Supplementary Material for "Instability in Generative Adversarial Imitation Learning with Deterministic Policy"

A Analysis of Exploding Gradients in DE-GAIL

A.1 Proof of Theorem 1

Theorem 1 *Let $\pi_h(\cdot|s)$ be the Gaussian stochastic policy with mean $h(s)$ and covariance $\boldsymbol{\Sigma}$. When the discriminator achieves its optimum $D^*(s,a)$ in Eq. (6), the gradient estimator of the policy loss with respect to the policy's parameter h satisfies $\|\hat{\nabla}_h D_{\mathrm{JS}}(\rho^{\pi_h}, \rho^{\pi_\mathrm{E}})\|_2 \to \infty$ with a probability of $\Pr(\|\boldsymbol{\Sigma}^{-1}(a_t - h(s_t))\|_2 \geq C$ for any $C > 0)$ as $\boldsymbol{\Sigma} \to \mathbf{0}$, where*

$$\hat{\nabla}_h D_{\mathrm{JS}}(\rho^{\pi_h}, \rho^{\pi_\mathrm{E}}) = \frac{d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t)}{2 d^{\pi_\mathrm{E}}(s_t) \pi_\mathrm{E}(a_t|s_t)} \log \frac{2 d^{\pi_h}(s_t) \pi_h(a_t|s_t)}{d^{\pi_h}(s_t) \pi_h(a_t|s_t) + d^{\pi_\mathrm{E}}(s_t) \pi_\mathrm{E}(a_t|s_t)},$$

and $\nabla_h \pi_h(a|s) = \pi_h(a|s) \kappa(s, \cdot) \boldsymbol{\Sigma}^{-1}(a - h(s))$.

Proof. Through importance sampling which transfers the learned state-action distribution to the expert demonstration distribution, the JS divergence can be rewritten from the definition in Eq. (4) as

$$
\begin{aligned}
&D_{\mathrm{JS}}(\rho^{\pi_h}, \rho^{\pi_\mathrm{E}}) \\
&= \frac{1}{2} D_{\mathrm{KL}}(\rho^{\pi_h}, \frac{\rho^{\pi_h} + \rho^{\pi_\mathrm{E}}}{2}) + \frac{1}{2} D_{\mathrm{KL}}(\rho^{\pi_\mathrm{E}}, \frac{\rho^{\pi_h} + \rho^{\pi_\mathrm{E}}}{2}) \\
&= \frac{1}{2} \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[\log \frac{2 \rho^{\pi_h}(s,a)}{\rho^{\pi_h}(s,a) + \rho^{\pi_\mathrm{E}}(s,a)} \right] + \frac{1}{2} \mathbb{E}_{(s,a) \sim \mathcal{D}^\star} \left[\log \frac{2 \rho^{\pi_\mathrm{E}}(s,a)}{\rho^{\pi_h}(s,a) + \rho^{\pi_\mathrm{E}}(s,a)} \right] \\
&= \frac{1}{2} \mathbb{E}_{(s,a) \sim \mathcal{D}^\star} \left[\frac{\rho^{\pi_h}(s,a)}{\rho^{\pi_\mathrm{E}}(s,a)} \log \frac{2 \rho^{\pi_h}(s,a)}{\rho^{\pi_h}(s,a) + \rho^{\pi_\mathrm{E}}(s,a)} + \log \frac{2 \rho^{\pi_\mathrm{E}}(s,a)}{\rho^{\pi_h}(s,a) + \rho^{\pi_\mathrm{E}}(s,a)} \right], \quad (9)
\end{aligned}
$$

where $\mathcal{D}^\star$ and $\mathcal{D}$ denote the expert demonstration and the replay buffer of π_h respectively. Then we can approximate the gradient of Eq. (9) with respect to h with

