Overcoming the Long Horizon Barrier for Sample-Efficient Reinforcement Learning with Latent Low-Rank Structure

Tyler Sam[1], Yudong Chen[2], and Christina Lee Yu[1]

[1]School of Operations Research and Information Engineering, Cornell University
[2]Department of Computer Sciences, University Wisconsin-Madison

February 28, 2022

Abstract

The practicality of reinforcement learning algorithms has been limited due to poor scaling with respect to the problem size, as the sample complexity of learning an ϵ-optimal policy is $\tilde{\Omega}\left(|S||A|H^3/\epsilon^2\right)$ over worst case instances of an MDP with state space S, action space A, and horizon H. We consider a class of MDPs that exhibit low rank structure, where the latent features are unknown. We argue that a natural combination of value iteration and low-rank matrix estimation results in an estimation error that grows doubly exponentially in the horizon H. We then provide a new algorithm along with statistical guarantees that efficiently exploits low rank structure given access to a generative model, achieving a sample complexity of $\tilde{O}\left(d^5(|S|+|A|)\mathrm{poly}(H)/\epsilon^2\right)$ for a rank d setting, which is minimax optimal with respect to the scaling of $|S|, |A|$, and ϵ. In contrast to literature on linear and low-rank MDPs, we do not require a known feature mapping, our algorithm is computationally simple, and our results hold for long time horizons. Our results provide insights on the minimal low-rank structural assumptions required on the MDP with respect to the transition kernel versus the optimal action-value function.

Contents

1 Introduction **3**

2 Related Work **4**

3 Preliminaries **5**

3.1 Doubly Exponential Growth in Long Horizons 6

4 Assumptions **6**

5 Algorithm **8**

5.1 Matrix Estimation Subroutine 9

5.2 Formal Algorithm Statement 10

6 Main Results **11**

7 Proof Sketch **12**

8 Conclusion **13**

A Extended Literature Discussion **18**

B Matrix Estimation Method **19**

C Omitted Proofs **24**

C.1 Proof of Lemma 1 24

C.2 Proof of Proposition 2 26

C.3 Supplement to Section 6 27

D Technical Propositions and Theorems **33**

1 Introduction

Reinforcement learning (RL) methods have been increasingly popular in sequential decision making tasks due to its empirical success, e.g., Atari Games [25], StarCraft II [34], and robotics [24]. However, the usage of RL is limited by the high computational resource costs consumed in the training process, resulting from poor scaling of RL algorithms with respect to the size of the state and action space. Many real-world problems when modeled as a Markov decision process (MDP) have very large state and action spaces. For example, the n-city Capacitated Vehicle Routing Problem (CVRP), a classical combinatorial problem from operations research, involves a state space $\{0,1\}^n$ and an action space being all partial permutations of $n-1$ cities [10]. Given a finite-horizon MDP with state space S, action space A, and horizon H, one needs $\tilde{\Omega}\left(|S||A|H^3/\epsilon^2\right)$ samples given a generative model to learn an optimal policy [29], which can be impractical when S and A are large.

The above tabular RL framework does not capture the fact that many real-world systems in fact have additional structure that if exploited should improve computational and statistical efficiency. The critical question becomes what structure is reasonable to assume, and how to design new algorithms and analyses to efficiently exploit it. In this work, we focus on the subclass of MDPs that exhibit latent low-dimensional structure with respect to the relationship between states and actions, which refers to the action-value function associated to the optimal policy, near-optimal policies having low rank (when viewed as a $|S|$-by-$|A|$ matrix), or the transition kernel having low Tucker rank (when viewed as a tensor). While the sample complexity under the fully general model scales as $|S||A|$, we would expect that the sample complexity under a rank-d model would scale as $d(|S|+|A|)$, as the low rank assumption on a matrix reduces the degrees of freedom from $|S||A|$ to $d(|S|+|A|)$. While this intuition holds true in the classical low rank matrix estimation setting, the additional dynamics of the MDP introduce complex dependencies that may amplify the error for long horizons.

Our Contributions. We study minimal low rank structural assumptions that allow for computationally and statistically efficient learning, reducing the sample complexity from scaling with $|S||A|$ to only $|S|+|A|$. First, we show that there are additional complexities that arise from MDPs with long horizons. In particular, we provide an example where the optimal action-value function Q^* is low rank, yet the error of policy evaluation for an optimal policy grows doubly exponentially in the horizon H under a natural combination of matrix estimation and value iteration suggested in [28]. This error amplification in *long horizons* illustrates that exploiting low rank structure in RL is significantly more involved than classical matrix estimation. We propose a new computationally simple model-free algorithm, referred to as Low Rank Monte Carlo Policy Iteration (LR-MCPI). Under the assumption that Q^* is low rank, by additionally assuming a constant suboptimality gap, we prove that LR-MCPI achieves the desired sample complexity, avoiding the exponential error amplification in the horizon. Additionally we prove that LR-MCPI also achieves the desired sample complexity when all ϵ-optimal policies π have low rank Q^π functions. Under the stronger assumption that the transition kernel and reward function have low rank, we show that the model-free algorithm in [28], which we refer to as Low Rank Empirical Value Iteration (LR-EVI), also achieves the desired sample complexity. The following table summarizes our sample complexity bounds in their corresponding settings, and compares them with existing results from literature in the tabular finite-horizon MDP setting; here d refers to the rank parameter.

MDP Setting	Sample Complexity
Low-rank Q_h^* & suboptimality gap $\Delta_{\min} > 0$ (Theorem 3)	$\tilde{O}\left(\frac{d^5(\lvert S\rvert+\lvert A\rvert)H^4}{\Delta_{\min}^2}\right)$
ϵ-optimal policies have low-rank Q_h^π (Theorem 4)	$\tilde{O}\left(\frac{d^5(\lvert S\rvert+\lvert A\rvert)H^6}{\epsilon^2}\right)$
Transition kernels and rewards are low-rank (Theorem 5)	$\tilde{O}\left(\frac{d^5(\lvert S\rvert+\lvert A\rvert)H^5}{\epsilon^2}\right)$
Low-rank Q_h^* & constant horizon [28]	$\tilde{O}\left(\frac{\lvert S\rvert+\lvert A\rvert}{\epsilon^2}\right)$
Tabular MDP with homogeneous rewards [29]	$\tilde{\Theta}\left(\frac{\lvert S\rvert\lvert A\rvert H^3}{\epsilon^2}\right)$

2 Related Work

As the body of work on RL theory is large, we present a focused literature review on works that also assume low rank structure on the action-value function when viewed as a matrix, or low Tucker rank of the transition kernel. In Appendix A we provide an extended discussion on the related literature.

[42, 28] consider the weakest low-rank assumption that only imposes low rank on Q^*, implying that the interaction between the states and actions decomposes. [42] show empirically that this assumption is satisfied in common stochastic control tasks. Their numerical experiments demonstrate that value iteration combined with existing matrix estimation methods requires significantly fewer samples compared to vanilla value iteration. [28] develops an algorithm that combines a novel matrix estimation method with value iteration to find an ϵ-optimal action-value function with $\tilde{O}\left(d(|S|+|A|)/\epsilon^2\right)$ samples for infinite-horizon γ-discounted MDPs assuming that Q^* has rank d. While this is a significant improvement over the tabular lower bound $\tilde{\Omega}\left(|S||A|/((1-\gamma)^3\epsilon^2)\right)$ [4], their results require strict assumptions. The primary limitation is they require the discount factor γ to be bounded above by a small constant, which effectively limits their results to short, *constant* horizons. Relaxing this assumption is left as an open question in their paper. In this work, we provide a concrete example that illustrates why long horizons may pose a challenge for using matrix estimation in RL. Subsequently we show that this long horizon barrier can be overcome by imposing additional structural assumptions. A second limitation of their result is that their algorithm relies on prior knowledge of special anchor states and and actions that span the entire space. We will show that under standard regularity conditions, randomly sampling states and actions will suffice.

[27] consider a setting in which the transition kernels have low Tucker Rank along all three modes, under which they propose an algorithm that learns low-dimensional state and action representations that can be used to estimate the transition kernel. This assumption on the transition kernel is similar to the structural assumptions from the Markov process state aggregation model, e.g., [15, 43]. These works only provide bounds on the estimation error of the latent feature estimates for the transition kernel without explicit sample complexity guarantees for finding an optimal policy. We consider a similar yet relaxed assumption which only imposes that the Tucker Rank is low along *two* modes rather than all three. Additionally, we develop a model-free approach that does not directly estimate the transition kernel.

Another class of works assumes a different type of low Tucker rank assumption, which they refer to as low-rank MDPs [2, 26, 33]. Their model is equivalent to the linear MDP setting where the feature representation is unknown. In particular, low rankness is imposed on the transition kernel with respect to the relationship between the destination state s' and the originating state-action pair (s, a), whereas we impose a decomposition between s and a. These works often require access to a strong optimization oracle or assume access to a finite class of potential latent representation functions $\mathcal{M}$. While the sample complexity bounds only depends logarithmically on $|\mathcal{M}|$, this will still be too costly if we consider a set $\mathcal{M}$ that covers all low rank representations [33]. A strength of their result however, is that they do not require access to a generative model, and can obtain guarantees in a reward-free setting [2, 26, 33].

3 Preliminaries

We consider a standard finite-horizon MDP given by (S, A, P, R, H) [31]. Here S and A are the state and action spaces, respectively, both of which are finite. $H \in \mathbb{Z}_+$ is the time horizon. $P = \{P_h\}_{h\in[H]}$ is the transition kernel, where $P_h(s'|s,a)$ is the probability of transitioning to state s' given upon taking action a in state s at step h. $R = \{R_h\}_{h\in[H]}$ is the reward function, where $R_h : \mathcal{S} \times \mathcal{A} \to \Delta([0,1])$ is the distribution of the reward for taking action a in state s at step h. We will use $r_h(s,a) := \mathbb{E}_{r\sim R_h(s,a)}[r]$ to denote the mean reward. A stochastic, time-dependent policy of an agent has the form $\pi = \{\pi_h\}_{h\in[H]}$ with $\pi_h : S \to \Delta(A)$, where the agent selects an action according to the distribution $\pi_h(s)$ at time step h when at state s.

For each policy π and $h \in [H]$, the value function and action-value function of π are defined as

$$V_h^\pi(s) := \mathbb{E}\left[\textstyle\sum_{t=h}^H R_t \;\middle|\; s_h = s\right], \tag{1}$$

$$Q_h^\pi(s,a) := r_h(s,a) + \mathbb{E}\left[\textstyle\sum_{t=h+1}^H R_t \;\middle|\; s_{h+1} \sim P_{t-1}(\cdot|s,a)\right], \tag{2}$$

where $R_t \sim R_t(s_t, a_t)$, $a_t \sim \pi_t(s_t)$, and $s_t \sim P_{t-1}(\cdot|s_{t-1}, a_{t-1})$. The optimal value and action-value functions are given by $V_h^*(s) := \sup_\pi V_h^\pi(s)$ and $Q_h^*(s,a) := \sup_\pi Q_h^\pi(s,a)$, respectively, for all $s \in S, h \in [H]$. These functions satisfy the Bellman equations

$$V_h^*(s) = \max_{a\in A} Q_h^*(s,a), \quad Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^*(s')], \quad \forall s,a,h \tag{3}$$

with $V_{H+1}^*(s) = 0$. For an MDP with finite spaces and horizon, there always exists an optimal policy π^* that satisfies $V_h^{\pi^*}(s) = V_h^*(s)$ for all s, h.

A general goal in this work is to find a near-optimal policy or action-value function. For $\epsilon > 0$, π is an ϵ-optimal policy if $|V_h^*(s) - V_h^\pi(s)| \leq \epsilon, \forall (s,h) \in S \times [H]$. Similarly, $Q = \{Q_h\}_{h\in[H]}$ is called an ϵ-optimal action-value function if $|Q_h^*(s,a) - Q_h(s,a)| \leq \epsilon, \forall (s,a,h) \in S \times A \times [H]$.

We will view Q_h^*, Q_h^π and R_h as $|S|$-by-$|A|$ matrices and $P_h(\cdot|\cdot,\cdot)$ as an $|S|$-by-$|S|$-by-$|A|$ tensor, for which various low-rank assumptions are considered. We sometimes use the shorthand $[P_h V_{h+1}](s,a) := \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}(s')]$ for the conditional expectation under P_h.

Throughout this paper, we assume access to a simulator (a.k.a. the generative model framework introduced by [22]), which takes as input a tuple $(s,a,h) \in S \times A \times [H]$ and outputs independent samples $s' \sim P_h(\cdot|s,a)$ and $r \sim R_h(s,a)$. This assumption is one of the stronger assumptions in reinforcement learning literature, but common in the line of work that studies sample complexity without directly addressing the issue of exploration, e.g., [29, 3].

Notation. Let $a \wedge b := \min(a,b)$, $a \vee b := \max(a,b)$, δ_a denote the distribution over A that puts probability 1 on action a, $\sigma_i(A)$ denote the i-th largest singular value of A, and A_i denote the i-th row. The n by n identity matrix is denoted by $I_{n\times n}$, and $[H] := \{1,\ldots,H\}$. We use several vector and matrix norms: Euclidean/ℓ_2 norm $\|\cdot\|_2$, spectral norm $\|A\|_{op} = \sigma_1(A)$, nuclear norm $\|A\|_*$, entrywise ℓ_∞ norm $\|A\|_\infty$ (largest absolute value of entries), and Frobenius norm $\|A\|_F$.

3.1 Doubly Exponential Growth in Long Horizons

While one may hope to learn the optimal action-value functions when only assuming that Q^*_h is low rank, we argue that the problem is more nuanced. Specifically, we present an example where, even for the seemingly easier problem of policy evaluation, a direct combination of value iteration and low-rank matrix estimation results in an estimation error of the action-value function growing doubly exponential with respect to H.

Consider an MDP (S, A, P, R, H) where $S = A = \{1, 2\}$ and Q^*_h is rank-1 for all $h \in [H]$, and π is an optimal policy to be evaluated. Suppose that we are given an inexact terminal estimate $\hat{V}^\pi_H$ and recurse backward in time $h = H, H-1, \ldots, 1$. For the state-action pairs $(s, a) \in \Omega := \{(1, 1), (1, 2), (2, 1)\}$, we compute an estimate of $Q^*_h(s, a)$ using the exact Bellman operator given $\hat{V}^\pi_{h+1}$ according to $\hat{Q}^\pi_h(s, a) = r_h(s, a) + [P_h \hat{V}^\pi_{h+1}](s, a), \forall (s, a) \in \Omega$. To estimate $Q^*_h(2, 2)$ under this rank-1 setting, we solve the natural least-squares formulation

$$\min_{u,v \in \mathbb{R}^2} \left(u_1 v_1 - \hat{Q}^\pi_h(1, 1)\right)^2 + \left(u_1 v_2 - \hat{Q}^\pi_h(1, 2)\right)^2 + \left(u_2 v_1 - \hat{Q}^\pi_h(2, 1)\right)^2 \tag{4}$$

and use the optimal (u, v) to form the estimate $\hat{Q}^\pi_h(2, 2) := u_2 v_2$. It is easy to see that the estimate is given in closed form by

$$\hat{Q}^\pi_h(2, 2) = \frac{\hat{Q}^\pi_h(1, 2)\hat{Q}^\pi_h(2, 1)}{\hat{Q}^\pi_h(1, 1)}. \tag{5}$$

This estimator corresponds to the unique rank-1 completion of $\hat{Q}^\pi_h \in \mathbb{R}^{2\times 2}$ given the entries in Ω, and also coincides with the estimator used in [28]. Finally, let $\hat{V}^\pi_h$ be the value function corresponding to $\hat{Q}^\pi_h$. We establish the following result, whose proof is deferred to Appendix C.

