Unlabeled Imperfect Demonstrations in Adversarial Imitation Learning

Yunke Wang[1], **Bo Du**[1*], **Chang Xu**[2*]

[1]School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China.

[2]School of Computer Science, Faculty of Engineering, The University of Sydney, Australia.

{yunke.wang, dubo}@whu.edu.cn, c.xu@sydney.edu.au

Abstract

Adversarial imitation learning has become a widely used imitation learning framework. The discriminator is often trained by taking expert demonstrations and policy trajectories as examples respectively from two categories (positive vs. negative) and the policy is then expected to produce trajectories that are indistinguishable from the expert demonstrations. But in the real world, the collected expert demonstrations are more likely to be imperfect, where only an unknown fraction of the demonstrations are optimal. Instead of treating imperfect expert demonstrations as absolutely positive or negative, we investigate *unlabeled* imperfect expert demonstrations as they are. A positive-unlabeled adversarial imitation learning algorithm is developed to dynamically sample expert demonstrations that can well match the trajectories from the constantly optimized agent policy. The trajectories of an initial agent policy could be closer to those non-optimal expert demonstrations, but within the framework of adversarial imitation learning, agent policy will be optimized to cheat the discriminator and produce trajectories that are similar to those optimal expert demonstrations. Theoretical analysis shows that our method learns from the imperfect demonstrations via a self-paced way. Experimental results on MuJoCo and RoboSuite platforms demonstrate the effectiveness of our method from different aspects.

Introduction

Reinforcement Learning (RL) (Sutton and Barto 2018; Kaelbling, Littman, and Moore 1996) provides an effective framework for solving sequential decision-making problems (Silver et al. 2016; Van Hasselt, Guez, and Silver 2016; Zha et al. 2021). It aims to learn a good policy by rewarding the agent's action during its interaction with the environment. A well-formulated reward can recover the best policy, yet this complex reward engineering (Amodei et al. 2016) in real-world tasks can make RL fail sometimes. By contrast, it could be more practical to introduce imitation learning (IL) (Hussein et al. 2017; Zheng et al. 2022): a popular learning paradigm to guide policy learning by directly mimicking expert behaviors. A basic approach of IL is Behavioral Cloning (BC) (Pomerleau 1988), in which the agent observes the action of the expert and learns a mapping from state to action via regression. However, this offline training manner may suffer from compounding errors (Brantley, Sun, and Henaff 2019; Xu, Li, and Yu 2020; Tu et al. 2022) when the agent executes the policy, leading it to drift to new and dangerous states. Instead, Adversarial Imitation Learning (AIL) encourages the agent to cover the distribution of the expert policy, which can result in a more precise policy. Generative Adversarial Imitation Learning (GAIL) (Ho and Ermon 2016) is the most prominent work of AIL and it inherits the framework of Generative Adversarial Nets (GAN) (Goodfellow et al. 2014). After GAIL, there are many variants (Li, Song, and Ermon 2017; Fu, Luo, and Levine 2018; Peng et al. 2018; Dadashi et al. 2020; Cai et al. 2021, 2019) to further enhance the performance of AIL from various aspects.

Existing imitation learning methods achieve promising results under the assumption that the given expert demonstrations are of high quality (Hussein et al. 2017). However, there is a fact that most of them would fail when injecting some non-optimal demonstrations into expert demonstrations, which results in the imperfect demonstrations issue in IL (Wu et al. 2019; Ross, Gordon, and Bagnell 2011). This issue is of practice since it could be costly to collect purely optimal demonstrations in the real world. Therefore, how to learn a good policy from a mixture of optimal and non-optimal demonstrations is crucial to bridging the applicable gap of IL from the simulator to real-world tasks.

Confidence-based methods are popular and effective to address imperfect demonstrations issue in imitation learning. The key lies in how to acquire proper confidence for each expert demonstration. In 2IWIL (Wu et al. 2019) and IC-GAIL (Wu et al. 2019), an annotator is employed to manually label confidence for a fraction of demonstrations. The former is a two-stage method, which predicts the confidence for the remaining unlabeled demonstrations first and then conducts a weighted imitation learning framework. The latter combines these two steps in a single objective function instead. WGAIL (Wang et al. 2021) successfully connects confidence estimation to the discriminator in GAIL, and BCND (Sasaki and Yamashina 2021) demonstrates confidence can be derived by the agent policy itself. Therefore, these two methods relax the assumption on the labeled confidence and can be conducted without exposure to prior information. The confidence estimation in these two methods largely relies on the model's training status itself, but there

*Corresponding author

might be some extreme situations where the model collapses and fails to predict informative confidence. For example, the high ratio of contamination in expert demonstrations could seriously hurt the training process of IL, which can further lead to the collapse of confidence estimation. Confidence-based methods under such cases might be hard to even outperform their baseline.

Instead of estimating the precise confidence, our thought is to adopt a better training scheme for adversarial imitation learning with imperfect demonstrations via positive-unlabeled learning. This results in our method UID, which is general and can be equipped with various adversarial imitation learning backbones. Specifically, the imperfect demonstrations in UID are treated as unlabeled data, in which there exists a fraction of demonstrations that can well match the agent demonstrations. The positive-unlabeled adversarial imitation learning process can therefore be formulated by dynamically sampling demonstrations that resemble the behavior of the constantly optimized agent policy. The agent policy might produce demonstrations similar to the non-optimal demonstrations at the early training stage, yet it will be optimized to cheat the discriminator and produces demonstrations resembling those optimal demonstrations within the framework of adversarial training. Theoretical analysis shows UID gradually makes the agent cover more samples in unlabeled demonstrations via a self-paced way. Experimental results in MuJoCo (Todorov, Erez, and Tassa 2012) and RoboSuite (Fan et al. 2018) demonstrate the effectiveness of UID from different aspects.

Related Work

In this section, we briefly review the existing researches on imitation learning with imperfect demonstrations. We roughly divided them into two categories, *i.e.*, confidence-based methods and preference-based methods.

