A Note on Generative Adversarial Imitation LearningHo and Ermon [2016]

L. C. Huang
Department of Computer Science
National Yang Ming Chiao Tung University
libao3128.cs08@nycu.edu.tw

1 Introduction

Since the conservation method to learn from expert behavior like inverse reinforcement learning or behavior cloning can be inefficient and indirect. The paper proposed a new general framework for directly extracting the policy from expert demonstration. They combine a model-free imitation learning algorithm and generative adversarial networks into generative adversarial imitation learning framework, which has outstanding performance gains over conservation model-free method in complex environment. Please provide a clear but brief overview of the selected paper. In my personal opinion, this framework can surely reduce the time and resource usage while learning from the expert.

2 Problem Formulation

2.1 Preliminaries

$\bar{\mathbb{R}}$ will denote the extended real numbers $\mathbb{R} \cup \{\infty\}$. S and A denote the finite state and action space. $\prod$ is the set of all stationary stochastic policies that take actions in A given states in S; successor states are drawn from the dynamics model $P(s'|s, a)$.

2.2 Inverse reinforcement learningZiebart et al. [2008]

Suppose we are given expert policy π_E that used to train an agent policy with IRL. The optimization problem is set as

$$\min_\pi \max_{c \in C} \mathbb{E}_\pi[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)]$$

where c is the cost function.

According to the theory of maximum entropy RL, we can expand this problem by a convex function ψ and discounted entropy $H(\pi)$. The original IRL for formula will become: $IRL_\psi(\pi_E) = argmax_{c \in \mathbb{R}^{S \times A}} - \psi(c) + (\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[c(s,a)]) - \mathbb{E}_{\pi_E}[c(s,a)]$ where $H(\pi) = \mathbb{E}_\pi[-\log \pi(a|s)]$ is the γ-discounted causal entropy of the policy π.

2.3 Occupancy Measure

Define the occupancy measure $\rho(s, a)$

$$\rho_\pi : S \times A \to \mathbb{R} \text{ as } \rho_\pi(s,a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi)$$

, which represent the time-discounted probability of state action pair (s, a) happened. According to the past research(Theorem 2 of Syed et al. [29]), the policy can be recovered from $\pi_\rho(a|s) = \rho(s,a)/\sum_{a'} \rho(s,a')$, and π_ρ is the only policy whose occupancy is ρ.

2.4 Convex Conjugate

For a function $f : \mathbb{R}^{S\times A} \to \bar{\mathbb{R}}$, its convex conjugate $f^* : \mathbb{R}^{S\times A} \to \bar{\mathbb{R}}$

2.5 Problem Setting

2.5.1 Theorem 1

$$RL \circ IRL_\psi(\pi_E) = argmin_{\pi\in\prod} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E})$$

2.5.2 Theorem 2

If psi is a constant function, $\tilde{c} \in IRL_\psi(\pi_E)$, and $\tilde{\pi} \in RL(\tilde{c})$, then $\rho_{\tilde{\pi}} = \rho_{\pi_E}$

2.5.3 Theorem 3

With certain settings of ψ, **Theorem 1 and Theorem 2** takes on the form of regularized variants of existing ***apprenticeship learning*** algorithms.

2.5.4 Theorem 4

Generative adversarial imitation learning algorithm can be derived from **Theorem 1, Theorem 2 and Theorem 3**.

3 Theoretical Analysis

3.1 Theorem 1: $RL \circ IRL_\psi(\pi_E) = argmin_{\pi\in\Pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E})$

3.1.1 Lamma 1.1

Let $\bar{H}(\rho) = -\sum_{s,a} \rho(s,a) \log(\rho(s,a)/\sum_{a'} \rho(s,a'))$. Then $\bar{H}(\rho)$ is strictly concave. And for all $\pi \in \Pi$ and $\rho \in D$, we have $H(\rho_\pi) = \bar{H}(\rho_\pi)$ and $\bar{H}(\rho) = H(\pi_\rho)$

Proof of $\bar{H}$ is concave.

