Appendix

This appendix is organized as follows. We begin with describing the formal statements for the extensions of Theorem 1 along with necessary algorithmic modifications in Appendix A. We then provide formal statements and proofs for several models with low Bellman rank introduced in Section 3 in Appendix B. In Appendix C, we provide the proofs of our main results, including a more general version of Theorem 1 which also covers the aforementioned extensions. In Appendix D we give the various technical lemmas required for our proofs and we present the main lower bound statements and proofs in Appendix F.

A. Extensions

We introduce four important extensions to the algorithm and analysis.

A.1. Unknown Bellman Rank

The first extension eliminates the need to know M in advance (note that Algorithm 1 requires M as an input parameter). A simple procedure, as described in Algorithm 2, can guess the value of M by a doubling schedule and handle this situation with no consequences to asymptotic sample complexity.[8]

Algorithm 2 GUESSM$(\mathcal{F}, \zeta, \epsilon, \delta)$

1: **for** $i = 1, 2, \ldots$ **do**
2: $M' \leftarrow 2^i$.
3: Call OLIVE$(\mathcal{F}, M', \epsilon, \frac{\delta}{i(i+1)})$ with parameters specified on Theorem 1.
4: Terminate the subroutine when $t > HM' \log\left(\frac{6H\sqrt{M'}\zeta}{\epsilon}\right) / \log(5/3)$ in Line 4 (the for-loop).
5: **if** a policy π is returned from OLIVE **then**
6: **Return** π.
7: **end if**
8: **end for**

Theorem 2. *For any $\epsilon, \delta \in (0, 1)$, any Contextual Decision Process and function class $\mathcal{F}$ that admit a Bellman factorization with parameters M, ζ, if we run* GUESSM$(\mathcal{F}, \epsilon, \delta)$*, then with probability at least $1 - \delta$,* GUESSM *halts and returns a policy which satisfies $V^{\hat{\pi}} \geq V^\star_{\mathcal{F}} - \epsilon$, and the number of episodes required is at most*

$$\tilde{\mathcal{O}}\left(\frac{M^2 H^3 K}{\epsilon^2} \log(N\zeta/\delta)\right).$$

We give some intuition about the proof here, with details in Appendix E.1. In Algorithm 2, M' is a guess for M which grows exponentially. When $M' \geq M$, analysis of the main algorithm shows that OLIVE$(\mathcal{F}, M', \epsilon, \frac{\delta}{i(i+1)})$ terminates and returns a near-optimal policy with high probability. The doubling schedule implies that the largest guess is at most $2M$, which has negligible effect on the sample complexity. On the other hand, OLIVE may not explore effectively when $M' < M$, because not enough samples (chosen according to M') are used to estimate the average Bellman errors in Line 13 of OLIVE. This worse accuracy does not guarantee sufficient progress in learning.

However, the high-probability guarantee that $f^\star$ is not eliminated is unaffected, because the threshold ϕ on Line 14 of OLIVE is set in accordance with the sample size n specified in Theorem 1, regardless of M. Consequently, if the algorithm ever terminates when $M' < M$, we still get a near-optimal policy. When $M' < M$ the OLIVE subroutine may not terminate, which the explicit termination on line 4 in Algorithm 2 addresses. Finally, by splitting the failure probability δ appropriately among all guesses of M', we obtain the same order of sample complexity as in Theorem 1.

[8] In Algorithm 2 we assume that ζ is fixed. In the examples provided in Proposition 1, 2, and 3, however, ζ grows with M in the form of $\zeta = 2\sqrt{M}$. In this case, we can compute $\zeta' = 2\sqrt{M'}$ and call OLIVE with ζ' instead of ζ. As long as ζ is a polynomial term and non-decreasing in M the same analysis applies and Theorem 2 holds.

A.2. Separation of Policy Class and V-value Class

So far, we have assumed that the agent has access to a class of Q-value functions $\mathcal{F} \subset \mathcal{X} \times \mathcal{A} \to [0,1]$. In this section, we show the algorithm allows separate representations of policies and *V-value* functions.

For every $f \in \mathcal{F}$, and any $x \in \mathcal{X}, a \neq \pi_f(x)$, we note that the value of $f(x,a)$ is not used by Algorithm 1, and changing it to arbitrary values does not affect the execution of the algorithm as long as $f(x,a) \leq f(x, \pi_f(x))$ (so that π_f does not change). In other words, the algorithm only interacts with f in two forms:

1. f's greedy policy π_f.
2. A mapping $g_f : x \mapsto f(x, \pi_f(x))$. We call such mappings V-value functions to contrast the previous use of Q-value functions.[9]

Hence, supplying $\mathcal{F}$ is equivalent to supplying the following space of (policy, V-value function) pairs:

$$\{(\pi_f, g_f) : f \in \mathcal{F}\}.$$

This observation provides further evidence that Definition 3 is significantly less restrictive than standard realizability assumptions. Validity of f means that (π_f, g_f) obeys the *Bellman Equations for Policy Evaluation* (i.e., g_f predicts the long-term value of following π_f), as opposed to the more common *Bellman Optimality Equations*. In MDPs, there are many ways to satisfy the policy evaluation equations at every state simultaneously, while $Q^\star$ is the only function that satisfies all optimality equations.

More generally, instead of using a Q-value function class, we can run OLIVE with a policy space $\Pi \subset \mathcal{X} \to \mathcal{A}$ and a V-value function class $\mathcal{G} \subset \mathcal{X} \to [0,1]$ where we assemble (policy,V-value function) pairs by taking the Cartesian product of Π and $\mathcal{G}$. OLIVE can be run here with the understanding that each Q-value function f in OLIVE is associated with a (π, g) pair, and the algorithm uses π instead of π_f and $g(x)$ instead of $f(x, \pi_f(x))$. All the analysis applies directly with this transformation, and the $\log|\mathcal{F}|$ dependence in sample complexity is replaced by $\log|\Pi| + \log|\mathcal{G}|$. Note also that the definition of Bellman factorization also extends naturally to this case, where f in Equation 3 corresponds to a (π, g) pair and f' corresponds to a roll-in policy, π'.

A.3. Infinite Hypothesis Classes

The arguments in Section 4 assume that $|\mathcal{F}| = N < \infty$. However, almost all commonly used function approximators are infinite classes, which restricts the applicability of the algorithm. On the other hand, the size of the function class appears in the analysis only through deviation bounds, so techniques from empirical process theory can be used to generalize the results to infinite classes. This section establishes parallel versions of those deviation bounds for function classes with finite combinatorial dimensions, and together with the rest of the original analysis we can show the algorithm enjoys similar guarantees when working with infinite hypothesis classes.

Specifically, we consider the setting where Π and $\mathcal{G}$ are given (see Appendix A.2), and they are infinite classes with finite combinatorial dimensions. We assume that Π has finite *Natarajan dimension* (Definition 6), and $\mathcal{G}$ has finite *pseudo dimension* (Definition 7). These two dimensions are standard extensions of VC-dimension to multi-class classification and regression respectively.

Definition 6 (Natarajan dimension (Natarajan, 1989))**.** *Suppose $\mathcal{X}$ is a feature space and $\mathcal{Y}$ is a finite label space. Given hypothesis class $\mathcal{H} \subset \mathcal{X} \to \mathcal{Y}$, its Natarajan dimension Ndim$(\mathcal{H})$ is defined as the maximum cardinality of a set $A \subseteq \mathcal{X}$ that satisfies the following: there exists $h_1, h_2 : A \to \mathcal{Y}$ such that (1) $\forall x \in A$, $h_1(x) \neq h_2(x)$, and (2) $\forall B \subseteq A$, $\exists h \in \mathcal{H}$ such that $\forall x \in B$, $h(x) = h_1(x)$ and $\forall x \in A \setminus B$, $h(x) = h_2(x)$.*

Definition 7 (Pseudo dimension (Haussler, 1992))**.** *Suppose $\mathcal{X}$ is a feature space. Given hypothesis class $\mathcal{H} \subset \mathcal{X} \to \mathbb{R}$, its pseudo dimension Pdim$(\mathcal{H})$ is defined as Pdim$(\mathcal{H})$ = VC-dim$(\mathcal{H}^+)$, where $\mathcal{H}^+ = \{(x,\xi) \mapsto \mathbf{1}[h(x) > \xi] : h \in \mathcal{H}\} \subset \mathcal{X} \times \mathbb{R} \to \{0,1\}$.*

The definition of pseudo dimension relies on that of VC-dimension, whose definition and basic properties are recalled in Appendix E.2. We state the final sample complexity result here. Since the algorithm parameters are somewhat com-

[9] In the MDP setting, such functions are also known as *state-value functions*.

plex expressions, we omit them in the theorem statement and provide specification in the proof, which is deferred to Appendix E.2.

Theorem 3. *Let $\Pi \subset \mathcal{X} \to \mathcal{A}$ with Ndim$(\Pi) \leq d_\Pi < \infty$ and $\mathcal{G} \subset \mathcal{X} \to [0,1]$ with Pdim$(\mathcal{G}) \leq d_\mathcal{G} < \infty$. For any $\epsilon, \delta \in (0,1)$, any Contextual Decision Process with policy space Π and function space $\mathcal{G}$ that admits a Bellman factorization with parameters M, ζ, if we run* OLIVE *with appropriate parameters, then with probability at least $1-\delta$,* OLIVE *halts and returns a policy $\hat{\pi}$ that satisfies $V^{\hat{\pi}} \geq V^\star_\mathcal{F} - \epsilon$, and the number of episodes required is at most*

$$\tilde{\mathcal{O}}\left(\frac{M^2H^3K^2}{\epsilon^2}\Big(d_\Pi + d_\mathcal{G} + \log(\zeta/\delta)\Big)\right). \tag{7}$$

Compared to Theorem 1, the sample complexity we get for infinite hypothesis classes has two differences: (1) $\log N$ is replaced by $d_\Pi + d_\mathcal{G}$, which is expected, based on the discussion in Appendix A.2, and (2) the dependence on K is quadratic as opposed to linear. In fact, in the proof of Theorem 1, we exploited the low-variance property of importance weights in Line 13 of OLIVE, and applied Bernstein's inequality to avoid a factor of K. With infinite hypothesis classes, the same approach does not apply directly. However, this may only be a technical issue, and a more refined analysis might recover the linear dependence (e.g., using tools from Panchenko (2002)).

