Supplementary Materials

A Q-sampling Algorithm

In this section, we provide the formal description for the algorithm EstQ in Algorithm 3, which returns an unbiased estimation of the state-action value function (Q-value).

Algorithm 3 EstQ (Zhang et al., 2019)

1: **Input:** s, a, θ. Initialize $\hat{Q} = 0, s_1^q = s, a_1^q = a$
2: Draw $T \sim \text{Geom}(1 - \gamma^{1/2})$
3: **for** $t = 1, 2, \ldots, T-1$ **do**
4: Collect reward $R(s_t^q, a_t^q)$ and update the Q-function $\hat{Q} \leftarrow \hat{Q} + \gamma^{t/2} R(s_t^q, a_t^q)$
5: Sample $s_{t+1}^q \sim \mathbb{P}(\cdot|s_t^q, a_t^q), a_{t+1}^q \sim \pi_\theta(\cdot|s_{t+1}^q)$
6: **end for**
7: Collect reward $R(s_T^q, a_T^q)$ and update the Q-function $\hat{Q} \leftarrow \hat{Q} + \gamma^{T/2} R(s_T^q, a_T^q)$
8: **Output:** $\hat{Q}^{\pi_\theta} \leftarrow \hat{Q}$

B Proof of Proposition 1

In this section, we first provide two useful lemmas, which establish the smoothness property of the visitation distribution and Q-function.

Lemma 1. *((Xu et al., 2020a, Lemma 3)) Consider the initial distribution $\xi(\cdot)$ and the transition kernel $\mathsf{P}(\cdot|s,a)$. Let $\xi(\cdot)$ be $\zeta(\cdot)$ or $\mathsf{P}(\cdot|\hat{s},\hat{a})$ for any given $\hat{s} \in \mathcal{S}, \hat{a} \in \mathcal{A}$. Denote $\nu_{\pi_\theta,\xi}$ as the state-action visitation distribution of MDP with policy π_θ and the initialization distribution ξ. Suppose Assumption 3 holds. Then we have, under direct parameterization for any $\theta_1, \theta_2 \in \Theta_p$,*

$$\left\|\nu_{\pi_\theta,\xi} - \nu_{\pi_{\theta'},\xi}\right\|_{TV} \leq C_\nu \left\|\theta_1 - \theta_2\right\|_2,$$

where $C_\nu = \frac{\sqrt{|\mathcal{A}|}}{2}\left(1 + \lceil \log_\rho C_M^{-1} \rceil + (1-\rho)^{-1}\right)$.

Lemma 2. *((Xu et al., 2020a, Lemma 4)) Suppose Assumptions 3 and 4 hold. Let Q_α^π denote the Q-function of policy π under the reward function r_α. For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\alpha \in \Lambda$ and $\theta_1, \theta_2 \in \Theta_p$ (under direct parameterization), we have*

$$|Q_\alpha^{\pi_{\theta_1}}(s,a) - Q_\alpha^{\pi_{\theta_2}}(s,a)| \leq L_Q \left\|\theta_1 - \theta_2\right\|_2,$$

where $L_Q = \frac{2C_r C_\alpha C_\nu}{1-\gamma}$ and C_ν is defined in Lemma 1.

Denote $d_\pi(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}\{s_t = s|\pi\}$ as the state visitation distribution induced by policy π. We next prove Proposition 1 to characterize the Lipschitz constants L_{11}, L_{12}, L_{21} and L_{22}, respectively.

Proof of Proposition 1. We consider the first inequality in Proposition 1:

$$\begin{aligned}\left\|\nabla_\theta F(\theta_1,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_2)\right\|_2 &= \left\|\nabla_\theta F(\theta_1,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_1) + \nabla_\theta F(\theta_2,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_2)\right\|_2 \\ &\leq \underbrace{\left\|\nabla_\theta F(\theta_1,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_1)\right\|_2}_{T_1} + \underbrace{\left\|\nabla_\theta F(\theta_2,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_2)\right\|_2}_{T_2}. \end{aligned} \tag{9}$$

Next, we upper-bound the terms T_1 and T_2 in eq. (9), respectively.

Upper-bounding T_1: For any given state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
&\left|\left(\nabla_\theta F(\theta_1,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_1)\right)_{s,a}\right| \\
&\overset{(i)}{=} \left| \frac{1}{1-\gamma}\left(d_{\pi_{\theta_1}}(s) Q_{\alpha_1}^{\pi_{\theta_1}}(s,a) - d_{\pi_{\theta_2}}(s) Q_{\alpha_1}^{\pi_{\theta_2}}(s,a)\right)\right| \\
&\leq \left| \frac{1}{1-\gamma}\left((d_{\pi_{\theta_1}}(s) - d_{\pi_{\theta_2}}(s)) Q_{\alpha_1}^{\pi_{\theta_1}}(s,a)\right)\right| + \left| \frac{1}{1-\gamma}\left(d_{\pi_{\theta_2}}(s) (Q_{\alpha_1}^{\pi_{\theta_1}}(s,a) - Q_{\alpha_1}^{\pi_{\theta_2}}(s,a))\right)\right| \\
&\overset{(ii)}{\leq} \frac{R_{max}}{(1-\gamma)^2} |d_{\pi_{\theta_1}}(s) - d_{\pi_{\theta_2}}(s)| + \frac{L_Q}{1-\gamma} d_{\pi_{\theta_2}}(s) \left\|\theta_1 - \theta_2\right\|_2,
\end{aligned} \tag{10}$$

where (i) follows from the fact that $\frac{\partial F(\theta,\alpha_1)}{\partial \theta_{s,a}} = -\frac{\partial V(\pi_\theta,\alpha_1)}{\partial \theta_{s,a}} = -\frac{1}{1-\gamma} d_{\pi_\theta}(s) Q_{\alpha_1}^{\pi_\theta}(s,a)$, and (ii) follows from Lemma 2.

Then, we proceed as follows:

$$\begin{aligned}
&\left\|\nabla_\theta F(\theta_1,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_1)\right\|_2 \\
&= \sqrt{\sum_{s,a} \left|\left(\nabla_\theta F(\theta_1,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_1)\right)_{s,a}\right|^2} \\
&\overset{(i)}{\leq} \sqrt{\sum_{s,a}\left(\frac{R_{max}}{(1-\gamma)^2}\left|d_{\pi_{\theta_1}}(s) - d_{\pi_{\theta_2}}(s)\right| + \frac{L_Q}{1-\gamma} d_{\pi_{\theta_2}}(s)\left\|\theta_1-\theta_2\right\|_2\right)^2} \\
&\leq \sqrt{2|\mathcal{A}|}\sqrt{\sum_s\left(\frac{R_{max}}{(1-\gamma)^2}\left|d_{\pi_{\theta_1}}(s) - d_{\pi_{\theta_2}}(s)\right|\right)^2} + \sqrt{2|\mathcal{A}|}\sqrt{\sum_s\left(\frac{L_Q}{1-\gamma} d_{\pi_{\theta_2}}(s)\left\|\theta_1-\theta_2\right\|_2\right)^2} \\
&\overset{(ii)}{\leq} \sqrt{2|\mathcal{A}|}\left(\sum_s \frac{R_{max}}{(1-\gamma)^2}\left|d_{\pi_{\theta_1}}(s) - d_{\pi_{\theta_2}}(s)\right| + \sum_s \frac{L_Q}{1-\gamma} d_{\pi_{\theta_2}}(s)\left\|\theta_1-\theta_2\right\|_2\right) \\
&\overset{(iii)}{\leq} \frac{2\sqrt{2}|\mathcal{A}|C_r C_\alpha}{(1-\gamma)^2}\left(1 + \lceil \log_\rho C_M^{-1}\rceil + (1-\rho)^{-1}\right)\left\|\theta_1-\theta_2\right\|_2,
\end{aligned}$$

where (i) follows from eq. (10), (ii) follows from the fact that $\|x\|_2 \leq \|x\|_1$, and (iii) follows from Lemma 1 and from the facts $R_{max} \leq C_r C_\alpha$ and

$$\sum_{s\in\mathcal{S}} \left|d_{\pi_{\theta_1}}(s) - d_{\pi_{\theta_2}}(s)\right| = 2\left\|d_{\pi_{\theta_1}} - d_{\pi_{\theta_2}}\right\|_{TV} \leq 2\left\|\nu_{\pi_{\theta_1}} - \nu_{\pi_{\theta_2}}\right\|_{TV}.$$

Upper-bounding T_2: For any given state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
\left|\left(\nabla_\theta F(\theta_2,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_2)\right)_{s,a}\right| &= \left|\frac{1}{1-\gamma}\left(d_{\pi_{\theta_2}}(s) Q_{\alpha_1}^{\pi_{\theta_2}}(s,a) - d_{\pi_{\theta_2}}(s) Q_{\alpha_2}^{\pi_{\theta_2}}(s,a)\right)\right| \\
&\overset{(i)}{=} \frac{1}{1-\gamma} d_{\pi_{\theta_2}}(s)\left|\frac{1}{1-\gamma}\sum_{\hat{s},\hat{a}} \nu_{\pi_{\theta_2},s,a}(\hat{s},\hat{a})(r_{\alpha_1}(\hat{s},\hat{a}) - r_{\alpha_2}(\hat{s},\hat{a}))\right| \\
&\overset{(ii)}{\leq} \frac{1}{(1-\gamma)^2} d_{\pi_{\theta_2}}(s) C_r \left\|\alpha_1 - \alpha_2\right\|_2,
\end{aligned}$$

where in (i) we denote $\nu_{\pi_{\theta_2},s,a}(\hat{s},\hat{a})$ as the visitation distribution of the Markov chain with initial distribution $\mathsf{P}(\cdot|s_0 = s, a_0 = a)$ and policy π_{θ_2}, and (ii) follows from the fact that $|r_{\alpha_1}(\hat{s},\hat{a}) - r_{\alpha_2}(\hat{s},\hat{a})| = |\langle \nabla_\alpha r_{\alpha'}(\hat{s},\hat{a}), \alpha_1 - \alpha_2\rangle| \leq \|\nabla_\alpha r_{\alpha'}(\hat{s},\hat{a})\|_2 \|\alpha_1 - \alpha_2\|_2 \leq C_r \|\alpha_1 - \alpha_2\|_2$, for some $\alpha' \in [\alpha_1, \alpha_2]$. The inequality above implies that

$$\left\|\nabla_\theta F(\theta_2,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_2)\right\|_2 = \sqrt{\sum_{s,a}\left|\left(\nabla_\theta F(\theta_2,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_2)\right)_{s,a}\right|^2}$$

$$\leq \sqrt{\sum_{s,a}\left(\frac{1}{(1-\gamma)^2} d_{\pi_{\theta_2}}(s) C_r \left\|\alpha_1-\alpha_2\right\|_2\right)^2}$$
$$= \frac{\sqrt{|\mathcal{A}|}C_r}{(1-\gamma)^2}\left\|\alpha_1-\alpha_2\right\|_2 \sqrt{\sum_s \left(d_{\pi_{\theta_2}}(s)\right)^2}$$
$$\overset{(i)}{\leq} \frac{\sqrt{|\mathcal{A}|}C_r}{(1-\gamma)^2}\left\|\alpha_1-\alpha_2\right\|_2,$$

where (i) follows from the fact that $\sqrt{\sum_s \left(d_{\pi_{\theta_2}}(s)\right)^2} \leq \left\|d_{\pi_{\theta_2}}\right\|_1 = 1$.

Therefore we obtain the upper bound of eq. (9) as follows:

$$\left\|\nabla_\theta F(\theta_1,\alpha_1)-\nabla_\theta F(\theta_2,\alpha_2)\right\|_2 \leq \frac{2\sqrt{2}|\mathcal{A}|C_r C_\alpha}{(1-\gamma)^2}\left(1+\lceil\log_\rho C_M^{-1}\rceil+(1-\rho)^{-1}\right)\left\|\theta_1-\theta_2\right\|_2 + \frac{\sqrt{|\mathcal{A}|}C_r}{(1-\gamma)^2}\left\|\alpha_1-\alpha_2\right\|_2,$$

which determines the constants L_{11} and L_{12}.

We then proceed to prove the second inequality in Proposition 1.

$$\begin{aligned}\left\|\nabla_\alpha F(\theta_1,\alpha_1)-\nabla_\alpha F(\theta_2,\alpha_2)\right\|_2 &\leq \left\|\nabla_\alpha F(\theta_1,\alpha_1)-\nabla_\alpha F(\theta_2,\alpha_1)+\nabla_\alpha F(\theta_2,\alpha_1)-\nabla_\alpha F(\theta_2,\alpha_2)\right\|_2\\ &\leq \underbrace{\left\|\nabla_\alpha F(\theta_1,\alpha_1)-\nabla_\alpha F(\theta_2,\alpha_1)\right\|_2}_{T_3}+\underbrace{\left\|\nabla_\alpha F(\theta_2,\alpha_1)-\nabla_\alpha F(\theta_2,\alpha_2)\right\|_2}_{T_4}. \end{aligned} \tag{11}$$

Next, we upper-bound T_3 and T_4 in eq. (11), respectively.

