When Will Generative Adversarial Imitation Learning Algorithms Attain Global Convergence

Ziwei Guan
Ohio State University

Tengyu Xu
Ohio State University

Yingbin Liang
Ohio State University

Abstract

Generative adversarial imitation learning (GAIL) is a popular inverse reinforcement learning approach for jointly optimizing policy and reward from expert trajectories. A primary question about GAIL is whether applying a certain policy gradient algorithm to GAIL attains a global minimizer (i.e., yields the expert policy), for which existing understanding is very limited. Such global convergence has been shown only for the linear (or linear-type) MDP and linear (or linearizable) reward. In this paper, we study GAIL under general MDP and for nonlinear reward function classes (as long as the objective function is strongly concave with respect to the reward parameter). We characterize the global convergence with a sublinear rate for a broad range of commonly used policy gradient algorithms, all of which are implemented in an alternating manner with stochastic gradient ascent for reward update, including projected policy gradient (PPG)-GAIL, Frank-Wolfe policy gradient (FWPG)-GAIL, trust region policy optimization (TRPO)-GAIL and natural policy gradient (NPG)-GAIL. This is the first systematic theoretical study of GAIL for global convergence.

1 Introduction

In reinforcement learning (RL), the reward function generally plays an important role to guide the design of policy optimization to attain the best long-term accumulative reward. However, a reward function may not be known in many situations, and imitation learning (Osa et al., 2018) aims to find a desirable policy in such cases, which produces behaviors as close as possible to expert demonstrations. Two popular classes of approaches for imitation learning have been developed. The first approach is behavioral cloning (BC) (Pomerleau, 1991), which directly provides a mapping strategy from the state space to the action space based on supervised learning to match expert demonstrations. The BC method often suffers from high sample complexity due to covariate shift (Ross and Bagnell, 2010; Ross et al., 2011) for achieving the desired performance, which is mitigated by improved algorithms such as DAgger (Ross et al., 2011) and Dart (Laskey et al., 2017) that require further interaction with the expert's demonstration. The second approach is the so-called inverse reinforcement learning (IRL) (Ng and Russell, 2000), which attempts to recover the unknown reward function based on the expert's trajectories, and then find an optimal policy by using such a reward function.

A popular IRL method has been developed in Finn et al. (2016); Ho and Ermon (2016); Fu et al. (2018), which leverages the connection of IRL to the training of generative adversarial networks (GANs) (Goodfellow et al., 2014). In particular, the generative adversarial imitation learning (GAIL) framework (Ho and Ermon, 2016) formulates a min-max optimization problem as in the GAN training. The maximization is over the reward function (which serves as a discriminator) to best distinguish between the trajectories generated by the expert and the learner, and the minimization is then over the learner's policy (which serves as a generator) to best match the expert's trajectories. Since the policy optimization in GAIL is nonconvex, its joint optimization with reward function in GAIL in general can be guaranteed to converge only to a stationary point. Such a type of result was recently established in Chen et al. (2020), which studied GAIL under general MDP model and reward function class, and showed that the gradient-decent and gradient-ascent algorithm converges to a stationary point (not necessarily the global minimum).

Proceedings of the 24^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

More recently, it has been shown that some popular policy gradient algorithms (Agarwal et al., 2019; Shani et al., 2020; Xu et al., 2020a) can converge to a globally optimal policy under certain policy parameterizations. Then a natural question to ask is whether such global convergence continues to hold in GAIL when these algorithms are further implemented in an alternating fashion with the reward optimization in GAIL. The global convergence does not necessarily hold in general, because the policy optimization is still over a nonconvex objective function, which can induce complicated and undesirable geometries jointly with the reward optimization as a min-max problem in GAIL. Thus, existing exploration on this topic in Cai et al. (2019); Zhang et al. (2020), which established global convergence for GAIL, requires restrictive conditions: (1) linear (but possibly infinite dimensional) MDP and (2) linear reward function or linearizable reward function such as overparameterized ReLU neural networks.

This paper aims to substantially expand the aforementioned global convergence results as follows.

- We allow general MDP models, not necessarily linear MDP. We study nonlinear reward functions as long as the resulting objective function is strongly concave with respect to the reward parameter. This is a much bigger class than linear reward, and is satisfied easily by incorporating a strongly concave regularizer which has been commonly used in GAIL practice.
- In addition to the projected gradient and NPG that have been studied in Cai et al. (2019); Zhang et al. (2020) for GAIL, we also study Frank-Wolfe policy gradient, which is easier to implement than projected policy gradient, and TRPO which is widely adopted in GAIL in practice.
- Existing convergence characterization for GAIL assumed that the samples are either identical and independently distributed (i.i.d.) as in Chen et al. (2020); Zhang et al. (2020) or follows the LQR dynamics as in Cai et al. (2019), whereas here we assume that samples follow a general Markovian distribution.

