Appendix for "Robust Adversarial Imitation Learning via Adaptively-Selected Demonstrations"

1 Proof for Theorem 1

Theorem 1. *Denote that the δ_k is the upper bound of the KL-constraint step in the k_{th} iteration, i.e $D_{KL}(\pi_\theta^k, \pi_\theta^{k+1}) \le \delta_k$ and r_φ is bounded by M. If $\sum_{k>1}\sqrt{\delta_k}$ has an upper bound, then SAIL can converge to an optimal solution.*

Proof. Rewrite the object function of SAIL as $\min_\theta \max_{w,\varphi} \mathcal{J}_{w,\varphi}(\theta)$, where

$$\begin{aligned}\mathcal{J}_{w,\varphi}(\theta) = \ & \mathbb{E}_{(s,a)\sim\rho_{\pi_E}}\left[w(s,a)r_\varphi(s,a) + f(w(s,a))\right] \\ & - \mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[r_\varphi(s,a)] + \lambda\Psi(r_\varphi). \end{aligned} \tag{1}$$

We first show that for any given w, φ, $\mathcal{J}_{w,\varphi}(\theta)$ can converge with respect to θ obtained from policy gradient step in TRPO. If $\mathcal{J}_{w,\varphi}(\theta)$ can converge, $|\mathcal{J}_{w,\varphi}(\theta_k) - \mathcal{J}_{w,\varphi}(\theta_{k+m})|$ has an upper bound should be satisfied according to Cauchy's convergence test. By absolute value inequality, we can easily have

$$|\mathcal{J}_{w,\varphi}(\theta_k) - \mathcal{J}_{w,\varphi}(\theta_{k+m})| \le \sum_{i=1}^{m} |\mathcal{J}_{w,\varphi}(\theta_k) - \mathcal{J}_{w,\varphi}(\theta_{k+i})| \tag{2}$$

Due to the definition of Total Variance, i.e. $D_{TV}(p,q) = \frac{1}{2}\sum_x |p(x) - q(x)|$, where p, q is the probability distribution of variable x . The right hand side of Eq. 2 can be further written as,

$$\begin{aligned}\sum_{i=1}^{m} |\mathcal{J}_{w,\varphi}(\theta_k) - \mathcal{J}_{w,\varphi}(\theta_{k+i})| &= \sum_{i=1}^{m} |\mathbb{E}_{(s,a)\sim\rho_{\pi_{\theta_k}}}[r_\varphi(s,a)] - \mathbb{E}_{(s,a)\sim\rho_{\pi_{\theta_{k+1}}}}[r_\varphi(s,a)]| && (3)\\ &= \sum_{i=1}^{m} |r_\varphi(s,a)\sum_{s,a}(\rho_{\pi_{\theta_k}}(s,a) - \rho_{\pi_{\theta_{k+1}}}(s,a))| && (4)\\ &\le 2M\sum_{i=1}^{m} D_{TV}(\rho_{\pi_{\theta_k}}, \rho_{\pi_{\theta_{k+1}}}) && (5)\end{aligned}$$

The inequality can be satisfied with the assumption that r_φ is bounded by M as stated above.. According to Pinsker's inequality, which states that $D_{TV}(p,q) \le \sqrt{\frac{1}{2}D_{KL}(p,q)}$, we can further extend Total Variance into KL-divergence and connects $|\mathcal{J}_{w,\varphi}(\theta_k) - \mathcal{J}_{w,\varphi}(\theta_{k+m})|$ with δ_k.

$$|\mathcal{J}_{w,\varphi}(\theta_k) - \mathcal{J}_{w,\varphi}(\theta_{k+m})| \le \sum_{i=1}^{m} 2MD_{TV}(\rho_{\pi_{\theta_k}}, \rho_{\pi_{\theta_{k+1}}}) \le \sum_{i=1}^{m} 2M\sqrt{\frac{1}{2}D_{KL}(\pi_{\theta_k}, \pi_{\theta_{k+1}})} \le M\sum_{i=0}^{m-1}\sqrt{2\delta_{k+i}} \tag{6}$$

