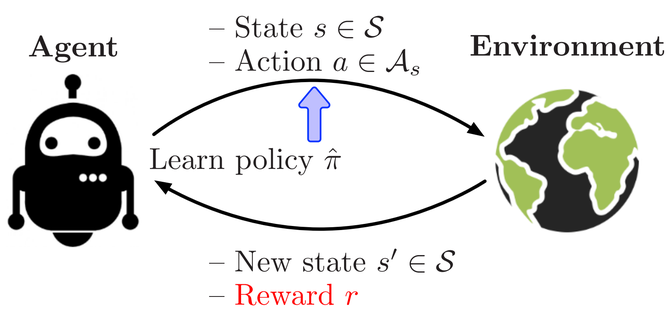
On Computation and Generalization of Generative Adversarial Imitation Learning

Minshuo Chen*, Yizhou Wang†, Tianyi Liu*, Zhuoran Yang‡, Xingguo Li‡, Zhaoran Wang◇, Tuo Zhao*
*Georgia Tech, †Xi'an Jiaotong University, ‡Princeton University, ◇Northwestern University

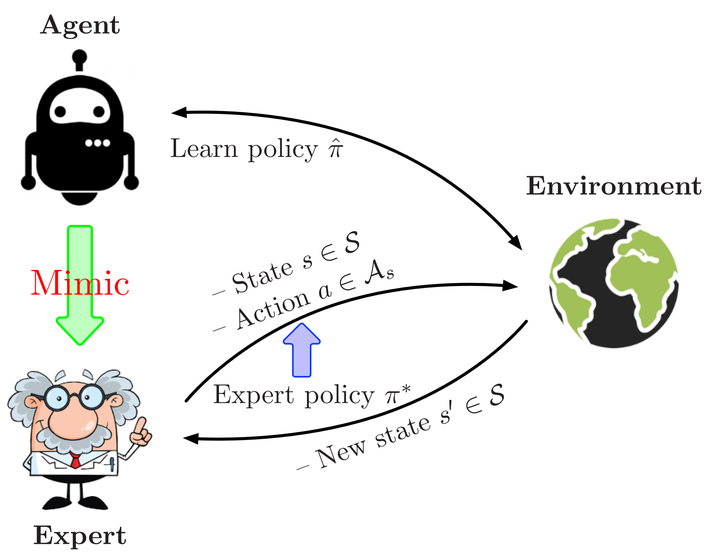
Background

▷ **Reinforcement Learning**



Reward functions are often **difficult** to describe in complex tasks.

▷ **Imitation Learning**



- Behavior Cloning
  - Supervised learning;
  - Mismatch in training and testing;
  - Poor generalization.
- Inverse Reinforcement Learning
  - Learn a reward function;
  - Bi-level optimization;
  - Computationally inefficient.

Generative Adversarial Imitation Learning

▷ $\mathcal{R}$**-distance**

$$d_{\mathcal{R}}(\pi, \pi') = \sup_{r \in \mathcal{R}} \mathbb{E}_{\pi}[r(s,a)] - \mathbb{E}_{\pi'}[r(s,a)].$$

- π, π': evaluation policies;
- $\mathcal{R}$: a symmetric class of reward functions;
- $\mathbb{E}_{\pi}[r]$, $\mathbb{E}_{\pi'}[r]$: expected average rewards.

Special examples:

- Wasserstein distance: $\mathcal{R} = \{1\text{-Lipschitz function}\}$;
- Total variation distance: $\mathcal{R} = \{\pm \mathbb{1}_A\}$.

▷ **Generative Adversarial Imitation Learning (GAIL)**

$$\widehat{\pi} \in \operatorname*{argmin}_{\pi} \max_{r \in \mathcal{R}} d_{\mathcal{R}}(\pi, \pi_n^*) = \operatorname*{argmin}_{\pi} \max_{r \in \mathcal{R}} \mathbb{E}_{\pi}[r(s,a)] - \mathbb{E}_{\pi_n^*}[r(s,a)].$$

- $\mathbb{E}_{\pi_n^*}[r(s,a)]$: expert empirical average reward;
- $\widehat{\pi}$ minimizes its discrepancy with π_n^* under the $\mathcal{R}$-distance.

