Imitation Learning from Imperfect Demonstration

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

1: National Taiwan University, 2: The University of Tokyo, 3: RIKEN Center for Advanced Intelligence Project

Introduction

- **Imitation Learning:** Learn the decision making strategy (**policy**) of experts from perfect demonstration.
- Issues:
  - Perfect demonstrations are **costly** when the task is difficult.
  - Confidence scores can reweight the demonstration distribution to the optimal one but **labeling all demonstrations is also expensive**.

⇒ We consider demonstrations **partially equipped with confidence** and propose two approaches to learning an optimal policy with **theoretical guarantee**.

Background

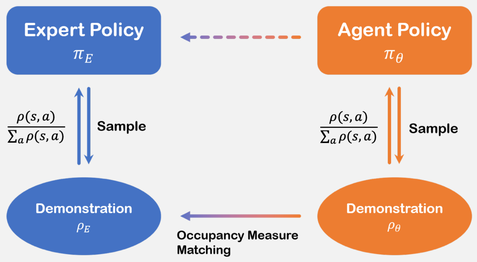
- **Policy (π):** decision making strategy given state
- **Demonstration:** a set of state-action pairs x

$$\left\{\left(\text{  }, \text{  }\right)\right\} \sim p_\pi(x)$$

- **Generative Adversarial Imitation Learning [1]:** Given perfect demonstration drawn from p_{opt}, utilize GAN structure to learn an optimal policy:

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

θ : parameters of an agent policy, D_w : discriminator.



- **Imperfect demonstration:** a **mixture** of optimal and non-optimal demonstrations with density

$$p(x) = \alpha p(x|y=+1) + (1-\alpha)p(x|y=-1)$$

$p(x|y=+1)$: $p_{\mathrm{opt}}(x)$, $p(x|y=-1)$: $p_{\mathrm{non}}(x)$, and α: $\Pr(y=+1)$.

- **Confidence score [2]:** $r(x) = \Pr(y=+1|x)$.
- **Confidence collection:**
  - crowdsourcing: $\frac{N(y=+1)}{N(y=+1)+N(y=-1)}$
  - digitized score: 0.0, 0.1, 0.2, . . . , 1.0

Problem Setting

A demonstration mixture partially labeled with confidence

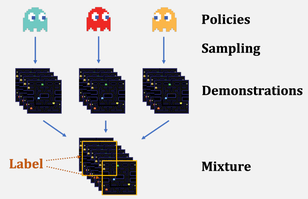
Human follows non-optimal policies when they make mistakes or are distracted.

Input:

Unlabeled demonstration $(s_i \to a_i) \sim p$

Demonstration with confidence $(s_i \to a_i, r(x_i)=0.7) \sim q$

Output: policy function



Two-Step Importance Weighting Imitation Learning

Step 1 **estimate confidence scores** for unlabeled demonstration by learning a confidence scoring function g. Unbiased risk estimator:

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

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

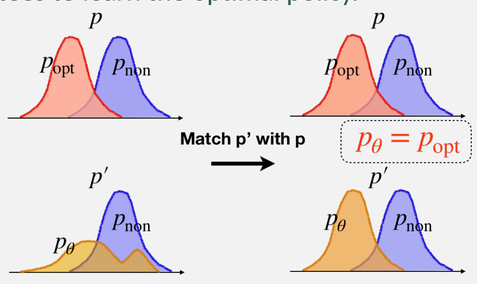
$$\min_\theta \max_w \mathbb{E}_{x\sim p_\theta}[\log D_w(x)] + \mathbb{E}_{x\sim p}\left[\frac{\hat{r}(x)}{\alpha}\log(1-D_w(x))\right]$$

Estimation error bound

$$\underbrace{R_{\mathrm{SC},\ell}(\hat{g}) - R_{\mathrm{SC},\ell}(g^*)}_{\text{estimation error of risk of empirical risk minimizer}} = O_p\left(\underbrace{n_{\mathrm{c}}^{-1/2}}_{\text{\# of conf}} + \underbrace{n_{\mathrm{u}}^{-1/2}}_{\text{\# of unlabeled}}\right)$$

GAIL with Imperfect Demonstration and Confidence

Mixing **the agent demonstration** (p_θ) with **the non-optimal one** (p_{non}) guarantees to learn the optimal policy.



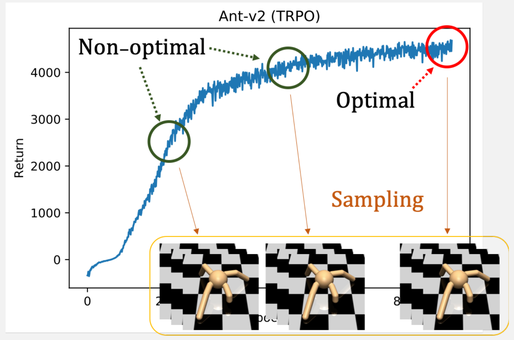
$$V(D_w) = \mathbb{E}_{x\sim p}[\log(1-D_w(x))] + \alpha\mathbb{E}_{x\sim p_\theta}[\log D_w(x)] + \mathbb{E}_{x,r\sim q}[(1-r)\log D_w(x)]$$

By optimizing $V(D_w)$, the discriminator recognizes p_θ and p_{opt} as the same class and p as the other.

With the same mixture weight α, p' is able to match p and meanwhile **benefit from the large amount of unlabeled data.**

Experiments

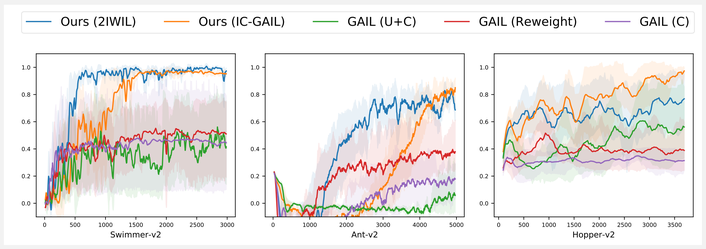
Data generation:



Demonstration Mixture

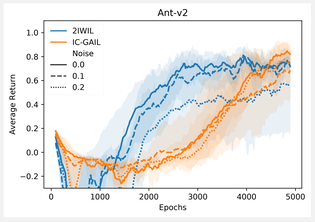
Confidence is given by a classifier trained with the demonstration mixture labeled as optimal ($y=+1$) and non-optimal ($y=-1$).

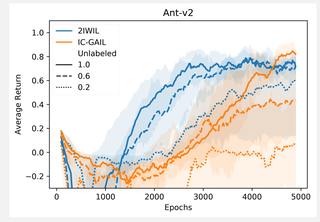
Results:



Robustness:

We also conduct experiment to investigate the influence of **noisy labelers** and **the number of unlabeled demonstration**





Our methods are **robust to noisy labelers** and unlabeled data plays an important role to learn a better policy.

References

[1] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NeurIPS*, pages 4565–4573, 2016.

[2] Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification from positive-confidence data. In *NeurIPS*, pages 5919–5930, 2018.