ICLR

PN-GAIL: Leveraging Non-optimal Information from Imperfect Demonstrations

Qiang Liu, Huiqiao Fu, Kaiqiang Tang, Daoyi Dong, Chunlin Chen

南京大學

Robotics & Reinforcement Learning Control

Motivation

Problem: Generative Adversarial Imitation Learning (GAIL) tends to **fail** when faced with data filled with **imperfect demonstrations**.

2IWIL Solution: 2IWIL reweights imitation learning based on **confidence**, and assigns **higher weights** to demonstrations with **higher confidence**, so as to prioritize learning of **high-quality demonstrations**.

Motivation

However, it is worth noting that this weighting behavior can be influenced by the **preferences inherent in** imperfect demonstrations.

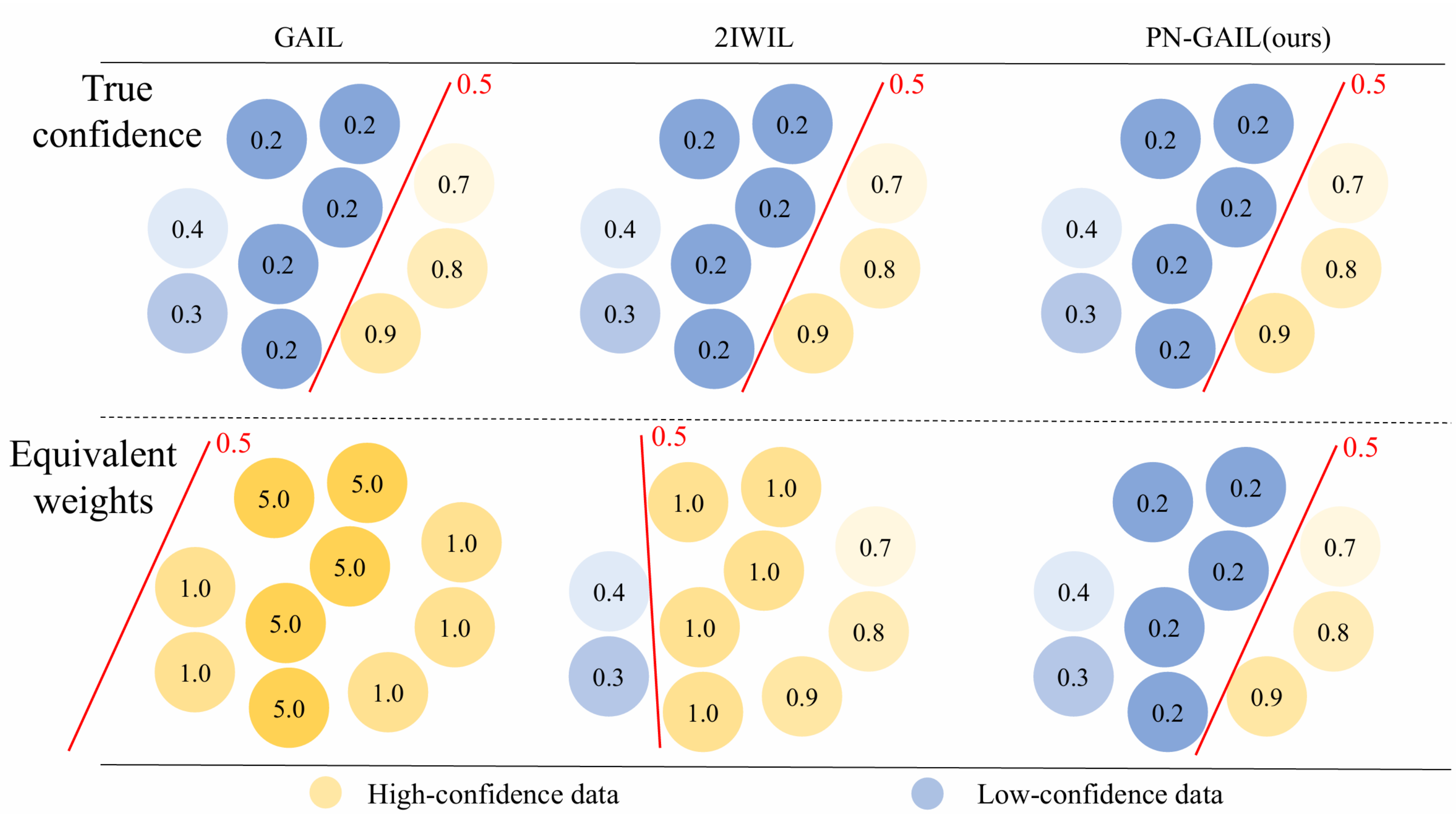
$$\text{GAIL}: \min_{\theta} \max_{w} \mathbb{E}_{x \sim p_\theta} [\log D_w(x)] + \mathbb{E}_{x \sim p} [\log(1 - D_w(x))]$$

