We would like to thank all reviewers for their valuable comments and time. Please see responses below:

Reviewer 1.

- Regarding the connections to Bellman rank and FLAMBE, since the algorithm style and underlying mechanisms of their papers are quite different from ours, to our best knowledge, we are not aware of direct connections based on our current analysis. However, our approach is model-based, which at a high level bears some similarity to the witness rank approach. We believe that a modified version of witness rank might be feasible for POMDPs, and a rank-based algorithm is also possible if combined with our current operator analysis. We agree this is an interesting future direction that is worth exploring.
- Proposition 1 shows that RL of POMDP is intractable in general if we do *not* assume undercompleteness. So the hard instance constructed in Proposition 1 is *overcomplete*. Proposition 2 shows that only assuming undercompleteness is also not enough. So the hard instance constructed there is *undercomplete*.
- Yes, the operator B defined in equation (1) has the property that its rank is at most S. We will clarify this.

Reviewer 2. Regarding the seven questions in the weakness section:

- Thanks for the suggestion about derivations. We will add more explanations about these derivations.
- For the equation between Lines 279 and 280, we note that π is a *deterministic* policy and $\Gamma(\pi, H)$ is a set of all the observation and action sequences of length H that could occur under policy π, i.e., for any $\tau_H = (o_H, \dots, a_1, o_1) \in \Gamma(\pi, H)$, we have $\pi(a_{H-1} \dots, a_1 \mid o_H, \dots, o_1) = 1$, and $\pi(a'_{H-1} \dots, a'_1 \mid o_H, \dots, o_1) = 0$ for any action sequence $(a'_{H-1} \dots, a'_1) \neq (a_{H-1} \dots, a_1)$. Therefore, for $\tau_H \in \Gamma(\pi, H)$, we have:

$$\mathbb{P}^\pi_\theta(o_H, \dots, o_1) = \textstyle\sum_{a'_{H-1} \in \mathscr{A}} \cdots \sum_{a'_1 \in \mathscr{A}} \mathbb{P}^\pi_\theta(o_H, a'_{H-1}, \dots, a'_1, o_1) = \mathbb{P}^\pi_\theta(o_H, a_{H-1}, \dots, a_1, o_1)$$
$$= [\textstyle\prod_{h=1}^{H-1} \pi(a_h \mid o_h, \dots, a_1, o_1)] \cdot [\prod_{h=1}^{H} \mathbb{P}_\theta(o_h \mid a_{h-1}, \dots, a_1, o_1)] = \prod_{h=1}^{H} \mathbb{P}_\theta(o_h \mid a_{h-1}, \dots, a_1, o_1)$$
$$= \mathbb{P}_\theta(o_H, \dots, o_1 | a_{H-1}, \dots, a_1).$$

- The inequality in Line 492 follows from standard vector-valued martingale concentration (e.g. see Corollary 7 in "A Short Note on Concentration Inequalities for Random Vectors with SubGaussian Norm" by Jin et al.). Here, we vectorize the tensor, then the Frobenius norm becomes the ℓ_2-norm of the vector. The upper bound on the Forbenius norm of the tensor is given at the beginning of the proof (Lines 490-491). The stated result is missing a $\log |O|$ factor. We will correct this and explain further in the final version.
- Please see Lines 7-11 in Algorithm 1, for *each* $(h, a, \tilde{a})$ triple, we *re-execute* the policy from the first step (i.e. start a new episode) to collect data (o_{h-1}, o_h, o_{h+1}) for $\mathbf{M}_h$ and $\mathbf{N}_h$. Therefore the samples for constructing $\mathbf{M}_h$/$\mathbf{N}_h$ and $\mathbf{M}_{h+1}$/$\mathbf{N}_{h+1}$ in Algorithm 1 are coming from completely different episodes. That is, other than the roll-in policy is the same, the samples are independent/have no undesirable correlation.
- Our result holds for any absolute constant probability $p < 1$. For general probably $1 - \delta$, directly applying our result will incur an additional $\text{poly}(1/\delta)$ factor in the sample complexity. However, one can easily improve this dependency to only $\text{polylog}(1/\delta)$ factor by computing $\log(1/\delta)$ independent policies that are near-optimal with only constant probability each, and pick the best one by evaluating each policy for $O(H^2/\epsilon^2)$ episodes.
- Yes, our savings of sample complexity is exponential compared to the naive $(OA)^H$. Our polynomial dependency is $\mathcal{O}(O^4 S^7 A^4 H^6/\alpha^4)$. We will provide this explicit dimension dependence in the final version.
- Thanks for pointing out this related work, we will add a discussion in the final version. The idea of using statistics o_{h-1}, o_h, o_{h+1} for learning parameters stems from the earlier HMM literature [11]. Azizzadenesheli et al. 2020 is different from us in that it considers only Markovian policies and does not address exploration.

Reviewer 3.

- Thanks for pointing out these typos! Yes, Line 565 should be $T_h(s \mid s', a)$.
- Thanks for the reference to Blai Bonet's paper. The main difference from our paper is that we assume deterministic initial state but stochastic emission process, while they assume stochastic initial state but deterministic emission process. In addition, their result is on planning while assuming the model is known, while our result requires the learning of the model. We will add more discussion on the related works of deterministic POMDPs.
- For the claim in Lines 578-579, since we assume the initial state is fixed (see definition in Line 237), once we have learned the underlying transition matrices from the stochastic observations (Algorithm 2), we can directly identify the current hidden state by looking only at the sequence of actions taken.