1.1 Main Contributions

In this paper, we establish the first *global* convergence guarantee for GAIL under the general MDP model and the nonlinear reward function class (as long as the objective function is strongly concave with respect to the reward parameter). We provide the convergence rate for three major types of algorithms, all of which alternate between gradient ascent (for reward update) and policy gradient descent (for policy update), respectively being (a) projected policy gradient (PPG)-GAIL and Frank-Wolfe policy gradient (FWPG)-GAIL (with direct policy parameterization); (b) trust region policy optimization (TRPO)-GAIL (with direct policy parameterization); and (c) natural policy gradient (NPG)-GAIL (with general non-linear policy parameterization). We show that all these alternating algorithms converge to the *global* minimum with a sublinear rate. We summarize our results on the convergence performance of the GAIL algorithms in Table 1. Comparing among these algorithms indicates that TRPO-GAIL with regularized MDP achieves the best convergence rate, and TRPO-GAIL with regularized and unregularized MDP outperform the other algorithms in terms of the overall sample complexity.

Technically, the global convergence guarantee for GAIL does not follow from the existing min-max optimization theory. In fact, the GAIL problem here falls into nonconvex-strongly-concave min-max optimization framework, for which existing optimization theory does not provide the global convergence in general. Thus, our establishment of global convergence for GAIL develops several new properties specially for GAIL. Furthermore, in contrast to conventional min-max optimization, which is under i.i.d. sampling by certain static distribution, GAIL is under Markovian sampling by time-varying distributions due to the policy update. Thus, the convergence analysis for GAIL is more challenging than that for min-max optimization.

1.2 Related Work

Due to the significant growth of studies in imitation learning, this section focuses only on those studies that are highly relevant to the theoretical analysis of the convergence for GAIL algorithms.

Theory for IRL via adversarial training: The idea of generative adversarial training (Goodfellow et al., 2014) has motivated a popular approach for IRL problems (Finn et al., 2016; Ho and Ermon, 2016; Fu et al., 2018). Among these studies, GAIL (Ho and Ermon, 2016) formulated a min-max problem for jointly optimizing the reward and policy, where reward and policy serve analogous roles as the discriminator and the generator in GANs. Naturally, such an approach has been explored via the divergence minimization perspective in Ke et al. (2019); Ghasemipour et al. (2019), and by leveraging GAN training (Nowozin et al., 2016). Moreover, the generalization performance and sample complexity have been studied for the setting where the expert's demonstrations include only the states but no actions.

Most relevant to our study is the recent studies (Cai et al., 2019; Chen et al., 2020; Zhang et al., 2020) on the convergence rate for the algorithms developed

Table 1: Comparison among GAIL algorithms studied in this paper

Algorithms	Convergence rate	Total Complexity[1,2]
PPG-GAIL	$\mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{T}}\right)$	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^4}\right)$
FWPG-GAIL	$\mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{T}}\right)$	$\tilde{\mathcal{O}}(\frac{1}{\epsilon^4})$
TRPO-GAIL (unregularized)	$\mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{T}}\right)$	$\tilde{\mathcal{O}}(\frac{1}{\epsilon^3})$
TRPO-GAIL (regularized)	$\tilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3 T}\right)$	$\tilde{\mathcal{O}}(\frac{1}{\epsilon^2})$
NPG-GAIL	$\mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{T}}\right)$	$\tilde{\mathcal{O}}(\frac{1}{\epsilon^4})$

[1] Total complexity refers to the total number of samples needed to achieve an ϵ-accurate globally optimal point.

[2] $\tilde{\mathcal{O}}(\cdot)$ does not include the logarithmic terms.

for GAIL. Among these studies, Chen et al. (2020) studied GAIL under the general MDP model and the reward function class, and showed that the gradient-decent and gradient-ascent algorithm converges to a stationary point (not necessarily the global minimum). Cai et al. (2019); Zhang et al. (2020) provided the global convergence result. More specifically, Cai et al. (2019) studied GAIL under linear quadratic regulator (LQR) dynamics and the linear reward function class, and showed that the alternating gradient algorithm converges to the unique saddle point. Zhang et al. (2020) studied GAIL under a type of linear but infinite dimensional MDP and with overparameterized neural networks for parameterizing the policy and reward function, and showed that the alternating algorithm between gradient-ascent (for reward update) and NPG (for policy update) converges to the neighborhood of a global optimal point, where the representation power of neural networks determines the convergence error. Our study here establishes global convergence for GAIL for general MDP and the nonlinear reward function class.

Difference from conventional min-max problems: Although the GAIL framework is formulated as a min-max optimization problem, the stochastic algorithms that we use for solving such a problem have the following major differences from the conventional min-max optimization problem. First, since these algorithms continuously update the policy, the samples that are used for iterations are sampled by time-varying policies; whereas the conventional min-max problem typically has a fixed sampling distribution. Second, since the samples are obtained following an MDP process, the samples are distributed with correlation rather than in the i.i.d. manner as in the conventional optimization. These two differences cause the convergence analysis to be more complicated for GAIL than the conventional min-max problem. Furthermore, the min-max problem that we encounter here for GAIL is nonconvex-strongly-concave, for which the conventional min-max optimization (Nouiehed et al., 2019; Lin et al., 2020) has been shown to converge only to a stationary point, whereas this paper exploits further properties in GAIL and establishes the global convergence guarantee.

Connection to policy gradient algorithms: In the GAIL framework, the policy optimization is jointly performed with the reward optimization via a min-max optimization. Thus, the variation of the reward function during the algorithm execution continuously change the objective function for the policy optimization. Hence, even if the policy gradient algorithms (running for a fixed objective function) converge globally, for example, PPG (Agarwal et al., 2019), NPG (Agarwal et al., 2019), and TRPO (Shani et al., 2020), the global convergence is generally not guaranteed if these algorithms are executed in an alternating fashion with reward iterations. Two special cases have been shown to retain such global convergence, namely, LQR model shown in Cai et al. (2019) and overparameterized neural networks for a linear type MDP (Zhang et al., 2020). This paper significantly expands such a set of cases by establishing the global convergence guarantee for more general MDP and reward class and a broader range of algorithms.

