Exploring Gradient Explosion in Generative Adversarial Imitation Learning: A Probabilistic Perspective

Wanying Wang[1,*], **Yichen Zhu**[2,*], **Yirui Zhou**[1], **Chaomin Shen**[3], **Jian Tang**[2], **Zhiyuan Xu**[2], **Yaxin Peng**[1,†], **Yangchun Zhang**[1,†]

[1]Department of Mathematics, College of Sciences, Shanghai University
[2]Midea Group
[3]School of Computer Science, East China Normal University
{wywang, zyr050798, yaxin.peng, zycstatis}@shu.edu.cn, cmshen@cs.ecnu.edu.cn,
{zhuyc25, xuzy70, tangjian22}@midea.com

Abstract

Generative Adversarial Imitation Learning (GAIL) stands as a cornerstone approach in imitation learning. This paper investigates the gradient explosion in two types of GAIL: GAIL with deterministic policy (DE-GAIL) and GAIL with stochastic policy (ST-GAIL). We begin with the observation that the training can be highly unstable for DE-GAIL at the beginning of the training phase and end up divergence. Conversely, the ST-GAIL training trajectory remains consistent, reliably converging. To shed light on these disparities, we provide an explanation from a theoretical perspective. By establishing a probabilistic lower bound for GAIL, we demonstrate that gradient explosion is an inevitable outcome for DE-GAIL due to occasionally large expert-imitator policy disparity, whereas ST-GAIL does not have the issue with it. To substantiate our assertion, we illustrate how modifications in the reward function can mitigate the gradient explosion challenge. Finally, we propose CREDO, a simple yet effective strategy that clips the reward function during the training phase, allowing the GAIL to enjoy high data efficiency and stable trainability.

Introduction

Imitation learning trains a policy directly from expert demonstrations without reward signals (Ng, Russell et al. 2000; Syed and Schapire 2007; Ho and Ermon 2016). It has been broadly studied under the twin umbrellas of behavioral cloning (Pomerleau 1991) and inverse reinforcement learning (IRL) (Ziebart et al. 2008). Generative adversarial imitation learning (GAIL) (Ho and Ermon 2016), established by the policy training of trust region policy optimization (Schulman et al. 2015), plugs the inspiration of generative adversarial networks (Goodfellow et al. 2014) into the maximum entropy IRL. The discriminator in GAIL aims to distinguish whether a state-action pair comes from the expert demonstration or is generated by the agent. Meanwhile, the learned policy generates interaction data to confuse the discriminator. GAIL is promising for many real-world scenarios where designing reward functions to learn the optimal control policies requires significant effort. It has made remarkable achievements in physical-world tasks, i.e., robot manipulation (Jabri 2021), mobile robot navigating (Tai et al. 2018), commodities search (Shi et al. 2019) and endovascular catheterization (Chi et al. 2020).

The GAIL can be bifurcated into two genres: stochastic policy algorithms and deterministic policy algorithms, namely DE-GAIL (Kostrikov et al. 2019; Zuo et al. 2020) and ST-GAIL (Ho and Ermon 2016; Zhou et al. 2022). The ST-GAIL with stochastic policy guarantees global convergence in high-dimensional environments, outperforming traditional Inverse Reinforcement Learning (IRL) methods (Ng, Russell et al. 2000; Ziebart et al. 2008; Boularias, Kober, and Peters 2011). Nevertheless, its application in real-world scenarios is limited due to low sample efficiency and excessive training times (Zuo et al. 2020). On the contrary, DE-GAIL has become a preferred approach due to its exceptional data efficiency. Typically, it outpaces GAIL with stochastic policy by a factor of more than ten, significantly accelerating the learning process.

However, while the DE-GAIL is much more data-efficient than ST-GAIL, it is not flawless: we observe a significant likelihood of generating near-zero rewards from the very beginning of the training stage. To elucidate this, we carried out experiments on three environments in Mujuco with multiple DE-GAIL and ST-GAIL algorithms. Among 11 experiments under uniform training settings, we observed DE-GAIL method frequently failed during the training phase. The average divergence rate is over 36% (Details can be found in Fig. 2).

Why does DE-GAIL have such a high probability of diverging? To shed light on this question, we prove a probabilistic lower bound that describes the gradient explosion in DE-GAIL. Our proof is built upon the policy disparity between expert demonstration and imitative action. In short, if the agent fails to mimic the expert's action, the expert has a large reward; then, the gradient could explode during training. We verify our conclusion by showing a simple manipulation of the reward function, i.e., switching to

[*] Equal contributions.
† Corresponding authors: Yangchun Zhang, Yaxin Peng.

