We thank all the reviewers for their insightful comments and for the acknowledgement of our contributions in this work. We intend to do our best to incorporate their feedback into our revision. Here we would like to use this opportunity to clarify some *important aspects* of our work first and respond to the *remaining points* raised by each reviewer.

1. Generative model (GM) [R#1, R#2]: We consider the setup with a GM in this work, which is common in theoretical RL literature. Our primary aim is to argue the utility of ME in exploiting low-rank structures for RL. By assuming a GM, we focus on providing key insights of the framework, while offering comparisons with structure-oblivious literature. However, we strongly believe our proposal is more broadly applicable, e.g., to the online setup. Indeed, we can still apply the "sample and pseudo-explore (via ME)" scheme with appropriate modifications. The most prominent challenge anticipated in online setup is that we are no longer able to sample "any" state-action pair freely and adaptively; the sampling needs to respect the exploration policy. This also implies that a more refined ME method needs to be designed to handle the difficulties caused by diminished sampling capability. We believe our structured RL can bring a similar complexity gain in general setups, and hope this can motivate further research in both RL and ME communities.

2. Discount factor [R#3, R#4]: It is true that our analysis requires γ to be small. We would like to clarify that the requirement stems from our analysis of ME method, and it does not necessarily indicate the limitation of the proposed framework. As a matter of fact, we used large values of γ in our experiments to give evidence that our algorithm can be effective beyond the range of γ allowed in analysis. The constraint on γ arises from requiring $\|Q^{(t)} - Q^*\|_\infty$ to decrease in the proof of Theorem 2. Our algorithm combines one-step lookahead and ME to update $Q^{(t)}$; the ME step "amplifies" the error by a factor of at most c_{me} with Assumption 1. In the end, $\gamma \mathsf{c}_{\mathsf{me}} < 1/2$ is required. However, we believe it is an artifact of the conservative nature of our decoupled analysis. Indeed, there are several ways to relax the restriction, e.g., by improving analysis to achieve a better ℓ_∞ guarantee or by devising novel ME methods for RL.

3. Anchor points [R#1, R#4]: Our proposed ME method relies on the existence and availability of a set of anchor states/actions. The existence of an anchor set is straightforward from linear algebra. Viewing $Q^*(\mathcal{S}, \mathcal{A})$ as a (possibly infinite-sized) matrix of rank r, there exists $\mathcal{S}^\sharp \subset \mathcal{S}$ s.t. $\{Q^*(s, \mathcal{A}) : s \in \mathcal{S}^\sharp\}$ spans the row space of $Q^*(\mathcal{S}, \mathcal{A})$ (likewise, $\exists \mathcal{A}^\sharp$ spanning the column space of $Q^*(\mathcal{S}^\sharp, \mathcal{A})$). Now there remains the algorithmic question. We did not discuss this point in the paper to avoid digression to secondary details, but we believe it won't be hard to find $\mathcal{S}^\sharp, \mathcal{A}^\sharp$ under mild assumptions. For example, if the principal components of Q^* are "incoherent" (i.e., $\|f_i\|_\infty / \|f_i\|_2$ and $\|g_i\|_\infty / \|g_i\|_2$ are small for $i \leq r$) and there is a sufficient "eigengap" (σ_r is well separated from 0), then a random sample of sufficient size would yield anchor sets with high probability, cf. discussions in Appendix G.2. Lastly, we remark that the anchor selection needs not be perfect in practice, due to the robustness implied by our results for approximate rank-r setup.

[Reviewer #1] Warm-up example. First, we would like to clarify that this is only a toy example meant to develop readers' intuition with elementary analysis. Indeed, positive reward is not really needed for our general results and in this toy case, one might also consider shifting the rewards by adding a constant to ensure c_{me} is not prohibitively large.

Sample complexity. The goal of this work is to study the sample complexity for RL with continuous $\mathcal{S}, \mathcal{A}$, and to understand if the dependence on the "size" of $\mathcal{S}$ and $\mathcal{A}$ can be improved by exploiting the structure. As the dependence on γ is of secondary interest to us, we treat γ as a constant and hide it in the big-O notation. We included the dependence on γ of the results from [3], [35], [36] in Table 1 for complete reference, however, we can omit them to avoid confusion.

[Reviewer #2] Interpolation methods. Note that we balance three error terms arising from Steps 2, 3, 4 of our RL-ME algorithm in the proof of Theorem 2 by choosing parameters appropriately. In this work, we only assume Lipschitz smoothness of Q^*; thus we don't expect order-wise gain from other interpolation methods and the eventual $\beta^{(t)}$-net needs to be as fine as $O(\epsilon)$. However, it could be beneficial to use a more refined method, e.g., local polynomial interpolation, if we assume higher-order smoothness or a parametric family of Q functions, such as neural networks.

Experimental section. We agree, and it was mainly due to the limited space. For the computational costs, nuclear norm minimization is known to be expensive for large matrices. Here we show our computational benefits: we summarize the ME runtime for Inverted Pendulum with a 2500×1000 matrix at one iteration. We will add related results in our revision.

ME	Soft-Imp.	Nuc. norm	**Ours**
Runtime (s)	41.5 ± 1.7	76.3 ± 8.2	**1.9 ± .6**

[Reviewer #3] Experiments. This work is primarily focused on theoretical introduction of the novel, powerful low-rank RL framework. While we validate it via classical tasks, we surely welcome suggestions for more practical examples.

Lower bound. This is achieved by viewing Q-function as a $(d_1 + d_2)$-dimensional function and interpreting the estimation of Q^* as a non-parametric regression problem. Classical minimax lower bound (e.g., Sec. 2.6 of [44]) would imply the result. Indeed, the lower bound for continuous $\mathcal{S}$ & finite $\mathcal{A}$ from [33] was obtained in the same manner.

[Reviewer #4] Nuclear norm minimization (NNM). Although NNM works well in practice and has been extensively studied, there is no satisfactory ℓ_∞ guarantee for it known so far. As we consider "provable" sample efficiency, we need a new ME method. Moreover, NNM is computationally expensive; the table above demonstrates our method is **40x** faster than NNM. Please find further discussions on the "failure" of existing ME methods in Appendix G.2. We believe these already highlight our contributions and address why technical advances for ME are needed in combining with RL.