{k+\frac{1}{2}}(s')\big].
\end{aligned}
$$

According to the expressions above, the following relation holds for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$:

$$
Q_k(s,a)-Q_{k+\frac{1}{2}}(s,a)=\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V_k(s')-V_{k+\frac{1}{2}}(s')\big]. \qquad (79)
$$

In order to measure the difference between $Q_{k+\frac{1}{2}}$ and Q_k, we need to bound the gap between $V_{k+\frac{1}{2}}$ and V_k. Here, we could define an auxiliary sequence $\{\tilde{\pi}_k\}_{k\geq 0}$ generated as below:

$$\tilde{\pi}_{k+1}(\cdot|s) \propto \exp\big(Q_k(s,\cdot)\big) \quad \forall s \in \mathcal{S}. \tag{80}$$

As a comparison, let us recall the approximated soft policy iteration (15):

$$\pi_{k+1}(a|s) \propto \exp\big(\widehat{Q}_k(s,a)\big), \text{ where } \|\widehat{Q}_k - Q_k\|_\infty \leq \epsilon_{\text{app}}. \tag{81}$$

Then for any $s \in \mathcal{S}$, we have the following series of relations:

$$\begin{aligned}
V_k(s) &\overset{(i)}{=} \mathbb{E}_{a\sim\pi_k(\cdot|s)}\big[-\log\pi_k(a|s) + Q_k(s,a)\big] \\
&= \mathbb{E}_{a\sim\pi_k(\cdot|s)}\big[-\log\tilde{\pi}_{k+1}(a|s) + Q_k(s,a)\big] + \mathbb{E}_{a\sim\pi_k(\cdot|s)}\big[\log\tilde{\pi}_{k+1}(a|s) - \log\pi_k(a|s)\big] \\
&\overset{(ii)}{=} \mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[\log\left(\sum_{\tilde{a}\in\mathcal{A}}\exp\big(Q_k(s,\tilde{a})\big)\right)\right] - D_{KL}\Big(\pi_k(\cdot|s) \;||\; \tilde{\pi}_{k+1}(\cdot|s)\Big) \\
&\overset{(iii)}{\leq} \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\left[\log\left(\sum_{\tilde{a}\in\mathcal{A}}\exp\big(Q_k(s,\tilde{a})\big)\right)\right] \\
&\overset{(iv)}{=} \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\Big[-\log\tilde{\pi}_{k+1}(a|s) + Q_k(s,a)\Big] \\
&= \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\Big[-\log\pi_{k+1}(a|s) + Q_k(s,a)\Big] + \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\Big[\log\pi_{k+1}(a|s) - \log\tilde{\pi}_{k+1}(a|s)\Big] \\
&\overset{(iv)}{\leq} \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\Big[-\log\pi_{k+1}(a|s) + Q_k(s,a)\Big] + \|\log\pi_{k+1} - \log\tilde{\pi}_{k+1}\|_\infty
\end{aligned} \tag{82}$$

where (i) follows the definition of the soft value function V_k; (ii) follows the fact in (80) that $\tilde{\pi}_{k+1}(a|s) := \frac{\exp Q_k(s,a)}{\sum_{\tilde{a}\in\mathcal{A}}\exp Q_k(s,\tilde{a})}$ and the definition of the KL divergence; (iii) follows the non-negativity of the KL divergence and the fact that $\log\left(\sum_{\tilde{a}\in\mathcal{A}}\exp\big(Q_k(s,\tilde{a})\big)\right)$ is independent of any action $a \in \mathcal{A}$; (iv) follows the fact that $-\log\tilde{\pi}_{k+1}(a|s) = \log\big(\sum_{\tilde{a}\in\mathcal{A}}\exp Q_k(s,\tilde{a})\big) - Q_k(s,a)$, which is derived from the definition of $\tilde{\pi}_{k+1}$ in (80); (iv) follows the definition of the infinity norm $\|\cdot\|_\infty$ such that $\|\log\pi_{k+1} - \log\tilde{\pi}_{k+1}\|_\infty = \max_{s\in\mathcal{S}}\max_{a\in\mathcal{A}}\big(\log\pi_{k+1}(a|s) - \log\tilde{\pi}_{k+1}(a|s)\big)$.

