

## I The Information of Detailed Proofs

### I.1 Proof of Theorem 1

**Theorem 1. (Generalization for policy classes of distribution error)**

Suppose Assumption 1 holds, and for any  $\hat{\epsilon} > 0$ , given  $D_I$  and  $\pi_E$  with

$$e(\hat{c}_{D_I}^{(m)} \rho^{\pi_E}, \hat{c}_{D_I}^{(m)} \rho^{\Pi}) \geq \sup_{D \in \mathcal{D}} \{e(\hat{c}_D^{(m)} \rho^{\pi_E}, \hat{c}_D^{(m)} \rho^{\Pi})\} - \hat{\epsilon},$$

then

$$\begin{aligned} & e(C_{D_I} \rho^{\pi_E}, C_{D_I} \rho^{\Pi}) \\ & \geq \underbrace{\sup_{D \in \mathcal{D}} \{e(\hat{c}_D^{(m)} \rho^{\pi_E}, \hat{c}_D^{(m)} \rho^{\Pi})\}}_{\text{Appr}(\mathcal{D}, m)} - \underbrace{2\hat{\mathfrak{A}}_{\mathbb{D}_I}^{(m)}(C_{D_I} \rho^{\Pi}) - 8B_{\Pi} \sqrt{\frac{\log(3/\delta)}{2m}}}_{\text{Estm}(\Pi, m, \delta)} - \hat{\epsilon} \end{aligned}$$

for all  $\delta \in (0, 1)$ , with a probability of at least  $1 - \delta$ .

*Proof.* For notational simplicity, we denote  $z_i = (s_{D_I}^{(i)}, a_{D_I}^{(i)})$ ,  $Z = (z_1, \dots, z_m)$ . Then

$$\begin{aligned} \mathbb{E}_{(s,a) \sim \hat{\mathbb{D}}_I} [C_{D_I} \rho^{\pi_E}(s, a)] &= \hat{\mathbb{E}}_Z [C_{D_I} \rho^{\pi_E}], \\ \mathbb{E}_{(s,a) \sim \hat{\mathbb{D}}_I} [C_{D_I} \rho^{\pi}(s, a)] &= \hat{\mathbb{E}}_Z [C_{D_I} \rho^{\pi}]. \end{aligned}$$

By Eq. (2) and Definition 2, we only need to prove

$$e(C_{D_I} \rho^{\pi_E}, C_{D_I} \rho^{\Pi}) - e(\hat{c}_{D_I}^{(m)} \rho^{\pi_E}, \hat{c}_{D_I}^{(m)} \rho^{\Pi})$$

has a lower bound. Specifically,

$$\begin{aligned} & e(C_{D_I} \rho^{\pi_E}, C_{D_I} \rho^{\Pi}) - e(\hat{c}_{D_I}^{(m)} \rho^{\pi_E}, \hat{c}_{D_I}^{(m)} \rho^{\Pi}) \\ &= C_{D_I} \inf_{\pi \in \Pi} \{ \mathbb{E}_{(s,a) \sim \mathbb{D}_I} [\rho^{\pi_E}(s, a) - \rho^{\pi}(s, a)] \} \\ & \quad - \hat{c}_{D_I}^{(m)} \inf_{\pi \in \Pi} \{ \mathbb{E}_{(s,a) \sim \hat{\mathbb{D}}_I} [\rho^{\pi_E}(s, a) - \rho^{\pi}(s, a)] \} \\ & \geq C_{D_I} \inf_{\pi \in \Pi} \{ \mathbb{E}_{(s,a) \sim \mathbb{D}_I} [\rho^{\pi_E}(s, a) - \rho^{\pi}(s, a)] - \mathbb{E}_{(s,a) \sim \hat{\mathbb{D}}_I} [\rho^{\pi_E}(s, a) - \rho^{\pi}(s, a)] \} \\ & \geq C_{D_I} \inf_{\pi \in \Pi} \{ \mathbb{E}_{(s,a) \sim \mathbb{D}_I} [\rho^{\pi_E}(s, a)] - \mathbb{E}_{(s,a) \sim \hat{\mathbb{D}}_I} [\rho^{\pi_E}(s, a)] \} \\ & \quad + C_{D_I} \inf_{\pi \in \Pi} \{ \mathbb{E}_{(s,a) \sim \hat{\mathbb{D}}_I} [\rho^{\pi}(s, a)] - \mathbb{E}_{(s,a) \sim \mathbb{D}_I} [\rho^{\pi}(s, a)] \} \\ &= \left( \mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^{\pi_E}(s, a)] - \hat{\mathbb{E}}_Z [C_{D_I} \rho^{\pi_E}] \right) \\ & \quad - \sup_{\pi \in \Pi} \{ \mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^{\pi}(s, a)] - \hat{\mathbb{E}}_Z [C_{D_I} \rho^{\pi}] \}. \end{aligned} \tag{13}$$

First, we show that  $\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^{\pi_E}(s, a)] - \mathbb{E}_{(s,a) \sim \hat{\mathbb{D}}_I} [C_{D_I} \rho^{\pi_E}(s, a)]$  has a lower bound. Let

$$\phi_{\pi_E}(Z) = \mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^{\pi_E}(s, a)] - \hat{\mathbb{E}}_Z [C_{D_I} \rho^{\pi_E}].$$

Let  $Z$  and  $Z'$  be two samples differing by exactly one point, say  $z_i \in Z$ ,  $z'_i \in Z'$ . Note that for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , we have  $C_{D_I} \rho^{\pi_E}(s, a) \leq B_\Pi$ . Then

$$\begin{aligned} & \phi_{\pi_E}(Z') - \phi_{\pi_E}(Z) \\ &= (\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^{\pi_E}(s, a)] - \hat{\mathbb{E}}_{Z'} [C_{D_I} \rho^{\pi_E}]) \\ & \quad - (\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^{\pi_E}(s, a)] - \hat{\mathbb{E}}_Z [C_{D_I} \rho^{\pi_E}]) \\ &= \frac{1}{m} (C_{D_I} \rho^{\pi_E}(z_i) - C_{D_I} \rho^{\pi_E}(z'_i)) \leq \frac{2}{m} B_\Pi. \end{aligned}$$

By a similar derivation, we obtain  $\phi_{\pi_E}(Z) - \phi_{\pi_E}(Z') \leq \frac{2}{m} B_\Pi$ . Therefore, we have  $|\phi_{\pi_E}(Z) - \phi_{\pi_E}(Z')| \leq \frac{2}{m} B_\Pi$ . According to McDiarmid's inequality [21], we have

$$\phi_{\pi_E}(Z) - \mathbb{E}[\phi_{\pi_E}(Z)] \geq -2B_\Pi \sqrt{\frac{\log(3/\delta)}{2m}}$$

with a probability of at least  $1 - \frac{\delta}{3}$ , where the outer expectation is taken over the random choice of  $\hat{\mathbb{D}}_I$  with  $m$  state-action pairs. By the fact that  $\mathbb{E}[\phi_{\pi_E}(Z)] = 0$ , we have

$$\phi_{\pi_E}(Z) \geq -2B_\Pi \sqrt{\frac{\log(3/\delta)}{2m}} \quad (14)$$

with a probability of at least  $1 - \frac{\delta}{3}$ .

