Imitation Learning from Imperfect Demonstration

Yueh-Hua Wu[1,2], Nontawat Charoenphakdee[3,2], Han Bao[3,2],
Voot Tangkaratt[2], Masashi Sugiyama[2,3]

[1]National Taiwan University

[2]RIKEN Center for Advanced Intelligence Project

[3]The University of Tokyo

Poster #47

Introduction

- Imitation learning
  - learning from **demonstration** instead of a reward function
  - useful when reward function is **sparse** or **hard to specify**
- Collected demonstration may be **imperfect**
  - Driving: speeding, traffic violation
  - Playing basketball: turnovers, technical foul

Motivation

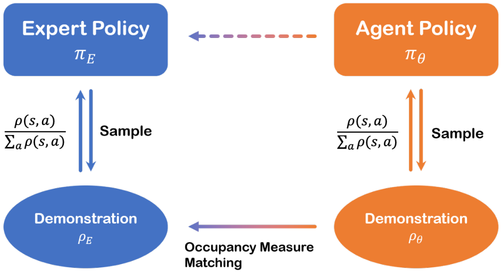
- A **semi-supervised** setting: demonstration **partially** equipped with **confidence**
- **Confidence**: a value between 0 and 1 that indicates the extent of a state-action pair (x) being optimal.
- How?
    - **crowdsourcing:** $N(1)/(N(1)+N(0))$. For example, $47/(47+53)=0.47$
    - **digitized score:** $0.0, 0.1, 0.2, \ldots, 1.0$
- **Robustness** to noisy labelers

Generative Adversarial Imitation Learning [1]

- Distribution of demonstration has a **one-to-one correspondence** with the policy [2]
- Utilize **generative adversarial training**

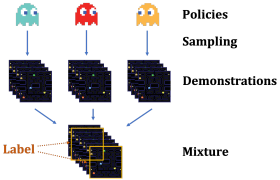
$$\min_{\theta} \max_{w} \mathbb{E}_{x \sim p_\theta}[\log D_w(x)] + \mathbb{E}_{x \sim p_{\mathrm{opt}}}[\log(1 - D_w(x))] \tag{1}$$

D_w: discriminator, p_{opt}: demonstration distribution of π_{opt}, and p_θ: trajectory distribution of agent π_θ



Problem Setting

Human switches to **non-optimal policies** when they **make mistakes** or **are distracted**



$$p(x) = \alpha \underbrace{p(x|y=+1)}_{p_{\mathrm{opt}}(x)} + (1-\alpha) \underbrace{p(x|y=-1)}_{p_{\mathrm{non}}(x)}$$

- **Confidence:** $r(x) \triangleq \Pr(y=+1|x)$
- **Unlabeled demonstration:** $\{x_i\}_{i=1}^{n_{\mathrm{u}}} \sim p$
- **Demonstration with confidence:** $\{(x_j, r_j)\}_{j=1}^{n_{\mathrm{c}}} \sim q$

Proposed Method 1: Two-Step Importance Weighting Imitation Learning

Step 1: **estimate confidence** by learning a confidence scoring function g

Step 2: employ **importance weighting** to reweight GAIL objective

- Unbiased risk estimator:

$$R_{\mathrm{SC},\ell}(g) = \underbrace{\mathbb{E}_{x,r\sim q}[r \cdot (\ell(g(x)))]}_{\text{Risk for optimal}} + \underbrace{\mathbb{E}_{x,r\sim q}[(1-r)\ell(-g(x))]}_{\text{Risk for non-optimal}}$$

- Importance weighting

$$\min_{\theta} \max_{w} \mathbb{E}_{x\sim p_\theta}[\log D_w(x)] + \mathbb{E}_{x\sim p}[\frac{\hat{r}(x)}{\alpha}\log(1-D_w(x))]$$

Theorem

For $\delta \in (0,1)$, with probability at least $1-\delta$ over repeated sampling of data for training $\hat{g}$,

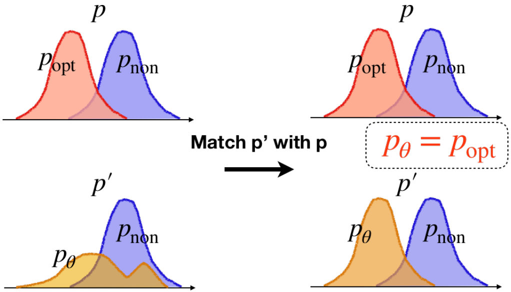
$$R_{\mathrm{SC},\ell}(\hat{g}) - R_{\mathrm{SC},\ell}(g^*) = \mathcal{O}_p(\underbrace{n_c^{-1/2}}_{\#\text{ of confidence}} + \underbrace{n_u^{-1/2}}_{\#\text{ of unlabeled}})$$