Lemma 1 (Doubly Exponential Growth)**.** *There exists an MDP in the above setting with horizon H and an optimal policy π, for which if the terminal estimate has error $\|\hat{V}^\pi_H - V^*_H\|_\infty = \epsilon > 0$, then for sufficiently large H, policy evaluation using the above algorithm results in*

$$\|\hat{V}^\pi_1 - V^*_1\|_\infty = \Omega\left((1 + \epsilon)^{(2^{H/2})}\right).$$

In the proof, we explicitly construct an MDP and show that the estimation error $\epsilon_h := \|\hat{V}^\pi_h - V^*_h\|_\infty$ satisfies the recursion $\epsilon_h = \epsilon_{h+1} + \epsilon^2_{h+1}$, resulting in a doubly exponential blow-up; moreover, this behavior is precisely due to the nonlinearity in the low-rank estimator (5). This example illustrates that reinforcement learning with low-rank structure is more nuanced than its linear counterpart, and that the constant horizon assumption in [28] is not merely an artifact of their analysis. This motivates us to consider additional assumptions, beyond Q^*_h being low rank, in order to achieve stable and sample-efficient learning with long horizons.

4 Assumptions

In this section, we present the key assumptions we use to show sample-efficient reinforcement learning under low-rank structural assumptions. We first present three low-rank settings which are increasing in the strength of the assumption.

Assumption 1 (Low-rank Q^*_h)**.** *For all $h \in [H]$, the rank of the matrix Q^*_h is d, i.e. it can be represented via its singular value decomposition $Q^*_h = U^{(h)}\Sigma^{(h)}(V^{(h)})^\top$, for a $|S| \times d$ matrix $U^{(h)}$, a $|A| \times d$ matrix $V^{(h)}$, and a $d \times d$ diagonal matrix $\Sigma^{(h)}$.*

Assumption 1 imposes that the optimal action-value function of the optimal policy is low rank. This assumption can be compared with another common structural assumption in the literature, namely linearly-realizable Q^*, meaning that $Q_h^*(s,a) = w_h^\top \phi_h(s,a)$ for some weight $w_h \in \mathbb{R}^d$ and *known* feature mapping $\phi_h : S \times A \to \mathbb{R}^d$ [37, 38]. In comparison, Assumption 1 decomposes ϕ_h into the product of separate feature mappings on the state space $U^{(h)}$ and the action space $V^{(h)}$. Hence, linearly-realizable Q^* does not imply low-rank Q_h^*. Furthermore, we assume the latent factors $U^{(h)}$ and $V^{(h)}$ are completely unknown, whereas the linear function approximation literature typically assumes ϕ_h is known or approximately known.

Assumption 1 only imposes low-rankness on Q_h^*, allowing for the Q^π function associated to non-optimal policies π to be full rank. Assumption 1 is likely too weak, as Lemma 1 illustrates a doubly exponential growth in policy evaluation error under only this assumption. Below we present three additional assumptions. Our results will show that Assumption 1 coupled with any single *one* of these three assumptions enables our main algorithm to achieve the desired sample complexity.

Assumption 2 (Suboptimality Gap)**.** *For each $(s,a) \in S \times A$, the suboptimality gap is defined as $\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a)$. Assume that there exists an $\Delta_{\min} > 0$ such that*

$$\min_{h\in[H], s\in S, a\in A} \{\Delta_h(s,a) : \Delta_h(s,a) > 0\} \geq \Delta_{\min}.$$

Assumption 2 stipulates the existence of a suboptimality gap. In the finite setting with $|S|, |A|, H < \infty$, there always exists a $\Delta_{\min} > 0$ for any non-trivial MDP in which there is at least one suboptimal action. This is an assumption commonly used in bandit and reinforcement learning literature.

Assumption 3 (ϵ-optimal Policies have Low-rank Q Functions)**.** *For all ϵ-optimal policies π, the associated Q_h^π matrices are rank-d for all $h \in [H]$, i.e., can be represented via $Q_h^\pi = U^{(h)}\Sigma^{(h)}(V^{(h)})^\top$ for some $|S| \times d$ matrix $U^{(h)}$, $|A| \times d$ matrix $V^{(h)}$, and $d \times d$ diagonal matrix $\Sigma^{(h)}$.*

Assumption 3 imposes that all ϵ-optimal policies π have low-rank Q_h^π. We have not seen this assumption in existing literature. Assumption 3 is implied by the stronger assumption that *all* policies have low-rank Q_h^π. The latter assumption is analogous to the property that Q^π is linear in the feature map ϕ for *all* policies, which is a key property of linear MDPs and commonly used in work on linear function approximation.

Assumption 4 (Low-rank Transition Kernels and Reward Functions)**.** *The transition kernel P_h has Tucker rank $(|S|, |S|, d)$ or $(|S|, d, |A|)$; for each $h \in [H]$, there exists a $|S| \times |S| \times d$ tensor $U^{(h)}$, an $|A| \times d$ matrix $V^{(h)}$, and an $|S| \times d$ matrix $W^{(h)}$ such that*

$$P_h(s'|s,a) = \sum_{i=1}^{d} U_{s',s,i}^{(h)} V_{a,i}^{(h)} \quad \text{and} \quad r_h(s,a) = \sum_{i=1}^{d} W_{s,i}^{(h)} V_{a,i}^{(h)}$$

for the Tucker rank $(|S|, |S|, d)$ case, or for each $h \in [H]$, there exists a $|S| \times |A| \times d$ tensor $V^{(h)}$, an $|S| \times d$ matrix $U^{(h)}$, and an $|A| \times d$ matrix $W^{(h)}$ such that

$$P_h(s'|s,a) = \sum_{i=1}^{d} U_{s,i}^{(h)} V_{s',a,i}^{(h)} \quad \text{and} \quad r_h(s,a) = \sum_{i=1}^{d} U_{s,i}^{(h)} W_{a,i}^{(h)}$$

for the Tucker rank $(|S|, d, |A|)$ case.

Assumption 4 imposes that the expected reward functions are low rank, and the transition kernels have low Tucker rank along one dimension. Assumption 4 is our strongest low-rank structural assumption as it implies that the Q_h^π functions associated with *any* policy π are low rank, which subsequently implies both Assumptions 3 and 1. In fact, Assumption 4 implies that for any value function estimate $\bar{V}_h$, the matrix $r_h + [P_h \bar{V}_{h+1}]$ is low rank, as stated in the following proposition.

Proposition 2. *If the transition kernel has Tucker rank* $(|S|, |S|, d)$ *or* $(|S|, d, |A|)$ *and the expected reward function has rank d with shared latent factors, i.e., Assumption 4 holds, then the matrix* $r_h + [P_h \bar{V}_{h+1}]$ *has rank at most d for any* $\bar{V}_{h+1} \in \mathbb{R}^{|S|}$.

Proposition 2 results from the fact that for any fixed h, the matrices corresponding to r_h and $P_h(s'|\cdot, \cdot)$ for all s' share either the same column or row space, which is critically used in the analysis of our Low Rank Empirical Value Iteration algorithm.

Next we present several regularity assumptions for matrix estimation. Even with the above low-rank assumptions and access to a simulator, we may still need $|S||A|$ samples if the reward matrices are sparse. For example, consider an MDP in the $H = 1$ setting where the reward is mostly a zeros matrix with d nonzero entries taking value 1. Since the locations of the nonzero entries are unknown, if we sample a subset of entries from the matrix, we will likely observe only zeros. While this example is primarily for illustration, the key idea is that each component of the low rank signal needs to be reasonably spread across the matrix in order to guarantee that observing a small subset of entries suffices to reconstruct the full matrix. This is formalized by the assumption of incoherence as presented in Assumption 5, which is standard in classical matrix estimation literature when assuming a random sampling model [5].

Assumption 5 (Incoherence)**.** *Let* $Q_h \in \mathbb{R}^{|S| \times |A|}$ *be a rank-d matrix with singular value decomposition* $Q_h = U \Sigma V^\top$ *with* $S \in \mathbb{R}^{|S| \times d}$ *and* $V \in \mathbb{R}^{|A| \times d}$*.* Q_h *is μ-incoherent if* $\max_{i \in [|S|]} \|U_i\|_2 \leq \sqrt{\mu d / |S|}$ *and* $\max_{j \in [|A|]} \|V_j\|_2 \leq \sqrt{\mu d / |A|}$*, where U_i denotes the i-th row of a matrix U.*

A small incoherence parameter μ ensures that the masses of U and V are not too concentrated in a couple of rows or columns, so that a randomly sampled subset of rows will span the row space, and a randomly sampled subset of columns will span the column space.

Assumption 6 (Bounded Condition Number)**.** *A rank-d matrix Q_h has bounded condition number if* $\kappa_{q_h} := \frac{\sigma_1(Q_h)}{\sigma_d(Q_h)}$ *is bounded above by a constant* $C < \infty$.

Assumptions 5 and 6 will be used in the analysis to show that the error amplification from the matrix estimation method does not scale with the size of the state or action spaces.

5 Algorithm

In this section, we describe the algorithm that admits sample-efficient reinforcement learning under the assumptions in the previous section.

The algorithm is a synthesis of matrix estimation with value iteration and Monte Carlo policy iteration. To motivate, we first describe the vanilla approximate dynamic programming algorithms for the general tabular MDP settings. Empirical value iteration simply replaces the expectation in the Bellman update in Equation (3) with empirical samples [17]. In particular, to estimate $Q_h^*(s, a)$, one collects N samples of one step transitions, which entails sampling a reward and next state from $R_h(s, a)$ and $P_h(\cdot|s, a)$. Let $\hat{r}_h(s, a)$ denote the empirical average reward of the N samples from $R_h(s, a)$. Let $\hat{P}_h(\cdot|s, a)$ denote the empirical distribution over N next states sampled from $P_h(\cdot|s, a)$.

Given an estimate $\hat{V}_{h+1}$ for the optimal value function at step $h+1$, the empirical Bellman update can be written as

$$\hat{Q}_h(s,a) = \hat{r}_h(s,a) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot|s,a)}[\hat{V}_{h+1}(s')], \quad \text{and} \quad \hat{V}_h(s) = \max_{a \in A} \hat{Q}_h(s,a). \tag{6}$$

Evaluating $\hat{Q}_h$ and $\hat{V}_h$ requires collecting N samples for each of the $|S||A|$ state action pairs (s,a).

Monte Carlo policy iteration for tabular MDPs approximates $Q_h^\pi(s,a)$ for a policy π by replacing the expectation in Equation (2) with empirical trajectory samples, which is similar to first-visit Monte Carlo policy evaluation except we use the generative model to start at a specified state-action pair and time step [31]. This involves sampling N independent trajectories starting from state-action pair (s,a) at step h and following a given policy π until the end of the horizon H. For a fixed policy π and state action pair (s,a), let the sequence of rewards along the i-th sampled trajectory be denoted $(R_h^i, R_{h+1}^i, \dots R_H^i)$. We will use $\hat{r}_h^{\text{cum}}(s,a)$ to denote the empirical average cumulative reward across the N trajectories, given by

$$\hat{r}_h^{\text{cum}}(s,a) := \frac{1}{N} \sum_{i=1}^{N} \sum_{t=h}^{H} R_t^i. \tag{7}$$

Given an estimate of the optimal policy for steps greater than h, denoted by $(\hat{\pi}_{h+1}, \hat{\pi}_{h+2}, \dots \hat{\pi}_H)$, the Monte Carlo estimate for the optimal action-value function and policy at step h would be

$$\hat{Q}_h(s,a) = \hat{r}_h^{\text{cum}}(s,a), \quad \text{and} \quad \hat{\pi}_h(s) = \delta_a \quad \text{for } a = \operatorname{argmax} \hat{Q}_h(s,a), \tag{8}$$

where the trajectories used to compute $\hat{r}_h^{\text{cum}}(s,a)$ are sampled by following the policy $(\hat{\pi}_{h+1}, \hat{\pi}_{h+2}, \dots \hat{\pi}_H)$, and recall δ_a denotes the distribution that puts probability 1 on action a. Computing $\hat{Q}_h$ and $\hat{\pi}_h$ involves sampling $|S||A|N$ trajectories, which are each of length $H-h$, which results in a sample complexity of $|S||A|N(H-h)$ individual transitions from the MDP.

The dependence on $|S||A|$ in the sample complexity for both of the classical algorithms described above is due to using empirical samples to evaluate $\hat{Q}_h$ for every state action pair $(s,a) \in S \times A$. The assumption that Q_h^* is at most rank d imposes constraints on the relationship between $Q_h^*(s,a)$ at different state-action pairs, such that by approximating Q_h^* using empirical samples at only $O(d|S| + d|A|)$ locations, we should intuitively be able to use the low rank constraint to predict the remaining entries. Let $\Omega_h \subset S \times A$ denote the subset of entries (s,a) for which we use empirical samples to approximate $\hat{Q}_h(s,a)$, computed via either (6) or (8). Given estimates of $\hat{Q}_h(s,a)$ at $(s,a) \in \Omega_h$, we can then use a low-rank matrix estimation subroutine to estimate the Q function for $(s,a) \notin \Omega$. This is the core concept of our algorithm, which we then combine with the two classical approaches of empirical value iteration and Monte Carlo policy iteration.

5.1 Matrix Estimation Subroutine

A critical piece to specify for the algorithm is how to choose the subset Ω_h, and what matrix estimation subroutine to use to predict the full Q_h function given $\hat{Q}_h(s,a)$ for $(s,a) \in \Omega_h$. The performance of any matrix estimation algorithm will depend both on the selected subset Ω_h, as well as the entrywise noise distribution on $\hat{Q}_h(s,a)$ relative to the "ground truth" low-rank matrix. As a result, the subset Ω_h should be determined jointly with the choice of matrix estimation algorithm.

