When Will Generative Adversarial Imitation Learning Algorithms Attain Global Convergence

Cheng-Chun Wu
Department of Computer Science
National Yang Ming Chiao Tung University
ricky310551149.cs10@nycu.edu.tw

1 Introduction

1.1 Introduction

Reinforcement learning is a field in machine learning that emphasizes how to act based on the environment to maximize returns. Therefore, the formulation of the reward function is the most special and important part of reinforcement learning. But in reality, there are many situations which cause we cannot effectively define the reward function. Imitation learning is to find an ideal policy in this kind of situation, so that this policy can be as close as possible to the demonstration of experts. Two imitation learning methods have been developed. The first method is Behavior Cloning, which uses supervised learning to match the mapping from state space to action space, but the BC method often suffers from high sample complexity due to covariate shift. The Dagger approach reduces the above problems by further interaction with experts.
The second method is inverse reinforcement learning, which tries to recover an unknown reward function based on the expert's activity trajectory, and then uses such a reward function to find the best policy. Regarding the method of IRL, the most popular one is generative adversarial imitation learning (GAIL). It uses the combination of IRL and generative adversarial network, and uses the min-max optimization problem similar to GAN as the main framework.
They tried to use a variety of commonly used gradient algorithms, all of which interact in a stochastic gradient ascent to update the reward. Including four methods, namely 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.
This paper aims to substantially expand the aforementioned global convergence results as follows.

- They allow general MDP models, not necessarily linear MDP. They 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, they 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 they assume that samples follow a general Markovian distribution.

1.2 Main Contributions

In this paper, they 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).
First, they provide the convergence rate for three major types of algorithms. (a) (PPG)-GAIL and (FWPG)-GAIL, (b) (TRPO)-GAIL, (c) (NPG)-GAIL.
Second, they show that all these alternating algorithms converge to the global minimum with a sublinear rate.
Furthermore, in contrast to conventional minmax 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.

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.

1.3 Related work

Theory for IRL via adversarial training:
Most relevant to their 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.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 gradientdecent 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. Their study here establishes global convergence for GAIL for general MDP and the nonlinear reward function class.

Difference from conventional min-max problems:
First, since these algorithms continuously update the policy, the samples that are used for iterations are sampled by timevarying 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.

Connection to policy gradient algorithms:
In the GAIL framework, the policy optimization is jointly performed with the reward optimization

via a minmax 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 the global convergence is generally not guaranteed if these algorithms are executed in an alternating fashion with reward iterations. 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 (S, A, P, r, γ). We assume that state space(S) and action space(A) is finite. We use $s \in S$ and $a \in A$ to denote a state and an action. The policy π is the conditional probability denoted by $\pi(a|s)$. The probability transition kernel is denoted by $P(s'|s,a)$, where the next state $s' \in S$ .The reward function $r_t = r(s,a)$ 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:

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

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

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

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

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

and policy π if the Markov chain is ergodic. Thus $\widetilde{P}$ 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 $\alpha \in \Lambda \subset R^d$, 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.,$\| \alpha_1 - \alpha_2 \|_2 \leq C_\alpha, \forall \alpha_1, \alpha_2 \in \Lambda$.
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 $\theta \in \Theta_p := \{\theta : \theta_{s,a} \geq 0, \sum_{a\in A} \theta_{s,a} = 1, \forall s \in S, a \in A\}$.
The GAIL framework is formulated as the following min-max optimization problem.

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

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

$$\alpha_t^{k+1} = P_\Lambda(\alpha_t^k + \beta \widetilde{\bigtriangledown}_\alpha F(\theta_t, \alpha_t^k)),$$

where the gradient estimator $\widetilde{\bigtriangledown}_\alpha F(\theta_t, \alpha_t^k)$ is obtained via a Markovian sample trajectory.

