We thank all reviewers for very helpful comments. This letter addresses several major questions raised by the reviewers.

A common question: numerical evaluation. We will add a series of numerical experiments to demonstrate the minimax optimality of the model-based approach studied herein. Fig. 1 depicts some numerical plots of this kind. Here, we adopt a (least favorable) example designed in the minimax bound in Azar et al. [2] (and also Wainwright [37]), which primarily focuses on the effect of the discount complexity $\frac{1}{1-\gamma}$ on the sample complexity. More specifically, consider the MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S} = \{0, 1\}$, $\mathcal{A}_0 = \{0\}, \mathcal{A}_1 = \{0, 1\}$, $P(0|0,0) = 1, P(1|1,0) = p, P(1|1,1) = q$ and $r(0,0) = 0, r(1,0) = r(1,1) = 1$. According to Azar et al., the quantities p and q are taken to be $p = \gamma + \frac{2\gamma(1-\gamma)^2\varepsilon}{(1+\gamma)^2}$ and $q = \gamma - \frac{2\gamma(1-\gamma)^2\varepsilon}{(1+\gamma)^2}$. As illustrated in Fig. 1, the numerical sample complexity per state-action pair N scales on the order of $\frac{1}{(1-\gamma)^3\varepsilon^2}$ for varying choices of accuracy level ε, which is consistent with our theory.

Specific questions by Reviewer 1: 1. *Necessesity of reward perturbation.* Indeed, reward perturbation is introduced merely to facilitate analysis. Given that we recommend extremely small perturbations, the difference between the perturbed and original MDPs is almost unnoticeable in the experiments. While our current analysis does not work without reward perturbation, it would be important to see whether this can be removed with more refined analysis.

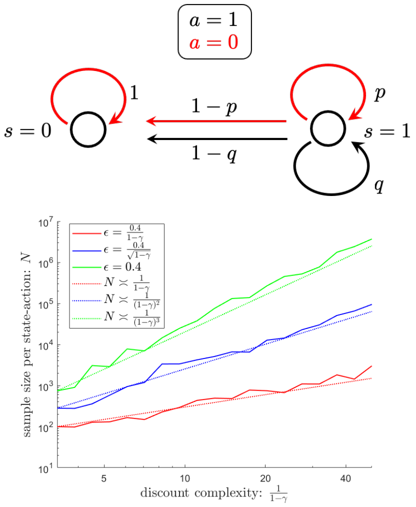

Figure 1: Numerical experiments.

Specific questions by Reviewer 2: 1. *Empirical illustration of minimax optimality.* We have conducted some illustrative numerical experiments to address this comment; see the response above for “Numerical evaluation”.

2. *Choice of α.* Indeed, our theory and analysis hold for any constant $\alpha > 1$. The current paper picks the specific choice $\alpha = 5$ only to convey that the perturbation $\frac{(1-\gamma)\varepsilon}{|\mathcal{S}|^\alpha |\mathcal{A}|^\alpha}$ can be very small; we shall make it more clear in the final version.

3. *Why the sample size barrier $N > 1/(1-\gamma)^2$ appeared in Agarwal et al. [1].* Take Section 4.3 of the Arxiv version of Agarwal et al. for example: the contraction factor $\gamma\sqrt{\frac{8\log(|\mathcal{S}||\mathcal{A}|/(1-\gamma)\delta)}{N}}\frac{1}{1-\gamma}$ needs to be smaller than 1, which requires the sample size per state-action pair to exceed $N > 1/(1-\gamma)^2$ (up to log factor). We shall explain it in the final paper.

Specific questions by Reviewer 3: 1. *Motivation for perturbed rewards and paper structure*: Thanks for the helpful suggestion. We will elucidate the motivation and intuition of reward perturbation earlier on in the revised paper.

2. *Empirical evaluation*: We have conducted some illustrative numerical experiments which will be added to the revised paper; see the response above on “Numerical evaluation”.

3. *Paper structure and broader impacts*: We will restructure the appendix (so as to reduce repetition) and add new statements about potential societal impacts as suggested by the reviewer.

Specific questions by Reviewer 4: 1. *Constraint on accuracy level ε.* In fact, our result (see Theorem 1) holds for an arbitrary choice of $\varepsilon \in (0, \frac{1}{1-\gamma}]$. This accommodates an arbitrary ε, and there is no lower bound on the accuracy level ε in our main theorem. To be a bit more specific, our sample complexity bound reads $\frac{|\mathcal{S}||\mathcal{A}}{(1-\gamma)^3\varepsilon^2}$, while prior work like Agarwal et al. reads $\max\left\{\frac{|\mathcal{S}||\mathcal{A}}{(1-\gamma)^3\varepsilon^2}, \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\right\}$ (up to some log factors). Here, the red text represents the sample size barrier, and is removed in our theory. We understand from the reviewer’s comment that there might be confusion in our current discussions/remarks, and we shall rephrase them in the final paper to clarify any such confusion.

2. *The role of reward perturbation and choice of α.* Reward perturbation is introduced mainly to break the ties (so as to ensure uniqueness of optimal policy). Fortunately, even an extremely small level of perturbation suffices for this purpose (as long as it is not exponentially small). Our analysis and theorems hold for any fixed constant $\alpha > 1$. Here, we pick $\alpha = 5$ only to emphasize that a fairly small level of reward perturbation suffices for our analysis to work (more specifically, a fairly small level of perturbation suffices in breaking the ties). This will be made clear in the final paper.

3. *About perturbed reward samples.* Here, reward perturbation is only enforced when running the planning algorithms on empirical MDPs (basically, we collect the true reward samples r, but use $r_{\mathrm{p}} = r + \zeta$ in the algorithm). In other words, it is not necessary to collect new samples for this purpose; we shall clarify this point in the revision.

4. *Missing state-action dependency in Eq. (10).* Here, N is defined as the sample size *per state-action pair* (so that the total sample size should be $|\mathcal{S}||\mathcal{A}|N$. We shall make it more clear in the revision.