$$
\begin{aligned}
&\hat{\nabla}_h D_{\mathrm{JS}}(\rho^{\pi_h}, \rho^{\pi_\mathrm{E}}) \\
&\overset{(i)}{=} \frac{1}{2} \nabla_h \left(\frac{\rho^{\pi_h}(s_t, a_t)}{\rho^{\pi_\mathrm{E}}(s_t, a_t)} \log \frac{2 \rho^{\pi_h}(s_t, a_t)}{\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t)} + \log \frac{2 \rho^{\pi_\mathrm{E}}(s_t, a_t)}{\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t)} \right) \\
&\overset{(ii)}{=} \frac{1}{2} \left(\frac{d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t)}{d^{\pi_\mathrm{E}}(s_t) \pi_\mathrm{E}(a_t|s_t)} \log \frac{2 d^{\pi_h}(s_t) \pi_h(a_t|s_t)}{d^{\pi_h}(s_t) \pi_h(a_t|s_t) + d^{\pi_\mathrm{E}}(s_t) \pi_\mathrm{E}(a_t|s_t)} \right. \\
&\quad + \frac{\rho^{\pi_h}(s_t, a_t)}{\rho^{\pi_\mathrm{E}}(s_t, a_t)} \cdot \frac{\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t)}{2 \rho^{\pi_h}(s_t, a_t)} \\
&\quad \cdot \frac{2 d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t) \left(\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t) \right) - 2 \rho^{\pi_h}(s_t, a_t) d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t)}{\left(\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t) \right)^2} \\
&\quad \left. - \frac{\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t)}{2 \rho^{\pi_\mathrm{E}}(s_t, a_t)} \cdot \frac{2 \rho^{\pi_\mathrm{E}}(s_t, a_t) d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t)}{\left(\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t) \right)^2} \right) \\
&\overset{(iii)}{=} \frac{1}{2} \left(\frac{d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t)}{d^{\pi_\mathrm{E}}(s_t) \pi_\mathrm{E}(a_t|s_t)} \log \frac{2 d^{\pi_h}(s_t) \pi_h(a_t|s_t)}{d^{\pi_h}(s_t) \pi_h(a_t|s_t) + d^{\pi_\mathrm{E}}(s_t) \pi_\mathrm{E}(a_t|s_t)} \right. \\
&\quad \left. + \frac{d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t)}{\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t)} - \frac{d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t)}{\rho^{\pi_h}(s_t, a_t) + \rho^{\pi_\mathrm{E}}(s_t, a_t)} \right) \\
&\overset{(iv)}{=} \frac{d^{\pi_h}(s_t) \nabla_h \pi_h(a_t|s_t)}{2 d^{\pi_\mathrm{E}}(s_t) \pi_\mathrm{E}(a_t|s_t)} \log \frac{2 d^{\pi_h}(s_t) \pi_h(a_t|s_t)}{d^{\pi_h}(s_t) \pi_h(a_t|s_t) + d^{\pi_\mathrm{E}}(s_t) \pi_\mathrm{E}(a_t|s_t)}, \quad (10)
\end{aligned}
$$

where (ii) comes from Eq. (1). By the fact that

$$\nabla_h \pi_h(a|s) = \pi_h(a|s) \nabla_h \log \pi_h(a|s) = \pi_h(a|s) \kappa(s, \cdot) \boldsymbol{\Sigma}^{-1}(a - h(s)), \quad (11)$$

Eq. (10) can be shown that

$$\|\hat{\nabla}_h D_{\mathrm{JS}}(\rho^{\pi_h}, \rho^{\pi_{\mathrm{E}}})\|_2$$
$$= \left\|\frac{d^{\pi_h}(s_t)\pi_h(a_t|s_t)\kappa(s_t,\cdot)\boldsymbol{\Sigma}^{-1}(a_t - h(s_t))}{2d^{\pi_{\mathrm{E}}}(s_t)\pi_{\mathrm{E}}(a_t|s_t)}\log\frac{2d^{\pi_h}(s_t)\pi_h(a_t|s_t)}{d^{\pi_h}(s_t)\pi_h(a_t|s_t) + d^{\pi_{\mathrm{E}}}(s_t)\pi_{\mathrm{E}}(a_t|s_t)}\right\|_2.$$

Then it follows that $\|\hat{\nabla}_h D_{\mathrm{JS}}(\rho^{\pi_h}, \rho^{\pi_{\mathrm{E}}})\|_2 \to \infty$ with a probability of $\Pr(\|\boldsymbol{\Sigma}^{-1}(a_t - h(s_t))\|_2 \geq C$ for any $C > 0)$ as $\boldsymbol{\Sigma} \to \mathbf{0}$. $\qquad\square$

A.2 Proof of Corollary 1

Corollary 1 *Let $\pi_h(\cdot|s)$ be the Gaussian stochastic policy with mean $h(s)$ and covariance $\boldsymbol{\Sigma}$. When the discriminator achieves its regularity $\tilde{D}(s,a)$ in Eq. (7), i.e., $\tilde{D}(s,a) \in (0,1)$, the gradient estimator of the policy loss with respect to the policy's parameter h satisfies*

$$\left\|\hat{\nabla}_h\left(\mathbb{E}_{\mathcal{D}^\star}[\log\tilde{D}(s,a)] + \mathbb{E}_{\mathcal{D}}[\log(1 - \tilde{D}(s,a))]\right)\right\|_2 \to \infty$$