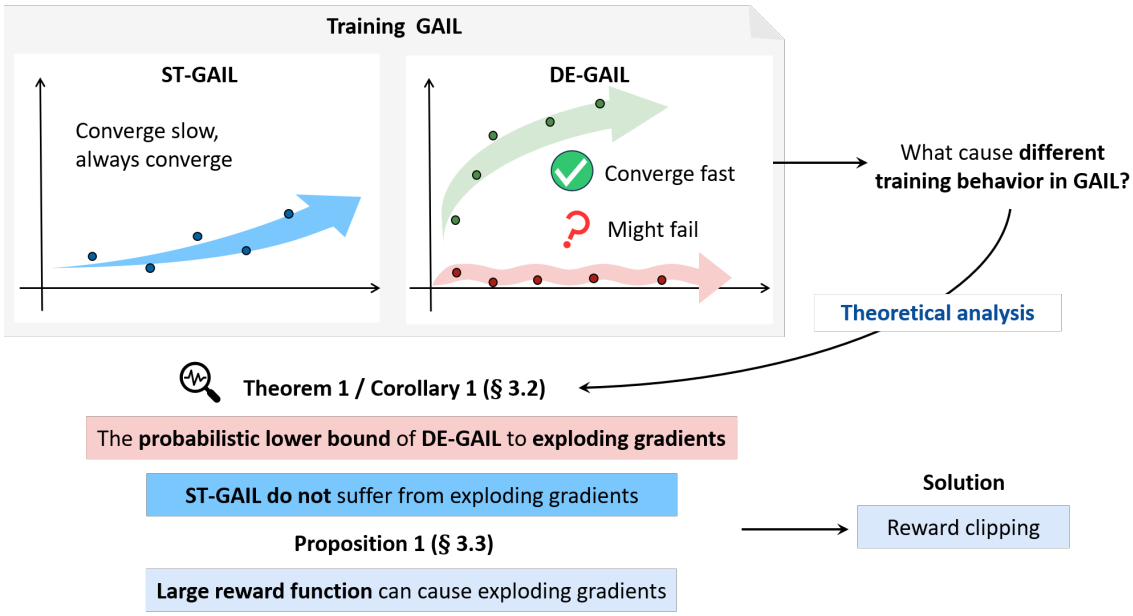

Figure 1: The overview of our analysis. Theorem 1 and Corollary 1 develop the probabilistic lower bound of DE-GAIL to quantify the gradient explosion. Proposition 1 connects the gradient explosion with the reward function. We further present a reward clipping technique to relieve the gradient explosion issue in DE-GAIL.

adversarial inverse reinforcement learning (AIRL) (Fu, Luo, and Levine 2018), which can alleviate the gradient explosion issue in GAIL. Nevertheless, since the DE-GAIL is much more data-efficient than ST-GAIL, we seek to resolve this issue by proposing CREDO. This reward clipping technique is both empirically effective and theoretically sound.

Our overall framework for analyzing gradient explosion in GAIL is shown in Fig. 1. In summary, our contributions are the follows:

- We conduct a comprehensive empirical study to show the fact that DE-GAIL is training unstable yet converges fast. On the other hand, the ST-GAIL is data inefficient yet ensures convergence.
- We develop a series of theoretical proofs to support our observation and conclude that reward function is the cause of the gradient explosion in DE-GAIL.
- We present a simple technique called CREDO which clips the reward function during training to relieve the gradient explosion problem in DE-GAIL.

Evidence of Gradient Explosion in GAIL

In this section, we perform a comprehensive study to examine the gradient explosion issue in GAIL. We reproduced three environments in Mujoco (Todorov, Erez, and Tassa 2012), Hopper-v2, HalfCheetah-v2, and Walker2d-v2, following the setup in two-stage stochastic gradient (TSSG) (Zhou et al. 2022). The expert trajectories were generated via the soft actor-critic (SAC) agent (Haarnoja et al. 2018). The expert demonstration has one million data points with a standard deviation of 0.01. We repeat our experiments 11 times across all environments, maintaining a consistent training setting, except for the number of random seeds.

Regarding the network architecture, we employ two-layer networks designed to approximate the kernel function (Arora et al. 2019) to train GAIL. The reward function is defined as $r(s,a) = -\log(1-D(s,a))$, which is referred to as Probability Logarithm Reward (PLR). Our evaluations encompass three variants of the DE-GAIL and two ST-GAIL methods. The DE-GAIL methods include deep deterministic policy gradient (Lillicrap et al. 2015) (DDPG-GAIL), twin delayed deep deterministic policy gradient (Fujimoto, Hoof, and Meger 2018) (TD3-GAIL), and softmax deep double deterministic policy gradients (Pan, Cai, and Huang 2020) (SD3-GAIL). The first two are recognized and widely adopted DE-GAIL algorithms, whereas SD3-GAIL represents a more recent and refined approach. For ST-GAIL, we utilize proximal policy optimization (Schulman et al. 2017) (PPO)-GAIL (Chen et al. 2020) and TSSG (Zhou et al. 2022), a method that integrate SAC into GAIL.

As shown in Fig. 2, we observed that all three DE-GAIL algorithms could potentially fail during training, irrespective of how advanced they are. This pattern is consistently seen across all three tasks. It illustrates that the training of DE-GAIL methods can be particularly unstable at the initial stages, reaching a point from which recovery becomes impossible as training progresses. This behavior sharply contrasts with the training curve of successful experiments, which, on the other hand, converge quickly and yield high returns, thus highlighting the solid data-efficiency characteristic of the DE-GAIL approach. Turning our attention to ST-GAIL, we noted that even though its convergence speed lags behind DE-GAIL, all experiments exhibited consistent

and successful convergence.

In summary, our observations reveal the following insights:

- The initial phase of DE-GAIL training can be remarkably unstable. However, once convergence is attained, it is often swift and results in higher return values, potentially leading to an elevated success rate.
- Conversely, ST-GAIL exhibits stability during the initial training process, yet its convergence is approximately ten times slower and it tends to achieve lower return values compared to advanced DE-GAIL algorithms.

These phenomena prompt us to delve into the root causes underpinning their differences. In the subsequent section, we offer a theoretical framework to support and deepen our understanding of these observations.

Gradients Explosion in GAIL: A Probability View

In this section, we will first introduce the necessary background information and notation for the forthcoming proof. Then, we provide a detailed analysis from a theoretical standpoint along with empirical evidence to unveil the mystery of gradient explosion in GAIL.

Background and Annotation

Markov Decision Process A discounted Markov Decision Process (MDP) in the conventional Reinforcement Learning (RL) context is defined by a quintuple $(\mathcal{S}, \mathcal{A}, r, p_M, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ represent the finite state space and action space respectively. The reward function, $r(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, denotes the reward obtained from executing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. The transition distribution is represented by $p_M(s'|s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$, and γ is the discount factor.

A stochastic policy, denoted as $\pi(a|s)$, can be characterized as a probability function mapping a state $s \in \mathcal{S}$ to a distribution of action $a \in \mathcal{A}$, expressed formally as $\mathcal{S} \times \mathcal{A} \to [0,1]$. In contrast, a deterministic policy, $\pi(s)$, is defined as a direct mapping from a state $s \in \mathcal{S}$ to a corresponding action $a \in \mathcal{A}$, formally written as $\mathcal{S} \to \mathcal{A}$.

