We would like to thank the reviewers for their thorough evaluations and for bringing to our attention some missing citations and typos, these will be corrected in the updated manuscript. We answer here specific questions raised by the reviewers and present requested additional experiments.

Additional MuJoCo experiments (R1, R2, R4). We focus on GAIL because AIRL claims to perform on par with GAIL on MuJoCo. We present in Figure 2 below additional results on the MuJoCo envs (as well as the additional Ant env) where GAIL has been re-tuned for further improvement and SQIL and ASAF-1 have been added. We see that even with careful tuning GAIL is outperformed by our method and that SQIL's instability is exacerbated on MuJoCo.

We also ran ASAF-1 on the Ant-v2 MuJoCo environment using various sets of 25 demonstrations ***(as requested by R2)***. These demonstrations were generated from a Soft Actor-Critic agent at various levels of performance during its training. Since at low-levels of performance the variance of episode return is high, we filtered collected demonstrations to lie in the targeted range of performance (e.g. return in [800, 1200] for the 1K set). Results in Figure 1 show that our algorithm succeeds at learning a policy that closely emulates various demonstrators (even when non-optimal).

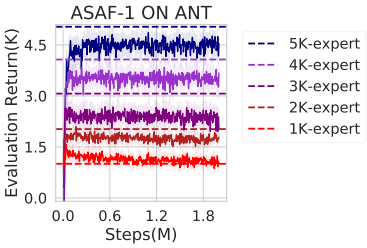

Figure 1: ASAF-1 on Ant-v2. Colors are 1K, 2K, 3K, 4K, 5K expert's performance.

Comparison of our methods vs Finn et al. [1] and Fu et al. [2] and Behavioral Cloning (BC) (R1, R3). Our work builds on [1], yet its novelty is to explicitly express the probability of a trajectory in terms of the policy in order to directly learn this latter when training the discriminator. In contrast, [2] considers a transition-wise discriminator with un-normalized probabilities which makes is closer to ASQF (Appendix B) than to ASAF-1. Additionally, AIRL [2] minimizes the Kullback-Leiber Divergence [3] between occupancy measures whereas ASAF minimizes the Jensen-Shanon Divergence between trajectories likelihood. Finally, BC uses the loss function from supervised learning (classification or regression) to match expert's actions given expert's states and suffers from compounding error due to co-variate shift [4] since it only learns on the expert state-action visitations (demonstrations) without environment interaction. Contrarily, ASAF-1 uses the binary cross entropy loss in Eq. (13) and does not suffer from compounding error as it learns on both generated and expert's trajectories.

Can windowed approach be used for GAIL and AIRL? (R1). GAIL's and AIRL's discriminators are updated based on transitions so that a step-wise reward can be learned and used in RL loops. Therefore, the windowed approach, which suits trajectory-wise formulations, could not be applied without major modification of their formulations.

Reward Acquisition from ASAF (R3). Although ASAF does not explicitly acquire reward during training, we can retrieve step-wise soft advantages $\log \pi(a|s)$ from learned agent's policy $\pi(a|s)$ which can be used as a reward function [2, 5].

Training Time (R4). Due to lack of room we cannot add here the equivalent of Figure 2 with wall clock time as x-axis but we will add it to the updated manuscript. Our observation is that ASAF-1 is always fastest to learn, e.g., 361.2s (ASAF-1), 473.1s (SQIL), 1561.0s (ASAF), 1763.6s (GAIL), 2079.1s (ASAF-w) were taken to reach a performance of 2000 in Hopper environment. Note however that reports of performance w.r.t wall-clock time should always be taken with a grain of salt as they are greatly influenced by hyper-parameters and implementation details.

No entropy regularization in loss (R3). We learn in the softmax policy class of Eq. (2) since it contains the expert's policy (See Section 3.2) but optimize Eq. (13) without entropy regularization as was done in the GAIL paper.

Additional Concerns. (To R3) We will update our related works with your recommendations. ***(To R4)*** It seems that our supplementary files have been correctly uploaded since R1 was able to read through our code. We are sorry that you couldn't access it. Unfortunately, we are not allowed to put an external link in this rebuttal. We suggest that you contact the AC about this issue.

References. [1] Finn et al., "A connection between generative ...," (2016) [2] Fu et al., "Learning robust rewards ...," (2017) [3] Ghasemipour et al., "A divergence minimization perspective ...," (2019) [4] Ross and Bagnell, "Efficient reductions for imitation learning," (2010) [5] Schulman et al., "High-dimensional continuous control ...," (2015)

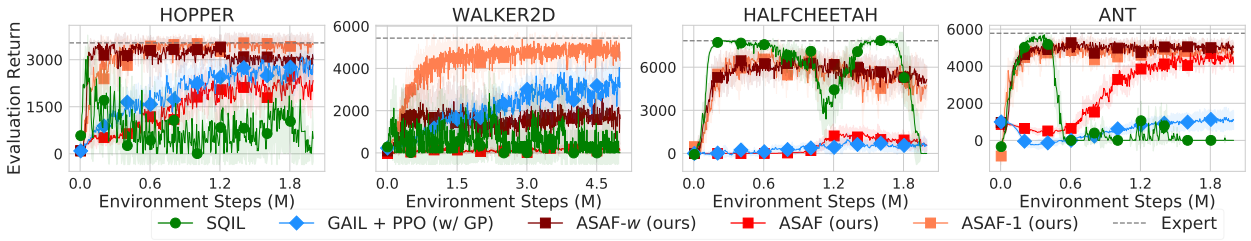

Figure 2: Results on Mujoco environments with improved tuning for GAIL, added SQIL and ASAF-1 and env (Ant)