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 $(s_i^\theta, a_i^\theta) \sim \tilde{\mathsf{P}}^{\pi_\theta 1}$

$$\widehat{\nabla}_\alpha F(\theta, \alpha) = \frac{1}{(1-\gamma)B} \sum_{i=0}^{B-1} \nabla_\alpha r_\alpha(s_i^E, a_i^E) - \frac{1}{(1-\gamma)B} \sum_{i=0}^{B-1} \nabla_\alpha r_\alpha(s_i^\theta, a_i^\theta) - \nabla_\alpha \psi(\alpha)$$

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

  end 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

2.3 Technical Preliminaries

Define the marginal-maximum function of $F(\theta, \alpha)$

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

Define the corresponding optimizer

$$\alpha_{op}(\theta) := \arg\max_{\alpha \in \Lambda} F(\theta, \alpha).$$

Definition 1:
Let $\theta^* = \arg\min_{\theta \in \Theta} g(\theta)$. The output $\overline{\theta}$ of an algorithm is said to attain an ϵ-global convergence if

$$g(\overline{\theta}) - g(\theta^*) \leq \epsilon$$

holds for a prescribed accuracy $\epsilon \in (0, 1)$.
As remarked in Zhang et al. (2020), ϵ-global convergence further implies

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

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

In this paper, we make the following standard assumptions for their analysis.
Assumption 1:
The regularizer function $\psi(\alpha)$ is differentiable with gradient Lipschitz constant L_ψ.

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

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

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

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

$$\sup_{s \in S} d_{TV}(P(s_t \in \cdot | s_0 = s), \chi_\theta) \leq C_{M\rho^t},$$

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

Assumption 4:
The reward parameterization satisfies Gradient Lipschitz condition, i.e., there exists $L_r \in R_+$, such that for all$s \in S$ and $a \in A$ and for all $\alpha_1, \alpha_2 \in \Lambda$, we have

$$\| \bigtriangledown_\alpha r_{\alpha_1}(s,a) - \bigtriangledown_\alpha r_{\alpha_2}(s,a) \|_2 \leq L_r \| \alpha_1 - \alpha_2 \|_2 .$$

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 $\forall \alpha_1, \alpha_2 \in \Lambda$,

$$\| \bigtriangledown_\theta F(\theta_1, \alpha_1) - \bigtriangledown_\theta F(\theta_2, \alpha_2) \| \leq L_{11} \| \theta_1 - \theta_2 \| + L_{12} \| \alpha_1 - \alpha_2 \|,$$

$$\| \bigtriangledown_\alpha F(\theta_1, \alpha_1) - \bigtriangledown_\alpha F(\theta_2, \alpha_2) \| \leq L_{21} \| \theta_1 - \theta_2 \| + L_{22} \| \alpha_1 - \alpha_2 \|,$$

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

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 $(s_i, a_i) \sim \widetilde{P}^{\pi_{\theta_t}}$ as follows.

$$\hat{\bigtriangledown} F(\theta_t, \alpha_t)(s,a) = -\frac{\hat{Q}(s,a)}{b(1-\gamma)} \sum_{i=0}^{b-1} 1\{s_i = s\}, \tag{3}$$

for all $s \in S$; $a \in A$, 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

$$\theta_{t+1} = P_{\Theta_p}(\theta_t - \eta \hat{\bigtriangledown}_\theta F(\theta_t, \alpha_t)) \tag{4}$$

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

$$\hat{v}_t = \arg\max_{\theta \in \Theta_p} \langle \theta, -\hat{\bigtriangledown} F(\theta_t, \alpha_t) \rangle, \theta_{t+1} = \theta_t + \eta(\hat{v}_t - \theta_t) \tag{5}$$

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

$$f(\theta) - f(\theta^*) \leq C \max_{\overline{\theta} \in \Theta} \langle \theta - \overline{\theta}, \bigtriangledown_\theta f(\theta) \rangle$$

for any given $\theta \in \Theta$, where $\theta^* := \arg\min_{\theta \in \Theta} f(\theta)$.
Proposition: 2
The function $g(\theta)$ given in eq. (2) satisfies the gradient dominance property.

3.1.1 global convergence of PPG-GAIL.

Theorem 1:
Suppose Assumptions 1 to 4 hold. Consider PPG-GAIL with the θ-update stepsize $\eta = (L_{11} + \frac{L_{12}L_{21}}{\mu})^{-1}$ 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

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

Theorem 1 implies that if we set $T = O(\frac{1}{\epsilon^2}), K = O(log(\frac{1}{\epsilon})), B = O(\frac{1}{\epsilon^2})$and $b = O(\frac{1}{\epsilon^2})$, then PPG-GAIL converges to an ϵ-accurate globally optimal value with an overall sample complexity $T(KB + b) = \widetilde{O}(\frac{1}{\epsilon^4})$

