Imitation Learning from Imperfect Demonstration

Yueh-Hua Wu\textsuperscript{1,2}, Nontawat Charoenphakdee\textsuperscript{3,2}, Han Bao\textsuperscript{3,2}, Voot Tangkaratt\textsuperscript{2}, Masashi Sugiyama\textsuperscript{2,3}

\textsuperscript{1}National Taiwan University
\textsuperscript{2}RIKEN Center for Advanced Intelligence Project
\textsuperscript{3}The University of Tokyo

Poster \#47
Imitation learning
- learning from demonstration instead of a reward function

Demonstration
- a set of decision makings (state-action pairs $x$)

Collected demonstration may be imperfect
- Driving: traffic violation
- Playing basketball: technical foul
Motivation

- **Confidence**: how optimal is state-action pair $x$ (between 0 and 1)
- A **semi-supervised** setting: demonstration **partially** equipped with confidence
- How?
  - **crowdsourcing**: $N(1)/(N(1) + N(0))$.
  - **digitized score**: 0.0, 0.1, 0.2, ..., 1.0
Generative Adversarial Imitation Learning [1]

- **One-to-one correspondence** between the policy $\pi$ and the distribution of demonstration [2]
- Utilize **generative adversarial training**

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

$D_w$: discriminator, $p_{opt}$: demonstration distribution of $\pi_{opt}$, and $p_{\theta}$: trajectory distribution of agent $\pi_{\theta}$
Problem Setting

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

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

- **Confidence**: \( r(x) \triangleq \text{Pr}(y = +1|x) \)
- **Unlabeled demonstration**: \( \{x_i\}_{i=1}^{n_u} \sim p \)
- **Demonstration with confidence**: \( \{(x_j, r_j)\}_{j=1}^{n_c} \sim q \)
Proposed Method 1: Two-Step Importance Weighting Imitation Learning

Step 1: **estimate confidence** by learning a confidence scoring function $g$
- Unbiased risk estimator (come to Poster #47 for details):

$$R_{SC,\ell}(g) = \mathbb{E}_{x, r \sim q}[r \cdot (\ell(g(x)))] + \mathbb{E}_{x, r \sim q}[(1 - r)\ell(-g(x))]$$

- Risk for optimal
- Risk for non-optimal

**Theorem**

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

$$R_{SC,\ell}(\hat{g}) - R_{SC,\ell}(g^*) = O_p\left(\begin{array}{c}n_c^{-1/2} \\ \text{# of confidence}\end{array}\right) + \begin{array}{c}n_u^{-1/2} \\ \text{# of unlabeled}\end{array}$$

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

$$\min_{\theta} \max_w \mathbb{E}_{x \sim p_\theta}[\log D_w(x)] + \mathbb{E}_{x \sim p}\left[\frac{\hat{f}(x)}{\alpha} \log(1 - D_w(x))\right]$$
**Proposed Method 2: GAIL with Imperfect Demonstration and Confidence**

- Mix **the agent demonstration** with **the non-optimal one**

\[
p' = \alpha p_\theta + (1 - \alpha)p_{\text{non}}
\]

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

- **Objective:**

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

- **Risk for P class**

\[
\mathbb{E}_{x \sim p}[\log(1 - D_w(x))]
\]

- **Risk for N class**

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

- **Diagram:**

Match \(p'\) with \(p\) where \(p_\theta = p_{\text{opt}}\).
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%

Optimal

Random

+46.3%
+47.1%
+41.3%
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 original unlabeled data.

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