The primary objective of Reinforcement Learning (RL) is to maximize the expected reward-to-go, represented as $\eta(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r\left(s_t, a_t\right) | s_0, a_0\right]$. Induced by a policy π, *the discounted stationary state distribution* is defined as $d^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}(s_t = s; \pi)$. Similarly, *the discounted stationary state-action distribution* is given by $\rho^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}(s_t = s, a_t = a; \pi)$. This distribution measures the cumulative "frequency" with which a state-action pair is visited under the policy π.

Let $\mathrm{P} \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ denote the transition matrix where $\mathrm{P}_{sa,s'} = p_M(s' \mid s,a)$, $\boldsymbol{\pi} \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|\times 1}$ denote the policy matrix where $(\boldsymbol{\pi})_{sa} = \pi(a|s)$, and $\boldsymbol{\pi}_{s_i} \in \mathbb{R}_+^{|\mathcal{A}|\times 1}$ the policy for the state s_i. Define the expanded matrix of $\boldsymbol{\pi}$ as

$$\Pi = \begin{bmatrix} \boldsymbol{\pi}_{s_1}^\top & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\pi}_{s_{|\mathcal{S}|}}^\top \end{bmatrix} \in \mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}. \tag{1}$$

We use $\boldsymbol{\pi}_h$ to denote a parameterized policy matrix with parameters h, then the policy Jacobian matrix $\mathrm{H}_h \in \mathbb{R}^{1\times|\mathcal{S}||\mathcal{A}|}$ is $(\mathrm{H}_h)_{sa} = \nabla_h (\boldsymbol{\pi}_h)_{sa} = \nabla_h \pi_h(a|s)$. The state-action distribution matrix and the state distribution matrix are $\boldsymbol{\rho}^h \in \mathbb{R}_+^{|\mathcal{S}\|\mathcal{A}|\times 1}$ and $\mathbf{d}^h \in \mathbb{R}_+^{|\mathcal{S}|\times 1}$, respectively, where $(\boldsymbol{\rho}^h)_{sa} = \rho^{\pi_h}(s,a)$ and $(\boldsymbol{d}^h)_s = d^{\pi_h}(s)$. More precisely, $\mathbf{d}^h = T\boldsymbol{\rho}^h$, here T is the marginalization matrix

$$T = \begin{bmatrix} \mathbf{1}_{|\mathcal{A}|}^\top & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{1}_{|\mathcal{A}|}^\top \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}, \tag{2}$$

where $\mathbf{1}$ is the vector of all ones and the subscript represents its dimensionality.

Generative Adversarial Imitation Learning (GAIL) The description of GAIL can be found in Supplementary Material D. Here, we define the discriminator $D(s,a)$, the imitative policy π, and the expert policy π_{E}. Given a state, the discriminator quantifies the distributional disparity between the expert's and imitative policies. This disparity can be interpreted as a reward for the agent. Consequently, the optimization problem for GAIL can be formulated as follows:

$$\min_\pi \max_{D\in(0,1)^{\mathcal{S}\times\mathcal{A}}} \mathbb{E}_{(s,a)\sim\rho^{\pi_{\mathrm{E}}}}[\log(D(s,a))] + \mathbb{E}_{(s,a)\sim\rho^{\pi}}[\log(1-D(s,a))], \tag{3}$$

where the policy π mimics the expert policy via the reward function $r(s,a) = -\log(1-D(s,a))$. When the discriminator reaches its optimum,

$$D^*(s,a) = \rho^{\pi_{\mathrm{E}}}(s,a)/(\rho^{\pi_{\mathrm{E}}}(s,a) + \rho^{\pi}(s,a)), \tag{4}$$

the optimization objective for the learned policy is formalized as minimizing the discrepancy in the state-action distribution between the imitated policy and the expert policy. This discrepancy is quantified using the Jensen-Shannon (JS) divergence,

$$\min_\pi D_{\mathrm{JS}}(\rho^\pi(s,a), \rho^{\pi_{\mathrm{E}}}(s,a)) := \frac{1}{2} D_{\mathrm{KL}}\left(\rho^\pi, \frac{\rho^\pi + \rho^{\pi_{\mathrm{E}}}}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(\rho^{\pi_{\mathrm{E}}}, \frac{\rho^\pi + \rho^{\pi_{\mathrm{E}}}}{2}\right). \tag{5}$$

Exploding Gradients in GAIL

We employ multivariate Gaussian policy to approximate deterministic policy (Paternain et al. 2020; Lever and Stafford 2015), where the learned policy π_h is defined as follow:

$$\pi_h(a|s) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp \frac{-(a-h(s))^\top \boldsymbol{\Sigma}^{-1}(a-h(s))}{2}, \tag{6}$$