A.4. Approximate Validity and Approximate Bellman Rank

Recall that the sample-efficiency guarantee of OLIVE relies on two major assumptions:

- We assumed that $\mathcal{F}$ contains valid functions (Definition 3). In practice, however, it is hard to specify a function class that contains strictly valid functions, as the notion of validity depends on the environment dynamics, which are unknown. A much more realistic situation is that some functions in $\mathcal{F}$ satisfy validity only *approximately*.
- We assumed that the average Bellman errors have an exact low-rank factorization (Definition 5). While this is true for a number of RL models (Section 3), it is worth keeping in mind that these are only *models* of the environments, which are different from and only approximations to the real environments themselves. Therefore, it is more realistic to assume that an *approximate* factorization exists when defining Bellman factorization.

In this section, we show that the algorithmic ideas of OLIVE are indeed robust against both types of approximation errors, and degrades gracefully as the two assumptions are violated. Below we introduce the approximate versions of Definition 3 and 5, give a slightly extended version of the algorithm, OLIVER (for Optimism-Led Iterative Value-function Elimination with Robustness, see Algorithm 3), and state its sample complexity guarantee in Theorem 4.

Definition 8 (Approximate validity of f)**.** *Given any CDP and function class $\mathcal{F}$, we say $f \in \mathcal{F}$ is θ-valid if for any $f' \in \mathcal{F}$ and any $h \in [H]$, $|\mathcal{E}(f, \pi_{f'}, h)| \leq \theta$.*

The approximation error θ introduced in Definition 8 allows the algorithm to compete against a broader range of functions; hence the notions of optimal function and value need to be re-defined accordingly.

Definition 9. *For a fixed θ, define $f^\star_\theta = \operatorname{argmax}_{f \in \mathcal{F}:\, f \text{ is } \theta\text{-valid}} V^{\pi_f}$, and $V^\star_{\mathcal{F},\theta} = V^{\pi_{f^\star_\theta}}$.*

By definition, $V^\star_{\mathcal{F},\theta}$ is non-decreasing in θ with Definition 3 being a special case where $\theta = 0$. When $\theta > 0$, we compete against some functions that do not obey Bellman equations, breaking an essential element of value-based RL. As a consequence, returning a policy with value close to $V^\star_{\mathcal{F},\theta}$ in a sample-efficient manner is very challenging, so the value that OLIVER can guarantee is suboptimal to $V^\star_{\mathcal{F},\theta}$ by a term that is proportional to θ and does not diminish with more data.

Definition 10 (Approximate Bellman rank)**.** *We say that a CDP $(\mathcal{X}, \mathcal{A}, H, P)$ and $\mathcal{F} \subset \mathcal{X} \times \mathcal{A} \to [0,1]$, admits a Bellman factorization with Bellman rank M, norm parameter ζ, and approximation error η, if there exists $\nu_h : \mathcal{F} \to \mathbb{R}^M, \xi_h : \mathcal{F} \to \mathbb{R}^M$ for each $h \in [H]$, such that for any $f, f' \in \mathcal{F}, h \in [H]$,*

$$|\mathcal{E}(f, \pi_{f'}, h) - \langle \nu_h(f'), \xi_h(f)\rangle| \leq \eta, \tag{8}$$

and $\|\nu_h(f')\|_2 \cdot \|\xi_h(f)\|_2 \leq \zeta < \infty$.

A modified version of OLIVE that deals with these approximation errors, OLIVER, is specified in Algorithm 3. Here, we use ϵ to denote the component of the suboptimality that diminishes as more data is collected, and the total suboptimality

Algorithm 3 OLIVER $(\mathcal{F}, \theta, M, \zeta, \eta, \epsilon, \delta)$

1: Let $\epsilon' = \epsilon + 2H(3\sqrt{M}(\theta+\eta)+\eta)$.
2: **Collect** n_{est} trajectories with actions taken in an arbitrary manner; save initial contexts $\{x_1^{(i)}\}_{i=1}^{n_{\text{est}}}$.
3: **Estimate** the predicted value for each $f \in \mathcal{F}$: $\hat{V}_f = \frac{1}{n_{\text{est}}} \sum_{i=1}^{n_{\text{est}}} f(x_1^{(i)}, \pi_f(x_1^{(i)}))$.
4: $\mathcal{F}_0 \leftarrow \mathcal{F}$.
5: **for** $t = 1, 2, \ldots$ **do**
6: **Choose policy** $f_t = \operatorname{argmax}_{f \in \mathcal{F}_{t-1}} \hat{V}_f$, $\pi_t = \pi_{f_t}$.
7: **Collect** n_{eval} trajectories $\{(x_1^{(i)}, a_1^{(i)}, r_1^i, \ldots, x_H^{(i)}, a_H^{(i)}, r_H^{(i)})\}_{i=1}^{n_{\text{eval}}}$ by following π_t (i.e. $a_h^{(i)} = \pi_t(x_h^{(i)})$ for all h, i).
8: **Estimate** $\forall h \in [H]$,

$$\tilde{\mathcal{E}}(f_t, \pi_t, h) = \frac{1}{n_{\text{eval}}} \sum_{i=1}^{n_{\text{eval}}} \left[f_t(x_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - f_t(x_{h+1}^{(i)}, a_{h+1}^{(i)}) \right]. \tag{9}$$

9: **if** $\sum_{h=1}^{H} \tilde{\mathcal{E}}(f_t, \pi_t, h) \le 5\epsilon'/8$ **then**
10: Terminate and ouptut π_t.
11: **end if**
12: Pick any $h_t \in [H]$ for which $\tilde{\mathcal{E}}(f_t, \pi_t, h_t) \ge 5\epsilon'/8H$ (One is guaranteed to exist).
13: Collect trajectories $\{(x_1^{(i)}, a_1^{(i)}, r_1^{(i)}, \ldots, x_H^{(i)}, a_H^{(i)}, r_H^{(i)})\}_{i=1}^{n}$ where $a_h^{(i)} = \pi_t(x_h^{(i)})$ for all $h \neq h_t$ and $a_{h_t}^{(i)}$ is drawn uniformly at random.
14: **Estimate**

$$\widehat{\mathcal{E}}(f, \pi_t, h_t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}[a_{h_t}^{(i)} = \pi_f(x_{h_t}^{(i)})]}{1/K} \Big(f(x_{h_t}^{(i)}, a_{h_t}^{(i)}) - r_{h_t}^{(i)} - f(x_{h_t+1}^{(i)}, \pi_f(x_{h_t+1}^{(i)})) \Big). \tag{10}$$

15: **Learn**

$$\mathcal{F}_t = \left\{ f \in \mathcal{F}_{t-1} : \left| \widehat{\mathcal{E}}(f, \pi_t, h_t) \right| \le \phi + \theta \right\}. \tag{11}$$

16: **end for**

that we can guarantee is ϵ plus a term proportional to θ and η (see Eq. (12) in Theorem 4). The algorithm is almost identical to OLIVE except in two places: (1) it uses ϵ' (defined on Line 1) in the termination condition (Line 9) as opposed to ϵ, and (2) it uses a higher threshold that depends on θ in Eq. (11) to avoid eliminating θ-valid functions. The approximation and sample complexity guarantees of OLIVER are stated in Theorem 4, with the proof deferred to Appendix D.

Theorem 4. *For any $\epsilon, \delta \in (0,1)$, any Contextual Decision Process and function class $\mathcal{F}$ that admit a Bellman factorization with parameters M, ζ, and η, suppose we run* OLIVER *with any $\theta \in [0,1]$, and $n_{est}, n_{eval}, n, \phi$ as specified in Theorem 1. Then with probability at least $1-\delta$,* OLIVER *halts and returns a policy $\hat{\pi}$ which is at most*

$$\epsilon + 8H\sqrt{M}(\theta + \eta) \tag{12}$$

suboptimal compared to $V^{\star}_{\mathcal{F},\theta}$ defined in Definition 9, and the number of episodes required is at most

$$\tilde{\mathcal{O}}\left(\frac{M^2 H^3 K}{\epsilon^2} \log(N\zeta/\delta) \right). \tag{13}$$

B. Models with Low Bellman Rank

B.1. Proof of Proposition 1

Let $M = |\mathcal{S}|$ and each element of $\nu_h(\cdot)$ and $\xi_h(\cdot)$ be indexed by $s \in \mathcal{S}$. We explicitly construct ν_h and ξ_h as follows: let $[\nu_h(f')]_s = \Pr[x_h = (s,h) \,|\, a_{1:h-1} \sim \pi_{f'}]$, and $[\xi_h(f)]_s = \mathbb{E}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1}) \,\middle|\, x_h = (s,h),\, a_{h:h+1} \sim \pi_f \right]$. In other words, $\nu_h(f')$ is the distribution over states induced by $\pi_{f'}$ at time step h, and the s-th element of ξ_h is the traditional notion of Bellman error for state s. It is easy to verify that Eq. (3) holds. For the norm constraint, since

$\|\nu_h(\cdot)\|_1 = 1$ and $\|\xi_h(\cdot)\|_\infty \le 2$, we have $\|\nu_h(\cdot)\|_2 \le 1$ and $\|\xi_h(\cdot)\|_2 \le 2\sqrt{M}$, hence $\zeta = 2\sqrt{M}$ is a valid upper bound on the product of vector norms.

