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 π_{E} with*

$$\boldsymbol{e}(\hat{c}_{D_{\mathrm{I}}}^{(m)}\rho^{\pi_{\mathrm{E}}}, \hat{c}_{D_{\mathrm{I}}}^{(m)}\rho^{\Pi}) \geq \sup_{D\in\mathcal{D}}\{\boldsymbol{e}(\hat{c}_{D}^{(m)}\rho^{\pi_{\mathrm{E}}}, \hat{c}_{D}^{(m)}\rho^{\Pi})\} - \hat{\epsilon},$$

then

$$\begin{aligned}
&\boldsymbol{e}(C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}, C_{D_{\mathrm{I}}}\rho^{\Pi}) \\
\geq &\underbrace{\sup_{D\in\mathcal{D}}\{\boldsymbol{e}(\hat{c}_{D}^{(m)}\rho^{\pi_{\mathrm{E}}}, \hat{c}_{D}^{(m)}\rho^{\Pi})\}}_{\mathrm{Appr}(\mathcal{D},m)} \underbrace{-2\hat{\mathfrak{R}}_{\mathbb{D}_{\mathrm{I}}}^{(m)}(C_{D_{\mathrm{I}}}\rho^{\Pi}) - 8B_{\Pi}\sqrt{\frac{\log(3/\delta)}{2m}}}_{\mathrm{Estm}(\Pi,m,\delta)} - \hat{\epsilon}
\end{aligned}$$

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

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

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim\hat{\mathbb{D}}_{\mathrm{I}}}\big[C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}(s,a)\big] &= \hat{\mathbb{E}}_Z\big[C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}\big], \\
\mathbb{E}_{(s,a)\sim\hat{\mathbb{D}}_{\mathrm{I}}}\big[C_{D_{\mathrm{I}}}\rho^{\pi}(s,a)\big] &= \hat{\mathbb{E}}_Z\big[C_{D_{\mathrm{I}}}\rho^{\pi}\big].
\end{aligned}$$

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

$$\boldsymbol{e}(C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}, C_{D_{\mathrm{I}}}\rho^{\Pi}) - \boldsymbol{e}(\hat{c}_{D_{\mathrm{I}}}^{(m)}\rho^{\pi_{\mathrm{E}}}, \hat{c}_{D_{\mathrm{I}}}^{(m)}\rho^{\Pi})$$

has a lower bound. Specifically,

$$\begin{aligned}
&\boldsymbol{e}(C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}, C_{D_{\mathrm{I}}}\rho^{\Pi}) - \boldsymbol{e}(\hat{c}_{D_{\mathrm{I}}}^{(m)}\rho^{\pi_{\mathrm{E}}}, \hat{c}_{D_{\mathrm{I}}}^{(m)}\rho^{\Pi}) \\
= &C_{D_{\mathrm{I}}}\inf_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\mathbb{D}_{\mathrm{I}}}\big[\rho^{\pi_{\mathrm{E}}}(s,a) - \rho^{\pi}(s,a)\big]\} \\
&\qquad\qquad - \hat{c}_{D_{\mathrm{I}}}^{(m)}\inf_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\hat{\mathbb{D}}_{\mathrm{I}}}\big[\rho^{\pi_{\mathrm{E}}}(s,a) - \rho^{\pi}(s,a)\big]\} \\
\geq &C_{D_{\mathrm{I}}}\inf_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\mathbb{D}_{\mathrm{I}}}\big[\rho^{\pi_{\mathrm{E}}}(s,a) - \rho^{\pi}(s,a)\big] - \mathbb{E}_{(s,a)\sim\hat{\mathbb{D}}_{\mathrm{I}}}\big[\rho^{\pi_{\mathrm{E}}}(s,a) - \rho^{\pi}(s,a)\big]\} \\
\geq &C_{D_{\mathrm{I}}}\inf_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\mathbb{D}_{\mathrm{I}}}\big[\rho^{\pi_{\mathrm{E}}}(s,a)\big] - \mathbb{E}_{(s,a)\sim\hat{\mathbb{D}}_{\mathrm{I}}}\big[\rho^{\pi_{\mathrm{E}}}(s,a)\big]\} \\
&\qquad\qquad + C_{D_{\mathrm{I}}}\inf_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\hat{\mathbb{D}}_{\mathrm{I}}}\big[\rho^{\pi}(s,a)\big] - \mathbb{E}_{(s,a)\sim\mathbb{D}_{\mathrm{I}}}\big[\rho^{\pi}(s,a)\big]\} \\
= &\Big(\mathbb{E}_{(s,a)\sim\mathbb{D}_{\mathrm{I}}}\big[C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}(s,a)\big] - \hat{\mathbb{E}}_Z\big[C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}\big]\Big) \\
&\qquad\qquad - \sup_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\mathbb{D}_{\mathrm{I}}}\big[C_{D_{\mathrm{I}}}\rho^{\pi}(s,a)\big] - \hat{\mathbb{E}}_Z\big[C_{D_{\mathrm{I}}}\rho^{\pi}\big]\}.
\end{aligned} \tag{13}$$