We further analyze the approximation error $\|\log\pi_{k+1} - \log\tilde{\pi}_{k+1}\|_\infty$ in (82). We first show that for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the following relations hold:

$$\big|\log\pi_{k+1}(a|s) - \log\tilde{\pi}_{k+1}(a|s)\big|$$

$$\begin{aligned}
&\overset{(i)}{=} \left| \log\left(\frac{\exp \widehat{Q}_k(s,a)}{\sum_{\tilde{a}\in\mathcal{A}} \exp \widehat{Q}_k(s,\tilde{a})}\right) - \log\left(\frac{\exp Q_k(s,a)}{\sum_{\tilde{a}\in\mathcal{A}} \exp Q_k(s,\tilde{a})}\right)\right| \\
&\overset{(ii)}{\leq} \left|\widehat{Q}_k(s,a) - Q_k(s,a)\right| + \left|\log\left(\sum_{\tilde{a}\in\mathcal{A}} \exp \widehat{Q}_k(s,\tilde{a})\right) - \log\left(\sum_{\tilde{a}\in\mathcal{A}} \exp Q_k(s,\tilde{a})\right)\right| \qquad (83)
\end{aligned}$$

where (i) follows (80) and (81); (ii) follows the triangle inequality. We further analyze the second term in (83).

We first define a short-handed notation $\log(\|\exp(v)\|_1) := \log(\|\sum_{\tilde{a}\in\mathcal{A}} \exp(v_{\tilde{a}})\|_1)$, where the vector $v \in \mathbb{R}^{|\mathcal{A}|}$ and $v = [v_1, v_2, \cdots, v_{|\mathcal{A}|}]$. Then for any $v', v'' \in \mathbb{R}^{|\mathcal{A}|}$, we have the following relation:

$$\begin{aligned}
\left|\log\left(\|\exp(v')\|_1\right) - \log\left(\|\exp(v'')\|_1\right)\right| &\overset{(i)}{=} \left|\left\langle v'-v'', \nabla_v \log\left(\|\exp(v)\|_1\right)|_{v=v^c}\right\rangle\right| \\
&\leq \|v'-v''\|_\infty \cdot \|\nabla_v \log\left(\|\exp(v)\|_1\right)|_{v=v^c}\|_1 \\
&\overset{(ii)}{=} \|v'-v''\|_\infty \qquad (84)
\end{aligned}$$

where (i) follows the mean value theorem and v_c is a solution lies between v' and v''; (ii) follows the following relations:

$$[\nabla_v \log\left(\|\exp(v)\|_1\right)]_i = \frac{\exp(v_i)}{\sum_{1\leq a\leq|\mathcal{A}|}\exp(v_a)}, \quad \|\nabla_v \log\left(\|\exp(v)\|_1\right)\|_1 = 1, \quad \forall\ v \in \mathbb{R}^{|\mathcal{A}|}.$$

Through plugging (84) into (83), it holds that

$$\left|\log \pi_{k+1}(a|s) - \log \tilde{\pi}_{k+1}(a|s)\right| \leq \left|\widehat{Q}_k(s,a) - Q_k(s,a)\right| + \max_{\tilde{a}\in\mathcal{A}} \left|\widehat{Q}_k(s,\tilde{a}) - Q_k(s,\tilde{a})\right|. \qquad (85)$$

Taking the infinity norm over $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, the following result holds:

$$\|\log \pi_{k+1} - \log \tilde{\pi}_{k+1}\|_\infty \leq 2\|\widehat{Q}_k - Q_k\|_\infty. \qquad (86)$$