Upper-bounding T_3: For any given $1 \leq i \leq q$, we have

$$\begin{aligned}&\left|(\nabla_\alpha F(\theta_1,\alpha_1)-\nabla_\alpha F(\theta_2,\alpha_1))_i\right|\\ &= \left|(\nabla_\alpha V(\pi_E,r_{\alpha_1})-\nabla_\alpha V(\pi_{\theta_1},r_{\alpha_1})-\nabla_\alpha\psi(\alpha_1)-(\nabla_\alpha V(\pi_E,r_{\alpha_1})-\nabla_\alpha V(\pi_{\theta_2},r_{\alpha_1})-\nabla_\alpha\psi(\alpha_1)))_i\right|\\ &= \left|(\nabla_\alpha V(\pi_{\theta_2},r_{\alpha_1})-\nabla_\alpha V(\pi_{\theta_1},r_{\alpha_1}))_i\right|\\ &= \frac{1}{1-\gamma}\left|\sum_{s,a}(\nu_{\pi_{\theta_1}}(s,a)-\nu_{\pi_{\theta_2}}(s,a))(\nabla_\alpha r_{\alpha_1})_i\right| \leq \frac{\left\|\nu_{\pi_{\theta_1}}-\nu_{\pi_{\theta_2}}\right\|_1\left\|\frac{\partial r_\alpha}{\partial\alpha_i}\right\|_\infty}{1-\gamma}\\ &\overset{(i)}{\leq} \frac{2C_\nu\left\|\theta_1-\theta_2\right\|_2\left\|\frac{\partial r_\alpha}{\partial\alpha_i}\right\|_\infty}{1-\gamma},\end{aligned}$$

where (i) follows from Lemma 1 and the fact that $\left\|p-q\right\|_1 = 2\left\|p-q\right\|_{TV}$. The inequality above further implies that

$$\begin{aligned}\left\|\nabla_\alpha F(\theta_1,\alpha)-\nabla_\alpha F(\theta_2,\alpha)\right\|_2 &\leq \frac{2C_\nu\left\|\theta_1-\theta_2\right\|_2}{1-\gamma}\sqrt{\sum_{i=1}^q\left\|\frac{\partial r_\alpha}{\partial\alpha_i}\right\|_\infty^2}\\ &\leq \frac{C_r\sqrt{|\mathcal{A}|}}{1-\gamma}\left(1+\lceil\log_\rho C_M^{-1}\rceil+(1-\rho)^{-1}\right)\left\|\theta_1-\theta_2\right\|_2.\end{aligned}$$

Upper-bounding T_4: We provide a proof for the general parameterization of policy, which includes the direct parameterization of policy as a special case and covers the last claim of Proposition 1. We proceed as follows:

$$\begin{aligned}&\left\|\nabla_\alpha F(\theta_2,\alpha_1)-\nabla_\alpha F(\theta_2,\alpha_2)\right\|_2\\ &\leq \left\|\nabla_\alpha V(\pi_E,r_{\alpha_1})-\nabla_\alpha V(\pi_{\theta_2},r_{\alpha_1})-\nabla_\alpha\psi(\alpha_1)-(\nabla_\alpha V(\pi_E,r_{\alpha_2})-\nabla_\alpha V(\pi_{\theta_2},r_{\alpha_2})-\nabla_\alpha\psi(\alpha_2))\right\|_2\\ &\leq \frac{1}{1-\gamma}\left(\left\|\int(\nabla_\alpha r_{\alpha_1}-\nabla_\alpha r_{\alpha_2})d\nu_{\pi_E}\right\|_2+\left\|\int(\nabla_\alpha r_{\alpha_1}-\nabla_\alpha r_{\alpha_2})d\nu_{\pi_\theta}\right\|_2\right)+\left\|\nabla_\alpha\psi(\alpha_1)-\nabla_\alpha\psi(\alpha_2)\right\|_2\\ &= \frac{1}{1-\gamma}\left(\sqrt{\sum_{i=1}^q\left(\int(\nabla_\alpha r_{\alpha_1}(s,a)-\nabla_\alpha r_{\alpha_2}(s,a))_i d\nu_{\pi_E}\right)^2}+\sqrt{\sum_{i=1}^q\left(\int(\nabla_\alpha r_{\alpha_1}(s,a)-\nabla_\alpha r_{\alpha_2}(s,a))_i d\nu_{\pi_{\theta_2}}\right)^2}\right)\end{aligned}$$

$$
\begin{aligned}
&+ \|\nabla_\alpha \psi(\alpha_1) - \nabla_\alpha \psi(\alpha_2)\|_2 \\
&\overset{(i)}{\leq} \left(\frac{2\sqrt{q} L_r}{1-\gamma} + L_\psi \right) \|\alpha_1 - \alpha_2\|_2 ,
\end{aligned}
$$

where (i) follows from Assumption 1 and further because for any (s, a) and i, we have

$$
|(\nabla_\alpha r_{\alpha_1}(s,a) - \nabla_\alpha r_{\alpha_2}(s,a))_i| \leq \|\nabla_\alpha r_{\alpha_1}(s,a) - \nabla_\alpha r_{\alpha_2}(s,a)\|_2 \leq L_r \|\alpha_1 - \alpha_2\|_2 .
$$

Therefore, we obtain the following upper bound in eq. (11)

$$
\begin{aligned}
&\|\nabla_\alpha F(\theta_1, \alpha_1) - \nabla_\alpha F(\theta_2, \alpha_2)\|_2 \\
&\leq \frac{C_r \sqrt{|\mathcal{A}|}}{1-\gamma} \left(1 + \lceil \log_\rho C_M^{-1} \rceil + (1-\rho)^{-1}\right) \|\theta_1 - \theta_2\|_2 + \left(\frac{2\sqrt{q} L_r}{1-\gamma} + L_\psi \right) \|\alpha_1 - \alpha_2\|_2 ,
\end{aligned}
$$

which determines L_{21} and L_{22}. □

C Proof of Proposition 2

We define $\theta_{op}(\alpha) := \mathrm{argmin}_{\theta \in \Theta_p} F(\theta, \alpha)$. If there exist multiple optimal points, then $\theta_{op}(\alpha)$ can be any optimal point.

We first provide a lemma, which characterizes the gradient dominance property for the function $F(\theta, \alpha)$ with a fixed reward parameter α.

Lemma 3. *((Agarwal et al., 2019, Lemma 4.1)) For any given $\alpha \in \Lambda$, $F(\theta, \alpha)$ defined in eq.* (1) *with direct parameterization satisfies,*

$$
F(\theta, \alpha) - F(\theta_{op}(\alpha), \alpha) \leq C_d \max_{\tilde{\theta} \in \Theta_p} \left\langle \theta - \tilde{\theta}, \nabla_\theta F(\theta, \alpha) \right\rangle ,
$$

where $C_d = \frac{1}{(1-\gamma)\min_s\{\zeta(s)\}}$.

We then provide the proof of Proposition 2.

Proof of Proposition 2. We proceed as follows:

$$
\begin{aligned}
g(\theta) - g(\theta^*) &= F(\theta, \alpha_{op}(\theta)) - F(\theta^*, \alpha_{op}(\theta^*)) \\
&= F(\theta, \alpha_{op}(\theta)) - F(\theta_{op}(\alpha_{op}(\theta)), \alpha_{op}(\theta)) + F(\theta_{op}(\alpha_{op}(\theta)), \alpha_{op}(\theta)) - F(\theta^*, \alpha_{op}(\theta^*)) \\
&\overset{(i)}{\leq} F(\theta, \alpha_{op}(\theta)) - F(\theta_{op}(\alpha_{op}(\theta)), \alpha_{op}(\theta)) \\
&\overset{(ii)}{\leq} C_d \max_{\bar{\theta} \in \Theta_p} \left\langle \theta - \bar{\theta}, \nabla_\theta F(\theta, \alpha_{op}(\theta)) \right\rangle \\
&\overset{(iii)}{=} C_d \max_{\bar{\theta} \in \Theta_p} \left\langle \theta - \bar{\theta}, \nabla g(\theta) \right\rangle ,
\end{aligned}
$$

where (i) follows from the fact that

$$
\begin{aligned}
&F(\theta_{op}(\alpha_{op}(\theta)), \alpha_{op}(\theta)) - F(\theta^*, \alpha_{op}(\theta^*)) \\
&= \underbrace{F(\theta_{op}(\alpha_{op}(\theta)), \alpha_{op}(\theta)) - F(\theta^*, \alpha_{op}(\theta))}_{\leq 0} + \underbrace{F(\theta^*, \alpha_{op}(\theta)) - F(\theta^*, \alpha_{op}(\theta^*))}_{\leq 0} \leq 0,
\end{aligned}
$$

(ii) follows from Lemma 3, and (iii) follows because $\nabla g(\theta) = \nabla_\theta F(\theta, \alpha)|_{\alpha = \alpha_{op}(\theta)}$. □

D Supporting Lemmas for GAIL Framework

In this section, we establish two supporting lemmas that are useful for the proof of our main theorems.

Lemma 4. *Suppose Assumption 3 holds. Consider the gradient approximation in the nested-loop GAIL framework (Algorithm 1). For any k and t, $0 \le k \le K-1$ and $0 \le t \le T-1$, we have*

$$\mathbb{E}\left[\left\|\widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t) - \nabla_\alpha F(\theta_t, \alpha_k^t)\right\|_2^2\right] \le \frac{16C_r^2}{1-\gamma}\left(1+\frac{C_M}{1-\rho}\right)\frac{1}{B}.$$

Proof of Lemma 4. We denote $d_\pi(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\{s_t = s\}$ as the state visitation distribution of the Markov chain with initial distribution $\zeta(\cdot)$, transition kernel $\mathsf{P}(\cdot|s,a)$ and policy π. Both trajectories $(s_0^E, a_0^E, s_1^E, a_1^E, \cdots, s_i^E, a_i^E)$ and $(s_0^\theta, a_0^\theta, s_1^\theta, a_1^\theta, \cdots, s_i^E, a_i^E)$ are sampled under the transition kernel $\widetilde{\mathsf{P}}(\cdot|s,a) = \gamma\mathsf{P}(\cdot|s,a) + (1-\gamma)\zeta(\cdot)$. Recall that it has been shown in Konda (2002) that the stationary distribution of the Markov chain with transition kernel and policy π is d_π.

By definition, we have,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t) - \nabla_\alpha F(\theta_t, \alpha_k^t)\right\|_2^2\right] \\
&\quad = \mathbb{E}\left[\left\|\tfrac{1}{(1-\gamma)B}\left(\textstyle\sum_{i=0}^{B-1} \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^\theta, a_i^\theta)\right) - \tfrac{1}{1-\gamma}\left(\mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right] - \mathbb{E}_{(s,a)\sim\nu_{\pi_{\theta_t}}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right)\right\|_2^2\right] \\
&\quad \le \frac{2}{(1-\gamma)^2B^2}\underbrace{\mathbb{E}\left[\left\|\sum_{i=0}^{B-1}\left(\nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right)\right\|_2^2\right]}_{T_1} \\
&\qquad + \frac{2}{(1-\gamma)^2B^2}\underbrace{\mathbb{E}\left[\left\|\sum_{i=0}^{B-1}\left(\nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^\theta, a_i^\theta) - \mathbb{E}_{(s,a)\sim\nu_{\pi_{\theta_t}}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right)\right\|_2^2\right]}_{T_2}.
\end{aligned} \tag{12}$$

We first provide an upper bound on the term T_1 in eq. (12), and proceed as follows:

$$\begin{aligned}
T_1 &= \textstyle\sum_{i=0}^{B-1}\mathbb{E}\left\|\nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\|_2^2 \\
&\quad + \textstyle\sum_{i\neq j}\mathbb{E}\left\langle \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right], \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_j^E, a_j^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\rangle \\
&\le 4BC_r^2 + \textstyle\sum_{i\neq j}\mathbb{E}\left\langle \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right], \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_j^E, a_j^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\rangle
\end{aligned} \tag{13}$$