Next, we show that  $\sup_{\pi \in \Pi} \{\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^\pi(s, a)] - \mathbb{E}_{(s,a) \sim \hat{\mathbb{D}}_I} [C_{D_I} \rho^\pi(s, a)]\}$  has an upper bound. Let

$$\phi_\pi(Z) = \sup_{\pi \in \Pi} \{\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^\pi(s, a)] - \hat{\mathbb{E}}_Z [C_{D_I} \rho^\pi]\}.$$

Note that for all  $\pi \in \Pi$ , we have  $\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \{C_{D_I} \rho^\pi(s, a)\} \leq B_\Pi$ . Then

$$\begin{aligned} & \phi_\pi(Z') - \phi_\pi(Z) \\ &= (\sup_{\pi \in \Pi} \{\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^\pi(s, a)] - \hat{\mathbb{E}}_{Z'} [C_{D_I} \rho^\pi]\}) \\ & \quad - (\sup_{\pi \in \Pi} \{\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^\pi(s, a)] - \hat{\mathbb{E}}_Z [C_{D_I} \rho^\pi]\}) \\ &\leq \sup_{\pi \in \Pi} \left\{ (\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^\pi(s, a)] - \hat{\mathbb{E}}_{Z'} [C_{D_I} \rho^\pi]) \right. \\ & \quad \left. - (\mathbb{E}_{(s,a) \sim \mathbb{D}_I} [C_{D_I} \rho^\pi(s, a)] - \hat{\mathbb{E}}_Z [C_{D_I} \rho^\pi]) \right\} \\ &= \sup_{\pi \in \Pi} \{\hat{\mathbb{E}}_Z [C_{D_I} \rho^\pi] - \hat{\mathbb{E}}_{Z'} [C_{D_I} \rho^\pi]\} \\ &= \sup_{\pi \in \Pi} \left\{ \frac{C_{D_I} \rho^\pi(z_i) - C_{D_I} \rho^\pi(z'_i)}{m} \right\} \leq \frac{2}{m} B_\Pi. \end{aligned}$$

By a similar derivation, we obtain that  $\phi_\pi(Z) - \phi_\pi(Z') \leq \frac{2}{m}B_\Pi$ . Therefore,  $|\phi_\pi(Z) - \phi_\pi(Z')| \leq \frac{2}{m}B_\Pi$ . According to McDiarmid's inequality,

$$\phi_\pi(Z) \leq \mathbb{E}[\phi_\pi(Z)] + 2B_\Pi \sqrt{\frac{\log(3/\delta)}{2m}} \quad (15)$$

with a probability of at least  $1 - \frac{\delta}{3}$ , where the outer expectation is taken over the random choice of  $\hat{\mathbb{D}}_I$  with  $m$  state-action pairs.

In what follows, we prove that the right-hand side of Eq. (15) has an upper bound. By the standard Rademacher complexity technique [21], we obtain

$$\begin{aligned} \mathbb{E}[\phi_\pi(Z)] &= \mathbb{E}\left[\sup_{\pi \in \Pi} \{\mathbb{E}_{Z''}[\hat{\mathbb{E}}_{Z''}[C_{D_1}\rho^\pi]] - \hat{\mathbb{E}}_Z[C_{D_1}\rho^\pi]\}\right] \\ &= \mathbb{E}\left[\sup_{\pi \in \Pi} \{\mathbb{E}_{Z''}[\hat{\mathbb{E}}_{Z''}[C_{D_1}\rho^\pi] - \hat{\mathbb{E}}_Z[C_{D_1}\rho^\pi]\}\right] \\ &\leq \mathbb{E}_{Z, Z''}\left[\sup_{\pi \in \Pi} \{\hat{\mathbb{E}}_{Z''}[C_{D_1}\rho^\pi] - \hat{\mathbb{E}}_Z[C_{D_1}\rho^\pi]\}\right] \\ &= \mathbb{E}_{Z, Z''}\left[\frac{1}{m} \sup_{\pi \in \Pi} \left\{\sum_{i=1}^m (C_{D_1}\rho^\pi(z_i'') - C_{D_1}\rho^\pi(z_i))\right\}\right] \\ &= \mathbb{E}_{\sigma, Z, Z''}\left[\frac{1}{m} \sup_{\pi \in \Pi} \left\{\sum_{i=1}^m \sigma_i (C_{D_1}\rho^\pi(z_i'') - C_{D_1}\rho^\pi(z_i))\right\}\right] \\ &\leq \mathbb{E}_{\sigma, Z''}\left[\frac{1}{m} \sup_{\pi \in \Pi} \left\{\sum_{i=1}^m \sigma_i C_{D_1}\rho^\pi(z_i'')\right\}\right] \\ &\quad + \mathbb{E}_{\sigma, Z}\left[\frac{1}{m} \sup_{\pi \in \Pi} \left\{-\sum_{i=1}^m \sigma_i C_{D_1}\rho^\pi(z_i)\right\}\right] \\ &= 2\mathfrak{R}_{\mathbb{D}_I}^{(m)}(C_{D_1}\rho^\Pi), \end{aligned} \quad (16)$$

where  $\sigma_i$  is the i.i.d. Rademacher random variable for  $i = 1, \dots, m$ .

According to McDiarmid's inequality,

$$\mathfrak{R}_{\mathbb{D}_I}^{(m)}(C_{D_1}\rho^\Pi) \leq \hat{\mathfrak{R}}_{\mathbb{D}_I}^{(m)}(C_{D_1}\rho^\Pi) + 2B_\Pi \sqrt{\frac{\log(3/\delta)}{2m}} \quad (17)$$

with a probability of at least  $1 - \frac{\delta}{3}$ , where

$$\hat{\mathfrak{R}}_{\mathbb{D}_I}^{(m)}(C_{D_1}\rho^\Pi) = \mathbb{E}_\sigma \left[ \frac{1}{m} \sup_{\pi \in \Pi} \left\{ \sum_{i=1}^m \sigma_i C_{D_1}\rho^\pi(z_i) \right\} \right].$$

Combining Eq. (15) with Eq. (16) and (17),

$$\phi_\pi(Z) \leq 2\hat{\mathfrak{R}}_{\mathbb{D}_I}^{(m)}(C_{D_1}\rho^\Pi) + 6B_\Pi \sqrt{\frac{\log(3/\delta)}{2m}} \quad (18)$$

with a probability of at least  $1 - \frac{2\delta}{3}$ . Combining Eq. (14) with Eq. (18), we have

$$\phi_{\pi_E}(Z) - \phi_{\pi}(Z) \geq -2\hat{\mathfrak{R}}_{\mathbb{D}_I}^{(m)}(C_{D_I}\rho^{\Pi}) - 8B_{\Pi}\sqrt{\frac{\log(3/\delta)}{2m}} \quad (19)$$

with a probability of at least  $1 - \delta$ . Combining Eq. (13) with Eq. (19) and the condition of Theorem 1, we complete the proof.  $\square$

## I.2 The lower bound of $e(C_{D_I}\rho^{\pi_E}, C_{D_I}\rho^{\Pi})$ in terms of the covering number

The covering number of the function class  $C_{D_I}\rho^{\Pi}$  under the  $\ell_{\infty}$  distance  $\|\cdot\|_{\infty}$  can be denoted as  $\mathcal{N}(C_{D_I}\rho^{\Pi}, \epsilon, \|\cdot\|_{\infty})$ .