Confidence-based methods Instance reweighting has been widely used in various machine learning problems (Zhang et al. 2020; Ren, Yeh, and Schwing 2020; Zhong, Du, and Xu 2021; Qiu et al. 2022; Yang, Qiu, and Fu 2022) and gains great success. 2IWIL (Wu et al. 2019) and IC-GAIL (Wu et al. 2019) first investigate the capacity of the weighting scheme in imitation learning and find it effective in dealing with imperfect demonstrations. However, the assumption that a fraction of demonstrations should be manually labeled with confidence is a strong prior and hard to satisfy in the real world. Additionally, different human annotators may have different judgments on the goodness of demonstrations. The following works (Wang et al. 2021; Zhang et al. 2021; Wang, Xu, and Du 2021; Chen et al. 2022) thus focus on how to relax the assumption when estimating the confidence. To name a few, CAIL (Zhang et al. 2021) considers to introduce a small fraction of ranked trajectories to help with the confidence estimation during the training. WGAIL (Wang et al. 2021) proves that the optimal confidence should be proportional to the exponential advantage function, and then connects advantage with agent policy and the discriminator in GAIL. An alternating interaction between weight estimation and GAIL training therefore holds. There are also some researches (Sasaki and Yamashina 2021; Kim et al. 2021; Xu et al. 2022; Liu et al. 2022) on addressing imperfect demonstrations issue in offline imitation learning. BCND (Sasaki and Yamashina 2021) is a weighted behavioral cloning method, with action distribution of learned policy as confidence. However, when sub-optimal demonstrations occupy the major mode within imperfect demonstrations, the confidence distribution is likely to drift to the sub-optimal demonstrations and assign higher confidence to them. DemoDICE (Kim et al. 2021) performs offline imitation learning with a KL constraint between the learned policy and supplementary imperfect demonstrations to efficiently utilize additional demonstration data.

Preference-based methods Preference-based methods have been proved to be effective in policy learning. (Christiano et al. 2017) firstly applied active preference learning to Atari games, asking the expert to select the best of two trajectories generated from an ensemble of policies. The policy is learned to maximize the reward defined by expert preference during the interaction. T-REX (Brown et al. 2019) aims to extrapolate a reward function by ranked trajectories. The learned reward function can well explain the rankings, and thus is informative to be used as feedback for the agent. T-REX only requires precise rankings of trajectories, yet does not set constraints on data quality. It can thus perform quite well even with no optimal trajectories. D-REX (Brown, Goo, and Niekum 2020) further relaxes T-REX's constraint on rankings. It learns a pre-trained policy by behavioral cloning first, and the ranked trajectories can be generated by injecting different noises into its action. SSRR (Chen, Paleja, and Gombolay 2021) fixes the possible error of rankings by defining a new structure of the reward function.

Preliminary

In this section, we briefly review the definition of Markov Decision Process (MDP) and adversarial imitation learning.

Markov Decision Process (MDP) MDP is popular to formulate reinforcement learning (RL) (Puterman 1994) and imitation learning (IL) problems. An MDP normally consists six basic elements $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu_0)$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the stochastic transition probability from current state s to the next state s', $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the obtained reward of agent when taking action a in a certain state s, $\gamma \in [0, 1]$ is the discounted rate and μ_0 denotes the initial state distribution. Given a trajectory $\tau = \{(s_t, a_t)\}_{t=0}^{T}$, the return $R(\tau)$ is defined as the discounted sum of rewards obtained by the agent over all episodes, $R(\tau) = \sum_{t=0}^{T} \gamma^k r(s_k, a_k)$ and T is the number of steps to reach an absorbing state. The goal of RL is thus to learn a policy that can maximize the expected return over all episodes during the interaction. For any policy $\pi : \mathcal{S} \to \mathcal{A}$, there is an one-to-one correspondence between π and its occupancy measure $\rho_\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$.

Adversarial Imitation Learning (AIL) Adversarial imitation learning addresses IL problems from the perspective of distribution matching. By minimizing the distance

between distributions of agent demonstrations and expert behaviors, AIL can thus recover the expert policy. Generative Adversarial Imitation Learning (GAIL) (Ho and Ermon 2016) is the most representative work of AIL, which directly applies the general GAN framework (Goodfellow et al. 2014) into adversarial imitation learning. Given a set of expert demonstrations $\mathcal{D}_e$ drawn from the expert policy π_e, GAIL aims to learn an agent policy π_θ by minimizing the Jensen-Shannon divergence between ρ_{π_θ} and ρ_{π_e} (Ke et al. 2020). In the implementation, a discriminator is introduced to distinguish demonstrations from expert policy and agent policy, yet the agent policy tries its best to 'fool' the discriminator. This results in a minimax adversarial objective as follow,

$$\min_{\theta}\max_{\psi} \mathbb{E}_{(s,a)\sim\rho_{\pi_e}}[\log D_\psi(s,a)] + \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[\log(1-D_\psi(s,a))]. \tag{1}$$

The agent is trained to minimize the outer objective function $\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[\log(1-D_\psi(s,a))]$, and therefore the output of $-\log(1-D_\psi(s,a))$ can be regarded as reward. Regular RL methods like TRPO (Schulman et al. 2015), PPO (Schulman et al. 2017) and SAC (Haarnoja et al. 2018) can be thus used to update the agent policy π_θ.

Methodology

Most adversarial imitation learning methods achieve promising results in benchmark tasks with a non-trivial assumption that the given expert demonstrations should be optimal. However, querying the expert for a large amount of optimal behaviors can be expensive in some real-world tasks. By contrast, it could be more realistic to collect mixed demonstrations with only a fraction of optimal samples. In this paper, we consider this practical setting and investigate how to ensure a promising performance when dealing with imperfect demonstrations.