Define the filtration $\mathcal{F}_i = \sigma(s_0^E, a_0^E, s_1^E, a_1^E, \cdots, s_i^E, a_i^E)$. We continue to bound the second term in eq. (13) as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right], \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_j^E, a_j^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\rangle\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right], \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_j^E, a_j^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\rangle\middle|\mathcal{F}_i\right]\right] \\
&\quad = \mathbb{E}\left[\left\langle \nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right], \mathbb{E}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s_j^E, a_j^E)\middle|\mathcal{F}_i\right] - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\rangle\right] \\
&\quad \le \mathbb{E}\left[\left\|\nabla_{\alpha_k^t} r_{\alpha_k^t}(s_i^E, a_i^E) - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\|_2\left\|\mathbb{E}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s_j^E, a_j^E)\middle|\mathcal{F}_i\right] - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\|_2\right] \\
&\quad \le 2C_r\mathbb{E}\left[\left\|\mathbb{E}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s_j^E, a_j^E)\middle|\mathcal{F}_i\right] - \mathbb{E}_{(s,a)\sim\nu_{\pi_E}}\left[\nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)\right]\right\|_2\right] \\
&\quad = 2C_r\mathbb{E}\left\|\textstyle\int_{s\sim\mathbb{P}(s_j\in\cdot|s_i^E, a_i^E), a\sim\pi_E(\cdot|s)} \nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)dsda - \int_{s\sim\chi_\theta, a\sim\pi_E(\cdot|s)} \nabla_{\alpha_k^t} r_{\alpha_k^t}(s,a)dsda\right\|_2 \\
&\quad = 2C_r\mathbb{E}\sqrt{\textstyle\sum_{l=1}^q\left(\int_{s\sim\mathbb{P}(s_j\in\cdot|s_i^E, a_i^E), a\sim\pi_E(\cdot|s)} \frac{\partial r_\alpha}{\partial\alpha_l}|_{\alpha=\alpha_k^t}(s,a)dsda - \int_{s\sim\chi_\theta, a\sim\pi_E(\cdot|s)} \frac{\partial r_\alpha}{\partial\alpha_l}|_{\alpha=\alpha_k^t}(s,a)dsda\right)^2}
\end{aligned}$$

$$\overset{(i)}{\leq} 2C_r\mathbb{E}\sqrt{\sum_{l=1}^{q}\left(\left\|\frac{\partial r_\alpha}{\partial \alpha_i}\right\|_\infty \mathrm{d}_{TV}\left(\mathbb{P}(s_j\in\cdot|s_i=s_i^E,a_i=a_i^E),\chi_{\pi_E}\pi_E\right)\right)^2}, \tag{14}$$

where (i) follows from the fact that $|\int fd\mu-\int fd\nu|\leq\|f\|_\infty\mathrm{d}_{TV}(\mu,\nu)$. We next derive a bound on the total variation distance in the above equation as follows.

$$\begin{aligned}\mathrm{d}_{TV}\left(\mathbb{P}(s_j\in\cdot,a_j\in\cdot|s_i=s_i^E,a_i=a_i^E),\chi_{\pi_E}\pi_E\right)&=\mathrm{d}_{TV}\left(\mathbb{P}(s_j\in\cdot|s_i=s_i^E,a_i=a_i^E),\chi_{\pi_E}\right)\\&=\mathrm{d}_{TV}\left(\int_s\mathbb{P}(s_j\in\cdot|s_{i+1}=s)d\tilde{\mathsf{P}}(s|s_i=s_i^E,a_i=a_i^E),\chi_{\pi_E}\right)\\&\leq\int_s\mathrm{d}_{TV}\left(\mathbb{P}(s_j\in\cdot|s_{i+1}=s),\chi_{\pi_E}\right)d\tilde{\mathsf{P}}(s|s_i=s_i^E,a_i=a_i^E)\\&\overset{(i)}{\leq}\int_s C_M\rho^{j-i-1}d\tilde{\mathsf{P}}(s|s_i=s_i^E,a_i=a_i^E)=C_M\rho^{j-i-1},\end{aligned}\tag{15}$$

where (i) follows from Assumption 3. Substituting eq. (15) into eq. (14) and then further into eq. (13) yields the following upper-bound on T_1

$$T_1\leq 4BC_r^2+2\sum_{i=0}^{B-2}\sum_{j=i+1}^{B-1}2C_MC_r^2\rho^{j-i-1}\leq 4BC_r^2(1+\frac{C_M}{1-\rho}). \tag{16}$$

By following steps similar to those from eqs. (13) to (16), we can show that

$$T_2\leq 4BC_r^2(1+\frac{C_M}{1-\rho}).$$

Therefore, we have

$$\mathbb{E}\left[\left\|\widehat{\nabla}_\alpha F(\theta_t,\alpha_k^t)-\nabla_\alpha F(\theta_t,\alpha_k^t)\right\|_2^2\right]\leq\frac{16C_r^2}{(1-\gamma)^2}\left(1+\frac{C_M}{1-\rho}\right)\frac{1}{B}.$$

□

Lemma 5. *Suppose Assumptions 3 and 4 hold. Consider Algorithm 1 with α-update stepsize $\beta=\frac{\mu}{4L_{22}^2}$. For any $0\leq t\leq T-1$, we have*

$$\mathbb{E}\left[\left\|\alpha_K^t-\alpha_{op}(\theta_t)\right\|_2^2\right]\leq C_\alpha^2e^{-\frac{\mu^2}{8L_{22}^2}K}+\frac{48C_r^2}{\mu^2(1-\gamma)^2}(1+\frac{C_M}{1-\rho})\frac{1}{B}.$$

Let $K\geq\frac{8L_{22}^2}{\mu^2}\log\frac{2C_\alpha^2}{\Delta_\alpha}$ and $B\geq\frac{96C_r^2}{\mu^2(1-\gamma)^2}\left(1+\frac{C_M}{1-\rho}\right)\frac{1}{\Delta_\alpha}$, we have $\mathbb{E}\left[\left\|\alpha_K^t-\alpha_{op}(\theta_t)\right\|_2^2\right]\leq\Delta_\alpha$. The expected total computational complexity is given by

$$KB=\mathcal{O}\left(\frac{1}{(1-\gamma)^2\Delta_\alpha}\log\left(\frac{1}{\Delta_\alpha}\right)\right).$$

Proof of Lemma 5. We proceed as follows:

$$\begin{aligned}\left\|\alpha_{k+1}^t-\alpha_{op}(\theta_t)\right\|_2^2&\overset{(i)}{\leq}\left\|\alpha_k^t+\beta\widehat{\nabla}_\alpha F(\theta_t,\alpha_k^t)-\alpha_{op}(\theta_t)\right\|_2^2\\&=\left\|\alpha_k^t-\alpha_{op}(\theta_t)\right\|_2^2+\beta^2\left\|\widehat{\nabla}_\alpha F(\theta_t,\alpha_k^t)\right\|_2^2+2\beta\left\langle\widehat{\nabla}_\alpha F(\theta_t,\alpha_k^t),\alpha_k^t-\alpha_{op}(\theta_t)\right\rangle\\&\overset{(ii)}{\leq}\left\|\alpha_k^t-\alpha_{op}(\theta_t)\right\|_2^2+2\beta^2\left\|\nabla_\alpha F(\theta_t,\alpha_k^t)\right\|_2^2+2\beta^2\left\|\widehat{\nabla}_\alpha F(\theta_t,\alpha_k^t)-\nabla_\alpha F(\theta_t,\alpha_k^t)\right\|_2^2\\&\quad+2\beta\left\langle\nabla_\alpha F(\theta_t,\alpha_k^t),\alpha_k^t-\alpha_{op}(\theta_t)\right\rangle+2\beta\left\langle\widehat{\nabla}_\alpha F(\theta_t,\alpha_k^t)-\nabla_\alpha F(\theta_t,\alpha_k^t),\alpha_k^t-\alpha_{op}(\theta_t)\right\rangle\\&\overset{(iii)}{\leq}(1-2\beta\mu+2\beta^2L_{22}^2)\left\|\alpha_k^t-\alpha_{op}(\theta_t)\right\|_2^2+2\beta^2\left\|\widehat{\nabla}_\alpha F(\theta_t,\alpha_k^t)-\nabla_\alpha F(\theta_t,\alpha_k^t)\right\|_2^2\end{aligned}$$

$$
\begin{aligned}
&+ 2\beta \left\langle \widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t) - \nabla_\alpha F(\theta_t, \alpha_k^t), \alpha_k^t - \alpha_{op}(\theta_t) \right\rangle \\
&\overset{(iv)}{\leq} (1 + 2\beta^2 L_{22}^2 - \mu\beta) \left\| \alpha_k^t - \alpha_{op}(\theta_t) \right\|_2^2 + (2\beta^2 + \beta/\mu) \left\| \widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t) - \nabla_\alpha F(\theta_t, \alpha_k^t) \right\|_2^2 \\
&\overset{(v)}{\leq} \left(1 - \frac{\mu^2}{8L_{22}^2}\right) \left\| \alpha_k^t - \alpha_{op}(\theta_t) \right\|_2^2 + \frac{3}{8L_{22}^2} \left\| \widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t) - \nabla_\alpha F(\theta_t, \alpha_k^t) \right\|_2^2, \qquad (17)
\end{aligned}
$$

where (i) follows from the non-expansive property of the projection operator, (ii) follows because $\|A + B\|_2^2 \leq 2\|A\|_2^2 + 2\|B\|_2^2$, (iii) follows from Proposition 1 and the fact $\langle \nabla_\alpha F(\theta_t, \alpha_k^t), \alpha_k^t - \alpha_{op}(\theta_t) \rangle \leq -\mu \|\alpha_k^t - \alpha_{op}(\theta_t)\|_2^2$, (iv) follows because

$$
\langle \widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t) - \nabla_\alpha F(\theta_t, \alpha_k^t), \alpha_k^t - \alpha_{op}(\theta_t) \rangle \leq \frac{\mu}{2} \left\| \alpha_k^t - \alpha_{op}(\theta_t) \right\|_2^2 + \frac{1}{2\mu} \|\widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t) - \nabla_\alpha F(\theta_t, \alpha_k^t)\|_2^2,
$$

and (v) follows by letting $\beta = \frac{\mu}{4L_{22}^2}$ and because $\mu \leq L_{22}$.

Applying eq. (17) recursively and using the fact $1 - x \leq e^{-x}$, we obtain

$$
\left\| \alpha_K^t - \alpha_{op}(\theta_t) \right\|_2^2 \leq e^{-\frac{\mu^2}{8L_{22}^2} K} \left\| \alpha_0^t - \alpha_{op}(\theta_t) \right\|_2^2 + \frac{3}{8L_{22}^2} \sum_{k=0}^{K-1} \left(1 - \frac{\mu^2}{8L_{22}^2}\right)^{K-1-k} \left\| \widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t) - \nabla_\alpha F(\theta_t, \alpha_k^t) \right\|_2^2.
$$

Then, taking expectation on both sides of above inequality and applying Lemma 4 yield

$$
\begin{aligned}
\mathbb{E}\left[\left\| \alpha_K^t - \alpha_{op}(\theta_t) \right\|_2^2 \right] &\leq C_\alpha^2 e^{-\frac{\mu^2}{8L_{22}^2} K} + \frac{3}{8L_{22}^2} \sum_{k=0}^{K-1} \left(1 - \frac{\mu^2}{8L_{22}^2}\right)^{K-1-k} \frac{16C_r^2}{(1-\gamma)^2} (1 + \frac{C_M}{1-\rho}) \frac{1}{B} \\
&\leq C_\alpha^2 e^{-\frac{\mu^2}{8L_{22}^2} K} + \frac{48C_r^2}{\mu^2(1-\gamma)^2} (1 + \frac{C_M}{1-\rho}) \frac{1}{B},
\end{aligned}
$$

which completes the proof. □

E Proof of Theorems 1 and 2: Global Convergence of PPG-GAIL and FWPG-GAIL

In this section, we provide the proof of Theorems 1 and 2. We first provide three supporting lemmas. Specifically, Lemmas 6 and 7 establish the smoothness condition of the global optimal $\alpha_{op}(\theta)$ and the gradient $\nabla g(\theta)$. Similar property has also been established in Nouiehed et al. (2019); Lin et al. (2020). Lemma 8 provides the upper bound on the bias and variance errors introduced by the stochastic gradient estimator of $\nabla_\theta F(\theta_t, \alpha_t)$.

E.1 Supporting Lemmas

Lemma 6. *Suppose Assumptions 1 to 4 holds and the policy takes the direct parameterization specified in Section 2.2. We have* $\|\alpha_{op}(\theta_1) - \alpha_{op}(\theta_2)\|_2 \leq \frac{L_{21}}{\mu} \|\theta_1 - \theta_2\|_2$, *where* $\alpha_{op}(\theta)$ *is the unique global optimal that satisfies* $\alpha_{op}(\theta) = \operatorname{argmax}_{\alpha \in \Lambda} F(\theta, \alpha)$.