B.2. Generalization of Li (2009)'s Setting

Li (2009, Section 8.2.3) considers the setting where the learner is given an abstraction ϕ that maps the large state space $\mathcal{S}$ in an MDP to some finite abstract state space $\bar{\mathcal{S}}$. $|\bar{\mathcal{S}}|$ is potentially much smaller than $|\mathcal{S}|$, and it is guaranteed that $Q^\star$ can be expressed as a function of $(\phi(s), a)$. Li shows that when delayed Q-learning is applied to this setting, the sample complexity has polynomial dependence on $|\bar{\mathcal{S}}|$ with no direct dependence on $|\mathcal{S}|$.

In the next proposition, we show that a similar setting for finite-horizon problems admits Bellman factorization with low Bellman rank. In particular, we subsume Li's setting by viewing it as a POMDP, where ϕ is a deterministic emission process that maps hidden state $s \in \mathcal{S}$ to discrete observations $\phi(s) \in \bar{\mathcal{S}} = \mathcal{O}$, and the candidate value functions are reactive so they depend on $\phi(s)$ but not directly on s or any previous state. More generally, Proposition 6 claims that for POMDPs with large hidden-state spaces and finite observation spaces, the Bellman rank is polynomial in the number of observations if the function class is reactive.

Proposition 6 (A generalization of (Li, 2009)'s setting)**.** *Consider a POMDP introduced in Example 2 with $|\mathcal{O}| < \infty$, and assume that rewards can only take C_R different discrete values.*[10] *The CDP $(\mathcal{X}, \mathcal{A}, H, P)$ induced by letting $\mathcal{X} = \mathcal{O} \times [H]$ and $x_h = (o_h, h)$, with any $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to [0, 1]$, admits a Bellman factorization with $M = |\mathcal{O}|^2 C_R K$ and $\zeta = 2|\mathcal{O}|K\sqrt{C_R}$.*

Proof. For any $f, f' \in \mathcal{F}, h \in [H]$, let $\nu_h(f')$ and $\xi_h(f)$ be vectors of length $|\mathcal{O}|^2 C_R K$. Let the entry of $\nu_h(f')$ indexed by (o_h, a_h, r_h, o_{h+1}) be

$$P[o_h, r_h, o_{h+1} \mid a_{1:h-1} \sim \pi_{f'}, a_h],$$

interpreted as the following: conditioned on the fact that the first $h-1$ actions are chosen according to $\pi_{f'}$, what is the probability of seeing a particular tuple of (o_h, r_h, o_{h+1}) when taking a particular action for a_h? For $\xi_h(f)$, let the corresponding entry be (with $x_h = (o_h, h)$ and $x_{h+1} = (o_{h+1}, h+1)$ as the corresponding contexts in the CDP)

$$\mathbf{1}[a_h = \pi_f(x_h)]\big(f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1}))\big).$$

It is not hard to verify that $\mathcal{E}(f, \pi_{f'}, h) = \langle \nu_h(f'), \xi_h(f) \rangle$. Since fixing a_h to any non-adaptive choice of action induces a valid distribution over (o_h, r_h, o_{h+1}), we have $\|\nu_h(f')\|_1 = K$ and $\|\nu_h(f')\|_2 \le K$. On the other hand, $\|\xi_h(f)\|_\infty \le 2$ but the vector only has $|\mathcal{O}|^2 C_R$ non-zero entries, so $\|\xi_h(f)\|_2 \le 2|\mathcal{O}|\sqrt{C_R}$. Together the norm bound follows. □

B.3. POMDP-like Models

Here we first state the formal version of Proposition 2, and prove Propositions 2 and 3 together by studying a slightly more general model (See Figure 1).

Proposition 7 (Formal version of Proposition 2)**.** *Consider an MDP introduced in Example 1. With a slight abuse of notation let Γ denote its transition matrix of size $|\mathcal{S} \times \mathcal{A}| \times |\mathcal{S}|$, whose element indexed by $((s, a), s')$ is $\Gamma(s'|s, a)$. Assume that there are two row-stochastic matrices $\Gamma^{(1)}$ and $\Gamma^{(2)}$ with sizes $|\mathcal{S} \times \mathcal{A}| \times M$ and $M \times |\mathcal{S}|$ respectively, such that $\Gamma = \Gamma^{(1)}\Gamma^{(2)}$. Recall that we convert an MDP into a CDP by letting $\mathcal{X} = \mathcal{S} \times [H]$, $x_h = (s_h, h)$. For any $\mathcal{F} \subset \mathcal{X} \times \mathcal{A} \to [0, 1]$, this model admits a Bellman factorization with Bellman rank M and $\zeta = 2\sqrt{M}$.*

The model in Figure 1 that we use to prove Propositions 2 and 3 simultaneously behaves like a POMDP except that both the transition function and the reward depends also on the observation, that is $\Gamma : \mathcal{S} \times \mathcal{O} \times \mathcal{A} \to \Delta(\mathcal{S})$ and $R : \mathcal{S} \times \mathcal{O} \times \mathcal{A} \to \Delta([0, 1])$. Clearly this model generalizes standard POMDPs, where the transition and reward are both assumed to be independent of the current observation.

This model also generalizes the MDP with low-rank dynamics described in Proposition 2: if the future hidden-state is independent of the current hidden-state conditioned on the observation (i.e., $\Gamma(s'|s, o, a)$ does not depend on s), the observations themselves become Markovian, and we can treat o as the observed state s in Proposition 7, and the hidden-state s

[10] The discrete reward assumption is made to simplify presentation and can be relaxed. For arbitrary rewards, we can always discretize the reward distribution onto a grid of resolution C_R, which incurs $\eta = O(1/C_R)$ approximation error in Definition 10.

as the low-rank factor in Proposition 7 (see Figure 1). Hence, Proposition 2 follows as a special case of the analysis for this more general model.

As in Proposition 3, we consider a class $\mathcal{F}$ reactive value functions. Observe that for the MDP with low rank dynamics, this provides essentially no loss of generality, since the optimal value function is reactive.

Proposition 8. *Let $(\mathcal{X}, \mathcal{A}, H, P)$ be the CDP induced by the model in Figure* **??** *which generalizes POMDPs, with $\mathcal{X} = \mathcal{O} \times [H]$ and $x_h = (o_h, h)$. Given any $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to [0, 1]$, the Bellman rank $M \leq |\mathcal{S}|$ with $\zeta = 2\sqrt{|\mathcal{S}|}$.*

Proof. For any $f, f' \in \mathcal{F}, h \in [H]$, consider

$$a_{1:h-1} \sim \pi_{f'}, \;\; a_{h:h+1} \sim \pi_f,$$

which is how actions are chosen in the definition of $\mathcal{E}(f, \pi_{f'}, h)$ (see Definition 2). Such a decision-making strategy induces a distribution over the following set of variables

$$(s_h, o_h, a_h, r_h, o_{h+1}, a_{h+1}).$$

We use $\mu_{f,f'}$ to denote this distribution, and the subscript emphasizes its dependence on both f and f'. Note that the marginal distribution of s_h only depends on f' and has no dependence on f, which we denote as $\mu_{f'}$. Then, sampling from $\mu_{f,f'}$ is equivalent to the following sampling procedure: (recall that $x_h = (o_h, h)$)

$$\begin{aligned} &s_h \sim \mu_{f'},\ o_h \sim D_{s_h},\ a_h = \pi_f(x_h),\ r_h \sim R(s_h, o_h, a_h), \\ &s_{h+1} \sim \Gamma(s_h, o_h, a_h),\ o_{h+1} \sim D_{s_{h+1}}, a_{h+1} = \pi_f(x_{h+1}). \end{aligned}$$

That is, we first sample s_h from the marginal $\mu_{f'}$, and then sample the remaining variables conditioned on s_h. Notice that once we condition on s_h, the sampling of the remaining variable has no dependence on f', so we denote the joint distribution over the remaining variables (conditioned on the value of s_h) $\mu_{f|s_h}$.

Finally, we express the factorization of $\mathcal{E}(f, \pi_{f'}, h)$ as follows:

$$\begin{aligned} \mathcal{E}(f, \pi_{f'}, h) &= \mathbb{E}_{\mu_{f,f'}}[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})] \\ &= \mathbb{E}_{s_h \sim \mu_{f'}} \mathbb{E}_{\mu_{f|s_h}}[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})] \\ &= \sum_{s \in \mathcal{S}} \mu_{f'}(s) \cdot \mathbb{E}_{\mu_{f|s}}[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})]. \end{aligned}$$

We define $\nu_h(\cdot)$ and $\xi_h(\cdot)$ explicitly with dimension $M = |\mathcal{S}|$: given f and f', we index the elements of $\nu_h(f')$ and those of $\xi_h(f)$ by $s \in \mathcal{S}$, and let $[\nu_h(f')]_s = \mu_{f'}(s)$, $[\xi_h(f)]_s = \mathbb{E}_{\mu_{f|s}}[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})]$. $\zeta = 2\sqrt{M}$ follows from the fact that $\|\nu_h(f')\|_1 = 1$ and $\|\xi_h(f)\|_\infty \leq 2$. □

B.4. Predictive State Representations

In this subsection we state and prove the formal version of Proposition 4. We first recall the definitions and some basic properties of PSRs, which can be found in Singh et al. (2004); Boots et al. (2011). Consider dynamical systems with discrete and finite observation space $\mathcal{O}$ and action space $\mathcal{A}$. Such systems can be fully specified by moment matrices $P_{\mathcal{T}|\mathcal{H}}$, where $\mathcal{H}$ is a set of *histories* (past events) and $\mathcal{T}$ is a set of *tests* (future events). Elements of $\mathcal{T}$ and $\mathcal{H}$ are sequences of alternating actions and observations, and the entry of $P_{\mathcal{T}|\mathcal{H}}$ indexed by $t \in \mathcal{T}$ on the row and $\tau \in \mathcal{H}$ on the column is $P_{t|\tau}$, the probability that the test t *succeeds* conditioned on a particular past τ. For example, if $t = aoa'o'$, success of t means seeing o and o' in the next two steps after τ is observed, if interventions a and a' *were* to be taken.