2 Problem Formulation and Preliminaries

2.1 Markov Decision Process

The imitation learning framework that we study is based on the Markov decision process (MDP) denoted by $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma)$. We assume that both the state space $\mathcal{S} \subset \mathbb{R}^d$ and the action space $\mathcal{A}$ are finite, and use $s \in \mathcal{S}$ and $a \in \mathcal{A}$ to denote a state and an action, respectively. A policy π describes the probability to take an action $a \in \mathcal{A}$ at each state $s \in \mathcal{S}$ in terms of the conditional probability $\pi(a|s)$. Then the system moves to a next state $s' \in \mathcal{S}$ governed by the probability transition kernel $\mathsf{P}(s'|s,a)$, and receives a reward $r_t = r(s,a)$, which is assumed to be bounded by $R_{\max}$.

Suppose the initial state takes a distribution ζ. For a given policy π and a reward function r, we define the average value function as:

$$\begin{aligned} V(\pi, r) &= \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \big| s_0 \sim \zeta, a_t \sim \pi(a_t|s_t), \dots \\ &\qquad\qquad s_{t+1} \sim \mathsf{P}(s_{t+1}|s_t, a_t)\big] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim \nu_\pi(s,a)}[r(s,a)], \end{aligned}$$

where $\gamma \in (0,1)$ is a discount factor and

$$\nu_\pi(s,a) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a)$$

is the state-action visitation distribution. It has been shown in (Konda, 2002) that $\nu_\pi(s,a)$ is the stationary distribution of the Markov chain with the transition kernel

$$\tilde{\mathsf{P}}(\cdot|s,a) := (1-\gamma)\zeta(\cdot) + \gamma \mathsf{P}(\cdot|s,a)$$

and policy π if the Markov chain is ergodic. Thus $\tilde{\mathsf{P}}$ is used in sampling for estimating the value function.

2.2 Generative Adversarial Imitation Learning (GAIL)

For imitation learning, in which the reward function is not known, GAIL (Ho and Ermon, 2016) is a framework to jointly learn the reward function and optimize the policy. We parameterize the reward function by $\alpha \in \Lambda \subset \mathbb{R}^q$, which takes the form $r_\alpha(s,a)$ at the state-action pair (s,a). We assume that Λ is a bounded closed set, i.e., $\|\alpha_1 - \alpha_2\|_2 \le C_\alpha$, $\forall \alpha_1, \alpha_2 \in \Lambda$.

We let π_E represent the expert policy, and let the learner's policy be parameterized by $\theta \in \Theta$ and be denoted as π_θ. In this paper, we consider two types of parameterization for the learner's policy. The first is the direct parameterization, where $\pi_\theta(a|s) = \theta_{s,a}$ and $\theta \in \Theta_p := \{\theta : \theta_{s,a} \ge 0, \sum_{a\in\mathcal{A}} \theta_{s,a} = 1, \ \text{ for all } \ s \in \mathcal{S}, a \in \mathcal{A}\}$. The second is the general nonlinear policy class, which satisfies certain smoothness conditions as given in Assumption 5.

The GAIL framework is formulated as the following min-max optimization problem.

$$\min_{\theta\in\Theta} \max_{\alpha\in\Lambda} F(\theta, \alpha) := V(\pi_E, r_\alpha) - V(\pi_\theta, r_\alpha) - \psi(\alpha), \tag{1}$$

where the objective function is given by the discrepancy of the accumulated rewards between the expert's and learner's policies, regularized by a function $\psi(\alpha)$ of the reward parameter. Thus, the maximization in eq. (1) aims to find the reward function that best distinguishes between the expert's and the learner's policies and the minimization aims to find the learner's policy that matches the expert's policy as close as possible. Such a formulation is analogous to the GANs, with the reward serving as a discriminator and the policy serving as a generator.

Algorithm 1 Nested-loop GAIL framework

Input: Outer loop length T, inner loop length K, stepsize η, β
for $t = 0, 1, ..., T-1$ **do**
  Randomly pick $\alpha_0^t \in \Lambda$
  for $k = 0, 1, ..., K-1$ **do**
    Query a length-B trajectory $(s_i^E, a_i^E) \sim \tilde{\mathsf{P}}^{\pi_E}$ and a length-B mini-batch $(s_i^\theta, a_i^\theta) \sim \tilde{\mathsf{P}}^{\pi_\theta}$[1]

$$\begin{aligned} &\widehat{\nabla}_\alpha F(\theta, \alpha) \\ &\quad = \frac{1}{(1-\gamma)B} \sum_{i=0}^{B-1} \nabla_\alpha r_\alpha(s_i^E, a_i^E) \\ &\qquad - \frac{1}{(1-\gamma)B} \sum_{i=0}^{B-1} \nabla_\alpha r_\alpha(s_i^\theta, a_i^\theta) - \nabla_\alpha \psi(\alpha) \\ &\alpha_{k+1}^t = P_\Lambda\left(\alpha_k^t + \beta \widehat{\nabla}_\alpha F(\theta_t, \alpha_k^t)\right) \end{aligned}$$

  end for
  $\alpha_t = \alpha_K^t$
  $\theta_{t+1} =$ **Options**: **PPG** in eq. (4); **FWPG** in eq. (5); **TRPO** in eq. (7); **NPG** in eq. (8)
end for

