Adversarial Soft Advantage Fitting: Imitation Learning without Policy Optimization

Yi-Chen Li
liyc@lamda.nju.edu.cn
LAMDA, Nanjing University

November 30, 2022

Authors and Institutes

Adversarial Soft Advantage Fitting: Imitation Learning without Policy Optimization

Paul Barde[*†]
Québec AI institute (Mila)
McGill University
bardepau@mila.quebec

Julien Roy[*†]
Québec AI institute (Mila)
Polytechnique Montréal
julien.roy@mila.quebec

Wonseok Jeon[*]
Québec AI institute (Mila)
McGill University
jeonwons@mila.quebec

Joelle Pineau[‡]
Québec AI institute (Mila)
McGill University
Facebook AI Research

Christopher Pal[‡]
Québec AI institute (Mila)
Polytechnique Montréal
Element AI

Derek Nowrouzezahrai
Québec AI institute (Mila)
McGill University

Barde P, Roy J, Jeon W, et al. Adversarial soft advantage fitting: Imitation learning without policy optimization[C]. Advances in Neural Information Processing Systems, 2020, 33: 12334-12344.

Table of Contents

Preliminaries

Motivation

Method

Experiments

Discussion

Table of Contents

Preliminaries

Motivation

Method

Experiments

Discussion

Preliminaries I

- Consider a T-horizon γ-discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{P}_0, \gamma, r, T \rangle$, where
  - $\mathcal{S}$ is the state space;
  - $\mathcal{A}$ is the action space;
  - $\mathcal{P}(s'|s, a) \in [0, 1]$ is the transition dynamics;
  - $\mathcal{P}_0(s_0)$ is the initial state distribution;
  - $\gamma \in [0, 1]$ is the discount factor;
  - $r(s, a) \in \mathbb{R}$ with r being bounded is the reward function;
  - $T \in \mathbb{N} \cup \{\infty\}$ is the horizon length.

  We suppose that $\mathcal{S}$ and $\mathcal{A}$ are both finite, and $T < \infty$ for $\gamma = 1$.
- For any trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T)$, define its probability $P_\pi(\tau)$ of being sampled on $\mathcal{M}$ as

$$P_\pi(\tau) \triangleq \mathcal{P}_0(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t)\mathcal{P}(s_{t+1}|s_t, a_t).$$

Preliminaries II

- For any policy π, we define its occupancy measure $\rho^\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ as

$$\rho^\pi(s,a) = \frac{1}{Z(\gamma,T)} \sum_{t=0}^{T-1} \gamma^t \left(\sum_{\tau:(s_t,a_t)=(s,a)} P_\pi(\tau) \right) = \frac{1}{Z(\gamma,T)} \sum_{t=0}^{T-1} \gamma^t \Pr(s_t = s)\pi(a|s),$$

  where $Z(\gamma,T) = \sum_t^{T-1} \gamma^t$.

- The expected sum of discounted rewards can be expressed in term of the occupancy measure as

$$J_\pi[r(s,a)] \triangleq \mathbb{E}_{\tau\sim P_\pi}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t,a_t)\right] = Z(\gamma,T)\mathbb{E}_{(s,a)\sim\rho^\pi}[r(s,a)].$$

Preliminaries III

- By properly regularizing the learned reward function r, (Ho & Ermon, 2016) associated the maximum entropy IRL problem with GAN, thus getting GAIL's objective:

$$\min_{\pi} \max_{D} \mathbb{E}_{(s,a)\sim\rho^{\pi_E}} \left[\log\left(D(s,a)\right)\right] + \mathbb{E}_{(s,a)\sim\rho^{\pi}} \left[\log\left(1 - D(s,a)\right)\right],$$

which is equivalent to the following Jensen-Shannon divergence minimization problem,

$$\min_{\pi} D_{\mathsf{JS}}(\rho_{\pi}, \rho_{\pi_E}).$$

Table of Contents

Preliminaries

Motivation

Method

Experiments

Discussion

Motivation

- GAIL performs not well in practice.
  1. The min-max optimization procedure of GAIL is brittle and unstable;
  2. The RL process is sample-inefficient and tricky.
- Can we instead imitate the expert without adversarial training and RL policy optimization?

Table of Contents

Preliminaries

Motivation

Method

Experiments

Discussion

Method I

Consider the following new objective

$$\min_{\pi_G} \max_{\tilde{\pi}} L(\tilde{\pi}, \pi_G) := \mathbb{E}_{\tau \sim P_{\pi_E}} \left[\log D_{\tilde{\pi},\pi_G}(\tau)\right] + \mathbb{E}_{\tau \sim P_{\pi_G}} \left[\log(1 - D_{\tilde{\pi},\pi_G}(\tau))\right], \tag{1}$$

with the *structured discriminator*:

$$D_{\tilde{\pi},\pi_G}(x) = \frac{P_{\tilde{\pi}}(\tau)}{P_{\tilde{\pi}}(\tau) + P_{\pi_G}(\tau)} = \frac{q_{\tilde{\pi}}(\tau)}{q_{\tilde{\pi}}(\tau) + q_{\pi_G}(\tau)}.$$

Here $q_\pi(\tau) = \prod_{t=0}^{T-1} \pi(a_t|s_t)$.