Among all such systems, we are concerned about those that have finite *linear dimension*, defined as $\sup_{\mathcal{T},\mathcal{H}} \text{rank}(P_{\mathcal{T}|\mathcal{H}})$. As an example, the linear dimension of a POMDP is bounded by the number of hidden-states. Systems with finite linear dimension have many nice properties, which allow them to be expressed by compact models, namely PSRs. In particular, fixing any $\mathcal{T}$ and $\mathcal{H}$ such that $\text{rank}(P_{\mathcal{T}|\mathcal{H}})$ is equal to the linear dimension (such $(\mathcal{H}, \mathcal{T})$ are called *core* histories and *core* tests), we have:

1. For any history $\tau \in (\mathcal{A} \times \mathcal{O})^*$, the conditional predictions of core tests $P_{\mathcal{T}|\{\tau\}}$ (we also write $P_{\mathcal{T}|\tau}$) is always a *state*, that is, a sufficient statistics of history. This gives rise to the name "predictive state representation".

2. Based on $P_{\mathcal{T}|\tau}$, the conditional prediction of any test t can be computed from a PSR model, parameterized by square matrices $\{B_{ao}\}$ and a vector b_∞ with dimension $|\mathcal{T}|$. Letting $t^{(i)}$ be the i-th (action, observation) pair in t, and $|t|$ be the number of such pairs, the prediction rule is

$$P_{t|\tau} = b_\infty^\top B_{t^{(|t|)}} \cdots B_{t^{(1)}} P_{\mathcal{T}|\tau}. \tag{14}$$

And these parameters can be computed as

$$B_{ao} = P_{\mathcal{T},ao,\mathcal{H}} P^\dagger_{\mathcal{T},\mathcal{H}}\,, \qquad b_\infty^\top = P_{\mathcal{H}}^\top P^\dagger_{\mathcal{T},\mathcal{H}} \tag{15}$$

where

- $P_{\mathcal{T},\mathcal{H}}$ is a matrix whose element indexed by $(t \in \mathcal{T}, \tau \in \mathcal{H})$ is $P_{\tau t|\varnothing}$, where τt is the concatenation of τ and t and $\varnothing$ is the null history.
- $P_{\mathcal{H}} = P_{\{\varnothing\},\mathcal{H}}$.
- $P_{\mathcal{T},ao,\mathcal{H}} = P_{\mathcal{T},\mathcal{H}_{ao}}$, where $\mathcal{H}_{ao} = \{\tau ao : \tau \in \mathcal{H}\}$.

Now we are ready to state and prove the formal version of Proposition 4.

Proposition 9 (Formal version of Proposition 4)**.** *Consider a partially observable system with observation space $\mathcal{O}$, and the induced CDP $(\mathcal{X}, \mathcal{A}, H, P)$ with $x_h = (o_h, h)$. To handle some subtleties, we assume that*

1. *$|\mathcal{O}| < \infty$ (classical PSR results assume discrete observations).*
2. *o_1 is deterministic (PSR trajectories always start with an action), and r_h is a deterministic function of o_{h+1} (reward is usually omitted or assumed to be part of the observation).*

If the linear dimension of the original system is at most L, then with any $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to [0, 1]$, this model admits a Bellman factorization with $M = LK$. Assuming further that the PSR's parameters are non-negative under some choice of core histories and tests $(\mathcal{H}, \mathcal{T})$ of size $|\mathcal{H}| = |\mathcal{T}| = L$, then we have $\zeta \le 2K^2L^3\sqrt{L}/\sigma^3_{\min}$, where $\sigma_{\min}$ is the minimal non-zero singular value of $P_{\mathcal{T},\mathcal{H}}$.

Proof. For any $f, f' \in \mathcal{F}, h \in [H]$, define

1. $\mu_{f',h}$ as the distribution vector over $(a_1, o_2, \ldots, o_{h-1}, a_{h-1}) \in (\mathcal{A} \times \mathcal{O})^{h-2} \times \mathcal{A}$ induced by $a_{1:h-1} \sim \pi_{f'}$. (Recall that o_1 is deterministic.)
2. $P_{2|h-1}$ as a moment matrix whose element with column index $(o_h, a_h, o_{h+1}) \in \mathcal{O} \times \mathcal{A} \times \mathcal{O}$ and row index $(a_1, o_2, \ldots, o_{h-1}, a_{h-1}) \in (\mathcal{A} \times \mathcal{O})^{h-2} \times \mathcal{A}$ is
$$P[o_h, o_{h+1} \parallel a_{h-1}, a_h \mid a_1, o_2, \ldots, o_{h-1}].$$[11]
3. $F_{f,h}$ as a vector whose element indexed by $(o_h, a_h, o_{h+1}) \in \mathcal{O} \times \mathcal{A} \times \mathcal{O}$ is (recall that $x_h = (o_h, h)$ and r_h is function of o_{h+1})
$$\mathbf{1}[a_h \neq \pi_f(x_h)] \big(f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1}))\big).$$

First we verify that

$$\mathcal{E}(f, \pi_{f'}, h) = \mu_{f',h}^\top P_{2|h-1} F_{f,h}.$$

To show this, first observe that $\mu_{f',h}^\top P_{2|h-1}$ is a row vector whose element indexed by (o_h, a_h, o_{h+1}) is

$$P[o_h, o_{h+1} \parallel a_h \mid a_{1:h-1} \sim \pi_{f'}].$$

Multiplied by $F_{f,h}$, we further get

$$\mathbb{E}[f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1})) \mid a_{1:h-1} \sim \pi_{f'}, a_h \sim \pi_f] = \mathcal{E}(f, \pi_{f'}, h).$$

[11] PSR literature often emphasizes the intervention aspect of the actions in tests via the uses "$\parallel$" symbol; mathematically they can be treated as the conditioning operator in most cases.

Next, we explicitly construct $\xi_h(f)$ and $\nu_h(f')$ by factorizing $P_{2|h-1} = P_1 \times P_2$, where both P_1 and P_2 have no dependence on either f or f'. Recall that for PSRs, any history $(a_1, o_2, \ldots, o_{h-1})$ has sufficient statistics $P_{\mathcal{T}|a_1,o_2,\ldots,o_{h-1}}$, that is a vector of predictions over the selected core tests $\mathcal{T}$ conditioned on the observed history. P_1 consists of row vectors of length LK, and for the row indexed by $(a_1, o_2, \ldots, o_{h-1}, a_{h-1})$ the vector is

$$\text{Pad}_{a_{h-1}}\big(P^\top_{\mathcal{T}|a_1,o_2,\ldots,o_{h-1}}\big),$$

where $\text{Pad}_a(\cdot)$ is a function that takes a L-dimensional vector, puts it in the a-th block of a vector of length LK, and fills the remaining entries with 0.

We construct P_2 to be a matrix whose column vector indexed by (o_h, a_h, o_{h+1}) is

$$\begin{bmatrix} B^\top_{a^{(1)},o_h} B^\top_{a_h,o_{h+1}} b_\infty \\ \cdots \\ B^\top_{a^{(K)},o_h} B^\top_{a_h,o_{h+1}} b_\infty \end{bmatrix},$$

where $\mathcal{A} = \{a^{(1)}, \ldots, a^{(K)}\}$. It is easy to verify that $P_{2|h-1} = P_1 \times P_2$ by recalling the prediction rules of PSRs in Eq. (14):

$$\begin{aligned} P[o_h, o_{h+1} \parallel a_{h-1}, a_h \mid a_1, o_2, \ldots, o_{h-1}] &= b^\top_\infty B_{a_h,o_{h+1}} B_{a_{h-1},o_h} P_{\mathcal{T}|a_1,o_2,\ldots,o_{h-1}} \\ &= P^\top_{\mathcal{T}|a_1,o_2,\ldots,o_{h-1}} (B^\top_{a_{h-1},o_h} B^\top_{a_h,o_{h+1}} b_\infty). \end{aligned}$$

Given this factorization, we can write

$$\mathcal{E}(f, \pi_{f'}, h) = (\mu^\top_{f',h} P_1) \times (P_2 F_{f,h}).$$

So we let $\nu_h(f') = P_1^\top \mu_{f',h}$ and $\xi_h(f) = P_2 F_{f,h}$. It remains to be shown that we can bound their norms. Notice that the entries of a state vector $P_{\mathcal{T}|(\cdot)}$ are predictions of probabilities, so $\|P_1\|_\infty \le 1$. Since $\mu_{f',h}$ is a probability vector, its dot product with every column in P_1 is bounded by 1, hence $\|\nu_h(f')\|_2 \le \sqrt{LK}$.