First, we show that $\mathbb{E}_{(s,a)\sim\mathbb{D}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}(s,a)\big] - \mathbb{E}_{(s,a)\sim\hat{\mathbb{D}}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}(s,a)\big]$ has a lower bound. Let

$$\phi_{\pi_\mathrm{E}}(Z) = \mathbb{E}_{(s,a)\sim\mathbb{D}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}(s,a)\big] - \hat{\mathbb{E}}_Z\big[C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}\big].$$

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 $C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}(s,a) \le B_\Pi$. Then

$$\begin{aligned}
&\phi_{\pi_\mathrm{E}}(Z') - \phi_{\pi_\mathrm{E}}(Z)\\
&= (\mathbb{E}_{(s,a)\sim\mathbb{D}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}(s,a)\big] - \hat{\mathbb{E}}_{Z'}\big[C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}\big])\\
&\qquad\qquad - (\mathbb{E}_{(s,a)\sim\mathbb{D}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}(s,a)\big] - \hat{\mathbb{E}}_{Z}\big[C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}\big])\\
&= \frac{1}{m}\big(C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}(z_i) - C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}(z_i')\big) \le \frac{2}{m}B_\Pi.
\end{aligned}$$

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

$$\phi_{\pi_\mathrm{E}}(Z) - \mathbb{E}\big[\phi_{\pi_\mathrm{E}}(Z)\big] \ge -2B_\Pi\sqrt{\frac{\log(3/\delta)}{2m}}$$

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

$$\phi_{\pi_\mathrm{E}}(Z) \ge -2B_\Pi\sqrt{\frac{\log(3/\delta)}{2m}} \tag{14}$$

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

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

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

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

$$\begin{aligned}
&\phi_\pi(Z') - \phi_\pi(Z)\\
&= (\sup_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\mathbb{D}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi}(s,a)\big] - \hat{\mathbb{E}}_{Z'}\big[C_{D_\mathrm{I}}\rho^{\pi}\big]\})\\
&\qquad\qquad - (\sup_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\mathbb{D}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi}(s,a)\big] - \hat{\mathbb{E}}_{Z}\big[C_{D_\mathrm{I}}\rho^{\pi}\big]\})\\
&\le \sup_{\pi\in\Pi}\Big\{\big(\mathbb{E}_{(s,a)\sim\mathbb{D}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi}(s,a)\big] - \hat{\mathbb{E}}_{Z'}\big[C_{D_\mathrm{I}}\rho^{\pi}\big]\big)\\
&\qquad\qquad - \big(\mathbb{E}_{(s,a)\sim\mathbb{D}_\mathrm{I}}\big[C_{D_\mathrm{I}}\rho^{\pi}(s,a)\big] - \hat{\mathbb{E}}_{Z}\big[C_{D_\mathrm{I}}\rho^{\pi}\big]\big)\Big\}\\
&= \sup_{\pi\in\Pi}\{\hat{\mathbb{E}}_Z\big[C_{D_\mathrm{I}}\rho^{\pi}\big] - \hat{\mathbb{E}}_{Z'}\big[C_{D_\mathrm{I}}\rho^{\pi}\big]\}\\
&= \sup_{\pi\in\Pi}\{\frac{C_{D_\mathrm{I}}\rho^{\pi}(z_i) - C_{D_\mathrm{I}}\rho^{\pi}(z_i')}{m}\} \le \frac{2}{m}B_\Pi.
\end{aligned}$$

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,

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

with a probability of at least $1 - \frac{\delta}{3}$, where the outer expectation is taken over the random choice of $\hat{\mathbb{D}}_{\mathrm{I}}$ 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

