

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 Σ . 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 $\|\hat{\nabla}_h D_{\text{JS}}(\rho^{\pi_h}, \rho^{\pi_E})\|_2 \rightarrow \infty$ with a probability of $\Pr(\|\Sigma^{-1}(a_t - h(s_t))\|_2 \geq C \text{ for any } C > 0) \text{ as } \Sigma \rightarrow \mathbf{0}$, where

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

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

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

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

where \mathcal{D}^* 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

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

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

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

Eq. (10) can be shown that

$$\begin{aligned} & \|\hat{\nabla}_h D_{\text{JS}}(\rho^{\pi_h}, \rho^{\pi_E})\|_2 \\ &= \left\| \frac{d^{\pi_h}(s_t) \pi_h(a_t|s_t) \kappa(s_t, \cdot) \Sigma^{-1}(a_t - h(s_t))}{2d^{\pi_E}(s_t) \pi_E(a_t|s_t)} \log \frac{2d^{\pi_h}(s_t) \pi_h(a_t|s_t)}{d^{\pi_h}(s_t) \pi_h(a_t|s_t) + d^{\pi_E}(s_t) \pi_E(a_t|s_t)} \right\|_2. \end{aligned}$$

Then it follows that $\|\hat{\nabla}_h D_{\text{JS}}(\rho^{\pi_h}, \rho^{\pi_E})\|_2 \rightarrow \infty$ with a probability of $\Pr(\|\Sigma^{-1}(a_t - h(s_t))\|_2 \geq C \text{ for any } C > 0) \text{ as } \Sigma \rightarrow \mathbf{0}$. \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 Σ . When the discriminator achieves its regularity in Eq. (??), 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

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

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

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

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

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

$$\begin{aligned} & \mathbb{E}_{(s,a) \sim \mathcal{D}^*} [\log(\tilde{D}(s, a))] + \mathbb{E}_{(s,a) \sim \mathcal{D}} [\log(1 - \tilde{D}(s, a))] \\ &= \mathbb{E}_{(s,a) \sim \mathcal{D}^*} \left[\log \frac{(1 + \epsilon_1) \rho^{\pi_E}(s_t, a_t)}{(1 + \epsilon_1) \rho^{\pi_E}(s_t, a_t) + (1 - \epsilon_2) \rho^{\pi}(s_t, a_t)} \right] \\ &+ \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[\log \frac{(1 - \epsilon_2) \rho^{\pi}(s_t, a_t)}{(1 + \epsilon_1) \rho^{\pi_E}(s_t, a_t) + (1 - \epsilon_2) \rho^{\pi}(s_t, a_t)} \right] \\ &= \mathbb{E}_{(s,a) \sim \mathcal{D}^*} \left[\log \frac{(1 + \epsilon_1) \rho^{\pi_E}(s, a)}{(1 + \epsilon_1) \rho^{\pi_E}(s, a) + (1 - \epsilon_2) \rho^{\pi_h}(s, a)} \right. \\ &\quad \left. + \frac{\rho^{\pi_h}(s, a)}{\rho^{\pi_E}(s, a)} \log \frac{(1 - \epsilon_2) \rho^{\pi_h}(s, a)}{(1 + \epsilon_1) \rho^{\pi_E}(s, a) + (1 - \epsilon_2) \rho^{\pi_h}(s, a)} \right]. \quad (12) \end{aligned}$$

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

$$\begin{aligned}
& \hat{\nabla}_h \left(\mathbb{E}_{(s,a) \sim \mathcal{D}^*} [\log(\tilde{D}(s,a))] + \mathbb{E}_{\mathcal{D}} [\log(1 - \tilde{D}(s,a))] \right) \\
&= \nabla_h \left(\log \frac{(1 + \epsilon_1)\rho^{\pi_E}(s,a)}{(1 + \epsilon_1)\rho^{\pi_E}(s,a) + (1 - \epsilon_2)\rho^{\pi_h}(s,a)} \right. \\
&\quad \left. + \frac{\rho^{\pi_h}(s,a)}{\rho^{\pi_E}(s,a)} \log \frac{(1 - \epsilon_2)\rho^{\pi_h}(s,a)}{(1 + \epsilon_1)\rho^{\pi_E}(s,a) + (1 - \epsilon_2)\rho^{\pi_h}(s,a)} \right) \\
&= - \frac{(1 + \epsilon_1)\rho^{\pi_E}(s,a) + (1 - \epsilon_2)\rho^{\pi_h}(s,a)}{(1 + \epsilon_1)\rho^{\pi_E}(s_t, a_t)} \cdot \frac{(1 + \epsilon_1)(1 - \epsilon_2)\rho^{\pi_E}(s_t, a_t)d^{\pi_h}(s_t)\nabla_h \pi_h(a_t|s_t)}{\left((1 + \epsilon_1)\rho^{\pi_E}(s_t, a_t) + (1 - \epsilon_2)\rho^{\pi_h}(s_t, a_t)\right)^2} \\
&\quad + \frac{d^{\pi_h}(s_t)\nabla_h \pi_h(a_t|s_t)}{d^{\pi_E}(s_t)\pi_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_E}(s_t)\pi_E(a_t|s_t) + (1 - \epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)} \\
&\quad + \frac{\rho^{\pi_h}(s_t, a_t)}{\rho^{\pi_E}(s_t, a_t)} \cdot \frac{(1 + \epsilon_1)\rho^{\pi_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)} \\
&\quad \cdot \frac{(1 - \epsilon_2)(1 + \epsilon_1)\rho^{\pi_E}(s_t, a_t)d^{\pi_h}(s_t)\nabla_h \pi_h(a_t|s_t)}{\left((1 + \epsilon_1)\rho^{\pi_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_E}(s_t)\pi_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_E}(s_t)\pi_E(a_t|s_t) + (1 - \epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)} \\
&\quad - \frac{(1 - \epsilon_2)d^{\pi_h}(s_t)\nabla_h \pi_h(a_t|s_t)}{(1 + \epsilon_1)\rho^{\pi_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_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_E}(s_t)\pi_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_E}(s_t)\pi_E(a_t|s_t) + (1 - \epsilon_2)d^{\pi_h}(s_t)\pi_h(a_t|s_t)} \\
&\quad + \frac{(\epsilon_1 + \epsilon_2)d^{\pi_h}(s_t)\nabla_h \pi_h(a_t|s_t)}{(1 + \epsilon_1)\rho^{\pi_E}(s_t, a_t) + (1 - \epsilon_2)\rho^{\pi_h}(s_t, a_t)}. \tag{13}
\end{aligned}$$

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

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

□

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

$$D^*(s_t, a_t) = \frac{\rho^{\pi_E}(s_t, a_t)}{\rho^{\pi_E}(s_t, a_t) + \rho^{\pi_h}(s_t, a_t)}.$$

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) . □

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

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

□

C Representative Reward Functions

Table 2: Representative Reward Functions.

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