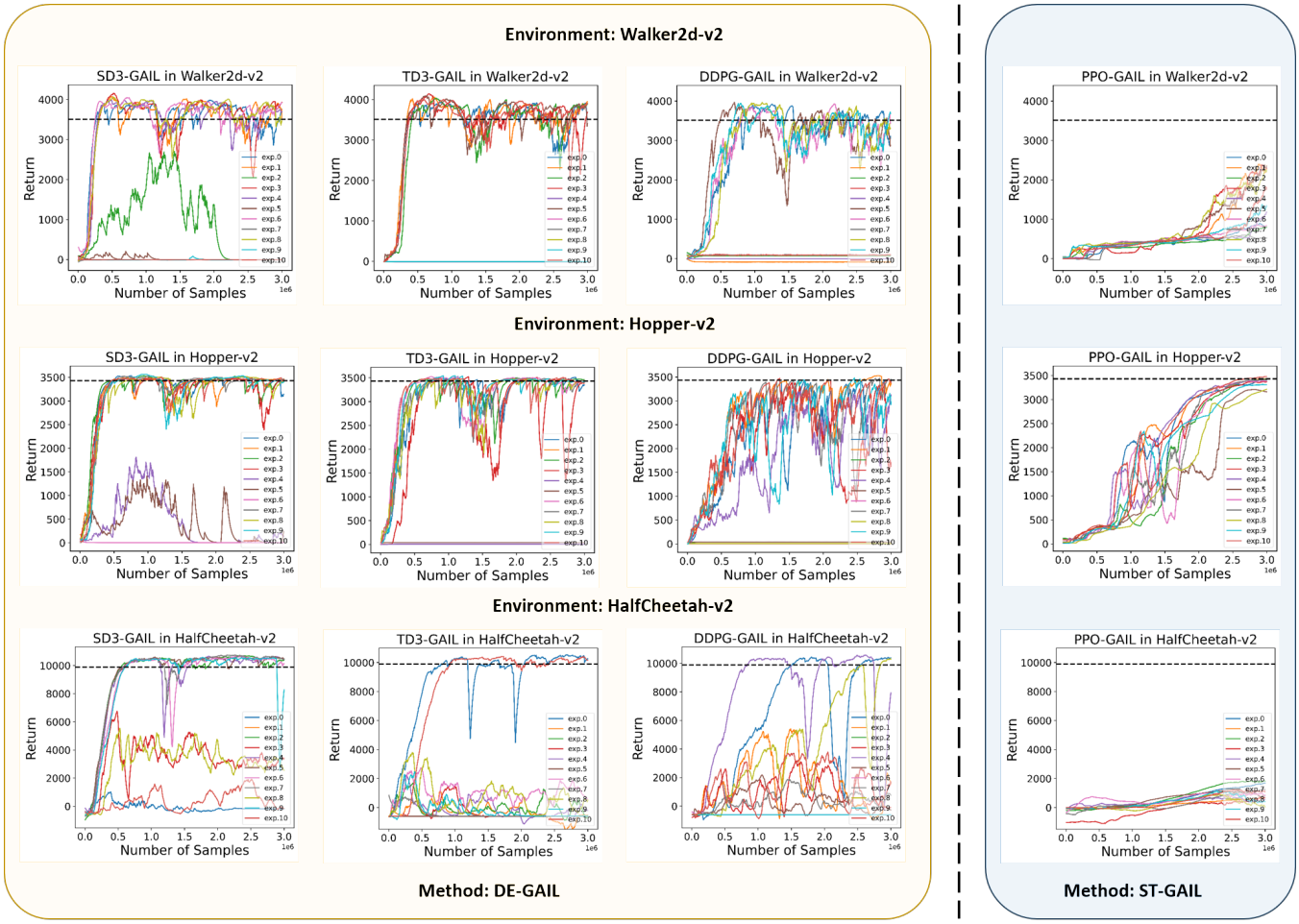

Figure 2: We conduct 11 experiments spread across three environments, with three DE-GAIL and ST-GAIL methods. We observe a clear tendency for DE-GAIL algorithms to struggle to reach convergence during the training phase in multiple experiments. In comparison, the training procedure utilizing the ST-GAIL method showcased significantly higher stability.

The above equation is parameterized by deterministic functions $h : \mathcal{S} \to \mathcal{A}$ and covariance matrix $\mathbf{\Sigma}$. The function $h(\cdot)$ is an element of reproducing kernel Hilbert space $\mathcal{H}_\kappa$, $h(\cdot) = \sum_i \kappa(s_i, \cdot)a_i \in \mathcal{H}_\kappa$, where $\kappa(s_i, s_j)$ is the kernel function, $s_i \in \mathcal{S}$ and $a_i \in \mathcal{A}$. Note that $\pi_h(a|s)$ can be regarded as an approximation to the Dirac's impulse via covariance matrix approaching zero, for instance,

$$\lim_{\mathbf{\Sigma}\to\mathbf{0}} \pi_h(a|s) = \delta(a - h(s)). \tag{7}$$

Eq. (7) means that when the covariance $\mathbf{\Sigma} \to \mathbf{0}$, the stochastic policy $\pi_h(a|s)$ approaches the deterministic policy $h(s)$. Therefore, we can substitute π with π_h and rewrite the optimization problem of GAIL under π_h is

$$\min_{\pi_h} \max_{D} \mathbb{E}_{(s,a)\sim\rho^{\pi_\mathrm{E}}}[\log(D(s,a))] + \mathbb{E}_{(s,a)\sim\rho^{\pi_h}}[\log(1 - D(s,a))], \tag{8}$$

the optimal discriminator is

$$D^*(s,a) = \rho^{\pi_\mathrm{E}}(s,a)/(\rho^{\pi_\mathrm{E}}(s,a) + \rho^{\pi_h}(s,a)), \tag{9}$$

and the policy optimization objective is

$$\min_{\pi_h} D_{\mathrm{JS}}(\rho^{\pi_h}(s,a), \rho^{\pi_\mathrm{E}}(s,a)). \tag{10}$$

Before jumping into our main result, we need the following definition.

Definition 1 (Expert-Imitator Policy Disparity) *Given the state s_t at time t, a_t and $h(s_t)$ are the actions induced by the expert policy and the imitated policy, respectively. If $\|h(s_t) - a_t\|_2 \geq C\|\mathbf{\Sigma}\|_2$ for any $C > 0$, we say that there exist policy disparity between the expert and the imitator. Otherwise, the $(s_t, h(s_t))$ perfectly matches the (s_t, a_t).*

Here, we utilize an event

$$\Xi = \{(s_t, h(s_t)) : \|h(s_t) - a_t\|_2 \geq C\|\mathbf{\Sigma}\|_2 \text{ for any } C > 0\} \tag{11}$$

to characterize the expert-imitator policy disparity. For convenience, we will use policy disparity to denote such behavior. Now we present the following theorem on the probability of exploding gradients in DE-GAIL.