$$\text{2IWIL}: \min_{\theta} \max_{w} \mathbb{E}_{x \sim p_\theta} [\log D_w(x)] + \mathbb{E}_{x \sim p} \left[\frac{r(x)}{\eta} \log(1 - D_w(x)) \right]$$

Expand the second expectation, that is:

$\sum \boldsymbol{p}(\boldsymbol{x}) \log(1 - D_w(x))$、 $\sum \boldsymbol{p}(\boldsymbol{x}) \frac{r(x)}{\eta} \log(1 - D_w(x))$

Motivation



According to Eq 3, the equivalent weight of x_1 will be **1.0** compared to others ($0.2 \times 5 = 1.0$). This means that the discriminator will consider x_1 to be **more likely the optimal demonstration than others** !

Method

PN-GAIL leverages **non-optimal information** from imperfect demonstrations, allowing the discriminator to comprehensively assess the positive and negative risks associated with these demonstrations.

We begin by focusing on the training of the **discriminator**:

$$R^{pn}_{D_w}(\mathcal{D}_{\pi_\theta}, \mathcal{D}) = R^{1}_{D_w}(\mathcal{D}_{\pi_\theta}) + R^{pn}_{D_w}(\mathcal{D})$$

$$R^{pn}_{D_w}(\mathcal{D}) = R^{pn}_{D_w}(\mathcal{D}_{\text{opt}}, \mathcal{D}_{\text{non}}) = \eta R^{0}_{D_w}(\mathcal{D}_{\text{opt}}) + (1-\eta) R^{1}_{D_w}(\mathcal{D}_{\text{non}})$$

Method

The overall risk of the discriminator can be rewritten as:

$$R^{pn}_{D_w}(\mathcal{D}, \mathcal{D}_{\pi_\theta}) = R^{1}_{D_w}(\mathcal{D}_{\pi_\theta}) + \eta R^{0}_{D_w}(\mathcal{D}_{\text{opt}}) + (1-\eta) R^{1}_{D_w}(\mathcal{D}_{\text{non}})$$

Replacing the loss function with the standard logistic loss and tidying up the statement, the objective of the discriminator becomes:

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

Method

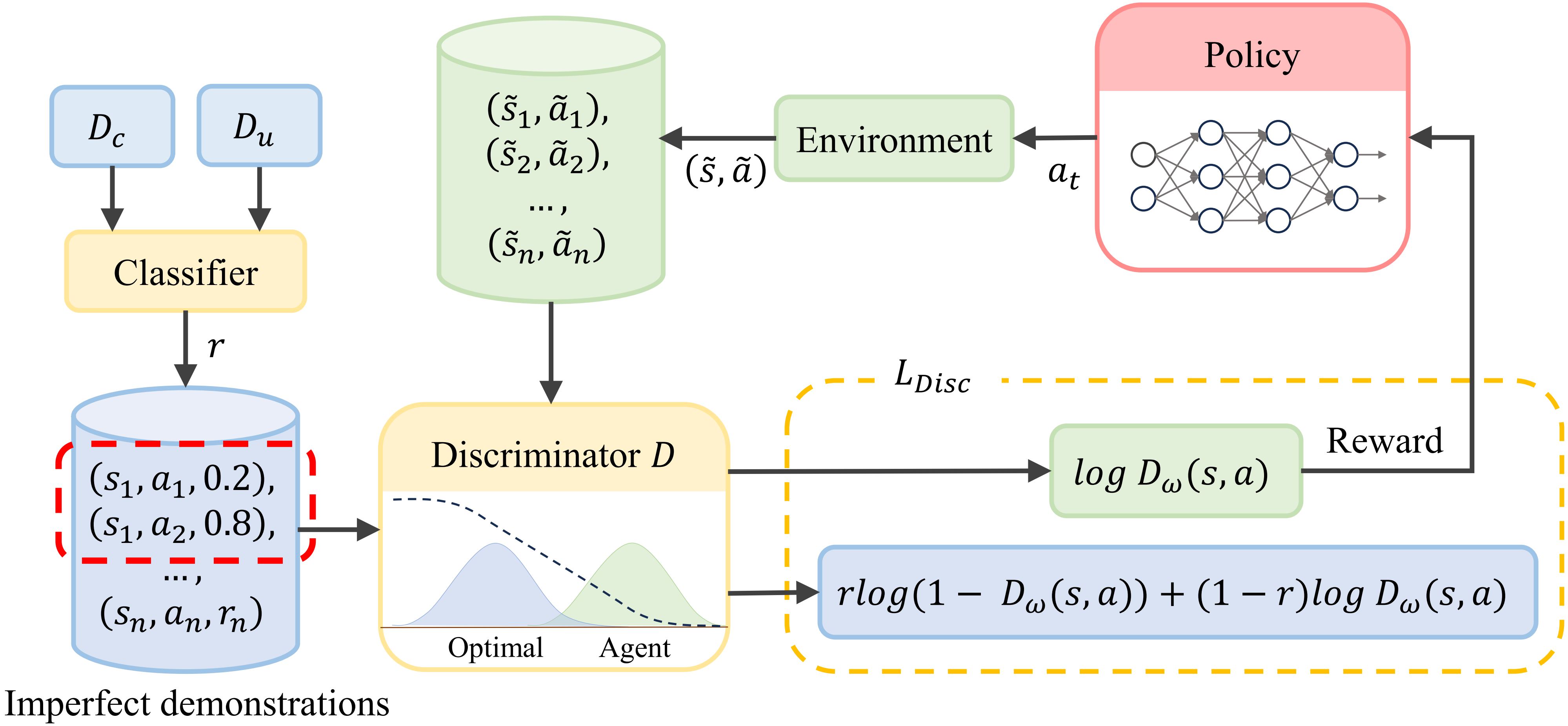
To get **more accurate** confidence scores, we refine the semi-conf (SC) classification proposed in 2IWIL, which is trained by minimizing the following risk:

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

We propose balanced semi-conf (BSC) classification. We introduce $\mathrm{E}_{x\sim p}[\alpha\ell(g(x)) - \mathrm{E}_{x\sim q}[\alpha\ell(g(x))$, the theoretical value of which is **0**.
And the final risk is as follows:

$$R_{\mathrm{BSC},\ell}(g) = \mathbb{E}_{x,r\sim q}\left[r\ell(g(x)) + (1-r)\ell(-g(x)) - \alpha\ell(g(x)) - \beta\ell(-g(x))\right] + \mathbb{E}_{x\sim p}[\alpha\ell(g(x)) + \beta\ell(-g(x))]$$

Method



D_c
D_u
Classifier
r
$(s_1, a_1, 0.2)$,
$(s_1, a_2, 0.8)$,
…,
(s_n, a_n, r_n)
Imperfect demonstrations
$(\tilde{s}_1, \tilde{a}_1)$,
$(\tilde{s}_2, \tilde{a}_2)$,
…,
$(\tilde{s}_n, \tilde{a}_n)$
$(\tilde{s}, \tilde{a})$
Environment
a_t
Policy
Discriminator D
Optimal
Agent
L_{Disc}
$log\ D_\omega(s, a)$
Reward
$rlog(1 - D_\omega(s, a)) + (1 - r)log\ D_\omega(s, a)$