Proposed Method: UID

In our setting, we have a mixture set of expert demonstrations $\mathcal{D}_e$ that contains both optimal demonstrations and non-optimal demonstrations. Since the specific information about the demonstrations' optimality is unknown, we consider to regard $\mathcal{D}_e$ as *unlabeled* demonstrations and label their categories dynamically based on the status of agent policy. Supposing $\rho_{\pi_{\widehat{\theta}}}$ represents the distribution of a fraction of unlabeled demonstrations that can well match agent demonstrations from ρ_{π_θ}, we model ρ_{π_e} as a mixture of distributions

$$\rho_{\pi_e}(s,a) = (1-\alpha)\rho_{\pi_\epsilon}(s,a) + \alpha\rho_{\pi_{\widehat{\theta}}}(s,a), \tag{2}$$

where $\alpha \in [0,1]$ is the mixing proportion of the matched distribution $\rho_{\pi_{\widehat{\theta}}}$, and ρ_{π_ϵ} can be regarded as the distribution of those remaining demonstrations in $\mathcal{D}_e$. We denote π_ϵ as the *residual* policy. In plain adversarial imitation learning, all unlabeled demonstrations are simply labeled as positives in discriminator training. However, this training scheme only makes sense when the labeled demonstrations are clean. When there exists some non-optimal demonstrations, the discriminator would equally treat both optimal demonstrations and non-optimal demonstrations. Hence, agent demonstrations that resemble those imperfect data would also be assigned with high reward, which results in sub-optimal agent behavior.

Our thought is to build an arbitrary discriminator $g : (s,a) \mapsto \mathbb{R}$ that has better discriminative ability among unlabeled demonstrations $\mathcal{D}_e$. Supposing the surrogate loss function $\phi : \mathbb{R}\times\{\pm 1\} \mapsto \mathbb{R}$ is a margin-based loss function for binary classification, the expectation of risk of discriminator g can be expressed as

$$R_{\pi_e}(g) = (1-\alpha)\mathbb{E}_{(s,a)\sim\rho_{\pi_\epsilon}}[\phi(g(s,a))] + \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_{\widehat{\theta}}}}[-\phi(g(s,a))]. \tag{3}$$

The residual policy π_ϵ is inaccessible in our setting, however, we have $\pi_{\widehat{\theta}}$ that is assumed to be approximating the agent policy π_θ. Therefore, we consider to replace $(1-\alpha)\rho_{\pi_\epsilon}$ with $(\rho_{\pi_e} - \alpha\rho_{\pi_{\widehat{\theta}}})$ and introduce the agent policy π_θ as an estimation of $\pi_{\widehat{\theta}}$. Then, the expected risk R_{π_e} can be estimated by π_θ and π_e, and the optimal discriminator g can be obtained by minimizing $R_{\pi_e}(g)$,

$$\min_g R_{\pi_e}(g) = \mathcal{T}\{0, \mathbb{E}_{(s,a)\sim\rho_{\pi_e}}[\phi(g(s,a))] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[\phi(g(s,a))]\} + \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[-\phi(g(s,a))], \tag{4}$$

where $\mathcal{T}\{\cdot\}$ is a flexible constraint, which makes the replacement $\mathbb{E}_{(s,a)\sim\rho_{\pi_e}}[\phi(g(s,a))] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[\phi(g(s,a))]$ have the same sign as the original loss function $\mathbb{E}_{(s,a)\sim\rho_{\pi_\epsilon}}[\phi(g(s,a))]$. Eq. (4) is an unbiased and consistent risk estimator of the true risk w.r.t all popular loss functions as mentioned in (Niu et al. 2016).

Considering the agent policy should also learn from this discriminator, it should be trained to produce trajectories that can 'fool' the judgment of the discriminator. We therefore set up an adversarial game between π_θ and g, and obtain the following objective function $\mathcal{J}(\theta, g)$,

$$\max_\theta\min_g \mathcal{J}(g,\theta) = \mathcal{T}\{0, \mathbb{E}_{(s,a)\sim\rho_{\pi_e}}[\phi(g(s,a))] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[\phi(g(s,a))]\} + \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[-\phi(g(s,a))]. \tag{5}$$

Eq. (5) is a general objective function with an unspecific loss function ϕ. However, since adversarial imitation learning methods are not always directly linked to a certain surrogate loss function, it is hard to straightly recover various AIL baselines by specifying a $\phi(g)$. By contrast, most adversarial imitation learning methods can be viewed as minimizing the different distances between occupancy measures of agent policy and expert policy. We therefore consider to connect margin-based loss function $\phi(g)$ with f-divergence and then write the general form of UID for various adversarial imitation learning methods. We summarize this process in the following theorem.

Theorem 1. *For any margin-based surrogate convex loss $\phi : \mathbb{R}\times\{\pm 1\} \mapsto \mathbb{R}$ in Eq. (4), there is a related f-*

divergence I_f such that

$$\min_g R_{\pi_e}(g,\phi) = -I_f(\mu,\nu) = -\int_{s,a} \mu(s,a) f(\frac{\mu(s,a)}{\nu(s,a)}) dsda, \tag{6}$$

where $\mu = \rho_{\pi_e} - \alpha\rho_{\pi_\theta}$, $\nu = \alpha\rho_{\pi_\theta}$ and $f : [0,\infty] \to \mathbb{R} \cup \{\infty\}$ is a continuous convex function. Then, by using variational approximation of f-divergence, $\min_g R_{\pi_e}(g)$ can be further written as

$$\max_T \min\{0, \mathbb{E}_{(s,a)\sim\rho_{\pi_e}}[T(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[T(s,a)]\} - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}} f^*[T(s,a)]. \tag{7}$$

where $T(s,a)$ is the decision function related to g. Different choices of convex function f can recover different objective function of UID adversarial imitation learning.

With the help of Theorem 1, we can now integrate the proposed method into various frameworks of AIL with different choices of f-divergence. This flexibility that combined with other models provides the proposed method a chance to get further improvement on existing adversarial imitation learning backbones.

UID-GAIL We provide a specific case by recovering GAIL, which is the most representative AIL methods. We consider to use Jensen-Shannon divergence and define $f(u) = -(u+1)\log\frac{u+1}{2} + u\log u$, $f^*(t) = -1 - \log(1-\exp(t))$. By replacing $T(s,a)$ with $\log[D(s,a)]$, the objective function of UID can be written as,

$$\begin{aligned}\min_\theta \max_\psi \mathcal{J}(\theta,\psi) = &\min\{0, \mathbb{E}_{(s,a)\sim\rho_{\pi_e}} \log[D_\psi(s,a)] \\ &- \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}} \log[D_\psi(s,a)]\} \\ &+ \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}} \log[1 - D_\psi(s,a)].\end{aligned} \tag{8}$$

The practical optimization of UID-GAIL is summarized in Algorithm 1.