Lemma 1.

For any stationary policy π, there is an one-to-one correspondence between π and q_π on $\mathcal{M}$.

Proof: For any trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T)$, we have that $P_\pi(\tau) = q_\pi \times \xi(\tau)$, with

$$\xi(\tau) := \mathcal{P}_0(s_0) \prod_{t=0}^{T-1} \mathcal{P}(s_{t+1}|s_t, a_t).$$

Method II

Since $\xi(\tau)$ is solely determined by the MDP $\mathcal{M}$, there is an one-to-one correspondence between q_π and P_π. And from the definition of ρ^π, we know that ρ^π can be computed directly from P_π. From Theorem 2 of (Syed et al., 2008), we know that ρ^π and π is one-to-one corresponded. Thus concluding the proof. □

Theorem 1.

The optimal $\tilde{\pi}^* = \arg\max_{\tilde{\pi}} L(\tilde{\pi}, \pi_G)$ *for any* π_G *in* $Eq.$ (1) *is such that* $q_{\tilde{\pi}^*} = q_{\pi_E}$, *and using* $\tilde{\pi}^*$ *as the generator policy* $\tilde{\pi}^*$ *minimizes* $L(\tilde{\pi}^*, \pi_G)$, *i.e,*

$$\tilde{\pi}^* \in \arg\min_{\pi_G} \max_{\tilde{\pi}} L(\tilde{\pi}, \pi_G) = \arg\min_{\pi_G} L(\tilde{\pi}^*, \pi_G).$$

Proof: Theorem 1 states that given $L(\tilde{\pi}, \pi_G)$ defined in Eq. (1):

- a $\tilde{\pi}^* = \arg\max_{\tilde{\pi}} L(\tilde{\pi}, \pi_G)$ satisfies $q_{\tilde{\pi}^*} = q_{\pi_E}$;
- b $\pi_G^* = \tilde{\pi}^* \in \arg\min_{\pi_G} L(\tilde{\pi}^*, \pi_G)$.

We will give the corresponding proofs of (a) and (b) below.

Method III

- By expanding Eq. (1), we have that

$$\begin{aligned}\arg\max_{\tilde{\pi}} L(\tilde{\pi}, \pi_G) &= \arg\max_{\tilde{\pi}} \sum_{\tau_i} P_{\pi_E}(\tau_i) \log D_{\tilde{\pi},\pi_G}(\tau_i) + P_{\pi_G}(\tau_i) \log(1 - D_{\tilde{\pi},\pi_G}(\tau_i)) \\ &= \arg\max_{\tilde{\pi}} \sum_{\tau_i} \xi(\tau_i) \left(q_{\pi_E}(\tau_i) \log D_{\tilde{\pi},\pi_G}(\tau_i) + q_{\pi_G}(\tau_i) \log(1 - D_{\tilde{\pi},\pi_G}(\tau_i))\right) \\ &= \arg\max_{\tilde{\pi}} \sum_{\tau_i} L_i(\tau_i).\end{aligned}$$

From Proposition 1 of (Goodfellow et al., 2014), we get

$$D^*_{\tilde{\pi},\pi_G} = \arg\max_{D_{\tilde{\pi},\pi_G}} L_i(\tau_i) = \frac{q_{\pi_E}(\tau_i)}{q_{\pi_E}(\tau_i) + q_{\pi_G}(\tau_i)}.$$

Thus from the definition and monotonicity of $D^*_{\tilde{\pi},\pi_G}$, we get $q_{\tilde{\pi}^*} = q_{\pi_E}$. Then by using Lemma 1, we conclude that $\tilde{\pi}^* = \pi_E$.

Method IV

ⓑ we use the conclusion from (a), and get

$$\begin{aligned}\pi_G^* &= \arg\min_{\pi_G} L(\tilde{\pi}^*, \pi_G) \\ &= \arg\min_{\pi_G} \mathbb{E}_{\tau\sim P_{\pi_E}}\left[\log \frac{P_{\pi_E}(\tau)}{P_{\pi_E}(\tau)+P_{\pi_G}(\tau)}\right] + \mathbb{E}_{\tau\sim P_{\pi_G}}\left[\log \frac{P_{\pi_G}(\tau)}{P_{\pi_E}(\tau)+P_{\pi_G}(\tau)}\right] \\ &= \arg\min_{\pi_G} -\log 4 + 2D_{\mathsf{JS}}(P_{\pi_E}\|P_{\pi_G}) \\ &= \pi_E.\end{aligned}$$

Conclude the proof.