Through plugging (86) into (82), it follows that

$$\begin{aligned}
V_k(s) &\leq \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\Big[-\log \pi_{k+1}(a|s) + Q_k(s,a)\Big] + \|\log \pi_{k+1} - \log \tilde{\pi}_{k+1}\|_\infty \\
&\overset{(i)}{\leq} \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s)}\Big[-\log \pi_{k+1}(a|s) + Q_k(s,a)\Big] + 2\|\widehat{Q}_k - Q_k\|_\infty \\
&\overset{(ii)}{=} \mathbb{E}_{a\sim\pi_{k+1}(\cdot|s), s'\sim P(\cdot|s,a)}\Big[-\log \pi_{k+1}(a|s) + r(s,a;\theta_k) + \gamma V_k(s')\Big] + 2\epsilon_{\text{app}} \qquad (87)
\end{aligned}$$

where (i) is from (86); (vi) follows the definition of the soft Q-function and the fact that the approximation error $\|\widehat{Q}_k - Q_k\|_\infty$ is bounded by ϵ_{app}. By recursively using the inequality (87), the following result hold for any $s \in \mathcal{S}$:

$$V_k(s) \leq \mathbb{E}_{\tau^{\text{A}} \sim \pi_{k+1}} \left[\sum_{t=0}^{\infty} \Big(-\log \pi_{k+1}(a_t|s_t) + r(s_t, a_t; \theta_k) \Big) \Big| s_0 = s \right] + \frac{2\epsilon_{\text{app}}}{1-\gamma} \overset{(i)}{=} V_{k+\frac{1}{2}}(s) + \frac{2\epsilon_{\text{app}}}{1-\gamma} \quad (88)$$

where (i) follows the definition of $V_{k+\frac{1}{2}}$ in (78). By plugging (88) into (79), we finish the proof. ■

20.2. Proof of Inequality (33)

Proof. For any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the following results hold

$$\begin{aligned}
&Q_{\theta_k}(s,a) - Q_{k+\frac{1}{2}}(s,a) \\
&\overset{(i)}{=} \left(r(s,a;\theta_k) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\log \big(\sum_{\tilde{a} \in \mathcal{A}} \exp Q_{\theta_k}(s', \tilde{a}) \big) \right] \right) \\
&\quad - \left(r(s,a;\theta_k) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi_{k+1}(\cdot|s')} \left[-\log \pi_{k+1}(a'|s') + Q_{k+\frac{1}{2}}(s',a') \right] \right) \\
&\overset{(ii)}{=} \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\log \big(\sum_{\tilde{a} \in \mathcal{A}} \exp Q_{\theta_k}(s', \tilde{a}) \big) \right] - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi_{k+1}(\cdot|s')} \left[-\log \left(\frac{\exp \widehat{Q}_k(s',a')}{\sum_{\tilde{a} \in \mathcal{A}} \exp \widehat{Q}_k(s', \tilde{a})} \right) + Q_{k+\frac{1}{2}}(s',a') \right] \\
&= \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\log \big(\sum_{\tilde{a} \in \mathcal{A}} \exp Q_{\theta_k}(s', \tilde{a}) \big) - \log \big(\sum_{\tilde{a} \in \mathcal{A}} \exp \widehat{Q}_k(s', \tilde{a}) \big) \right] \\
&\quad - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi_{k+1}(\cdot|s')} \left[Q_{k+\frac{1}{2}}(s',a') - \widehat{Q}_k(s',a') \right] \\
&\overset{(iii)}{\leq} \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{\tilde{a} \in \mathcal{A}} \big| Q_{\theta_k}(s', \tilde{a}) - \widehat{Q}_k(s', \tilde{a}) \big| \right] - \gamma \min_{s \in \mathcal{S}, a \in \mathcal{A}} \left(Q_{k+\frac{1}{2}}(s,a) - \widehat{Q}_k(s,a) \right) \\
&\overset{(iv)}{\leq} \gamma \big\| Q_{\theta_k} - \widehat{Q}_k \big\|_\infty - \gamma \min_{s \in \mathcal{S}, a \in \mathcal{A}} \left(Q_{k+\frac{1}{2}}(s,a) - \widehat{Q}_k(s,a) \right) \\
&\leq \gamma \|\widehat{Q}_k - Q_k\|_\infty + \gamma \|Q_{\theta_k} - Q_k\|_\infty - \gamma \min_{s \in \mathcal{S}, a \in \mathcal{A}} \left(Q_{k+\frac{1}{2}}(s,a) - \widehat{Q}_k(s,a) \right) \\
&\leq \gamma \epsilon_{\text{app}} + \gamma \|Q_{\theta_k} - Q_k\|_\infty - \gamma \min_{s \in \mathcal{S}, a \in \mathcal{A}} \left(Q_{k+\frac{1}{2}}(s,a) - \widehat{Q}_k(s,a) \right) \qquad (89)
\end{aligned}$$

