Thank reviewers for the comments. Please find our responses below, *with reference indices consistent with the paper.*

<u>**To reviewer 3.**</u> **Q3-1: Experimental choices.** Imitation learning (IL) aims to match the state-action distributions between the learner and the expert, rather than the goal feature(s). The expert state-action distributions from the data are high-dimensional and stochastic. Taking CartPole as an example, Fig. 1 (right) shows the expert (stochastic) policy on two sample states, where the states are from a 4-dimensional continuous space (see Appendix B in [18]). Moreover, Fig. 1 (left) shows that in the expert demonstrations, the angle (feature) is NOT simply a constant (over 250 trajectories in CartPole). In addition, for fair comparisons, our evaluation tasks/settings are consistent with SOTA [13,18,20]. **Q3-2: Sample efficiency.** Results on sample efficiency are presented in Fig. 5 in the paper, where f-GAIL outperforms baselines over different sample sizes (Sec 4.2). Moreover, it is true that f-GAIL can more accurately estimate f-divergence from a limited number of samples. Fig. 2 below shows that given a task, the learned f^* functions are consistent with different sample sizes. In fact, the choice of f-divergence matters. The better divergence estimation accuracy enables f-GAIL to examine and compare f-divergence choices, which is why f-GAIL consistently outperforms baselines.

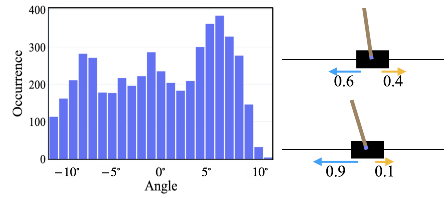

Figure 1: Angle dynamics (left) and policy (right) in CartPole from expert data.

Q3-3: Meaning of the "best" f-divergence. Our f-GAIL is defined as a minimax optimization problem in eq.(5) in the paper. The best f-divergence is searched in the "max" inner-loop *given the current learned policy* π learned from the "min" outer-loop, eventually leading to a stable solution of (π, f^*). **Q3-4: The optimality depends on the divergence and context?** The notion of optimality exists and depends on the expert demonstration data, rather than the divergence, namely, the optimality refers to the smallest discrepancy of behavior distributions (in state-action pairs) between the learner and the expert. Given an expert demonstration dataset, a better divergence can measure the discrepancy more precisely than other divergences, thus enable training a learner with closer behaviors to the expert. In the example of whether mode-seeking or -covering makes sense, it depends on the expert demonstration data (NOT the context), i.e., whether the expert was performing mode-seeking or -covering when generating the data. **Q3-5: Meaning of the learned divergence? The variance of the learned divergence?** It is nontrivial to find an analytical close-form function to express the learned f^*, due to the huge convex function space with $f(1) = 0$. We leave this as our future work. Fig. 2 above shows that given a task, the learned f^* functions are consistent (small variance) for different sample sizes. **Q3-6: Comparison with BC.** We agree that BC minimizes the policy KL divergence as what we noted in Sec. 4 (line 200). We included BC as a baseline for completeness, namely, a comprehensive comparison with SOTA. **Q3-7: Notation of $\mathcal{P}$ in line 62.** Our notation represents a probability distribution of transitioning from (s, a) to a next state s', thus the outcome is in $[0, 1]$. It is consistent with the literature, e.g., Sec. 2 in [Yu et al. arXiv:1909.09314].

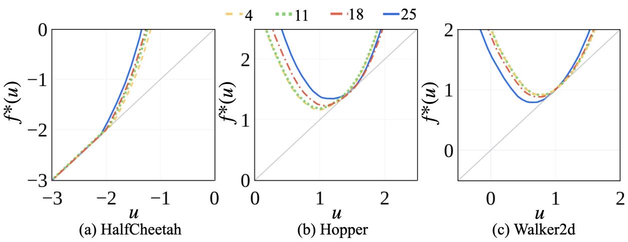

Figure 2: The learned $f^*_\phi(u)$ with different sample sizes

<u>**To reviewer 2.**</u> **Q2-1: Necessity and sufficiency of two constraints.** The f-divergence definition requires the generator f function to be convex and $f(1) = 0$ [11,23,24]. Convex and zero-gap constraints are necessary and sufficient conditions to guarantee an f-divergence, based on $f^{**} = f$ (see §3.3.2 in [9]) for convex functions, i.e., $f(1) = f^{**}(1) = \max_u\{u - f^*(u)\} = 0$. **Q2-2: Implementation details.** We used 5 random seeds with mean and variance calculated over 50 trajectories (see Sec. 4 and Appendix B). These settings are consistent with SOTA [11,23,24]. **Q2-3: Baselines.** All baselines were implemented with their original models [13,18,20], rather than $f^*(T(s, a))$. In fact, as shown in Tab. 1 (with GAIL as the original model and GAIL$_f$ as $f^*(T(s, a))$), the baseline results are similar, when implemented using original models vs $f^*(T(s, a))$. **Q2-4:** The input state distributions were sampled from expert demonstrations. **Q2-5: Divergence evaluation.** Following your suggestion, Fig. 3 below shows the training curve of f-divergence wrt. epochs where it converges to less than 0.02 for HalfCheetah after 450 epochs. Similar results were observed in other tasks.

Table 1: Baseline performances in HalfCheetah.

Datasize	GAIL	GAIL$_f$
4	4047±344	4055±257
25	4340±185	4472±166

<u>**To reviewer 1.**</u> **Q1-1: Novelty:** We are the *first* to model imitation learning with a learnable f-divergence measure (using the proposed f^*-network), rather than a predefined divergence, which yields better learner policies than the literature on GAIL [13,18,20]. **Q1-2: More complex tasks:** We evaluated f-GAIL on tasks consistent with SOTA [13,18,20], including Humanoid, with the high state dimension of 376. We plan to evaluate f-GAIL on more complex tasks, e.g., Simitate [Memmesheimer et al. arXiv:1905.06002].

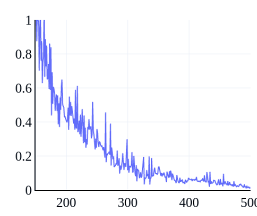

Figure 3: f-divergence curve in HalfCheetah.

<u>**To reviewer 5.**</u> **Training objective:** In our Alg 1, all three networks are trained with the same objective in eq.(5), using *adversarial training*. The updating gradients (∇_ω and ∇_ϕ) are obtained by taking the derivative of eq.(5) wrt. ω and ϕ, respectively, while fixing π_θ. The objective for policy π_θ is the same as eq.(5) with T_ω and f^*_ϕ fixed. We will add such details in the final paper.