At last, we consider bounding the norm of $P_2 F_{f,h}$. We upper bound each entry of $P_2 F_{f,h}$ by providing an ℓ_1 bound on the row vectors of P_2, and then applying the Hölder's inequality with $\|F_{f,h}\|_\infty \le 2$. Since we assumed that all model parameters of the PSRs are non-negative, P_2 is a non-negative matrix, and bounding the ℓ_1 norm of its row vectors is equivalent to bounding each entry of the vector $P_2\,\mathbf{1}$, where $\mathbf{1}$ is an all-1 vector. This vector is equal to

$$P_2\,\mathbf{1} = \begin{bmatrix} \sum_{(o_h,a_h,o_{h+1})} B^\top_{a^{(1)},o_h} B^\top_{a_h,o_{h+1}} b_\infty \\ \cdots \\ \sum_{(o_h,a_h,o_{h+1})} B^\top_{a^{(K)},o_h} B^\top_{a_h,o_{h+1}} b_\infty \end{bmatrix} = \begin{bmatrix} \left(\sum_{o_h} B^\top_{a^{(1)},o_h}\right) \left(\sum_{(a_h,o_{h+1})} B^\top_{a_h,o_{h+1}}\right) b_\infty \\ \cdots \\ \left(\sum_{o_h} B^\top_{a^{(K)},o_h}\right) \left(\sum_{(a_h,o_{h+1})} B^\top_{a_h,o_{h+1}}\right) b_\infty \end{bmatrix}. \tag{16}$$

Since we care about the ℓ_∞ norm of this vector, we can bound the ℓ_∞ norm of each component vector. Using the PSR learning equations, we have

$$\sum_{a,o} B_{ao} = \sum_{a,o} P_{\mathcal{T},ao,\mathcal{H}} P^\dagger_{\mathcal{T},\mathcal{H}} = \left(\sum_{a,o} P_{\mathcal{T},ao,\mathcal{H}}\right) P^\dagger_{\mathcal{T},\mathcal{H}}.$$

Note that for any fixed $a = a^{(i)}$, every entry of $\sum_o P_{\mathcal{T},ao,\mathcal{H}}$ is the probability that the event $t \in \mathcal{T}$ happens after $h \in \mathcal{H}$ happens with a one step delay in the middle, where a is intervened in that delayed time step. Such entries are predicted probabilities of events, hence lie in $[0, 1]$. Consequently, $\|\sum_{a,o} P_{\mathcal{T},ao,\mathcal{H}}\|_\infty \le K$, and we can upper bound the matrix ℓ_2 norm by Frobenius norm: $\|\sum_{a,o} P_{\mathcal{T},ao,\mathcal{H}}\|_2 \le \|\sum_{a,o} P_{\mathcal{T},ao,\mathcal{H}}\|_F \le KL$. Hence,

$$\left\|\sum_{a,o} B_{ao}\right\|_2 \le \left\|\sum_{a,o} P_{\mathcal{T},ao,\mathcal{H}}\right\|_2 \cdot \left\|P^\dagger_{\mathcal{T},\mathcal{H}}\right\|_2 \le KL/\sigma_{\min}.$$

Using a similar argument, for any fixed $a = a^{(i)}$, $\|\sum_o B_{ao}\|_2 \leq L/\sigma_{\min}$. We also recall the definition of b_∞ and bound its norm similarly:

$$\|b_\infty\|_2 = \left\|P_{\mathcal{H}}^\top P_{\mathcal{T},\mathcal{H}}^\dagger\right\|_2 \leq \sqrt{L}/\sigma_{\min}.$$

Finally, we have

$$\begin{aligned}
\|P_2\,\mathbf{1}\|_\infty &\leq \max_{a\in\mathcal{A}} \left\| \left(\sum_{o_h} B_{a,o_h}^\top\right) \left(\sum_{(a_h,o_{h+1})} B_{a_h,o_{h+1}}^\top\right) b_\infty \right\|_\infty && \text{(Eq. (16))} \\
&\leq \max_{a\in\mathcal{A}} \left\| \left(\sum_{o_h} B_{a,o_h}^\top\right) \left(\sum_{(a_h,o_{h+1})} B_{a_h,o_{h+1}}^\top\right) b_\infty \right\|_2 \\
&\leq \left(\max_{a\in\mathcal{A}} \left\|\sum_o B_{ao}\right\|_2\right) \left\|\sum_{a,o} B_{ao}\right\|_2 \|b_\infty\|_2 \leq KL^2\sqrt{L}/\sigma_{\min}^3.
\end{aligned}$$

So each row of P_2 has ℓ_1 norm bounded by the above expression. Applying Hölder's inequality we have each entry of $P_2 F_{f,h}$ bounded by $2KL^2\sqrt{L}/\sigma_{\min}^3$, hence $\|\xi_h(f)\|_2 = \|P_2 F_{f,h}\|_2 \leq 2L^3K\sqrt{K}/\sigma_{\min}^3$. Combined with the bound on $\|\nu_h(f')\|_2$ the proposition follows. □

B.5. Linear Quadratic Regulators

In this subsection we prove that Linear Quadratic Regulators (LQR) (See e.g., Anderson & Moore (2007) for a standard reference) admit Bellman factorization with low Bellman rank. We study a finite-horizon, discrete-time LQR, governed by the equations:

$$x_1 = \epsilon_0, \qquad x_{h+1} = Ax_h + Ba_h + \epsilon_h, \qquad \text{and} \qquad c_h = x_h^\top Q x_h + a_h^\top a_h + \tau_h,$$

where $x_h \in \mathbb{R}^d$, $a_h \in \mathbb{R}^K$ and the noise variables are centered with $\mathbb{E}[\epsilon_h \epsilon_h^\top] = \Sigma$, and $\mathbb{E}\tau_h^2 = \sigma^2$. We operate with *costs* c_h, and the goal is to minimize the cumulative cost. We assume that all parameters $A, B, \Sigma, Q, \sigma^2$ are bounded in spectral norm by some $\Theta \geq 1$, that $\lambda_{\min}(B^\top B) \geq \kappa > 0$, and that Q is strictly positive definite. Other formulations of LQR replace $a_h^\top a_h$ in the cost with $a_h^\top R a_h$ for a positive definite matrix R, which can be accounted for by a change of variables. Generalization to non-stationary parameters is straightforward.

This model describes an MDP with continuous state and action spaces, and the corresponding CDP has context space $\mathbb{R}^d \times [H]$, although we always explicitly write both parts of the context in this section. It is well known that in a discrete time LQR, the optimal policy is a non-stationary linear policy $\pi^\star(x,h) = P_{\star,h}x$ (Anderson & Moore, 2007), where $P_{\star,h} \in \mathbb{R}^{K\times d}$ is an h-dependent control matrix. Moreover, if all of the parameters are known to have spectral norm bounded by Θ then the optimal policy has matrices with bounded spectral norm as well, as we will see in the proof.

The arguments for LQR use decoupled policy and value function classes as in Appendix A.2. We use a policy class and value function class defined below for parameters B_1, B_2, B_3 that we set in the proof.

$$\Pi = \{\pi_{\vec{P}} : \pi_{\vec{P}}(x,h) = P_h x, \vec{P} \in \prod_{i=1}^H \mathbb{R}^{K\times d}, \|P_h\|_2 \leq B_1\}$$

$$\mathcal{G} = \{f_{\vec{\Lambda},\vec{O}} : f_{\vec{\Lambda},\vec{O}}(x,h) = x^\top \Lambda_h x + O_h, \vec{\Lambda} \in \prod_{i=1}^H \mathbb{R}^{d\times d}, \|\Lambda_h\|_2 \leq B_2, \vec{O} \in \mathbb{R}^H, |O_h| \leq B_3\}$$

The policy class consists of linear non-stationary policies, while the value functions are nonstationary quadratics with constant offset.

Proposition 10 (Formal version of Proposition 5)**.** *Consider an LQR under the assumptions above.*

Let $\mathcal{G}$ be a class of non-stationary quadratic value functions with offsets and let Π be a class of linear non-stationary policies, defined above. Then, at level h, for any (π, g) pair and any roll-in policy $\pi' \in \Pi$, the average Bellman error can

be written as

$$\mathcal{E}(g, \pi, \pi', h) = \langle \xi_h(\pi, g), \nu_h(\pi') \rangle,$$

where $\nu, \xi \in \mathbb{R}^{d^2+1}$. *If* $\Pi, \mathcal{G}$ *are defined as above with bounds* B_1, B_2, B_3 *and if all problem parameters have spectral norm at most* Θ*, then*

$$\|\xi_h(\pi, g)\|_2^2 \leq d\left(B_2 + \Theta + B_1^2 - (\Theta + \Theta B_1)^2 B_2\right) + 4B_3^2 + d^2\Theta^2 B_2^2$$
$$\|\nu_h(\pi')\|_2^2 \leq d^{H+1}\Theta(\Theta B_1)^{2H} + 1.$$

Hence, the problem admits Bellman factorization with Bellman rank at most $d^2 + 1$ *and* ζ *that is exponential in* H *but polynomial in all other parameters. Moreover, if we set* B_1, B_2, B_3 *as,*

$$B_1 = \Theta^2/\kappa, B_2 = \left(\frac{6\Theta^6}{\kappa^2}\right)^H \Theta, B_3 = \left(\frac{6\Theta^6}{\kappa^2}\right)^H dH\Theta^2,$$

then the optimal policy and value function belong to $\Pi, \mathcal{G}$ *respectively.*

We prove the proposition in several components. First, we study the relationship between policies and value functions, showing that linear policies induce quadratic value functions. Then, we turn to the structure of the optimal policy, showing that it is linear. Next, we derive bounds on the parameters B_1, B_2, B_3, which ensure that the optimal policy and value function belong to $\Pi, \mathcal{G}$. Lastly, we demonstrate the Bellman factorization.

The next lemma derives a relationship between linear policies and quadratic value functions.

Lemma 2. *If* π *is a linear non-stationary policy,* $\pi_h(x) = P_{\pi,h}x$*, then* $V^\pi(x, h) = x^\top \Lambda_{\pi,h} x + O_{\pi,h}$ *where* $\Lambda_{\pi,h} \in \mathbb{R}^{d\times d}$ *depends only on* π *and* h *and* $O_{\pi,h} \in \mathbb{R}$*. These parameters are defined inductively by,*

$$\Lambda_{\pi,H} = Q + P_{\pi,H}^\top P_{\pi,H}, \qquad O_{\pi,H} = 0$$
$$\Lambda_{\pi,h} = Q + P_{\pi,h}^\top P_{\pi,h} + (A + BP_{\pi,h})^\top \Lambda_{\pi,h+1}(A + BP_{\pi,h})$$
$$O_{\pi,h} = \mathrm{tr}(\Lambda_{\pi,h+1}\Sigma) + O_{\pi,h+1},$$

where we recall that Σ *is the covariance matrix of the* ϵ_h *random variables.*

Proof. The proof is by backward induction on h, starting from level H. Clearly,

$$V^\pi(x, H) = x^\top Q x + \pi_H(x)^\top \pi_H(x) = x^\top Q x + x^\top P_{\pi,H}^\top P_{\pi,H} x \triangleq x^\top \Lambda_{\pi,H} x,$$

so $V^\pi(\cdot, H)$ is a quadratic function.