A limitation of a majority of results in the classical matrix estimation literature is that they do not admit entrywise bounds on the estimation error, and the analyses may be sensitive to the distribution of the observation error. Many standard analyses of RL algorithms rely upon the construction of entrywise confidence sets for the estimates of the Q function. To address these limitations, [28] proposed a new matrix estimation algorithm that provides a simple explicit formula for its estimates along with accompanying entrywise error bounds. Their algorithm uses a specific sampling pattern, in which Ω_h is constructed according to $\Omega_h = (S^\# \times A) \cup (S \times A^\#)$, where $S^\#$ and $A^\#$ denote a set of so-called anchor states and actions. Given estimates $\hat{Q}_h(s,a)$ for all $(s,a) \in \Omega_h$, their algorithm estimates the Q function at all state action pairs according to

$$\bar{Q}_h(s,a) = \hat{Q}_h(s, A^\#) \left[\hat{Q}_h(S^\#, A^\#)\right]^\dagger \hat{Q}_h(S^\#, a), \tag{9}$$

where $M^\dagger$ denotes the pseudoinverse of M, and $\bar{Q}$ is the output of the matrix estimation algorithm.

We will utilize their matrix estimation algorithm for our results. We additionally show that under incoherence conditions on Q_h, randomly selecting the set of anchor states $S^\#$ and anchor actions $A^\#$ will be sufficient, where the size of the set of anchor states and actions scales as $\Theta(d \log |S|)$ and $\Theta(d \log |A|)$ respectively, such that the size of Ω_h scales as $\Theta(d|S| \log |A| + d|A| \log |S|)$. This reduces the sample complexity from scaling with $\Omega(|S||A|)$ in the general tabular setting to only scaling with $\tilde{\Omega}(|S| + |A|)$ when the rank d is constant. As a rank d matrix has $d(|S| + |A|)$ degrees of freedom, it follows that this is optimal up to logarithmic factors with respect to $|S|$ and $|A|$.

5.2 Formal Algorithm Statement

We present two algorithms that follow the approach of using matrix estimation to reduce the number samples in either empirical value iteration or Monte Carlo policy iteration. Both algorithms approximate the optimal value function and associated optimal policy via a backward induction, and thus we present them together. We use "Low Rank Empirical Value Iteration" (LR-EVI) to refer to the algorithm which uses option (a) for Step 1 below, and we use "Low Rank Monte Carlo Policy Iteration" (LR-MCPI) to refer to the algorithm which uses option (b) for Step 1.

Hyperparameters: p_1, p_2, and $\{N_h\}_{h \in [H]}$

Initialize: Set $\hat{V}_{H+1}(s) = 0$ for all s, and let $\hat{\pi}^{H+1}$ be any arbitrary policy.

For each $h \in \{H, H-1, H-2, \dots 1\}$ in descending order,

- **Step 0:** Construct the set of anchor states $S_h^\#$ and anchor actions $A_h^\#$ by random sampling, including each state in $S_h^\#$ independently with probability p_1, and including each action in $A_h^\#$ independently with probability p_2. Let $\Omega_h = (S_h^\# \times A) \cup (S \times A_h^\#)$.
- **Step 1:** For each $(s,a) \in \Omega_h$, compute $\hat{Q}_h(s,a)$ using empirical estimates according to either (a) empirical value iteration or (b) Monte Carlo policy evaluation.
  (a) **Empirical Value Iteration:** Collect N_h samples of a single transition starting from state s and action a at step h. Use the samples to estimate $\hat{Q}_h(s,a)$ according to
  $$\hat{Q}_h(s,a) = \hat{r}_h(s,a) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot|s,a)}[\hat{V}_{h+1}(s')],$$
  where $\hat{r}_h(s,a)$ denotes the empirical average reward of the N_h samples from $R_h(s,a)$, and $\hat{P}_h(\cdot|s,a)$ denotes the empirical distribution over the N_h states sampled from $P_h(\cdot|s,a)$.
  (b) **Monte Carlo Policy Evaluation:** Collect N_h independent full trajectories starting from state s and action a at step h until the end of the horizon H, where actions are chosen according to the estimated policy $(\hat{\pi}_{h+1}, \hat{\pi}_{h+2}, \dots \hat{\pi}_H)$. Let $\hat{Q}_h(s,a) = \hat{r}_h^{\text{cum}}(s,a)$,

where $\hat{r}_h^{\text{cum}}(s,a)$ denotes the empirical average cumulative reward across the N_h trajectories starting from (s,a) at step h. If $(R_h^i, R_{h+1}^i, \dots R_H^i)$ denotes the sequence of rewards along the i-th sampled trajectory from (s,a) at step h, then

$$\hat{Q}_h(s,a) = \hat{r}_h^{\text{cum}}(s,a) := \frac{1}{N_h} \sum_{i=1}^{N_h} \sum_{t=h}^{H} R_t^i.$$

- **Step 2:** Predict the action-value function for all $(s,a) \in S \times A$ according to

$$\bar{Q}_h(s,a) = \hat{Q}_h(s, A^{\#}) \left[\hat{Q}_h(S^{\#}, A^{\#})\right]^{\dagger} \hat{Q}_h(S^{\#}, a).$$

- **Step 3:** Compute the estimates of the value function and the optimal policy according to

$$\hat{V}_h(s) = \max_{a \in A} \bar{Q}_h(s,a) \quad \text{and} \quad \hat{\pi}_h(s) = \delta_{\operatorname{argmax} \bar{Q}_h(s,a)}.$$

The tabular MDP variant of the algorithm proposed in [28] is equivalent to LR-EVI where anchor states and actions are given. Furthermore, LR-EVI is equivalent to a modification of the algorithm in [42] with a different choice for the matrix estimation algorithm used in Step 2 and the corresponding sample set Ω_h constructed in Step 1.

6 Main Results

In this section, we present our main theorems that establish sample complexity guarantees for Low Rank Monte Carlo Policy Iteration and Low Rank Empirical Value Iteration under different low-rank assumptions, beginning from the weakest assumption to the strongest. Under low rank Q^* (Assumption 1) and constant suboptimality gap (Assumption 2), Theorem 3 states that Low Rank Monte Carlo Policy Iteration finds an optimal policy with $\tilde{O}((|S|+|A|)H^4/\Delta_{\min}^2)$ samples.

Theorem 3. *Assume that Q_h^* is rank d, has suboptimality gap $\Delta_{\min}$, and is μ-incoherent with condition number bounded by κ (Assumptions 1, 2, 5, and 6). Let $p_1 = O(d\mu \log(|S|))$, $p_2 = O(d\mu \log(|A|))$, and $N_h = \tilde{O}\left((H-h)^2 |S_h^{\#}|^2 |A_h^{\#}|^2 / \Delta_{\min}^2\right)$. Low Rank Monte Carlo Policy Iteration returns an optimal policy with probability at least $1 - \delta - 6H(|S| \wedge |A|)^{-10}$ with a sample complexity of $\tilde{O}\left(d^5 \mu^5 \kappa^4 (|S|+|A|) H^4 / \Delta_{\min}^2\right)$.*

In the tabular setting, there always exists a $\Delta_{\min} > 0$. Thus, as long as Q_h^* is low rank, incoherent, and well conditioned, LR-MCPI learns an optimal policy in $\tilde{O}\left((|S|+|A|)/\Delta_{\min}^2\right)$ samples with high probability. This sample complexity improves upon $|S||A|$ when $\Delta_{\min}$ is greater than $|S|^{-1/2} \wedge |A|^{-1/2}$. When $\Delta_{\min}$ is small, if stronger low-rank assumptions also hold, then the results in Theorems 4 and 5 may provide stronger bounds. Under the assumption that the action-value function is low rank for all ϵ-optimal policies, Theorem 4 states that Low Rank Monte Carlo Policy Iteration learns an ϵ-optimal policy with $\tilde{O}((|S|+|A|)H^6/\epsilon^2)$ samples, independent from $\Delta_{\min}$.

Theorem 4. *For some $\epsilon > 0$, assume that for all ϵ-optimal policies π, Q_h^π is low rank, μ-incoherent, and has condition number bounded by κ (Assumptions 3, 5 and 6). Let $p_1 = O(d\mu \log(|S|)), p_2 = O(d\mu \log(|A|))$, and $N_h = \tilde{O}\left((H-h)^2 |S_h^{\#}|^2 |A_h^{\#}|^2 H^2/\epsilon^2\right)$. Low Rank Monte Carlo Policy Iteration returns an ϵ-optimal policy with probability at least $1-\delta-6H(|S| \wedge |A|)^{-10}$ with a sample complexity of $\tilde{O}\left(d^5 \mu^5 \kappa^4 (|S|+|A|) H^6/\epsilon^2\right)$.*

The strongest assumption that the transition kernel has low Tucker rank and the reward function is low rank, implies that Q^π associated to all policies π is low rank. As such, the result in Theorem 4 also implies an efficient sample complexity guarantee for LR-MCPI under Assumption 4. Additionally, when learning ϵ-optimal Q_h functions, we can remove a factor of H by using Low Rank Empirical Value Iteration (LR-EVI) instead. Empirical Value Iteration as stated in Step 1(a) reduces the sample complexity by a factor of H since it does not require sampling a full rollout of the policy to the end of the horizon, as required for the Monte Carlo estimates presented in Step 1(b). Theorem 5 states that LR-EVI learns ϵ-optimal action-value functions with $\tilde{O}((|S| + |A|)H^5/\epsilon^2)$ samples.

Theorem 5. *Assume that for any ϵ-optimal value function $\hat{V}_{h+1}$, the matrix corresponding to $[r_h + [P_h\hat{V}_{h+1}]]$ is rank d, μ-coherent, and has condition number bounded by κ (Assumptions 4, 5 and 6). Let $p_1 = O(d\mu\log(|S|)), p_2 = O(d\mu\log(|A|))$, and $N_h = \tilde{O}\left((H-h)^2|S_h^\#|^2|A_h^\#|^2H^2/\epsilon^2\right)$. For $\epsilon > 0$, Low Rank Empirical Value Iteration returns ϵ-optimal Q function estimates $\bar{Q}_h$ for all $h \in [H]$ with probability at least $1-\delta-6H(|S|\wedge|A|)^{-10}$ with a sample complexity of $\tilde{O}\left(d^5\mu^5\kappa^4(|S|+|A|)H^5/\epsilon^2\right)$.*

From Proposition 2, under Assumption 4 (low-rank transition kernel and expected rewards), the matrix corresponding to $[r_h + [P_h\hat{V}_{h+1}]]$ has rank d for any value function estimate $\hat{V}_{h+1}$. This is critical to the analysis of LR-EVI as it guarantees that the expectation of the matrix $\bar{Q}_h$ constructed from Empirical Value Iteration in Step 1(a) is low rank. This property is not satisfied by Assumptions 3 and 1, and as such the analysis for Theorem 5 does not extend to these weaker settings. Additionally, this property eliminates the need for constructing estimates with rollouts, which removes a factor of H in the sample complexity compared to LR-MCPI under Assumption 3. While a naive construction of an ϵ-optimal policy from the near-optimal Q_h obtained with LR-EVI will increase the sample complexity by a factor of H^2, incorporating the Monotonicity Technique from [29] into our algorithm may eliminate this increase in sample complexity.

In all three settings, Theorems 3, 4, and 5 show that for carefully chosen p_1, p_2, and N_h, LR-MCPI or LR-EVI learn near-optimal polices or action-value functions, respectively, in a sample efficient manner. Specifically, in both algorithms, under the assumption that the matrix estimation regularity conditions hold, incorporating a matrix estimation subroutine decreases the sample complexity's dependence on S and A from $|S||A|$, as seen in [29], to $|S| + |A|$, which is minimax optimal with respect to $|S|$ and $|A|$.

7 Proof Sketch

In this section, we outline the key ideas needed for Theorems 3, 4, and 5.

Error Amplification from Matrix Estimation. A critical part of the analysis is to understand how the matrix estimation method amplifies the error on the observation set Ω_h. Recall that Ω_h is constructed according to $\Omega_h = (S^\# \times A) \cup (S \times A^\#)$, where $S^\#$ and $A^\#$ denote a set of anchor states and actions. Proposition 8 ensures that if the submatrix $Q_h(S_h^\#, A_h^\#)$ has rank d and $\max_{(s,a)\in\Omega_h} |\hat{Q}_h(s,a) - Q_h^*(s,a)| \leq \eta$, then the ℓ_∞ error of the matrix estimation subroutine over the whole state-action space is bounded by

$$\max_{(s,a)\in S\times A} |\bar{Q}_h(s,a) - Q_h(s,a)| \in O\left(\left(\|Q_h\|_\infty/\sigma_d(Q_h(S_h^\#, A_h^\#)\right)^2 |S_h^\#||A_h^\#|\eta\right).$$

We additionally prove that under the assumptions that Q_h is μ-incoherent, when $S_h^\#$ and $A_h^\#$ are randomly sampled, the rank of $Q_h(S_h^\#, A_h^\#)$ will be equal to the rank of Q_h with high probability. Furthermore, if the condition number of Q_h is bounded by a constant, we prove that $\|Q_h\|_\infty/\sigma_d(Q_h(S^\#, A^\#)) = O(1)$. Thus, the amplification of the error due to the matrix estimation step is $\tilde{O}(|S_h^\#||A_h^\#|\eta)$. This results relaxes the assumption, introduced in [28], that the algorithm has access to d special states $\mathcal{S}$ and actions $\mathcal{A}$ such that $Q_h(\mathcal{S}, \mathcal{A})$ has rank d.

Provided that the error of our initial estimates on the observation set Ω_h are small enough, the above error guarantee eliminates the need to sample transitions from every state-action pair at each time step, as $|\Omega_h| = O(d(|S|\log(|A|) + |A|\log(|S|))$. For Theorem 3, we will apply this analysis to bound the error of $\bar{Q}_h - Q_h^*$. For Theorem 4, we will apply this analysis to bound the error of $\bar{Q}_h - Q_h^\pi$ for some ϵ-optimal policy π. For Theorem 5, we will apply this analysis to bound the error of $\bar{Q}_h - (r_h + [P_h\hat{V}_{h+1}])$ for some value function estimate $\hat{V}_{h+1}$. See Appendix B for formal details regarding our guarantees for the matrix estimation method.

Identifying ϵ-optimal $\hat{\pi}_{H-t}$. As the proof of the theorems in all three settings are very similar, we will provide intution by discussing the key step in the proof for Theorem 4. To prove the desired sample complexity bounds, we show that $\hat{\pi}_{H-t}$ is $(t+1)\epsilon/H$-optimal assuming that $\hat{\pi}_{H-t+1}$ is $t\epsilon/H$-optimal by recursing backwards through the horizon, i.e., $t \in \{0, \dots, H-1\}$.