Questions

- Is $\widehat{\pi}$ close to π^* under the $\mathcal{R}$-distance? $\Longrightarrow$ Generalization theory;
- Can $\widehat{\pi}$ be obtained efficiently? $\Longrightarrow$ Computation theory.

Generalization of GAIL

Assumption 1 (β-Mixing). *Expert demonstrations $(s_t, a_t)_{t=0}^{T-1}$ forms an exponentially β-mixing Markov chain, i.e.,*

$$\sup_m \mathbb{E}_{B \in \sigma_0^m} \sup_{A \in \sigma_{m+k}^{\infty}} |\mathbb{P}(A|B) - \mathbb{P}(A)| \leq \beta_0 \exp(-\beta_1 k^{\alpha}),$$

Assumption 2 (Bounded Feature Vectors). *There exist feature vectors $\psi_s \in \mathbb{R}^{d_{\mathcal{S}}}$ and $\psi_a \in \mathbb{R}^{d_{\mathcal{A}}}$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$, respectively, satisfying $\|\psi_s\|_2 \leq 1$ and $\|\psi_a\|_2 \leq 1$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

Assumption 3 (Bounded Reward). *The reward function class is bounded, i.e., $\|r\|_{\infty} \leq B_r$ for any $r \in \mathcal{R}$.*

Theorem 1 (Generalization of GAIL). *Suppose Assumptions 1-3 hold, and $\widehat{\pi}$ satisfies $d_{\mathcal{R}}(\pi_n^*, \widehat{\pi}) - \inf_{\pi} d_{\mathcal{R}}(\pi_n^*, \pi) < \epsilon$. With probability at least $1-\delta$ over the expert demonstrations $\{(a_t^{(i)}, s_t^{(i)})_{t=0}^{T-1}\}_{i=1}^n$, we have*

$$d_{\mathcal{R}}(\pi^*, \widehat{\pi}) - \inf_{\pi} d_{\mathcal{R}}(\pi^*, \pi) \leq O\left(\frac{B_r}{\sqrt{nT/\zeta}} \sqrt{\log \mathcal{N}\left(\mathcal{R}, \sqrt{\tfrac{\zeta}{nT}}, \|\cdot\|_{\infty}\right)} + B_r \sqrt{\frac{\log(1/\delta)}{nT/\zeta}} \right) + \epsilon \quad \textit{with } \zeta = (\beta_1^{-1} \log \tfrac{\beta_0 T}{\delta})^{\frac{1}{\alpha}}.$$

ζ accounts for the **dependence** in the expert demonstrations.

▷ **Reproducing Kernel Reward Function**

$$r(s,a) = \theta^{\top} g(\psi_s, \psi_a), \quad \text{with } \theta \in \mathbb{R}^q, \|\theta\|_2 \leq B_\theta \text{ and } g \text{ being } \rho_g\text{-Lipschitz}, g(0,0) = 0.$$

Corollary 2. *With probability at least $1-\delta$ over the joint distribution of $\{(a_t^i, s_t^i)_{t=0}^{T-1}\}_{i=1}^n$, we have*

$$d_{\mathcal{R}}(\pi^*, \widehat{\pi}) - \inf_{\pi} d_{\mathcal{R}}(\pi^*, \pi) \leq O\left(\frac{\rho_g B_\theta}{\sqrt{nT/\zeta}} \sqrt{q \log\left(\rho_g B_\theta \sqrt{nT/\zeta}\right)} + \rho_g B_\theta \sqrt{\frac{\log(1/\delta)}{nT/\zeta}} \right) + \epsilon.$$

▷ **Neural Network Reward Function**

$$r(s,a) = W_D^{\top} \sigma(W_{D-1} \sigma(\cdots \sigma(W_1 [\psi_a^{\top}, \psi_s^{\top}]^{\top}))), \quad \text{with } W_i \in \mathbb{R}^{d_i \times d_{i-1}}, \|W_i\|_2 \leq 1 \text{ and Lipschitz activation } \sigma.$$