$-(\lambda\rho(s,a) + (1-\lambda)\rho'(s,a))log\frac{\lambda\rho(s,a)+(1-\lambda)\rho'(s,a)}{(\lambda\rho(s,a')+(1-\lambda)\rho'(s,a'))}$

$= -(\lambda\rho(s,a) + (1-\lambda)\rho'(s,a)) \log \frac{\lambda\rho(s,a)+(1-\lambda)\rho'(s,a)}{\rho\sum_{a'}\rho(s,a')+(1-\lambda)\sum_{a'}\rho'(s,a')}$

Applying log sum inequality $\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b}$

$\geq -\lambda\rho(s,a) \log \frac{\lambda\rho(s,a)}{\lambda\sum_{a'}\rho(s,a')} - (1-\lambda)\rho'(s,a) \log \frac{(1-\lambda)\rho'(s,a')}{(1-\lambda)\sum_{a'}\rho'(s,a')}$

$= \lambda(-\rho(s,a) \log \frac{\rho(s,a)}{\sum_{a'}\rho(s,a')}) + (1-\lambda)(-\rho'(s,a)log\frac{\rho'(s,a)}{\sum_{a'}\rho'(s,a')})$

with equality if and only if $\pi_\rho = \rho(s,a)/\sum_{a'}\rho(s,a') = \rho'(s,a)/\sum_{a'}\rho'(s,a') = \pi_{\rho'}$.

Proof of $H(\rho_\pi) = \bar{H}(\rho_\pi)$ and $\bar{H}(\rho) = H(\pi_\rho)$.

$$
\begin{aligned}
H(\pi) &= \mathbb{E}_\pi[-log\pi(a|s)] \\
&= -\sum_{s,a} \rho_\pi(s,a) \log \pi(a|s) \\
&= -\sum_{s,a} \rho_\pi(s,a) \log \frac{\rho_\pi(s,a)}{\sum_{a'} \rho_\pi(s,a')} \\
&= \bar{H}(\rho_\pi) \\
\bar{H}(\rho) &= -\sum_{s,a} \rho(s,a) \log \frac{\rho(s,a)}{\sum_{a'} \rho(s,a')} \\
&= -\sum_{s,a} \rho_{\pi_\rho}(s,a) \log \pi_\rho(a|s) \\
&= \mathbb{E}_{\pi_\rho}[-\log \pi_\rho(a|s)] \\
&= H(\pi_\rho)
\end{aligned}
$$

Lemma 1.1 is proved.

3.1.2 Lemma 1.2.

If $L(\pi, c) = -H(\pi) + \mathbb{E}_\pi[c(s,a)]$ and $\bar{L}(\rho, c) = -\bar{H}(\rho) + \sum_{s,a} \rho(s,a)c(s,a)$, then, for all cost functions c, $L(\pi, c) = \bar{L}(\rho\pi, c)$ for all policies $\pi \in \Pi$, and $\bar{L}(\rho, c) = L(\pi\rho, c)$ for all occupancy measure $\rho \in D$.

Define $\bar{L}(\rho, c) = -\bar{H}(\rho) + \sum_{s,a} c(s,a)(\rho(s,a) - \rho_E(s,a)) - \psi$(c), where psi is a constant function. Known that the $-\bar{H}(\rho)$ is convex(through **Lemma 1**) and the summation part is convex and concave. We can have a conclusion:

$$
\begin{aligned}
&\bar{L}(\rho, \bullet) \text{ is concave and} \\
&\bar{L}(\bullet, c) \text{ is convex}
\end{aligned}
$$

Let $\pi_A \in argmin_\pi - H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E} = argmin_\pi max_c - H(\pi) - \psi(c) + \sum_{s,a}(\rho_\pi(s,a) - \rho_{\pi_E}(s,a))c(s,a)$ and $\tilde{c} \in IRL_\psi(\pi_E), \tilde{\pi} \in RL(\tilde{c}) = RL \circ IRL_\psi(\pi_E)$. Through **Theorem 1** definition and the lemma above:

$$
\begin{aligned}
\rho_A &\in argmin_{\rho \in D} max_c \bar{L}(\rho, c), \\
\tilde{c} &\in argmax_c min_{\rho \in D} \bar{L}(\rho, c), \\
\tilde{\rho} &\in argmin_{\rho \in D} \bar{L}(\rho, \tilde{c}).
\end{aligned}
$$

By applying duality optimum:

$$
min_{\rho \in D} max_{c \in C} \bar{L}(\rho, c) = max_{c \in C} min_{\rho \in D} \bar{L}(\rho, c)
$$

, we know that both $(\rho_A, \tilde{c})$ and $(\tilde{\rho}, \tilde{c})$ are the saddle points of function $\tilde{L}$. Due to the concave property, the saddle point of function $\tilde{L}$ is unique. Therefore, $\rho_A = \tilde{\rho}$ and the theorem is proved.