UID-WAIL We also show the flexibility of UID with other popular AIL methods, *i.e.*, WAIL (Xiao et al. 2019). Recall that Theorem 1 makes it possible to recover specific adversarial imitation learning baselines by defining different f-divergence functions, however, the Wasserstein distance metric used in WAIL is not strictly an f-divergence. Therefore, we begin with Total Variation (TV), which is a kind of f-divergence that is related to Wasserstein distance. The f function in total variation is defined as $f(u) = \frac{1}{2}|u-1|$, therefore we have $f^*(t) = t$. By defining critic $r_\psi(s,a) = T(s,a)$, we then re-write Eq. (7) as,

$$\max_\psi \min\{0, \mathbb{E}_{(s,a)\sim\rho_{\pi_E}}[r_\psi(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[r_\psi(s,a)]\} - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[r_\psi(s,a)]. \tag{9}$$

TV can be regarded as the Wasserstein distance with respect to 1-Lipschitz constraint on r_ψ. We then add this regularization on r_ψ and obtain the final objective function of UID-WAIL,

$$\min_\theta \max_\psi \min\big\{0, \mathbb{E}_{(s,a)\sim\rho_{\pi_e}}[r_\psi(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[r_\psi(s,a)]\big\} - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[r_\psi(s,a)] + \lambda\Psi(r_\psi), \tag{10}$$

Algorithm 1: UID-GAIL

Require:
Unlabeled demonstrations $\mathcal{D}_e = \{s_i, a_i\}_{i=1}^n \sim \rho_{\pi_e}$;
Total iterations N;
Ensure:
The agent policy π_θ; The discriminator D_ψ;
Initialize D_ψ and π_θ;
for iter = 1 to N **do**
  Sample trajectories $\{s^\theta, a^\theta\} \sim \rho_{\pi_\theta}$, $\{s^e, a^e\} \sim \mathcal{D}_e$;
  Update D_ψ by maximizing $\mathcal{J}(\theta,\psi)$
  Update π_θ by TRPO with reward $-\log[1-D_\psi(s,a)]$;
end for

where the critic r_ψ serves as the reward function and $\Psi(r_\psi) = -\mathbb{E}_{(s,a)\sim\rho_{\hat{\pi}}}(||\nabla r_\psi(\hat{s},\hat{a})||_2 - 1)^2$ is the regularization term to satisfy the Lipschitz constraint.

Theoretical Results of UID

Since $\rho_{\pi_{\hat{\theta}}}$ dynamically samples from ρ_{π_e} to approximate ρ_{π_θ} during training, the PU discriminator will thus make the agent produce demonstrations that resemble the residual policy π_ϵ. As π_ϵ changes during training as well, the target of the optimization of agent policy π_θ changes accordingly.

Remark 1. *At the early training stage, $\pi_{\hat{\theta}}$ is of bad quality and represents the relatively bad part in unlabeled imperfect demonstrations. This makes the residual policy π_ϵ occupy the optimal mode within unlabeled demonstrations. Under such cases, agent policy π_θ is imitating the optimal demonstrations.*

Theorem 2. *For the agent policy π_θ fixed, the optimal discriminator $D^*_\psi(s,a)$ can be written as*

$$D^*_\psi(s,a) = \frac{\rho_{\pi_\epsilon}(s,a)}{\rho_{\pi_\epsilon}(s,a) + \frac{1-\alpha}{\alpha}\rho_{\pi_\theta}(s,a)}, \tag{11}$$

With the optimal discriminator $D^*_\psi(s,a)$ fixed, the optimization of π_θ is equivalent to minimize

$$C + (1-\alpha)KL(\rho_{\pi_\epsilon}||\rho_{\pi_e}) + \alpha KL(\rho_{\pi_\theta}||\rho_{\pi_e}), \tag{12}$$

where $C = (1-\alpha)\log(1-\alpha) + \alpha\log\alpha$. The global minimum of the proposed objective function is achieved if and only if $\rho_{\pi_\theta} = \rho_{\pi_\epsilon} = \rho_{\pi_e}$. At that point, the objective achieves the value $(1-\alpha)\log(1-\alpha) + \alpha\log\alpha$, and $D^*_\psi(s,a)$ achieves the value α.

From Theorem 2, we prove that UID approaches Nash equilibrium when $\rho_{\pi_\theta} = \rho_{\pi_e}$. This illustrates that UID makes the agent imitate π_e finally. Recall that we also show that π_θ is imitating the optimal demonstrations at the early training stage in Remark 1. Therefore, we conclude that UID makes π_θ imitate optimal demonstrations within unlabeled demonstrations firstly and then gradually covers more demonstrations in unlabeled imperfect demonstrations. This actually leads UID to relate to curriculum learning (Bengio et al. 2009) and self-paced learning (Kumar, Packer, and Koller 2010), which also make the model learn from good samples to other samples gradually. This connection provides a theoretical guarantee of UID's advantage compared

to plain GAIL. The empirical study in the experiment part identifies the analysis above.

Discussion

Connection with PU Learning The discriminator training scheme above is related to non-negative positive-unlabeled learning (Du Plessis, Niu, and Sugiyama 2014; Kiryo et al. 2017; Xu et al. 2017, 2019). In positive-unlabeled classification, two sets of data are sampled independently from positive data distribution $p_p(x)$ and unlabeled data distribution $p_u(x)$ as $\mathcal{X}_p = \{x_i^P\}_{i=1}^{n_p} \sim p_p(x)$ and $\mathcal{X}_u = \{x_i^U\}_{i=1}^{n_u} \sim p_u(x)$, and a classifier $g(x)$ needs to be trained to distinguish samples from $\mathcal{X}_p$ and $\mathcal{X}_u$. Regarding $\rho_{\pi_{\hat{\theta}}}$ as the known positive distribution p_p and ρ_{π_e} as the unlabeled mixture data distribution p_u, we find that the process of discriminator training can be exactly viewed as a special example of PU learning. Moreover, we investigate the compatibility of PU learning with the adversarial imitation learning framework and show it can well handle imperfect demonstrations issue in adversarial imitation learning.