For the inductive step, consider level h and assume that for all $x, V^\pi(x, h+1) = x^\top \Lambda_{\pi,h+1} x + O_{\pi,h+1}$. Then, expanding definitions

$$\begin{aligned}
V^\pi(x, h) &= x^\top Q x + \pi_h(x)^\top \pi_h(x) + \mathbb{E}_{x'\sim(x,\pi_h(x))} V^\pi(x', h+1) \\
&= x^\top Q x + x^\top P_{\pi,h}^\top P_{\pi,h} x + \mathbb{E}_{x'\sim(x,\pi_h(x))}\left[(x')^\top \Lambda_{\pi,h+1}(x') + O_{\pi,h+1}\right] \\
&= x^\top Q x + x^\top P_{\pi,h}^\top P_{\pi,h} x + \mathbb{E}_{\epsilon_h}\left[(Ax + B\pi_h(x) + \epsilon_h)^\top \Lambda_{\pi,h+1}(Ax + B\pi_h(x) + \epsilon_h) + O_{\pi,h+1}\right] \\
&= x^\top Q x + x^\top P_{\pi,h}^\top P_{\pi,h} x + \mathbb{E}_{\epsilon_h}\left[(Ax + BP_{\pi,h}x + \epsilon_h)^\top \Lambda_{\pi,h+1}(Ax + BP_{\pi,h}x + \epsilon_h) + O_{\pi,h+1}\right] \\
&= x^\top Q x + x^\top P_{\pi,h}^\top P_{\pi,h} x + x^\top (A + BP_{\pi,h})^\top \Lambda_{\pi,h+1}(A + BP_{\pi,h})x + \mathbb{E}_{\epsilon_h} \epsilon_h^\top \Lambda_{\pi,h+1}\epsilon_h + O_{\pi,h+1} \\
&= x^\top Q x + x^\top P_{\pi,h}^\top P_{\pi,h} x + x^\top (A + BP_{\pi,h})^\top \Lambda_{\pi,h+1}(A + BP_{\pi,h})x + \mathrm{tr}(\Lambda_{\pi,h+1}\Sigma) + O_{\pi,h+1}.
\end{aligned}$$

Thus, setting

$$\Lambda_{\pi,h} = Q + P_{\pi,h}^\top P_{\pi,h} + (A + BP_{\pi,h})^\top \Lambda_{\pi,h+1}(A + BP_{\pi,h})$$
$$O_{\pi,h} = \mathrm{tr}(\Lambda_{\pi,h+1}\Sigma) + O_{\pi,h+1},$$

we have shown that $V^\pi(x, h)$ is a quadratic function of x. □

The next lemma shows that the optimal policy is linear.

Lemma 3. *In an LQR, the optimal policy $\pi^\star$ is a non-stationary linear policy given by $\pi^\star(x,h) = P_{\star,h}x$, with parameter matrices $P_{\star,h} \in \mathbb{R}^{K\times d}$ at each level h. The optimal value function $V^\star$ is a non-stationary quadratic function given by $V^\star(x,h) = x^\top \Lambda_{\star,h} x + O_{\star,h}$ with parameter matrix $\Lambda_{\star,h} \in \mathbb{R}^{d\times d}$ and offset $O_{\star,h} \in \mathbb{R}$. The optimal parameters are defined recursively by,*

$$\begin{aligned}
P_{\star,H} &= 0 \qquad \Lambda_{\star,H} = Q \qquad O_{\star,H} = 0 \\
P_{\star,h} &= (I + B^\top \Lambda_{\star,h+1} B)^{-1} B^\top \Lambda_{\star,h+1} A \\
\Lambda_{\star,h} &= Q + P_{\star,h}^\top P_{\star,h} + (A + BP_{\star,h})^\top \Lambda_{\star,h+1}(A + BP_{\star,h}) \\
O_{\star,h} &= \mathrm{tr}(\Lambda_{\star,h+1}\Sigma) + O_{\star,h+1}.
\end{aligned}$$

Proof. We explicitly calculate the optimal policy $\pi_\star$ and demonstrate that it is linear. Then we instantiate these matrices in Lemma 2 to compute the optimal value function.

For the optimal policy, we use backward induction on H. At the last level, we have,

$$\pi^\star(x,H) = \operatorname*{argmin}_a \{x^\top Q x + a^\top a\} = 0.$$

Recall that we are working with costs, so the optimal policy minimizes the expected cost. Thus $P_{\star,H} = 0 \in \mathbb{R}^{K\times d}$ and $\pi^\star(x,H)$ is a linear function of x.

Plugging into Lemma 2 the value function has parameters

$$\Lambda_{\star,H} = Q, \qquad O_{\star,H} = 0.$$

For the induction step, assume that $\pi^\star(x,h+1) = P_{\star,h+1}x$ is linear and $V^\star(x,h+1)$ is quadratic with parameter $\Lambda_{\star,h+1} \succ 0$ and $O_{\star,h+1}$. We then have,

$$\begin{aligned}
\pi^\star(x,h) &= \operatorname*{argmin}_a x^\top Q x + a^\top a + \mathbb{E}_{x'\sim(x,a)} V^\star(x', h+1) \\
&= \operatorname*{argmin}_a x^\top Q x + a^\top a + \mathbb{E}_{\epsilon_h}(Ax + Ba + \epsilon_h)^\top \Lambda_{\star,h+1}(Ax + Ba + \epsilon_h) + O_{\star,h+1} \\
&= \operatorname*{argmin}_a a^\top (I + B^\top \Lambda_{\star,h+1} B) a + 2\langle \Lambda_{\star,h+1} Ax, Ba\rangle.
\end{aligned}$$

This follows by applying definitions and eliminating terms that are independent of a. Since $R, \Lambda_{\star,h+1} \succ 0$ by assumption and using the inductive hypothesis we can analytically minimize. Setting the derivative equal to zero gives,

$$a = (I + B^\top \Lambda_{\star,h+1} B)^{-1} B^\top \Lambda_{\star,h+1} Ax.$$

Thus $P_{\star,h} = (I + B^\top \Lambda_{\star,h+1} B)^{-1} B^\top \Lambda_{\star,h+1} A$. □

As a consequence, we can now derive bounds on the policy and value function parameters. Recall that we assume that all system parameters are bounded in spectral norm by $\Theta \geq 1$ and that $(B^\top B)^{-1}$ has minimum eigenvalue at least κ.

Corollary 1. *With Θ and κ defined above, we have*

$$\|P_{\star,h}\|_f \leq \frac{\Theta^2}{\kappa}, \qquad \|\Lambda_{\star,h}\| \leq \left(\frac{6\Theta^6}{\kappa^2}\right)^{H-h} \Theta, \qquad |O_{\star,h}| \leq (H-h)\left(\frac{6\Theta^6}{\kappa^2}\right)^{H-h} d\Theta^2.$$

Proof. Again we proceed by backward induction, using Lemma 3. Clearly $\|P_{\star,H}\|_F = 0$, $\|\Lambda_{\star,H}\|_F \leq \Theta$, $|O_{\star,H}| = 0$.

For the inductive step we can actually compute $P_{\star,h}$ without any assumption on $\Lambda_{\star,h+1}$, except for the fact that it is symmetric positive definite, which follows from Lemma 3. First, we consider just the matrix $B^\top \Lambda_{\star,h+1} A$. Diagonalizing $\Lambda_{\star,h+1} = U^\top D U$ where U is orthonormal and D is diagonal, gives,

$$\begin{aligned}
B^\top \Lambda_{\star,h+1} A &= (UB)^\top D (UA) = (UB)^\top D(UB)(B^\top U^\top UB)^{-1}(UB)^\top (UA) \\
&= (UB)^\top D(UB)(B^\top B)^{-1} B^\top A = B^\top \Lambda_{\star,h+1} \Pi_B A.
\end{aligned}$$

Here $\Pi_B = B(B^\top B)^{-1}B^\top$ is an orthogonal projection operator. This derivation uses the fact that since $(UB)^\top D$ has rows in the column space of UB, we can right multiply by the projector onto UB. We also use that $U^\top U = I$ since U has orthonormal rows and columns.

Thus, by the submultiplicative property of the spectral norm, we obtain

$$\begin{aligned}\|(I + B^\top \Lambda_{\star,h+1} B)^{-1} B^\top \Lambda_{\star,h+1} A\|_2 &\leq \|(I + B^\top \Lambda_{\star,h+1} B)^{-1} B^\top \Lambda_{\star,h+1} B\|_2 \|(B^\top B)^{-1} B^\top A\|_2 \\ &\leq \|(B^\top B)^{-1} B^\top A\|_2 \leq \Theta^2/\kappa.\end{aligned}$$

Here κ is a lower bound on the minimum eigenvalue of $B^\top B$.

Using this bound on $\|P_{\star,h}\|$, we can now bound the optimal value function as

$$\|\Lambda_{\star,h}\| \leq \Theta + \Theta^4/\kappa^2 + (\Theta + \Theta^3/\kappa)^2 \|\Lambda_{\star,h+1}\| \leq 6\Theta^6/\kappa^2 \|\Lambda_{\star,h+1}\|.$$

The last bound uses the fact we apply a bound for $\|\Lambda_{\star,h+1}\|_2$ that is larger than one, so the last term dominates. We also use the inequalities $\Theta^2/\kappa \geq 1$ and $\Theta \geq 1$. This recurrence yields

$$\|\Lambda_{\star,h}\|_2 \leq \left(\frac{6\Theta^6}{\kappa^2}\right)^{H-h} \Theta.$$

A naive upper bound on $O_{\star,h}$ gives,

$$O_{\star,h} \leq \|\Lambda_{\star,h+1}\| \operatorname{tr}(\Sigma) + |O_{\star,h+1}| \leq (H-h)\left(\frac{6\Theta^6}{\kappa^2}\right)^{H-h} d\Theta^2. \qquad \square$$

The final component of the proposition is to demonstrate the Bellman factorization.

