We thank all the reviewers for their constructive feedback. We address the key questions and concerns below.

The role of ρ-projection and relationship to $p(\tau)$ (R1, R3 and R4): To clarify, ρ-projection in Eq. 6 is *not* an approximation of $p(\tau)$, despite the similar forms. $\mathcal{F}(\tau)$ in the denominator of ρ-projection is sampled to have the same starting point and length as τ; as such, it may not cover the space of all trajectories and hence, does not approximate $Z(\theta)$ even with large M. The appealing property of ρ-projection is that the partition function is cancelled off from the numerator and denominator, thereby eliminating the need to approximate it. This is shown in Eq. 1 below.

$$\rho_\tau(\theta) \triangleq \frac{p_\theta(\tau)}{p_\theta(\tau) + \sum_{\tau' \in \mathcal{F}(\tau)} p_\theta(\tau')} = \frac{e^{R_\theta(\tau)}/Z(\theta)}{e^{R_\theta(\tau)}/Z(\theta) + \sum_{\tau' \in \mathcal{F}(\tau)} e^{R_\theta(\tau')}/Z(\theta)} = \frac{e^{R_\theta(\tau)}}{e^{R_\theta(\tau)} + \sum_{\tau' \in \mathcal{F}(\tau)} e^{R_\theta(\tau')}} \quad (1)$$

Handling other forms of policy invariance (R2) Scaling of reward is not a valid form of policy invariance as it changes the corresponding optimal *stochastic* policy and therefore, our projection will *correctly* map them to different points. Therefore, this is not a valid counterexample to ρ-projection's handling of other forms of policy invariance.

Additional empirical comparisons (R1, R3 and R4): Regarding comparisons to GAIL and Deep ME-IRL: GAIL was not compared against because it is an imitation learning algorithm that retrieves policies directly and does not return a reward function. Deep ME-IRL is only applicable to the two discrete environments, out of which Borlange is infeasible due to the large state space (>20,000 states). When applied to Gridworld, Deep ME-IRL performance was very poor; it converged to ≈18% of expert's ESOR.

The ESOR values in Table 1 shows the number of iterations taken to reach expert's ESOR. For point mass maze, the success rates and ESOR values show the compared algorithms failed to reach the expert's ESOR *within* the set budget (100 iterations for AIRL & GCL; 50 for BO-IRL). Increasing the limit to 1000 for AIRL & GCL and 100 for BO-IRL results in higher success rates (Table 1 below). Table 1 also includes BIRL as suggested by **R1** using a) Mean of the samples collected thus far (BIRL-Mean) and b) the Policy Walk algorithm. BO-IRL with ρ-RBF outperforms the rest in success rates and the iterations required in most scenarios. Even though BIRL (Policy Walk) has a higher success rate in Gridworld, it comes at the cost of significantly higher iterations compared to our method. We will include the above results in the revision.

Relationship to Active Learning (AL) in IRL (R1). At first glance, BO-IRL might look similar to AL since they both use uncertainity measures to calculate the next query. However, they differ in the type of query used. AL queries the most informative states (actions) for additional trajectories while BO-IRL queries the likelihood of a reward function given a fixed set of demonstrations.

Time-complexity (R3): K only affects the calculation of covariance and is linear in complexity.

Run-time comparison (R1) In our experiments, the GAN-based methods currently have a faster run-time per iteration since they apply approximate policy evaluation. However, the core advantage of our method is the efficient exploration of the reward function space in order to identify multiple valid reward functions. GAN-based methods require numerous runs with random initialization to identify multiple valid reward functions which will increase their runtime.

Significance of theoretical contributions (R4): Our main contribution is a Bayesian Optimization approach to IRL. We provide Theorem 2 to support our approach—Theorem 2 is important as it formalizes the key idea that the ρ-projection maps PBRS-based policy invariant rewards to a single point. We do not claim Theorem 1 as a contribution and we state explicitly in the paper (lines 147 and 152) that it is from [20]. We re-state the theorem and Corollary 1 in the paper for the reader's convenience.

Definition 1 is not ill-posed since $\mathcal{F}(\tau)$ is a deterministic hyper-parameter rather than of a random variable; the value of $\mathcal{F}(\tau)$ is fixed in each run rather than sampled throughout.

		Gridworld		Börlange		Point mass maze	
Algorithm	Kernel	SR	Iterations	SR	Iterations	SR	Iterations
BO-IRL	ρ-RBF	70%	**16.0**±15.6	**100%**	**2.0**±1.1	**80%**	**51.4**±23.1
	RBF	50%	30.0±34.4	80%	9.5±6.3	20%	28.0±4
	Matérn	60%	22.2±12.2	100%	5.6±3.8	20%	56±29
BIRL (Mean)		60%	560.3±206.4	0%	-	N.A	
BIRL (Policy Walk)		**80%**	630.5±736.9	80%	98±167.4	N.A	
Deep ME-IRL		0%	–	Intractable		N.A	
AIRL		70%	70.4±23.1	100%	80±36.3	80%	90.0±70.4
GCL		40%	277.5±113.1	80%	375±68.7	0%	–