A major issue in [28] is their estimates of Q^* do not have zero-mean noise, which results in the error scaling exponentially with H. Under Assumption 3, our estimates $\hat{Q}_{H-t}$ are unbiased estimates of $Q_{H-t}^{\hat{\pi}}$, an $(t+1)\epsilon/H$-optimal Q_{H-t} function. Hence, we choose N_{H-t} large enough to ensure $\max_{(s,a)\in\Omega_{H-t}} |\hat{Q}_{H-t}(s,a) - Q_{H-t}^{\hat{\pi}}(s,a)|$ is small enough to offset the amplification in error from the matrix estimation method. Specifically, for our choice of N_{H-t}, we ensure that $\max_{(s,a)\in S\times A} |\bar{Q}_{H-t}(s,a) - Q_{H-t}^{\hat{\pi}}(s,a)| \leq \epsilon/(2H)$. Using the fact that $\hat{\pi}_{H-t+1}$ is $t\epsilon/H$-optimal and the matrix estimation guarantee, we show choosing an action greedily with respect to $\bar{Q}_{H-t}$ ensures that $\|V_{H-t}^* - V_{H-t}^{\hat{\pi}}\|_\infty \leq (t+1)\epsilon/H$.

The proof of Theorem 3 relies on a similar key step, except one identifies an optimal action at each time step $H-t$. Hence, in the suboptimality gap setting, we do not require the error over Ω_{H-t} to depend on H, which decreases N_h by H^2. The proof of Theorem 5 hinges on a similar step, but under Assumption 13, we additionally show that at each time step $r_{H-t} + [P_{H-t}\hat{V}_{H-t+1}]$, the expected value of $\hat{Q}_{H-t}$ for LR-EVI, is close to Q_{H-t}^*.

Finally, for our choice of N_h, the sample complexity is

$$\tilde{O}\left(\frac{(|S||S^\#|^2|A^\#|^3 + |A||S^\#|^3|A^\#|^2)\text{poly}(H)}{\epsilon^2}\right).$$

Under the assumption of a suboptimality gap, ϵ is replaced with $\Delta_{\min}$. Using Bernstein's inequality ensures that $|S_h^\#|, |A_h^\#| \in \tilde{O}(\mu d)$ with high probability. A union bound asserts that the algorithms are correct with their respective sample complexity with high probability.

8 Conclusion

In this work, we provide answers to open questions proposed in [28]; we prove novel sample complexity bounds using matrix estimation methods for MDPs with long time horizons. Building upon our results, there are potential future directions in low-rank reinforcement learning. We have not optimized the algorithms' sample complexity with respect to d and H and do not employ any of the techniques in existing tabular reinforcement learning literature used to decrease the dependence on the time horizon. We also believe relaxing the generative model assumption is worthwhile as having access to a simulator is one of the stronger assumptions in reinforcement learning due to the difficulties of exploration. In sum, we believe this work has shown the benefit of low-rank reinforcement learning setting and hope that many will continue to study this perspective.

References

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

[2] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20095–20107. Curran Associates, Inc., 2020.

[3] Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 67–83. PMLR, 09–12 Jul 2020.

[4] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1707–1714, Madison, WI, USA, 2012. Omnipress.

[5] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[6] Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

[7] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30:3098–3121, 01 2020.

[8] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[9] Mark A. Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[10] Arthur Delarue, Ross Anderson, and Christian Tjandraatmadja. Reinforcement learning with combinatorial actions: An application to vehicle routing. In *NeurIPS*, 2020.

[11] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66(11):7274–7301, 2020.

[12] Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In Jacob Abernethy

and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1554–1557. PMLR, 09–12 Jul 2020.

[13] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR, 18–24 Jul 2021.

[14] Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

[15] Yaqi Duan, Tracy Ke, and Mengdi Wang. State aggregation learning from markov transition data. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[16] Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvari, and Mengdi Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. In *ICML*, 2021.

[17] William B. Haskell, Rahul Jain, and Dileep Kalathil. Empirical dynamic programming. *Mathematics of Operations Research*, 41(2):402–429, 2016.

[18] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[19] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017.

[20] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[21] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020.

[22] Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1999.

[23] Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021.

[24] Yuxi Li. Reinforcement learning applications. *arXiv preprint arXiv:1908.06973*, 2019.

[25] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[26] Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *CoRR*, abs/2102.07035, 2021.

[27] Chengzhuo Ni, Anru R Zhang, Yaqi Duan, and Mengdi Wang. Learning good state and action representations via tensor decomposition. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1682–1687. IEEE, 2021.

[28] Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample efficient reinforcement learning via low-rank matrix estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12092–12103. Curran Associates, Inc., 2020.

[29] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[30] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *COLT*, 2019.

[31] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, second edition, 2018.

[32] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, Aug 2011.

[33] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

[34] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, pages 1–5, 2019.

[35] M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[36] Bingyan Wang, Yuling Yan, and Jianqing Fan. Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34, 2021.

[38] Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

[39] Kunhe Yang, Lin F. Yang, and Simon Shaolei Du. Q-learning with logarithmic regret. In *AISTATS*, 2021.

[40] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10746–10756. PMLR, 13–18 Jul 2020.

[41] Lin F. Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *ICML*, 2019.

[42] Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. Harnessing structures for value-based planning and reinforcement learning. In *International Conference on Learning Representations*, 2020.

[43] Anru Zhang and Mengdi Wang. Spectral state compression of markov processes. *IEEE Transactions on Information Theory*, 66:3202–3231, 2020.

A Extended Literature Discussion

In this section, we provide an extended discussion on the related literature.

Tabular Reinforcement Learning. Sample complexity bounds for reinforcement learning algorithms in the tabular MDP setting have been studied extensively, e.g., [39, 3, 39, 8, 23]. [4] show that even with a generative model, $\Omega\left(\frac{|S||A|}{\epsilon^2(1-\gamma)^3}\right)$ samples are required to estimate an ϵ-optimal action-value function. To match the lower bound, [29] utilize their Monotonicity, Variance Reduction, and Total Variance Techniques to prove a matching upper bound (up to logarithmic factors) $\tilde{O}\left(\frac{|S||A|}{\epsilon^2(1-\gamma)^3}\right)$ on the sample complexity of finding an ϵ-optimal action-value function and policy. The sample complexities in all of these works are at least linear in both the size of the state space and action space and polynomial in the effective horizon. Our work focuses on decreasing the sample complexity's dependence on $|S|$ and $|A|$ from $|S||A|$ to $|S| + |A|$ under models with a low-rank structure.

Complexity Measures for General Function Approximation. The search for the most general types of structure that allows for sample-efficient reinforcement learning has resulted in newly defined complexity measures of Bellman rank [19, 12], witness rank [30], Bellman Eluder dimension [20], and Bilinear Class [13], which impose low rankness on the Bellman error, or any estimator of it. The Bilinear Class framework contains both the Bellman and witness frameworks while low Bellman error implies low Bellman Eluder dimension. While similar neither the Bilinear Class or Bellman Eluder dimension frameworks capture each other. For the class of MDPs with the Bilinear Class having low rank d or low Bellman Eluder dimension d, the sample complexity of finding an ϵ-optimal policy reduces to $\tilde{O}\left(\frac{\text{poly}(d,H)}{\epsilon^2}\right)$. Unfortunately, these complexity measures are so broad that the resulting algorithms that achieve sample efficiency are not polynomial time computable, and rely on strong optimization oracles in general, e.g. assuming that we can solve a high dimensional non-convex optimization problem. We remark that our setting, with the assumption that the transition kernels and reward functions are low-rank, has low Bellman rank.

Linear Function Approximation. To combat the curse of dimensionality, there has been significant research in reinforcement learning consider the simplest function class of linear function approximation. The weakest assumption only imposes that the action-value function associated to the optimal policy, denoted Q^*, is linearly-realizable with respect to a known low dimensional feature representation, given by a known feature extractor $\phi : S \times A \to \mathbb{R}^d$ for $d \ll |S|, |A|$. However, there have been a sequence of works that imply that only imposing linear with respect to Q^* may be too weak to admit sample efficient algorithms. In particular, [37] provide an example of an MDP with a linearly-realizable Q^* and constant suboptimality gap, which still requires an exponential number of samples in the minimum of the dimension d or the time horizon H. [38] show a similar result holds even when the algorithm has access to a generative model. These lower bounds highlight that the dynamics due to the MDP introduce additional complexities for linear function approximation in RL settings.

Since lower bounds show that the linearly-realizable Q^* setting is difficult, there has been a subsequent literature on the linear MDP model, which additionally imposes that the transition kernels and reward functions are linear with respect to a known low dimensional feature extractor ϕ [21, 41, 40, 36, 16]. Since the feature extractor is known, one only needs to estimate the $d|S|$ dimensional coefficients of the linear transition kernel, so the sample complexities or regret guarantees of these works are polynomial in d with no dependence on $|S|$ or $|A|$. In practice, the feature extractor is often not known, and thus another algorithms such as deep learning are employed to learn the representation in advance, with no theoretical guarantees. Furthermore, [14] showed that if the learned feature extractor is not exact, that is the entrywise error is $\Omega\left(\sqrt{\frac{H}{d}}\right)$, then an exponential number of samples is required to find a near-optimal policy in both the linear Q^* and linear MDP settings. Du et al. remark that if the ℓ_1 error of the approximated transition kernel is bounded, then there are polynomial sample complexity algorithms given by [41, 40, 21].

The limitation of the linear function approximation approach is that it requires exact or approximate knowledge of the unknown feature mapping ϕ. While it may be reasonable to have access to features of the states and actions, it is a strong requirement that the model behaves linear with respect to the given features, entirely bypassing the effort of feature engineering, which itself can be costly and non-trivial. Under only low rank structural assumptions, one only imposes that there exists some low dimensional representation under which the model exhibits linearity, but this representation is latent or unknown. As a result the degrees of freedom of a low-rank model is significantly larger than a linear model, as the unknown parameters also must specify the latent feature representations of the states and actions. The implications towards algorithm design and statistical complexity critically depend on where one is imposing the low rank structure. As stated in Section 2, examples of class of works that impose low-rank structure without knowledge of the features are low-rank MDPs, MDPs with transition kernels having low Tucker rank, and Q^* having low rank.

Matrix Estimation. Low-rank matrix estimation methods focus on recovering the missing entries of a partially observed low-rank matrix with noise. The field has been studied extensively with provable recovery guarantees; see the surveys [6, 9]. However, the majority of recovery guarantees of matrix estimation are in the Frobenius norm instead of an entry-wise/ℓ_∞ error bound, whereas a majority of common analyses for reinforcement learning algorithms rely upon constructing entrywise confidence sets for the estimated values. Matrix estimation methods with entry-wise error bounds are given in [7, 11, 1], but all require strict distributional assumptions on the noise, e.g., mean-zero sub-Gaussian/Gaussian error. In the infinite-horizon discounted MDP setting with bounded rewards, estimates of the action-value function will have bounded error. Thus, we can guarantee that the noise is sub-Gaussian [35]. While our estimates have sub-Gaussian noise, we cannot ensure that our estimates have zero mean using existing matrix estimation methods; to the best of our knowledge, the existing matrix estimation methods with entry-wise error bounds have no guarantees on the mean of their estimators. In contrast, the matrix estimation method proposed in [28] provides entry-wise error guarantees in this harsh setting and is the method we use in our algorithms. With additional assumptions, the matrix comprised of the estimates of our algorithm is low-rank. Hence, our input to the matrix estimation method has mean-zero and bounded noise and enables us to relax the small discount factor requirement.

B Matrix Estimation Method

In this section, we provide the formal details of the matrix estimation method and its guarantees. In contrast to assuming knowledge of anchor states and actions, we show that uniform sampling of $O(d\log(|S|))$ states and $O(d\log(|A|))$ actions gives us anchor states and actions with high probability assuming that Q_h is incoherent, Assumption 5. Our sampling method is as follows: we sample states and actions using the Bernoulli model. Let $\tilde{U} \in \mathbb{R}^{|S| \times d}, \tilde{V} \in \mathbb{R}^{|A| \times d}$ such that

$$\tilde{U}_i = \begin{cases} U_i \text{ with probability } p_1, \\ 0 \text{ otherwise} \end{cases}, \qquad \tilde{V}_i = \begin{cases} V_i \text{ with probability } p_2, \\ 0 \text{ otherwise} \end{cases}$$

Let $\tilde{Q}_h := \tilde{U} \Sigma \tilde{V}^\top \in \mathbb{R}^{|S| \times |A|}$, so $\tilde{Q}_h$ contains a $p_1 |S| \times p_2 |A|$ submatrix in expectation and is zero elsewhere. The sampled anchor states and actions are the states corresponding to the non-zero rows and columns, respectively. We remark that the Bernoulli model is chosen for convenience and similar results hold if we sample with replacement. We now present our lemma that shows $p_1^{-1/2} \tilde{U}$ and $p_2^{-1/2} \tilde{V}$ have near orthonormal columns, which implies that $\tilde{U}$ and $\tilde{V}$ have full column rank, with high probability.

