Technical Appendix for "Unlabeled Imperfect Demonstrations in Adversarial Imitation Learning"

Proof for Theorem 1.

Theorem 1. *For any margin-based surrogate convex loss $\phi : \mathbb{R} \times \{\pm 1\} \mapsto \mathbb{R}$ in Eq. (8), there is a related f-divergence I_f such that*

$$\min_g R_{\pi_e}(g, \phi) = -I_f(\mu, \nu) = -\int_{s,a} \mu(s,a) f(\frac{\mu(s,a)}{\nu(s,a)}) dsda, \tag{13}$$

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

$$\max_T \min\{0, \mathbb{E}_{(s,a)\sim\rho_{\pi_e}}[T(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[T(s,a)]\} - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}} f^*[T(s,a)]. \tag{14}$$

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.

Proof. An optimal discriminator g recovers the minimum expected risk. By defining $\gamma = g(s,a)$, we have

$$-\inf_{\gamma\in\mathbb{R}} R_{\pi_e} = -\sum_{s,a} \inf_{\gamma\in\mathbb{R}} \Big[\phi(\gamma)\rho_{\pi_e}(s,a) - \alpha\phi(\gamma)\rho_{\pi_\theta}(s,a) + \alpha\rho_{\pi_\theta}(s,a)\phi(-\gamma)\Big] \tag{15}$$

$$= -\sum_{s,a} \inf_{\gamma\in\mathbb{R}} \Big[\big(\rho_{\pi_e}(s,a) - \alpha\rho_{\pi_\theta}(s,a)\big)\phi(\gamma) + \alpha\rho_{\pi_\theta}(s,a)\phi(-\gamma)\Big] \tag{16}$$

$$= \sum_{s,a} \alpha\rho_{\pi_\theta}(s,a)\Big(- \inf_{\gamma\in\mathbb{R}} \Big[\phi(-\gamma) + \phi(\gamma)u\Big]\Big), \tag{17}$$

where $u = \big(\rho_{\pi_e}(s,a) - \alpha\rho_{\pi_\theta}(s,a)\big)/\alpha\rho_{\pi_\theta}(s,a)$ and $\gamma = g(s,a)$. The infimum of $\big[\phi(-\gamma) + \phi(\gamma)u\big]$ is concave as a function of u since the minimum of a collection of linear functions is concave. Therefore, $-\inf_{\gamma\in\mathbb{R}} \Big[\phi(-\gamma) + \phi(\gamma)u\Big]$ is a convex function. We further define this convex function as $f(u)$,

$$f(u) = -\inf_{\gamma\in\mathbb{R}} \Big[\phi(-\gamma) + \phi(\gamma)u\Big] \tag{18}$$

$$= \sup_{\gamma\in\mathbb{R}} \Big[- \phi(-\gamma) - \phi(\gamma)u\Big] \tag{19}$$

$$= \sup_{\gamma\in\mathbb{R}} \Big[- \phi(-\phi^{-1}(\beta)) - \beta u\Big] \tag{20}$$

$$= \sup_{\beta\in\mathbb{R}} \big[- \beta u - \varpi(\beta)\big] \tag{21}$$

$$= \varpi^*(-u) \tag{22}$$

By defining $\beta = \phi(\gamma)$, γ can be directly expressed by the inverse form of β, *i.e.*, $\gamma = \phi^{-1}(\beta)$. $\varpi(\beta) : \beta \to \phi(-\phi^{-1}(\beta))$ is a convex function since it is a concave function followed by a nonincreasing convex function. Eq. (22) can be satisfied due to the convex conjugate of $f(u)$, *i.e.*, $f^*(t) = \sup_{u\in dom_f}\{ut - f(u)\}$. Therefore, we have $\varpi(\beta) = \varpi^{**}(\beta) = f^*(-\beta)$ and Eq. (17) can be written as,

$$-\inf_{\gamma\in\mathbb{R}} R_{\pi_e} = -\int_{s,a} \alpha\rho_{\pi_\theta}(s,a) \inf_{\gamma} \Big[\phi(-\gamma) + \phi(\gamma)u\Big] dsda \tag{23}$$