Another related method is PU-GAIL, which also adopts a PU-based classifier in adversarial imitation learning (Guo et al. 2020). Under the assumption that the agent policy produces diverse demonstrations during training, PU-GAIL treats agent demonstrations as unlabeled data while regarding expert demonstrations as positive data to form this PU classifier. PU-GAIL can be regarded as a regularization technology for the discriminator to prevent overfitting problems (Orsini et al. 2021), which may help to stabilize the adversarial training. But PU-GAIL would fail when dealing with imperfect demonstrations, since it still lets the agent imitate all demonstrations equally all the time. By contrast, UID views expert demonstrations as unlabeled data and learns from the demonstrations via a self-paced way. Empirical results in the experiment show that UID has a better discriminative ability within unlabeled demonstrations and can achieve better performance with imperfect demonstrations.

Experiments

In this section, we conduct experiments to verify the effectiveness of UID in various benchmarks (*i.e.*, MuJoCo (Todorov, Erez, and Tassa 2012) and Robosuite (Zhu et al. 2020)) under different settings. The experimental results demonstrate the advantage of UID from different aspects.[1]

Experimental Setting We evaluate UID on three MuJoCo (Todorov, Erez, and Tassa 2012) locomotion tasks (*i.e.*, Ant-v2, HalfCheetah-v2 and Walker2d-v2) firstly. The agent performance in MuJoCo can be measured by both the average cumulative rewards along trajectories and the final location of the agent (*i.e.*, higher the better). We evaluate the agent every 5,000 transitions in training and the reported results are the average of the last 100 evaluations. We repeat experiments for 5 trials with different random seeds. To verify the robustness of UID with real-world human operation demonstrations, we also conduct experiments on a robot control task in Robosuite (Zhu et al. 2020).

[1]https://github.com/yunke-wang/UID

Source of Demonstrations We collect a mixture of optimal and non-optimal demonstrations to conduct experiments. To form these unlabeled demonstrations, an optimal expert policy π_o trained by TRPO is used to sample optimal demonstrations $\mathcal{D}_o$, and then 3 non-optimal expert policies π_n are used to sample non-optimal demonstrations $\mathcal{D}_n$. Following existing works, we use two different kinds of π_n to sample non-optimal demonstrations.

- **D1**: We save 3 checkpoints during the RL training as 3 non-optimal expert policies π_n.
- **D2**: We add Gaussian noise ξ to the action distribution a^* of π_o to form non-optimal expert π_n. The action of π_n is modeled as $a \sim \mathcal{N}(a^*, \xi^2)$ and we choose $\xi = [0.25, 0.4, 0.6]$ in these 3 non-optimal policies.

Equal demonstrations are sampled from each policy. The unlabeled expert demonstrations $\mathcal{D}_e$ is formed by mixing the sampled optimal demonstrations $\mathcal{D}_o$ and non-optimal demonstrations $\mathcal{D}_n$. The data quality and the detailed implementation are deferred to the supplementary material.

Results on MuJoCo

Varying Ratios of Optimal Demonstrations We firstly investigate the capacity of UID when dealing with varying ratios of optimal demonstrations in Ant-v2 task. We begin with 50% (1:1) optimal demonstrations, and gradually decrease the ratio of optimal data to around 16.7% (1:5). The compared method are two state-of-the-art confidence-based methods WGAIL (Wang et al. 2021) and BCND (Sasaki and Yamashina 2021) that do not require any prior information when estimating weight.

As claimed in (Sasaki and Yamashina 2021), BCND needs a "50% optimal data" assumption on the mixed demonstrations to ensure a promising performance. If non-optimal demonstrations occupy the major mode, the confidence distribution is likely to drift to the non-optimal part. We observe a similar phenomenon in our experiment. As shown in Figure 1, when given 50% optimal demonstrations, BCND can still outperform BC by a clear margin. However, when the ratio of optimal demonstrations decreases, the performance of BCND drops and starts to inferior to BC with less than 25% optimal demonstrations. Online imitation learning methods (*i.e.*, UID, WGAIL, and GAIL) perform generally better than offline imitation learning methods. WGAIL performs best at 50% optimal demonstrations point, however, its performance decreases rapidly and achieves similar performance with GAIL when given

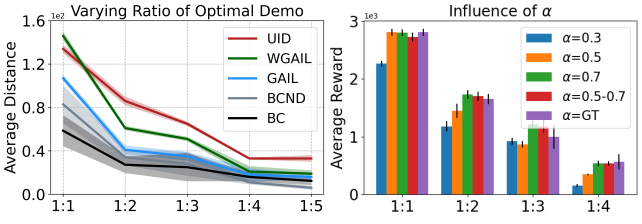

Figure 1: Performance with varying ratios of optimal demonstrations.

Method	D1			D2		
	Ant-v2	HalfCheetah-v2	Walker2d-v2	Ant-v2	HalfCheetah-v2	Walker2d-v2
WAIL (Xiao et al. 2019)	1348±120	2282±58	2180±46	2039±48	3124±334	2656±170
UID-WAIL (Ours)	**1709±118**	**2569±157**	**2359±43**	**2490±59**	**3582±340**	**3364±104**
GAIL (Ho and Ermon 2016)	1179±158	2159±139	1873±115	1797±137	2758±205	2786±262
UID-GAIL (Ours)	**1674±142**	**3276±114**	**2482±65**	**2426±110**	**3983±179**	**3343±180**
2IWIL (Wu et al. 2019)	1591±71	2704±129	2204±66	2317±123	2656±261	2749±258
IC-GAIL (Wu et al. 2019)	**1974±41**	2779±92	2002±54	1883±90	3087±226	2429±166
T-REX (Brown et al. 2019)	-556±83	2223±255	1866±296	-22±2	1399±499	1622±165
D-REX (Brown, Goo, and Niekum 2020)	-1751±194	470±86	529±91	-27±20	2588±75	1433±104
PU-GAIL (Xu and Denil 2021)	310±86	1136±332	1469±379	1734±140	2413±505	2652±112

Table 1: Performance of proposed methods and compared methods in MuJoCo tasks with both stage 1 and stage 2 demonstrations, which is measured by the average and standard variance of ground-truth cumulative reward along 10 trajectories, *i.e.*, higher average value is better. The value in Bold denotes the best value between UID and its baseline.

less than 25% optimal demonstrations point. By contrast, the curve of UID is clearly above GAIL as the data quality decreases. We therefore conclude that UID can have a better performance than WGAIL and BCND with limited optimal demonstrations.