In this paper, we study four GAIL algorithms, all of which follow the nested-loop framework described in Algorithm 1. Namely, at each time step t (associated with one outer loop), there is an entire inner loop updates of the reward parameter α_t to a certain accuracy and one update step of the policy parameter θ_t. Specifically, α_t is updated by the stochastic projected gradient ascent given by

$$\alpha_t^{k+1} = P_\Lambda\left(\alpha_t^k + \beta \widehat{\nabla}_\alpha F(\theta_t, \alpha_t^k)\right),$$

where the gradient estimator $\widehat{\nabla}_\alpha F(\theta_t, \alpha_t^k)$ is obtained via a Markovian sample trajectory. Then the policy parameter θ_t is updated for one step, determined by any of the four policy gradient algorithms, namely,

[1]The samples are obtained over a single trajectory path for the entire algorithm execution.

PPG in eq. (4), FWPG in eq. (5), TRPO in eq. (7) and NPG in eq. (8).

2.3 Technical Preliminaries

For the GAIL problem in eq. (1) to be well posed, we assume that $\max_{\alpha\in\Lambda} F(\theta,\alpha)$ exists for any $\theta\in\Theta$, and define the marginal-maximum function of $F(\theta,\alpha)$

$$g(\theta) := \max_{\alpha\in\Lambda} F(\theta,\alpha). \tag{2}$$

We further define the corresponding optimizer $\alpha_{op}(\theta) := \mathrm{argmax}_{\alpha\in\Lambda} F(\theta,\alpha)$. If there exists more than one optimizer, $\alpha_{op}(\theta)$ denotes the elements of the corresponding optimizer set.

Definition 1. *Let $\theta^* = \mathrm{argmin}_{\theta\in\Theta}\, g(\theta)$. The output $\bar{\theta}$ of an algorithm is said to attain an ϵ-global convergence if*

$$g(\bar{\theta}) - g(\theta^*) \le \epsilon$$

holds for a prescribed accuracy $\epsilon \in (0,1)$.

As remarked in Zhang et al. (2020), ϵ-global convergence further implies

$$\max_{\alpha\in\Lambda}[V(\pi_E, r_\alpha) - V(\pi_{\bar{\theta}}, r_\alpha)] \le \max_{\alpha\in\Lambda}\psi(\alpha) + \epsilon.$$

Hence, as long as $\psi(\alpha)$ is chosen properly (for example, with a small regularization coefficient), $\pi_{\bar{\theta}}$ is guaranteed to be sufficiently close to the expert policy.

In this paper, we make the following standard assumptions for our analysis.

Assumption 1. *The regularizer function $\psi(\alpha)$ is differentiable with gradient Lipschitz constant L_ψ.*

Assumption 1 captures the property for designing a regularizer and can be easily attained.

Assumption 2. *For any given θ, the objective function $F(\theta,\alpha)$ in eq. (1) is μ-strongly concave on α.*

Assumption 2 includes the linear function class as a special case. In practice, a strongly convex regularizer $\psi(\alpha)$ is often used to guarantee the strong concavity of $F(\theta,\alpha)$.

Assumption 3 (Ergodicity)**.** *For any policy parameter $\theta\in\Theta$, consider the MDP with policy π_θ and transition kernel $\mathsf{P}(\cdot|s,a)$ or*

$$\tilde{\mathsf{P}}(\cdot|s,a) = \gamma\mathsf{P}(\cdot|s,a) + (1-\gamma)\zeta(\cdot).$$

There exist constants $C_M > 0$ and $0<\rho<1$ such that $\forall t \ge 0$,

$$\sup_{s\in\mathcal{S}} \mathrm{d}_{TV}\left(\mathbb{P}(s_t\in\cdot|s_0=s), \chi_\theta\right) \le C_M\rho^t,$$

where χ_θ is the stationary distribution of the given transition kernel $\mathsf{P}(\cdot|s,a)$ or $\tilde{\mathsf{P}}(\cdot|s,a)$ under policy π_θ and $\mathrm{d}_{TV}(\cdot,\cdot)$ is the total variation distance.

Assumption 3 holds for any time-homogeneous Markov chain with finite state space or any uniformly ergodic Markov chain with general state space.

Assumption 4. *The reward parameterization satisfies Gradient Lipschitz condition, i.e., there exists $L_r \in \mathbb{R}_+$, such that for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$ and for all $\alpha_1,\alpha_2\in\Lambda$, we have*

$$\left\|\nabla_\alpha r_{\alpha_1}(s,a) - \nabla_\alpha r_{\alpha_2}(s,a)\right\|_2 \le L_r\left\|\alpha_1-\alpha_2\right\|_2.$$

Clearly, under Assumption 4, the reward function also has a bounded gradient. There exists $C_r\in\mathbb{R}_+$ such that, for all $\alpha\in\Lambda$, we have

$$\left\|\nabla_\alpha r_\alpha\right\|_{\infty,2} := \sqrt{\sum_{i=1}^{q}\left\|\frac{\partial r_\alpha}{\partial\alpha_i}\right\|_\infty^2} \le C_r,$$

where $\|\cdot\|_\infty$ is taken over the state-action space $\mathcal{S}\times\mathcal{A}$.