$$\begin{aligned}
\mathbb{E}\big[\phi_\pi(Z)\big] &= \mathbb{E}\big[\sup_{\pi\in\Pi}\{\mathbb{E}_{Z''}\Big[\hat{\mathbb{E}}_{Z''}\big[C_{D_{\mathrm{I}}}\rho^\pi\big]\Big] - \hat{\mathbb{E}}_Z\big[C_{D_{\mathrm{I}}}\rho^\pi\big]\}\big] \\
&= \mathbb{E}\big[\sup_{\pi\in\Pi}\{\mathbb{E}_{Z''}\Big[\hat{\mathbb{E}}_{Z''}\big[C_{D_{\mathrm{I}}}\rho^\pi\big] - \hat{\mathbb{E}}_Z\big[C_{D_{\mathrm{I}}}\rho^\pi\big]\Big]\}\big] \\
&\leq \mathbb{E}_{Z,Z''}\Big[\sup_{\pi\in\Pi}\{\hat{\mathbb{E}}_{Z''}\big[C_{D_{\mathrm{I}}}\rho^\pi\big] - \hat{\mathbb{E}}_Z\big[C_{D_{\mathrm{I}}}\rho^\pi\big]\}\Big] \\
&= \mathbb{E}_{Z,Z''}\Big[\frac{1}{m}\sup_{\pi\in\Pi}\{\sum_{i=1}^m \big(C_{D_{\mathrm{I}}}\rho^\pi(z_i'') - C_{D_{\mathrm{I}}}\rho^\pi(z_i)\big)\}\Big] \\
&= \mathbb{E}_{\sigma,Z,Z''}\Big[\frac{1}{m}\sup_{\pi\in\Pi}\{\sum_{i=1}^m \sigma_i\big(C_{D_{\mathrm{I}}}\rho^\pi(z_i'') - C_{D_{\mathrm{I}}}\rho^\pi(z_i)\big)\}\Big] \\
&\leq \mathbb{E}_{\sigma,Z''}\Big[\frac{1}{m}\sup_{\pi\in\Pi}\{\sum_{i=1}^m \sigma_i C_{D_{\mathrm{I}}}\rho^\pi(z_i'')\}\Big] \\
&\qquad\qquad + \mathbb{E}_{\sigma,Z}\Big[\frac{1}{m}\sup_{\pi\in\Pi}\{-\sum_{i=1}^m \sigma_i C_{D_{\mathrm{I}}}\rho^\pi(z_i)\}\Big] \\
&= 2\mathfrak{R}_{\mathbb{D}_{\mathrm{I}}}^{(m)}(C_{D_{\mathrm{I}}}\rho^\Pi),
\end{aligned} \tag{16}$$

where σ_i, $i = 1, ..., m$, are i.i.d. Rademacher random variables.

According to McDiarmid's inequality,

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

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

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

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

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

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

$$\phi_{\pi_{\mathrm{E}}}(Z)-\phi_{\pi}(Z)\geq -2\hat{\mathfrak{R}}_{\mathbb{D}_{\mathrm{I}}}^{(m)}(C_{D_{\mathrm{I}}}\rho^{\Pi})-8B_{\Pi}\sqrt{\frac{\log(3/\delta)}{2m}} \tag{19}$$

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

I.2 The lower bound of $\boldsymbol{e}(C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}, C_{D_{\mathrm{I}}}\rho^{\Pi})$ in terms of the covering number

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

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

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

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

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

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

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

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

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

I.3 Proof of Corollary 1

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

$$
\begin{aligned}
\boldsymbol{e}(C_{D_{\mathrm{I}}}\rho^{\pi_{\mathrm{E}}}, C_{D_{\mathrm{I}}}\rho^{\Pi}) \ge \sup_{D\in\mathcal{D}} & \{\boldsymbol{e}(\hat{c}_D^{(m)}\rho^{\pi_{\mathrm{E}}}, \hat{c}_D^{(m)}\rho^{\Pi})\} \\
& -\frac{8}{m} - \frac{24B_\Pi}{\sqrt{m}}\sqrt{p\log(1+2\sqrt{2m}B_\theta L_h)} - 8B_\Pi\sqrt{\frac{\log(3/\delta)}{2m}} - \hat{\epsilon}
\end{aligned}
$$

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

Proof. $C_{D_{\mathrm{I}}}\rho^{\pi}(s,a)$ can be bounded by

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

where the first inequality comes from Cauchy-Schwartz inequality.

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

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

where (i) comes from Cauchy-Schwartz inequality, (ii) comes from the Lipschitz continuity of h, and (iii) comes from the boundedness of ψ_s and ψ_a.

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

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

Plugging Eq. (20) into Proposition 1, we complete the proof. □

I.4 Proof of Theorem 2

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

$$
V_{\pi_{\mathrm{E}}} - \sup_{\pi\in\Pi} V_\pi \ge \frac{1}{1-\gamma}\big(\mathrm{Appr}(\mathcal{D},m) + \mathrm{Estm}(\Pi,m,\delta) - \hat{\epsilon}\big)
$$

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

Proof. Combining the definition of V_π in Eq. (1)

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

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

$$\begin{aligned}
&V_{\pi_\mathrm{E}} - \sup_{\pi\in\Pi} V_\pi \\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho^{\pi_\mathrm{E}}}\big[r(s,a)\big] - \sup_{\pi\in\Pi}\{\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho^\pi}\big[r(s,a)\big]\} \\
&= \frac{1}{1-\gamma}\inf_{\pi\in\Pi}\{\mathbb{E}_{(s,a)\sim\rho^{\pi_\mathrm{E}}}\big[-\log(1-D_\mathrm{I}(s,a))\big] \\
&\qquad\qquad - \mathbb{E}_{(s,a)\sim\rho^\pi}\big[-\log(1-D_\mathrm{I}(s,a))\big]\} \\
&= \frac{1}{1-\gamma}\boldsymbol{e}(C_{D_\mathrm{I}}\rho^{\pi_\mathrm{E}}, C_{D_\mathrm{I}}\rho^{\Pi}).
\end{aligned} \tag{21}$$

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

II Convergence Analysis of TSSG

The convergence of soft policy iteration in SAC algorithms has been proved in [15]. Therefore, for a fixed ϕ, the SAM submodule converges. Then we only need to prove the convergence of Algorithm 2.