Impact of α We conduct ablation studies on α to find how different α could influence the final results of UID. We evaluate the performance of UID with varying ratios of optimal demonstrations with different α (*i.e.*, $\alpha = 0.3, 0.5, 0.7, 0.5-0.7$). We also provide a result by heuristically setting α as the real ratio of optimal demonstrations. The results are summarized in Figure 1. The 'red' rectangle denotes that we set α as the real ratio of non-optimal demonstrations. We find that UID enjoys a relatively considerable tolerance of α. Generally, setting $\alpha = 0.7$ results in the best performance in most cases. We therefore consider directly treating α as a hyperparameter and UID can also be regarded as a method that does not require prior information.

Performance on various AIL frameworks Since UID can be extended into more adversarial imitation learning frameworks by defining different f-divergence in Theorem 1, we test the capacity of UID with two AIL baselines, *i.e.*, GAIL and WAIL. The results are shown in Table 1. We observe that UID beats vanilla AIL with both D1 and D2 demonstrations in all three baselines with a clear improvement. This illustrates the effectiveness of UID when dealing with different kinds of mixed imperfect demonstrations. We also conduct student's t-test on the results and the null hypothesis is the performance of UID is similar or worse than the GAIL baseline. The result is shown in Table 2, from which we can observe that there is a statistical significance between the performance of UID and GAIL since most p-values are clearly below 0.05. We also provide screenshots in MuJoCo software to observe the performance of the agent from the visual perspective, as shown in Figure 2. We find that the agent learned by UID runs fast and can be successfully qualified for these tasks. Additionally, we compare UID with several preference-based methods (*i.e.*, T-REX and D-REX) and confidence-based methods (*i.e.*, 2IWIL and IC-GAIL). The rankings of trajectories in T-REX are given

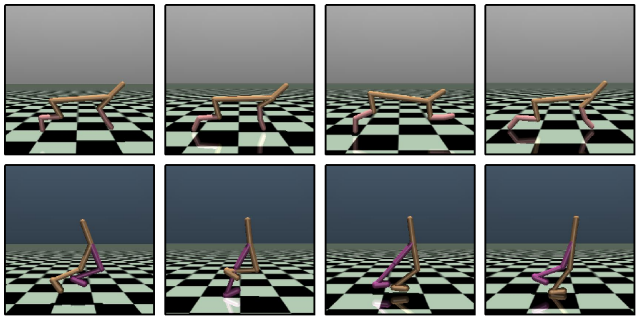

Figure 2: Visualization of the agent trained by UID with class 1 demonstrations. Time step increases from the leftmost figure (t=25) to the rightmost figure (t=100).

as a prior and we use the normalized reward of each checkpoint as the confidence for each demonstration. Generally speaking, preference-based methods do not perform well in most cases, yet the confidence-based methods 2IWIL and IC-GAIL perform clearly better. Especially in Ant-v2 and Walker-v2, we find that the performance of 2IWIL in these two environments is only slightly inferior to UID. However, 2IWIL requires strong prior information on the confidence of demonstrations that may not be easily obtained.

As discussed in the methodology, PU-GAIL also introduces a PU classifier into a generative adversarial imitation learning framework. While treating agent demonstrations as unlabeled samples, PU-GAIL learns a better discriminator by considering the increasing ratio of good samples produced by agent policy. This training scheme is more sound than GAIL training and might be helpful to stable GAN training and avoid local minima, however, this does

p-value	Ant-v2	HalfCheetah-v2	Walker2d-v2
(D1)	0.0702	0.0005	0.0032
(D2)	0.0126	0.0038	0.1556

Table 2: The p-value between UID and its baseline GAIL.

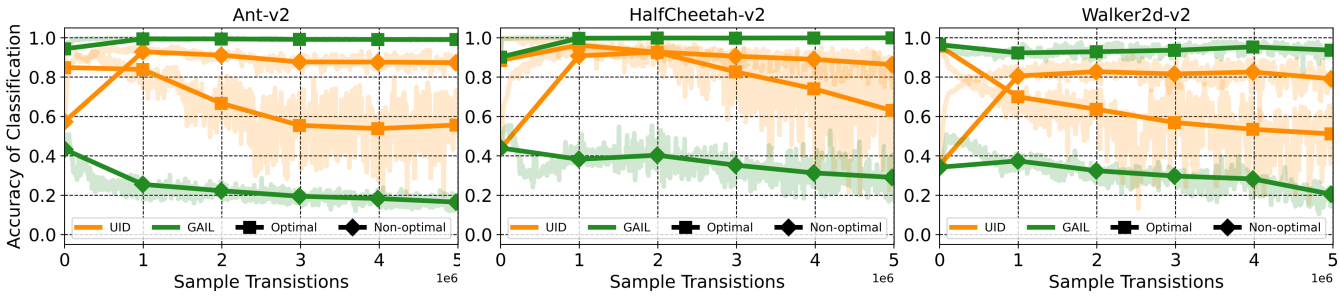

Figure 3: The accuracy of D_ψ in classifying optimal demonstrations $\mathcal{D}_o$ and non-optimal demonstrations $\mathcal{D}_n$ within unlabeled demonstrations $\mathcal{D}_e$ during UID and GAIL training. We provide smooth version of the initial learning curve (the shade part) for better observation.

not change its actual upper bound of performance since PU-GAIL still regards all given expert demonstrations as positives. When given imperfect demonstrations, PU-GAIL can only learn an inferior performance. In Table 1, we observe the performance of PU-GAIL is similar or sometimes inferior to its baseline. This illustrates that PU-GAIL can not well handle imperfect demonstrations in imitation learning.