Theorem 1 *Let $\pi_h(\cdot|s)$ be the Gaussian stochastic policy with mean $h(s)$ and covariance $\mathbf{\Sigma}$. When the discriminator achieves its optimum $D^*(s,a)$ in Eq. (9), 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 at least $\Pr(\|\mathbf{\Sigma}^{-1}(a_t - h(s_t))\|_2 \geq C$ for any $C > 0)$ as $\mathbf{\Sigma} \to \mathbf{0}$, where*

$$\hat{\nabla}_h D_{\mathrm{JS}}(\rho^{\pi_h}, \rho^{\pi_{\mathrm{E}}}) = \frac{H_h \Delta\left(T^\top \mathbf{d}^h\right)(\mathrm{I} - \gamma \mathrm{P}\Pi_h)^{-1}\mathbf{e}_{s_t,a_t}}{2\rho^{\pi_{\mathrm{E}}}(s_t,a_t)} \cdot \log \frac{2\rho^{\pi_h}(s_t,a_t)}{\rho^{\pi_h}(s_t,a_t) + \rho^{\pi_{\mathrm{E}}}(s_t,a_t)},$$

and $(\mathrm{H}_h)_{sa} = \pi_h(a|s)\kappa(s,\cdot)\mathbf{\Sigma}^{-1}(a - h(s))$, $\Delta(\cdot)$ maps a vector to a diagonal matrix with its elements on the main diagonal, $\mathbf{e}_{s_t,a_t} = [0, \cdots, 1_{s_t,a_t}, \cdots, 0]^\top \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times 1}$.

Proof See Supplementary Material A.1 in our arxiv version (Wang et al. 2023).

The result establishes the probability of exploding gradients in DE-GAIL. Due to the compatibility of norms, we have

$$\begin{aligned} &\Pr(\|\mathbf{\Sigma}^{-1}(a_t - h(s_t))\|_2 \geq C \text{ for any } C > 0) \\ &\geq \Pr(\|a_t - h(s_t)\|_2 \geq C\|\mathbf{\Sigma}\|_2 \text{ for any } C > 0) \\ &= \Pr(\Xi). \end{aligned} \tag{12}$$

Therefore, the probability of policy disparity $\Pr(\Xi)$ constitutes a probabilistic lower bound of exploding gradients in DE-GAIL. Note that $\Pr(\Xi)$ is nontrivial as $\mathbf{\Sigma} \to \mathbf{0}$.

Remark 1 *Theorem 1 implies that when the discriminator achieves its optimal state, DE-GAIL will suffer from exploding gradients with the probabilistic lower bound $\Pr(\Xi) > 0$.*

In contrast, for a Gaussian stochastic policy (fixed $\mathbf{\Sigma}$), we have that $\|\hat{\nabla}_h D_{\mathrm{JS}}(\rho^{\pi_h}, \rho^{\pi_{\mathrm{E}}})\|_2$ is bounded referring to the proof strategy of Theorem 1. Thus, when the discriminator achieves its optimal state, the Gaussian stochastic policy in GAIL will not suffer from exploding gradients.

Theorem 1 indicates that when the discriminator attains its optimal state, the policy loss can encounter a gradient explosion issue. However, this represents an idealized scenario. In practical applications, the discriminator seldom reaches this optimum. Hence, we adapt our findings to a broader context using a "non-optimum" discriminator derived from data. Here, we name such "non-optimum" discriminators as imperfect discriminators and define them by the following:

$$\tilde{D}(s_t,a_t) = \frac{(1+\epsilon_1)\rho^{\pi_{\mathrm{E}}}(s_t,a_t)}{(1+\epsilon_1)\rho^{\pi_{\mathrm{E}}}(s_t,a_t) + (1-\epsilon_2)\rho^{\pi}(s_t,a_t)}, \quad \text{where} \begin{cases} \epsilon_1 > -1, \epsilon_2 < 1 \\ \epsilon_1 < -1, \epsilon_2 > 1 \end{cases} \tag{13}$$

The explanation of the imperfect discriminator $\tilde{D}(s_t,a_t)$ and its properties are as follows:

- ϵ_1 and ϵ_2 can be regarded as fluctuations in the optimal discriminator.
- The imperfect discriminator generalize the fixed $D^*(s_t,a_t)$ to a ranges within $(0,1)$ stemmed from Eq. (3).
- $\tilde{D}(s_t,a_t)$ degenerates to 0 when $\epsilon_1 = -1$ and degenerates to 1 when $\epsilon_2 = 1$.
- $\tilde{D}(s_t,a_t)$ reaches its optimum when ϵ_1 and ϵ_2 are 0.

We next state the exploding gradients on the imperfect discriminator $\tilde{D}(s_t,a_t)$.

Corollary 1 *Let $\pi_h(\cdot|s)$ be the Gaussian stochastic policy with mean $h(s)$ and covariance $\mathbf{\Sigma}$. When the discriminator is in the format of Eq. (13), i.e., $\tilde{D}(s,a) \in (0,1)$, the gradient estimator of the policy loss concerning 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 at least

$$\Pr(\|\mathbf{\Sigma}^{-1}(a_t - h(s_t))\|_2 \geq C \text{ for any } C > 0)$$

as $\mathbf{\Sigma} \to \mathbf{0}$, where $\mathcal{D}^\star$ and $\mathcal{D}$ denote the expert demonstration and the replay buffer of π_h respectively,

$$\begin{aligned} &\hat{\nabla}_h \left(\mathbb{E}_{\mathcal{D}^\star}[\log(\tilde{D}(s,a))] + \mathbb{E}_{\mathcal{D}}[\log(1 - \tilde{D}(s,a))]\right) \\ &= \frac{H_h \Delta\left(T^\top \mathbf{d}^h\right)(\mathrm{I} - \gamma \mathrm{P}\Pi_h)^{-1}\mathbf{e}_{s_t,a_t}}{\rho^{\pi_{\mathrm{E}}}(s_t,a_t)} \cdot \log \frac{(1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)}{(1+\epsilon_1)\rho^{\pi_{\mathrm{E}}}(s_t,a_t) + (1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)} \\ &\quad + \frac{(\epsilon_1+\epsilon_2)\nabla_h \rho^{\pi_h}(s_t,a_t)}{(1+\epsilon_1)\rho^{\pi_{\mathrm{E}}}(s_t,a_t) + (1-\epsilon_2)\rho^{\pi_h}(s_t,a_t)}, \end{aligned}$$

