We thank all the reviewers for their valuable feedback and appreciating our contributions. Please find our response to each individual reviewer below.

—— To Reviewer #2 ——

Line 13 of the algorithm. As correctly observed by Reviewer #1, in Line 13 of Algorithm 2, we clear the dataset and retake samples to guarantee the independence of samples in the dataset. Our current analysis (Lemma B.2) critically relies on such independence, and it is unclear if the algorithm is still correct after removing this step. We will add more discussion on this step in the next version.

Avoid $\mathrm{poly}(H)$ dependence in deterministic environment. In deterministic systems, if a state-action pair (s, a) is reachable, then there exists a policy that always visits (s, a), and in order to find the optimal policy, it suffices for the agent to visit each reachable state-action pair once. Therefore, a simple algorithm would be the following: whenever there exists a state-action pair (s, a) that has not been visited, sample a trajectory to try to visit (s, a), observe the reward value r and the transition s', and mark (s', a') to be unvisited for all actions a'. Clearly, after sampling $|\mathcal{S}| \times |\mathcal{A}|$ trajectories, the agent should have visited all reachable state-action pairs, at which point the agent could output the optimal policy by planning on the learned model.

—— To Reviewer #3 ——

What if transition probability varies as horizon changes. If the transition operator varies as the horizon changes, our algorithm can no longer achieve $\mathrm{poly}(\log H)$ dependency. To name one reason, the size of the ε-net defined in Section 5.1 now has exponential dependency on H. However, in such a setting, one can prove a lower bound of $\Omega(H)$ on the number of episodes. Such a lower bound can be proved by, e.g., using the standard combination lock environment. We will add more discussion on this in the next version.

Typos. Thanks for pointing out. We will fix these typos in the next version.

—— To Reviewer #4 ——

Comparison with previous results. First of all, we do not claim our result strictly improves existing results. The current discussions on previous results in Section 3 primarily focus on their dependency on H. We totally agree that when ε is sufficiently small (e.g. $\varepsilon \ll 1/H$), the sample complexity of our algorithm is worse than that of previous algorithms. In the next version, we will make this point explicit in the introduction, and provide more comparisons between the sample complexity of our algorithm and that of previous algorithms to make the complexity landscape clearer. Meanwhile, as mentioned by Reviewer #2, this paper is the first one that proposes an algorithm to achieve $\mathrm{poly}(\log H)$ dependence on the number of episodes.

Answering the open problem in [Jiang and Agarwal, 2018]. In terms of answering the open problem in [Jiang and Agarwal, 2018], their problem statement is *"Can we prove a lower bound that depends polynomially in H?"* Furthermore, [Jiang and Agarwal, 2018] mentioned explicitly that the relevant setting is when $\varepsilon \gg 1/H$ (see the paragraph before Section 2.2 in [Jiang and Agarwal, 2018]). In this sense, our work resolves the open problem in [Jiang and Agarwal, 2018] with a negative answer.

The key aspects of the algorithm that enable this result would be very helpful to highlight in the main paper. Right now we have provided a technical overview in Section 4.1. We are happy to add more discussion on the novel aspects of the analysis in the next version. Thanks for the suggestion!

Why a new set of trajectories are needed every time the rollout is executed. As correctly observed by Reviewer #2, in Line 13 of Algorithm 2, we clear the dataset and retake samples to guarantee the independence of samples in the dataset. See the discussion above regarding this for more details.

Section 5.1 could be kept in the appendix. We decided to keep Section 5.1 in the main text since it could be of interest to a broad audience. For example, Reviewer #2 mentioned that "the idea of epsilon net is novel and new to me". However, we are glad to move parts of Section 5.1 to the appendix and add more discussion on other novel aspects of our analysis.