3.2 Theorem 2

If ψ is a constant function, $\tilde{c} \in IRL_\psi(\pi_E)$, and $\tilde{\pi} \in RL(\tilde{c})$, then $\rho_{\tilde{\pi}} = \rho_{\pi_E}$ Let $\bar{L}(\rho, c) = -\bar{H}(\rho) + \sum_{s,a} c(s,a)(\rho(s,a) - \rho_E(s,a))$, we have:

$$
\begin{aligned}
&\tilde{c} \in IRL_\psi(\pi_E) \\
&\quad = argmax_{c \in \mathbb{R}^{S \times A}} min_{\pi \in \Pi} - H(\pi) + \mathbb{E}_\pi[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)] + const \\
&\quad = argmax_{c \in \mathbb{R}^{S \times A}} min_{\rho \in D} - \bar{H}(\rho) + \sum_{s,a} \rho(s,a)c(s,a) - \sum_{s,a} \rho_E(s,a)c(s,a) \\
&\quad = argmax_{c \in \mathbb{R}^{S \times A}} min_{\rho \in D} \bar{L}(\rho, c)
\end{aligned}
$$

The optimization problem above is the duality of

$$minimize_{\rho \in D} - \bar{H}(\rho) \text{ subject to } \rho(s,a) = \rho_E(s,a) \forall s \in S, a \in A$$

and the corresponding Lagrange multiplier is $c(s,a)$. The original problem can be solved as $\tilde{\rho} = argmin_{\rho \in D} \bar{L}(\rho, c) = \rho_E$.

3.3 Theorem 3: Apprenticeship Learning

The original apprenticeship learning optimization problem can be wrote as:

$$minimize_{\pi} - H(\pi) + max_{c \in C} \mathbb{E}_{\pi}[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)]$$

Let:

$$\delta_c : \mathbb{R}^{S \times A} \rightarrow \bar{\mathbb{R}}$$
$$\delta_C(C) = 0 \text{ if} \in C, otherwise + \infty$$

We have:

$$\max_{c \in C} \mathbb{E}_{\pi}[c(s,a)] - \max_{c \in C} \mathbb{E}_{\pi_E}[c(s,a)] =$$
$$\max_{c \in \mathbb{R}^{S \times A}} -\delta_C(c) + \sum s, a(\rho_p i(s,a) - \rho_{\pi_E}(s,a))c(s,a) = \delta_C^*(\rho_\pi - \rho_{\pi_E})$$

Therefore, we see that entropy regularized apprenticeship learning is equivalent to performing RL following IRL with cost regularize $\psi = \delta_C$.

$$minimize_{\pi} - H(\pi) + max_{c \in C} \mathbb{E}_{\pi}[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)] = -H(\pi) + \delta_c^*$$

Note that we can scale the policy's entropy regularization strength in equation above by scaling C by a constant α as $\alpha c : c \in C$, recovering the original apprenticeship objective by taking $\alpha \rightarrow \infty$.

3.4 Theorem 4: Generative adversarial imitation learning

Let:

$$g_\phi(x) = \begin{cases} -x + \phi(-\phi^{-1}(-x)) & x \in T \\ +\infty & otherwise \end{cases}$$

$$\psi_\phi(c) = \begin{cases} \sum_{s,a} \rho_{\pi_E}(s,a) g_\phi(c(s,a)) & c(s,a) \in T \forall s, a \\ +\infty & otherwise \end{cases}$$

where ϕ can be any strictly decreasing and convex function. In the paper, $\log(1 + \exp^{-x})$ is used. Let $R_\phi(\pi, \pi_E) = \sum \min_{\gamma \in \mathbb{R}} \rho_\pi(s,a)\phi(\gamma) + \rho_{\pi_E}(s,a)\phi(-\gamma)$, we have:

