We thank all the reviewers for their valuable feedback and appreciating our contributions. We first address some common concerns.

The proof applies only to deterministic systems / Deterministic systems seem highly restrictive. Despite deterministic systems seem restrictive in theory, in practice, lots of RL problems are indeed deterministic. Moreover, all known algorithms that work under the assumption that the optimal Q-function is linear require deterministic or near-deterministic systems [Wen and Van Roy, 2013, Du et al., 2019].

The proof depends on the assumption on the gap optimality / The model has to be very correct. In this paper, we show that unless the gap $\rho = \Omega(\sqrt{\dim_E}\delta)$ where δ is the approximation error, any algorithm requires exponential number of samples even just to find a near-optimal policy (see Proposition 1.2). Therefore, such an assumption is necessary for any algorithm with polynomial sample complexity.

Please find our response to each individual reviewer below.

—— To Reviewer #1 ——

The algorithm for the general case requires an oracle. When the number of actions is finite (as in Atari games), the agent can possibly enumerate all actions and find $f_1, f_2 \in \mathcal{F}$ separately for each action by running continuous optimization algorithms that can handle constraints (e.g. projected gradient ascent). When the action space is continuous (as in control tasks), the agent could directly optimize a, f_1, f_2 by running continuous optimization algorithms (as done in practice). Moreover, we would like to note that our paper is concerned with the statistical efficiency, and the oracle does not require any new sample (it solves an optimization problem based on existing samples).

Compared to the "Know-What-It-Knows" oracle, our uncertainty oracle just requires solving an optimization problem, while it is even unclear whether the "Know-What-It-Knows" oracle can be implemented statistically efficiently for general function classes. We will make the comparison clearer in the next version.

The range of the return is assumed to lie in [0,1]. This is a standard regularity assumption in RL theory, and is required to make sure that the empirical mean of the reward values concentrates around their expectation by taking enough samples. Such assumption is required for the algorithm in the supplementary material (Section D). In general, if the summation of the reward values is in $[0, C]$, then the sample complexity of the algorithm in Section D will be increased by a factor of C^2. Note that the required assumption that $\rho = \Omega(\sqrt{\dim_E}\delta)$ keeps unchanged even if one changes the range of the reward values. E.g., if one scales all reward values by a factor of C, then the ratio between ρ and δ remains unchanged.

—— To Reviewer #2 ——

We would like to thank the reviewer for the positive feedbacks.

—— To Reviewer #3 ——

Cannot operate in the scenario where there exists more than one optimal policy. We disagree that our algorithm does not work in the scenario where there exists more than one optimal policy. Consider the case that for some state s, there are three actions a_1, a_2 and a_3. If $Q^*(s, a_1) = Q^*(s, a_2) = 1$ and $Q^*(s, a_3) = 0$, then by Definition 3.1, the gap would be 1 and our algorithm still works. However in this case, it is clear that there could be more than one optimal policy.