Notice that in the second inequality in Eq. (6), we replace ρ_π with π since the state distribution $d(s)$ of $\rho_{\pi_{\theta_k}}$ and $\rho_{\pi_{\theta_{k+1}}}$ is approximated to be the same in each TRPO step. The third inequality is also satisfied since $D_{KL}(\pi_{\theta_k}, \pi_{\theta_{k+1}})$ is bounded by δ_i in TRPO. According to Eq. (6), we can conclude that $\mathcal{J}_{w,\varphi}(\theta)$ is Cauchy so it can converge.

As stated above, $\mathcal{J}_{w,\varphi}(\theta)$ can converge with respect to θ, which means for any fixed w and φ we have $\mathcal{J}_{w,\varphi}(\theta_k)$ converges uniformaly to $\mathcal{J}_{w,\varphi}(\theta^*)$. On the other hand, $\mathcal{J}_{w,\varphi}(\theta)$ is a continuous and concave function of φ, so we have $\varphi_k^* = \arg\sup_\varphi \mathcal{J}_{w,\varphi}(\theta_k)$ also converges to $\varphi^* = \arg\sup_\varphi \mathcal{J}_{w,\varphi}(\theta^*)$. These results imply that there is a saddle point (θ^*, φ^*) in the proposed object function of SAIL and both θ and r_φ can converge to their optimal solutions alternately by gradient descent in Algorithm 1. As for weight w, since it is largely determined by the r_φ dynamically, it could also converge as r_φ reaches convergence. Another intuitive explanation on convergence of weight can be stated from self-paced learning, the weight could reach convergence since it is set to 1 at the end of the training.

2 Additional Experiments

Implementation Details

We use the neural network which has two 100×100 fully connected layers and followed by Tanh as an activation layer to parameterize policy, discriminator and value function in both GAIL-based methods. 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. Also, in Wasserstein GAIL, the output of the reward function r_φ is constrained to be in the interval $[0, 5]$ by sigmoid function. This is also consistent with the assumption that r_φ is bounded by M in Theorem 1.

Data Collection

To collect imperfect demonstrations, we firstly train an optimal policy π^* by TRPO, then we add different Gaussian noise ϵ to the action output by π^*, therefore different noisy policy can be formed. We choose $\epsilon = [0.01, 0.25, 0.4, 0.6, 0.8, 1.0]$ to form six different noisy policies (i.e. $a_i \sim \mathcal{N}(a^*, \epsilon_i^2)$, where a_i and a^* are the actions from i_{th} noisy policy and optimal policy), the quality of which is provided in Table 1. We consider using two different stages of demonstrations for training. Stage 1 demonstrations are formed by trajectories from π_1, π_2, π_3 and π_4 while stage 2 demonstrations are formed by trajectories from π_1, π_2, π_5, π_6, so the quality of stage 1 demonstrations is obviously better.

Table 1: Quality of different demonstrators, measured by the average cumulative reward of trajectories.

	Ant	HalfCheetah	Walker2d	Swimmer
π_1	3985	4311	4434	95.24
π_2	3514	1853	4362	76.55
π_3	227	1090	467	56.41
π_4	-73	567	523	43.65
π_5	-979	-45	283	28.23
π_6	-203	-177	249	17.85
π^*	4349	4624	4963	100.78
π_0	995	-0.58	131	-3.96

Hyper-parameter chosen

We use the same hyper-parameters in different tasks for the common part of GAIL-based methods. For discriminator D_ψ or critic r_φ, the learning rate is set to 3×10^{-4}, 5 training epochs are used with batch size 256. For the value function, the learning rate is set to 3×10^{-4}, 3 training epochs are used with batch size 128. We conduct TRPO as RL step in our training, the learning rate is set to 3×10^{-4} with batch size 1000. The discount rate γ of the sampled trajectory is set to 0.995. The τ (GAE parameter) is set to 0.97. In SAIL and InfoSAIL, m is set to 3 for updating the threshold K.