We next provide the following Lipschitz properties, which are vital for the analysis of convergence, and were often taken as assumptions in the literature of min-max optimization (Jin et al., 2019; Nouiehed et al., 2019).

Proposition 1. *Suppose Assumptions 1, 3 and 4 hold. Then the GAIL min-max problem in eq. (1) with direct parameterization satisfies the following Lipschitz conditions: $\forall\theta_1,\theta_2\in\Theta$ and $\forall\alpha_1,\alpha_2\in\Lambda$,*

$$\begin{aligned}\|\nabla_\theta F(\theta_1,\alpha_1) - \nabla_\theta F(\theta_2,\alpha_2)\| &\le L_{11}\|\theta_1-\theta_2\| \\ &\quad + L_{12}\|\alpha_1-\alpha_2\|, \\ \|\nabla_\alpha F(\theta_1,\alpha_1) - \nabla_\alpha F(\theta_2,\alpha_2)\| &\le L_{21}\|\theta_1-\theta_2\| \\ &\quad + L_{22}\|\alpha_1-\alpha_2\|,\end{aligned}$$

where $L_{11} = \frac{2\sqrt{2}|\mathcal{A}|C_rC_\alpha}{(1-\gamma)^2}(1+\lceil\log_\rho C_M^{-1}\rceil + (1-\rho)^{-1})$, $L_{12} = \frac{\sqrt{|\mathcal{A}|}C_r}{(1-\gamma)^2}$, $L_{21} = \frac{C_r\sqrt{|\mathcal{A}|}}{1-\gamma}(1+\lceil\log_\rho C_M^{-1}\rceil + (1-\rho)^{-1})$, and $L_{22} = \frac{2\sqrt{q}L_r}{1-\gamma} + L_\psi$.

Furthermore, if $\theta_1=\theta_2$, the above second bound holds with a general parameterization for the policy.

3 Global Convergence of GAIL Algorithms

In this section, we provide the global convergence guarantee for four GAIL algorithms.

3.1 PPG-GAIL and FWPG-GAIL Algorithms

In this section, we study the PPG-GAIL and FWPG-GAIL algorithms, both of which take the general framework in Algorithm 1, and update the policy parameter θ respectively based on projected policy gradient (PPG) and Frank-Wolfe policy gradient (FWPG).

We take the direct parameterization for the policy. At each time t of the outer loop, both PPG-GAIL and FWPG-GAIL first estimate the stochastic policy gradient by drawing a minibatch sample trajectory with length b as $(s_i, a_i) \sim \tilde{\mathsf{P}}^{\pi_{\theta_t}}$ as follows.

$$\widehat{\nabla}_\theta F(\theta_t, \alpha_t)(s,a) = -\frac{\hat{Q}(s,a)}{b(1-\gamma)} \sum_{i=0}^{b-1} \mathbb{1}\{s_i = s\}, \quad (3)$$

for all $s \in \mathcal{S}, a \in \mathcal{A}$, where $\hat{Q}(s,a)$ applies EstQ in Zhang et al. (2019) (also given in Supplementary materials, Section A) with the reward function $r_{\alpha_t}(s,a)$. Then, PPG-GAIL updates θ_t as

$$\theta_{t+1} = P_{\Theta_p}\left(\theta_t - \eta \widehat{\nabla}_\theta F(\theta_t, \alpha_t)\right), \quad (4)$$

where Θ_p is the probability simplex defined in Section 2.2.

Differently from PPG-GAIL, FWPG-GAIL updates θ_t based on the Frank-Wolfe gradient as given by

$$\hat{v}_t = \underset{\theta \in \Theta_p}{\operatorname{argmax}} \left\langle \theta, -\widehat{\nabla}_\theta F(\theta_t, \alpha_t) \right\rangle,$$
$$\theta_{t+1} = \theta_t + \eta\left(\hat{v}_t - \theta_t\right). \quad (5)$$

To analyze the convergence, we first define the gradient dominance property.

Definition 2. *A function $f(\theta)$ satisfies the gradient dominance property, if there exists a positive C, such that*

$$f(\theta) - f(\theta^*) \le C \max_{\bar{\theta} \in \Theta} \langle \theta - \bar{\theta}, \nabla_\theta f(\theta) \rangle$$

for any given $\theta \in \Theta$, where $\theta^* := \operatorname{argmin}_{\theta \in \Theta} f(\theta)$.

The following proposition facilitates to prove global convergence for PPG-GAIL and FWPG-GAIL.

Proposition 2. *The function $g(\theta)$ given in eq.* (2) *satisfies the gradient dominance property.*

The following theorem characterizes the global convergence of PPG-GAIL.

