We thank all reviewers for their comments and insightful reviews. We will address the concerns as follows. Typos will be corrected in the final version.

Reviewer 1 Thank you for your comments. We hope following explanations can answer your questions.

Related work Model-based algorithms with function approximation have been indeed studied in prior works. What we want to claim is that we are the first to study the *plug-in* solver approach with function approximation. We believe this is a nontrivial contribution to the RL community.

Novelty Although we have applied the anchor-state model from [Yang and Wang] and our technique has certain similarity to [Agarwal et al., 2019], we stress that neither of the two papers can be directly applied in our setting for a plug-in approach of model-based RL with function approximation. In fact, our technique is highly non-trivial and inspires several new understandings in terms of model-based RL with features. Moreover, our results can also be applied to the setting without anchor-state by restricting the plug-in solver to value iteration. In addition, turn-based stochastic game, a multi-agent extension of MDP, is analyzed by using the optimal response policy in the auxiliary MDP to approximate empirical optimal response policy.

More specifically, the model-free algorithm in [Yang and Wang] relies on monotonicity and variance reduction, which cannot be applied to model-based setting with a plug-in solver. The absorbing MDP in [Agarwal et al., 2019] cannot be applied to linear MDP as it destroys the linear dependence, which inspired us to invent a *new auxiliary MDP*. This auxiliary MDP fixes the transition of an anchor state-action and changes the entire transition kernel via linear dependence (Definition 2). The auxiliary MDP technique we propose can also be applied to tabular MDP, which covers absorbing MDP. Another critical disadvantage of the absorbing MDP in [Agarwal et al., 2019] is that it can only recover state values, while our technique can recover the entire state-action values (Lemma 6).

Model-based planning First, we want to emphasize that arbitrary planning algorithm is suitable in our algorithm, so the computational complexity of the planning algorithm is not our main focus. It is known that the exactly optimal policy in MDP and TBSG can be obtained in strong polynomial time $\widetilde{O}(\mathrm{poly}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-1}))$ by policy iteration/strategy iteration [Ye, 2011, Hansen et al., 2013]. Approximate dynamic programming methods like LSVI/FQI can utilize the features to achieve $\widetilde{O}(\mathrm{poly}(K(1-\gamma)^{-1}\epsilon^{-1}))$ computational complexity . In addition, one can use the learning algorithm 'Optimal Phased Parametric Q-Learning' in [Yang and Wang] to do planning, which has computational complexity of $\widetilde{O}(K(1-\gamma)^{-3}\epsilon^{-2})$ (i.e. same to the sample complexity result in our work and immediately achieves minimax computational complexity).

Anchor states and pseudo-MDP The analysis in [Zanette et al., 2019] requires $\|\lambda\|_1 \leq 1+\frac{1}{H}$ so that the error will not amplify exponentially and fails when $\|\lambda\|_1$ is larger than this threshold. Our result shows that empirical value iteration (EVI) is a sample efficient algorithm for bounded $\|\lambda\|_1$ with sample complexity $\widetilde{O}(K\max_{s,a}\|\lambda(s,a)\|_1^2\mathrm{poly}((1-\gamma))\epsilon^{-2})$, which demonstrate that EVI is sample efficient for $\|\lambda\|_1 > 1+\frac{1}{H}$. Note that [Zanette et al., 2019] assumes linear representation of Q^*, which is different from our assumption. To our best knowledge, the minimax sample complexity in linear MDP without anchor state assumption is still unknown.

Correctness We apologize for the typos in the appendix (proof of Lemma 6). The correct and more detailed version is $\widetilde{Q}^{\pi}_{u^\pi} = (I-\gamma\widetilde{P}^{\pi})^{-1}(r+\Phi^{s,a}u^\pi) = (I-\gamma\widetilde{P}^{\pi})^{-1}(r+\Phi^{s,a}\gamma(\widehat{P}(s,a)-P(s,a))\widehat{V}^{\pi}) = (I-\gamma\widetilde{P}^{\pi})^{-1}((I-\gamma\widehat{P}^{\pi})\widehat{Q}^{\pi}+\gamma\Phi(\widehat{P}_{\mathcal{K}}-\widetilde{P}_{\mathcal{K}})\widehat{V}^{\pi}) = (I-\gamma\widetilde{P}^{\pi})^{-1}((I-\gamma\widehat{P}^{\pi})\widehat{Q}^{\pi}+\gamma(\widehat{P}-\widetilde{P})\widehat{V}^{\pi}) = (I-\gamma\widetilde{P}^{\pi})^{-1}(I-\gamma\widetilde{P}^{\pi})\widehat{Q}^{\pi} = \widehat{Q}^{\pi}$. Note that $\widehat{P}_{\mathcal{K}}-\widetilde{P}_{\mathcal{K}}$ has all zero rows except row (s,a) by the definition of auxiliary MDP.

Reviewer 2 & Reviewer 3 & Reviewer 4 & Reviewer 5 Thank you for your appreciation. We will fix typos and clarify some of the confusions in the next version. Below, we address the common concern about the assumption in this paper.

Strong Assumptions Our assumptions on linear MDP are widely used in literature as discussed in Section 4. Anchor state assumption indeed appears a strong assumption, however this is rather general: this assumption essentially assumes all the features vectors lie in a convex hull, which is without loss of generality. The number of vertices of the convex hull is the number of anchor-state-action pairs. The number of vertices can be small in many cases (see e.g., [Blum et al., 2019] "Sparse approximation via generating point sets" and reference therein). Moreover, our results also apply to the approximate model setting (Theorem 2).

Furthermore, we show that the anchor state assumption is essential to obtain an eligible empirical model by showing an hard instance (Proposition 3). Our work also gives a minimax sample complexity algorithm for $\|\lambda\|_1 = 1$ (anchor state assumption) and efficient algorithm for $\|\lambda\|_1 > 1$ but bounded.

Generative model is a meaningful oracle which receives much attention (see Section 2 for a detailed review) as it separates the subtle exploration questions from learning. In many realistic settings we also have a simulator to generate samples from arbitrary state-action pair. For instance, learning in physical simulators allows this kind of sampling. Moreover, games like Go and chess are turn-based stochastic game that can be viewed as generative model.

Table of previous results (R4) Due to limited space, we cannot put the table in this rebuttal. A table of previous results will be added in the final submission. We will also move some discussion of estimating the transition kernel P in the appendix to the main paper.