Proof of Lemma 6. Since $F(\theta_1, \alpha)$ is strongly concave on α, the following two inequalities hold for all $\alpha \in \Lambda$,

$$
F(\theta_1, \alpha_{op}(\theta_1)) - F(\theta_1, \alpha) \geq \frac{\mu}{2} \|\alpha - \alpha_{op}(\theta_1)\|_2^2, \qquad (18)
$$

$$
F(\theta_1, \alpha_{op}(\theta_1)) - F(\theta_1, \alpha) \leq \frac{\|\nabla_\alpha F(\theta_1, \alpha)\|_2^2}{2\mu}. \qquad (19)
$$

In eqs. (18) and (19), letting $\alpha = \alpha_{op}(\theta_2)$ and using the gradient Lipschitz condition established in Proposition 1, we have

$$
\frac{\mu}{2} \|\alpha_{op}(\theta_2) - \alpha_{op}(\theta_1)\|_2^2 \leq \frac{\|\nabla_\alpha F(\theta_1, \alpha_{op}(\theta_2))\|_2^2}{2\mu} \leq \frac{L_{21}^2 \|\theta_2 - \theta_2\|_2^2}{2\mu},
$$

which implies $\|\alpha_{op}(\theta_1) - \alpha_{op}(\theta_2)\|_2 \leq \frac{L_{21}}{\mu} \|\theta_1 - \theta_2\|_2$. □

Lemma 7. *Suppose Assumptions 1 to 4 hold and the policy takes the direct parameterization specified in Section 2.2. Then we have*

$$\nabla_\theta g(\theta) = \nabla_\theta F(\theta, \alpha)|_{\alpha=\alpha_{op}(\theta)},$$

and for any $\theta_1, \theta_2 \in \Theta_p$,

$$\|\nabla_\theta g(\theta_1) - \nabla_\theta g(\theta_2)\|_2 \leq (L_{11} + (L_{12}L_{21})/\mu) \|\theta_1 - \theta_2\|_2 ,$$

where L_{11}, L_{12} and L_{21} are defined in Proposition 1.

Proof of Lemma 7. Taking the directional derivative of $g(\theta)$ with respect to the direction ℓ, we have

$$\begin{aligned}\frac{\partial g(\theta)}{\partial \ell} &= \lim_{\epsilon\to 0} \frac{g(\theta+\epsilon\ell) - g(\theta)}{\epsilon} = \lim_{\epsilon\to 0} \frac{F(\theta+\epsilon\ell, \alpha_{op}(\theta+\epsilon\ell)) - F(\theta, \alpha_{op}(\theta))}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{F(\theta+\epsilon\ell, \alpha_{op}(\theta+\epsilon\ell)) - F(\theta+\epsilon\ell, \alpha_{op}(\theta)) + F(\theta+\epsilon\ell, \alpha_{op}(\theta)) - F(\theta, \alpha_{op}(\theta))}{\epsilon} \\ &\overset{(i)}{=} \lim_{\epsilon\to 0} \ell^\top \nabla_\alpha F(\theta, \alpha'_\epsilon) + \ell^\top \nabla_\theta F(\theta, \alpha_{op}(\theta)) \\ &\overset{(ii)}{=} \ell^\top \nabla_\theta F(\theta, \alpha_{op}(\theta)),\end{aligned} \tag{20}$$

where α'_ϵ in (i) is a point between $\alpha_{op}(\theta+\epsilon\ell)$ and $\alpha_{op}(\theta)$, and (ii) follows from Lemma 6 and hence we have $\lim_{\epsilon\to 0} \nabla_\alpha F(\theta, \alpha'_\epsilon) = \nabla_\alpha F(\theta, \alpha_{op}(\theta)) = 0$. Since eq. (20) holds for all directions ℓ, we have $\nabla_\theta g(\theta) = \nabla_\theta F(\theta, \alpha_{op}(\theta))$.

We then proceed to prove the gradient Lipschitz condition of $g(\theta_t)$. For any given $\theta_1, \theta_2 \in \Theta_p$, we have

$$\begin{aligned}\|\nabla_\theta g(\theta_1) - \nabla_\theta g(\theta_2)\|_2 &= \|\nabla_\theta F(\theta_1, \alpha_{op}(\theta_1)) - \nabla_\theta F(\theta_2, \alpha_{op}(\theta_2))\|_2 \\ &= \|\nabla_\theta F(\theta_1, \alpha_{op}(\theta_1)) - \nabla_\theta F(\theta_1, \alpha_{op}(\theta_2)) + \nabla_\theta F(\theta_1, \alpha_{op}(\theta_2)) - \nabla_\theta F(\theta_2, \alpha_{op}(\theta_2))\|_2 \\ &\leq \|\nabla_\theta F(\theta_1, \alpha_{op}(\theta_1)) - \nabla_\theta F(\theta_1, \alpha_{op}(\theta_2))\|_2 + \|\nabla_\theta F(\theta_1, \alpha_{op}(\theta_2)) - \nabla_\theta F(\theta_2, \alpha_{op}(\theta_2))\|_2 \\ &\leq L_{12} \|\alpha_{op}(\theta_1) - \alpha_{op}(\theta_2)\|_2 + L_{11} \|\theta_1 - \theta_2\|_2 \\ &\overset{(i)}{\leq} (L_{11} + \frac{L_{12}L_{21}}{\mu}) \|\theta_1 - \theta_2\|_2 ,\end{aligned}$$

where (i) follows from Lemma 6. □

Lemma 8. *Suppose Assumption 3 holds. For the policy gradient estimation specified in eq.* (3)*, in each iteration t, $0 \leq t \leq T-1$, we have*

$$\mathbb{E}\left[\left\|\widehat{\nabla}_\theta F(\theta_t, \alpha_t) - \nabla_\theta F(\theta_t, \alpha_t)\right\|_2^2\right] \leq \frac{4|\mathcal{A}|R_{max}^2}{(1-\gamma^{1/2})^2(1-\gamma)^2}\left(1 + \frac{2C_M\rho}{1-\rho}\right)\frac{1}{b}.$$

Let the sample trajectory size $b \geq \frac{4|\mathcal{A}|R_{max}^2}{(1-\gamma^{1/2})^2(1-\gamma)^2}\left(1 + \frac{2C_M\rho}{1-\rho}\right)\frac{1}{\Delta_\theta}$, we have $\mathbb{E}\left[\left\|\widehat{\nabla}_\theta F(\theta_t, \alpha_t) - \nabla_\theta F(\theta_t, \alpha_t)\right\|_2^2\right] \leq \Delta_\theta$.

Proof of Lemma 8. We define the vector $g_i \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ with each entry given by $(g_i)_{s,a} = -\frac{\hat{Q}(s,a)}{1-\gamma}\mathbb{1}\{s_i = s\}$. Then, we proceed as follows:

$$\begin{aligned}&\mathbb{E}\left[\left\|\widehat{\nabla}_\theta F(\theta_t, \alpha_t) - \nabla_\theta F(\theta_t, \alpha_t)\right\|_2^2\right] = \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=0}^{b-1}(g_i - \nabla_\theta F(\theta_t, \alpha_t))\right\|_2^2\right] \\ &= \frac{1}{b^2}\mathbb{E}\left[\sum_{i=0}^{b-1}\mathbb{E}\|g_i - \nabla_\theta F(\theta_t, \alpha_t)\|_2^2 + \sum_{i\neq j}\mathbb{E}\langle g_i - \nabla_\theta F(\theta_t, \alpha_t), g_j - \nabla_\theta F(\theta_t, \alpha_t)\rangle\right] \\ &\overset{(i)}{\leq} \frac{4|\mathcal{A}|R_{max}^2}{b(1-\gamma^{1/2})^2(1-\gamma)^2} + \frac{2}{b^2}\sum_{i=1}^{b-2}\sum_{j=i+1}^{b-1}\underbrace{\mathbb{E}\left[\langle g_i - \nabla_\theta F(\theta_t, \alpha_t), g_j - \nabla_\theta F(\theta_t, \alpha_t)\rangle\right]}_{T_1},\end{aligned} \tag{21}$$

where (i) follows from the facts that $\|g_i\|_2 = \left|\frac{\sqrt{|\mathcal{A}|}\hat{Q}(s_i,a_i)}{1-\gamma}\right| \leq \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma^{1/2})(1-\gamma)}$ and $\|\nabla_\theta F(\theta_t,\alpha_t)\|_2 \leq \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \leq \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma^{1/2})(1-\gamma)}$.

Define the filtration $\mathcal{F}_i = \sigma(s_0, s_1, \cdots, s_i)$. For the term T_1 in eq. (21) with $i < j$, we have

$$
\begin{aligned}
\mathbb{E}\left[\langle g_i - \nabla_\theta F(\theta_t,\alpha_t), g_j - \nabla_\theta F(\theta_t,\alpha_t)\rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle g_i - \nabla_\theta F(\theta_t,\alpha_t), g_j - \nabla_\theta F(\theta_t,\alpha_t)\rangle|\mathcal{F}_i\right]\right] \\
&= \mathbb{E}\left[\langle g_i - \nabla_\theta F(\theta_t,\alpha_t), \mathbb{E}\left[g_j - \nabla_\theta F(\theta_t,\alpha_t)|\mathcal{F}_i\right]\rangle\right] \\
&\leq \mathbb{E}\left[\|g_i - \nabla_\theta F(\theta_t,\alpha_t)\|_2 \left\|\mathbb{E}\left[g_j - \nabla_\theta F(\theta_t,\alpha_t)|\mathcal{F}_i\right]\right\|_2\right] \\
&\leq \frac{2R_{max}\sqrt{|\mathcal{A}|}}{(1-\gamma)(1-\gamma^{1/2})}\mathbb{E}\left\|\mathbb{E}\left[g_j|\mathcal{F}_i\right] - \nabla_\theta F(\theta_t,\alpha_t)\right\|_2 \\
&\leq \frac{2R_{max}\sqrt{|\mathcal{A}|}}{(1-\gamma)(1-\gamma^{1/2})}\mathbb{E}\left\|\sqrt{\sum_{s,a}\left(\mathbb{P}\left\{s_j = s|s_i\right\}\frac{Q(s,a)}{1-\gamma} - d_{\pi_{\theta_t}}(s)\frac{Q(s,a)}{1-\gamma}\right)^2}\right\|_2 \\
&\leq \frac{2R_{max}^2\sqrt{|\mathcal{A}|}}{(1-\gamma)^3(1-\gamma^{1/2})}\sqrt{\sum_{s,a}\left(\mathbb{P}\left\{s_j = s|s_i\right\} - d_{\pi_{\theta_t}}(s)\right)^2} \\
&\overset{(i)}{=} \frac{2R_{max}^2|\mathcal{A}|}{(1-\gamma)^3(1-\gamma^{1/2})}\left\|\mathbb{P}\left\{s_j = \cdot|s_i\right\} - \chi_{\pi_{\theta_t}}\right\|_2 \\
&\overset{(ii)}{\leq} \frac{4C_M R_{max}^2|\mathcal{A}|}{(1-\gamma)^3(1-\gamma^{1/2})}\rho^{j-i},
\end{aligned} \tag{22}
$$

where (i) follows because $\chi_{\pi_{\theta_t}} = d_{\pi_{\theta_t}}$, and (ii) follows from Assumption 3 and because $d_{\pi_{\theta_t}} = \chi_{\theta_t}$ and

$$
\left\|\mathbb{P}\left\{s_j = \cdot|s_i\right\} - d_{\pi_{\theta_t}}\right\|_2 \leq \left\|\mathbb{P}\left\{s_j = \cdot|s_i\right\} - d_{\pi_{\theta_t}}\right\|_1 = 2\mathrm{d}_{TV}\left(\mathbb{P}\left\{s_j = \cdot|s_i\right\}, d_{\pi_{\theta_t}}\right).
$$

Substituting eq. (22) into eq. (21), we obtain

$$
\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\right\|_2^2\right] &\leq \frac{4|\mathcal{A}|R_{max}^2}{b(1-\gamma^{1/2})^2(1-\gamma)^2} + \frac{2}{b^2}\sum_{i=1}^{b-2}\sum_{j=i+1}^{b-1}\frac{4C_M|\mathcal{A}|R_{max}^2}{(1-\gamma^{1/2})^2(1-\gamma)^2}\rho^{j-i} \\
&\leq \frac{4|\mathcal{A}|R_{max}^2}{b(1-\gamma^{1/2})^2(1-\gamma)^2}\left(1 + \frac{2C_M\rho}{1-\rho}\right)\frac{1}{b}.
\end{aligned}
$$

The second claim can be easily checked. □

E.2 Proof of Theorem 1

Based on the projection property, we have

$$
\left\langle \theta_t - \eta\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \theta_{t+1}, \theta - \theta_{t+1}\right\rangle \leq 0, \quad \forall\theta \in \Theta. \tag{23}
$$

Next we use eq. (23) to upper bound on $\mathbb{E}\left[\|\theta_{t+1} - \theta_t\|_2^2\right]$. Letting $\theta = \theta_t$ and rearranging eq. (23) yield

$$
\left\langle \widehat{\nabla}_\theta F(\theta_t,\alpha_t), \theta_{t+1} - \theta_t\right\rangle \leq -\eta^{-1}\|\theta_{t+1} - \theta_t\|_2^2. \tag{24}
$$

According to the gradient Lipschitz condition established in Lemma 7, we have

$$
\begin{aligned}
g(\theta_{t+1}) &\leq g(\theta_t) + \langle\nabla_\theta g(\theta_t), \theta_{t+1} - \theta_t\rangle + \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)\|\theta_{t+1} - \theta_t\|_2^2 \\
&= g(\theta_t) + \left\langle\widehat{\nabla}_\theta F(\theta_t,\alpha_t), \theta_{t+1} - \theta_t\right\rangle - \langle\nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t), \theta_{t+1} - \theta_t\rangle
\end{aligned}
$$

$$
\begin{aligned}
&- \left\langle \widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t), \theta_{t+1}-\theta_t \right\rangle + \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)\|\theta_{t+1}-\theta_t\|_2^2 \\
&\overset{(i)}{\leq} g(\theta_t) - \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)\|\theta_{t+1}-\theta_t\|_2^2 - \langle \nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t), \theta_{t+1}-\theta_t\rangle \\
&- \left\langle \widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t), \theta_{t+1}-\theta_t \right\rangle,
\end{aligned}
$$

where (i) follows from eq. (24) and the fact that $\eta = \left(L_{11} + \frac{L_{12}L_{21}}{\mu}\right)^{-1}$.

