Appendix for ”Learning to Weight Imperfect Demonstrations”

1. Proof

1.1. Proof of Theorem 1

Proof. We have the constraints $0 \leq \widetilde{\pi}(a|s) \leq 1$ and $\sum_a \widetilde{\pi}(a|s) = 1$. Since $\widetilde{\pi}(a|s)$ is strictly positive, if $\sum_a \widetilde{\pi}(a|s) = 1$ is satisfied, then $0 \leq \widetilde{\pi}(a|s) \leq 1$ can be also satisfied. So we can convert object function into an unconstrained optimization problem by writing its Lagrange function as follows,

$$\mathcal{L}(\widetilde{\pi}, \lambda_s) = \sum_s d_\pi(s) \sum_a \widetilde{\pi}(a|s) A_\pi(s,a) - \beta D_f(\rho_{\widetilde{\pi}}||\rho_\pi) - \lambda_s[\sum_a \widetilde{\pi}(a|s) - 1] \tag{1}$$

where λ_s is the Lagrange multiple. The necessary condition for the extremum with respect to $\widetilde{\pi}$ is that derivative of $\widetilde{\pi}(a|s)$ to $\mathcal{L}$ equals to 0,

$$\frac{\partial \mathcal{L}(\widetilde{\pi}, \lambda_s)}{\partial \widetilde{\pi}(a|s)} = d_\pi(s) A_\pi(s,a) - \beta \frac{\partial D_f(\rho_{\widetilde{\pi}}||\rho_\pi)}{\partial \widetilde{\pi}(a|s)} - \lambda_s = 0. \tag{2}$$

Recall that $d_{\widetilde{\pi}}(s)$ is hard to sample and we sample from $d_\pi(s)$ instead, thus we have $d_\pi(s) \approx d_{\widetilde{\pi}}(s)$. Denote occupancy measure ρ_π can be also defined as $\rho_\pi = d_\pi(s)\pi(a|s)$, the derivative of $D_f(\rho_{\widetilde{\pi}}||\rho_\pi)$ can be calculated as,

$$\begin{aligned}
\frac{\partial D_f(\rho_{\widetilde{\pi}}||\rho_\pi)}{\partial \widetilde{\pi}(a|s)} &= \frac{\partial \sum_s d_\pi(s) \sum_a \pi(a|s) f\big(\frac{\widetilde{\pi}(a|s)}{\pi(a|s)}\big)}{\partial \widetilde{\pi}(a|s)} \\
&= d_\pi(s)\pi(a|s) \frac{\partial f\big(\frac{\widetilde{\pi}(a|s)}{\pi(a|s)}\big)}{\partial \widetilde{\pi}(a|s)} \\
&= d_\pi(s) \frac{\partial f(u)}{\partial u} \\
&= d_\pi(s) f'(u)
\end{aligned} \tag{3}$$

where $u = \widetilde{\pi}(a|s)/\pi(a|s)$ and f' is the derivative function of f. Apply derivative of $D_f(\widetilde{\pi}||\pi)$ into Eq. (2), we have

$$d_\pi(s)(A_\pi(s,a) - \beta f'(u)) - \lambda_s = 0 \tag{4}$$

Thus $f'(u)$ can be expressed as,

$$f'(u) = \frac{1}{\beta}\big[A_\pi(s,a) - \frac{\lambda_s}{d(s)}\big] \tag{5}$$

According to the definition of Legendre transformation, we have

$$f^*(p) = pu - f(u), \tag{6}$$

where $p = f'(u)$ should be satisfied. Then by applying derivative of p on both sides in Eq. (6), we have $f'_*(p) = u$. Combined with Eq. (5), u can be finally written as,

$$u = f'_*\big(\frac{1}{\beta}\big[A_\pi(s,a) - \frac{\lambda_s}{d(s)}\big]\big) \tag{7}$$

The solution of $\widetilde{\pi}$ can be exactly expressed as

$$\widetilde{\pi}(a|s) = \pi(a|s) f'_*\big(\frac{1}{\beta}(A_\pi(s,a) + C(s))\big) \tag{8}$$

where f'_* is the derivative of the conjugate function and $C(s) = -\lambda_s/d(s)$ is a constraint to ensure $\sum_a \widetilde{\pi}(a|s) = 1$. □

1.2. Proof of Theorem 2

Proof. We start from the known formula as follows,

$$\widetilde{\pi}(a|s) = \pi(a|s)\exp(\frac{1}{\beta}(A_\pi(s,a) + C(s))) \tag{9}$$

Take logarithm function on both sides, we have

$$\log\widetilde{\pi}(a|s) - \log\pi(a|s) = \frac{1}{\beta}(A_\pi(s,a) + C(s)) \tag{10}$$

Consider two action sets $\mathcal{A}_1 = \{a \in \mathcal{A} \mid \widetilde{\pi}(a|s) \geq \pi(a|s)\}$ and $\mathcal{A}_2 = \{a \in \mathcal{A} \mid \widetilde{\pi}(a|s) \leq \pi(a|s)\}$, for $\forall a_1 \in \mathcal{A}_1$, $\forall a_2 \in \mathcal{A}_2$ we have

$$\begin{aligned}\frac{1}{\beta}(A_\pi(s,a_1) + C(s)) &= \log\widetilde{\pi}(a_1|s) - \log\pi(a_1|s)\\ &\geq \log\widetilde{\pi}(a_2|s) - \log\pi(a_2|s)\\ &= \frac{1}{\beta}(A_\pi(s,a_2) + C(s))\end{aligned} \tag{11}$$