with a probability of $\Pr(\|\boldsymbol{\Sigma}^{-1}(a_t - h(s_t))\|_2 \geq C$ for any $C > 0)$ as $\boldsymbol{\Sigma} \to \mathbf{0}$, where $\mathcal{D}^\star$ and $\mathcal{D}$ denote the expert demonstration and the replay buffer of π_h respectively,

$$\hat{\nabla}_h\left(\mathbb{E}_{\mathcal{D}^\star}[\log(\tilde{D}(s,a))] + \mathbb{E}_{\mathcal{D}}[\log(1 - \tilde{D}(s,a))]\right)$$
$$= \frac{d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{d^{\pi_{\mathrm{E}}}(s_t)\pi_{\mathrm{E}}(a_t|s_t)}\log\frac{(1 - \epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)}{(1 + \epsilon_1)d^{\pi_{\mathrm{E}}}(s_t)\pi_{\mathrm{E}}(a_t|s_t) + (1 - \epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)}$$
$$+ \frac{(\epsilon_1 + \epsilon_2)d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{(1 + \epsilon_1)\rho^{\pi_{\mathrm{E}}}(s_t,a_t) + (1 - \epsilon_2)\rho^{\pi_h}(s_t,a_t)},$$

and $\nabla_h\pi_h(a|s) = \pi_h(a|s)\kappa(s,\cdot)\boldsymbol{\Sigma}^{-1}(a - h(s))$.

Proof. Referring to the proof strategy of Theorem 1, the learned state-action distribution can be transferred to the expert demonstration distribution by importance sampling. Thus when the discriminator achieves its regularity $\tilde{D}(s,a)$, we can write the policy objective from the optimization problem in Eq. (2) as

$$\mathbb{E}_{(s,a)\sim\mathcal{D}^\star}[\log(\tilde{D}(s,a))] + \mathbb{E}_{(s,a)\sim\mathcal{D}}[\log(1 - \tilde{D}(s,a))]$$
$$= \mathbb{E}_{(s,a)\sim\mathcal{D}^\star}\left[\log\frac{(1 + \epsilon_1)\rho^{\pi_{\mathrm{E}}}(s_t,a_t)}{(1 + \epsilon_1)\rho^{\pi_{\mathrm{E}}}(s_t,a_t) + (1 - \epsilon_2)\rho^\pi(s_t,a_t)}\right]$$
$$+ \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\log\frac{(1 - \epsilon_2)\rho^\pi(s_t,a_t)}{(1 + \epsilon_1)\rho^{\pi_{\mathrm{E}}}(s_t,a_t) + (1 - \epsilon_2)\rho^\pi(s_t,a_t)}\right]$$
$$= \mathbb{E}_{(s,a)\sim\mathcal{D}^\star}\left[\log\frac{(1 + \epsilon_1)\rho^{\pi_{\mathrm{E}}}(s,a)}{(1 + \epsilon_1)\rho^{\pi_{\mathrm{E}}}(s,a) + (1 - \epsilon_2)\rho^{\pi_h}(s,a)}\right.$$
$$+ \left.\frac{\rho^{\pi_h}(s,a)}{\rho^{\pi_{\mathrm{E}}}(s,a)}\log\frac{(1 - \epsilon_2)\rho^{\pi_h}(s,a)}{(1 + \epsilon_1)\rho^{\pi_{\mathrm{E}}}(s,a) + (1 - \epsilon_2)\rho^{\pi_h}(s,a)}\right]. \tag{12}$$

2

Then the gradient of Eq. (12) can be approximated with

$$
\hat{\nabla}_h \left(\mathbb{E}_{(s,a)\sim\mathcal{D}^\star}[\log(\tilde{D}(s,a))] + \mathbb{E}_{\mathcal{D}}[\log(1 - \tilde{D}(s,a))] \right)
$$

$$
= \nabla_h \Bigg(\log \frac{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s,a)}{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s,a) + (1-\epsilon_2)\rho^{\pi_h}(s,a)}
$$

$$
+ \frac{\rho^{\pi_h}(s,a)}{\rho^{\pi_\mathrm{E}}(s,a)} \log \frac{(1-\epsilon_2)\rho^{\pi_h}(s,a)}{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s,a) + (1-\epsilon_2)\rho^{\pi_h}(s,a)} \Bigg)
$$

$$
= -\frac{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s,a) + (1-\epsilon_2)\rho^{\pi_h}(s,a)}{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s_t,a_t)} \cdot \frac{(1+\epsilon_1)(1-\epsilon_2)\rho^{\pi_\mathrm{E}}(s_t,a_t)d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{\left((1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s,a) + (1-\epsilon_2)\rho^{\pi_h}(s,a)\right)^2}
$$