**Proposition 1.** *Under the same assumption of Theorem 1, then*

$$\begin{aligned} & e(C_{D_I}\rho^{\pi_E}, C_{D_I}\rho^{\Pi}) \\ & \geq \sup_{D \in \mathcal{D}} \{e(\hat{c}_D^{(m)}\rho^{\pi_E}, \hat{c}_D^{(m)}\rho^{\Pi})\} - \frac{8}{m} - \frac{24B_{\Pi}}{\sqrt{m}} \sqrt{\log(\mathcal{N}(C_{D_I}\rho^{\Pi}, \frac{1}{\sqrt{m}}, \|\cdot\|_{\infty}))} \\ & \quad - 8B_{\Pi}\sqrt{\frac{\log(3/\delta)}{2m}} - \hat{\epsilon} \end{aligned}$$

with a probability of at least  $1 - \delta$ .

*Proof.* We apply Dudley's entropy integral [8] to connect  $\hat{\mathfrak{R}}_{\mathbb{D}_I}^{(m)}(C_{D_I}\rho^{\Pi})$  with the covering number. Specifically, we have

$$\begin{aligned} \hat{\mathfrak{R}}_{\mathbb{D}_I}^{(m)}(C_{D_I}\rho^{\Pi}) & \leq \frac{4}{m} + \frac{12}{m} \int_{\frac{1}{\sqrt{m}}}^{\sqrt{m}B_{\Pi}} \sqrt{\log(\mathcal{N}(C_{D_I}\rho^{\Pi}, \epsilon, \|\cdot\|_{\infty}))} d\epsilon \\ & = \frac{4}{m} + \frac{12}{m} \sqrt{\log(\mathcal{N}(C_{D_I}\rho^{\Pi}, \xi_{\pi}, \|\cdot\|_{\infty}))} (\sqrt{m}B_{\Pi} - \frac{1}{\sqrt{m}}) \\ & \leq \frac{4}{m} + \frac{12B_{\Pi}}{\sqrt{m}} \sqrt{\log(\mathcal{N}(C_{D_I}\rho^{\Pi}, \frac{1}{\sqrt{m}}, \|\cdot\|_{\infty}))}, \end{aligned}$$

where  $\xi_{\pi} \in (\frac{1}{\sqrt{m}}, \sqrt{m}\frac{B_{\Pi}}{C_{D_I}})$ . Plugging this into Theorem 1, we obtain

$$\begin{aligned} & e(C_{D_I}\rho^{\pi_E}, C_{D_I}\rho^{\Pi}) \\ & \geq \sup_{D \in \mathcal{D}} \{e(\hat{c}_D^{(m)}\rho^{\pi_E}, \hat{c}_D^{(m)}\rho^{\Pi})\} - \frac{8}{m} - \frac{24B_{\Pi}}{\sqrt{m}} \sqrt{\log(\mathcal{N}(C_{D_I}\rho^{\Pi}, \frac{1}{\sqrt{m}}, \|\cdot\|_{\infty}))} \\ & \quad - 8B_{\Pi}\sqrt{\frac{\log(3/\delta)}{2m}} - \hat{\epsilon} \end{aligned}$$

with a probability of at least  $1 - \delta$ .  $\square$

### I.3 Proof of Corollary 1

**Corollary 1.** *Suppose Assumption 2 holds and  $\|\theta\|_2 \leq B_\theta$ . Then*

$$\begin{aligned} e(C_{D_I}\rho^{\pi_E}, C_{D_I}\rho^\Pi) &\geq \sup_{D \in \mathcal{D}} \{e(\hat{c}_D^{(m)}\rho^{\pi_E}, \hat{c}_D^{(m)}\rho^\Pi)\} \\ &\quad - \frac{8}{m} - \frac{24B_\Pi}{\sqrt{m}} \sqrt{p \log(1 + 2\sqrt{2m}B_\theta L_h)} - 8B_\Pi \sqrt{\frac{\log(3/\delta)}{2m}} - \hat{\epsilon} \end{aligned}$$

with a probability of at least  $1 - \delta$ .

*Proof.*  $C_{D_I}\rho^\pi(s, a)$  can be bounded by

$$|C_{D_I}\rho^\pi(s, a)| = |\theta^\top h(\psi_s, \psi_a)| \leq \|\theta\|_2 \|h(\psi_s, \psi_a)\|_2 \leq \sqrt{2}B_\theta L_h,$$

where the first inequality comes from Cauchy-Schwartz inequality.

To compute the covering number, we exploit the Lipschitz continuity of  $C_{D_I}\rho^\pi(s, a)$  with respect to the parameter  $\theta$ . Specifically, for two different parameters  $\theta$  and  $\theta'$ , we have

$$\begin{aligned} &\|C_{D_I}\rho^{\tilde{\pi}_\theta}(\psi_s, \psi_a) - C_{D_I}\rho^{\tilde{\pi}_{\theta'}}(\psi_s, \psi_a)\|_\infty \\ &= \|(\theta - \theta')^\top h(\psi_s, \psi_a)\|_\infty \\ &\stackrel{(i)}{\leq} \|\theta - \theta'\|_2 \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \{\|h(\psi_s, \psi_a)\|_2\} \\ &\stackrel{(ii)}{\leq} \|\theta - \theta'\|_2 L_h \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \{\sqrt{\|\psi_s\|_2^2 + \|\psi_a\|_2^2}\} \\ &\stackrel{(iii)}{\leq} \sqrt{2}L_h \|\theta - \theta'\|_2, \end{aligned}$$

where (i) comes from Cauchy-Schwartz inequality, (ii) comes from the Lipschitz continuity of  $h$ , and (iii) comes from the boundedness of  $\psi_s$  and  $\psi_a$ .

Denote  $\Theta = \{\theta : \|\theta\|_2 \leq B_\theta\}$ . By the standard argument of the volume ratio, we have

$$\begin{aligned} \mathcal{N}(C_{D_I}\rho^\Pi, \frac{1}{\sqrt{m}}, \|\cdot\|_\infty) &\leq \mathcal{N}(\Theta, \frac{1}{\sqrt{2m}L_h}, \|\cdot\|_2) \\ &\leq \left(1 + \frac{2B_\theta}{\frac{1}{\sqrt{2m}L_h}}\right)^p \\ &= (1 + 2\sqrt{2m}B_\theta L_h)^p. \end{aligned} \tag{20}$$

Plugging Eq. (20) into Proposition 1, we complete the proof.  $\square$

### I.4 Proof of Theorem 2

**Theorem 2. (GAIL Generalization for policy classes)** *Under the same assumption of Theorem 1, we have*

$$V_{\pi_E} - \sup_{\pi \in \Pi} V_\pi \geq \frac{1}{1 - \gamma} (\text{Appr}(\mathcal{D}, m) + \text{Estm}(\Pi, m, \delta) - \hat{\epsilon})$$

with a probability of at least  $1 - \delta$ .

