Thanks for the reviews! Response to individual reviewers are below.

R1. You are right that the contribution is theoretical, but our formulation of the representation learning problem and the computational oracles assumed are heavily motivated by practice. The FLAMBE algorithm is not difficult to implement with modern deep learning libraries. In addition, modern RL algorithms can be employed, in the planning step (Eq. (2) in Algorithm 2, see e.g. the recent preprint `https://arxiv.org/abs/2007.08459`), as well as in the maximum likelihood step (e.g., using VAEs as in recent model-based RL works such as `https://arxiv.org/abs/2005.05960`). To summarize, our theory directly motivates new strategies for representation learning and exploration that we believe will be empirically effective, and we are excited about experimenting with these approaches in future work.

R2. We believe the algorithm is actually quite precisely written and intuitive. Perhaps the simpler algorithm (and simpler analysis) for the simplex features setting provided in the appendix is useful for added intuition. The algorithm at a high-level can be seen as doing MLE to estimate features, and using an algorithm for low-rank MDPs with known features to compute a good policy. Inducting these two at each step $h = 0, \dots, H$ with some careful attention to details gives the FLAMBE algorithm. For specifics:

- We believe you may have missed the role of the exploratory policy ρ_h which is computed in iteration $h-1$ of the algorithm (via a call to Algorithm 2) and is used to collect the data for the MLE step in iteration h. Concretely, our model-based planning step (call to Alg 2) returns an exploratory policy ρ_h^{pre}, which is executed for $h-1$ steps, followed by *two* uniform actions at steps h and $h+1$ to collect the samples for MLE estimation at level $h+1$. We need two random actions as planning stays one step behind model learning, as briefly discussed in lines 291–295. We *do not* assume access to a generative model; we are in the online exploration setting and collect data by the learned exploratory policies ρ_h. The use of such exploratory policies to cover states well is common to many provable methods (e.g., Du et al, 2019b, Misra et al., 2019).
- Loop over episodes missing – The only interaction with the environment happens in the line "Collect n triples ..." in Algorithm 1. To collect one triple, we execute a full episode, so the total number of episodes is nH.
- Objective for elliptical potential – This quadratic objective originated in the linear bandits literature and is quite standard in linear RL problems with known features, e.g., it appears in the analysis of Jin et al.
- The sampling oracle is used in Algorithm 2, to optimize Eq (2) in the learned model. This can be done with dynamic programming-type algorithms, such as Optimistic LSVI, which is quite straightforward and described in Appendix C, Lemma 5. We *do not* say that Algorithm 2 corresponds to optimistic LSVI.

Thanks for the references. We agree the count-based exploration work is relevant. Note that the other works mentioned do not consider representation learning in the context of exploration, which is our focus, and so they are less relevant. We will add more discussion to address some of the confusions above,

R3. Assumption 1 in block MDPs – You are right, the realizability assumptions in our work and block MDP results are not equivalent, but there is some subtlety here. We consider "model-based realizability" while a weaker notion may be simply that $\phi^\star \in \Phi$, with no assumption on $\mu^\star$. The realizability assumption for block MDP results is between these two: Du et al.'s and Misra et al.'s assumption is equivalent to realizability of $\phi^\star$ and *the support* of $\mu^\star$ (since the next state is also decodable). We do not see a natural analog to this intermediate assumption for the general low rank setting, but we agree that considering the weaker assumption (only $\phi^\star \in \Phi$) is a nice question for future work. Note also that when we discuss expressiveness, we are focusing only on the dynamics assumptions, and not on realizability requirements. We can expand on this in the final version.

We can add some discussions around practical issues, comparison to other works, and on limitations of low rank MDPs.

R5. Thanks for the pointers to bisimulation! We can discuss more in the final version. Briefly, our learned representations are related to a state abstraction notion, called Kinematic Inseparability (from Misra et al. 2019), which is finer than bisimulation, but remains meaningful in the reward-free setting. Note that in the absence of rewards, every MDP admits a trivial bisimulation that aggregates all states together. Indeed, it is not possible to learn a bisimulation with polynomial sample complexity while exploring in a sparse reward problem, as formally proved in Modi et al. 2020 (Proposition B.1). Thus our abstraction is less coarse, but learnable in the exploration setting.

Regarding reward-free exploration: You are right that system identification may be overkill for easy RL problems (e.g., dense rewards, local exploration suffices, etc.). However, we are interested in *provable* sample efficiency, which means we must consider *hard* RL problems. In this case, the distinction between reward-free and reward-sensitive learning is less significant, since even in the reward-sensitive setting the agent must explore the entire environment so it can certify that there is no "hidden rewards" anywhere.

You are also hinting at a separation between model-based and value/policy-based algorithms. This seems plausible, and we agree that developing provable model-free algorithms for low rank MDPs is an exciting direction for future work.