Rearranging the above inequality, we obtain

$$
\begin{aligned}
\|\theta_{t+1}-\theta_t\|_2^2 \leq& \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)^{-1}(g(\theta_t) - g(\theta_{t+1})) \\
&- \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)^{-1}\langle \nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t), \theta_{t+1}-\theta_t\rangle \\
&- \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)^{-1}\left\langle \widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t), \theta_{t+1}-\theta_t \right\rangle \\
\overset{(i)}{\leq}& \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)^{-1}(g(\theta_t) - g(\theta_{t+1})) \\
&+ \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)^{-2}\|\nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t)\|_2^2 + \frac{1}{4}\|\theta_{t+1}-\theta_t\|_2^2 \\
&+ \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right)^{-2}\|\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\|_2^2 + \frac{1}{4}\|\theta_{t+1}-\theta_t\|_2^2,
\end{aligned}
$$

where (i) follows from Young's inequality.

Taking expectation on both sides of the above inequality yields

$$
\begin{aligned}
\mathbb{E}\left[\|\theta_{t+1}-\theta_t\|_2^2\right] \overset{(i)}{\leq}& \frac{4\mu}{\mu L_{11} + L_{12}L_{21}}\mathbb{E}\left[g(\theta_t) - g(\theta_{t+1})\right] + \frac{8\mu^2 L_{22}^2}{(\mu L_{11} + L_{12}L_{21})^2}\mathbb{E}\left[\|\alpha_t - \alpha_{op}(\theta_t)\|_2^2\right] \\
&+ \frac{8\mu^2}{(\mu L_{11} + L_{12}L_{21})^2}\mathbb{E}\left[\left\|\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\right\|_2^2\right],
\end{aligned}
\tag{25}
$$

where (i) follows from the gradient Lipschitz condition established in Proposition 1

Next, rearranging eq. (23), we obtain

$$
\begin{aligned}
\langle \theta_t - \theta_{t+1}, \theta - \theta_{t+1}\rangle \leq& \eta\left\langle \widehat{\nabla}_\theta F(\theta_t,\alpha_t), \theta - \theta_{t+1}\right\rangle \\
=& \eta\left\langle \widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t), \theta - \theta_{t+1}\right\rangle + \eta\langle \nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t), \theta - \theta_{t+1}\rangle \\
&+ \eta\langle \nabla_\theta g(\theta_t,\alpha_t), \theta - \theta_t\rangle + \eta\langle \nabla_\theta g(\theta_t,\alpha_t), \theta_t - \theta_{t+1}\rangle.
\end{aligned}
$$

Letting $\eta = \left(L_{11} + \frac{L_{12}L_{21}}{\mu}\right)^{-1}$ and rearranging the above inequality yield

$$
\begin{aligned}
\langle \nabla_\theta g(\theta_t), \theta - \theta_t\rangle \geq& \left(L_{11} + \frac{L_{12}L_{21}}{\mu}\right)\langle \theta_t - \theta_{t+1}, \theta - \theta_{t+1}\rangle - \langle \nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t), \theta - \theta_{t+1}\rangle \\
&- \left\langle \widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t), \theta - \theta_{t+1}\right\rangle - \langle \nabla_\theta g(\theta_t), \theta_t - \theta_{t+1}\rangle \\
\overset{(i)}{\geq}& -\left(L_{11} + \frac{L_{12}L_{21}}{\mu}\right)\|\theta_t - \theta_{t+1}\|_2 \cdot 2R - \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2}\|\theta_{t+1}-\theta_t\|_2 \\
&- 2R(\|\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\|_2 + \|\nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t)\|_2),
\end{aligned}
\tag{26}
$$

where (i) follows from the Cauchy-Schwartz inequality and the boundness properties of Θ_p ($R := \max_{\theta\in\Theta_p}\{\|\theta\|_2\}$) and because $\|\nabla_\theta g(\theta_t)\|_2 = \|\nabla_\theta F(\theta_t, \alpha_{op}(\theta_t))\|_2 \leq \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2}$.

Applying the gradient dominance property of $g(\theta)$ established in Proposition 2, we obtain

$$\begin{aligned}g(\theta_t) - g(\theta^*) &\le C_d \max_{\theta\in\Theta} \langle \nabla_\theta g(\theta_t), \theta_t - \theta\rangle \\ &\overset{(i)}{\le} C_d \left(\frac{2(\mu L_{11} + L_{12}L_{21})R}{\mu} + \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \right) \|\theta_t - \theta_{t+1}\|_2 \\ &\quad + 2RC_d \|\widehat{\nabla}_\theta F(\theta_t, \alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\|_2 + 2RC_d \|\nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t)\|_2,\end{aligned}$$

where (i) follows by multiplying -1 on both sides of eq. (26) and taking the maximum over all $\theta \in \Theta_p$.

Taking expectation on both sides of above inequality and telescoping, we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \\ &\le C_d \left(\frac{2(\mu L_{11} + L_{12}L_{21})R}{\mu} + \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \right) \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\theta_t - \theta_{t+1}\|_2\right] \\ &\quad + 2RC_d \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\|_2\right] + 2RC_d \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t)\|_2\right] \\ &\overset{(i)}{\le} C_d \left(\frac{2(\mu L_{11} + L_{12}L_{21})R}{\mu} + \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \right) \sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \|\theta_t - \theta_{t+1}\|_2^2\right]} \\ &\quad + 2RC_d \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\|_2\right] + 2RC_d \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t)\|_2\right] \\ &\overset{(ii)}{\le} \left(\frac{2(\mu L_{11} + L_{12}L_{21})R}{\mu} + \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \right) C_d \sqrt{\frac{4\mu}{\mu L_{11} + L_{12}L_{21}} \frac{\mathbb{E}\left[g(\theta_0) - g(\theta_T)\right]}{T}} \\ &\quad + \left(\frac{2(\mu L_{11} + L_{12}L_{21})R}{\mu} + \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \right) C_d \sqrt{\frac{8\mu^2 L_{22}^2}{(\mu L_{11} + L_{12}L_{21})^2} \mathbb{E}\left[\|\alpha_t - \alpha_{op}(\theta_t)\|_2^2\right]} \\ &\quad + \left(\frac{2(\mu L_{11} + L_{12}L_{21})R}{\mu} + \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \right) C_d \sqrt{\frac{8\mu^2}{(\mu L_{11} + L_{12}L_{21})^2} \mathbb{E}\left[\left\|\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\right\|_2^2\right]} \\ &\quad + 2RC_d \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\widehat{\nabla}_\theta F(\theta_t,\alpha_t) - \nabla_\theta F(\theta_t,\alpha_t)\|_2\right] + 2RC_d \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla_\theta F(\theta_t,\alpha_t) - \nabla_\theta g(\theta_t)\|_2\right] \\ &\overset{(iii)}{\le} \left(\frac{2(\mu L_{11} + L_{12}L_{21})R}{\mu} + \frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \right) C_d \sqrt{\frac{4\mu}{\mu L_{11} + L_{12}L_{21}} \frac{R_{max}}{(1-\gamma)T}} \\ &\quad + \left(\frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \frac{2\mu}{\mu L_{11} + L_{12}L_{21}} + 5R \right) 2L_{22}C_d \sqrt{C_\alpha^2 e^{-\frac{\mu^2}{8L_{22}^2}K} + \frac{48C_r^2}{\mu^2(1-\gamma)^2}(1 + \frac{C_M}{1-\rho})\frac{1}{B}} \\ &\quad + \left(\frac{\sqrt{|\mathcal{A}|}R_{max}}{(1-\gamma)^2} \frac{2\mu}{\mu L_{11} + L_{12}L_{21}} + 5R \right) 2C_d \sqrt{\frac{4|\mathcal{A}|R_{max}^2}{b(1-\gamma^{1/2})^2(1-\gamma)^2}} \left(1 + \frac{2C_M\rho}{1-\rho}\right)\frac{1}{b} \\ &\overset{(iv)}{\le} \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{T}}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2 K}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{B}}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{b}}\right),\end{aligned}$$

where (i) follows because $\mathbb{E}[X] \le \sqrt{\mathbb{E}[X^2]}$ holds for any random variable X, (ii) follows by telescoping eq. (25) and further because $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ holds, for all $a, b > 0$, (iii) follows from Lemmas 5 and 8 and because $\mathbb{E}[X] \le \sqrt{\mathbb{E}[X^2]}$ holds for any random variable X, and (iv) follows because $L_{11} = \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$, $L_{12} = \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$, $L_{21} = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$, $L_{22} = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$, $C_d = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$ and $\mathcal{O}\left(\frac{1}{1-\gamma^{1/2}}\right) \le \mathcal{O}\left(\frac{1}{1-\gamma}\right)$.

E.3 Proof of Theorem 2

By the gradient Lipschitz condition (established in Lemma 7) of $g(\theta)$, we have

$$\begin{aligned}
g(\theta_{t+1}) &\leq g(\theta_t) + \langle \nabla_\theta g(\theta_t), \theta_{t+1} - \theta_t \rangle + \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right) \|\theta_{t+1} - \theta_t\|_2^2 \\
&= g(\theta_t) + \eta \langle \nabla_\theta g(\theta_t), \hat{v}_t - \theta_t \rangle + \left(\frac{L_{11}}{2} + \frac{L_{12}L_{21}}{2\mu}\right) \eta^2 \|\hat{v}_t - \theta_t\|_2^2 \\
&\overset{(i)}{\leq} g(\theta_t) + \eta \left\langle \widehat{\nabla}_\theta F(\theta_t, \alpha_t), \hat{v}_t - \theta_t \right\rangle + \eta \left\langle \nabla_\theta g(\theta_t) - \widehat{\nabla}_\theta F(\theta_t, \alpha_t), \hat{v}_t - \theta_t \right\rangle \\
&\quad + \left(2L_{11} + \frac{2L_{12}L_{21}}{\mu}\right) \eta^2 R^2 \\
&\overset{(ii)}{\leq} g(\theta_t) + \eta \left\langle \widehat{\nabla}_\theta F(\theta_t, \alpha_t), v_t - \theta_t \right\rangle + \eta \left\langle \nabla_\theta g(\theta_t) - \widehat{\nabla}_\theta F(\theta_t, \alpha_t), \hat{v}_t - \theta_t \right\rangle \\
&\quad + \left(2L_{11} + \frac{2L_{12}L_{21}}{\mu}\right) \eta^2 R^2 \\
&= g(\theta_t) + \eta \langle \nabla_\theta g(\theta_t), v_t - \theta_t \rangle + \eta \left\langle \nabla_\theta g(\theta_t) - \widehat{\nabla}_\theta F(\theta_t, \alpha_t), \hat{v}_t - v_t \right\rangle \\
&\quad + \left(2L_{11} + \frac{2L_{12}L_{21}}{\mu}\right) \eta^2 R^2, \qquad (27)
\end{aligned}$$