*Proof.* Combining the definition of  $V_\pi$  in Eq. (1)

$$V_\pi = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim \rho^\pi} [r(s,a)]$$

with the reward function of GAIL  $r_{D_I}(s,a) = -\log(1 - D_I(s,a))$  [16], we obtain

$$\begin{aligned} & V_{\pi_E} - \sup_{\pi \in \Pi} V_\pi \\ &= \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim \rho^{\pi_E}} [r(s,a)] - \sup_{\pi \in \Pi} \left\{ \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim \rho^\pi} [r(s,a)] \right\} \\ &= \frac{1}{1-\gamma} \inf_{\pi \in \Pi} \left\{ \mathbb{E}_{(s,a) \sim \rho^{\pi_E}} [-\log(1 - D_I(s,a))] \right. \\ &\quad \left. - \mathbb{E}_{(s,a) \sim \rho^\pi} [-\log(1 - D_I(s,a))] \right\} \\ &= \frac{1}{1-\gamma} e(C_{D_I} \rho^{\pi_E}, C_{D_I} \rho^\Pi). \end{aligned} \tag{21}$$

Plugging Eq. (21) into the conclusion of Theorem 1, we complete the proof.  $\square$

## II Convergence Analysis of TSSG

The convergence of soft policy iteration in the SAC algorithm has been proved in [15]. Therefore, for a fixed  $\phi$ , the SAM submodule converges. Then we only need to prove the convergence of Alg. 2.

First, the definition of stationary point and two assumptions are given:

**Definition 3. (*Stationary point for TSSG*)** Suppose  $(w^*, \theta^*, \alpha^*)$  is a stationary point for the SAM submodule,  $\theta^*$  is a stationary point for TSSG if

$$\theta^* - \text{SAM}(\theta^*; w^*, \alpha^*, \phi^*(\theta^*)) = 0.$$

The stationarity is a necessary condition for optimality. Since each iteration of Alg. 2 contains  $T$  iterations of Alg. 1, the formula below can be used to measure the sub-stationarity of the TSSG algorithm at the  $N$ -th iteration:

$$\begin{aligned} I_N &= \min_{0 \leq k \leq N-1} \left( \mathbb{E} \|\theta^{(k)} - \text{SAM}(\theta^{(k)}; w_0^{(k)}, \alpha_0^{(k)}, \phi^*(\theta^{(k)}))\|_2^2 \right) \\ &= \min_{0 \leq k \leq N-1} \left( \mathbb{E} \left\| \sum_{t=0}^{T-1} \nabla_\theta F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \right\|_2^2 \right). \end{aligned}$$

**Assumption 4.** For any  $\theta, w, \alpha, \hat{\phi}(\theta)$  and  $k, t \in \mathbb{N}$ , there exist constants

$$M_\theta, M_w, M_\alpha > 0,$$

subject to:

Unbiasedness:

$$\mathbb{E} [\nabla_\theta \tilde{f}^{(j)}(\theta; w, \hat{\phi}(\theta), \alpha)] = \nabla_\theta F(\theta; w, \hat{\phi}(\theta), \alpha),$$

*Gradient boundedness:*

$$\begin{aligned}\mathbb{E}[\|\nabla_{\theta} \tilde{f}^{(j)}(\theta; w, \hat{\phi}(\theta), \alpha)\|_2^2] &\leq M_{\theta}, \\ \mathbb{E}[\|\nabla_w \tilde{J}_Q^{(j)}(w; \theta, \hat{\phi}(\theta), \alpha)\|_2^2] &\leq M_w, \\ \mathbb{E}[\|\nabla_{\alpha} \tilde{J}_t^{(k)}(\alpha; \theta)\|_2^2] &\leq M_{\alpha}.\end{aligned}$$

Assumption 4 is mild and satisfied by first-order optimization algorithms [8], e.g., the greedy stochastic gradient algorithm.

**Assumption 5.** (1) *There exists a constant  $L_Q > 0$ , subject to for any  $w$  and  $w'$ , we have*

$$\|Q_w^{\text{soft}} - Q_{w'}^{\text{soft}}\|_{\infty} \leq L_Q \|w - w'\|_2.$$

(2) *There exists constants  $S_g, S_Q > 0$ , subject to for any  $\theta$  and  $\theta'$ , we have*

$$\begin{aligned}\|\mathbb{E}_{\rho^{\pi_{\theta}}} [g(\psi_s, \psi_a)] - \mathbb{E}_{\rho^{\pi_{\theta'}}} [g(\psi_s, \psi_a)]\|_2 &\leq S_g \|\theta - \theta'\|_2, \\ \|\nabla_{\theta} \mathbb{E}_{\rho^{\pi_{\theta}}} [Q_w^{\text{soft}}(\psi_s, \psi_a; \theta, \phi)] - \nabla_{\theta} \mathbb{E}_{\rho^{\pi_{\theta'}}} [Q_w^{\text{soft}}(\psi_s, \psi_a; \theta', \phi)]\|_2 &\leq S_Q \|\theta - \theta'\|_2.\end{aligned}$$

(3) *There exists constants  $B_H, L_{\alpha}, S_H > 0$ , subject to for any  $\alpha, \alpha', \theta$  and  $\theta'$ , we have*

$$\begin{aligned}\mathbb{E}_{s \sim d^{\pi_{\theta}}} [\alpha \mathbb{H}(\pi_{\theta}(\cdot|s))] &\leq B_H, \\ |\mathbb{E}_{s \sim d^{\pi_{\theta}}} [\alpha \mathbb{H}(\pi_{\theta}(\cdot|s))] - \mathbb{E}_{s \sim d^{\pi_{\theta}}} [\alpha' \mathbb{H}(\pi_{\theta}(\cdot|s))]| &\leq L_{\alpha} \|\alpha - \alpha'\|_2, \\ \|\nabla_{\theta} \mathbb{E}_{s \sim d^{\pi_{\theta}}} [\alpha \mathbb{H}(\pi_{\theta}(\cdot|s))] - \nabla_{\theta} \mathbb{E}_{s \sim d^{\pi_{\theta'}}} [\alpha \mathbb{H}(\pi_{\theta'}(\cdot|s))]\|_2 &\leq S_H \|\theta - \theta'\|_2.\end{aligned}$$

(1) of Assumption 5 is the Lipschitz continuity of the soft Q-function with respect to its parameter  $w$ . (2) characterizes some Lipschitz continuity conditions with respect to the parameter  $\theta$ . (3) states some common regularity conditions for entropy [13,14,15]. The convergence of the TSSG algorithm is analyzed as follows.

**Theorem 3.** *Suppose Assumptions 3,4,5 hold. Given  $T$ , in the condition that the SAM submodule converges, for any  $\epsilon > 0$ , we take*

$$\begin{aligned}\eta_{\theta} &= \frac{\epsilon}{2T^2 M_{\theta} (2S_{\theta} + TS_g^2/\mu)}, \\ \eta_w &= \frac{\epsilon^2}{8T^4 L_Q M_{\theta} \sqrt{M_w} (2S_{\theta} + TS_g^2/\mu)}, \\ \eta_{\alpha} &= \frac{\epsilon^2}{8T^4 L_{\alpha} M_{\theta} \sqrt{M_{\alpha}} (2S_{\theta} + TS_g^2/\mu)},\end{aligned}$$

where  $S_{\theta} = S_H + S_Q$ . Then at most

$$N = \tilde{O}\left(\frac{T^3 B_F M_{\theta} (S_{\theta} + TS_g^2/\mu)}{\epsilon^2}\right)$$

iterations such that  $I_N \leq \epsilon$ , where  $B_F = \frac{\sqrt{2\kappa} L_g (2-\gamma)}{1-\gamma} + \frac{B_H}{1-\gamma} + \frac{\mu}{2} \kappa^2$ .

