We thank all reviewers for spending their valuable time reviewing our paper. We appreciate their insightful comments, and will incorporate them in the revision. We now answer some specific question in detail.

Relaxation of the notion of equivalent subMDPs: The definition of "equivalent subMDPs" (Definition 2) requires that (1) there is a *bijection* between state spaces and (2) through which the subMDPs have the *same* transition/reward models at internal states. As discussed in the paper, (2) can be relaxed to *similar* transition/reward models. The bijection assumption (1) is essential for the planning algorithm (Algorithm 2) to work in its current form, and hence, for the computational efficiency results. For the statistical efficiency results, this assumption could be relaxed, e.g. if a clustering of the state space into clusters with similar rewards and transition distributions exists. We agree that finding a different approach for Algorithm 2, not relying on the bijection, would be an interesting direction for future work.

Automatically discovering subproblems: We fully agree that how to automatically discover the subMDPs is an important open problem. However, it is beyond the scope of this paper and we aim to address it in future work. The results we provide are a necessary first step before tackling discovery.

Connections to prior work: We will add a more explicit discussion about the comparison to Mann et al. (2015) to the paper. To summarize, they discuss and analyze two algorithms: *Option-Fitted Value Iteration (OFVI)* and *Landmark-Approximate Value iteration (LAVI)*. The OFVI analysis relies on the discounted-average concentrability of the future state distributions in the semi-MDP defined by options, so it is a very different-flavor result. LAVI relies on options that go to designated landmark states, and which are computed by solving a deterministic relaxation of the semi-MDP in a neighborhood of landmarks. In our terminology, such options have a single exit state, and LAVI then solves the problem that jumps between landmarks. There is no repeating structure in this approach; in fact, each option only applies in a small neighborhood of state space around a landmark. Our result could be applied to the LAVI setup directly, but it would be hard to compare to their bound directly due to the very different quantities involved.

Theorem 1 in this paper is partially motivated by Osband et al. (2013); however, we consider a very different setting and our result is technically more complex. Specifically, (1) Theorem 1 considers hierarchical structure while Osband et al. (2013) does not; (2) Theorem 1 allows sub-optimal planning while Osband et al. (2013) does not; (3) Theorem 1 allows a random time horizon τ while Osband et al. (2013) is restricted to a fixed time horizon. These major differences make several key steps in the analysis both different and more challenging (e.g. the step to bound the probability that the sampled MDP $\mathcal{M}^t$ is not in the confidence set $\mathbb{M}_t$ at each episode t) . We will further highlight these differences in the revision.

About the "Acyclicity" assumption: We chose to make the "acyclicity" assumption (Assumption 1) mainly to simplify the exposition of our computational efficiency results. This assumption ensures that value iteration (VI) will terminate in a finite number of steps, a fact we use in Proposition 1. This assumption can be relaxed: by using the *weighted sup-norm contraction* under VI in *stochastic shortest path problems* (see Section 3.3 of Bertsekas (2015)), we can obtain a similar computational efficiency result without this assumption, but it is mathematically more complicated and harder for readers to digest, and so we opted for the cleaner version. We will add an observation to the paper noting that the "acyclicity" assumption is not strictly necessary, thanks for pointing that out!

To Reviewer #1: Assuming that (1) the agent knows $\mathcal{S}$ and (2) a fixed initial state is mainly to simplify the exposition. We agree this does not represent the most encompassing formulation of RL, however, we believe that these simplifying assumptions are without loss of generality and the insights obtained from this paper apply more broadly. We will further clarify this point in the revision.

To Reviewer #2: We will clarify in the revision that this paper focuses on model-based RL algorithms.

It is not completely clear what is meant by "the case where the subproblems are not precisely defined". If you mean the case where the subproblems need to be discovered, or where the current notion of equivalent subMDPs needs to be relaxed, please see the discussion above.

To Reviewer #3: The exit profile of a subMDP can be any vector that assigns a real number to each exit state. Different exit profiles will induce different policies in the subMDP. If an exit profile is close to (or equal to) the optimal state values at the exit states, then it will induce a near-optimal (or optimal) policy. The notion of exit profiles in this paper is different from *"exits"* defined by Hengst (ICML 2002). Specifically, the exit profiles in this paper are vectors (see Definition 3), while the "exits" in Hengst (ICML 2002) are state-action pairs (see Definition 1 in their paper).

We will indicate in the main paper that the proof for Theorem 1 is provided in Appendix A of the supplementary material. Thanks for the reminder!

To Reviewer #4: We believe that we have addressed all of your concerns above.

To all reviewers: Thanks again for reviewing our paper and reading this author response!