where (i) follows because $\|\hat{v}_t - \theta_t\|_2 \leq 2R$, and (ii) follows by definition of $\hat{v}_t$ in eq. (5) ($\hat{v}_t := \text{argmax}_{\theta \in \Theta_p} \langle \theta, -\widehat{\nabla}_\theta F(\theta_t, \alpha_t) \rangle$), and further we define $v_t := \text{argmax}_{\theta \in \Theta} \langle \theta, -\nabla_\theta g(\theta_t) \rangle$. We continue the proof as follows:

$$\begin{aligned}
\max_{\theta \in \Theta} \langle \nabla_\theta g(\theta_t), \theta_t - \theta \rangle &\overset{(i)}{=} \langle \nabla_\theta g(\theta_t), \theta_t - v_t \rangle \\
&\overset{(ii)}{\leq} \eta^{-1} \left(g(\theta_t) - g(\theta_{t+1})\right) + \left(2L_{11} + \frac{2L_{12}L_{21}}{\mu}\right) \eta R^2 \\
&\quad + \langle \nabla_\theta g(\theta_t) - \nabla_\theta F(\theta_t, \alpha_t), \hat{v}_t - v_t \rangle + \left\langle \nabla_\theta F(\theta_t, \alpha_t) - \widehat{\nabla}_\theta F(\theta_t, \alpha_t), \hat{v}_t - v_t \right\rangle \\
&\leq \eta^{-1} \left(g(\theta_t) - g(\theta_{t+1})\right) + \left(2L_{11} + \frac{2L_{12}L_{21}}{\mu}\right) \eta R^2 \\
&\quad + 2R \|\nabla_\theta g(\theta_t) - \nabla_\theta F(\theta_t, \alpha_t)\|_2 + 2R \left\|\nabla_\theta F(\theta_t, \alpha_t) - \widehat{\nabla}_\theta F(\theta_t, \alpha_t)\right\|_2, \qquad (28)
\end{aligned}$$

where (i) follows by definition $v_t := \text{argmax}_{\theta \in \Theta} \langle \theta, -\nabla_\theta g(\theta_t) \rangle$, and (ii) follows by rearranging eq. (27).

Finally, we complete the proof as follows:

$$\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \\
&\overset{(i)}{\leq} C_d \cdot \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\max_{\theta \in \Theta} \langle \nabla_\theta g(\theta_t), \theta_t - \theta \rangle\right] \\
&\overset{(ii)}{\leq} \frac{C_d \mathbb{E}\left[g(\theta_0) - g(\theta_T)\right]}{\eta T} + C_d \left(2L_{11} + \frac{2L_{12}L_{21}}{\mu}\right) \eta R^2 + \frac{2RC_d}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla_\theta g(\theta_t) - \nabla_\theta F(\theta_t, \alpha_t)\|_2 \\
&\quad + \frac{2RC_d}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\|\nabla_\theta F(\theta_t, \alpha_t) - \widehat{\nabla}_\theta F(\theta_t, \alpha_t)\right\|_2 \\
&\overset{(iii)}{\leq} C_d \cdot \frac{R_{max} + 2(1-\gamma)^3 \left(L_{11} + L_{12}L_{21}\mu^{-1}\right) R^2}{(1-\gamma)^2 \sqrt{T}} + 2RC_d \sqrt{\frac{4|\mathcal{A}| R_{max}^2}{b(1-\gamma^{1/2})^2 (1-\gamma)^2} \left(1 + \frac{2C_M \rho}{1-\rho}\right) \frac{1}{b}} \\
&\quad + 2RC_d L_{22} \sqrt{C_\alpha^2 e^{-\frac{\mu^2}{8L_{22}^2} K} + \frac{48 C_r^2}{(1-\gamma)^2 \mu^2} (1 + \frac{C_M}{1-\rho}) \frac{1}{B}}
\end{aligned}$$

$$\overset{(iv)}{\leq} \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{T}}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2 K}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{B}}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{b}}\right),$$

where (i) follows from Proposition 2, (ii) follows from telescoping eq. (28), (iii) follows from Lemmas 5 and 8 and because $\eta = \frac{1-\gamma}{\sqrt{T}}$ and $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$ holds for any random variable X, and (iv) follows because $L_{11} = \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$, $L_{12} = \mathcal{O}\left(\frac{1}{(1-\gamma)^2}\right)$, $L_{21} = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$, $L_{22} = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$, $C_d = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$ and $\mathcal{O}\left(\frac{1}{1-\gamma^{1/2}}\right) \leq \mathcal{O}\left(\frac{1}{1-\gamma}\right)$.

F Proof of Theorems 3 and 4: Global Convergence of TRPO-GAIL

In this section, we add the subscript λ to the notations of the Q-function $Q_\alpha^\pi(s,a)$, the value function $V(\pi, r_\alpha)$, the objective function $F(\theta,\alpha)$ and $g(\theta)$ in order to emphasize that these functions are derived under λ-regularized MDP.

F.1 Supporting Lemmas

In this subsection, we introduce several useful lemmas.

Lemma 9. *((Beck, 2017, Lemma 9.1)) Consider a proper closed convex function $\omega\colon E \to (-\infty, \infty]$. Let dom$(\partial\omega)$ denote the subset of E where ω is differentiable and dom(ω) denote the subset of E where the value of ω is finite. Assume $a, b \in$ dom$(\partial\omega)$ and $c \in$ dom(ω). Then the following inequality holds:*

$$\langle \nabla\omega(b) - \nabla\omega(a), c - a\rangle = B_\omega(c,a) + B_\omega(a,b) - B_\omega(c,b),$$

where $B_\omega(\cdot,\cdot)$ denotes the Bregman distance associated with $\omega(\cdot)$.

Lemma 10. *((Shani et al., 2020, Lemma 25)) Consider the Q-function estimation in Algorithm 3. For any $t \in \{0, 1, \cdots, T-1\}$, we have*

$$\left\| -\hat{Q}_{\lambda,\alpha_t}^{\pi_{\theta_t}}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)) \right\|_\infty \leq C_\omega(t;\lambda),$$

where $\hat{Q}_{\lambda,\alpha_t}^{\pi_{\theta_t}}$ is the Q-function estimated under the reward function r_{α_t} and policy π_{θ_t}, and $C_\omega(t;\lambda) \leq \mathcal{O}\left(\frac{C_r C_\alpha (1+\mathbb{1}\{\lambda\neq 0\}\log t)}{1-\gamma^{1/2}}\right)$.

Lemma 11. *For any policy $\pi, \pi' \in \Delta_{\mathcal{A}}$ and $\alpha \in \Lambda$, the following equality holds,*

$$(V_\lambda(\pi, r_\alpha) - V_\lambda(\pi', r_\alpha))(1-\gamma) = \sum_{s\in\mathcal{S}} d_{\pi'}(s)\left(\left\langle -Q_{\lambda,\alpha}^\pi(s,\cdot) + \lambda\nabla\omega(\pi(\cdot|s)), \pi'(\cdot|s) - \pi(\cdot|s)\right\rangle + \lambda B_\omega(\pi'(\cdot|s), \pi(\cdot|s))\right),$$

where $V_\lambda(\pi, r_\alpha)$ is the average value function under λ-regularized MDP with the reward function r_α and $d_{\pi'}$ is the state visitation distribution of π'.

Proof of Lemma 11. Following from (Shani et al., 2020, Lemma 24), for any $s \in \mathcal{S}$, we have

$$\left\langle -Q_{\lambda,\alpha}^\pi(s,\cdot) + \lambda\nabla\omega(\pi(\cdot|s)), \pi'(\cdot|s) - \pi(\cdot|s)\right\rangle = -(T_\lambda^{\pi'} V_{\lambda,\alpha}^\pi(s) - V_{\lambda,\alpha}^\pi(s)) - \lambda B_\omega(\pi'(\cdot|s), \pi(\cdot|s)), \tag{29}$$

where $T_\lambda^{\pi'}$ is the Bellman operator under λ-regularized MDP, i.e.,

$$T_\lambda^{\pi'} V_{\lambda,\alpha}^\pi(s) = \sum_{a\in\mathcal{A}} \left(\pi'(a|s) r_{\alpha,\lambda}(s,a) + \sum_{s'\in\mathcal{S}} \mathsf{P}(s'|s,a) V_{\lambda,\alpha}^\pi(s')\right).$$

Furthermore, we have

$$\begin{aligned}
V_\lambda(\pi', r_\alpha) - V_\lambda(\pi, r_\alpha) &= \sum_s \zeta(s)(V_{\lambda,\alpha}^{\pi'}(s) - V_{\lambda,\alpha}^\pi(s)) \\
&\overset{(i)}{=} \frac{1}{(1-\gamma)} \sum_{s\in\mathcal{S}} d_{\pi'}(s)(T_\lambda^{\pi'} V_{\lambda,\alpha}^\pi(s) - V_{\lambda,\alpha}^\pi(s)) \\
&\overset{(ii)}{=} -\frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d_{\pi'}(s)\left(\left\langle -Q_{\lambda,\alpha}^\pi(s,\cdot) + \lambda\nabla\omega(\pi(\cdot|s)), \pi'(\cdot|s) - \pi(\cdot|s)\right\rangle + \lambda B_\omega(\pi'(\cdot|s), \pi(\cdot|s))\right),
\end{aligned}$$

where (i) follows from (Shani et al., 2020, Lemma 29) and (ii) follows by multiplying eq. (29) by $d_{\pi'}(s)$ and take the summation over $\mathcal{S}$. ☐

F.2 Proof of Theorems 3 and 4

Since the unregularized MDP can be viewed as a special case of the regularized MDP, i.e., $\lambda = 0$, in this subsection, we first develop our proof for the general regularized MDP up to a certain step, and then specialize to the case with $\lambda = 0$ for proving Theorem 3 and continue to keep λ general for proving Theorem 4.

To we start the proof, recall that the update of θ_t specified in eq. (7) satisfies,

$$\pi_{\theta_{t+1}}(\cdot|s) \in \underset{\pi\in\Delta_{\mathcal{A}}}{\operatorname{argmin}}(\underbrace{\left\langle -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)), \pi - \pi_{\theta_t}(\cdot|s)\right\rangle + \eta_t^{-1}B_\omega(\pi, \pi_{\theta_t}(\cdot|s))}_{:=f_0(\pi)}).$$

Following from the first-order optimality condition, we have

$$\nabla_\pi f_0(\pi_{\theta_{t+1}}(\cdot|s))^\top(\pi - \pi_{\theta_{t+1}}(\cdot|s)) \geq 0, \forall \pi \in \Delta_{\mathcal{A}},$$

which together with the fact

$$\nabla_\pi f_0(\pi) = -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)) + \eta_t^{-1}(\nabla\omega(\pi) - \nabla\omega(\pi_{\theta_t}(\cdot|s))),$$

implies that

$$\left\langle -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)) + \eta_t^{-1}(\nabla\omega(\pi_{\theta_{t+1}}(\cdot|s)) - \nabla\omega(\pi_{\theta_t}(\cdot|s))), \pi - \pi_{\theta_{t+1}}(\cdot|s)\right\rangle \geq 0 \tag{30}$$

holds for any π.

Taking $\pi = \pi_{\theta^*}(\cdot|s)$ in eq. (30), we obtain

$$\begin{aligned}
0 &\leq \eta_t \left\langle -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_t}(\cdot|s)\right\rangle \\
&\quad + \eta_t \left\langle -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)), \pi_{\theta_t}(\cdot|s) - \pi_{\theta_{t+1}}(\cdot|s)\right\rangle \\
&\quad + \left\langle \nabla\omega(\pi_{\theta_{t+1}}(\cdot|s)) - \nabla\omega(\pi_{\theta_t}(\cdot|s)), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_{t+1}}(\cdot|s)\right\rangle \\
&\overset{(i)}{\leq} \eta_t \left\langle -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_t}(\cdot|s)\right\rangle \\
&\quad + \frac{\eta_t^2 \left\| -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s))\right\|_\infty^2}{2} + \frac{\left\|\pi_{\theta_t}(\cdot|s) - \pi_{\theta_{t+1}}(\cdot|s)\right\|_1^2}{2} \\
&\quad + B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s)) - B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_{t+1}}(\cdot|s)) - B_\omega(\pi_{\theta_{t+1}}(\cdot|s), \pi_{\theta_t}(\cdot|s)) \\
&\overset{(ii)}{\leq} \eta_t \left\langle -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_t}(\cdot|s)\right\rangle + \frac{\eta_t^2 C_\omega(t;\lambda)^2}{2} \\
&\quad + B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s)) - B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_{t+1}}(\cdot|s)),
\end{aligned} \tag{31}$$

where (i) follows from Hölder's inequality and Lemma 9, and (ii) follows from the Lemma 10 and Pinsker's inequality given by

$$\frac{\left\|\pi_{\theta_t}(\cdot|s) - \pi_{\theta_{t+1}}(\cdot|s)\right\|_1^2}{2} \leq \mathrm{KL}\left(\pi_{\theta_{t+1}}(\cdot|s) \middle\| \pi_{\theta_t}(\cdot|s)\right) = B_\omega(\pi_{\theta_{t+1}}(\cdot|s), \pi_{\theta_t}(\cdot|s)),$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the KL-divergence.