Experiments

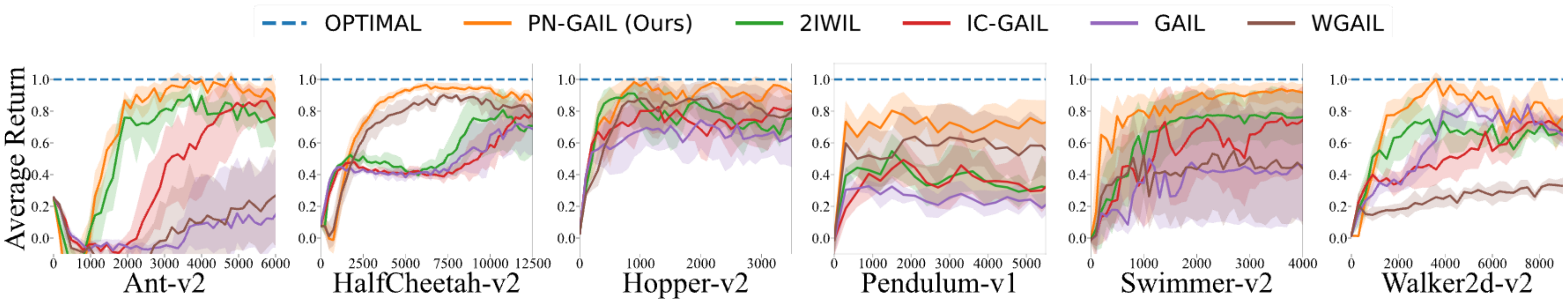
We validate PN-GAIL by conducting experiments on six control tasks, including Pendulum-v1 and five challenging MuJoCo environments.

We aim to answer three questions:

- Is 2IWIL influenced by the preferences **inherent in** imperfect demonstrations, and can our method **alleviate such influence**?
- Does our proposed BSC **outperform** the SC proposed in 2IWIL?
- How **robust** is our method?

Experiments

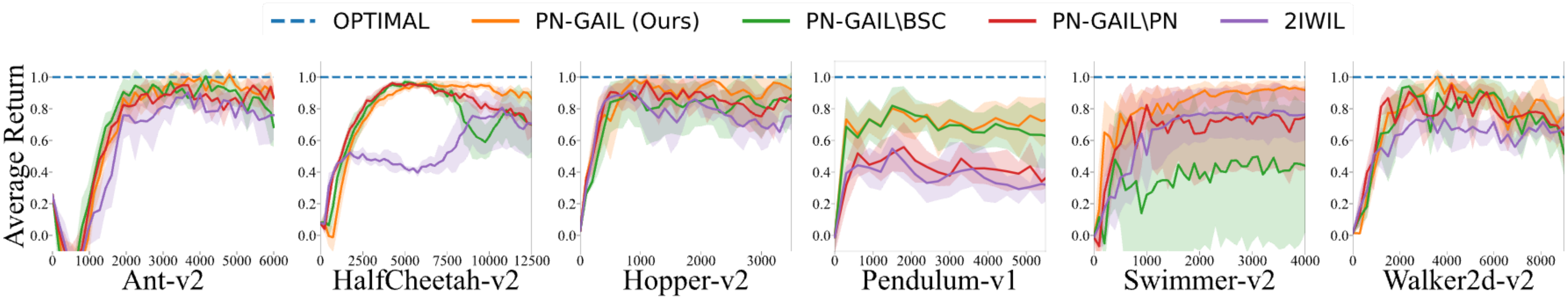
Performance



The results of the experiments with other baseline methods have shown that PN-GAIL is **superior to** other methods when dealing with imperfect demonstrations.

Experiments

Performance



The results of the ablation experiments have shown that both of the proposed improvements **contribute to** the final performance.

Experiments

Accuracy of classifier

Table 1: Accuracy of classifier measured by MAE and RMSE.