and $(\mathrm{H}_h)_{sa} = \pi_h(a|s)\kappa(s,\cdot)\mathbf{\Sigma}^{-1}(a - h(s))$, $\Delta(\cdot)$ maps a vector to a diagonal matrix with its elements on the main diagonal, $\mathbf{e}_{s_t,a_t} = [0, \cdots, 1_{s_t,a_t}, \cdots, 0]^\top \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times 1}$.

Proof See Supplementary Material A.2.

Analogous to Theorem 1, Corollary 1 suggests that when the discriminator adopts the perturbation form given by Eq. (13), DE-GAIL is susceptible to gradient explosions. Conversely, ST-GAIL remains unaffected by such explosions as long as the discriminator values lie within the interval $(0,1)$.

Remark 2 *In the implementation of deterministic policy, it requires exploration that adds noise Σ' to the output action. Concurrently, as the deterministic policy h progressively updated, the covariance of Gaussian stochastic policy $\Sigma \to 0$. Notably, the stochastic factors are taken into consideration by using*

$$\begin{aligned} \Xi_1 = \{&(s_t, h(s_t)) : \|h(s_t) + \mathcal{N}(0,\Sigma') - a_t\|_2 \geq C\|\Sigma\|_2 \\ &\text{for any } C > 0\} \end{aligned} \tag{14}$$

to characterize policy disparity in practice. During $\Sigma \to 0$, the proofs of Theorem 1 and Corollary 1 only depend on $\Sigma \to 0$, Theorem 1 and Corollary 1 both hold for Ξ and Ξ_1 regardless of Σ'.

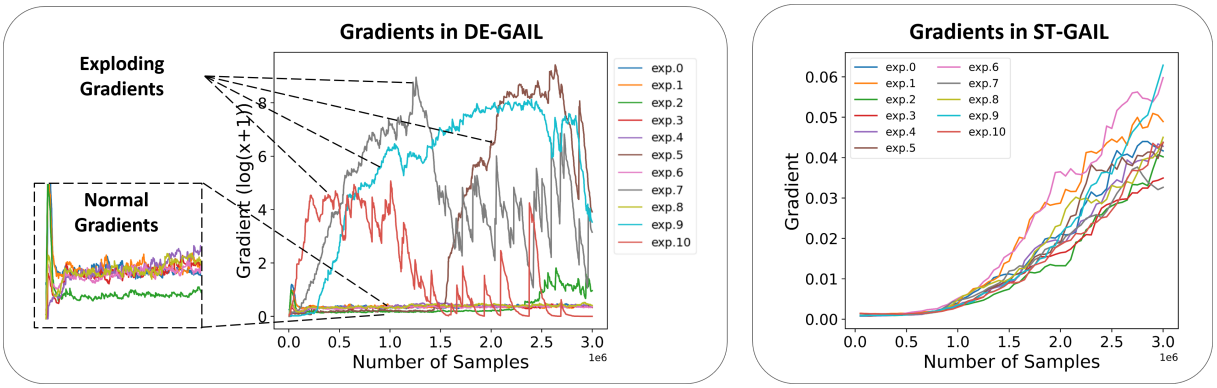

Figure 3: The absolute gradients of SD3-GAIL and PPO-GAIL policy networks in Walker2d-v2. Each algorithm is repeatedly run 11 times. We observe that four experiments in SD3-GAIL display exploding gradients. Note that we use $log(x+1)$ to rescale the y-axis in the left figure.

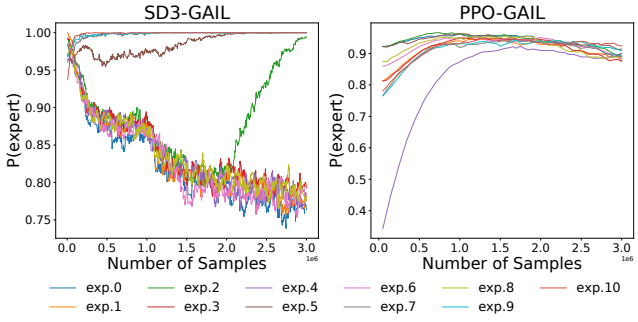

Figure 4: The y-axis denotes the probability of the imitator's policy being classified as the expert's policy, recorded as P(expert). We observe that DE-GAIL can have P(expert) $\approx$ 1, which leads to gradient explosion during training.

In brief, our Theorem 1 and Corollary 1 hold in the practical setting.

Empirical evidence. Our experimental results support our theoretical analysis. Specifically, in Fig 3, we record the absolute gradient in the training phase. Notably, four experiments in SD3-GAIL have gradient explosions. The total percent of failure cases is over 36%.

In Fig.4, we document the probability P(expert) of the expert's demonstration being classified to expert policy by the discriminator. It's crucial to understand that when P(expert) ≈ 1, there's potential for gradient explosion. Our observations indicate that the P(expert) of ST-GAIL never attains a value of one. On the contrary, in several experiments in DE-GAIL, P(expert) gravitates exceedingly close to one.

Meanwhile, degenerated discriminator behaviors in SD3-GAIL (left of Fig. 4) are consistent with the cases where returns are zero ($r \to 0$) in Fig. 2. In comparison, the gradients of PPO-GAIL maintain their training stability.

Relieving Exploding Gradients with Reward Modification

The AIRL (Kostrikov et al. 2019) modified the reward function $r_2(s_t, a_t) = \log(D(s_t, a_t)) - \log(1 - D(s_t, a_t))$ into DE-GAIL. This reward function can empirically mitigate the training instability of DE-GAIL. Here we give a theory that supports their experiments and corroborate our analysis.