Corollary 3. *With probability at least $1-\delta$ over the joint distribution of $\{(a_t, s_t)_{t=0}^{T-1}\}_{i=1}^n$, we have*

$$d_{\mathcal{R}}(\pi^*, \widehat{\pi}) - \inf_{\pi} d_{\mathcal{R}}(\pi^*, \pi) \leq O\left(\frac{1}{\sqrt{nT/\zeta}} \sqrt{d^2 D \log\left(D\sqrt{dnT/\zeta}\right)} + \sqrt{\frac{\log(1/\delta)}{nT/\zeta}} \right) + \epsilon \quad \textit{with } d = \max_i d_i.$$

Trade-off

$$d_{\mathcal{R}}(\hat{\pi}, \pi^*) \leq \inf_{\pi} d_{\mathcal{R}}(\pi, \pi^*) + \text{generalization gap}$$

class of π NOT too small $\Rightarrow$ $\mathcal{R}$ NOT too small ***V.S.*** $\mathcal{R}$ NOT too large

Computation of GAIL

We consider kernel reward function $r_\theta(s,a)$ and policy parameterized by ω. A regularized GAIL objective function:

$$\min_{\omega} \max_{\|\theta\|_2 \leq \kappa} F(\theta, \omega) = \mathbb{E}_{\pi_\omega}[r_\theta(s,a)] - \mathbb{E}_{\pi^*}[r_\theta(s,a)] - \lambda H(\pi_\omega) - \frac{\mu}{2}\|\theta\|_2^2, \quad \text{with } \lambda, \mu \text{ tuning parameters.}$$

▷ **Alternating Stochastic Gradient Descent Ascent**

$$\theta^{(t+1)} = \Pi_\kappa\left(\theta^{(t)} + \eta_\theta \tfrac{1}{q_\theta} \textstyle\sum_{j \in \mathcal{M}_\theta^{(t)}} \nabla_\theta f_j(\omega^{(t)}, \theta^{(t)})\right) \quad \text{and} \quad \omega^{(t+1)} = \omega^{(t)} - \eta_\omega \tfrac{1}{q_\omega} \textstyle\sum_{j \in \mathcal{M}_\omega^{(t)}} \nabla_\omega \widetilde{f}_j(\omega^{(t)}, \theta^{(t+1)}),$$

- η_θ, η_ω: learning rates;
- $\mathcal{M}_\theta^{(t)}, \mathcal{M}_\omega^{(t)}$ mini-batches;
- $\nabla f_j, \nabla \widetilde{f}_j$: independent stochastic approximations of ∇F.

Assumption 4 (Unbiased and Bounded Stochastic Gradients). *There are two positive constants M_ω and M_θ such that*

$$\textit{Unbiased}: \ \mathbb{E}\nabla f_j(\omega,\theta) = \mathbb{E}\nabla \widetilde{f}_j(\omega,\theta) = \nabla F(\omega,\theta),$$
$$\textit{Bounded}: \ \mathbb{E}\|\nabla_\omega \widetilde{f}_j(\omega,\theta) - \nabla_\omega F(\omega,\theta)\|_2^2 \leq M_\omega \quad \textit{and} \quad \mathbb{E}\|\nabla_\theta f_j(\omega,\theta) - \nabla_\theta F(\omega,\theta)\|_2^2 \leq M_\theta.$$

Assumption 5 (Bounded and Smooth Regularizer). *There exist constants B_H and S_H such that*

$$H(\pi_\omega) \leq B_H \quad \textit{and} \quad \|\nabla_\omega H(\widetilde{\pi}_\omega) - \nabla_\omega H(\widetilde{\pi}_{\omega'})\|_2 \leq S_H \|\omega - \omega'\|_2.$$

Theorem 4 (Computation of GAIL). *Suppose Assumptions 1-5 hold, and certain regularity conditions on π_ω. With properly chosen η_ω and η_θ, for any given $\epsilon > 0$, alternating stochastic gradient descent ascent converges to a fist-order stationary point in at most*

$$N = \eta(C_0 + 4\sqrt{2}\rho_g\kappa + \mu\kappa^2 + 2\lambda B_H)\epsilon^{-1}$$

iterations. Here C_0 depends on the initialization, η depends on η_ω and η_θ, and the mini-batch size is $\widetilde{O}(1/\epsilon)$.

Empirical Convergence of GAIL