3.1.2 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

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

Theorem 2 implies that if we set $T = O(\frac{1}{\epsilon^2}), K = O(log(\frac{1}{\epsilon})), B = O(\frac{1}{\epsilon^2})$and $b = O(\frac{1}{\epsilon^2})$, then FWPG-GAIL converges to an ϵ-accurate globally optimal value with an overall sample complexity $T(KB + b) = \widetilde{O}(\frac{1}{\epsilon^4})$

3.2 TRPO-GAIL Algorithm

TRPO-GAIL adopts the update rule in Shani et al. (2020) for updating θ_t as follows:

$$\pi_{\theta_{t+1}}(\cdot|s) \in \arg\min_{\pi\in\Delta_A}\{\langle -\hat{Q}^{\pi_{\theta_t}}_{\lambda,\alpha_t}(s,\cdot), \pi-\pi_{\theta_t}(\cdot|s)\rangle+\lambda\langle\bigtriangledown\omega(\pi_{\theta_t}(\cdot|s)), \pi-\pi_{\theta_t}(\cdot|s)\rangle+\eta_t^{-1}B_\omega(\pi,\pi_{\theta_t}(\cdot|s))\},$$

where $\hat{Q}^{\pi_t}_{\lambda,\alpha_t}$ denotes the estimation of the Q-function based on EstQ.
the regularized reward $r_{\lambda,\alpha_t}(s,a) := r_{\alpha_t}(s,a) + \lambda\omega(\pi_\theta(\cdot|s))$,
the negative entropy function

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

and the Bregman distance

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

associated with $\omega(x)$, which is the KL-divergence here.
For each $(s,a) \in S \times A$,

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

where $Z_t(s)$ is the normalization factor,

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

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,

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

Theorem 3 implies that if we set $T = O(\frac{1}{\epsilon^2}), K = O(log(\frac{1}{\epsilon})), andB = O(\frac{1}{\epsilon})$, then TRPO-GAIL with unregularized MDP converges to an ϵ-accurate globally optimal value with an overall sample complexity $TKB = \widetilde{O}(\frac{1}{\epsilon^3})$.

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 $\beta = \frac{\mu}{L_{22}^2}$,where L_{22} is given in Proposition 1. Then we have,

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

Theorem 4 implies that if we set $T = O(\frac{1}{\epsilon}), K = O(log(\frac{1}{\epsilon})), andB = O(\frac{1}{\epsilon})$, then TRPO-GAIL with regularized MDP converges to an ϵ-accurate globally optimal value with an overall sample complexity $TKB = \widetilde{O}(\frac{1}{\epsilon^2})$.

3.3 NPG-GAIL Algorithm

As algorithm 2, NPG-GAIL ideally should update θ_t via a regularized natural gradient

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

where

$$F(\theta) = \mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}[\bigtriangledown_\theta log(\pi_\theta(a|s)) \bigtriangledown_\theta log(\pi_\theta(a|s))^T]$$

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.

$$\min_{\omega\in R^d} \mathbb{E}_{(s,a)\sim\nu_{\pi_\theta}}[\bigtriangledown_\theta log(\pi_\theta(a|s))^T\omega - A_\alpha^{\pi_\theta}(s,a)]^2$$

where $A_\alpha^{\pi_\theta}(s,a) := Q_\alpha^{\pi_\theta}(s,a) - V_\alpha^{\pi_\theta}(s)$ Suppose such an algorithm provides an output ω. Then the policy parameter is updated as

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

Assumption 5:

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

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

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

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

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

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

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

where

$$\zeta' = \max_{\theta\in\Theta,\alpha\in\Lambda} \min_{\omega\in R^d} \sqrt{\mathbb{E}_{\nu_{\pi_\theta}}[\bigtriangledown_\theta log(\pi_{\pi_\theta}(a|s))^T\omega - A_\alpha^{\pi_\theta}(s,a)]^2}$$

and T_c and M are defined in Algorithm 2.

Theorem 5 implies that if we set $T = O(\frac{1}{\epsilon^2}), K = O(log(\frac{1}{\epsilon})), B = O(\frac{1}{\epsilon}), T_c = O(log(\frac{1}{\epsilon})), \lambda = O(\zeta')$, and $M = O(\frac{1}{\epsilon^2})$, then NPG-GAIL converges to an $(\epsilon + O(\zeta'))$-accurate globally optimal value with an overall sample complexity $T(KB + T_cM = \widetilde{O}(\frac{1}{\epsilon^4})$.

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: $\theta_{t+1} = \theta_t - \eta w_t$

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.
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.
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.

References