The reward function of AIRL is defined as a combination reward function (CR) (Wang and Li 2021). For convenience, DE-GAIL with PLR and CR are called PLR-DE-GAIL and CR-DE-GAIL, respectively. We study whether CR-DE-GAIL can have a lower probability of gradient explosion compared to PLR-DE-GAIL. Note that Theorem 1 shows that the policy disparity causes the gradient explosion. Unlike the discriminator, which can be defined within a finite interval of $(0, 1)$, the expert-imitator policy disparity is vaguely defined. Therefore, the following proposition provides a concrete understanding of policy disparity.

Proposition 1 *When the discriminator achieves its optimum $D^*(s, a)$ in Eq. (9), we have*

$$D^*(s_t, a_t) \approx 1 \Leftrightarrow$$
$$h(s_t) \textit{ unequaled } a_t \textit{ under the event } \Xi \textit{ in Eq. (11).}$$

Proof See Supplementary Material B.1.

Proposition 1 indicates that exploding gradients can either depend on the distance between the discriminator's value and value 1, or the degree of $r_i(s_t, a_t)$, where $i = 1, 2$ that goes to infinity. This is due to the monotonicity of both $r_1(s_t, a_t)$ and $r_2(s_t, a_t)$. As $D(s_t, a_t) \approx 1$, we obtain

$$r_1(s_t, a_t) \approx \infty \text{ and } r_2(s_t, a_t) \approx \infty.$$

Intuitively, we want to prevent exploding gradients. As such, we make the constraints that $r_i(s_t, a_t) \leq c$, $i = 1, 2$, for some appropriate constant c. In contrast, the outliers of the discriminator can also be characterized as $r_i(s_t, a_t) > c$ for $i = 1, 2$, which represents the situation of gradient explosion. We define such a state as follows:

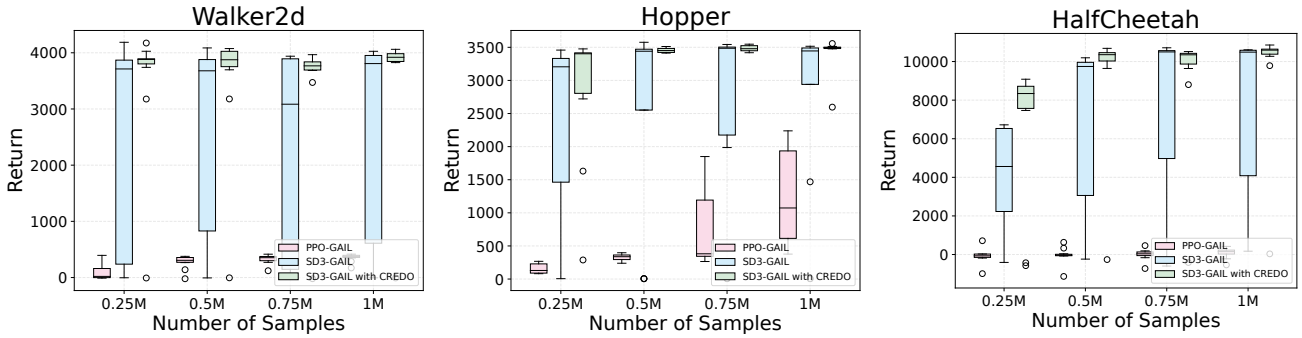

Figure 5: Box-plot of PPO-GAIL, SD3-GAIL, and SD3-GAIL with CREDO in three environments. The clipping technique significantly enhances the training stability of SD3-GAIL, resulting in a high successful rate in terms of convergence.

Definition 2 *When the discriminator achieves its optimum $D^*(s,a)$ in Eq. (9), the outliers of the discriminator are defined in $[\alpha, 1]$ such that $r_1(s_t, a_t) \geq c$. Similarly, under the same upper bound c, the outliers of the discriminator are defined in $[\beta, 1]$ for $r_2(s_t, a_t)$.*

We note that the training process will suffer from exploding gradients when the discriminator falls into the $[\alpha, 1]$ range. The next proposition describes how to relieve the gradient explosion in CR-DE-GAIL.

Proposition 2 *When the discriminator achieves its optimum $D^*(s,a)$ in Eq. (9), we have $\beta \geq \alpha$.*

Proof See Supplementary Material B.2.

Proposition 2 reveals that the discriminator in CR-DE-GAIL exhibits a smaller interval of outliers than that in PLR-DE-GAIL, which decreases the probability of gradient explosion.

Clipping Reward of Discriminator Outlier

The analysis from the preceding section demystified the phenomenon of gradient explosion in GAIL. This examination uncovered that the intrinsic limitations of DE-GAIL can occasionally lead to inevitable divergence. Nonetheless, when DE-GAIL does converge, its data efficiency notably surpasses that of ST-GAIL. As a result, we are driven to develop a robust approach that identifies and alleviates the gradient explosion issue in DE-GAIL while preserving its admirable training efficiency.

Building on the insights from Proposition 1, we highlighted the pivotal role of the reward function in the gradient explosion issue. Inspired by this understanding, we aim to mitigate the likelihood of exploding gradients in DE-GAIL. To achieve this, we apply a clipping mechanism to the rewards that align with the discriminator's outliers in DE-GAIL, enforcing constraints such that $r(s,a) \leq c$ for some appropriate constant c to reduce the outliers. We term this clipping strategy CREDO, an acronym for Clipping REward of Discriminator Outlier.

It's worth noting that our method is versatile and can be applied across all DE-GAIL algorithms. We selected SD3-GAIL as our baseline and integrated CREDO owing to its standout performance. To maintain consistency and to showcase the resilience of our approach, we performed 11 independent experiments for each methodology, ensuring uniform training settings for all, except for the random seed. We maintained identical hyperparameters across all experiments. In particular, we set the update frequency at 64 and established a threshold $c = 5$. Additional training details, including pseudocode and hyperparameter settings, and additional experimental results in Humanoid-v2 and Ant-v2 are shown in Supplementary Material F.