Theorem 1. *Suppose Assumptions 1 to 4 hold. Consider PPG-GAIL with the θ-update stepsize $\eta = \left(L_{11} + \frac{L_{12}L_{21}}{\mu}\right)^{-1}$ and the α-update stepsize $\beta = \frac{\mu}{4L_{22}^2}$, where L_{11}, L_{12}, L_{21} and L_{22} are given in Proposition 1. Then we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \le \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{T}}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2 K}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{B}}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{b}}\right). \quad (6)$$

Theorem 1 implies that if we set $T = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, $K = \mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$, $B = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ and $b = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, then PPG-GAIL converges to an ϵ-accurate *globally* optimal value with an overall sample complexity $T(KB + b) = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^4}\right)$. Due to the Markovian sampling for updating both the reward and policy parameters α and θ, our analysis bounds the two corresponding bias error terms by $\mathcal{O}(\frac{1}{\sqrt{B}})$ and $\mathcal{O}(\frac{1}{\sqrt{b}})$ as shown in eq. (6). Hence, the choices for the mini-batch sizes B and b trade off between the convergence error and the computational complexity. To achieve a given accuracy ϵ, the tradeoff yields the overall complexity of $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^4}\right)$. We also note that the result here provides the first convergence rate for projected stochastic gradient with non-i.i.d. sampling.

We next provide the following theorem, which characterizes the global convergence of FWPG-GAIL.

Theorem 2. *Suppose Assumptions 1 to 4 hold. Consider FWPG-GAIL with the θ-update stepsize $\eta = \frac{1-\gamma}{\sqrt{T}}$ and α-update stepsize $\beta = \frac{\mu}{4L_{22}^2}$, where L_{22} is given in Proposition 1. Then we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \le \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{T}}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2 K}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{B}}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^3\sqrt{b}}\right).$$

Theorem 2 implies that if we let $T = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, $K = \mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$, $B = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ and $b = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, then FWPG-GAIL converges to an ϵ-accurate *globally* optimal value with overall sample complexity $T(KB + b) = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^4}\right)$, which is the same as that of PPG-GAIL. The analysis of FWPG-GAIL also needs to bound the two bias terms due to the Markovian sampling for updating the reward and policy parameters. This is the first analysis that provides the convergence rate for stochastic Frank-Wolfe gradient with non-i.i.d. sampling.

3.2 TRPO-GAIL Algorithm

In this section, we study the TRPO-GAIL algorithm, which takes the general framework in Algorithm 1 and updates the policy parameter θ based on TRPO under λ-regularized MDP. At each time t of the outer loop, TRPO-GAIL adopts the update rule in Shani et al. (2020) for updating θ_t as follows:

$$\pi_{\theta_{t+1}}(\cdot|s) \in \underset{\pi \in \Delta_{\mathcal{A}}}{\operatorname{argmin}} \left\{ \left\langle -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot), \pi - \pi_{\theta_t}(\cdot|s) \right\rangle \ldots + \lambda \left\langle \nabla\omega(\pi_{\theta_t}(\cdot|s)), \pi - \pi_{\theta_t}(\cdot|s) \right\rangle \ldots + \eta_t^{-1} B_\omega(\pi, \pi_{\theta_t}(\cdot|s)) \right\},$$

where $\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}$ denotes the estimation of the Q-function based on EstQ (Zhang et al., 2019) (also given in Supplementary materials, Section A), the regularized reward $r_{\lambda,\alpha_t}(s,a) := r_{\alpha_t}(s,a)+\lambda\omega(\pi_\theta(\cdot|s))$, the negative entropy function

$$\omega(\pi(\cdot|s)) := \sum_{a\in\mathcal{A}} \pi(\cdot|s)\log\pi(\cdot|s) + \log|\mathcal{A}|,$$

and the Bregman distance

$$B_\omega(x,y) := \omega(x) - \omega(y) - \langle\nabla\omega(y), x-y\rangle$$

associated with $\omega(x)$, which is the KL-divergence here. We consider the direct parameterization for the policy, and hence the update for the policy parameter θ can be analytically computed (Shani et al., 2020) as follows. For each $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\theta_{t+1}(s,a) = \frac{1}{Z_t(s)}\theta_t(s,a)\exp\left(\eta_t(\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,a) - \lambda\log\theta_t(s,a))\right), \tag{7}$$

where $Z_t(s)$ is the normalization factor and it equals to

$$\sum_{a'\in\mathcal{A}} \theta_t(s,a')\exp\left(\eta_t(\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,a') - \lambda\log\theta_t(s,a'))\right).$$

The following theorem provides the global convergence of TRPO-GAIL under the unregularized MDP, where we have $\lambda = 0$.

Theorem 3. *Suppose Assumptions 1 to 4 hold. Consider unregularized TRPO-GAIL ($\lambda = 0$) with θ-update stepsize $\eta_t = \frac{1-\gamma}{\sqrt{T}}$ and α-update stepsize $\beta = \frac{\mu}{4L_{22}^2}$, where L_{22} is given in Proposition 1. Then we have,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \le \mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{T}}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2K}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^4B}\right).$$

We further consider the regularized MDP, where we have $\lambda > 0$.

Theorem 4. *Suppose Assumptions 1 to 4 hold. Consider regularized TRPO-GAIL ($\lambda > 0$) with θ-update stepsize $\eta_t = \frac{1}{\lambda(t+2)}$ and α-update stepsize $\beta = \frac{\mu}{4L_{22}^2}$, where L_{22} is given in Proposition 1. Then we have,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \le \tilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3T}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2K}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^4B}\right).$$

Theorem 3 indicates that if we set $T = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, $K = \mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$ and $B = \mathcal{O}\left(\frac{1}{\epsilon}\right)$, then TRPO-GAIL with unregularized MDP converges to an ϵ-accurate *globally* optimal value with a total sample complexity $TKB = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^3}\right)$. Theorem 4 indicates that if we let $T = \tilde{\mathcal{O}}(\frac{1}{\epsilon})$, $K = \mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$, and $B = \mathcal{O}\left(\frac{1}{\epsilon}\right)$, then TRPO-GAIL with regularized MDP converges to an ϵ-accurate *globally* optimal value with an overall sample complexity $TKB = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right)$. The regularized MDP changes the objective function with λ-regularized perturbation and yields orderwisely better sample complexity. Moreover, the sample complexity here is with respect to the convergence in expectation, which improves that in high-probability convergence in Shani et al. (2020) by a factor of $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right)$.