Classifier	Metrics	Ant-v2	HalfCheetah-v2	Hopper-v2	Pendulum-v1	Swimmer-v2	Walker2d-v2
SC	MAE	0.213 ± 0.023	0.184 ± 0.011	0.307 ± 0.025	0.126 ± 0.014	0.362 ± 0.049	0.132 ± 0.015
	RMSE	0.345 ± 0.033	0.272 ± 0.009	0.519 ± 0.022	0.164 ± 0.013	0.595 ± 0.040	0.246 ± 0.032
BSC	MAE	**0.056 ± 0.011**	**0.057 ± 0.012**	**0.169 ± 0.126**	**0.097 ± 0.006**	**0.286 ± 0.179**	**0.014 ± 0.002**
	RMSE	**0.212 ± 0.026**	**0.175 ± 0.013**	**0.371 ± 0.138**	**0.138 ± 0.005**	**0.472 ± 0.188**	**0.101 ± 0.010**

The MAE and RMSE of the BSC classifier are **notably lower** than those of the SC classifier, indicating that the predictions of the BSC classifier are **closer to the ground truth**.

Experiments

Robustness of PN-GAIL

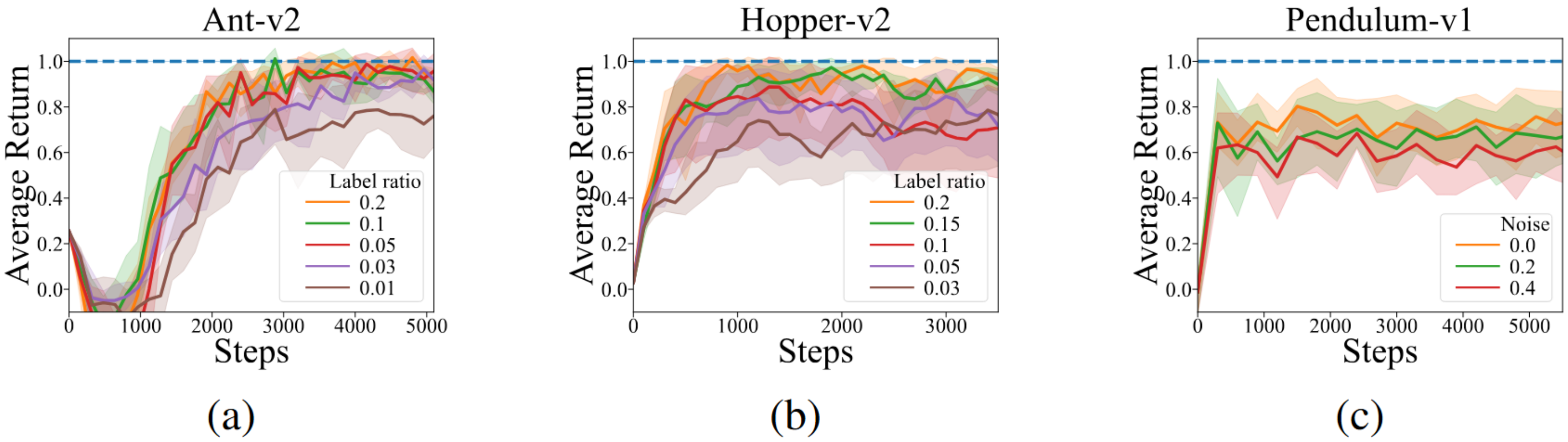


Figure 4: (a) Ant-v2 experiments with different label ratios. (b) Hopper-v2 experiments with different label ratios. (c) Pendulum-v1 experiments with different standard deviations of Gaussian noise.

As the label ratio decreases, PN-GAIL exhibits **only a marginal decline** in performance.
Even when confidence scores are subject to noise, PN-GAIL still demonstrates **satisfactory performance**, indicating its **robustness** to noisy confidence scores.

Thank You

Contact: qiangliu@smail.nju.edu.cn

南京大學

Robotics & Reinforcement Learning Control