We thank the reviewers for their valuable feedback and suggestions for improving the paper. We are glad the reviewers believe this paper presents a useful contribution on developing more sample efficient online planning algorithms. Our responses to the reviewers' comments are below.

To R1. The specific choice of λ_t is made to guarantee the convergence rate in the tree case. It guarantees that at each node of the tree MENTS performs sufficient exploration for all actions so that the softmax value can still be efficiently estimated even under the drift condition (Theorem 4). In equation (7), the decay rate is decided by the total number of simulations of s, which is $N(s) = \sum_a N(s,a)$. Table 1 presents the final performance of MENTS and UCT when using 500 simulations to generate a move. All results are averaged over 10 environment restarts. We will present the standard errors and exact values of all hyperparameters in the final version of the paper. In Figure 3 ($k = 8, d = 5$), the decrease of error of UCT stops because it does not propose any reasonable strategy given the simulation budgets.

To R2. Thanks for your suggestions. We did not compare MENTS with other variants of UCT, such as UCT with TD(λ) backups, since it has not been shown that with those variants the convergence property of vanilla UCT still holds. We will add this comparison in the updated paper.

To R3. We agree that entropy regularization introduces a different learning objective compared with existing methods. However, in both theoretical analysis (Theorem 5) and experiments (Figure 3), we directly compare MENTS with UCT in terms of the efficiency of finding the best action at the root. The reason we propose to minimize the MSE of the estimated softmax value in the *stochastic softmax bandit* setting is because it is difficult to directly apply the regret minimization objective in this setting. In the proposed lower bound (Theorem 1), we show that minimizing the MSE is equal to finding the optimal softmax policy. We would also like to point out that MENTS achieves a better convergence rate under weaker assumptions. Note that UCT assumes the value estimation convergence property of the internal nodes under drift condition (see Section 2.4 of [2]), while MENTS only needs the sub-Gaussianness on the leaves. This is because of the benefit of using the entropy regularization in the tree case, that the softmax value of the internal node is guaranteed to satisfy the sub-Gaussianness (Theorem 4). Furthermore, our proposed method and analysis can be directly and completely applied to the setting with stochastic transitions and rewards using standard techniques, such as those in [1]. For example, we can use empirical estimates of rewards and transitions in the search tree. The convergence rate of the empirical estimations follow from Hoeffding's inequality and therefore will not affect our results. In Figure 3 ($k = 8, d = 5$), the decrease in error of UCT stops because it does not propose any reasonable strategy given the simulation budgets.

We thank all reviewers for pointing out the need to strengthen the discussion, particularly in the experimental section. We will add more discussion about the effect of entropy regularization and analysis of the online planning experiment results in the revised version of the paper. We will also conduct more experiments, including sensitivity test of parameter τ and comparison with UCT variants.

References

[1] Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine learning*, 49(2-3):193–208, 2002.

[2] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.