Proof of Proposition 10. Fix h and a value function g parametrized by matrices Λ and offset O at time h and Λ', O' at time $h+1$. Also fix π which uses operator P_π at time h.

$$\begin{aligned}\mathcal{E}(\pi, g, \pi', h) &= \mathbb{E}_{x\sim(\pi',h)} x^\top \Lambda x + O - x^\top Q x - x^\top P_\pi^\top P_\pi x - \mathbb{E}_{x'\sim(x,\pi(x))}(x')^\top \Lambda' x' + O' \\ &= \mathbb{E}_{x\sim(\pi',h)} x^\top \Lambda x + O - x^\top Q x - x^\top P_\pi^\top P_\pi x - \mathbb{E}_\epsilon (Ax + BP_\pi x + \epsilon)^\top \Lambda' (Ax + BP_\pi x + \epsilon) + O' \\ &= \operatorname{tr}\left[\left(\Lambda - Q - P_\pi^\top P_\pi - (A + BP_\pi)^\top \Lambda' (A + BP_\pi)\right) \mathbb{E}_{x\sim(\pi',h)} x x^\top\right] + O - O' - \operatorname{tr}(\Lambda\Sigma).\end{aligned}$$

Thus we may write $\xi_h(\pi, g) = \operatorname{vec}(\Lambda - Q - P_\pi^\top P_\pi - (A + BP_\pi)^\top \Lambda' (A + BP_\pi))$ in the first d^2 coordinates and $O - O' - \operatorname{tr}(\Lambda\Sigma)$ in the last coordinate. We also write $\nu_h(\pi') = \operatorname{vec}(\mathbb{E}_{x\sim(\pi',h)} x x^\top)$ in the first d^2 coordinates and 1 in the last coordinate.

The norm bound on ξ is straightforward, since all terms in its decomposition have an exponential in H bound.

For ν, since the distribution is based on applying a bounded policy π' at level $h-1$ iteration, we can write $x = A\tilde{x} + BP_{\pi'}\tilde{x} + \epsilon$ where $\tilde{x}$ is obtained by rolling in with π' for $h-1$ steps. If $(\pi', h-1)$ denotes the distribution at the previous level, this gives

$$\begin{aligned}\|\mathbb{E}_{x\sim(\pi',h)} x x^\top\|_F &\leq \|\Sigma\|_F + \operatorname{tr}\left((A + BP)^\top (A + BP) \mathbb{E}_{\tilde{x}\sim(\pi',h-1)} \tilde{x}\tilde{x}^\top\right) \\ &\leq \|\Sigma\|_F + d(\Theta + \Theta B_1)^2 \|\mathbb{E}_{\tilde{x}\sim(\pi',h-1)} \tilde{x}\tilde{x}^\top\|_F.\end{aligned}$$

Since at level one we have that the norm is at most $\|\Sigma\|_F$, we obtain a recurrence which produces a bound at level h of

$$\|\mathbb{E}_{x\sim(\pi',h)} x x^\top\|_F \leq \|\Sigma\|_F \sum_{i=1}^{h} d^{i-1}(\Theta + \Theta B_1)^{2(i-1)} \leq \|\Sigma\|_F H d^H (\Theta B_1)^{2H},$$

if $\Theta, B_1 \geq 1$, which is the regime of interest. $\square$

C. Proofs of Main Results

In this section, we provide the main ideas as well as the key lemmas involved in proving Theorem 1. We also show how the lemmas are assembled to prove the theorem. Detailed proofs of the lemmas are in Appendix D.

The proof follows an *explore-or-terminate* argument common to existing sample-efficient RL algorithms. We argue that the optimistic policy chosen in Line 5 of Algorithm 1 is either approximately optimal, or visits a context distribution under which its associated value function has a large Bellman error. This implies that using this policy for exploration leads to learning on a new context distribution. For sample efficiency, we then need to establish that this event cannot happen too many times. This is done by leveraging the Bellman factorization of the process and arguing that the number of times an ϵ sub-optimal policy is found can be no larger than $\tilde{\mathcal{O}}(MH)$. Combining with the number of samples collected for every sub-optimal policy, this immediately yields the PAC learning guarantee.

C.1. Key Lemmas for Theorem 1

We begin by recalling Lemma 1.

Lemma (*Restatement of Lemma 1 from main text for convenience*) With $V_f = \mathbb{E}[f(x_1, \pi_f(x_1))]$, we have

$$V_f - V^{\pi_f} = \sum_{h=1}^{H} \mathcal{E}(f, \pi_f, h). \tag{17}$$

The structure of this lemma is similar to many existing results in RL that upper-bound the loss of following an approximate value function greedily using the function's Bellman errors (e.g., Singh & Yee, 1994). However, most existing results are inequalities that use max-norm relaxations to deal with mismatch in distributions; hence, they are likely to be loose. This lemma, on the other hand, is an equality, thanks to the fact that we are comparing V^{π_f} to V_f, not $V^\star$. As the remaining analysis shows, this simple equation allows us to relate policy loss (the LHS) with the average Bellman error (the RHS) that we use to drive exploration. In particular, this lemma implies an explore-or-terminate behavior for the algorithm.

Lemma 4 (Optimism drives exploration)**.** *Suppose the estimates $\hat{V}_f$ and $\tilde{\mathcal{E}}(f_t, \pi_t, h)$ in Line 2 and 7 always satisfy*

$$|\hat{V}_f - V_f| \le \epsilon/8, \qquad \text{and} \qquad |\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le \frac{\epsilon}{8H} \tag{18}$$

throughout the execution of the algorithm. Assume further that $f^\star$ is never eliminated. Then in any iteration t, one of the following two statements holds:

(i) the algorithm does not terminate and

$$\mathcal{E}(f_t, \pi_t, h_t) \ge \frac{\epsilon}{2H}, \tag{19}$$

(ii) the algorithm terminates and the output policy π_t satisfies $V^{\pi_t} \ge V^\star_{\mathcal{F}} - \epsilon$.

The lemma guarantees that the policy π_t used at iteration t in OLIVE has sufficiently large Bellman error on at least one of the levels, provided that the two conditions in Equation (18) are met. These conditions require that (1) we have reasonably accurate value function estimates from Line 1, and (2) we collect enough samples in Line 6 to form reliable Bellman error estimates under f_t at each level h. The result of Theorem 1 can then be obtained using two further ingredients. First, we need to make sure that the first case in Lemma 4 does not happen too many times. Second, we need to collect enough samples in Lines 1 and 6 to ensure the preconditions in Equation (18). We first establish a bound on the number of iterations using the *Bellman rank* of the problem, before moving on to sample complexity questions.

Lemma 5 (Iteration complexity)**.** *If $\widehat{\mathcal{E}}(f, \pi_t, h_t)$ in Line 13 always satisfies*

$$|\widehat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| \le \phi \tag{20}$$

throughout the execution of the algorithm (ϕ is the threshold in the elimination criterion), then $f^\star$ is never eliminated. Furthermore, for any particular level h, if whenever $h_t = h$ we have

$$|\mathcal{E}(f_t, \pi_t, h_t)| \ge 6\sqrt{M\phi}, \tag{21}$$

then the number of iterations that $h_t = h$ is at most $M \log\left(\frac{\zeta}{2\phi}\right) / \log\frac{5}{3}$.

Precondition (20) simply posits that we collect enough samples for reliable Bellman error estimation in Line 12. Intuitively, since $f^\star$ has no Bellman error, this is sufficient to ensure that it is never eliminated. Precondition (21) is naturally satisfied by the exploration policies π_t given Lemma 4, when ϕ is chosen appropriately according to ϵ, M, and H. Given this, the above lemma bounds the number of iterations at which we can find a large Bellman error at any particular level.

The intuition behind this claim is most clear in the POMDP setting of Proposition 3. In this case, $\nu_h(f')$ in Definition 5 corresponds to the distribution over hidden states induced by $\pi_{f'}$ at level h. At iteration t, the exploration policy π_{f_t} induces such a hidden-state distribution $p = \nu_h(f_t)$ at the chosen level $h = h_t$, which results in the elimination of all functions that have large Bellman error on p. Thanks to the Bellman factorization, this corresponds to the elimination of all f with a large $|p^\top \xi_h(f)|$, where $\xi_h(f)$ is also defined in Definition 5. In this case, it can be easily shown that $\xi_h(f) \in [-2, 2]^M$, so the space of all such vectors $\{\xi_h(f) : f \in \mathcal{F}\}$ at each level h is originally contained in an ℓ_∞ ball in M dimensions with radius 2, and, whenever $h_t = h$, we intersect this set with two parallel halfspaces. Via a geometric argument adapted from Todd (1982), we show that each such intersection reduces the volume of the space by a multiplicative factor of $3/5$. We also show that the volume is bounded from below, hence volume reduction cannot occur indefinitely. Together, these two facts lead to the iteration complexity in Lemma 5. The mathematical techniques used here are analogous to the analysis of the Ellipsoid method in linear programming (see e.g. Bland et al. (1981)).

Finally, we need to ensure that the number of samples collected in each of Lines 1, 6, and 12 of OLIVE can be upper bounded, which yields the overall PAC learning result in Theorem 1. The next three lemmas present precisely the deviation bounds required for this argument. The first two follow from simple applications of Hoeffding's inequality.

Lemma 6 (Deviation bound for $\hat{V}_f$)**.** *With probability at least* $1 - \delta$*,*

$$|\hat{V}_f - V_f| \le \sqrt{\frac{1}{2n_{est}} \log \frac{2N}{\delta}}$$

holds for all $f \in \mathcal{F}$ *simultaneously. Hence, we can set* $n_{est} \ge \frac{32}{\epsilon^2} \log \frac{2N}{\delta}$ *to guarantee that* $|\hat{V}_f - V_f| \le \epsilon/8$*.*

This controls the number of samples required in Line 1.