$$= \int_{s,a} \alpha\rho_{\pi_\theta}(s,a) \sup_{\beta\in\mathbb{R}} \big[- \beta u - f^*(-\beta)\big] dsda \tag{24}$$

$$\geq \sup_{T\in\mathcal{T}} \Bigg(\int_{s,a} \rho_{\pi_e}(s,a)T(s,a)dsda - \int_{s,a} \alpha\rho_{\pi_\theta}(s,a)T(s,a)dsda - \int_{s,a} \alpha\rho_{\pi_\theta}(s,a) f^*[T(s,a)]dsda \Bigg) \tag{25}$$

$$= \sup_{T\in\mathcal{T}} \Bigg(\mathbb{E}_{(s,a)\sim\rho_{\pi_e}(s,a)}[T(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(s,a)}[T(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(s,a)} f^*[T(s,a)] \Bigg) \tag{26}$$

Notice that we use a function $T(s,a)$ as an approximation of $-\beta$ and the loss $\beta = \phi(\gamma)$ is expected to be an non-negative value. Therefore we have the final results as follows,

$$\max_{T} \min\{0, \mathbb{E}_{(s,a)\sim\rho_{\pi_e}(s,a)}[T(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(s,a)}[T(s,a)]\} - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}(s,a)} f^*[T(s,a)]. \tag{27}$$

Proof of Theorem 2.

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

$$D^*(s,a) = \frac{\rho_{\pi_\epsilon}(s,a)}{\rho_{\pi_\epsilon}(s,a) + \frac{1-\alpha}{\alpha}\rho_{\pi_\theta}(s,a)}, \tag{28}$$

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

$$C + (1-\alpha)KL(\rho_{\pi_\epsilon}||\rho_{\pi_e}) + \alpha KL(\rho_{\pi_\theta}||\rho_{\pi_e}), \tag{29}$$

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 $D^*(s,a)$ achieves the value α.

Proof. For simplicity, we make a assumption that $\mathbb{E}_{(s,a)\sim\rho_{\pi_e}}\log[D(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}\log[D(s,a] \leq 0$ always holds and relax the $\min(\cdot)$ constraint in the objective function. The proof can be divided into two parts. In the first part, we treat π_θ as a constant and take the derivation of D to find the optimal discriminator. With the optimal D fixed, we can find the saddle point of UID objective function in the second part.

The outer optimization of Eq. (8) can be re-written as $\max_D V(D)$, where

$$\begin{aligned} V(D) &= \mathbb{E}_{(s,a)\sim\rho_{\pi_e}}[\log D(s,a)] - \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[\log D(s,a)] + \alpha\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[\log(1-D(s,a))] && (30)\\ &= \int \rho_{\pi_e}(s,a)\log D(s,a)dsda - \alpha\int \rho_{\pi_\theta}(s,a)\log[D(s,a)]dsda \\ &\qquad + \alpha\int \rho_{\pi_\theta}(s,a)\log(1-D(s,a))dsda && (31)\\ &= (1-\alpha)\int \rho_{\pi_\epsilon}(s,a)\log D(s,a)dsda + \alpha\int \rho_{\pi_\theta}(s,a)\log(1-D(s,a))dsda && (32) \end{aligned}$$

Taking the derivation of D_ψ in both sides, and $V(D_\psi)$ achieves its maximum when $\frac{\partial V(D)}{\partial D_\psi} = 0$. Therefore, we have

$$\frac{\partial V(D_\psi)}{\partial D_\psi} = \int \rho_{\pi_\epsilon}(s,a)\frac{1-\alpha}{D_\psi(s,a)}dsda + \int \rho_{\pi_\theta}(s,a)\frac{\alpha}{D_\psi(s,a)-1}dsda = 0 \tag{33}$$