Taking expectation conditioned on $\mathcal{F}_t = \sigma(\theta_0, \theta_1, \cdots, \theta_t)$ over eq. (31), we have

$$\begin{aligned}
0 &\leq \eta_t \left\langle -Q^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda\nabla\omega(\pi_{\theta_t}(\cdot|s)), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_t}(\cdot|s)\right\rangle + \frac{\eta_t^2 C_\omega(t;\lambda)^2}{2} \\
&\quad + B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s)) - \mathbb{E}\left[B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_{t+1}}(\cdot|s)) \middle| \mathcal{F}_t\right].
\end{aligned} \tag{32}$$

Since eq. (32) holds for any state, we multiply it by $d_{\pi_{\theta^*}}(s)$ for each state s and take the summation over $\mathcal{S}$. Then we rearrange the resulting bound and obtain

$$\frac{\eta_t^2 C_\omega(t;\lambda)^2}{2} + \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s) B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s)) - \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s) \mathbb{E}\left[B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_{t+1}}(\cdot|s)) \middle| \mathcal{F}_t\right]$$

$$
\begin{aligned}
&\geq -\eta_t \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s) \left\langle -Q^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot) + \lambda \nabla \omega(\pi_{\theta_t}(\cdot|s)), \pi_{\theta^*}(\cdot|s) - \pi_{\theta_t}(\cdot|s) \right\rangle \\
&\overset{(i)}{=} \eta_t(1-\gamma)(V_\lambda(\pi_{\theta^*}, r_{\alpha_t}) - V_\lambda(\pi_{\theta_t}, r_{\alpha_t})) + \eta_t \lambda \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s) B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s)),
\end{aligned} \tag{33}
$$

where (i) follows from applying Lemma 11 with $\pi = \pi_{\theta_t}$ and $\pi' = \pi_{\theta^*}$. Rearranging eq. (33), we obtain

$$
\begin{aligned}
V_\lambda(\pi_{\theta^*}, r_{\alpha_t}) - V_\lambda(\pi_{\theta_t}, r_{\alpha_t}) \leq& \frac{1}{\eta_t(1-\gamma)} \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s)(1-\lambda\eta_t)\mathbb{E}\left[B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s))\right] \\
&- \frac{1}{\eta_t(1-\gamma)} \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s)\mathbb{E}\left[B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_{t+1}}(\cdot|s))\right] + \frac{\eta_t C_\omega(t,\lambda)^2}{2(1-\gamma)}.
\end{aligned} \tag{34}
$$

Furthermore, we proceed the proof as follows:

$$
\begin{aligned}
\mathbb{E}\left[g_\lambda(\theta_t)\right] - g_\lambda(\theta^*) &= \mathbb{E}\left[g_\lambda(\theta_t) - F_\lambda(\theta_t,\alpha_t)\right] + \mathbb{E}\left[F_\lambda(\theta_t,\alpha_t) - g_\lambda(\theta^*)\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[g_\lambda(\theta_t) - F_\lambda(\theta_t,\alpha_t)\right] + \mathbb{E}\left[F_\lambda(\theta_t,\alpha_t) - F_\lambda(\theta^*,\alpha_t)\right] \\
&\overset{(ii)}{=} \mathbb{E}\left[g_\lambda(\theta_t) - F_\lambda(\theta_t,\alpha_t)\right] + \mathbb{E}\left[V_\lambda(\pi_{\theta^*},\alpha_t) - V_\lambda(\pi_{\theta_t},\alpha_t)\right] \\
&\overset{(iii)}{\leq} L_{22}^2 \mathbb{E}\left[\|\alpha_t - \alpha_{op}(\theta_t)\|_2^2\right] + \frac{1}{\eta_t(1-\gamma)} \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s)(1-\lambda\eta_t)\mathbb{E}\left[B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_t}(\cdot|s))\right] \\
&\quad - \frac{1}{\eta_t(1-\gamma)} \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s)\mathbb{E}\left[B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_{t+1}}(\cdot|s))\right] + \frac{\eta_t C_\omega(t,\lambda)^2}{2(1-\gamma)},
\end{aligned} \tag{35}
$$

where (i) follows because $g_\lambda(\theta^*) \geq F_\lambda(\theta^*, \alpha_{op}(\theta_t))$, (ii) follows from the definition of $F_\lambda(\theta,\alpha)$, and (iii) follows from the gradient Lipschitz condition of α in Proposition 1 and eq. (34).

Next, to prove Theorem 3, we let $\lambda = 0$ and recall $\eta_t = \frac{1-\gamma}{\sqrt{T}}$. Telescoping eq. (35), we obtain

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \leq& \frac{1}{(1-\gamma)^2\sqrt{T}} \sum_{s\in\mathcal{S}} d_{\pi_{\theta^*}}(s)\mathbb{E}\left[B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_0}(\cdot|s)) - B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_T}(\cdot|s))\right] \\
&+ \frac{L_{22}^2}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\alpha_t - \alpha_{op}(\theta_t)\|_2^2\right] + \frac{C_\omega^2}{2\sqrt{T}} \\
\overset{(i)}{\leq}& L_{22}^2 C_\alpha^2 e^{-\frac{\mu^2}{8L_{22}^2}K} + \frac{48 C_r^2 L_{22}^2}{\mu^2(1-\gamma)^2}(1+\frac{C_M}{1-\rho})\frac{1}{B} + \frac{(1-\gamma)^2 C_\omega^2 + 2\log|\mathcal{A}|}{2(1-\gamma)^2\sqrt{T}} \\
\overset{(ii)}{\leq}& \mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{T}}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2 K}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^4 B}\right),
\end{aligned}
$$

where (i) follows from Lemma 5 and because $0 \leq B_\omega(\pi_1,\pi_2) \leq \log|\mathcal{A}|$ for any θ_1,θ_2 and (ii) follows because $L_{22} = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$ and $C_\omega = \mathcal{O}\left(\frac{1}{1-\gamma^{1/2}}\right) \leq \mathcal{O}\left(\frac{1}{1-\gamma}\right)$. This completes the proof of Theorem 3.

To prove the Theorem 4, let $\eta_t = \frac{1}{\lambda(t+2)}$. Then, telescoping eq. (35) and applying Lemma 5, we obtain

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[g_\lambda(\theta_t)\right] - g_\lambda(\theta^*) \leq& L_{22}^2 C_\alpha^2 e^{-\frac{\mu^2}{8L_{22}^2}K} + \frac{48 C_r^2 L_{22}^2}{\mu^2(1-\gamma)^2}(1+\frac{C_M}{1-\rho})\frac{1}{B} + \frac{C_\omega^2(T,\lambda)}{2(1-\gamma)\lambda}\frac{\log(T+1)}{T} \\
&+ \frac{\lambda\sum_s d_{\pi_{\theta^*}(s)}\mathbb{E}[B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_0}(\cdot|s)) - (T+1)B_\omega(\pi_{\theta^*}(\cdot|s), \pi_{\theta_T}(\cdot|s))]}{(1-\gamma)T} \\
\overset{(i)}{\leq}& \mathcal{O}\left(\frac{1}{(1-\gamma)^3 T}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2 K}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^4 B}\right),
\end{aligned}
$$

where (i) follows because $0 \leq B_\omega(\pi_1,\pi_2) \leq \log(|\mathcal{A}|)$ for any π_1,π_2, $L_{22} = \mathcal{O}\left(\frac{1}{1-\gamma}\right)$ and $C_\omega(T,\lambda) = \tilde{\mathcal{O}}\left(\frac{1}{1-\gamma^{1/2}}\right) \leq \tilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\right)$. This completes the proof of Theorem 4.

G Proof of Theorem 5: Global Convergence of NPG-GAIL

To prove the theorem, we first define some notations. Let $\lambda_P := \min_{\theta\in\Theta}\{\lambda_{min}(F(\theta)+\lambda I)\}$,

$$W_{\theta,\alpha}^{\lambda *} := (F(\theta)+\lambda I)^{-1}\mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\left[A_\alpha^{\pi_\theta}(s,a)\nabla_\theta\log\pi_\theta(a|s)\right]$$

and

$$W_{\theta,\alpha}^{*} := F(\theta)^{\dagger}\mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\left[A_\alpha^{\pi_\theta}(s,a)\nabla_\theta\log\pi_\theta(a|s)\right].$$

For brevity, we denote $W_t^{\lambda *} = W_{\theta_t,\alpha_t}^{\lambda *}$ and $W_t^{*} = W_{\theta_t,\alpha_t}^{*}$.

G.1 Supporting Lemmas

In this subsection, we give several useful lemmas.

Lemma 12. *((Agarwal et al., 2019, Lemma 3.2)) For any policy π and π' and reward function r_α, we have*

$$V(\pi,r_\alpha)-V(\pi',r_\alpha) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim\nu_\pi(s,a)}\left[A_\alpha^{\pi'}(s,a)\right].$$

Lemma 13. *((Xu et al., 2020a, Lemma 6)) For any θ and α, we have $\left\|W_{\theta,\alpha}^{\lambda *}-W_{\theta,\alpha}^{*}\right\|_2 \le C_\lambda\lambda$, where $0<C_\lambda<\infty$ is a constant only depending on the policy class.*

Lemma 14. *Suppose Assumptions 3 and 5 hold. Consider the policy update of NPG-GAIL (Algorithm 2) with $\beta_W = \frac{\lambda_P}{4(C_\phi^2+\lambda)^2}$. Then, for all $t=0,1,\cdots,T-1$, we have*

$$\begin{aligned}\mathbb{E}[\left\|w_t - W_t^{\lambda *}\right\|_2^2] \le \exp\left\{-\frac{\lambda_P^2 T_c}{16(C_\phi^2+\lambda)^2}\right\}\frac{R_{max}^2 C_\phi^2}{\lambda_P^2(1-\gamma)^2}\\ +\left(\frac{1}{\lambda_P}+\frac{\lambda_P}{2(C_\phi^2+\lambda)^2}\right)\frac{98R_{max}^2C_\phi^2[(C_\phi^2+\lambda)^2+4\lambda_P^2][1+(C_M-1)\rho]}{(1-\rho)(1-\gamma)^2\lambda_P^3 M}.\end{aligned}$$

Proof of Lemma 14. At iteration t, $W_0, W_1,\cdots,W_{T_c}$ follows the linear SA iteration rule defined in (Xu et al., 2020a, eq. (3)) with $\alpha=\beta_W$, $A=-(F(\theta_t)+\lambda I)$, $b=\mathbb{E}_{(s,a)\sim\nu_{\pi_{\theta_t}}}\left[A_{\alpha_t}^{\pi_{\theta_t}}(s,a)\nabla_{\theta_t}\log\pi_{\theta_t}(a|s)\right]$ and $\theta^*=-A^{-1}b = W_t^{\lambda *}$ with $\left\|W_t^{\lambda *}\right\|_2 \le R_\theta = \frac{2C_\phi R_{max}}{\lambda_A(1-\gamma)}$. It is easy to check that the Assumption 3 in Xu et al. (2020a) holds. Namely, (i), $\|A\|_F \le C_\phi^2+\lambda$ and $\|b\|_2 \le \frac{2R_{max}C_\phi}{1-\gamma}$; (ii), for any $w\in\mathbb{R}^d$, $\left\langle w-W_t^{\lambda *}, A(w-W_t^{\lambda *})\right\rangle \le -\lambda_p\left\|w-W_t^{\lambda *}\right\|_2^2$; (iii), The ergodicity of MDP is assumed here. Thus, applying (Xu et al., 2020a, Theorem 4) completes the proof. □

G.2 Proof of Theorem 5

Define $D(\theta) = \mathbb{E}_{s\sim d_{\pi_{\theta^*}}}[\mathrm{KL}(\pi_{\theta^*}(\cdot|s)\|\pi_\theta(\cdot|s))]$. Then we have

$$\begin{aligned}D(\theta_t)-D(\theta_{t+1}) &= \mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\log(\pi_{\theta_{t+1}}(\cdot|s))-\log(\pi_{\theta_t}(\cdot|s))\right]\\ &\overset{(i)}{\ge} \mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta\log(\pi_{\theta_t}(a|s))\right]^\top(\theta_{t+1}-\theta_t) - \frac{L_\phi^2}{2}\left\|\theta_{t+1}-\theta_t\right\|_2^2,\end{aligned}$$

where (i) follows from the gradient Lipschitz condition on $\log(\pi_\theta(\cdot|s))$ in Assumption 5.