Proposed Method 2: GAIL with Imperfect Demonstration and Confidence

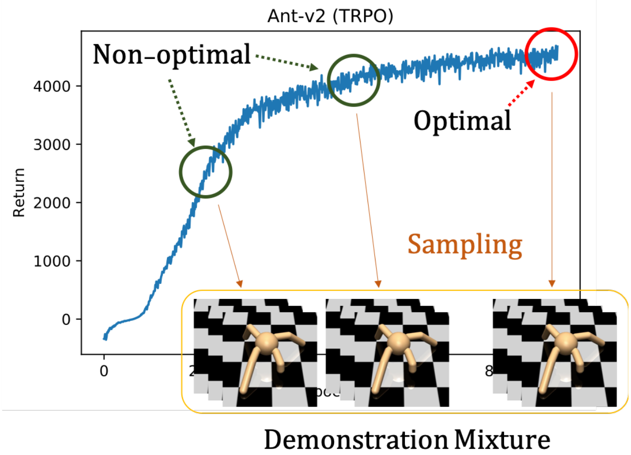
- Mixing **the agent demonstration** with **the non-optimal** one guarantees to learn the optimal policy
- Objective:

$$V(\theta, D_w) = \underbrace{\mathbb{E}_{x\sim p}[\log(1 - D_w(x))]}_{\text{Risk for P class}} + \underbrace{\alpha\mathbb{E}_{x\sim p_\theta}[\log D_w(x)] + \mathbb{E}_{x,r\sim q}[(1-r)\log D_w(x)]}_{\text{Risk for N class}}$$

- Matching p' with p enables $p_\theta = p_{\text{opt}}$ and meanwhile **benefits from the large amount of unlabeled data.**



Setup

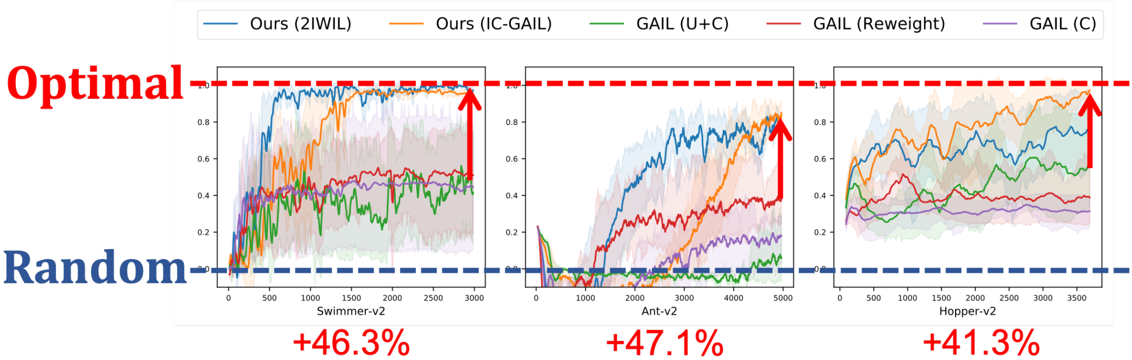


Confidence is given by a classifier trained with the demonstration mixture labeled as optimal ($y = +1$) and non-optimal ($y = -1$)

Results: Higher Average Return of the Proposed Methods

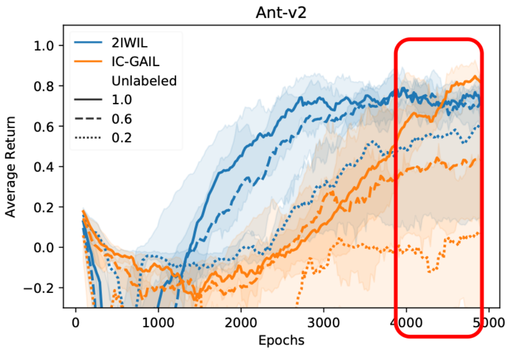
Environment: Mujoco

Proportion of labeled data: 20%

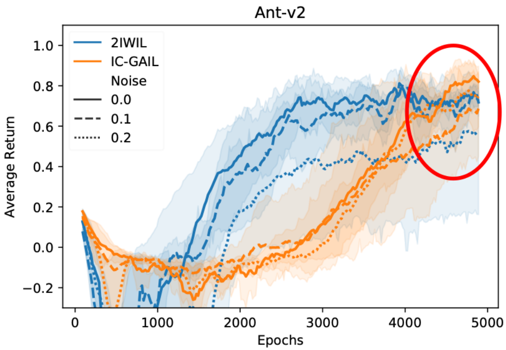


Results: Unlabeled Data Helps

- **More unlabeled data** results in **lower variance** and **better performance**
- proposed methods are **robust** to noise



(a) Number of unlabeled data. The number in the legend indicates **proportion** of orignal unlabeled data.



(b) Noise influence. The number in the legend indicates **standard deviation** of Gaussian noise.

Conclusion

- Two approaches that utilize **both unlabeled and confidence data** are proposed
- Our methods are **robust to labelers with noise**
- The proposed approaches can be generalized to other IL and IRL methods

Poster #47

Reference

[1] Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." Advances in Neural Information Processing Systems. 2016.

[2] Syed, Umar, Michael Bowling, and Robert E. Schapire. "Apprenticeship learning using linear programming." Proceedings of the 25th international conference on Machine learning. ACM, 2008.