Therefore, we obtain the optimal $D^*_\psi(s,a)$ is equal to $\frac{\rho_{\pi_\epsilon}(s,a)}{\rho_{\pi_\epsilon}(s,a)+\frac{\alpha}{1-\alpha}\rho_{\pi_\theta}}$. By fixing the optimal $D^*(s,a)$, we consider to find the saddle point of this minimax objective function. The inner optimization for policy training can be expressed as,

$$\begin{aligned} \min_\theta V(\pi_\theta) &= \int \rho_{\pi_e}\log[\frac{\rho_{\pi_\epsilon}}{\rho_{\pi_\epsilon}+\frac{\alpha}{1-\alpha}\rho_{\pi_\theta}}]dsda - \alpha\int \rho_{\pi_\theta}\log[\frac{\rho_{\pi_\epsilon}}{\rho_{\pi_\epsilon}+\frac{\alpha}{1-\alpha}\rho_{\pi_\theta}}]dsda \\ &\quad + \alpha\int \rho_{\pi_\theta}\log[1-\frac{\rho_{\pi_\epsilon}}{\rho_{\pi_\epsilon}+\frac{\alpha}{1-\alpha}\rho_{\pi_\theta}}]dsda && (34)\\ &= (1-\alpha)\int \rho_{\pi_e}\log[\frac{(1-\alpha)\rho_{\pi_\epsilon}}{(1-\alpha)\rho_{\pi_\epsilon}+\alpha\rho_{\pi_\theta}}]dsda + \alpha\int \rho_{\pi_\theta}\log[\frac{\alpha\rho_{\pi_\theta}}{(1-\alpha)\rho_{\pi_\epsilon}+\alpha\rho_{\pi_\theta}}]dsda && (35)\\ &= C + (1-\alpha)KL(\rho_{\pi_\epsilon}||\rho_{\pi_e}) + \alpha KL(\rho_{\pi_\theta}||\rho_{\pi_e}), && (36) \end{aligned}$$

where $C = (1-\alpha)\log(1-\alpha) + \alpha\log\alpha$. Since the KL divergence is always an non-negative value and $KL(a||b) = 0$ is satisfied only if $a = b$, so the saddle point is achieved when $\rho_{\pi_\theta} = \rho_{\pi_e}$ and $\rho_{\pi_e} = \rho_{\pi_\epsilon}$. At this point, the optimal discriminator $D^*_\psi(s,a)$ achieves the value α.

Experimental Details

Implementation Details

Our code is available at the **Code & Data Appendix**. We use the neural network that has two 100×100 fully connected layers and uses Tanh as the activation layer to parameterize policy, discriminator and value function in all adversarial imitation learning methods. To output continuous action, agent policy adopts a Gaussian strategy, hence the policy network outputs the mean and standard deviation of action. The continuous action is sampled from the normal distribution that is formulated with the action's mean and standard deviation.

Hyper-parameter chosen We use the same hyper-parameters in different tasks for the common part of adversarial imitation learning methods. For the discriminator D_ψ and the critic r_ψ, the learning rate is set to 3×10^{-4}. Five updates on the discriminator follow with one update on the policy network in one iteration. For the value function, the learning rate is set to 3×10^{-4} and three training updates are used in one iteration. We conduct on-policy method TRPO (Schulman et al. 2015) as RL step in our training, the learning rate is set to 3×10^{-4} with batch size 5000. The discount rate γ of the sampled trajectory is set to 0.995. The τ (GAE parameter) is set to 0.97.

Data Collection We provide two different kinds of imperfect demonstrations data (*i.e.,* **D1** and **D2**) to evaluate the performance of UID. We firstly train an optimal policy π_o by TRPO and π_o is used to sample optimal demonstrations $\mathcal{D}_o$. To collect imperfect demonstrations, 3 non-optimal demonstrators π_n are used. π_n in **D1** is obtained by saving 3 checkpoints with increasing quality during the RL training. In **D2**, we add different Gaussian noise ξ to the action distribution a^* of π_o to form non-optimal policy 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 (*i.e.,* π_{n_3}, π_{n_2} and π_{n_1}). The quality of each demonstrator is provided in Table 1.

