We thank the reviewers for their detailed comments and positive evaluation of our work. The questions primarily concerned (1) evaluation in an environment with a larger action space, (2) performance of ME-TRPO, (3) analysis of the gap between MBPO and the baselines, and (4) discussion of the tightness of the theoretical bound.

(R1) Code release. Code for reproducing our experiments is now available on GitHub. To preserve anonymity, we do not link directly to the repository.

(R1) Larger environment. We provide results on the Humanoid environment, requested by R1, in Figure 1. We will add these to the final version.

(R1) Intuitions for the theory. We will expand the discussion in Section 4 to present a better intuition for the practical implications of the theory. The theory suggests that: (1) short-horizon rollouts may be beneficial in some settings; (2) incorporating the model can allow for larger policy changes while still achieving monotonic improvement, but only if the model generalizes well to changes in the policy – the worst-case generalization does not achieve this, but we empirically find that real models on MuJoCo benchmark tasks generalize substantially better.

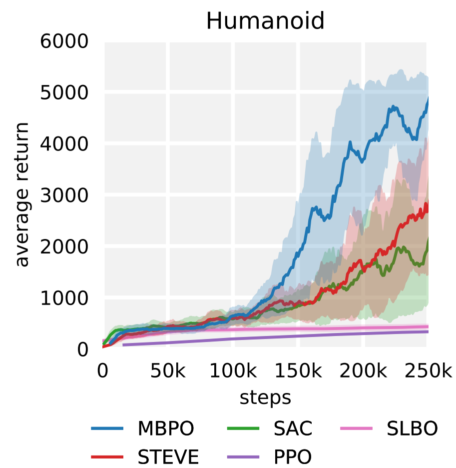

Figure 1: Results on Humanoid-v2. MBPO results are averaged over four seeds. The short rebuttal period did not allow for running all baselines to convergence, but we will add them to the final.

(R1, R2) Analysis of comparative performance. R1 and R2 asked about the sources of improvement of our method over the baselines. Here we elaborate on our choice of ablations and how they address this question.

1. The 500-length rollout ablation in the paper's Figure 3 is the suggested ME-SAC baseline, as it uses the ensemble to generate model rollouts with lengths on the order of the task horizon for consumption by SAC. We conclude from this result that truncated rollouts are a primary source of the performance difference between our method and **ME-TRPO**.
2. To better understand the difference between using model data directly for training and for improved target value estimates (as done in **MVE** and **STEVE**), we implemented the value expansion technique on top of SAC to control for the underlying model-free algorithm. This comparison is found in the paper's Figure 3.
3. We do not include a separate ablation for **PETS** because the comparison between MBPO and PETS is already well-controlled. The model ensembles are the same in both methods, so the difference in performance is attributable to the different ways the model is used: planning by sampling from a fixed prior in PETS and policy optimization in MBPO.

(R2) ME-TRPO baseline. R2 raised concerns about the relatively poor performance of ME-TRPO on the full-length HalfCheetah. The ME-TRPO paper evaluates on modified tasks, with horizons of 100 or 200, making their reported results not representative of the standard benchmarks. Our results use the authors' code and are representative of the actual performance of the method. An independent benchmarking of model-based RL algorithms, released after the NeurIPS deadline, reported the same results from ME-TRPO on the full-length environments [Wang et al., 2019].

(R2) Elaboration on branched rollouts. The branching rollouts use the marginal distribution from a previous policy as an initial state distribution for truncated model rollouts. In practice, this amounts to sampling a state $s \sim \mathcal{D}$ from the environment replay buffer, rolling out under the model for at most k steps using the current policy, and using these model predictions for policy optimization.

(R3) Tightness of the bound. Our bound is tightest in MDPs in which a single differing action or transition leads two trajectories to permanently diverge, as in the binary tree MDP in Figure 2. A crucial step in proving Theorem 1 is that if two agents select differing actions with ϵ probability, then their state marginals diverge by ϵt in total variation (Lemma B2). In Figure 2, the amount of divergence is exactly $1-(1-\epsilon)^t$, which is close to ϵt when ϵ is small. We are not aware of a way to create a tighter bound while still handling this pathological MDP. Similar proof techniques are used to analyze the TRPO and CPI algorithms.

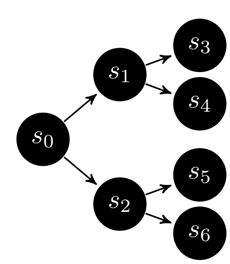

Figure 2: An example MDP where the bound is presented in Theorem 1 is nearly tight.

References

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019.