Lemma 6. *Let $Q_h, U, \tilde{U}, \Sigma, V,$ and $\tilde{V}$ be defined as above. Let Q_h be μ-incoherent. Then, with probability at least $1 - 4(|S| \wedge |A|)^{-10}$, we have*

$$\|p_1^{-1} \tilde{U}^\top \tilde{U} - I_{d \times d}\|_{op} \leq \sqrt{\frac{40 \mu d \log(|S|)}{p_1 |S|}} + \frac{40 \mu d \log(|S|)}{p_1 |S|}$$
$$\|p_2^{-1} \tilde{V}^\top \tilde{V} - I_{d \times d}\|_{op}, \leq \sqrt{\frac{40 \mu d \log(|A|)}{p_2 |A|}} + \frac{40 \mu d \log(|A|)}{p_2 |A|}.$$

Proof of Lemma 6. For each $i \in [|S|]$, let $Z^{(i)} \in \mathbb{R}^{|S| \times d}$ be the matrix obtained from U by zeroing out all but the i-th row. Let $\delta_1, \ldots, \delta_{|S|}$ be i.i.d. Bernoulli(p_1) random variables. We can express

$$U = \sum_{i \in [|S|]} Z^{(i)} \text{ and } \tilde{U} = \sum_{i \in [|S|]} \delta_i Z^{(i)}$$

Note that

$$\tilde{U}^\top \tilde{U} = \sum_{i \in [|S|]} \sum_{j \in [|S|]} \delta_i \delta_j Z^{(i)\top} Z^{(j)} \tag{10}$$
$$= \sum_{i \in [|S|]} \delta_i^2 Z^{(i)\top} Z^{(i)} \tag{11}$$

by construction of $Z^{(i)}$ and $Z^{(j)}$. Hence,

$$\begin{aligned} \mathbb{E}[\tilde{U}^\top \tilde{U}] &= p_1 \sum_{i \in [|S|]} Z^{(i)^\top} Z^{(i)} \\ &= p_1 \sum_{i \in [|S|]} \sum_{j \in [|S|]} Z^{(i)^\top} Z^{(j)} \\ &= p_1 U^\top U \\ &= p_1 I_{d \times d} \end{aligned} \tag{12}$$

where the last equality is due to U having orthonormal columns. For each $i \in [|S|]$, we define the following the mean-zero matrices

$$X^{(i)} := (\delta_i^2 - \mathbb{E}[\delta_i^2]) Z^{(i)\top} Z^{(i)} = (\delta_i - p_1) Z^{(i)\top} Z^{(i)}.$$

Since Q_h^* is $\mu-$incoherent,

$$\|X^{(i)}\|_{op} \leq |\delta_i - p_1| \|Z^{(i)^\top} Z^{(i)}\|_{op} \leq \|Z^{(i)^\top} Z^{(i)}\|_{op} = \|U_{i-}\|_2^2 \leq \frac{\mu r d}{|S|} \quad \text{surely.}$$

Furthermore,

$$\begin{aligned}
\sum_{i\in[|S|]} \mathbb{E}[X^{(i)^\top} X^{(i)}] = \sum_{i\in[|S|]} \mathbb{E}[X^{(i)} X^{(i)^\top}] &= \sum_{i\in[|S|]} \mathbb{E}[(\delta_i - p)^2] Z^{(i)^\top} Z^{(i)} Z^{(i)^\top} Z^{(i)} \\
&= p_1(1-p_1) \sum_{i\in[|S|]} \|U_{i-}\|_2^2 Z^{(i)^\top} Z^{(i)} \\
&\preceq p_1 \cdot \frac{d\mu}{|S|} \sum_{i\in[|S|]} Z^{(i)^\top} Z^{(i)} \\
&= \frac{d\mu p_1}{|S|} U^T U \\
&= \frac{d\mu p_1}{|S|} I_{d\times d}.
\end{aligned}$$

Thus,

$$\| \sum_{i\in[|S|]} \mathbb{E}[X^{(i)^\top} X^{(i)}] \|_{op} = \| \sum_{i\in[|S|]} \mathbb{E}[X^{(i)} X^{(i)^\top}] \|_{op} \leq \frac{d\mu p_1}{|S|}$$

From the matrix Bernstein inequality (Theorem 16), we have

$$\begin{aligned}
\mathbb{P}\left(\|\tilde{U}^\top \tilde{U} - p_1 I_{d\times d}\|_{op} \geq t \right) &= \mathbb{P}\left(\left\| \sum_{i\in[|S|]} \left((\delta_i^2 - p_1) Z^{(i)^\top} Z^{(i)} \right) \right\|_{op} \geq t \right) \\
&= \mathbb{P}\left(\left\| \sum_{i\in[|S|]} X^{(i)} \right\|_{op} \geq t \right) \\
&\leq 2|S| \exp\left(- \frac{t^2/2}{\frac{p_1 \mu d}{|S|} + \frac{\mu d}{3|S|} t} \right) \\
&\leq 2|S| \exp\left(- \frac{t^2}{\frac{2p_1 \mu d}{|S|} + \frac{2\mu d}{|S|} t} \right)
\end{aligned}$$

where the first equality follows from equations 11 and 12. For $t = \sqrt{\frac{40 p_1 \mu d \log(|S|)}{|S|}} + \frac{40 \mu d \log(|S|)}{|S|}$, we have

$$\left\| \tilde{U}^\top \tilde{U} - p_1 I_{d\times d} \right\|_{op} \leq \sqrt{\frac{40 p_1 \mu d \log(|S|)}{|S|}} + \frac{40 \mu d \log(|S|)}{|S|}$$

with probability at least $1 - 2|S|^{-10}$. Dividing both sides by p_1 yields the first inequality in the lemma. The corresponding bound for $\tilde{V}$ holds from a similar argument. Taking a union bound over the two events proves the lemma. □

Now, we present the theorem that states this uniformly sampled submatrix ($\tilde{O}(d)$ by $\tilde{O}(d)$ in expectation) has rank-d with its smallest non-zero singular value bounded away from zero.

Proposition 7. *Let $p_1 = \frac{\mu d \log(|S|)}{320|S|}$ and $p_2 = \frac{\mu d \log(|A|)}{320|A|}$. Under the event in Lemma 6, we have*

$$\sigma_d((p_1 \vee p_2)^{-1}\tilde{Q}) \geq \frac{1}{2}\sigma_d(Q_h).$$

Proof Of Proposition 7. Under the assumption that $p_1 = \frac{\mu d \log(|S|)}{320|S|}$ and $p_2 = \frac{\mu d \log(|A|)}{320|A|}$ and the event in Lemma 6, we have $\|p_1^{-1}\tilde{U}^\top \tilde{U} - I_{d\times d}\|_{op} \leq \frac{1}{2}$. From Weyl's inequality, we have $\sigma_d(p_1^{-1}\tilde{U}^\top \tilde{U}) \geq \frac{1}{2}$, which implies $\sigma_d(p_1^{-1/2}\tilde{U}) \geq \frac{1}{\sqrt{2}}$. From a similar argument, $\sigma_d(p_1^{-1/2}\tilde{V}) \geq \frac{1}{\sqrt{2}}$. Let $p = p_1 \vee p_2$, from the singular value version of the Courant-Fischer minimax theorem (Theorem 7.3.8 [18]), we have

$$\begin{aligned}
\sigma_d(p^{-1}\tilde{Q}) &= \max_{S:dim(S)=d} \min_{x\in S, x\neq 0} \frac{\|p^{-1}\tilde{U}\Sigma\tilde{V}^\top x\|_2}{\|x\|_2} \\
&= \max_{S:dim(S)=d} \min_{x\in S, x\neq 0} \frac{\|(p^{-1/2}\tilde{U})\Sigma(p^{-1/2}\tilde{V}^\top)x\|_2}{\|\Sigma(p^{-1/2}\tilde{V}^\top)x\|_2} \frac{\|\Sigma(p^{-1/2}\tilde{V}^\top)x\|_2}{\|p^{-1/2}\tilde{V}^\top x\|_2} \frac{\|p^{-1/2}\tilde{V}^\top x\|_2}{\|x\|_2} \\
&\geq \max_{S:dim(S)=d} \min_{x\in S, x\neq 0} \frac{\|(p^{-1/2}\tilde{U})\Sigma(p^{-1/2}\tilde{V}^\top)x\|_2}{\|(p^{-1/2}\tilde{U})^\dagger\|_{op}\|(p^{-1/2}\tilde{U})\Sigma(p^{-1/2}\tilde{V}^\top)x\|_2} \\
&\qquad \cdot \frac{\|\Sigma(p^{-1/2}\tilde{V}^\top)x\|_2}{\|\Sigma^{-1}\|_{op}\|\Sigma(p^{-1/2}\tilde{V}^\top)x\|_2} \frac{\|p^{-1/2}\tilde{V}^\top x\|_2}{\|x\|_2} \\
&= \sigma_d(p^{-1/2}\tilde{U}) \cdot \sigma_d(\Sigma) \max_{S:dim(S)=d} \min_{x\in S, x\neq 0} \frac{\|p^{-1/2}\tilde{V}^\top x\|_2}{\|x\|_2} \\
&= \sigma_d(p^{-1/2}\tilde{U}) \cdot \sigma_d(\Sigma)\sigma_d(p^{-1/2}\tilde{V}^\top) \\
&\geq \sigma_d(p_1^{-1/2}\tilde{U}) \cdot \sigma_d(\Sigma)\sigma_d(p_2^{-1/2}\tilde{V}^\top) \\
&\geq \frac{1}{\sqrt{2}}\sigma_d(Q_h)\frac{1}{\sqrt{2}} \\
&= \frac{1}{2}\sigma_d(Q_h)
\end{aligned}$$

where the first inequality comes from properties of the operator norm and inverses/pseudo-inverses and the second inequality comes from replacing $p = p_1 \vee p_2$ with either p_1 or p_2. □

We next present the matrix estimation method and guarantee used in [28] with our randomly sampling anchor states and actions and our bound on the smallest non-zero singular value.

Matrix Estimation Method. For each time step $h \in [H]$, we first uniformly sample anchor states $S^\#$ with probability p_1 and actions $A^\#$ with probability p_2. Specifically, for all $s \in S$ and $a \in A$, $s \in S^\#$ and $a \in A^\#$ with probability p_1 and p_2, respectively, for $p_1 = \frac{\mu d \log(|S|)}{320|S|}$ and $p_2 = \frac{\mu d \log(|A|)}{320|A|}$. We then define the observation set as $\Omega_h = \{(s,a) \in S \times A | s \in S^\# \text{ or } a \in A^\#\}$. Given estimates of Q_h over the observation set, $\hat{Q}_h(s,a)$ for all $(s,a) \in \Omega_h$, the matrix estimate is

$$\bar{Q}_h(s,a) = \hat{Q}_h(s, A^\#)[\hat{Q}_h(S^\#, A^\#)]^\dagger \hat{Q}_h(S^\#, a) \quad \forall (s,a) \in S \times A, \forall h \in [H]$$

where $[X]^\dagger$ is the pseudo inverse of matrix X. We now present our version of the proposition in citeserl that controls the error amplification of the matrix estimator adapted to the finite-horizon setting assuming that Q_h is incoherent and well conditioned.

Proposition 8 (Proposition 13 [28] with Bounded Condition Number)**.** *Let $S^\#, A^\#, \Omega_h$, and $\bar{Q}_h$ be as defined above. Let Assumption 6 hold on Q_h, that is $\kappa_{Q_h} \leq C$ for some $C > 1$. Under the event in Lemma 6, for any $\eta \leq \frac{1}{2\sqrt{|S_h^\#||A_h^\#|}}\sigma_d(Q_h^*(S^\#, A^\#))$, and if $\max_{(s,a)\in\Omega_h} |\hat{Q}_h(s,a) - Q_h(s,a)| \leq \eta$, then*

$$\max_{(s,a)\in S\times A} |\bar{Q}_h(s,a) - Q_h(s,a)| \leq c_h'|S_h^\#||A_h^\#|\eta$$

for all $h \in [H]$ where $c_h' = \left(6\sqrt{2}\left(\frac{640\kappa_{Q_h}}{\log(|S|\wedge|A|)}\right) + 2(1+\sqrt{5})\left(\frac{640\kappa_{Q_h}}{\log(|S|\wedge|A|)}\right)^2\right)$.

Proof of Proposition 8. Under the event in Lemma 6 and from Proposition 7, we have $rank(Q_h^*(S^\#, A^\#)) = d$. We follow the same argument as the proof of Proposition 13 in [28] except we upperbound equations (22) and (23) with $\|Q_h\|_\infty$ instead of $V_{\max}$. Following the steps in [28], for all $(s,a) \in S \times A$,

$$\begin{aligned}|\bar{Q}_h(s,a) - Q_h(s,a)| \leq \sqrt{2}\left\|[\hat{Q}_h(S^\#, A^\#)]^\dagger\right\|_{op}\left\|\hat{Q}_h(S^\#, a)\hat{Q}_h(s, A^\#) - Q_h(S^\#, a)Q_h(s, A^\#)\right\|_F \\ + \left\|[\hat{Q}_h(S^\#, A^\#)]^\dagger - [Q_h(S^\#, A^\#)]^\dagger\right\|_{op}\left\|Q_h(S^\#, a)Q_h(s, A^\#)\right\|_F.\end{aligned}$$

With the same logic in the proof of Proposition 13,

$$\left\|[\hat{Q}_h(S^\#, A^\#)]^\dagger\right\|_{op} \leq \frac{2}{\sigma_d(Q_h(S^\#, A^\#))}$$

$$\left\|[\hat{Q}_h(S^\#, A^\#)]^\dagger - [Q_h(S^\#, A^\#)]^\dagger\right\|_{op} \leq 2(1+\sqrt{5})\frac{\eta\sqrt{|S^\#||A^\#|}}{\sigma_d(Q_h(S^\#, A^\#))^2}.$$

Since for all $s, s' \in S$ and $a, a'A$,

$$\begin{aligned}\left|\hat{Q}_h(s',a)\hat{Q}_h(s,a') - Q_h(s',a)Q_h(s,a')\right| &\leq \left|(Q_h(s',a)+\eta)(Q_h(s,a')+\eta) - Q_h(s',a)Q_h(s,a')\right| \\ &\leq \eta|Q_h(s',a)| + \eta|Q_h(s,a')| + \eta^2 \\ &\leq 2\eta\|Q_h\|_\infty + \eta^2,\end{aligned}$$

then, $\left\|\hat{Q}_h(S^\#, a)\hat{Q}_h(s, A^\#) - Q_h(S^\#, a)Q_h(s, A^\#)\right\|_F \leq (2\eta\|Q_h\|_\infty + \eta^2)\sqrt{|S^\#||A^\#|}$. Because $|Q_h(s',a)Q_h(s,a')| \leq \|Q_h\|_\infty^2$ for all $s, s' \in S$ and $a, a'A$, clearly $\left\|Q_h(S^\#, a)Q_h(s, A^\#)\right\|_F \leq \|Q_h\|_\infty^2\sqrt{|S^\#||A^\#|}$. Using these inequalities gives that for all $(s,a) \in S \times A$,

$$|\bar{Q}_h(s,a) - Q_h(s,a)| \leq \left(6\sqrt{2}\left(\frac{\|Q_h\|_\infty}{\sigma_d(Q_h(S^\#, A^\#))}\right) + 2(1+\sqrt{5})\left(\frac{\|Q_h\|_\infty}{\sigma_d(Q_h(S^\#, A^\#))}\right)^2\right)|S^\#||A^\#|\eta \tag{13}$$

since $\eta \leq \|Q_h\|_\infty$. Next, we upper bound $\frac{\|Q_h\|_\infty}{\sigma_d(Q_h(S^\#, A^\#))}$. Let the singular value decomposition of the rank d matrix Q_h be $Q_h = U\Sigma V^\top$. For $(s,a) \in S \times A$,

$$\begin{aligned}|Q_h(s,a)| &= |U_s\Sigma V_a| \\ &\leq \|\Sigma\|_{op}|U_sV_a| \\ &\leq \sigma_1(Q_h)\|U_s\|_2|V_a\|_2 \\ &\leq \sigma_1(Q_h)\sqrt{\frac{\mu d}{|S|}}\sqrt{\frac{\mu d}{|A|}} \\ &= \frac{d\sigma_1(Q_h)\mu}{\sqrt{|S||A|}}\end{aligned}$$

where the third inequality comes from Q_h being μ incoherent. Hence,

$$\begin{aligned}\frac{\|Q_h\|_\infty}{\sigma_d(Q_h(S^\#, A^\#))} &\leq \frac{d\sigma_1(Q_h)\mu}{\sigma_d(Q_h(S^\#, A^\#))\sqrt{|S||A|}} \\ &\leq \frac{d\sigma_1(Q_h)\mu}{\sigma_d(Q_h(S^\#, A^\#))(|S| \wedge |A|)} \\ &= \frac{320\sigma_1(Q_h)}{\sigma_d((p_1 \vee p_2)^{-1}Q_h(S^\#, A^\#)) \log(|S| \wedge |A|)} \\ &= \frac{640\sigma_1(Q_h)}{\sigma_d(Q_h) \log(|S| \wedge |A|)} \\ &= \frac{640\kappa_{Q_h}}{\log(|S| \wedge |A|)}\end{aligned}$$

where the third line comes from the definition of p_1 and p_2 and the fourth line comes from Proposition 7. Applying this inequality to Equation 13 proves the proposition. □

C Omitted Proofs

In this section, we present the proofs of several technical results from the main text.

