We thank the reviewers for carefully reviewing our paper and providing constructive feedback.

Responses to Reviewer 1. Our analysis assumes discrete action space but the bound scales as $\log(A)$, which means that in theory can handle large action spaces. Continuous action spaces can also be handled as long as the KL divergence of the initial action distribution (e.g.,$\pi^0(.|s)$) and the comparator's action distribution (e.g., $\pi^\star(.|s)$) is bounded. We will elaborate this point in the revised version.

Responses to Reviewer 2. We agree with the reviewer that our current PAC bound is probably not tight and there is much room to potentially improve the bound. Part of the reason is that our approach is on-policy and model-free which makes the dependence on parameters worse than those methods which re-use off-policy data, to perform either LSVI or model-based VI, for instance.

The upshot of our on-policy model-free approach is the robustness to modeling errors that we establish, such as Theorem 3.6 for the classical problem of imperfect state aggregation (see Theorem 3.6). So we believe that there is a trade-off between the best bounds under strong modeling assumptions and more broadly robust techniques. Getting the best of both using a single method is a fascinating direction for future research. We thank the reviewer for recognizing the high-level motivation behind this work though, which is the development of a theoretically sound approach amenable to use in conjunction with practical deep learning and PG methods.

Regarding reset: unlike most policy optimization approaches' analysis, we only assume that we can reset to a fixed initial state (results extend to resetting to a fixed initial distribution), which in our perspective, is equivalent to the common episodic finite horizon setting where agent is also reset at the end of each episode.

Responses to Reviewer 3. We are happy to include a more detailed discussion of our approach versus POLITEX (indeed we have already done so in our revision). Note that POLITEX does not explicitly address exploration. Instead it assumes *every* policy is able to visit every state (e.g., Assumption 4 in POLITEX) which is a strong assumption for the underlying MDPs. In other words, POLITEX is not a PAC algorithm for general tabular MDPs and is similar to other works we cite such as [1, 9, 22, 32, 54] in that respect, while EPOC is.

Regarding the sample complexity of EPOC, we agree with the reviewer that there is some data waste. One of the reasons is that we are aiming for a model-free and on-policy algorithm, which potentially wastes samples (as we do not re-use off-policy data from previous rounds), but we get more robustness result under model-misspecification, as we demonstrated in the classic state-abstraction setting (Theorem 3.6). See also response to R2 above.

Thanks for pointing out the slide issue with the learning rate in B.3. We typically assume the number of iterations T is large and at least no smaller than $\log(A)$, which we will clarify in the revised version.

Additional clarifications for the final version: A few clarifications and corrections are worth explicitly mentioning to avoid any ambiguities, which we will include in the final version. The critic estimation (Line 6 Alg 2) in the final version will be revised so that $Q^\pi(s, a; r + b^n)$ will be changed to to $Q^\pi(s, a; r + b^n) - b^n$ because this is a convenient change for the special case of linear MDPs, as the shifted Q-value is always linear function of $\phi(s, a)$ (since the reward as well as Bellman backup of *any function* are always linear in $\phi(s, a)$), without any need to augment the features as we current did for linear MDPs. The change does not affect the sample complexity or algorithmic properties for the other cases which we study (such as state aggregation, tabular results, and the general agnostic result). This also adresses a minor misspecification in statement of the algorithm (and proof) for linear MDP special case, due to that the current known set definition may be ambiguous to the reader.