We thank all the reviewers for their thoughtful feedback. We highlight that all the reviews were positive with a few specific questions, which we hope to address in our response below.

Reviewer # 2

1. Hyper-parameters : We will include an ablation study in the final version of paper.
2. How to collect "good" data for a new task : We first wish to clarify that in offline RL, the primary motivation is to use existing offline datasets. As a consequence, we typically treat the dataset as fixed and given to the agent; as opposed to the agent (or researcher) having the luxury of choosing the dataset. Having said that, an ideal dataset for offline RL would be one that not only has high support overlap with the optimal policy, but also one that enables a large hitting time, as suggested by our proposition and lower bound. Designing exploratory policies for purposes of data collection is an exciting direction for future work but outside the scope of current submission.
3. Choice of NPG : We used NPG for its conceptual and implementation simplicity. A number of prior papers have successfully used NPG and shown impressive results. Furthermore, our MOReL framework is modular and has a clear separation between learning the P-MDP and optimizing a policy in the P-MDP. We conjecture that most algorithms (e.g. PPO, SAC etc) for optimizing the policy in the P-MDP would yield similar results.

Reviewer # 5

1. Contributions of our work : To our knowledge, our work is the first to study a model-based approach to offline RL, apart from Ross et al. which provided negative results for a naive algorithm. While there has been extensive work on model-based RL and offline RL individually, their intersection has been explored only sparsely. As most reviewers concurred, offline RL is an important learning paradigm that can expand the applicability of RL. We develop a new framework for offline RL that utilizes learned models and show that it is mini-max optimal. We also demonstrate state of the art experimental results.
   Our survey of related work is extensive with 86 citations (kindly also see expanded related works in appendix). We are also happy to cite and discuss any additional related work that the reviewers may point out.
2. Theoretical insights : While it is intuitively clear that if a policy does not drive too quickly towards unknown states, Theorem 1 presents a *precise, quantitative* bound using *hitting times*. Corollary 3 further bounds this in terms of mismatch in the support of state-action visitation distributions. In contrast, prior works only consider settings where there is no support mismatch. Proposition 4 shows that the bound in Corollary 3 is best possible up to logarithmic factors, demonstrating minimax optimality of MOReL.
3. Use standard errors in table : Thank you for the suggestion. We followed common practice to report standard deviations, but we are happy to report standard errors if it is more appropriate in the reviewer's opinion. Note however that our claim of SOTA results in 12 out of 20 environments is based on our average scores, which remains unaffected by choice of error bars. We also highlight that prior work does not report any error bars, and also tune hyper-parameters on a seed-specific basis. In contrast, we use the same hyper-parameters across all seeds.
4. Proposition 4 : The value of 0.95 comes from requiring γ to satisfy certain inequalities in Lines 579 and 581 in Appendix A. Since our goal is to show that the $(1-\gamma)^{-2}$ dependence in Corollary 3 is optimal, it is okay to assume that $\gamma \in [0.95, 1]$ (since, if $\gamma < 0.95$, then $(1-\gamma)^{-2}$ is bounded by a constant).

Reviewer # 6

1. Alternate ways to penalize uncertain states : Our particular approach to penalizing unknown states enables detailed theoretical analysis while also demonstrating SOTA experimental results on well studied domains that require function approximation. It would make for an interesting future work to study if similar results (theoretical and/or experimental) can be obtained with alternate approaches, but is outside the scope of our submission.
2. Choice of α : Note that there are two competing terms in the sub-optimality bound (Corollary 1). Decreasing α decreases $(1-\gamma)^{-2} \cdot (4\gamma R_{\max}) \cdot \alpha$, but also has the effect of decreasing the number of "known" states. This in-turn reduces the hitting time, and increases $(1-\gamma)^{-1} \cdot 2R_{\max} \cdot \mathbb{E}\left[\gamma^{T_U^{\pi^*}}\right]$ in the bound. Thus, an appropriate choice of α that balances the two terms is required, and can be treated as a hyper-parameter.
3. Hyper-parameters : We will include an ablation study of hyper-parameters in the final version of the paper.
4. Comparison to more recent work : Thank you for the pointers to these very interesting papers! First, we wish to highlight that these are *very recent* papers making comparisons with them difficult – especially at the time of NeurIPS submission. Furthermore, these papers report results on non-standard domains compared to most prior work. For example, ABM uses DeepMind control suite while BOPAH uses a different data logging policy. This makes a direct comparison with published results impossible. In this submission, we used identical setups to most prior papers for a fair and transparent comparison.