The experimental results are illustrated in Figure 5. We employ box plots to showcase the returns at various training phases, gauged by the number of samples. Notably, the blue box plot, symbolizing the vanilla SD3-GAIL, displays outlier points with zero return values. These points correspond to experiments that failed to converge, due to gradient explosion. In contrast, the integration of CREDO into SD3-GAIL (as represented by the green box plot) significantly reduces the box plot's span and reduces the outliers. Such outcomes are consistently observed across all environments. When juxtaposed with the ST-GAIL approach, our CREDO method boasts data efficiency that is an order of magnitude higher than that of PPO-GAIL.

Conclusion

This paper delves into the issue of gradient explosion in Generative Adversarial Imitation Learning (GAIL). The journey begins by examining the existence of divergence in GAIL. Among the two types of GAIL, namely deterministic policy and stochastic policy, we observe that the former has a non-negligible probability of divergence, whereas the latter exhibits successful convergence. To gain an in-depth comprehension of this phenomenon, we analyze it from a theoretical standpoint, explicitly considering the structure of the reward function. Subsequently, we introduce an example featuring a modified reward function, demonstrating that such alterations can effectively mitigate the gradient explosion issue. To further alleviate this problem in DE-GAIL, we propose a novel technique, the efficacy of which is substantiated through experimental evidence. Overall, our analysis of exploding gradients fosters a new understanding of GAIL in terms of training schemes.

Acknowledgments

This work was supported in part by the Key Technologies Research and Development Program under Grant 2021ZD0140300, the National Natural Science Foundation of China under Grants 12301351, 62225308 and 11771276, the Shanghai Sailing Program, Shanghai Association for Science and Technology under Grant 21YF1413500, "Shuguang Program" under Grant 20SG40 supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission, and the Science and Technology Innovation Action Plan of Shanghai under Grant 22511105400.

References

Arora, S.; Du, S. S.; Hu, W.; Li, Z.; Salakhutdinov, R. R.; and Wang, R. 2019. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, volume 32, 8139–8148.

Boularias, A.; Kober, J.; and Peters, J. 2011. Relative entropy inverse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 182–189.

Chen, M.; Wang, Y.; Liu, T.; Yang, Z.; Li, X.; Wang, Z.; and Zhao, T. 2020. On computation and generalization of generative adversarial imitation learning. In *International Conference on Learning Representations (ICLR). Preprint retrieved from arXiv:2001.02792*.

Chi, W.; Dagnino, G.; Kwok, T. M.; Nguyen, A.; Kundrat, D.; Abdelaziz, M. E.; Riga, C.; Bicknell, C.; and Yang, G.-Z. 2020. Collaborative robot-assisted endovascular catheterization with generative adversarial imitation learning. In *IEEE International Conference on Robotics and Automation*, 2414–2420.

Fu, J.; Luo, K.; and Levine, S. 2018. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR). Preprint retrieved from arXiv:1710.11248*.

Fujimoto, S.; Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, 1587–1596.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2672—-2680.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 1861–1870.

Ho, J.; and Ermon, S. 2016. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, 4565–4573.

Jabri, M. K. 2021. Robot manipulation learning using generative adversarial imitation learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*, 4893–4894.

Kostrikov, I.; Agrawal, K. K.; Dwibedi, D.; Levine, S.; and Tompson, J. 2019. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *International Conference on Learning Representations (ICLR). Preprint retrieved from arXiv:1809.02925*.

Lever, G.; and Stafford, R. 2015. Modelling policies in MDPs in reproducing kernel hilbert space. In *Artificial Intelligence and Statistics*, 590–598.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Ng, A. Y.; Russell, S. J.; et al. 2000. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 663–670.

Pan, L.; Cai, Q.; and Huang, L. 2020. Softmax deep double deterministic policy gradients. In *Advances in Neural Information Processing Systems*, volume 33, 11767–11777.

Paternain, S.; Bazerque, J. A.; Small, A.; and Ribeiro, A. 2020. Stochastic policy gradient ascent in reproducing kernel hilbert spaces. *IEEE Transactions on Automatic Control*, 66(8): 3429–3444.

Pomerleau, D. A. 1991. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1): 88–97.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International Conference on Machine Learning*, 1889–1897.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shi, J.-C.; Yu, Y.; Da, Q.; Chen, S.-Y.; and Zeng, A.-X. 2019. Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4902–4909.

Syed, U.; and Schapire, R. E. 2007. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems*, volume 20, 1449–1456.

Tai, L.; Zhang, J.; Liu, M.; and Burgard, W. 2018. Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In *IEEE International Conference on Robotics and Automation*, 1111–1117.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, 5026–5033. IEEE.

Wang, W.; Zhu, Y.; Zhou, Y.; Shen, C.; Tang, J.; Xu, Z.; Peng, Y.; and Zhang, Y. 2023. Exploring Gradient Explosion in Generative Adversarial Imitation Learning: A Probabilistic Perspective. arXiv:2312.11214.

Wang, Y.; and Li, X. 2021. Reward function shape exploration in adversarial imitation learning: An empirical study. In *IEEE International Conference on Artificial Intelligence and Computer Applications*, 52–57.

Zhou, Y.; Zhang, Y.; Liu, X.; Wang, W.; Che, Z.; Xu, Z.; Tang, J.; and Peng, Y. 2022. Generalization and computation for policy classes of generative adversarial imitation learning. In *International Conference on Parallel Problem Solving from Nature*, 385–399.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; Dey, A. K.; et al. 2008. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 8, 1433–1438.

Zuo, G.; Chen, K.; Lu, J.; and Huang, X. 2020. Deterministic generative adversarial imitation learning. *Neurocomputing*, 388: 60–69.