We thank the reviewers for their feedback. We took great care to precisely state our research question and situate it amongst the prior work, and appreciate that the reviewers found this effort useful. We will definitely use the reviewers' feedback to improve the clarity further. We will now try to address some of the specific concerns, and add these clarifications to the paper.

Core states: We agree with all the reviewers that the assumption of core states is quite strong, and we are highly interested in generalizing and weakening this condition. In particular, we would like to: (1) "localize" the condition, requiring only core states that cover the successor states of the current planning state, rather than the entire state space; (2) weaken the condition by considering "approximate" core states, whose convex hull covers "most" of the probable next states. Then one would expect a similar approximation error, with additional dependence on the probability of visiting states whose features lie outside the core set convex hull.

We hope that these questions will be the subject of future work. Nevertheless, we feel that the contribution is useful in its present form. In particular, as we discuss in the Related Work section, it is still not known how to avoid exponential blow-up of computation in the planning horizon without such strong assumptions. Furthermore, as we also discuss in the paper, unlike other assumptions that have been made in prior work (e.g. linear transition models, low Bellman rank, etc.) ours is a purely geometric condition on the feature representation, *not* a restriction on the transition model of the MDP. We foresee two basic approaches to satisfying this assumption in practice: (1) adapt exploration methods to discover core states; (2) in practice, large scale reinforcement learning usually requires representation learning — the learned representations could be made to satisfy the core state assumption *by design*. This option is not available when restrictions are placed on the MDP itself, rather than the feature representation. Both of these approaches are quite intriguing, require careful design and analysis, and will be the subject of future work. We believe that our results in this paper lay the foundations for more exhaustive research in this direction, and will discuss these possibilities in the paper.

Novelty of the optimization approach (Reviewer 1): The linear programming approach to control in RL has, of course, a long history. It is also quite standard in the optimization literature to use saddle point methods to solve linear programs. However, it remains a tricky problem to actually *use* the solutions of an approximate saddle-point solver for planning and control in RL. We hint at some of the challenges in the main paper, and will definitely add more discussion about this along with our techniques. This algorithmic issue is important and not addressed by Lakshminarayanan et al.

The solutions produced by approximate saddle point algorithms are not just suboptimal — they violate the constraints of the original linear program. This is especially problematic with function approximation, where the interpretation of the dual LP solution as an occupancy measure is no longer meaningful. Bas-Serrano and Neu [arxiv:1909.10904] have recently pointed out that, without an additional (strong) "coherence" assumption, the policies extracted from approximate saddle point solutions can be arbitrarily suboptimal. They also argue that other prior works suffer from this issue. The problem is also difficult because saddle solvers needed bounded constraint sets. We will add this discussion to the paper.

One of our contributions in this paper (which we will expand upon) is to show that the core state assumption not only bounds the approximation error of the LP, but also solves the algorithmic problem of using approximate saddle point solutions to identify good actions. We hope that our techniques will be useful for future algorithm designers.

Achievable error (Reviewer 1): It is an interesting question whether polynomial-time planning under function approximation is possible without the extra H^2 factor in the error bound. We believe that Theorem 1 of Tsitsiklis and Van Roy [doi:10.1007/BF00114724] shows that the H^2 factor is unavoidable, and will expand on this in the paper.

In our work the expected $\sqrt{d}$ factor is replaced by $\sqrt{m}$, where m is the number of core states. Importantly, if the feature matrix is full-rank then $m \geq d$, so we are not avoiding the dimension-dependence. However, we believe that improving the optimization method can improve the dependence on $\sqrt{m}$ to $\sqrt{d}$ and hope to achieve this in the future. The number of core states m is related to the geometry of the feature-set, regardless of its cardinality; m can be finite for infinite state spaces (**Reviewer 4**).

Uniform approximation (Reviewer 1): $\varepsilon_{\text{approx}}$ is an ∞-norm error bound and thus measures approximation error uniformly over all states. Prior works have considered weighted norms (e.g. Lakshminarayanan et al.) and we believe those results can be transferred to our work, at the price of additional assumptions (the Lyapunov stability condition). We opted not to make this extension to keep the paper focused, but will note the possibility in the discussion.

Experimental results (Reviewer 2): Although the focus of this work was theoretical, we will endeavour to provide some representative numerical results, possibly in the Supplementary Material. We will also refer readers to the simulations of Lakshminarayanan et al., which address the quality of the relaxed ALP solutions, if not the efficiency of the optimization algorithms presented in this paper.