We thank the reviewers for their insightful comments and questions, which we answer below.

(R1) High-level explanation The high-level intuition behind MDP-GapE is that unlike a purely optimistic policy, we do not aim at minimizing regret while learning the best action to take. MDP-GapE indeed explores much more at depth 1 compared to a UCRL type algorithm. The UGapE rule used at depth 1 is crucial for quickly achieving the stopping condition that we propose: stop when one of the confidence intervals on the value at depth 1 is larger than and separated from the others. We will make sure to improve the high-level explanation of the algorithm in our revision. Note that using a greedy (optimistic) policy at depth 1 would require a different stopping rule, inspired perhaps by lil'UCB [Jamieson et al. 2013], and a completely different analysis. Hence we did not explore this path.

(R1,R2) Lower bounds The only lower bound on the sample complexity of MCTS planning that we are aware of it the $(1/\varepsilon)^{1/\log(1/\gamma)}$ worse-case bound of [Kearns et al. 02]. As for problem-dependent lower bounds, there exists some for $H = 1$, which corresponds to ε-best-arm identification in a multi-armed bandit. In that case, the lower bound of [Mannor and Tsitsiklis 04] indeed scales with the gaps in step 1. We will add a paragraph on lower bounds in our revision, in which we will leave the design of problem-dependent lower bounds for $H \geq 2$ as an important future work.

(R1) Experiments It is true that in experiments, we use tighter threshold functions than those prescribed by theory. However, we did not *tune* these functions to perform well on the studied problems. Their choice is rather inspired by our theoretical results (for their scaling in $n_h^t(s,a)$), un-doing a few union bounds that were found to be conservative. Albeit questionable, using threshold functions "slightly smaller than theory" mimics what is sometimes done in the bandit literature, and is also quite common for UCT users. Regarding Sparse Sampling (SS), we will remove n_{SS} from Table 4 and propose instead the following discussion. The sample complexity of SS with parameter C (number of calls to the generative model in each node) is $(K^{H+1} - K)/(K-1)$ for $C = 1$ and of order $\sum_{h=0}^{H-1}[(KC) \times (K(\min(B,C)))^h]$ for larger values of C. Thus, beyond very small C, the runtime of SS is prohibitively too large to try the algorithm in our setting (larger than 10^H). For $C = 1$, the sample complexity of SS is 2.0×10^4, 4.9×10^5 and 1.2×10^7 in the 3 experiments in Table 4, which is larger than the maximal sample complexity observed for MDP-GapE. Still, SS has a larger simple regret (e.g. we observed $\max_n r_n = 0.43$ for $\varepsilon = 0.5, H = 8$).

(R1) Size of search tree We believe that we can indeed replace $(BK)^{H-1}$ with SA in the bound. Yet planning algorithms are usually intended for the case $(BK)^{H-1} \ll SA$, this is why we focus on the scaling in $(BK)^{H-1}$. Although this does not hold in our experiments, none of the algorithms that we implemented exploits the knowledge of S (only that of the support B) and we propose to perform experiments with much larger state spaces in the revision.

(R1) Gaps of all state-action pairs For deterministic transitions, Theorem 2 actually provides an upper bound that involves the gaps in all (reachable) state-action pairs. Yet, beyond this case, we did not manage to get a tight bound of this flavor. We remark that the bound of Simchowitz and Jamieson (which is on the regret, and is therefore hard to compare to our sample complexity bound), features a sum over all inverse gaps but also includes a term that is inversely proportional to the minimum gap among *all* state-action pairs, which can be arbitrarily smaller than the depth 1 gaps.

(R2) Novelties in MDP-GapE The analysis of MDP-GapE is much more sophisticated than that of UGapE-MCTS. The crucial difference is that for stochastic transitions we need to construct confidence intervals on the expected values, whose diameter is harder to relate to the number of visits in the tree and required the introduction of *pseudo-counts*. The proof of Theorem 1 relies on a completely new proof technique that sums the local confidence bounds across episodes.

(R2) KL confidence sets One could replace the KL confidence sets with the L1 confidence sets typically used in UCRL, and obtain the same guarantees (our current analysis uses Pinsker's inequality and does not fully exploit the KL confidence sets). Yet as the KL confidence sets are tighter, their practical use leads to earlier stopping.

(R2) Dependence on ε Improving the dependence on ε is mentioned as a possible future work in our conclusion.

(R2) Benign/hard planning problems Our best intuition is that "hiding" the reward very deep in the tree will result in hard problems, while providing intermediate reward along the optimal path will cause the algorithm to stop more quickly. In our revision, we will try to explicit the scaling of our bounds in these two cases.

(R2) UGapE in every step? Our analysis crucially depends on being optimistic at depth > 1, and does not permit to analyze an algorithm using UGapE in every step. Our intuition is that this approach would not be more efficient for finding the best action at depth 1, although it might be beneficial for finding a good policy for the next steps as well.

(R3) Depth-dependent dynamics A standard MDP would indeed have the same dynamics at every depth., i.e. the transition probabilities and rewards are depth-independent. However, since most planning algorithms explore the search tree up to a fixed horizon, they can also handle depth-dependent transition probabilities and rewards without incurring additional computational complexity, so we include them in order to make the algorithm as general as possible.

(R3) Trajectory-based planning As explained in the introduction, we do not assume that we have access to an arbitrary generative model, but only to a *forward model* that can sample actions *in the current state*, starting from the root state.