We thank all reviewers for their valuable time and feedback. We believe that we have addressed the main concerns about the assumptions, related work, and evaluation.

CONTENT

R2, learning the features. The linear MDP assumption $\Pi\Phi = \Phi M$ and $\mathbf{r} = \Phi w$ implies that the transition matrix Π is low-rank and that rewards are linear in the features, so these assumptions can be used to guide feature learning. For example, for discrete actions we can set $\phi(s,a) = \phi(s) \otimes \delta(a)$, and make $\phi(s|\theta)$ the output of a neural network. The objective would be to minimize $L(\theta, K, w) := E_{(s,a,r,s')\sim\mathcal{D}}\left[c_1\|\phi(s'|\theta) - K\phi(s,a|\theta)\|^2 + c_2(r - \phi(s,a|\theta)^\top w)^2\right]$. Note that multiple recent works (offline and online) simply assume a linear MDP with known features in analysis.

R2, violation of assumptions. In terms of exploration, we do not require the logging policy to cover all states, but just to span the features (feature covariance matrix should be full-rank, assumption A3). If this is not the case, the error would depend on the missing subspace. In terms of linearity, for non-linear MDPs our method would incur an approximation error. Separating the linearity of dynamics and rewards (rather than implicitly assuming both with a linear Q-function) at least allows us to drop the linear-reward assumption if we learn the full distribution (simulated in Figure 1, middle).

R3, R4, necessity of entropy. Our finite-sample guarantees for J_π actually hold for any distribution satisfying the constraint (see Remark 1) as long as the MDP is linear, and we can also use the simple estimate in Equation (10). Empirically, we find entropy to be a good regularizer when learning the full distribution $d_\pi(s,a) = \mu_\pi(s) \otimes \pi(a|s)$. Learning d_π is useful in the case of non-linear rewards, or $E_d[r(s,a)]$ cannot be computed in closed form. Appendix D gives a justification of the maximum-entropy objective for MDPs with sufficiently random dynamics. More generally, maximizing entropy is equivalent to minimizing KL-divergence to the uniform distribution. We can use the same KL-divergence formulation to impose different distribution priors when available.

R3, comparison to bounds in Duan and Wang (2020). These bounds are expressed in terms of χ^2 divergence, and for linear f, they are a function of the spectrum of the feature covariance matrix Σ. This also the case with our bound - it scales inversely with the smallest eigenvalue of Σ. The results are admittedly different otherwise (we will clarify this in the paper). Note also that the bounds of Duan and Wang (2020) scale with the horizon / effective horizon, and are thus infinite in our setting. Furthermore, their discounted infinite-horizon bound scales as $N^{-1/2}$ where N is the number of trajectories, whereas our bound scales as $T^{-1/2}$ where T is the number of transitions (possibly in a single trajectory).

EXPERIMENTS

R1, R3, we agree the the methods are not always well-separated in the sense of 95% confidence intervals, and BRM and FQI can perform well. However, note that some of the evaluated environments violate our assumptions (Taxi, Acrobot), and on these our method performs at least as well as the baselines. The only evaluated environment satisfying linearity and ergodicity is LQ control (Figure 2 left), where our approach is clearly better. Concretely, the model-based method is clearly not sensitive to ε_1 (controlling level of suboptimality of the target policy), while BRM and FQI deteriorate with larger ε_1, and the curves are well-separated for $\varepsilon_1 \geq 0.2$.

R2, for Acrobot, it is an interesting observation that MaxEnt and Model often underestimate the expected reward (actually so does FQI). However, for smaller true J_π, the same methods overestimate. This may be an artifact of covariance regularization for linear regression in the model / Q-function.

R3, we chose not to explicitly demonstrate divergence of FQI and biasednesss of BRM since these issues are well known (see references in Section 4).

STYLE. **R2**, thank you very much for the suggestions. We agree about Section 3.3 and in retrospect should have saved a discussion of policy improvement for future work. We will remove it and move some of the Appendix into the paper.

RELATED WORK

R2, thank you for the suggestions. We will add references to maximum-entropy approaches in RL and IRL. In RL, entropy is typically a regularizer on the policy, rather than the state-action distribution. IRL approaches are more similar to ours, maximizing the entropy of distribution over paths s.t. feature expectations match demonstration data.

R4, thank you, we will improve the discussion. We actually do refer to the "Breaking the Curse of the Horizon" work (lines 24,65, 204-208). Their formulation also relies on the relaxed feature expectation constraint. They minimize the constraint violation while maximizing over features, and have a consistency guarantee. We assume oracle features and a linear MDP, satisfy an approximate constraint (given by the model estimate), and show finite-sample guarantees.

R3, R4. We will add a reference to GenDICE. In the paper, we refer to "RL via Fenchel-Rockafellar Duality" by Nachum and Dai (2020), which provides a unified view of the DICE papers, including GenDICE. Section 7 there discusses the undiscounted setting via different DICE methods. We compare to the basic Lagrangian formulation in our paper (lines 195-205), and will add a note on GenDICE (regularized Lagrangian).