We sincerely thank all the reviewers, and feel really honored to receive such positive and constructive comments. We are in the process of incorporating many of the changes into the final version.

Reviewer 1 **1)** We will improve the broader impact section by emphasizing the implications of our theoretical results on applications. **2)** We will mention total variation distance in the appendix, and correct the typo on "Corollary 3.3". Thanks for the careful reading.

Reviewer 2 **1)** Note that the smooth planning oracle is not needed throughout the paper, and is thus not the "primary concept" in our paper. It is only used in Sec. 3.2 for finding ϵ-NE policies with near-optimal sample complexity. We have justified this in lines 209-231, and will add some demonstration on simple games in the final version. **2)** Yes, by saying "rarely", R-MAX is definitely one of the very few ones. We have discussed R-MAX in lines 82-83. We will change the wording in the final version. **3)** By saying "especially model-free ones..." this sentence, we simply meant that the convergence techniques for single-agent Q-learning/model-free methods cannot be applied directly in MARL. The works on Q-learning in games you mentioned exactly conquered this issue, with non-trivial efforts. We will add more clarifications on this. **4)** We will address all the grammatical comments/typos in the final version. Thanks a lot for the careful examining.

Reviewer 3 **1)** We will add more comparisons with [4,17] on the lower bound proof, in the final version. **2)** We thought the "generative model" setting is a standard one. Thanks for pointing this, and we will add the clarification. We believe an adaptive policy guided sampling may indeed decrease the sample complexity of "model-based" methods, if more structure about the "model" is known beforehand, e.g., which state transition is more significant. Otherwise, for general models without special structure, estimating the whole transition model (approximately) seems inevitable, and our matching upper-bound seems to be hardly improvable.

Reviewer 4 On the smooth planning oracle: First, technically, our "leave-one-out" proof technique requires some "smoothness" of the change of either the "value-function" or the "policy" (for proving ϵ-approximate NE value and ϵ-NE policy, respectively), when one state is made "absorbing". When showing the near-optimal sample complexity to find ϵ-approximate NE value, the change in value function is indeed smooth, so no "smooth planning oracle" is needed; when showing the near-optimal sample complexity to find ϵ-NE policy, the small change in the Markov game cannot guarantee the change of the "best-response" policy of the opponent to be small, too. In fact, a small change in the value may correspond to a drastic change in the best-response policy. Thus, we introduce the smooth planning oracle to address this. Second, computationally, as we have discussed in lines 209-231, such an oracle can be readily satisfied in practice, especially in the regularized setting. Regularized setting makes solving the matrix game a strongly-convex-strongly-concave problem, and thus improves the computational efficiency, over the un-regularized ones. Whether it is possible to achieve "near-optimal sample complexity" for achieving "ϵ-NE policy" without the smooth oracle may require totally different proof techniques, and is left as our future work.