Analysis on the discriminator During the training of UID on unlabeled demonstrations $\mathcal{D}_e$, we investigate the performance of the discriminator by testing its classification accuracy on optimal demonstrations $\mathcal{D}_o$ and non-optimal demonstrations $\mathcal{D}_n$. The accurate classification is defined as treating $\mathcal{D}_o$ as positive and treating $\mathcal{D}_n$ as negative. In Figure 3, there is a clear trend that the discriminator in both methods reaches high accuracy in classifying $\mathcal{D}_o$. However, when it comes to $\mathcal{D}_n$, the accuracy of the discriminator is generally low in GAIL. This shows that the discriminator in GAIL equally regards $\mathcal{D}_n$ and $\mathcal{D}_o$ as 'positive', while the discriminator in UID obviously has discriminative ability on these two kinds of demonstrations within unlabeled demonstrations $\mathcal{D}_e$. This is due to introducing the idea of PU classification in UID.

Another trend is that the discriminator in UID has a decreased classification accuracy on optimal demonstrations $\mathcal{D}_o$ during training. Since the output of the discriminator is proportional to the reward, agent demonstrations that are classified as 'positive' by the discriminator will be assigned with a high reward in the RL step. At the beginning of the training, the obtained reward of agent demonstrations that resemble $\mathcal{D}_o$ would be thus relatively higher. This can be exactly viewed as encouraging the agent to learn from $\mathcal{D}_o$ at first. As the training progresses, the reward of agent demonstrations that are close to $\mathcal{D}_o$ will decline accordingly. This enables a chance for those non-optimal demonstrations $\mathcal{D}_u$ to participate in and guide the agent training. The empirical results here identify our analysis on the connection between UID and self-paced learning.

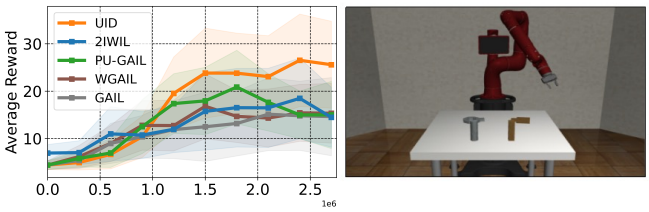

Figure 4: Performance of UID in RoboSuite tasks.

Results on RoboSuite Platform

We also evaluate the robustness of UID on the RoboSuite platform (Zhu et al. 2020) with real-world demonstrations. We consider a "Nut Assembly" task in Saywer, in which two colored pegs and two colored nuts are mounted on the tabletop, as shown in the right of Figure 4. The robot aims to fit the nut into its related peg. We use real-world demonstrations by human operators from RoboTurk website[2]. The demonstrations contain 10 trajectories with approaching length and the overall number of demonstrations is 5000. Based on the accumulative reward of trajectories, only three trajectories are regarded as optimal demonstrations. We therefore expect to test the performance of UID in imperfect demonstrations from the real world. Figure 4 shows the performance of UID with 3 million transition samples for RL training. We find that UID performs best over the other 4 compared methods. This experiment further identifies the robustness of UID with human demonstrations.

Conclusion

In this paper, we propose a general framework called UID to address the unlabeled imperfect demonstrations problem in adversarial imitation learning. Instead of treating all imperfect demonstrations as absolutely positive in plain GAIL, we regard imperfect demonstrations as unlabeled data and adopt a more efficient scheme to make the agent learn from them. With a fraction of unlabeled demonstrations separated to match the agent demonstrations, we develop a positive-unlabeled adversarial imitation learning framework. We also make this technology compatible with various adversarial imitation learning baselines. The final experimental results on MuJoCo and RoboSuite platforms demonstrate the advantage of UID in dealing with imperfect demonstrations over other compared methods.

[2]https://roboturk.stanford.edu/dataset_sim.html

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62225113, 41871243, 62141112, the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, the Australian Research Council under Project DP210101859, and the University of Sydney Research Accelerator (SOAR) Prize.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Brantley, K.; Sun, W.; and Henaff, M. 2019. Disagreement-regularized imitation learning. In *International Conference on Learning Representations*.

Brown, D.; Goo, W.; Nagarajan, P.; and Niekum, S. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, 783–792. PMLR.

Brown, D. S.; Goo, W.; and Niekum, S. 2020. Better-than-Demonstrator Imitation Learning via Automatically-Ranked Demonstrations. In *Conference on Robot Learning*, 330–359.

Cai, X.-Q.; Ding, Y.-X.; Chen, Z.-X.; Jiang, Y.; Sugiyama, M.; and Zhou, Z.-H. 2021. Seeing Differently, Acting Similarly: Heterogeneously Observable Imitation Learning. *arXiv preprint arXiv:2106.09256*.

Cai, X.-Q.; Ding, Y.-X.; Jiang, Y.; and Zhou, Z.-H. 2019. Imitation learning from pixel-level demonstrations by hashreward. *arXiv preprint arXiv:1909.03773*.

Chen, L.; Paleja, R.; and Gombolay, M. 2021. Learning from Suboptimal Demonstration via Self-Supervised Reward Regression. In *Conference on Robot Learning*, 1262–1277. PMLR.

Chen, Z.-X.; Cai, X.-Q.; Jiang, Y.; and Zhou, Z.-H. 2022. Anomaly Guided Policy Learning from Imperfect Demonstrations. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 244–252.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

Dadashi, R.; Hussenot, L.; Geist, M.; and Pietquin, O. 2020. Primal Wasserstein Imitation Learning. In *International Conference on Learning Representations*.

Du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2014. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27.

Fan, L.; Zhu, Y.; Zhu, J.; Liu, Z.; Zeng, O.; Gupta, A.; Creus-Costa, J.; Savarese, S.; and Fei-Fei, L. 2018. Surreal: Open-source reinforcement learning framework and robot manipulation benchmark. In *Conference on Robot Learning*, 767–782. PMLR.

