We thank the reviewers for their comments.

We feel that the contributions of this work and the motivation for it are well understood by the reviewers: We investigate the convergence properties of an online planning algorithm, given access to a lookahead policy oracle. Indeed, we study the performance of such algorithms, both in the exact form and in several approximate settings, and compare them to their approximate dynamic programming (ADP) counterparts. To the best of our knowledge, there is no theoretical analysis regarding guarantees of lookahead policies in online planning algorithms. Furthermore, we believe the generality of the presented techniques may be found useful in the analysis of and development of online planning algorithms.

There is no dispute on the importance of empirical work. However, we believe that the theoretical results provided in this work are important on their own. Our analysis spans not only the exact, but also three approximate settings, and provide detailed comparison to the performance of ADP. Furthermore, unlike the scarcity of theoretical results in online planning with lookahead policies, there are many works that study the empirical performance of different online planning algorithms that are based on lookahead policy. Yet, the existing empirical works are heuristic; this stresses the importance of theoretical results on online planning algorithms with lookahead policies. We hope the rigorous approach pursued in our work will stimulate further theoretical research on the interplay between the lookahead horizon and the performance of the online planning algorithm. There are, as always, interesting and important theoretical questions to be solved.

As mentioned in response to Reviewer 3 and also clearly highlighted by Reviewer 4, a thorough empirical comparison of RTDP, MCTS and ADP is very important and useful for the community, and may shed light on several unanswered questions about the MCTS algorithm. However, due to the extent of our theoretical results and the length of current paper, we feel a thorough empirical study of these algorithms is outside the scope of this work.

R1. For the comment on the needed empirical work, please see the above paragraphs that have been written for all the reviewers. The second question that you raised is definitely of interest. In fact, in our opinion, answering such a question deserves a work on its own, as there are probably several answers for it, such as which assumptions should be made? What are the relevant structural properties? Can we choose it in an online manner? In a somewhat different context, this question is equivalent to asking how to choose a hyperparameter of an algorithm? Indeed, it is an important question, however it is outside the scope of the current work.

In the ICML-2020 paper “Multi-step Greedy Reinforcement Learning Algorithms" by Tomar, Efroni, and Ghavamzadeh, the authors consider the framework of *model-free RL*, whereas we focus on online planning. For this reason, the works are not very much related (we shall clarify this in the text, thanks!). Furthermore, this empirical paper is a follow up to two theoretical papers (’Beyond the one step greedy approach in RL’ ICML 18’ and ’Multiple-Step Greedy Policies in Approximate and Online RL’ NeurIPS 19’). In our opinion, this clearly shows the importance of theoretical results that can guide the design of practical algorithms.

R2. We would like to thank the reviewer for the supportive review. It definitely encourages us to keep investigating online planning algorithms and expand our current understanding. We will revise the sentence on line 207.

Previous results which analyzed the performance of the UCT algorithm showed its worst-case sample complexity depends exponentially (or even worst) on to the horizon of the problem. Furthermore, the sample complexity to find an ϵ optimal action using the UCT usually scales as $O(1/\epsilon^2)$. Unlike the UCT, the sample complexity of RTDP depends on the size of the state space (or the abstract state space as we showed in this work) and does not depend exponentially on the horizon, only polynomially. The dependence on ϵ scales as $O(1/\epsilon)$. These results definitely implies on a possible superiority of RTDP over MCTS from a theoretical perspective. We will emphasize it in the discussion part.

R3. We agree with the reviewer about the need for a thorough empirical work that compares RTDP, MCTS and ADP. Such work is very important to the community in our opinion and might resolve the hardness of tuning the hyper-parameters of the MCTS algorithm. Due to the extent of the theoretical results supplied in the paper, that comprehensively study h-RTDP in its exact form and in three approximate settings, we feel a thorough empirical study of these algorithms is outside the scope of the current paper.

We now address the additional feedback the reviewer gave. -) It is the same motivation as in RTDP. We will discuss this in the final version of the paper. -) We tried to stress this as much as we can. We will add it to the abstract to avoid confusions. -) Due to lack of space, we will add a table to the appendix to fully specify this computational complexity. -) Thanks! we will fix this.

R4. We would like to thank the reviewer for the positive feedback. We also thank you for the minor comments which helps us to improve the work. We will fix them all in the final version of the paper.