$$\begin{aligned}
\psi_\phi^*(\pi - \rho_{\pi_E} &= \max_{c \in C} \sum_{s,a} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a))c(s,a) - \sum_{s,a} \rho_{\pi_E}(s,a) g_\phi(c(s,a)) \\
&= \sum_{s,a} \max_{c \in T} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a))c - \rho_{\pi_E}(s,a)[-c + \phi(-\phi^{-1}(-c))] \\
&= \sum_{s,a} \max_{c \in T} (\rho_\pi(s,a) - \rho_{\pi_E}(s,a)\phi(-\phi^{-1}(-c)) \\
&= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a)(-\phi(\gamma)) - \rho_{\pi_E}(s,a)\phi(-\phi^{-1}(\phi(\gamma))) \\
&= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a)(-\phi(\gamma)) - \rho_{\pi_E}(s,a)\phi(-gamma) \\
&= -R_\phi(\rho_\pi, \rho_{\pi_E})
\end{aligned}$$

While $\phi(x) = \log(1 + \exp^{-x})$, we have:

$$
\begin{aligned}
\psi^*_{GA}(\rho_\pi - \rho_{\pi_E}) &= -R_\psi(\rho_\pi, \rho_{\pi_E}) \\
&= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_p i(s,a) \log(\frac{1}{1+\exp^{-\gamma}}) + \rho_{\pi_E}(s,a) \log(\frac{1}{1+\exp^{\gamma}} \\
&= \sum_{s,a} = \max \gamma \in \mathbb{R} \rho_p i(s,a) \log(\frac{1}{1+\exp^{-\gamma}} + \rho_{\pi_E}(s,a) \log(1 - \frac{1}{1+\exp^{-\gamma}} \\
&= \sum_{s,a} \max_{\gamma \in \mathbb{R}} \rho_\pi(s,a) \log(\sigma(\gamma)) + \rho_{\pi_E}(s,a) \log(1 - \sigma(\gamma)) \\
&= \sum_{s,a} \max_{d \in (0,1} \rho_\pi(s,a) \log d + \rho_{\pi_E}(s,a) \log(1-d) \\
&= \max D \in (0,1)^{S \times A} \sum_{s,a} \rho_\pi(s,a) \log(D(s,a)) + \rho_{\pi_E}(s,a) \log(1 - D(s,a))
\end{aligned}
$$

Finally, we have the optimization problem of GAIL algorithm:

$$
minimize_\pi \psi^*_{GA}(\rho_\pi - \rho_{\pi_E}) - \lambda H(\pi) = D_{JS}(\rho_\pi, \rho_{\pi_E}) - \lambda H(\pi)
$$

The pseudo code of the algorithm looked like:

Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories τ_E π_E, initial policy and discriminator parameters θ_0, ω_0
2: **for** $i = 0, 1, 2, ..$ **do**
3: Sample trajectories τ_i π_{θ_i}
4: Update the discriminator parameters from ω_i to ω_{i+1} with the gradient $\hat{\mathbb{E}}_{\tau_i}[\delta_\omega \log(D_\omega(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\delta_\omega \log(1 - D_\omega(s,a))]$
5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_\omega i + 1(s,a))$. Specifically, take a KL-constrained natural gradient step with $\hat{\mathbb{E}}_{\tau_i}[\delta_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \delta_\theta H(\pi_t heta a)$, where $Q(\bar{s}, \bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{\omega_{i+1}}(s,a) | s_0 = \bar{s}, a_0 = \bar{a}$
6: **end for**

4 Conclusion

GAIL algorithm is a very efficient and strong framework that the agent can learn from expert demonstration. There are many derivated research regarding GAIL. Some of them addressed some potential issue in GAIL framework like:

- Reward bias.Wang and Li [2021]
- Sample inefficient.Kostrikov et al. [2018]
- The agent cannot query for more information from the expert in an iterative manner.Jena et al. [2020]
- and so on...

In my personal study experience, GAIL framework will also encounter Nash equilibrium in multi-agent environment.Song et al. [2018] Therefore, although GAIL is an efficient to learn from the expert, there most be more further research to improve the performance.

References

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Yawei Wang and Xiu Li. Reward function shape exploration in adversarial imitation learning: an empirical study. In *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 52–57. IEEE, 2021.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *arXiv preprint arXiv:1809.02925*, 2018.

Rohit Jena, Changliu Liu, and Katia Sycara. Augmenting gail with bc for sample efficient imitation learning. *arXiv preprint arXiv:2001.07798*, 2020.

Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. *Advances in neural information processing systems*, 31, 2018.