Method V

□

From Theorem 1, we find that optimizing the inner $\tilde{\pi}$ will give us the optimal outer π_G. We thus get ASAF, a new imitation learning algorithm with pseudocode shown in Algorithm 1.

Algorithm 1: ASAF

```
Input: expert trajectories $\mathcal{D}_E = \{\tau_i^{(E)}\}_{i=1}^{N_E}$
randomly initialize $\tilde{\pi}$ and set $\pi_G \leftarrow \tilde{\pi}$;
for steps $m = 0$ to $M$ do
    Collect trajectories $\mathcal{D}_G = \{\tau_i^{(G)}\}_{i=1}^{N_G}$ using $\pi_G$ ;
    // ASAF Update
    Update $\tilde{\pi}$ by minimizing Eq. (2) ;
    $\pi_G \leftarrow \tilde{\pi}$
end
```

$$\mathcal{L}_{BCE}(\mathcal{D}_E, \mathcal{D}_G, \tilde{\pi}) \approx -\frac{1}{n_E}\sum_{i=1}^{n_E} \log D_{\tilde{\pi},\pi_G}(\tau_i^{(E)}) - \frac{1}{n_G}\sum_{i=1}^{n_G} \log\left(1 - D_{\tilde{\pi},\pi_G}(\tau_i^{(G)})\right), \qquad (2)$$

Method VI

where $\tau_i^{(E)} \sim \mathcal{D}_E$, $\tau_i^{(G)} \sim \mathcal{D}_G$ and

$$D_{\tilde{\pi},\pi_G}(\tau) = \frac{\prod_{i=1}^{T-1} \tilde{\pi}(a_t|s_t)}{\prod_{i=1}^{T-1} \tilde{\pi}(a_t|s_t) + \prod_{i=1}^{T-1} \pi_G(a_t|s_t)}. \tag{3}$$

Remark: ASAF uses the full-length trajectory, which may be sample-inefficient. The authors also considered using a window size of w or even transition-wise variants, and named them ASAF-w and ASAF-1, respectively.

Table of Contents

Preliminaries

Motivation

Method

Experiments

Discussion

Experiments I

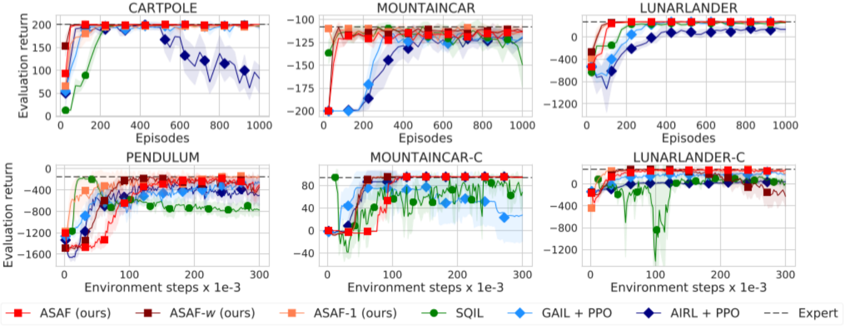

Figure 1: Results on classic control and Box2D tasks for 10 expert demonstrations. First row contains discrete actions environments, second row corresponds to continuous control.

Experiments II

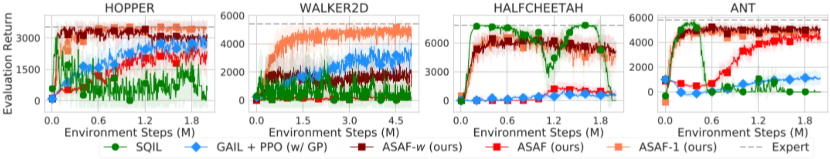

Figure 2: Results on MuJoCo tasks for 25 expert demonstrations.

Table of Contents

Preliminaries

Motivation

Method

Experiments

Discussion

Discussion

- Model learning

$$D_{\tilde{\mathcal{P}},\mathcal{P}_G}(\tau) = \frac{\tilde{\mathcal{P}}(s_0)\prod_{i=1}^{T-1}\tilde{\mathcal{P}}(a_t|s_t)}{\prod_{i=1}^{T-1}\tilde{\mathcal{P}}(a_t|s_t) + \prod_{i=1}^{T-1}\mathcal{P}_G(a_t|s_t)}.$$

- Reduction to transition-wise scenario: The author also discussed a novel transition-wise objective (connected to (Fu et al., 2017)), which is only suit for discrete action space.
- We can use any policy to sample fake trajectories.

The End

Thanks!

Q & A

References

[1] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *CoRR*, abs/1710.11248, 2017.

[2] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[3] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4565–4573, Barcelona, Spain, 2016.

[4] Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *International conference on Machine learning*, pp. 1032–1039, 2008.