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

Definition 3. ***(stationary point)*** *Suppose there exists T such that the SAM submodule converges and $(\theta^\star, w_0^\star, \alpha_0^\star)$ is a stationary point for the SAM submodule, then $\theta^\star$ is a stationary point for TSSG if*

$$\sum_{t=0}^{T-1}\nabla_\theta F(\theta_t^\star; w_{t+1}^\star, \phi^\star(\theta^\star), \alpha_t^\star) = 0,$$

where $\theta_0^\star = \theta^\star$.

The stationarity is a necessary condition for optimality. Accordingly, the formula below can be used to measure the sub-stationarity of the algorithm at the $N-$th iteration:

$$I_N = \min_{0\le k\le N-1}\Big(\mathbb{E}\|\sum_{t=0}^{T-1}\nabla_\theta F(\theta_t^{(k)}; w_{t+1}^{(k)}, \phi^\star(\theta^{(k)}), \alpha_t^{(k)})\|_2^2\Big).$$

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:

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

Gradient boundedness:

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

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*

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

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

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

(3) There exists constants $B_H, L_\alpha, S_H > 0$, subject to for any α, α', θ and θ', we have

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

(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 θ. (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*

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

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

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

iterations such that $I_N \le \epsilon$, 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 $\|Q_w^{\mathrm{soft}}\|_\infty \le B_Q$, 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

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

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

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

□

II.2 Lipschitz properties of the gradients

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

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

where $S_\theta = S_H + S_Q$.

Proof. By [14,15], we have

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

Therefore,

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

□

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* θ, w, ϕ, α*, we have* $|F(\theta; w, \phi, \alpha)| \leq B_F$*.*

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

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

□

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*

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

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

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

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

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

where

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

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

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

Dividing both sides by η_θ and rearranging the terms in Eq. (24), we get

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

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

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

Therefore, we obtain

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

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

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

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

$$\phi^\star(\theta^{(k)})=\frac{1}{\mu}[G(\pi_{\mathrm{E}})-G(\pi_{\theta^{(k)}})]$$

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]

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

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}\Big[\mathbb{E}_{(s,a)\sim\rho^{\pi_{\mathrm{E}}}}\Big[r_{\phi^\star(\theta^{(k+1)})}(s,a) - r_{\phi^\star(\theta^{(k)})}(s,a)\Big] \\ &\quad + \mathbb{E}_{s_t\sim\mathcal{D}_{\mathrm{I}}}\Big[\mathbb{E}_{a_t\sim\pi_{\theta^{(k+1)}}}\big[- Q^{\text{soft}}_{w_1^{(k+1)}}(s_t, a_t; \theta^{(k+1)}, \phi^\star(\theta^{(k+1)})) \\ &\quad + Q^{\text{soft}}_{w_1^{(k+1)}}(s_t, a_t; \theta^{(k+1)}, \phi^\star(\theta^{(k)}))\big]\Big]\Big] - \frac{\mu}{2}(\|\phi^\star(\theta^{(k+1)})\|_2^2 - \|\phi^\star(\theta^{(k)})\|_2^2) \\ &= \mathbb{E}\Big[\big(\mathbb{E}_{(s,a)\sim\rho^{\pi_{\mathrm{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)]\big) \\ &\quad - \big(\mathbb{E}_{(s,a)\sim\rho^{\pi_{\mathrm{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)]\big)\Big] \\ &\quad - \frac{\mu}{2}(\|\phi^\star(\theta^{(k+1)})\|_2^2 - \|\phi^\star(\theta^{(k)})\|_2^2) \\ &= \mathbb{E}\langle G(\pi_{\mathrm{E}}) - G(\pi_{\theta^{(k+1)}}), \phi^\star(\theta^{(k+1)}) - \phi^\star(\theta^{(k)})\rangle \\ &\qquad\qquad - \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}\|\frac{1}{\mu}(G(\pi_{\theta^{(k+1)}}) - G(\pi_{\theta^{(k)}}))\|_2^2. \end{aligned}$$

Under Assumptions 5 and 4, we get

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

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

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

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

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

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

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

By Lemma 3, we have

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

Then we obtain

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

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

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

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

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.