where (i) follows (61b), (62a) and the definition of the soft Q-function $Q_{k+\frac{1}{2}}$ in (75); (ii) is from the definition of π_{k+1} in (15); (iii) follows (84); (iv) follows the definition of the infinity norm such that $\big\| Q_{\theta_k} - \widehat{Q}_k \big\|_\infty = \max_{s \in \mathcal{S}} \max_{a \in \mathcal{A}} |Q_{\theta_k}(s,a) - \widehat{Q}_k(s,a)|$.

Given the fact that Q_{θ_k} is the soft Q-function under the reward function $r(\cdot, \cdot; \theta_k)$ and the optimal policy π_{θ_k}, we have the relation that

$$Q_{\theta_k}(s,a) - Q_{k+\frac{1}{2}}(s,a) \geq 0, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \qquad (90)$$

This implies that

$$\max_{s\in\mathcal{S},a\in\mathcal{A}}\{Q_{\theta_k}(s,a)-Q_{k+\frac{1}{2}}(s,a)\}=\|Q_{\theta_k}-Q_{k+\frac{1}{2}}\|_\infty. \tag{91}$$

Combining (90) and (89), we bound the absolute difference between Q_{θ_k} and $Q_{k+\frac{1}{2}}$ as below:

$$\|Q_{\theta_k}-Q_{k+\frac{1}{2}}\|_\infty\le\gamma\epsilon_{\text{app}}+\gamma\|Q_{\theta_k}-Q_k\|_\infty-\gamma\min_{s\in\mathcal{S},a\in\mathcal{A}}\left(Q_{k+\frac{1}{2}}(s,a)-\widehat{Q}_k(s,a)\right). \tag{92}$$

Then we further bound the last term above. For any $s\in\mathcal{S}$ and $a\in\mathcal{A}$, note that:

$$\begin{aligned}
&Q_{k+\frac{1}{2}}(s,a)-\widehat{Q}_k(s,a)\\
&=\big(Q_{k+\frac{1}{2}}(s,a)-Q_k(s,a)\big)+\big(Q_k(s,a)-\widehat{Q}_k(s,a)\big)\\
&\ge-\frac{2\gamma\epsilon_{\text{app}}}{1-\gamma}-\epsilon_{\text{app}}
\end{aligned}$$

where the last inequality follows (32) and the definition of the approximation error ϵ_{app}.

Through plugging the inequality above into (92), it holds that:

$$\|Q_{\theta_k}-Q_{k+\frac{1}{2}}\|_\infty\le\gamma\|Q_{\theta_k}-Q_k\|_\infty+\gamma\epsilon_{\text{app}}+\gamma\big(\epsilon_{\text{app}}+\frac{2\gamma\epsilon_{\text{app}}}{1-\gamma}\big)=\gamma\|Q_{\theta_k}-Q_k\|_\infty+\frac{2\gamma\epsilon_{\text{app}}}{1-\gamma}. \tag{93}$$

This completes the proof of the lemma. ■