Thus we have $A_\pi(s,a_1) \geq A_\pi(s,a_2)$, which means there exsits a constant value ϵ satisfies $Q_\pi(s,a_1) \geq \epsilon \geq Q_\pi(s,a_2)$. Consider

$$\begin{aligned}&\sum_a \widetilde{\pi}(a|s)Q_\pi(s,a) - \sum_a \pi(a|s)Q_\pi(s,a)\\ =&\sum_{a\in\mathcal{A}_1}(\widetilde{\pi}(a|s) - \pi(a|s))Q_\pi(s,a) + \sum_{a\in\mathcal{A}_2}(\widetilde{\pi}(a|s) - \pi(a|s))Q_\pi(s,a)\\ \geq&\sum_{a\in\mathcal{A}_1}(\widetilde{\pi}(a|s) - \pi(a|s))\epsilon + \sum_{a\in\mathcal{A}_2}(\widetilde{\pi}(a|s) - \pi(a|s))\epsilon\\ =&\epsilon\sum_a \widetilde{\pi}(a|s) - \epsilon\sum_a \pi(a|s)\\ =&\epsilon - \epsilon = 0\end{aligned} \tag{12}$$

Define a two-stage value function $V_l(s)$ as follows,

$$V_l(s) = \begin{cases}\mathbb{E}_{a\sim\widetilde{\pi}(a|s)}\big(\mathbb{E}_{s',r|s,a}(r + \gamma V_{l-1}(s'))\big), & l \geq 1\\ V_\pi(s), & l = 0\end{cases} \tag{13}$$

which means $V_l(s)$ follows $\widetilde{\pi}$ in the first l steps and then follow π in the subsequent steps. When $l = 0$, $V_1(s) \geq V_0(s)$ for $\forall s \in \mathcal{S}$ has been proved in Eq. (12).

We use mathematical induction and assume $V_l(s) \geq V_{l-1}(s)$ for $\forall s \in \mathcal{S}$, we have

$$V_{l+1}(s) = \mathbb{E}_{a\sim\widetilde{\pi}(a|s)}\big(\mathbb{E}_{s',r|s,a}(r + \gamma V_l(s'))\big) \tag{14}$$

$$V_l(s) = \mathbb{E}_{a\sim\widetilde{\pi}(a|s)}\big(\mathbb{E}_{s',r|s,a}(r + \gamma V_{l-1}(s'))\big) \tag{15}$$

thus $V_{l+1}(s) \geq V_l(s)$ is proved, $\forall s \in \mathcal{S}$. So we can say that for

$$V_{\widetilde{\pi}}(s) \geq V_\pi(s), \forall s \in \mathcal{S} \tag{16}$$

□

2. Additional Experiment Details

2.1. Experiment Setting with Mujoco

Network Structure We use the neural network which has two fully connected layers and followed by Tanh as an activation layer to parameterize policy, discriminator and value function in weighted GAIL and GAIL. To output continuous action, agent policy adopts a gaussian strategy, hence the policy network outputs mean and standard deviation of action. The continuous action is sampled from the normal distribution with action's mean and standard deviation.

Hyper-parameters Selection We use the same hyper-parameters in different tasks for WGAIL and GAIL. The batch size of training is set to 5000. The discount rate γ of the sampled trajectory is set to 0.995. The learning rate of value function and discriminator are set to 3×10^{-4} and 1×10^{-3}. We also conduct early stop to weight estimation task since we need to control the discriminator in a near-optimal condition. The initial weight is set to 1 at the early training step and then update every 50 iterations to stabilize the GAIL training procedure. To choose proper β, we conduct experiments on Ant-v2 (Stage 2). The result is shown in Table 1, and as a result β is set to 1 in the Mujoco experiment. We use $1/\beta$ in the exponent for the first computation of weight.

Table 1. Performance with different β.

β	0	1	2	5	10
	119.49	182.00	123.03	106.35	109.71

Data Quality The quality of three policy checkpoints used in the experiment is available at Table 2. We suppose the demonstrations sampled from the same checkpoint share the same confidence score. The confidence score is the normalized reward of each checkpoint.

Table 2. The quality of checkpoints in different Mujoco tasks, which is measured by the average cumulative (normalized) reward of trajectories.

Task	$\mathcal{S}$	$\mathcal{A}$	**Random**	**Ckpt1**	**Ckpt2**	**Ckpt3 (Optimal)**
Ant-v2	$\mathbb{R}^{111}$	$\mathbb{R}^{8}$	992.18 (0.00)	1892.77 (0.29)	2813.69 (0.58)	4145.89 (1.00)
HalfCheetah-v2	$\mathbb{R}^{17}$	$\mathbb{R}^{6}$	-1.08 (0.00)	1437.14 (0.41)	2773.33 (0.78)	3546.63 (1.00)
Hopper-v2	$\mathbb{R}^{11}$	$\mathbb{R}^{3}$	73.21 (0.00)	789.96 (0.22)	2153.75 (0.64)	3323.99 (1.00)
Walker2d-v2	$\mathbb{R}^{17}$	$\mathbb{R}^{6}$	249.50 (0.00)	1704.23 (0.39)	3467.96 (0.85)	4018.19 (1.00)

2.2. Experiment Setting with Atari

A 3-layer conventional neural network in DQN is used in the policy network, with last 4 stacked frames as input. Notice that in our setting, we treat the end of the game instead of losing agent's life as the termination of an episode in Atari. To accelerate the training process, we adopt 8 CPU workers to sample demonstrations to fill in rollouts parallelly and a GPU server is responsible for updating π_θ and D_ψ with batched demonstrations provided by rollouts. The batch size of training is set to 1024. The discount rate γ is set to 0.99. The learning rate of value function and discriminator is set to 2.5×10^{-4} and 1×10^{-3}. β is set to 2 in Atari tasks. GAIL shares the same setting and hyper-parameters with weighted GAIL.