**Response to Reviewer 1** Thanks for your useful suggestion. We will add more discussion about the generalization bounds of other imitation learning methods and fix the mistake in the proof of Lemma 8 in the future version.

**C1**: We have provided a PAC result for BC in Corollary 10 in the Appendix for a direct comparison. Please note that the PAC bound is based on sample-approximation, hence the dependency on the effective horizon does not change.

**C2:** Rademacher complexity in this paper refers to its empirical version and we will clarify this in the future version.

**Response to Reviewer 2** Thanks for your insightful comments.

**Q1:** The choice of 3 expert trajectories and its effect?

**A1:** We choose 3 trajectories to better demonstrate that when the estimation error in PAC bounds is large, the $\gamma$-dependency of GAIL is better than BC. Consistent with prior works [20, 27], we do observe that when more expert trajectories are provided (e.g., 10 trajectories), BC achieves comparable performance to GAIL since the estimation errors for both methods are very small and the $\gamma$-dependency does not dominate.

**Q2:** Results of model learning experiment?

**A2:** We use policy evaluation errors to evaluate the quality of model learning. The results in Section 6.2 indicate that the environment recovered by GAIL could be better.

**C1:** The shading on plots refers to the standard deviation over 3 random seeds. We will clarify this in the future version.

**Response to Reviewer 3** Thanks for your helpful advice. The word "generation" means that the policy could perform well on unseen states in the training MDP, and we will clarify this in the future version.

**Q1:** The difference regarding the horizon settings between theoretical analysis and experiments.

**A1**: We cannot run infinite steps in an infinite-horizon MDP if there are no absorbing states. Therefore, we choose to use a finite horizon MDP with a large horizon $H = 1/(1-\gamma)$, such that there is only a small difference to the infinite horizon. To see this, assuming the reward function is bounded within $[0, 1]$, we have $|\sum_{t=H}^{\infty} \gamma^t r(s_t, a_t)| \leq \epsilon \Longrightarrow H \geq \frac{1}{1-\gamma} \log(\frac{1}{(1-\gamma)\epsilon})$.

**Others:** For your choice of "No" to the reproducibility evaluation, we would like to point out that the proof, source code and relevant dataset are provided in the supplementary material to help reproduce the work.

**Response to Reviewer 4** We thank you for your insightful suggestion. We will enlarge the picture in the future version.

**Q1:** The assumption on why [35], as well as [1] and [46] as proposed in [21], are used instead of these other approaches (MaxEnt IRL)? The comparisons provided in the paper are sufficient?

**A1:** Since our main contribution is the theoretical results, the conducted experiments are used to verify the theory. Note that GAIL [20] is closely related to MaxEnt IRL by incorporating the $\gamma$-discounted causal entropy (see Equation (1) in [20]). To remedy your concern, here we provide an additional experiment that considers the SOTA MaxEnt IRL method called AIRL [Fu et al., ICLR 2018], which will be included in the paper. We would like to consider the error bounds of MaxEnt IRL in the future.

Compared to prior works [20, 26, 27], we have additionally considered the BC-like method DAgger [39] and we think the empirical comparisons are sufficient, though the empirical comparisons are not the main scope of this paper.

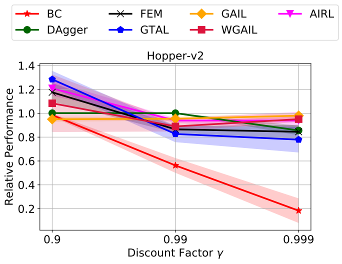

Figure 1: Performance on Hopper-v2.

| $\gamma$ | Expert | BC | GAIL | AIRL |
|---|---|---|---|---|
| 0.9 | $10.85 \pm 0.00$ | $10.69 \pm 0.10$ | $10.31 \pm 0.19$ | $13.11 \pm 0.65$ |
| 0.99 | $275.81 \pm 0.00$ | $155.19 \pm 16.27$ | $263.17 \pm 3.47$ | $258.96 \pm 4.50$ |
| 0.999 | $2223.49 \pm 0.00$ | $408.25 \pm 222.42$ | $2177.76 \pm 43.64$ | $2087.75 \pm 154.99$ |

Table 1: Discounted returns of learned policies on Hopper-v2 (performance of other methods can be found in the original Appendix). We use ± to denote the standard deviation.

**Q2:** The paper feels incomplete. The authors mention "...".

**A2:** The main contribution of this paper is to provide error bounds on imitating policies and environments and to give insights for future algorithm designs. For MBRL, our theoretical results suggest that environment-learning by GAIL could be better. Combining the environment-learning with all kinds of policy optimization algorithms, while is very interesting, is beyond the main scope of this paper and is left for future work.