We thank all the reviewers for their valuable feedback and appreciating our contributions. First we would like to emphasize the technical novelty of our upper bound and lower bound as Reviewer #1, Reviewer # 3 and Reviewer #4 commented on the technical novelty of our theoretical results.

Technical novelty of the upper bound. In the exploration phase, Jin et al. [2020] set reward to be 1 for significant states and 0 for other states. Note their technique cannot be used in Linear MDPs because there are possibly infinitely many states and thus one needs to take the structure of linear functions into account. In this paper, we use UCB bonus *as the reward signal* in the exploration phase. To our knowledge, this idea is new in the literature. We also would like to thank Reviewer # 4 for a detailed description of our main algorithmic ideas.

Technical novelty of the lower bound. We have discussed the differences between our lower bound and that in [Du et al. 2020] in Line 276 - 279. We acknowledge that in our hard instance, we use a similar feature extractor as that in [Du et al. 2020]. However, all other aspects of the hard instance construction are significantly different from that in [Du et al. 2020]. For example, for the hard instance in [Du et al. 2020], only a single state-action pair has non-zero reward value, which is not case in our hard instance. Note that such distinction is crucial, since in our hard instance the optimal Q-function is *exactly* linear, whereas the the optimal Q-function is only *approximately* linear in the hard instance in [Du et al. 2020]. Moreover, we focus on the reward-free setting while Du et al. [2020] focused on the standard RL setting.

Below we address specific concerns from each reviewer.

—— To Reviewer #1 ——

Lack of rigor. We have introduced necessary background on MDP in Section 2.1, including the state space, the action space, the transition operator, the reward distribution, the Q-function, etc. We have also provided necessary definitions related to linear function approximation in Section 2.2. Our descriptions mostly follow existing works. We will expand this part to make the paper clearer.

Q^* **is linear on the suboptimal action.** In our construction, when defining the reward functions, we first define the optimal Q-function (Q^*) as a specific linear function (see Line 292), and then define the reward values according to the Bellman equations (see Line 296). Therefore, the optimal Q function must be linear for both optimal actions and suboptimal actions in our hard instances.

Relation to prior work. We will discuss the suggested paper in the next version. Thanks for the suggestion.

—— To Reviewer #2 ——

Extension to more general settings. Even in the standard RL setting, going beyond linear MDPs is hard. See the open problems in [Du et al. 2020]. Therefore, we believe it is highly non-trivial to obtain more general results.

—— To Reviewer #3 ——

More emphasize on the lower bound. Thanks for the suggestion. We will emphasize more on the lower bound and the implied conceptual messages in the final version.

Why do you need optimism in the planning phase. Optimism in the planning phase is used when we prove Lemma 3.3. It also guarantees the correctness of the first inequality in Line 247-248.

—— To Reviewer #4 ——

We would like to thank the reviewer for the detailed description of our key ideas in our algorithm. The understanding is correct.

Experiments. Thanks for the suggestion. We will consider adding empirical results in the next version.

Related work. Thanks for the references. We will add more discussion in the next version.

Agent just gets samples from the reward function. If we only have samples, we can change Line 6 in Algorithm 2 to $w_h \leftarrow (\Lambda_h)^{-1} \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)(V_{h+1}(s_{h+1}^\tau) + r_h^\tau(s_h^\tau, a_h^\tau))$ where $r_h^\tau(s_h^\tau, a_h^\tau)$ is the *sampled* reward value, and remove $r_h(\cdot, \cdot)$ from Line 7. Our theoretical results still hold after this modification, and we will add a discussion on this.

Linearity approximately holds. This is an interesting question and we will list it as a future direction.

Line 182-183. This is correct.

The effect of increasing/decreasing c_β. c_β needs to be larger than a universal constant in order to guarantee optimism. Once c_β is larger than that constant, the sample complexity decreases as c_β decreases.