3.3 NPG-GAIL Algorithm

In this section, we study the NPG-GAIL algorithm, which takes the general framework in Algorithm 1 and updates the policy parameter θ based on natural policy gradient (NPG).

We consider the general nonlinear parameterization for the policy, so that the state space may not be finite and for example can be $\mathbb{R}^d$. At each time t of the outer loop, NPG-GAIL ideally should update θ_t via a regularized natural gradient

$$-(F(\theta_t)+\lambda I)^{-1}\nabla_\theta V(\pi_{\theta_t}, r_{\alpha_t}),$$

where

$$F(\theta) = \mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\left[\nabla_\theta\log(\pi_\theta(a|s))\nabla_\theta\log(\pi_\theta(a|s))^\top\right]$$

is the Fisher-information matrix, and λ is the regularization coefficient for avoiding singularity. In practice, we estimate such a natural gradient via solving the problem

$$\min_{w\in R^d}\mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}\left[\nabla_\theta\log(\pi_\theta(a|s))^\top w - A^{\pi_\theta}_\alpha(s,a)\right]^2$$

using the mini-batch linear stochastic approximation (SA) algorithm over a Markovian sampled trajectory, where $A^{\pi_\theta}_\alpha(s,a) := Q^{\pi_\theta}_\alpha(s,a) - V^{\pi_\theta}_\alpha(s)$ is the advance function under reward r_α. More details are provided

Algorithm 2 Policy update in NPG-GAIL

Input: Policy parameter θ_t, reward parameter α_t, stepsize β_W, policy stepsize η, batch-size M and trajectory length T_c
for $i = 0, \cdots, MT_c$ **do**
$s_i \sim \widetilde{\mathsf{P}}(\cdot|s_{i-1}, a_{i-1})$
Sample a_i and a_i' independently from $\pi_{\theta_t}(\cdot|s_i)$
end for
Initialize $W_0 = 0$
for $k = 0, \cdots, T_c - 1$ **do**
for $i = kM, \cdots, (k+1)M - 1$ **do**
Obtain Q-function estimation $\widehat{Q}(s_i, a_i)$ with reward function r_{α_t} by EstQ (Zhang et al., 2019) (also given in Supplementary material, Section A).

$$\begin{aligned}\hat{g}_i = &-\nabla_{\theta_t}\log(\pi_{\theta_t}(a_i|s_i))^\top W_k \nabla_{\theta_t}\log(\pi_{\theta_t}(a_i|s_i)) \\ &+ \widehat{Q}(s_i, a_i)\nabla_{\theta_t}\log(\pi_{\theta_t}(a_i|s_i)) \\ &- \widehat{Q}(s_i, a_i)\nabla_{\theta_t}\log(\pi_{\theta_t}(a_i'|s_i)) - \lambda W_k\end{aligned}$$

end for
$\hat{G}_k = \frac{1}{M}\sum_{i=kM}^{(k+1)M-1}\hat{g}_i$
$W_{k+1} = W_k + \beta_W \hat{G}_k$
end for
$w_t = W_{T_c}$
Return: $\theta_{t+1} = \theta_t - \eta w_t$

in Algorithm 2. Suppose such an algorithm provides an output w_t. Then the policy parameter is updated as

$$\theta_{t+1} = \theta_t - \eta w_t. \tag{8}$$

Since we take the general nonlinear parameterization for the policy, we make the following assumptions for the policy parameterization, which are standard in the literature (Kumar et al., 2019; Zhang et al., 2019; Agarwal et al., 2019; Xu et al., 2020b).

Assumption 5. *For any $\theta, \theta' \in \Theta$, and any state-action pair $(s, a) \in \mathcal{S}\times\mathcal{A}$, there exist positive constants L_π, L_ϕ, C_ϕ and C_π, such that the following bounds hold:*

(1) $$\|\nabla_\theta \log(\pi_\theta(a|s)) - \nabla_\theta \log(\pi_{\theta'}(a|s))\|_2 \leq L_\phi \|\theta - \theta'\|_2$$

(2) $\|\nabla_\theta \log(\pi_\theta(a|s))\|_2 \leq C_\phi$,

(3) $\|\pi_\theta(\cdot|s) - \pi_{\theta'}(\cdot|s)\|_{TV} \leq C_\pi \|\theta - \theta'\|_2$,

where $\|\cdot\|_{TV}$ denotes the total-variation norm.

Next, we provide the following theorem, which characterizes the global convergence of NPG-GAIL.