Here  $\tilde{O}$  hides high dependence on  $T$  and linear or quadratic dependence on some constants in Assumptions 3-5. Next, we prove Theorem 3.

### II.1 Boundedness of soft Q-function

**Lemma 1.** *For any  $w$ , we have  $\|Q_w^{\text{soft}}\|_\infty \leq B_Q$ , where  $B_Q = \frac{\sqrt{2}\kappa L_g}{1-\gamma} + \frac{\gamma B_H}{1-\gamma}$ .*

*Proof.* The reward function  $r_\phi(s, a)$  can be bounded by

$$|r_\phi(s, a)| \leq \|\phi\|_2 \cdot \|g(\psi_s, \psi_a)\|_2 \leq \sqrt{2}\kappa L_g.$$

By the definition of soft Q-function [13], we have

$$\begin{aligned} Q_w^{\text{soft}}(s_t, a_t; \theta, \phi) &= r_\phi(s_t, a_t) + \mathbb{E}_{(s_{t+1}, \dots) \sim d^{\pi_\theta}} \left[ \sum_{l=1}^{\infty} \gamma^l (r_\phi(s_{t+l}, a_{t+l}) + \alpha \mathbb{H}(\pi_\theta(\cdot | s_{t+l}))) \right] \\ &\leq \sum_{l=0}^{\infty} \gamma^l \sqrt{2}\kappa L_g + \sum_{l=1}^{\infty} \gamma^l B_H \\ &= \frac{\sqrt{2}\kappa L_g}{1-\gamma} + \frac{\gamma B_H}{1-\gamma}. \end{aligned}$$

□

### II.2 Lipschitz properties of the gradients

**Lemma 2.** *Suppose Assumption 5 holds. For any  $\theta, \theta', w, \phi, \alpha$ , we have*

$$\|\nabla_\theta F(\theta; w, \phi, \alpha) - \nabla_{\theta'} F(\theta'; w, \phi, \alpha)\|_2 \leq S_\theta \|\theta - \theta'\|_2,$$

where  $S_\theta = S_H + S_Q$ .

*Proof.* By [14,15], we have

$$\begin{aligned} \nabla_\theta F(\theta; w, \phi, \alpha) &= \mathbb{E}_{s_t \sim d^{\pi_\theta}} \left[ \mathbb{E}_{a_t \sim \pi_\theta} [\alpha \nabla_\theta \log(\pi_\theta(a_t | s_t)) \right. \\ &\quad \left. + (\alpha \nabla_{a_t} \log(\pi_\theta(a_t | s_t)) - \nabla_{a_t} Q_w^{\text{soft}}(s_t, a_t; \theta, \phi)) \nabla_\theta a_t] \right] \\ &= -\nabla_\theta \mathbb{E}_{s_t \sim d^{\pi_\theta}} [\alpha \mathbb{H}(\pi_\theta(\cdot | s_t))] - \nabla_\theta \mathbb{E}_{(s_t, a_t) \sim \rho^{\pi_\theta}} [Q_w^{\text{soft}}(s_t, a_t; \theta, \phi)]. \end{aligned}$$

Therefore,

$$\begin{aligned} &\|\nabla_\theta F(\theta; w, \phi, \alpha) - \nabla_{\theta'} F(\theta'; w, \phi, \alpha)\|_2 \\ &\leq \|\nabla_\theta \mathbb{E}_{s_t \sim d^{\pi_\theta}} [\alpha \mathbb{H}(\pi_\theta(\cdot | s_t))] - \nabla_{\theta'} \mathbb{E}_{s_t \sim d^{\pi_{\theta'}}} [\alpha \mathbb{H}(\pi_{\theta'}(\cdot | s_t))]\|_2 \\ &\quad + \|\nabla_\theta \mathbb{E}_{(s_t, a_t) \sim \rho^{\pi_\theta}} [Q_w^{\text{soft}}(s_t, a_t; \theta, \phi)] - \nabla_{\theta'} \mathbb{E}_{(s_t, a_t) \sim \rho^{\pi_{\theta'}}} [Q_w^{\text{soft}}(s_t, a_t; \theta', \phi)]\|_2 \\ &\leq (S_H + S_Q) \|\theta - \theta'\|_2. \end{aligned}$$

□



### II.3 Boundedness of $F$

**Lemma 3.** *Under Assumption 5, there exists  $B_F = \frac{\sqrt{2}\kappa L_g(2-\gamma)}{1-\gamma} + \frac{B_H}{1-\gamma} + \frac{\mu}{2}\kappa^2$  such that for any  $\theta, w, \phi, \alpha$ , we have  $|F(\theta; w, \phi, \alpha)| \leq B_F$ .*

*Proof.* By the definition of  $F(\theta; w, \phi, \alpha)$ , we have

$$\begin{aligned}
& |F(\theta; w, \phi, \alpha)| \\
& \leq \mathbb{E}_{(s,a) \sim \rho^{\pi_E}} [r_\phi(s, a)] + \mathbb{E}_{s_t \sim d^{\pi_\theta}} [\alpha \mathbb{H}(\pi_\theta(\cdot | s_t))] + \|Q_w^{\text{soft}}\|_\infty + \frac{\mu}{2} \|\phi\|_2^2 \\
& \leq \sqrt{2}\kappa L_g + B_H + B_Q + \frac{\mu}{2} \kappa^2 \\
& = \frac{\sqrt{2}\kappa L_g(2-\gamma)}{1-\gamma} + \frac{B_H}{1-\gamma} + \frac{\mu}{2} \kappa^2.
\end{aligned}$$

□

### II.4 Proof of Theorem 3

**Theorem 3.** *Suppose Assumptions 3,4,5 hold. Given  $T$ , in the condition that the SAM submodule converges, for any  $\epsilon > 0$ , we take*

$$\begin{aligned}
\eta_\theta &= \frac{\epsilon}{2T^2 M_\theta (2S_\theta + TS_g^2/\mu)}, \\
\eta_w &= \frac{\epsilon^2}{8T^4 L_Q M_\theta \sqrt{M_w} (2S_\theta + TS_g^2/\mu)}, \\
\eta_\alpha &= \frac{\epsilon^2}{8T^4 L_\alpha M_\theta \sqrt{M_\alpha} (2S_\theta + TS_g^2/\mu)},
\end{aligned}$$

where  $S_\theta = S_H + S_Q$ . Then at most

$$N = \tilde{O}\left(\frac{T^3 B_F M_\theta (S_\theta + TS_g^2/\mu)}{\epsilon^2}\right)$$

iterations such that  $I_N \leq \epsilon$ , where  $B_F = \frac{\sqrt{2}\kappa L_g(2-\gamma)}{1-\gamma} + \frac{B_H}{1-\gamma} + \frac{\mu}{2}\kappa^2$ .