We visualize the agent body of both π_o and π_n in the first 200 timesteps and use them to form animated GIFs with 20 FPS. From these GIFs, we can observe that agents with non-optimal policies can also move forward but merely slow compared to the agent with the optimal policy. This suggests that most imperfect demonstrations used in the experiment can be viewed as *reasonable-but-not-best* demonstrations. This kind of imperfect demonstrations are mostly common in the real world. The GIFs can be found in the **Multimedia Appendix**.

Task	Ant-v2	HalfCheetah-v2	Walker2d-v2
π_o	4349	4624	4963
π_{n_3}	2743	2587	2717
π_{n_2}	1570	1491	1699
π_{n_1}	194	-226	576

Task	Ant-v2	HalfCheetah-v2	Walker2d-v2
π_o	4349	4624	4963
π_{n_3}	3514	1853	4362
π_{n_2}	227	1090	467
π_{n_1}	-73	567	523

Table 1: Quality of different demonstrators in MuJoCo, measured by the average cumulative reward of trajectories. The left table is the data quality of **D1** demonstrations, the right table is the quality of **D2** demonstrations.

Compared Methods

We mainly compare UID with several *state-of-the-art* imitation learning with imperfect demonstrations methods. Specifically, several confidence-based methods (*i.e.*, 2IWIL (Wu et al. 2019), IC-GAIL (Wu et al. 2019) and WGAIL (Wang et al. 2021),) and preference-based methods (*i.e.*, T-REX (Brown et al. 2019) and D-REX (Brown, Goo, and Niekum 2020)) are compared. The rankings of trajectories in T-REX are given as a prior based on the demonstrator quality and we use the normalized cumulative reward of each demonstrator as the confidence for each demonstration. We briefly review the details of methods compared against UID in our experiments below. The RL part in these methods uses the same setting and share the same hyper-parameters.

2IWIL and IC-GAIL (Wu et al. 2019) We re-implement 2IWIL/IC-GAIL based on the official implementation[1]. In 2IWIL and IC-GAIL, a fraction of imperfect expert demonstrations are labeled with confidence (*i.e.*, $\mathcal{D}_l = \{(s_i, a_i), r_i\}_i^{n_l}$), while the remaining demonstrations are unlabeled (*i.e.*, $\mathcal{D}_u = \{(s_i, a_i)\}_i^{n_u}$). Since we have no access to the confidence score of the state-action pair in our setting, we use the normalized reward of the demonstrator as the confidence score for their related demonstrations. In the experiment, we choose 20% labeled demonstrations to train a semi-supervised classifier and then predict confidence for other 80% unlabeled demonstrations.

T-REX (Brown et al. 2019) and D-REX (Brown, Goo, and Niekum 2020) We re-implement T-REX and D-REX based on the official PyTorch code[2]. In T-REX, the preference of trajectories is based on the quality of the checkpoint. For example, the trajectory from π_o should be regarded as a better trajectory than that from π_n. In D-REX, the preference label of two trajectories is determined by the noise level of their corresponding demonstrators. With **D1** and **D2** demonstrations, we collect 5,000 random pairs of partial trajectories of length 50. We train the reward network using the Adam optimizer with a learning rate of 1e-4 and a batch size of 64 for 5000 iterations.

WGAIL (Wang et al. 2021) WGAIL is proposed to estimate confidence in GAIL framework without auxiliary information. The confidence $w(s, a)$ of each demonstration is calculated by $[(1/D^*_\psi(s, a) - 1)\pi_\theta(a|s)]^{\frac{1}{\beta+1}}$. The confidence estimation and GAIL training interacts during the training. Followed with the official implementation[3], β is set to be 1 and the early stop of confidence estimation is set at around 10% of total interactions.

[1]https://github.com/kristery/Imitation-Learning-from-Imperfect-Demonstration

[2]https://dsbrown1331.github.io/CoRL2019-DREX/

[3]https://github.com/yunke-wang/WGAIL

References

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.

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

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.