C.1 Proof of Lemma 1

We prove Lemma 1 by explicitly constructing an MDP exhibiting the doubly exponential blow-up.

Recall that $S = A = \{1, 2\}$. We define the reward function as $R_h(s, a) = 0$ for all $(s, a, h) \in S \times A \times [H-1]$ with terminal reward $R_H(s, a) = 1/2$ for all $(s, a) \in S \times A$. The transition kernel is

$$P_h(\cdot|s, a) = \begin{cases} \delta_s, & \text{if } s = a, \\ \text{uniform}(S), & \text{if } s \neq a \end{cases} \qquad \forall h \in [H],$$

where δ_s denotes the Dirac distribution at s. We remark that all policies are optimal with value and action-value functions

$$Q_h^*(s, a) = V_h^*(s) = \frac{1}{2}, \qquad \forall (s, a, h) \in S \times A \times [H].$$

Consider policy evaluation for one of the optimal policies, $\pi = \{\pi_h\}_{h \in [H]}$, where for all $h \in [H]$,

$$\pi_h(s) := \begin{cases} 1, & \text{if } s = 1, \\ 2, & \text{if } s = 2. \end{cases}$$

For $h = H-1, \ldots, 1$, we use the algorithm described in Section 3.1 to compute the estimates $\hat{Q}_h^\pi \in \mathbb{R}^{2\times 2}$ and $\hat{V}_h^\pi \in \mathbb{R}^2$ of the value and action-value functions of π. For this algorithm, we note that a global optimum (u, v) of the least-squares formulation (4) is

$$u_1 = v_1 = \sqrt{\hat{Q}_h^\pi(1, 1)}, \quad u_2 = \frac{\hat{Q}_h^\pi(2, 1)}{\sqrt{\hat{Q}_h^\pi(1, 1)}}, \quad \text{and} \quad v_2 = \frac{\hat{Q}_h^\pi(1, 2)}{\sqrt{\hat{Q}_h^\pi(1, 1)}},$$

which leads to formula for $\hat{Q}_h^\pi(2, 2)$ given in Equation (5).

We state a lemma that shows if our estimate at step $h + 1$ is not perfect, then the error at step h increases quadratically.

Lemma 9. *Let π and the MDP be defined as above. For each $h \in [H-1]$, suppose that the value function estimate of V^{π}_{h+1} is in the form*

$$\hat{V}^{\pi}_{h+1} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} + 2\epsilon_{h+1} \end{bmatrix}.$$

Then the algorithm described in Section 3.1 results in the estimate

$$\hat{V}^{\pi}_{h} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} + 2\epsilon_{h} \end{bmatrix}$$

where $\epsilon_h = \epsilon_{h+1} + \epsilon^2_{h+1}$.

Proof of Lemma 9. We compute the first row and column of $\hat{Q}^{\pi}_h$ using the exact Bellman operator given $\hat{V}^{\pi}_{h+1}$:

$$\hat{Q}^{\pi}_h(1,1) = R_h(1,1) + \sum_{s'=1}^{2} P_h(s'|1,1)\hat{V}^{\pi}_{h+1}(s') = \hat{V}^{\pi}_{h+1}(1) = \frac{1}{2},$$

$$\hat{Q}^{\pi}_h(1,2) = R_h(1,2) + \sum_{s'=1}^{2} P_h(s'|1,2)\hat{V}^{\pi}_{h+1}(s') = \frac{1}{2}(\hat{V}^{\pi}_{h+1}(1) + \hat{V}^{\pi}_{h+1}(2)) = \frac{1}{2} + \epsilon_{h+1},$$

$$\hat{Q}^{\pi}_h(2,1) = R_h(2,1) + \sum_{s'=1}^{2} P_h(s'|2,1)\hat{V}^{\pi}_{h+1}(s') = \frac{1}{2}(\hat{V}^{\pi}_{h+1}(1) + \hat{V}^{\pi}_{h+1}(2)) = \frac{1}{2} + \epsilon_{h+1}.$$

Using the estimator in Equation (5) to compute $\hat{Q}^{\pi}_h(2,2)$:

$$\hat{Q}^{\pi}_h(2,2) = \frac{\hat{Q}^{\pi}_h(1,2)\hat{Q}^{\pi}_h(2,1)}{\hat{Q}^{\pi}_h(1,1)} = \frac{1}{2} + 2(\epsilon_{h+1} + \epsilon^2_{h+1}).$$

Hence, following policy π_h results in the value function estimate:

$$\hat{V}^{\pi}_{h} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} + 2\epsilon_{h} \end{bmatrix}$$

for $\epsilon_h = \epsilon_{h+1} + \epsilon^2_{h+1}$. □

Using Lemma 9, we can show that if the estimate for the terminal value function V^{π}_H is not be perfect, then the estimation error blows up quickly as we recurse backward to compute an estimate of V^{π}_1. In particular, we establish the following lemma, which immediately implies the desired Lemma 1.

Lemma 10 (Doubly Exponential Growth)**.** *Let π and the MDP be defined as above. Suppose that the estimate for the terminal value function is of the form*

$$\hat{V}^{\pi}_{H} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} + 2\epsilon_{H} \end{bmatrix}$$

for some $\epsilon_H > 0$. Then, the algorithm described in Section 3.1 results in the estimate

$$\hat{V}^{\pi}_{1} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} + \Omega\left(\left((1+\epsilon_H)^{c'}\right)^{2^{H - \frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)} - 1 - c'}} \right) \end{bmatrix}$$

for $H > 1 + \frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)} + c'$ and $c' = \left\lceil \frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)} \right\rceil - \frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)}$.

Proof of Lemma 10. Let $H > 1 + \left\lceil \frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)} \right\rceil$, $c' = \left\lceil \frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)} \right\rceil - \frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)}$, and $t = H - 1 - \frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)} - c'$. Recursively applying Lemma 9 gives:

$$\hat{V}_t^\pi = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} + 2\epsilon_t \end{bmatrix}$$

where $\epsilon_h = \epsilon_{h+1}(1+\epsilon_{h+1})$ for $h \in \{t, \dots, H-1\}$. Since $\epsilon_h > \epsilon_{h+1}$, it follows that $\epsilon_h > (1+\epsilon_H)\epsilon_{h+1}$ for $h \in [H]$. Thus, $\epsilon_t > (1+\epsilon_H)^{H-t-1}\epsilon_H$. Plugging in the value for t gives

$$\epsilon_t > (1+\epsilon_H)^{\frac{\ln(1/\epsilon_H)}{\ln(1+\epsilon_H)} + c'} \epsilon_H = (1+\epsilon_H)^{c'}.$$

Recursively applying Lemma 9 again gives:

$$\hat{V}_1^\pi = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} + 2\epsilon_1 \end{bmatrix}$$

where $\epsilon_h = \epsilon_{h+1} + \epsilon_{h+1}^2$ for $h \in [t-1]$. Lower bounding ϵ_1 it terms of $\epsilon_t = (1+\epsilon_H)^{c'}$ gives

$$\hat{V}_1^\pi = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} + \Theta\left(\epsilon_t^{2^t}\right) \end{bmatrix}.$$

Using the lower bound on ϵ_t and plugging in our value for t proves the lemma. □

C.2 Proof of Proposition 2

We next present the proof of Proposition 2, which shows that if the reward function and transition kernel are low rank, then for any value function estimate $\hat{V}_{h+1}$, $r_h + [P_h \hat{V}_{h+1}]$ has rank upperbounded by d.

Let MDP $M = (S, A, P, r, H)$ satisfy Assumption 4 (specifically, P_h has Tucker rank $(|S|, |S|, d)$. It follows that for each $h \in [H]$, there exists an $|S| \times |S| \times d$ tensor $U^{(h)}$, an $|A| \times d$ matrix $V^{(h)}$, and an $|S| \times d$ matrix $W^{(h)}$ such that

$$P_h(s'|s,a) = \sum_{i=1}^d U_{s',s,i}^{(h)} V_{a,i}^{(h)} \quad \text{and} \quad r_h(s,a) = \sum_{i=1}^d W_{s,i}^{(h)} V_{a,i}^{(h)}.$$

Hence, for any value function estimate $\bar{V}_{h+1}$,

$$\begin{aligned} r_h(s,a) + P_h\hat{V} &= \sum_{i=1}^d W_{s,i}^{(h)} V_{a,i}^{(h)} + \sum_{s' \in S} \hat{V}(s') P_h(s'|s,a) \\ &= \sum_{i=1}^d W_{s,i}^{(h)} V_{a,i}^{(h)} + \sum_{s' \in S} \hat{V}(s') \sum_{i=1}^d U_{s',s,i}^{(h)} V_{a,i}^{(h)} \\ &= \sum_{i=1}^d V_{a,i}^{(h)} \left(W_{s,i}^{(h)} + \sum_{s' \in S} \hat{V}(s') U_{s',s,i}^{(h)} \right). \end{aligned}$$

Since $W_{:,:}^{(h)} + \sum_{s' \in S} \hat{V}(s') U_{s',:,:}^{(h)}$ is an $|S| \times d$ matrix, $r_h(s,a) + P_h\hat{V}$ has rank at most d. The same result holds when P_h has Tucker rank $(|S|, d, |A|)$ from a similar argument.

C.3 Supplement to Section 6

We next present the missing proofs of Theorems 3, 4, and 5 and related lemmas.

Lemma 11. *Let* $p_1 = \frac{\mu d \log(|S|)}{320|S|}, p_2 = \frac{\mu d \log(|A|)}{320|A|}$, *and* $N_{H-t} = \frac{2(t+1)^2(c'_{H-t})^2|S^{\#}_{H-t}|^2|A^{\#}_{H-t}|^2\log(2H|S||A|/\delta)}{\Delta^2_{\min}}$, *where* c'_{H-t} *is defined as in Proposition 8. Then the policy* $\hat{\pi}_{H-t}$, *as defined in step 3 of LR-MCPI, is an optimal policy with probability at least* $1-\delta(t+1)/H - 4(t+1)(|S|\wedge|A|)^{-10}$ *for each* $t \in \{0,\ldots,H-1\}$.

Proof of Lemma 11. Let p_1, p_2, and N_{H-t} be as defined in Lemma 11. We prove this lemma with strong induction on t. The base case occurs at step $t=0$ in which case our estimates, $\hat{Q}_H(s,a) = \frac{1}{N_H}\sum_{i=1}^{N_H} r_{H,i}(s,a)$ over Ω_H, are only averages of realizations $r_{H,i}$ of $R_H(s,a)$. Since $R_H(s,a)$ is a bounded random variable for all $(s,a) \in S\times A$, from Hoeffding's inequality (Theorem 14) with $N_H = \frac{2(c'_H)^2|S^{\#}_H|^2|A^{\#}_H|^2\log(2H|S||A|/\delta)}{\Delta^2_{\min}}$ (c'_H defined as in Proposition 8),

$$|\hat{Q}_H(s,a) - Q^*_H(s,a)| \leq \frac{\Delta_{\min}}{2c'_H|S^{\#}_H||A^{\#}_H|}$$

for all $(s,a)\in\Omega_H$ with probability at least $1-\delta/H$ because $|\Omega| \leq |S||A|$. With probability at least $1-4(|S|\wedge|A|)^{-10}$, Step 1 of LR-MCPI samples $S^{\#}_H$ and $A^{\#}_H$, which contain d anchor states and actions, respectively. Conditioned on this event, Step 2 of Algorithm gives

$$|\bar{Q}_H(s,a) - Q^*_H(s,a)| \leq \frac{\Delta_{\min}}{2}$$

for all $(s,a)\in S\times A$ from Proposition 8. From the union bound, the above inequality holds with probability at least $1-\delta/H - 4(|S|\wedge|A|)^{-10}$. From Step 3 of LR-MCPI, the identified policy is $\hat{\pi}_H(s) = \operatorname{argmax}_{a\in A}\bar{Q}_H(s,a)$. Assume for sake of contradiction that there exists an $s\in S$ such that $Q^*_H(s,\hat{\pi}_H(s)) < Q^*_H(s,\pi^*_H(s))$. Let $\hat{\pi}_H(s) = a, \pi^*_H(s) = a^*$. Hence,

$$\begin{aligned} Q^*_H(s,a^*) - Q^*_H(s,a) &= Q^*_H(s,a^*) - \bar{Q}_H(s,a) + \bar{Q}_H(s,a) - Q^*_H(s,a) \\ &\leq Q^*_H(s,a^*) - \bar{Q}_H(s,a^*) + \frac{\Delta_{\min}}{2} \\ &\leq \Delta_{\min} \end{aligned}$$

where the first inequality comes from how $\hat{\pi}_H(s)$ is defined and the matrix estimation step. Hence, we reach a contradiction since $Q^*_H(s,a^*) - Q^*_H(s,a)$ is less than the suboptimality gap. Thus, $\hat{\pi}_H(s)$ is an optimal policy. Hence, the base case holds.

Next, let $x\in\{0,\ldots,H-1\}$. Assume that the inductive hypothesis, the policy $\hat{\pi}_{H-x'}$ found in Step 4 of LR-MCPI is an optimal policy with probability at least $1-\delta(x'+1)/H - 4(x'+1)(|S|\wedge|A|)^{-10}$, holds for all $0\leq x'\leq x$. From the union bound, it follows that with probability at least $1-\delta(x+1)/H - 4(x+1)(|S|\wedge|A|)^{-10}$, LR-MCPI has constructed an optimal policy for time steps $\{H-x,\ldots,H\}$. At step $t+1$, under the event in Lemma 6, $S^{\#}_{H-x-1}$ and $A^{\#}_{H-x-1}$ contain d anchor states and actions, which occurs with probability at least $1-4(|S|\wedge|A|)^{-10}$.

