[General Response] We thank all reviewers for the detailed and valuable feedback. We will fix the typos and improve the draft carefully based on your comments.

Q1.Limitation of the deterministic setting and extension to stochastic settings: Our main contribution is designing a simple value iteration style algorithm to get the interval estimation. For the sake of clarity, we choose to focus on the deterministic transition and reward settings. One way to extend our results to stochastic settings is to follow the Appendix C proof's technique in [1], by decomposing the empirical operator $\hat{\mathcal{B}}^\pi$ as $\mathcal{B}^\pi + (\hat{\mathcal{B}}^\pi - \mathcal{B}^\pi)$, where we can bound the latter operator $(\hat{\mathcal{B}}^\pi - \mathcal{B}^\pi)$ via Rademacher complexity of the Lipschitz function class. To avoid digressing from the primary focus of the current paper, we decide to leave them as our future work.

[Reviewer #1] We thank reviewer #1 for your valuable suggestions and detailed writing corrections.

Q2.Misleading claim on non i.i.d. assumption: Even when assuming deterministic transitions, typical approaches based on concentration inequalities still require i.i.d. conditions on the transition (s_i, a_i) pairs, or can only work on the trajectory level as in [2] (which has a smaller effective sample size). We plan to have a new section to compare the concentration inequality approach with our method side by side to further clarify the raised concerns.

[Reviewer #2] We thank reviewer #2 for your insightful questions and valuable experimental references.

Q3.Empirical comparison to existing approaches: We have a comparison with [2] in Table 1 of Appendix D. In general, [2] views each trajectory as one sample while we view each transition pair as one sample. The result in Appendix D shows that our method is better than that of [2] when the number of trajectories is small. Moreover, like other trajectory-based IS methods, [2] also suffers from the curse of horizon. We will add more discussion in the revision.

Q4. Reason to choose Lipschitz function class: We had a brief discussion on this in line 128-129. We choose the Lipschitz function class to make a good balance between expressiveness and tightness of the bounds. More specifically, it includes a very rich set of functions that could cover the true value function with high probability, while allowing us to get practical bounds with a simple algorithm. We will add more discussion in the final draft.

Q5.Distance function and high dimensional state space: We find out L_2 distance is enough for the low-dimension environment. In high-dimension data, we may need to find a better distance measure to capture the underlying low-dimension manifold of the data, which seems to be an exciting direction to explore. We will leave it as future work.

[Reviewer #3] We thank reviewer #3 for your valuable comments and suggestions.

Q6.Quadratic dependency on sample size and random sample method seems not ideal: We avoid the quadratic dependency by adopting the random sub-sampling technique, which may sacrifice the tightness of the interval bounds for reducing the computation burden. Moreover, as the sub-sampling bounds are still provably correct bounds of the true R^π (despite being less tight), the sub-sampling interval bounds can still guide the end-user for decision-making. In real world applications, we can trade off the tightness and the computation complexity conditioning on the available computation resource.

Q7.Require knowledge of Lipschitz constant: We agree that Lipschitz constant is crucial for the success of a valid interval estimation. We emphasize and discuss it in section 4.2.

Q8.Related literature on estimating distribution min/max form which can improve equation (14): Thanks for pointing out the reference. It sounds very interesting! Could you kindly send the references in the revised rebuttal?

[Reviewer #4] We thank reviewer #4 for your valuable comments and suggestions.

Q9.The sample complexity appears to be exponential in the effective dimension: We agree that the sample complexity is exponential and the main reason to choose Lipschitz function class is because it strikes a good balance between richness and simplicity. In line 180 we pointed out that: "it is possible to choose smaller space sets (such as RKHS) to obtain smaller gaps, it would scarify other properties such as capacity and simplicity."

Q10.Theorem 3.2 for the specified initial points: We can start with an arbitrary initial point and still achieve linear convergence as in Proposition 3.3, which means that when the algorithm converges we will get a pair of provably correct bounds. However, with the specified initial point in Theorem 3.2, we can have a stronger guarantee on **anytime** bounds, which means that whenever we stop the algorithm (before it converges), the upper and lower bounds we get is guaranteed to include R^π. Moreover, we believe that the calculation of the initial point is not difficult (see Eq (12)).

[References]

[1] Mousavi, Li, Liu, Zhou. Black-box Off-policy Estimation for Infinite-Horizon Reinforcement Learning, ICLR 2020.

[2] Thomas, Theocharous, Ghavamzadeh. High-confidence off-policy evaluation. AAAI 2015