*Proof.* Employing Lemma 2 and the Mean Value Theorem, we have

$$\begin{aligned}
& \sum_{t=0}^{T-1} \left( F(\theta_{t+1}^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) - F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \right) \\
& - \sum_{t=0}^{T-1} \langle \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}), \theta_{t+1}^{(k)} - \theta_t^{(k)} \rangle \\
& = \sum_{t=0}^{T-1} \left( \langle \nabla_{\theta} F(\tilde{\theta}_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}), \theta_{t+1}^{(k)} - \theta_t^{(k)} \rangle \right. \\
& \quad \left. - \langle \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}), \theta_{t+1}^{(k)} - \theta_t^{(k)} \rangle \right) \\
& \leq \sum_{t=0}^{T-1} \left\| \nabla_{\theta} F(\tilde{\theta}_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \right. \\
& \quad \left. - \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \right\|_2 \|\theta_{t+1}^{(k)} - \theta_t^{(k)}\|_2 \\
& \leq S_{\theta} \sum_{t=0}^{T-1} \|\tilde{\theta}_t^{(k)} - \theta_t^{(k)}\|_2 \|\theta_{t+1}^{(k)} - \theta_t^{(k)}\|_2 \leq S_{\theta} \sum_{t=0}^{T-1} \|\theta_{t+1}^{(k)} - \theta_t^{(k)}\|_2^2, \tag{22}
\end{aligned}$$

where  $\tilde{\theta}_t^{(k)}$  is some interpolation between  $\theta_{t+1}^{(k)}$  and  $\theta_t^{(k)}$ . Note that

$$\begin{aligned}
& \mathbb{E} \langle \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}), \theta_{t+1}^{(k)} - \theta_t^{(k)} \rangle \\
& \stackrel{(i)}{=} \mathbb{E} \langle \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}), -\eta_{\theta} (\nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_t^{(k)}) + \xi_{\theta_t}^k) \rangle \\
& \stackrel{(ii)}{=} \mathbb{E} \langle \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}), -\eta_{\theta} \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_t^{(k)}) \rangle \\
& \stackrel{(iii)}{=} \mathbb{E} \langle \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}), -\eta_{\theta} \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \rangle \\
& \stackrel{(iv)}{=} -\eta_{\theta} \mathbb{E} \|\nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)})\|_2^2, \tag{23}
\end{aligned}$$

where

$$\xi_{\theta_t}^k = \frac{1}{n_{\theta}} \sum_{j \in D_{\theta}^k} \nabla_{\theta} \tilde{f}^{(j)}(\theta_t^{(k)}; w_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_t^{(k)}) - \nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_t^{(k)}).$$

Here (ii) comes from the unbiased property of  $\tilde{f}^{(j)}$ , and (iii) comes from the unbiased property of  $\hat{\phi}(\theta^{(k)})$  and the linearity of  $\nabla_{\theta} F$  in  $\phi$ . Now taking the

expectation on both sides of Eq. (22) and plugging Eq. (23) in, we obtain

$$\begin{aligned}
& \sum_{t=0}^{T-1} \left( \mathbb{E}F(\theta_{t+1}^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) - \mathbb{E}F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \right) \\
& + \eta_\theta \sum_{t=0}^{T-1} (\mathbb{E} \|\nabla_\theta F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)})\|_2^2) \\
& \leq S_\theta \left( \sum_{t=0}^{T-1} \mathbb{E} \|\theta_{t+1}^{(k)} - \theta_t^{(k)}\|_2^2 \right) \\
& = S_\theta \eta_\theta^2 \left( \sum_{t=0}^{T-1} \frac{1}{n_\theta^2} \mathbb{E} \left\| \sum_{j \in D_\theta^t} (\nabla_\theta \tilde{f}^{(j)}(\theta_t^{(k)}; w_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_t^{(k)})) \right\|_2^2 \right) \\
& \leq S_\theta \eta_\theta^2 \left( \sum_{t=0}^{T-1} \left( \frac{1}{n_\theta} \sum_{j \in D_\theta^t} \mathbb{E} \|\nabla_\theta \tilde{f}^{(j)}(\theta_t^{(k)}; w_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_t^{(k)})\|_2^2 \right) \right) \\
& \leq S_\theta \eta_\theta^2 T M_\theta. \tag{24}
\end{aligned}$$

Dividing both sides by  $\eta_\theta$  and rearranging the terms in Eq. (24), we get

$$\begin{aligned}
& \sum_{t=0}^{T-1} (\mathbb{E} \|\nabla_\theta F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)})\|_2^2) \\
& \leq \frac{\sum_{t=0}^{T-1} \left( \mathbb{E}F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) - \mathbb{E}F(\theta_{t+1}^{(k)}; w_{t+2}^{(k)}, \phi^*(\theta^{(k)}), \alpha_{t+1}^{(k)}) \right)}{\eta_\theta} \\
& + \frac{\sum_{t=0}^{T-1} \left( \mathbb{E}F(\theta_{t+1}^{(k)}; w_{t+2}^{(k)}, \phi^*(\theta^{(k)}), \alpha_{t+1}^{(k)}) - \mathbb{E}F(\theta_{t+1}^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \right)}{\eta_\theta} \\
& + S_\theta \eta_\theta T M_\theta \\
& = \frac{\mathbb{E}F(\theta^{(k)}; w_1^{(k)}, \phi^*(\theta^{(k)}), \alpha_0^{(k)}) - \mathbb{E}F(\theta^{(k+1)}; w_{T+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_T^{(k)})}{\eta_\theta} \\
& + \frac{\sum_{t=0}^{T-1} \left( \mathbb{E}F(\theta_{t+1}^{(k)}; w_{t+2}^{(k)}, \phi^*(\theta^{(k)}), \alpha_{t+1}^{(k)}) - \mathbb{E}F(\theta_{t+1}^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \right)}{\eta_\theta} \\
& + S_\theta \eta_\theta T M_\theta. \tag{25}
\end{aligned}$$