Following Step 1 of LR-MCPI, we have $\hat{Q}_{H-x-1}(s,a) = \hat{r}^{\text{cum}}_{H-x-1}(s,a)$, which is an unbiased estimate of $Q^*_{H-x-1}(s,a)$. Furthermore, $\hat{r}^{\text{cum}}_{H-x-1}(s,a)\in[0,x+2]$ is a bounded random variable because of bounded rewards. Hence, from Hoeffding's inequality (Theorem 14), with the choice of $N_{H-x-1} = \frac{2(x+2)^2(c'_{H-x-1})^2|S^{\#}_{H-x-1}|^2|A^{\#}_{H-x-1}|^2\log(2H|S||A|/\delta)}{\Delta^2_{\min}}$, it follows that

$$|\hat{Q}_{H-x-1}(s,a) - Q^*_{H-x-1}(s,a)| \leq \frac{\Delta_{\min}}{2c'_{H-x-1}|S^{\#}_{H-x-1}||A^{\#}_{H-x-1}|} \quad \forall (s,a)\in\Omega_{H-x-1}$$

with probability $1 - \frac{\delta}{H|S||A|}$ for each $(s, a) \in \Omega_{H-x-1}$. Step 2 of LR-MCPI gives

$$|\bar{Q}_{H-x-1} - Q^*_{H-x-1}|_\infty \leq \frac{\Delta_{\min}}{2}$$

from Proposition 8. The union bound asserts that the above error guarantee holds with probability at least $1 - \delta(x+2)/H - 4(x+2)(|S| \wedge |A|)^{-10}$. From Step 3 of LR-MCPI, the identified policy is $\hat{\pi}_{H-x-1}(s) = \operatorname{argmax}_{a \in A} \bar{Q}_{H-x-1}(s, a)$. Assume for sake of contradiction that there exists an $s \in S$ such that $Q^*_{H-x-1}(s, \hat{\pi}_{H-x-1}(s)) < Q^*_{H-x-1}(s, \pi^*_{H-x-1}(s))$. Let $\hat{\pi}_{H-x-1}(s) = a, \pi^*_{H-x-1}(s) = a^*$. Hence,

$$\begin{aligned}
&Q^*_{H-x-1}(s, a^*) - Q^*_{H-x-1}(s, a)\\
&\qquad = Q^*_{H-x-1}(s, a^*) - \bar{Q}_{H-x-1}(s, a) + \bar{Q}_{H-x-1}(s, a) - Q^*_{H-x-1}(s, a)\\
&\qquad \leq Q^*_{H-x-1}(s, a^*) - \bar{Q}_{H-x-1}(s, a^*) + \frac{\Delta_{\min}}{2}\\
&\qquad \leq \Delta_{\min}
\end{aligned}$$

where the first inequality comes from how $\hat{\pi}_{H-x-1}(s)$ is defined and the matrix estimation step. Hence, we reach a contradiction since $Q^*_{H-x-1}(s, a^*) - Q^*_{H-x-1}(s, a)$ is less than the suboptimality gap. Thus, $\hat{\pi}_{H-x-1}(s)$ is an optimal policy, and the inductive step holds for $x + 1$. Hence, from mathematical induction, the lemma holds. □

Proof of Theorem 3. Recall $N_{H-t} = \frac{2(t+1)^2(c'_{H-t})^2|S^\#_{H-t}|^2|A^\#_{H-t}|^2 \log(2H|S||A|/\delta)}{\Delta^2_{\min}}$. From Lemma 11 for $t = H - 1$, we have proven that the algorithm correctly identifies an optimal policy with probability at least $1 - \delta - 4H(|S| \wedge |A|)^{-10}$. The number of samples used is upper bounded by

$$\sum_{t=0}^{H-1} (|S^\#_{H-t}||A| + |S||A^\#_{H-t}|) N_{H-t}(t+1)$$

where the $t+1$ comes from the length of the rollout. With our choice of $p_1, p_2,$ and N_{H-t}, it follows that

$$\begin{aligned}
\sum_{t=0}^{H-1} |S^\#_{H-t}| N_{H-t}(t+1) &= \sum_{t=0}^{H-1} \frac{2(t+1)^3(c'_{H-t})^2|S^\#_{H-t}|^3|A^\#_{H-t}|^2 \log(2H|S||A|/\delta)}{\Delta^2_{\min}}\\
&= \frac{2{c'_{\max}}^2 \log(2H|S||A|/\delta)}{\Delta^2_{\min}} \sum_{t=0}^{H-1} |S^\#_{H-t}|^3 |A^\#_{H-t}|^2 (t+1)^3
\end{aligned}$$

where $c_{max} \geq c'_i$ for all $i \in [H]$. Similarly,

$$\sum_{t=0}^{H-1} |A^\#_{H-t}| N_{H-t}(t+1) = \frac{2{c'_{\max}}^2 \log(2H|S||A|/\delta)}{\Delta^2_{\min}} \sum_{t=0}^{H-1} |S^\#_{H-t}|^2 |A^\#_{H-t}|^3 \ (t+1)^3.$$

From the one-sided Bernstein's inequality, Proposition 15, for $t \in \{0, \ldots, H-1\}$ and $C'' = \frac{25600}{3\mu d}$,

$$\begin{aligned}
\mathbb{P}\left(|S^\#_{H-t}| - \mathbb{E}[|S^\#_{H-t}|] \geq C'' p_1 |S|\right) &\leq \exp\left(-\frac{p_1^2 (C'')^2 |S|}{2(p_1 + \frac{p_1 C''}{3})}\right)\\
&\leq \exp\left(-\frac{\mu d C''}{640(1 + \frac{1}{3})} \log(|S|)\right)\\
&= |S|^{-10}.
\end{aligned}$$

With a similar argument,

$$\mathbb{P}\left(|A^{\#}_{H-t}| - \mathbb{E}[|A^{\#}_{H-t}|] \geq C'' p_2 |A|\right) \leq |A|^{-10}$$

for $t \in \{0, \ldots, H-1\}$. From our definition of p_1, p_2, it follows that $\mathbb{E}\left[|S^{\#}_{H-t}|\right] = O\left(d\mu \log(|S|)\right)$ and $\mathbb{E}\left[|A^{\#}_{H-t}|\right] = O\left(d\mu \log(|A|)\right)$ for all $t \in \{0, \ldots, H-1\}$. A union bound over all $t \in \{0, \ldots, H-1\}$ asserts that

$$|S^{\#}_{H-t}| \in O\left(d\mu \log(|S|)\right), \quad |A^{\#}_{H-t}| \leq O\left(d\mu \log(|A|)\right)$$

with probability at least $1 - 2H(|S| \wedge |A|)^{-10}$. From another union bound, with probability at least $1 - \delta - 6H(|S| \wedge |A|)^{-10}$, the sample complexity of LR-MCPI under Assumption 2 is upper bounded by

$$\begin{aligned}
&|S| \sum_{t=0}^{H-1} |A^{\#}_{H-t}| N_{H-t}(t+1) + |A| \sum_{t=0}^{H-1} |S^{\#}_{H-t}| N_{H-t}(t+1) \\
&\qquad \leq \frac{2{c'_{\max}}^2 \log(2H|S||A|/\delta)}{\Delta^2_{\min}} \sum_{t=0}^{H-1} \left(|S||S^{\#}_{H-t}|^2 |A^{\#}_{H-t}|^3] + |A||S^{\#}_{H-t}|^3 |A^{\#}_{H-t}|^2\right) (t+1)^3 \\
&\qquad \in \tilde{O}\left(\frac{d^5 \mu^5 \kappa^4 (|S| + |A|) H^4}{\Delta^2_{\min}}\right)
\end{aligned}$$

since $c'_{\max} \in \tilde{O}(\kappa^2)$. □

Lemma 12. *Let* $p_1 = \frac{\mu d \log(|S|)}{320|S|}, p_2 = \frac{\mu d \log(|A|)}{320|A|}$, *and* $N_{H-t} = \frac{2(t+1)^2 (c'_{H-t})^2 H^2 |S^{\#}_{H-t}|^2 |A^{\#}_{H-t}|^2 \log(2H|S||A|/\delta)}{\epsilon^2}$, *where* c'_{H-t} *is defined as in Proposition 8. Then the policy* $\hat{\pi}_{H-t}$ *and action-value function estimate* $\bar{Q}_{H-t}$, *as defined in Step 3 of LR-MCPI, are* $(t+1)\epsilon/H$*-optimal with probability* $1 - \delta(h+1)/H - 4(h+1)(|S| \wedge |A|)^{-10}$ *for each* $t \in \{0, \ldots, H-1\}$.

Proof of Lemma 12. Let p_1, p_2, and N_{H-t} be as defined in Lemma 12. This proof follows the steps in the proof of Lemma 11 but identifies a near optimal action instead of the optimal action at each time step with high probability. We prove this lemma with strong induction on t. The base case occurs at step $t = 0$ in which case our estimates, $\hat{Q}_H(s,a) = \frac{1}{N_H} \sum_{i=1}^{N_H} r_{H,i}(s,a)$ over Ω_H, are only averages of realizations $r_{H,i}$ of $R_H(s,a)$. Since $R_H(s,a)$ is a bounded random variable for all $(s,a) \in S \times A$, from Hoeffding's inequality (Theorem 14) with $N_H = \frac{2(c'_H)^2 H^2 |S^{\#}_H|^2 |A^{\#}_H|^2 \log(2H|S||A|/\delta)}{\epsilon^2}$ (c'_H defined as in Proposition 8),

$$|\hat{Q}_H(s,a) - Q^*_H(s,a)| \leq \frac{\epsilon}{2c'_H |S^{\#}_H||A^{\#}_H| H}$$

for all $(s,a) \in \Omega_H$ with probability at least $1 - \delta/H$ because $|\Omega_H| \leq |S||A|$. With probability at least $1 - 4(|S| \wedge |A|)^{-10}$, Step 0 of LR-MCPI samples $S^{\#}$ and $A^{\#}$, which contain d anchor states and actions, respectively. Conditioned on this event, step 1 of Algorithm gives

$$|\bar{Q}_H(s,a) - Q^*_H(s,a)| \leq \frac{\epsilon}{2H}$$

for all $(s,a) \in S \times A$ from Proposition 8. From the union bound, the above inequality holds with probability at least $1 - \delta/H - 4(|S| \wedge |A|)^{-10}$. From step 3 of LR-MCPI, the identified policy is

$\hat{\pi}_H(s) = \text{argmax}_{a\in A}\,\bar{Q}_H(s,a)$. Assume for sake of contradiction that there exists an $s \in S$ such that $Q^*_H(s,\hat{\pi}_H(s)) < Q^*_H(s,\pi^*_H(s)) - \epsilon/H$. Let $\hat{\pi}_H(s) = a, \pi^*_H(s) = a^*$. Hence,

$$\begin{aligned} Q^*_H(s,a^*) - Q^*_H(s,a) &= Q^*_H(s,a^*) - \bar{Q}_H(s,a) + \bar{Q}_H(s,a) - Q^*_H(s,a) \\ &\leq Q^*_H(s,a^*) - \bar{Q}_H(s,a^*) + \frac{\epsilon}{2H} \\ &\leq \frac{\epsilon}{H} \end{aligned}$$

where the first inequality comes from how $\hat{\pi}_H(s)$ is defined and the matrix estimation step. Hence, we reach a contradiction since $Q^*_H(s,a^*) - Q^*_H(s,a)$ is less ϵ/H. Thus, $\bar{Q}_H$ and $\hat{\pi}_H$ are both ϵ/H-optimal, and the base case holds.

Next, let $x \in \{0,\ldots,H-1\}$. Assume that the inductive hypothesis, the policy $\hat{\pi}_{H-x'}$ and action-value function estimate $\bar{Q}_{H-x'}$ found in Step 3 of LR-MCPI are $\epsilon(x'+1)/H$-optimal with probability at least $1-\delta(x'+1)/H - 4(x'+1)(|S|\wedge|A|)^{-10}$, holds for all $0 \leq x' \leq x$. From the union bound, it follows that with probability at least $1-\delta(x+1)/H - 4(x+1)(|S|\wedge|A|)^{-10}$, LR-MCPI has constructed $\epsilon(x'+1)/H$-optimal policies and action-value function estimates for time steps $x' \in \{H-x,\ldots,H\}$. At step $x+1$, under the event in Lemma 6, $S^{\#}_{H-x-1}$ and $A^{\#}_{H-x-1}$ contain d anchor states and actions, which occurs with probability at least $1-4(|S|\wedge|A|)^{-10}$.

Following Step 1 from LR-MCPI, we have $\hat{Q}_{H-x-1}(s,a) = \hat{r}^{\text{cum}}_{H-x-1}(s,a)$, an unbiased estimate of $Q^{\hat{\pi}}(s,a)$ for $\hat{\pi} = \{\hat{\pi}_h\}_{H-x\leq h\leq H}$, which is an ϵ-optimal policy. Furthermore, $\hat{r}^{\text{cum}}_{H-x-1}(s,a) \in [0,x+2]$ is a bounded random variable because of bounded rewards. Hence, from Hoeffding's inequality (Theorem 14), with the choice of $N_{H-x-1} = \frac{2(x+2)^2(c'_{H-x-1})^2H^2|S^{\#}_{H-x-1}|^2|A^{\#}_{H-x-1}|^2\log(2H|S||A|/\delta)}{\epsilon^2}$, it follows that

$$|\hat{Q}_{H-x-1}(s,a) - Q^{\hat{\pi}}_{H-x-1}(s,a)| \leq \frac{\epsilon}{2c'_{H-x-1}H|S^{\#}_{H-x-1}||A^{\#}_{H-x-1}|} \qquad \forall (s,a) \in \Omega_{H-x-1}$$

with probability $1-\frac{\delta}{H|S||A|}$ for each $(s,a)\in\Omega_{H-x-1}$. Step 2 of LR-MCPI gives

$$\|\bar{Q}_{H-x-1} - Q^{\hat{\pi}}_{H-x-1}\|_\infty \leq \frac{\epsilon}{2H}$$

from Proposition 8. The union bound asserts that the above error guarantee holds with probability at least $1-\delta(x+2)/H - 4(x+2)(|S|\wedge|A|)^{-10}$.