Recall that the update rule in NPG-GAIL (Algorithm 2) is given by $\theta_{t+1}=\theta_t-\eta w_t$. Then we have

$$\begin{aligned}D(\theta_t)-D(\theta_{t+1}) &\ge \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta\log(\pi_{\theta_t}(a|s))\right]^\top w_t - \frac{L_\phi^2\eta^2}{2}\left\|w_t\right\|_2^2\\ &= \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[A_{\alpha_t}^{\pi_{\theta_t}}(s,a)\right] + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta\log(\pi_{\theta_t}(a|s))^\top W_t^* - A_{\alpha_t}^{\pi_{\theta_t}}(s,a)\right]\\ &\quad + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta\log(\pi_{\theta_t}(a|s))\right]^\top(W_t^{\lambda *}-W_t^*) + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta\log(\pi_{\theta_t}(a|s))\right]^\top(w_t-W_t^{\lambda *})\\ &\quad - \frac{L_\phi^2\eta^2}{2}\left\|w_t\right\|_2^2\end{aligned}$$

$$
\begin{aligned}
&\overset{(i)}{=} (1-\gamma)\eta\left(V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t})\right) + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))^\top W_t^* - A_{\alpha_t}^{\pi_{\theta_t}}(s,a)\right] \\
&\quad + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))\right]^\top (W_t^{\lambda*} - W_t^*) + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))\right]^\top (w_t - W_t^{\lambda*}) \\
&\quad - \frac{L_\phi^2\eta^2}{2}\|w_t\|_2^2 \\
&\overset{(ii)}{\geq} (1-\gamma)\eta\left(V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t})\right) - \frac{L_\phi^2\eta^2}{2}\|w_t\|_2^2 \\
&\quad + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))\right]^\top (W_t^{\lambda*} - W_t^*) + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))\right]^\top (w_t - W_t^{\lambda*}) \\
&\quad - \eta\sqrt{\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[(\nabla_\theta \log(\pi_{\theta_t}(a|s))^\top W_t^* - A_{\alpha_t}^{\pi_{\theta_t}}(s,a))^2\right]} \\
&\overset{(iii)}{\geq} (1-\gamma)\eta\left(V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t})\right) - \frac{L_\phi^2\eta^2}{2}\|w_t\|_2^2 \\
&\quad + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))\right]^\top (W_t^{\lambda*} - W_t^*) + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))\right]^\top (w_t - W_t^{\lambda*}) \\
&\quad - \eta\sqrt{C_d\mathbb{E}_{\nu_{\pi_{\theta_t}}}\left[(\nabla_\theta \log(\pi_{\theta_t}(a|s))^\top W_t^* - A_{\alpha_t}^{\pi_{\theta_t}}(s,a))^2\right]},
\end{aligned} \tag{36}
$$

where (i) follows from Lemma 12, (ii) follows from the concavity of $f(x) = \sqrt{x}$ and Jensen's inequality, and (iii) follows from the fact that $(\nabla_\theta \log(\pi_{\theta_t}(a|s))^\top W_t^* - A_{\alpha_t}^{\pi_{\theta_t}}(s,a))^2 \geq 0$ and $\left\|\frac{\nu_{\pi_{\theta^*}}}{\nu_{\pi_{\theta_t}}}\right\|_\infty \leq \frac{1}{(1-\gamma)\min\{\zeta(s)\}} := C_d$.

Continuing to bound eq. (36), we have

$$
\begin{aligned}
D(\theta_t) - D(\theta_{t+1}) &\overset{(i)}{\geq} (1-\gamma)\eta\left(V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t})\right) - \frac{L_\phi^2\eta^2}{2}\|w_t\|_2^2 - \eta\sqrt{C_d}\zeta' \\
&\quad + \eta\mathbb{E}_{\nu_{\pi_{\theta^*}}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))\right]^\top (W_t^{\lambda*} - W_t^*) + \eta\mathbb{E}_{\nu_{\pi_E}}\left[\nabla_\theta \log(\pi_{\theta_t}(a|s))\right]^\top (w_t - W_t^{\lambda*}) \\
&\overset{(ii)}{\geq} (1-\gamma)\eta\left(V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t})\right) - \eta\sqrt{C_d}\zeta' - \eta C_\phi C_\lambda \lambda \\
&\quad - \eta C_\phi \left\|w_t - W_t^{\lambda*}\right\|_2 - \frac{L_\phi^2\eta^2}{2}\|w_t\|_2^2 \\
&\overset{(iii)}{\geq} (1-\gamma)\eta\left(V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t})\right) - \eta\sqrt{C_d}\zeta' - \eta C_\phi C_\lambda \lambda \\
&\quad - \eta C_\phi \left\|w_t - W_t^{\lambda*}\right\|_2 - L_\phi^2\eta^2\left\|w_t - W_t^{\lambda*}\right\|_2^2 - L_\phi^2\eta^2\left\|W_t^{\lambda*}\right\|_2^2 \\
&\overset{(iv)}{\geq} (1-\gamma)\eta\left(V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t})\right) - \eta\sqrt{C_d}\zeta' - \eta C_\phi C_\lambda \lambda \\
&\quad - \eta C_\phi \left\|w_t - W_t^{\lambda*}\right\|_2 - L_\phi^2\eta^2\left\|w_t - W_t^{\lambda*}\right\|_2^2 - \frac{L_\phi^2\eta^2}{\lambda_P^2}\|\nabla_\theta V(\theta_t, r_{\alpha_t})\|_2^2,
\end{aligned} \tag{37}
$$

where (i) follows from the definition of ζ' in the statement of Theorem 5, (ii) follows from the upper bound on $\|\nabla_\theta \pi_\theta(a|s)\|_2$ in Assumption 5, Lemma 13 and Cauchy-Schwartz inequality, (iii) follows from the fact $\|A+B\|_2^2 \leq 2\|A\|_2^2 + 2\|B\|_2^2$, and (iv) follows from the definition of $W_t^{\lambda*}$ and because $\lambda_P I \preceq F(\theta_t) + \lambda I$.

Rearranging eq. (37), we obtain

$$
\begin{aligned}
V(\pi_{\theta^*}, r_{\alpha_t}) - V(\pi_{\theta_t}, r_{\alpha_t}) &\leq \frac{D(\theta_t) - D(\theta_{t+1})}{\eta(1-\gamma)} + \frac{\sqrt{C_d}\zeta'}{1-\gamma} + \frac{C_\phi C_\lambda \lambda}{1-\gamma} + \frac{C_\phi}{1-\gamma}\left\|w_t - W_t^{\lambda*}\right\|_2 \\
&\quad + \frac{L_\phi^2\eta}{1-\gamma}\left\|w_t - W_t^{\lambda*}\right\|_2^2 + \frac{L_\phi^2\eta}{\lambda_P^2(1-\gamma)}\|\nabla_\theta V(\theta_t, r_{\alpha_t})\|_2^2.
\end{aligned} \tag{38}
$$

Finally, we complete the proof as follows:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \\
&\quad = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[g(\theta_t) - F(\theta_t, \alpha_t)\right] + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[F(\theta_t, \alpha_t) - g(\theta^*)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[g(\theta_t)-F(\theta_t,\alpha_t)\right]+\frac{1}{T}\sum_{t=0}^{T-1}(F(\theta_t,\alpha_t)-F(\theta^*,\alpha_t))\\
&= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[g(\theta_t)-F(\theta_t,\alpha_t)\right]+\frac{1}{T}\sum_{t=0}^{T-1}(V(\pi_{\theta^*},r_{\alpha_t})-V(\pi_{\theta_t},r_{\alpha_t}))\\
&\overset{(ii)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[g(\theta_t)-F(\theta_t,\alpha_t)\right]+\frac{D(\theta_0)-D(\theta_T)}{(1-\gamma)\eta T}+\frac{\sqrt{C_d}\zeta'}{1-\gamma}+\frac{C_\phi C_\lambda\lambda}{1-\gamma}\\
&\quad+\frac{C_\phi}{(1-\gamma)T}\sum_{t=0}^{T-1}\left\|w_t-W_t^{\lambda*}\right\|_2+\frac{L_\phi^2\eta}{(1-\gamma)T}\sum_{t=0}^{T-1}\left\|w_t-W_t^{\lambda*}\right\|_2^2+\frac{L_\phi^2\eta R_{max}^2C_\phi^2}{(1-\gamma)^3\lambda_P^2}\\
&\overset{(iii)}{\leq} L_{22}^2C_\alpha^2e^{-\frac{\mu^2}{8L_{22}^2}K}+\frac{48C_r^2L_{22}^2}{\mu^2(1-\gamma)^2}(1+\frac{\rho C_M}{1-\rho})\frac{1}{B}+\frac{\mathbb{E}\left[D(\theta_0)-D(\theta_T)\right]}{(1-\gamma)^2\sqrt{T}}+\frac{\sqrt{C_d}\zeta'}{1-\gamma}+\frac{C_\phi C_\lambda\lambda}{1-\gamma}\\
&\quad+\frac{C_\phi}{(1-\gamma)T}\sum_{t=0}^{T-1}\left\|w_t-W_t^{\lambda*}\right\|_2+\frac{L_\phi^2}{T^{3/2}}\sum_{t=0}^{T-1}\left\|w_t-W_t^{\lambda*}\right\|_2^2+\frac{L_\phi^2R_{max}^2C_\phi^2}{(1-\gamma)^2\lambda_P^2\sqrt{T}}\\
&\overset{(iv)}{\leq} L_{22}^2C_\alpha^2e^{-\frac{\mu^2}{8L_{22}^2}K}+\frac{48C_r^2L_{22}^2}{\mu^2(1-\gamma)^2}(1+\frac{\rho C_M}{1-\rho})\frac{1}{B}+\frac{\mathbb{E}\left[D(\theta_0)-D(\theta_T)\right]}{(1-\gamma)^2\sqrt{T}}+\frac{\sqrt{C_d}\zeta'}{1-\gamma}+\frac{C_\phi C_\lambda\lambda}{1-\gamma}\\
&\quad+\frac{C_\phi}{(1-\gamma)}\sqrt{\exp\left\{-\frac{\lambda_P^2T_c}{16(C_\phi^2+\lambda)^2}\right\}\frac{R_{max}^2C_\phi^2}{\lambda_P^2(1-\gamma)^2}+\left(\frac{1}{\lambda_P}+\frac{\lambda_P}{2(C_\phi^2+\lambda)^2}\right)\frac{98R_{max}^2C_\phi^2[(C_\phi^2+\lambda)^2+4\lambda_P^2][1+(C_M-1)\rho]}{(1-\rho)(1-\gamma)^2\lambda_P^3M}}\\
&\quad+\frac{L_\phi^2}{\sqrt{T}}\left(\exp\left\{-\frac{\lambda_P^2T_c}{16(C_\phi^2+\lambda)^2}\right\}\frac{R_{max}^2C_\phi^2}{\lambda_P^2(1-\gamma)^2}+\left(\frac{1}{\lambda_P}+\frac{\lambda_P}{2(C_\phi^2+\lambda)^2}\right)\frac{98R_{max}^2C_\phi^2[(C_\phi^2+\lambda)^2+4\lambda_P^2][1+(C_M-1)\rho]}{(1-\rho)(1-\gamma)^2\lambda_P^3M}\right)\\
&\quad+\frac{L_\phi^2R_{max}^2C_\phi^2}{(1-\gamma)^2\lambda_P^2\sqrt{T}}\\
&\overset{(v)}{\leq} \mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{T}}\right)+\mathcal{O}\left(e^{-(1-\gamma)^2K}\right)+\mathcal{O}\left(\frac{1}{(1-\gamma)^4B}\right)\\
&\quad+\mathcal{O}\left(\frac{\zeta'}{(1-\gamma)^{3/2}}\right)+\mathcal{O}\left(\frac{\lambda}{1-\gamma}\right)+\mathcal{O}\left(e^{-T_c}\right)+\mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{M}}\right),
\end{aligned}
$$

where (i) follows because $g(\theta^*)=F(\theta^*,\alpha_{op}(\theta^*))\geq F(\theta^*,\alpha_t)$ and (ii) follows from eq. (38) and because $\|\nabla_\theta V(\theta_t,\alpha_t)\|_2\leq\frac{R_{max}C_\phi}{1-\gamma}$, (iii) follows from Proposition 1 and Lemma 5, and the fact $\eta=\frac{1-\gamma}{\sqrt{T}}$, (iv) follows from Lemma 14, and (v) follows because $L_{22}=\mathcal{O}\left(\frac{1}{1-\gamma}\right)$ and $C_d=\mathcal{O}\left(\frac{1}{1-\gamma}\right)$.