By Assumption 5, for any  $\theta, w, w', \theta', \alpha, \alpha'$ , we have

$$\begin{aligned}
& |F(\theta; w, \phi^*(\theta'), \alpha) - F(\theta; w', \phi^*(\theta'), \alpha')| \\
& = |\mathbb{E}_{s_t \sim \mathcal{D}_1} [\mathbb{E}_{a_t \sim \pi_\theta} [\alpha \log(\pi_\theta(a_t | s_t)) - Q_w^{\text{soft}}(s_t, a_t; \theta, \phi^*(\theta'))]] \\
& \quad - \mathbb{E}_{s_t \sim \mathcal{D}_1} [\mathbb{E}_{a_t \sim \pi_\theta} [\alpha' \log(\pi_\theta(a_t | s_t)) - Q_{w'}^{\text{soft}}(s_t, a_t; \theta, \phi^*(\theta'))]]| \\
& \leq |\mathbb{E}_{s_t \sim \mathcal{D}_1} [\alpha \mathbb{H}(\pi_\theta(\cdot | s_t))] - \mathbb{E}_{s_t \sim \mathcal{D}_1} [\alpha' \mathbb{H}(\pi_\theta(\cdot | s_t))]| + \|Q_w^{\text{soft}} - Q_{w'}^{\text{soft}}\|_\infty \\
& \leq L_\alpha \|\alpha - \alpha'\| + L_Q \|w - w'\|_2.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
& \sum_{t=0}^{T-1} \left( \mathbb{E}F(\theta_{t+1}^{(k)}; w_{t+2}^{(k)}, \phi^*(\theta^{(k)}), \alpha_{t+1}^{(k)}) - \mathbb{E}F(\theta_{t+1}^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)}) \right) \\
& \leq \frac{\sum_{t=0}^{T-1} \left( L_\alpha \mathbb{E} \|\alpha_{t+1}^{(k)} - \alpha_t^{(k)}\|_2 + L_Q \mathbb{E} \|w_{t+2}^{(k)} - w_{t+1}^{(k)}\|_2 \right)}{\eta_\theta} \\
& = \sum_{t=0}^{T-1} \left( L_\alpha \eta_\alpha \mathbb{E} \|\nabla_\alpha \tilde{J}_t^{(k)}(\alpha_t^{(k)}; \theta_{t+1}^{(k)})\|_2 \right. \\
& \quad \left. + L_Q \eta_w \frac{\sum_{j \in D_w^t} \mathbb{E} \|\nabla_w \tilde{J}_Q^{(j)}(w_{t+1}^{(k)}; \theta_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_{t+1}^{(k)})\|_2}{n_w} \right) \\
& \leq \sum_{t=0}^{T-1} \left( L_\alpha \eta_\alpha \sqrt{\mathbb{E} \|\nabla_\alpha \tilde{J}_t^{(k)}(\alpha_t^{(k)}; \theta_{t+1}^{(k)})\|_2^2} \right. \\
& \quad \left. + L_Q \eta_w \frac{\sum_{j \in D_w^t} \sqrt{\mathbb{E} \|\nabla_w \tilde{J}_Q^{(j)}(w_{t+1}^{(k)}; \theta_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_{t+1}^{(k)})\|_2^2}}{n_w} \right) \\
& \leq TL_\alpha \sqrt{M_\alpha} \eta_\alpha + TL_Q \sqrt{M_w} \eta_w. \tag{26}
\end{aligned}$$

Plugging Eq. (26) into Eq. (25), we get

$$\begin{aligned}
& \sum_{t=0}^{T-1} \left( \mathbb{E} \|\nabla_\theta F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)})\|_2^2 \right) \\
& \leq \frac{\mathbb{E}F(\theta^{(k)}; w_1^{(k)}, \phi^*(\theta^{(k)}), \alpha_0^{(k)}) - \mathbb{E}F(\theta^{(k+1)}; w_{T+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_T^{(k)})}{\eta_\theta} \\
& \quad + TL_\alpha \sqrt{M_\alpha} \frac{\eta_\alpha}{\eta_\theta} + TL_Q \sqrt{M_w} \frac{\eta_w}{\eta_\theta} + TS_\theta M_\theta \eta_\theta \\
& = \frac{\mathbb{E}F(\theta^{(k)}; w_1^{(k)}, \phi^*(\theta^{(k)}), \alpha_0^{(k)}) - \mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^*(\theta^{(k+1)}), \alpha_0^{(k+1)})}{\eta_\theta} \\
& \quad + \frac{\mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^*(\theta^{(k+1)}), \alpha_0^{(k+1)}) - \mathbb{E}F(\theta^{(k+1)}; w_{T+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_T^{(k)})}{\eta_\theta} \\
& \quad + TL_\alpha \sqrt{M_\alpha} \frac{\eta_\alpha}{\eta_\theta} + TL_Q \sqrt{M_w} \frac{\eta_w}{\eta_\theta} + TS_\theta M_\theta \eta_\theta, \tag{27}
\end{aligned}$$

where  $w_{T+1}^{(k)} = w_1^{(k+1)}$ .

For notational simplicity, we define a vector function

$$G(\pi) = \mathbb{E}_{\rho^\pi} [g(\psi_s, \psi_a)].$$

Given a fixed  $\theta^{(k)}$ , by the definition of  $G$ , the optimal

$$\phi^*(\theta^{(k)}) = \frac{1}{\mu} [G(\pi_E) - G(\pi_{\theta^{(k)}})]$$

can be obtained.

Now consider

$$\mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^\star(\theta^{(k+1)}), \alpha_0^{(k+1)}) - \mathbb{E}F(\theta^{(k+1)}; w_{T+1}^{(k)}, \phi^\star(\theta^{(k)}), \alpha_T^{(k)}),$$

by the definition of soft Q-function [13]

$$\begin{aligned} & Q_w^{\text{soft}}(s_t, a_t; \theta, \phi) \\ &= r_\phi(s_t, a_t) + \mathbb{E}_{(s_{t+1}, \dots) \sim d^{\pi_\theta}} \left[ \sum_{l=1}^{\infty} \gamma^l (r_\phi(s_{t+l}, a_{t+l}) + \alpha \mathbb{H}(\pi_\theta(\cdot | s_{t+l}))) \right], \end{aligned}$$

we have

$$\begin{aligned} & \mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^\star(\theta^{(k+1)}), \alpha_0^{(k+1)}) - \mathbb{E}F(\theta^{(k+1)}; w_{T+1}^{(k)}, \phi^\star(\theta^{(k)}), \alpha_T^{(k)}) \\ &= \mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^\star(\theta^{(k+1)}), \alpha_0^{(k+1)}) - \mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^\star(\theta^{(k)}), \alpha_T^{(k)}) \\ &= \mathbb{E} \left[ \mathbb{E}_{(s,a) \sim \rho^{\pi_E}} \left[ r_{\phi^\star(\theta^{(k+1)})}(s, a) - r_{\phi^\star(\theta^{(k)})}(s, a) \right] \right. \\ & \quad + \mathbb{E}_{s_t \sim \mathcal{D}_I} \left[ \mathbb{E}_{a_t \sim \pi_{\theta^{(k+1)}}} \left[ -Q_{w_1^{(k+1)}}^{\text{soft}}(s_t, a_t; \theta^{(k+1)}, \phi^\star(\theta^{(k+1)})) \right. \right. \\ & \quad \left. \left. + Q_{w_1^{(k+1)}}^{\text{soft}}(s_t, a_t; \theta^{(k+1)}, \phi^\star(\theta^{(k)})) \right] \right] - \frac{\mu}{2} (\|\phi^\star(\theta^{(k+1)})\|_2^2 - \|\phi^\star(\theta^{(k)})\|_2^2) \\ &= \mathbb{E} \left[ \left( \mathbb{E}_{(s,a) \sim \rho^{\pi_E}} [r_{\phi^\star(\theta^{(k+1)})}(s, a)] - \mathbb{E}_{(s,a) \sim \rho^{\pi_{\theta^{(k+1)}}}} [r_{\phi^\star(\theta^{(k+1)})}(s, a)] \right) \right. \\ & \quad \left. - \left( \mathbb{E}_{(s,a) \sim \rho^{\pi_E}} [r_{\phi^\star(\theta^{(k)})}(s, a)] - \mathbb{E}_{(s,a) \sim \rho^{\pi_{\theta^{(k+1)}}}} [r_{\phi^\star(\theta^{(k)})}(s, a)] \right) \right] \\ & \quad - \frac{\mu}{2} (\|\phi^\star(\theta^{(k+1)})\|_2^2 - \|\phi^\star(\theta^{(k)})\|_2^2) \\ &= \mathbb{E} \langle G(\pi_E) - G(\pi_{\theta^{(k+1)}}), \phi^\star(\theta^{(k+1)}) - \phi^\star(\theta^{(k)}) \rangle \\ & \quad - \frac{\mu}{2} \mathbb{E} \langle \phi^\star(\theta^{(k+1)}) + \phi^\star(\theta^{(k)}), \phi^\star(\theta^{(k+1)}) - \phi^\star(\theta^{(k)}) \rangle \\ &= \mathbb{E} \langle \mu \phi^\star(\theta^{(k+1)}) - \frac{\mu}{2} (\phi^\star(\theta^{(k+1)}) + \phi^\star(\theta^{(k)})), \phi^\star(\theta^{(k+1)}) - \phi^\star(\theta^{(k)}) \rangle \\ &= \frac{\mu}{2} \mathbb{E} \|\phi^\star(\theta^{(k+1)}) - \phi^\star(\theta^{(k)})\|_2^2 \\ &= \frac{\mu}{2} \mathbb{E} \left\| \frac{1}{\mu} (G(\pi_{\theta^{(k+1)}}) - G(\pi_{\theta^{(k)}})) \right\|_2^2. \end{aligned}$$