Fu, J.; Luo, K.; and Levine, S. 2018. Learning Robust Rewards with Adverserial Inverse Reinforcement Learning. In *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Guo, T.; Xu, C.; Huang, J.; Wang, Y.; Shi, B.; Xu, C.; and Tao, D. 2020. On positive-unlabeled classification in GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8385–8393.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Ho, J.; and Ermon, S. 2016. Generative adversarial imitation learning. In *Advances in neural information processing systems*, 4565–4573.

Hussein, A.; Gaber, M. M.; Elyan, E.; and Jayne, C. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2): 1–35.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4: 237–285.

Ke, L.; Choudhury, S.; Barnes, M.; Sun, W.; Lee, G.; and Srinivasa, S. 2020. Imitation learning as f-divergence minimization. In *International Workshop on the Algorithmic Foundations of Robotics*, 313–329. Springer.

Kim, G.-H.; Seo, S.; Lee, J.; Jeon, W.; Hwang, H.; Yang, H.; and Kim, K.-E. 2021. DemoDICE: Offline Imitation Learning with Supplementary Imperfect Demonstrations. In *International Conference on Learning Representations*.

Kiryo, R.; Niu, G.; Du Plessis, M. C.; and Sugiyama, M. 2017. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In *Advances in neural information processing systems*, 1189–1197.

Li, Y.; Song, J.; and Ermon, S. 2017. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, 3812–3822.

Liu, L.; Tang, Z.; Li, L.; and Luo, D. 2022. Robust Imitation Learning from Corrupted Demonstrations. *arXiv preprint arXiv:2201.12594*.

Niu, G.; du Plessis, M. C.; Sakai, T.; Ma, Y.; and Sugiyama, M. 2016. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. *Advances in neural information processing systems*, 29.

Orsini, M.; Raichuk, A.; Hussenot, L.; Vincent, D.; Dadashi, R.; Girgin, S.; Geist, M.; Bachem, O.; Pietquin, O.; and

Andrychowicz, M. 2021. What matters for adversarial imitation learning? *Advances in Neural Information Processing Systems*, 34.

Peng, X. B.; Kanazawa, A.; Toyer, S.; Abbeel, P.; and Levine, S. 2018. Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow. In *International Conference on Learning Representations*.

Pomerleau, D. A. 1988. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley. ISBN 978-0-47161977-2.

Qiu, Z.; Yang, Q.; Wang, J.; and Fu, D. 2022. Dynamic Graph Reasoning for Multi-person 3D Pose Estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3521–3529.

Ren, Z.; Yeh, R.; and Schwing, A. 2020. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21786–21797.

Ross, S.; Gordon, G.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 627–635.

Sasaki, F.; and Yamashina, R. 2021. Behavioral Cloning from Noisy Demonstrations. In *International Conference on Learning Representations*.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International conference on machine learning*, 1889–1897.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587): 484–489.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. IEEE.

Tu, S.; Robey, A.; Zhang, T.; and Matni, N. 2022. On the sample complexity of stability constrained imitation learning. In *Learning for Dynamics and Control Conference*, 180–191. PMLR.

Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Wang, Y.; Xu, C.; and Du, B. 2021. Robust Adversarial Imitation Learning via Adaptively-Selected Demonstrations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 3155–3161.

Wang, Y.; Xu, C.; Du, B.; and Lee, H. 2021. Learning to Weight Imperfect Demonstrations. In *International Conference on Machine Learning*, 10961–10970. PMLR.

Wu, Y.-H.; Charoenphakdee, N.; Bao, H.; Tangkaratt, V.; and Sugiyama, M. 2019. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*, 6818–6827. PMLR.

Xiao, H.; Herman, M.; Wagner, J.; Ziesche, S.; Etesami, J.; and Linh, T. H. 2019. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*.

Xu, D.; and Denil, M. 2021. Positive-Unlabeled Reward Learning. In *Conference on Robot Learning*, 205–219. PMLR.

Xu, H.; Zhan, X.; Yin, H.; and Qin, H. 2022. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *International Conference on Machine Learning*, 24725–24742. PMLR.

Xu, T.; Li, Z.; and Yu, Y. 2020. Error bounds of imitating policies and environments. *Advances in Neural Information Processing Systems*, 33: 15737–15749.

Xu, Y.; Wang, Y.; Chen, H.; Han, K.; Xu, C.; Tao, D.; and Xu, C. 2019. Positive-unlabeled compression on the cloud. *Advances in Neural Information Processing Systems*, 32.

Xu, Y.; Xu, C.; Xu, C.; and Tao, D. 2017. Multi-Positive and Unlabeled Learning. In *IJCAI*, 3182–3188.

Yang, Z.; Qiu, Z.; and Fu, D. 2022. DMIS: Dynamic Mesh-based Importance Sampling for Training Physics-Informed Neural Networks. *arXiv preprint arXiv:2211.13944*.

Zha, D.; Xie, J.; Ma, W.; Zhang, S.; Lian, X.; Hu, X.; and Liu, J. 2021. Douzero: Mastering doudizhu with self-play deep reinforcement learning. In *International Conference on Machine Learning*, 12333–12344. PMLR.

Zhang, J.; Zhu, J.; Niu, G.; Han, B.; Sugiyama, M.; and Kankanhalli, M. 2020. Geometry-aware Instance-reweighted Adversarial Training. In *International Conference on Learning Representations*.

Zhang, S.; Cao, Z.; Sadigh, D.; and Sui, Y. 2021. Confidence-Aware Imitation Learning from Demonstrations with Varying Optimality. *Advances in Neural Information Processing Systems*, 34: 12340–12350.

Zheng, B.; Verma, S.; Zhou, J.; Tsang, I. W.; and Chen, F. 2022. Imitation learning: Progress, taxonomies and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 1–16.

Zhong, Y.; Du, B.; and Xu, C. 2021. Learning to reweight examples in multi-label classification. *Neural Networks*, 142: 428–436.

Zhu, Y.; Wong, J.; Mandlekar, A.; and Martín-Martín, R. 2020. robosuite: A Modular Simulation Framework and Benchmark for Robot Learning. In *arXiv preprint arXiv:2009.12293*.