From step 3 of LR-MCPI, the identified policy is $\hat{\pi}_{H-x-1}(s) = \text{argmax}_{a\in A}\,\bar{Q}_{H-x-1}(s,a)$. For all $(s,a) \in S\times A$,

$$\begin{aligned} |\bar{Q}_{H-x-1}(s,a) - Q^*_{H-x-1}(s,a)| &\leq |\bar{Q}_{H-x-1}(s,a) - Q^{\hat{\pi}}_{H-x-1}(s,a)| \\ &\quad + |Q^{\hat{\pi}}_{H-x-1}(s,a) - Q^*_{H-x-1}(s,a)| \\ &\leq \frac{\epsilon}{2H} + \left|\mathbb{E}_{s'\sim P_{H-x-1}(\cdot|s,a)}\left[V^{\hat{\pi}}_{H-x}(s') - V^*_{H-x}(s')\right]\right| \\ &\leq \frac{\epsilon}{2H} + \left|\mathbb{E}_{s'\sim P_{H-x-1}(\cdot|s,a)}\left[(x+1)\epsilon/H\right]\right| \\ &= \frac{(2x+3)\epsilon}{2H}. \end{aligned}$$

Thus, $\bar{Q}_{H-x-1}$ is $\frac{\epsilon(x+2)}{H}$-optimal. It follows from the construction of $\hat{\pi}_{H-x-1}(s)$ that

$$\bar{Q}_{H-x-1}(s,\hat{\pi}_{H-x-1}(s)) \geq \bar{Q}_{H-x-1}(s,a'),$$

where $a' = \arg\max_a Q^*_{H-x-1}(s,a)$. Hence, for all $s \in S$,

$$\begin{aligned}|V^*_{H-x-1}(s) - V^{\hat{\pi}}_{H-x-1}(s)| &\leq |Q^*_{H-x-1}(s,a') - \bar{Q}_{H-x-1}(s, \hat{\pi}_{H-x-1}(s))| \\ &\quad + |\bar{Q}_{H-x-1}(s, \hat{\pi}_{H-x-1}(s)) - Q^{\hat{\pi}}_{H-x-1}(s, \hat{\pi}_{H-x-1}(s))| \\ &\leq \frac{(2x+3)\epsilon}{2H} + \frac{\epsilon}{2H} \\ &= \frac{(x+2)\epsilon}{H}.\end{aligned}$$

Thus, $\hat{\pi}_{H-x-1}(s)$ and $\bar{Q}_{H-x-1}$ are $(x+2)\epsilon/H$-optimal, and the inductive step holds for $x+1$. Hence, from mathematical induction, the lemma holds. □

Proof of Theorem 4. The proof of this theorem is the same as the proof of Theorem 3 except with different N_h. Recall $N_{H-t} = \frac{2(t+1)^2(c'_{H-t})^2H^2|S^{\#}_{H-t}|^2|A^{\#}_{H-t}|^2\log(2H|S||A|/\delta)}{\epsilon^2}$. From Lemma 12 for $t = H-1$, we have proven that the algorithm correctly identifies an ϵ-optimal policy and action-value function with probability at least $1 - \delta - 4H(|S| \wedge |A|)^{-10}$. Following the same argument as in the proof of Theorem 3, the number of samples used is upper bounded by

$$\frac{2{c'_{\max}}^2H^2\log(2H|S||A|/\delta)}{\epsilon^2}\sum_{t=0}^{H-1}(|A||S^{\#}_{H-t}|^3|A^{\#}_{H-t}|^2 + |S||S^{\#}_{H-t}|^2|A^{\#}_{H-t}|^3)(t+1)^3$$

with our choice of $p_1, p_2,$ and N_{H-t}. Following in the same argument used in the Proof of Theorem 3, from the one-sided Bernstein's inequality, Proposition 15, for $t \in \{0, \ldots, H-1\}$,

$$\mathbb{P}\left(|S^{\#}_{H-t}| - \mathbb{E}[|S^{\#}_{H-t}|] \geq C''p_1|S|\right) \leq |S|^{-10}$$
$$\mathbb{P}\left(|A^{\#}_{H-t}| - \mathbb{E}[|A^{\#}_{H-t}|] \geq C''p_2|A|\right) \leq |A|^{-10}$$

where $C'' = \frac{25600}{3\mu d}$. A union bound over all $t \in \{0, \ldots, H-1\}$ asserts that

$$|S^{\#}_{H-t}| \leq O\left(\mu d \log(|S|)\right), \quad |A^{\#}_{H-t}| \in O\left(\mu d \log(|A|)\right)$$

with probability at least $1 - 2H(|S| \wedge |A|)^{-10}$. From one last union bound, with probability at least $1 - \delta - 6H(|S| \wedge |A|)^{-10}$, the sample complexity of LR-MCPI under Assumption 3 is upper bounded by

$$\begin{aligned}&\frac{2{c'_{\max}}^2H^2\log(2H|S||A|/\delta)}{\epsilon^2}\sum_{t=0}^{H-1}\left(|S||S^{\#}_{H-t}|^2|A^{\#}_{H-t}|^3] + |A||S^{\#}_{H-t}|^3|A^{\#}_{H-t}|^2\right)(t+1)^3 \\ &\qquad \in \tilde{O}\left(\frac{d^5\mu^5\kappa^2(|S|+|A|)H^6}{\epsilon^2}\right)\end{aligned}$$

since $c'_{\max} \in \tilde{O}(\kappa^2)$. □

Lemma 13. *Let* $p_1 = \frac{\mu d \log(|S|)}{320|S|}, p_2 = \frac{\mu d \log(|A|)}{320|A|}$, *and* $N_{H-t} = \frac{(t+1)^2(c'_{H-t})^2|S^{\#}_{H-t}|^2|A^{\#}_{H-t}|^2H^2\log(2H|S||A|/\delta)}{2\epsilon^2}$, *where* c'_{H-t} *is defined as in Proposition 8. Then, the action-value function estimate* $\bar{Q}_{H-t}$, *as defined in step 3 of LR-EVI, is* $(t+1)\epsilon/H$*-optimal with probability at least* $1 - \delta(t+1)/H - 4(t+1)(|S| \wedge |A|)^{-10}$ *for each* $t \in \{0, \ldots, H-1\}$.

Proof of Lemma 13. This proof follows the steps in the proof of the previous two lemmas but shows that the identified Q function is near optimal at each time step instead of building a near-optimal policy. Let p_1, p_2, and N_{H-t} be as defined in Lemma 13. We prove this lemma with strong induction on t. The base case occurs at step $t = 0$ in which case our estimates, $\hat{Q}_H(s,a) = \frac{1}{N_H}\sum_{i=1}^{N_H} r_{H,i}(s,a)$ over Ω_H, are only averages of realizations $r_{H,i}$ of $R_H(s,a)$ since $\hat{V}_{H+1} = \vec{0}$. Since $R_H(s,a)$ is a bounded random variable for all $(s,a) \in S \times A$, from Hoeffding's inequality (Theorem 14) with $N_H = \frac{(c'_H)^2|S^{\#}_H|^2|A^{\#}_H|^2H^2\log(2H|S||A|/\delta)}{2\epsilon^2}$ (c'_H defined as in Proposition 8),

$$|\hat{Q}_H(s,a) - Q^*_H(s,a)| \leq \frac{\epsilon}{c'_H|S^{\#}_H||A^{\#}_H|H}$$

for all $(s,a) \in \Omega_H$ with probability at least $1 - \delta/H$ because $|\Omega_H| \leq |S||A|$. With probability at least $1 - 4(|S| \wedge |A|)^{-10}$, Step 1 of LR-MCPI samples $S^{\#}_H$ and $A^{\#}_H$, which contain d anchor states and actions, respectively. Conditioned on this event, Step 2 of LR-EVI gives

$$|\bar{Q}_H(s,a) - Q^*_H(s,a)| \leq \frac{\epsilon}{H}$$

for all $(s,a) \in S \times A$ from Proposition 8. From the union bound, the above inequality holds with probability at least $1 - \delta/H - 4(|S| \wedge |A|)^{-10}$. Hence, the base case holds.

Next, let $x \in \{0, \ldots, H-1\}$. Assume that the inductive hypothesis, the action-value function estimates $\bar{Q}_{H-x'}$ are $(x'+1)\epsilon/H$-optimal with probability at least $1 - \delta(x'+1)/H - 4(t'+1)(|S| \wedge |A|)^{-10}$ holds for all $0 \leq x' \leq x$. From the union bound, it follows that with probability at least $1 - \delta(x+1)/H - 4(x+1)(|S| \wedge |A|)^{-10}$, LR-EVI has constructed $(x'+1)\epsilon/H$-optimal action-value functions for time steps $x' \in [x]$. At step $x+1$, under the event in Lemma 6, $S^{\#}_{H-x-1}$ and $A^{\#}_{H-x-1}$ contain d anchor states and actions, which occurs with probability at least $1 - 4(|S| \wedge |A|)^{-10}$. Following Step 1 from LR-EVI, we have

$$\hat{Q}_{H-x-1}(s,a) = \hat{r}_{H-x-1}(s,a) + \mathbb{E}_{s' \sim \hat{P}_{H-x-1}(\cdot|s,a)}[\hat{V}_{H-x}(s')],$$

an unbiased estimate of $Q'_{H-x-1}(s,a) = r_{H-x-1}(s,a) + \mathbb{E}_{s' \sim P_{H-x-1}(\cdot|s,a)}[\hat{V}_{H-x}(s')]$. Furthermore, $\hat{Q}_{H-x-1}(s,a) \in [0, x+2]$ is a bounded random variable because of bounded rewards. Hence, from Hoeffding's inequality (Theorem 14), with the choice of

$$N_{H-x-1} = \frac{(x+2)^2(c'_{H-x-1})^2|S^{\#}_{H-x-1}|^2|A^{\#}_{H-x-1}|^2H^2\log(2H|S||A|/\delta)}{2\epsilon^2},$$

it follows that

$$|\hat{Q}_{H-x-1}(s,a) - Q'_{H-x-1}(s,a)| \leq \frac{\epsilon}{c'_{H-x-1}|S^{\#}_{H-x-1}||A^{\#}_{H-x-1}|H} \quad \forall (s,a) \in \Omega_{H-x-1}$$

with probability $1 - \frac{\delta}{H|S||A|}$ for each $(s,a) \in \Omega_{H-x-1}$. Step 2 of LR-EVI gives

$$|\bar{Q}_{H-x-1} - Q'_{H-x-1}|_\infty \leq \frac{\epsilon}{H}$$

from Proposition 8. The union bound asserts that the above error guarantee holds with probability at least $1-\delta(x+2)/H-4(x+2)(|S|\wedge|A|)^{-10}$. Hence, for all $(s,a)\in S\times A$,

$$\begin{aligned}|\bar{Q}_{H-x-1}(s,a)-Q^*_{H-x-1}(s,a)| &\leq |\bar{Q}_{H-x-1}(s,a)-Q'_{H-x-1}(s,a)|+|Q'_{H-x-1}(s,a)-Q^*_{H-x-1}(s,a)|\\ &\leq \frac{\epsilon}{H}+|\mathbb{E}_{s'\sim P_{H-x-1}(\cdot|s,a)}[\max_{a\in A}\bar{Q}_{H-x}(s',a')-V^*_{H-x}(s')]|\\ &\leq \frac{\epsilon}{H}+|\mathbb{E}_{s'\sim P_{H-x-1}(\cdot|s,a)}[(x+1)\epsilon/H]|\\ &= \frac{(x+2)\epsilon}{H}\end{aligned}$$

Thus, $\bar{Q}_{H-x-1}$ is $(x+2)\epsilon/H$-optimal, and the inductive step holds for $x+1$. Hence, from mathematical induction, the lemma holds. □

Proof of Theorem 5. The proof of this theorem is essentially the same as the proofs of the previous two theorems. Recall $N_{H-t}=\frac{(t+1)^2(c'_{H-t})^2|S^{\#}_{H-t}|^2|A^{\#}_{H-t}|^2H^2\log(2H|S||A|/\delta)}{2\epsilon^2}$. From Lemma 13 for $t=H-1$, we have proven that the algorithm finds ϵ-optimal action-value functions with probability at least $1-\delta-4H(|S|\wedge|A|)^{-10}$. Following the same argument as in the proof of Theorem 3, the number of samples used is upper bounded by

$$\frac{{c'_{\max}}^2H^2\log(2H|S||A|/\delta)}{2\epsilon^2}\sum_{t=0}^{H-1}(|A||S^{\#}_{H-t}|^3|A^{\#}_{H-t}|^2+|S||S^{\#}_{H-t}|^2|A^{\#}_{H-t}|^3)(t+1)^2$$

with our choice of $p_1,p_2,$ and N_{H-t}. Following in the same argument used in the Proofs of Theorem 3 and 4, from the one-sided Bernstein's inequality, Proposition 15, for $t\in\{0,\ldots,H-1\}$,

$$\mathbb{P}\left(|S^{\#}_{H-t}|-\mathbb{E}[|S^{\#}_{H-t}|]\geq C''p_1|S|\right)\leq|S|^{-10}$$
$$\mathbb{P}\left(|A^{\#}_{H-t}|-\mathbb{E}[|A^{\#}_{H-t}|]\geq C''p_2|A|\right)\leq|A|^{-10}$$

where $C''=\frac{25600}{3\mu d}$. A union bound over all $t\in\{0,\ldots,H-1\}$ asserts that

$$|S^{\#}_{H-t}|\leq O\left(d\mu\log(|S|)\right),\quad |A^{\#}_{H-t}|\in O\left(d\mu\log(|A|)\right)$$

with probability at least $1-2H(|S|\wedge|A|)^{-10}$. From one last union bound, with probability at least $1-\delta-6H(|S|\wedge|A|)^{-10}$, the sample complexity of LR-EVI under Assumption 4 is upper bounded by

$$\begin{aligned}&\frac{{c'_{\max}}^2H^2\log(2H|S||A|/\delta)}{2\epsilon^2}\sum_{t=0}^{H-1}\left(|S||S^{\#}_{H-t}|^2|A^{\#}_{H-t}|^3]+|A||S^{\#}_{H-t}|^3|A^{\#}_{H-t}|^2\right)(t+1)^2\\ &\qquad\in\tilde{O}\left(\frac{d^5\mu^5\kappa^4(|S|+|A|)H^5}{\epsilon^2}\right).\end{aligned}$$

since $c'_{\max}\in\tilde{O}(\kappa^2)$. □

D Technical Propositions and Theorems

We present the following lemmas, propositions, and theorems for the readers' convenience.

Theorem 14 (Hoeffding's Inequality [35])**.** *Let $X_1, \ldots, X_n$ be independent, and X_i have mean μ_i and sub-Gaussian parameter σ_i. Then, for all $t \geq 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

Proposition 15 (Proposition 2.14 (One-sided Bernstein's Inequality) [35])**.** *Given n independent random variables such that $X_i \leq b$ almost surely, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq cn\right) \leq \exp\left(-\frac{nc^2}{2(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] + \frac{bc}{3})}\right).$$

Theorem 16 (Matrix Bernstein [32])**.** *Let $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^{d_1 \times d_2}$ be independent zero-mean matrices satisfying*

$$\|X^{(i)}\|_{op} \leq b, \quad a.s.$$

$$\max\{\|\sum_{i=1}^{n}\mathbb{E}[X^{(i)^\top}X^{(i)}]\|_{op}, \|\sum_{i=1}^{n}\mathbb{E}[X^{(i)}X^{(i)^\top}]\|_{op}\} \leq n\sigma^2.$$

Then

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}X^{(i)}\right\|_{op} \geq t\right) \leq (d_1 + d_2)\exp\left(-\frac{t^2}{2(n\sigma^2 + \frac{bt}{3})}\right).$$