Under Assumptions 5 and 4, we get

$$\begin{aligned}
& \mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^*(\theta^{(k+1)}), \alpha_0^{(k+1)}) - \mathbb{E}F(\theta^{(k+1)}; w_{T+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_T^{(k)}) \\
& \leq \frac{S_g^2}{2\mu} \mathbb{E}\|\theta^{(k+1)} - \theta^{(k)}\|_2^2 \leq \frac{TS_g^2}{2\mu} \left( \sum_{t=0}^{T-1} \mathbb{E}\|\theta_{t+1}^{(k)} - \theta_t^{(k)}\|_2^2 \right) \\
& \leq \frac{TS_g^2\eta_\theta^2}{2\mu n_\theta^2} \left( \sum_{t=0}^{T-1} (n_\theta \sum_{j \in D_\theta^t} \mathbb{E}\|\nabla_\theta \tilde{f}^{(j)}(\theta_t^{(k)}; w_{t+1}^{(k)}, \hat{\phi}(\theta^{(k)}), \alpha_t^{(k)})\|_2^2) \right) \\
& \leq \frac{T^2 S_g^2 M_\theta \eta_\theta^2}{2\mu}. \tag{28}
\end{aligned}$$

Plugging Eq. (28) into Eq. (27), we obtain

$$\begin{aligned}
& \sum_{t=0}^{T-1} (\mathbb{E}\|\nabla_\theta F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)})\|_2^2) \\
& \leq \frac{\mathbb{E}F(\theta^{(k)}; w_1^{(k)}, \phi^*(\theta^{(k)}), \alpha_0^{(k)}) - \mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^*(\theta^{(k+1)}), \alpha_0^{(k+1)})}{\eta_\theta} \\
& \quad + T L_\alpha \sqrt{M_\alpha} \frac{\eta_\alpha}{\eta_\theta} + T L_Q \sqrt{M_w} \frac{\eta_w}{\eta_\theta} + (T S_\theta M_\theta + \frac{T^2 S_g^2 M_\theta}{2\mu}) \eta_\theta. \tag{29}
\end{aligned}$$

Summing the equation Eq. (29) up, we have

$$\begin{aligned}
& \sum_{k=0}^{N-1} \left( \sum_{t=0}^{T-1} (\mathbb{E}\|\nabla_\theta F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)})\|_2^2) \right) \\
& \leq \frac{1}{\eta_\theta} \sum_{k=0}^{N-1} (\mathbb{E}F(\theta^{(k)}; w_1^{(k)}, \phi^*(\theta^{(k)}), \alpha_0^{(k)}) \\
& \quad - \mathbb{E}F(\theta^{(k+1)}; w_1^{(k+1)}, \phi^*(\theta^{(k+1)}), \alpha_0^{(k+1)})) + N T L_\alpha \sqrt{M_\alpha} \frac{\eta_\alpha}{\eta_\theta} \\
& \quad + N T L_Q \sqrt{M_w} \frac{\eta_w}{\eta_\theta} + N (T S_\theta M_\theta + \frac{T^2 S_g^2 M_\theta}{2\mu}) \eta_\theta.
\end{aligned}$$

Dividing both sides of the above equation by  $N$ , we get

$$\begin{aligned}
& \min_{0 \leq k \leq N-1} \sum_{t=0}^{T-1} (\mathbb{E}\|\nabla_\theta F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)})\|_2^2) \\
& \leq \frac{|F(\theta^{(0)}; w_1^{(0)}, \phi^*(\theta^{(0)}), \alpha_0^{(0)}) - \mathbb{E}F(\theta^{(N)}; w_1^{(N)}, \phi^*(\theta^{(N)}), \alpha_0^{(N)})|}{N \eta_\theta} \\
& \quad + T L_\alpha \sqrt{M_\alpha} \frac{\eta_\alpha}{\eta_\theta} + T L_Q \sqrt{M_w} \frac{\eta_w}{\eta_\theta} + (T S_\theta M_\theta + \frac{T^2 S_g^2 M_\theta}{2\mu}) \eta_\theta.
\end{aligned}$$

By Lemma 3, we have

$$|F(\theta^{(0)}; w_1^{(0)}, \phi^*(\theta^{(0)}), \alpha_0^{(0)}) - \mathbb{E}F(\theta^{(N)}; w_1^{(N)}, \phi^*(\theta^{(N)}), \alpha_0^{(N)})| \leq 2B_F.$$

Then we obtain

$$\begin{aligned} I_N &\leq T \min_{0 \leq k \leq N-1} \sum_{t=0}^{T-1} (\mathbb{E} \|\nabla_{\theta} F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^*(\theta^{(k)}), \alpha_t^{(k)})\|_2^2) \\ &\leq \frac{2TB_F}{N\eta_{\theta}} + T^2 L_{\alpha} \sqrt{M_{\alpha}} \frac{\eta_{\alpha}}{\eta_{\theta}} + T^2 L_Q \sqrt{M_w} \frac{\eta_w}{\eta_{\theta}} + T^2 M_{\theta} (S_{\theta} + \frac{TS_g^2}{2\mu}) \eta_{\theta}, \end{aligned}$$

where  $B_F = \frac{\sqrt{2}\kappa L_g(2-\gamma)}{1-\gamma} + \frac{B_H}{1-\gamma} + \frac{\mu}{2}\kappa^2$ .

Given any  $\epsilon > 0$ , take

$$\begin{aligned} \eta_{\theta} &= \frac{\epsilon}{2T^2 M_{\theta} (2S_{\theta} + TS_g^2/\mu)}, \\ \eta_w &= \frac{\epsilon^2}{8T^4 L_Q M_{\theta} \sqrt{M_w} (2S_{\theta} + TS_g^2/\mu)}, \\ \eta_{\alpha} &= \frac{\epsilon^2}{8T^4 L_{\alpha} M_{\theta} \sqrt{M_{\alpha}} (2S_{\theta} + TS_g^2/\mu)}, \end{aligned}$$

then we need at most

$$N = \frac{8TB_F}{\epsilon\eta_{\theta}} = \tilde{O}\left(\frac{T^3 B_F M_{\theta} (S_{\theta} + TS_g^2/\mu)}{\epsilon^2}\right)$$

iterations such that  $I_N \leq \epsilon$ . □

Theorem 3 discloses that despite the min-max computational formulation of GAIL without convex-concave structure, the TSSG algorithm can still converge to a stationary point.