Theorem 5. *Suppose Assumptions 1 to 5 hold. Consider NPG-GAIL with θ-update stepsize $\eta = \frac{1-\gamma}{\sqrt{T}}$, α-update stepsize $\beta = \frac{\mu}{4L_{22}^2}$, and the SA-update stepsize $\beta_W = \frac{\lambda_P}{4(C_\phi^2+\lambda)^2}$, where L_{22} is given in Proposition 1. Then we have*

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[g(\theta_t)\right] - g(\theta^*) \\ &\leq \mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{T}}\right) + \mathcal{O}\left(e^{-(1-\gamma)^2 K}\right) \\ &\quad + \mathcal{O}\left(\frac{1}{(1-\gamma)^4 B}\right) + \mathcal{O}\left(e^{-T_c}\right) + \mathcal{O}\left(\frac{\lambda}{1-\gamma}\right) \\ &\quad + \mathcal{O}\left(\frac{\zeta'}{(1-\gamma)^{3/2}}\right) + \mathcal{O}\left(\frac{1}{(1-\gamma)^2\sqrt{M}}\right).\end{aligned}$$

where

$$\zeta' = \max_{\theta\in\Theta,\alpha\in\Lambda}\min_{w\in R^d} \sqrt{\mathbb{E}_{\nu_{\pi_\theta}}\left[\nabla_\theta \log(\pi_{\pi_\theta}(a|s))^\top w - A_\alpha^{\pi_\theta}(s,a)\right]^2}$$

and T_c and M are defined in Algorithm 2.

Theorem 5 indicates that if we let $T = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, $K = \mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$, $B = \mathcal{O}\left(\frac{1}{\epsilon}\right)$, $T_c = \mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$, $\lambda = \mathcal{O}\left(\zeta'\right)$ and $M = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, then NPG-GAIL converges to an $(\epsilon + \mathcal{O}(\zeta'))$-accurate *globally* optimal value with an overall sample complexity of $T(KB + T_c M) = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^4}\right)$, which is the same as PPG-GAIL and FWPG-GAIL. Comparison of Theorem 3 and Theorem 5 indicates that TRPO-GAIL has a better sample complexity than NPG-GAIL, mainly because TRPO can update the policy parameter based on an analytical form, which saves the samples that NPG uses for estimating the natural gradient by solving the quadratic optimization problem.

4 Conclusion

In this paper, we study four GAIL algorithms, each of which is implemented in an alternating fashion between a popular policy gradient algorithm for the policy update and a gradient ascent for the reward update. Our focus is on investigating whether incorporation of these policy gradient algorithms to the GAIL framework will still have global convergence guarantee. We show that all these GAIL algorithms converge globally as long as the objective function is properly regularized (to be strongly concave) with respect to the reward parameter. We also anticipate that the analysis tools that we develop here will benefit the future theoretical studies of similar problems including GANs, min-max optimization, and bi-level optimization algorithms.

Acknowledgements

The work was supported partially by the U.S. National Science Foundation under the grants CCF-1909291 and CCF-1900145.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2019). Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*.

Cai, Q., Hong, M., Chen, Y., and Wang, Z. (2019). On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*.

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 *Proc. International Conference on Learning Representations (ICLR)*.

Finn, C., Christiano, P., Abbeel, P., and Levine, S. (2016). A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*.

Fu, J., Luo, K., and Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. In *Proc. International Conference on Learning Representations (ICLR)*.

Ghasemipour, S. K. S., Zemel, R., and Gu, S. (2019). A divergence minimization perspective on imitation learning methods. In *Proc. Conference on Robot Learning (CoRL)*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems (NIPS)*.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Proc. Advances in Neural Information Processing Systems (NIPS)*.

Jin, C., Netrapalli, P., and Jordan, M. I. (2019). What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*.

Ke, L., Barnes, M., Sun, W., Lee, G., Choudhury, S., and Srinivasa, S. (2019). Imitation learning as f-divergence minimization. *arXiv preprint arXiv:1905.12888*.

Konda, V. (2002). *Actor-critic algorithms*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Kumar, H., Koppel, A., and Ribeiro, A. (2019). On the sample complexity of actor-critic method for reinforcement learning with function approximation. *arXiv preprint arXiv:1910.08412*.

Laskey, M., Lee, J., Hsieh, W., Liaw, R., Mahler, J., Fox, R., and Goldberg, K. (2017). Iterative noise injection for scalable imitation learning. In *Proc. 1st Conference on Robot Learning (CoRL)*.

Lin, T., Jin, C., and Jordan, M. (2020). Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*.

Ng, A. Y. and Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. (2019). Solving a class of non-convex min-max games using iterative first order methods. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. In *Proc. Advances in Neural Information Processing Systems (NIPS)*.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. (2018). An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179.

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

Ross, S. and Bagnell, D. (2010). Efficient reductions for imitation learning. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Ross, S., Gordon, G. J., and Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Shani, L., Efroni, Y., and Mannor, S. (2020). Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*.

Xu, T., Wang, Z., and Liang, Y. (2020a). Improving sample complexity bounds for actor-critic algorithms. *arXiv preprint arXiv:2004.12956*.

Xu, T., Wang, Z., Zhou, Y., and Liang, Y. (2020b). Reanalysis of variance reduced temporal difference learning. *arXiv preprint arXiv:2001.01898*.

Zhang, K., Koppel, A., Zhu, H., and Başar, T. (2019). Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*.

Zhang, Y., Cai, Q., Yang, Z., and Wang, Z. (2020). Generative adversarial imitation learning with neural networks: Global optimality and convergence rate. *arXiv preprint arXiv:2003.03709*.