$$
+ \frac{d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{d^{\pi_\mathrm{E}}(s_t)\pi_\mathrm{E}(a_t|s_t)} \log \frac{(1-\epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)}{(1+\epsilon_1)d^{\pi_\mathrm{E}}(s_t)\pi_\mathrm{E}(a_t|s_t) + (1-\epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)}
$$

$$
+ \frac{\rho^{\pi_h}(s_t,a_t)}{\rho^{\pi_\mathrm{E}}(s_t,a_t)} \cdot \frac{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s_t,a_t) + (1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)}{(1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)}
$$

$$
\cdot \frac{(1-\epsilon_2)(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s_t,a_t)d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{\left((1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s_t,a_t) + (1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)\right)^2}
$$

$$
= \frac{d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{d^{\pi_\mathrm{E}}(s_t)\pi_\mathrm{E}(a_t|s_t)} \log \frac{(1-\epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)}{(1+\epsilon_1)d^{\pi_\mathrm{E}}(s_t)\pi_\mathrm{E}(a_t|s_t) + (1-\epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)}
$$

$$
- \frac{(1-\epsilon_2)d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s_t,a_t) + (1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)} + \frac{(1+\epsilon_1)d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s_t,a_t) + (1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)}
$$

$$
= \frac{d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{d^{\pi_\mathrm{E}}(s_t)\pi_\mathrm{E}(a_t|s_t)} \log \frac{(1-\epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)}{(1+\epsilon_1)d^{\pi_\mathrm{E}}(s_t)\pi_\mathrm{E}(a_t|s_t) + (1-\epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)}
$$

$$
+ \frac{(\epsilon_1 + \epsilon_2)d^{\pi_h}(s_t)\nabla_h\pi_h(a_t|s_t)}{(1+\epsilon_1)\rho^{\pi_\mathrm{E}}(s_t,a_t) + (1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)}. \tag{13}
$$

Plugging Eq. (11) into Eq. (13), when $\|\mathbf{\Sigma}^{-1}(a_t - h(s_t))\|_2 \geq C$ for any $C > 0$, we have

$$
\left\| \hat{\nabla}_h \left(\mathbb{E}_{(s,a)\sim\mathcal{D}^\star}[\log(\tilde{D}(s,a))] + \mathbb{E}_{\mathcal{D}}[\log(1 - \tilde{D}(s,a))] \right) \right\|_2 \to \infty.
$$

$\square$

B Analysis of Relieving Exploding Gradients

B.1 Proof of Proposition 1

Proposition 1 *When the discriminator achieves its optimum $D^*(s,a)$ in Eq. (6), we have*
$$
D^*(s_t, a_t) \approx 1 \Leftrightarrow h(s_t) \text{ mismatches } a_t.
$$

Proof. The optimal discriminator of (s_t, a_t) can be denoted by
$$
D^*(s_t, a_t) = \frac{\rho^{\pi_\mathrm{E}}(s_t, a_t)}{\rho^{\pi_\mathrm{E}}(s_t, a_t) + \rho^{\pi_h}(s_t, a_t)}.
$$
We can derive that the necessary and sufficient condition of $D^*(s_t, a_t) \approx 1$ is that $\rho^{\pi_h}(s_t, a_t) \approx 0$, i.e., $(s_t, h(s_t))$ mismatches (s_t, a_t). $\square$

B.2 Proof of Proposition 2

Proposition 2 *When the discriminator achieves its optimum $D^*(s,a)$ in Eq. (6), we have $\beta \geq \alpha$.*

Proof. When $r_i(s_t, a_t) = c$, $i = 1, 2$, we obtain $\log \beta - \log(1-\beta) = -\log(1-\alpha)$, which is followed by

$$
\beta - \alpha = \frac{\alpha^2 - 2\alpha + 1}{2 - \alpha} \geq 0.
$$

$\square$

C Representative Reward Functions

Table 2: Representative Reward Functions.

Reward	Function Shape
$r_1(s,a)$	$-\log(1 - D(s,a))$
$r_2(s,a)$	$\log D(s,a) - \log(1 - D(s,a))$
$r_3(s,a)$	$\log D(s,a)$
$r_4(s,a)$	$D(s,a)$
$r_5(s,a)$	$e^{D(s,a)}$
$r_6(s,a)$	$-1/D(s,a)$
$r_7(s,a)$	$D(s,a)^2$
$r_8(s,a)$	$\sqrt{D(s,a)}$