Lemma 7 (Deviation bound for $\tilde{\mathcal{E}}(f_t, \pi_t, h)$)**.** *For any fixed* f_t*, with probability at least* $1 - \delta$*,*

$$|\widehat{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le 3\sqrt{\frac{1}{2n_{eval}} \log \frac{2H}{\delta}}$$

holds for all $h \in [H]$ *simultaneously. Hence, we can set* $n_{eval} \ge \frac{288H^2}{\epsilon^2} \log \frac{2H}{\delta}$ *to guarantee that* $|\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le \frac{\epsilon}{8H}$*.*

This lemma can be seen as the sample complexity at each iteration in Line 6. Note that no union bound over $\mathcal{F}$ is needed here, since Line 6 only estimates the average Bellman error for a single function, which is fixed before data is collected. Finally, we bound the sample complexity of the learning step.

Lemma 8 (Deviation bound for $\widehat{\mathcal{E}}(f, \pi_t, h_t)$)**.** *For any fixed* π_t *and* h_t*, with probability at least* $1 - \delta$*,*

$$|\widehat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| \le \sqrt{\frac{8K \log \frac{2N}{\delta}}{n}} + \frac{2K \log \frac{2N}{\delta}}{n}$$

holds for all $f \in \mathcal{F}$ *simultaneously. Hence, we can set* $n \ge \frac{32K}{\phi^2} \log \frac{2N}{\delta}$ *to guarantee that* $|\widehat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| \le \phi$ *as long as* $\phi \le 4$*.*

This lemma uses Bernstein's inequality to exploit the small variance of the importance weighted estimates.

C.2. Proof of Theorem 1

Suppose the preconditions of Lemma 4 (Eq. (18)) and Lemma 5 (Eq. (20)) hold; we show them via concentration inequalities later. Applying Lemma 4, in every iteration t before the algorithm terminates,

$$\mathcal{E}(f_t, \pi_t, h_t) \ge \frac{\epsilon}{2H} = 6\sqrt{M}\phi,$$

due to the choice of ϕ. For level $h = h_t$, Eq. (21) is satisfied. According to Lemma 5, the event $h_t = h$ can happen at most $M \log\left(\frac{\zeta}{2\phi}\right) / \log\frac{5}{3}$ times for every $h \in [H]$. Hence, the total number of iterations in the algorithm is at most

$$HM \log\left(\frac{\zeta}{2\phi}\right) / \log\frac{5}{3} = HM \log\left(\frac{6H\sqrt{M}\zeta}{\epsilon}\right) / \log\frac{5}{3}.$$

Now we are ready to apply the concentration inequalities to show that Eq. (18) and (20) hold with high probability. We split the total failure probability δ among the following estimation events:

1. Estimation of $\hat{V}_f$ (Lemma 6; only once): $\delta/3$.
2. Estimation of $\tilde{\mathcal{E}}(f_t, \pi_t, h)$ (Lemma 7; every iteration): $\delta / \left(3HM \log\left(\frac{6H\sqrt{M}\zeta}{\epsilon}\right) / \log\frac{5}{3}\right)$.
3. Estimation of $\widehat{\mathcal{E}}(f, \pi_t, h_t)$ (Lemma 8; every iteration): same as above.

Since these events happen in a particular sequence, the proof actually bounds the probability of these failure events conditioned on all previous events succeeding. This imposes no technical challenge as fresh data is collected for every event, so it effectively reduces to a standard union bound.

Applying Lemmas 6, 7, and 8 with the above failure probabilities, we can verify that the choices of n_{est}, n_{eval}, and n in the algorithm statement satisfy the preconditions of Lemmas 4 and 5. Finally, we upper bound the total number of episodes as

$$\begin{aligned}
&n_{\text{est}} + n_{\text{eval}} \cdot HM \log\left(\frac{6H\sqrt{M}\zeta}{\epsilon}\right) / \log\frac{5}{3} + n \cdot HM \log\left(\frac{6H\sqrt{M}\zeta}{\epsilon}\right) / \log\frac{5}{3} \\
&= \tilde{O}\left(\frac{\log(N/\delta)}{\epsilon^2} + \frac{MH^3}{\epsilon^2}\log(\zeta/\delta) + \frac{M^2H^3K}{\epsilon^2}\log(N\zeta/\delta)\right) = \tilde{O}\left(\frac{M^2H^3K}{\epsilon^2}\log(N\zeta/\delta)\right). \qquad \square
\end{aligned}$$

D. Auxiliary Proofs of the Main Lemmas

In this appendix we give the full proofs of the lemmas sketched in Appendix C. Note that OLIVER (Algorithm 3) with parameters $\theta = 0$ and $\eta = 0$ is precisely OLIVE, and the two analyses are identical. To avoid repetition, in this appendix we analyze OLIVER (Algorithm 3) and prove the versions of the lemmas that can be used for Theorem 4. Readers can easily recover the detailed proofs of the lemmas in Appendix C for OLIVE by letting $\theta = 0$, $\eta = 0$, $\epsilon' = \epsilon$, $f^\star_\theta = f^\star$, $V^\star_{\mathcal{F},\theta} = V^\star_{\mathcal{F}}$.

To facilitate understanding we break up the proofs into 3 parts. The main proofs appear in D.1, and two types of technical lemmas are invoked from there: (1) a series of lemmas that adapt the work of Todd (1982) for the purpose, which are given in D.2; (2) deviation bounds, which are given in D.3.

D.1. Main Proofs

Proof of Lemma 1. Recall from Definition 2 that the average Bellman errors are defined as

$$\mathcal{E}(f, \pi, h) = \mathbb{E}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1}) \;\middle|\; a_{1:h-1} \sim \pi,\ a_{h:h+1} \sim \pi_f\right].$$

Expanding RHS of Eq. (17), we get

$$\sum_{h=1}^{H} \mathbb{E}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1}) \;\middle|\; a_{1:h-1} \sim \pi_f,\ a_{h:h+1} \sim \pi_f\right].$$

Since all H expected values share the same distribution over trajectories, which is the one induced by $a_{1:H} \sim \pi_f$, the above expression is equal to

$$\begin{aligned}
&\sum_{h=1}^{H} \mathbb{E}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1}) \;\middle|\; a_{1:H} \sim \pi_f\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H} \Big(f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})\Big) \;\middle|\; a_{1:H} \sim \pi_f\right] \\
&= \mathbb{E}\left[f(x_1, \pi_f(x_1))\right] - \mathbb{E}\left[r_h \;\middle|\; a_{1:H} \sim \pi_f\right] = V_f - V^{\pi_f}. \qquad \square
\end{aligned}$$

Lemma 9 (Optimism drives exploration, analog of Lemma 4)**.** *If the estimates $\hat{V}_f$ and $\tilde{\mathcal{E}}(f_t, \pi_t, h)$ in Line 3 and 8 of Algorithm 3 always satisfy*

$$|\hat{V}_f - V_f| \le \epsilon'/8, \qquad |\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le \frac{\epsilon'}{8H} \tag{22}$$

throughout the execution of the algorithm (recall that ϵ' is defined on Line 1), and $f^\star_\theta$ is never eliminated, then in any iteration t, either the algorithm does not terminate and

$$\mathcal{E}(f_t, \pi_t, h_t) \ge \frac{\epsilon'}{2H}, \tag{23}$$

or the algorithm terminates and the output policy π_t satisfies $V^{\pi_t} \ge V^\star_{\mathcal{F},\theta} - \epsilon' - H\theta$.

Proof. Eq. (23) follows directly from the termination criterion and Eq. (22). Suppose the algorithm terminates in iteration t. Let $f_{\max} := \operatorname{argmax}_{f \in \mathcal{F}_{t-1}} V_f$, and we have

$$\begin{aligned}
V^{\pi_t} &= V_{f_t} - \sum_{h=1}^{H} \mathcal{E}(f_t, \pi_t, h) && \text{(Lemma 1)} \\
&\ge \hat{V}_{f_t} - \sum_{h=1}^{H} \tilde{\mathcal{E}}(f_t, \pi_t, h) - \epsilon'/4 && \text{(Eq. (22))} \\
&\ge \hat{V}_{f_t} - 7\epsilon'/8 && \text{(termination criterion)} \\
&\ge \hat{V}_{f_{\max}} - 7\epsilon'/8 && (f_t \text{ is the maximizer of } \hat{V}_f) \\
&\ge V_{f_{\max}} - \epsilon' \ge V_{f^\star_\theta} - \epsilon' && (f^\star_\theta \text{ is not eliminated}) \\
&\ge V^\star_{\mathcal{F},\theta} - H\theta - \epsilon'. && \text{(Lemma 1)} \quad \square
\end{aligned}$$

The last inequality uses Lemma 1 on $V_{f^\star_\theta}$ and the definition of $V^\star_{\mathcal{F},\theta}$, which is the value of policy $\pi_{f^\star_\theta}$. Lemma 1 relates these two quantities to the average Bellman errors, which are upper bounded by θ at each level since $f^\star_\theta$ is θ-valid.

Lemma 10 (Volumetric argument, analog of Lemma 5)**.** *If $\widehat{\mathcal{E}}(f, \pi_t, h_t)$ in Eq. (10) always satisfies*

$$|\widehat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| \le \phi \tag{24}$$

throughout the execution of the algorithm (ϕ is the threshold in the elimination criterion), then $f^\star_\theta$ is never eliminated. Furthermore, for any particular level h, if whenever $h_t = h$, we have

$$|\mathcal{E}(f_t, \pi_t, h_t)| \ge 3\sqrt{M}(2\phi + \theta + \eta) + \eta, , \tag{25}$$

then the number of iterations that $h_t = h$ is at most

$$M \log \frac{\zeta}{2\phi} / \log \frac{5}{3}. \tag{26}$$

Proof. The first claim that $f_\theta^\star$ is never eliminated follows directly from the fact $|\mathcal{E}(f_\theta^\star, \pi_t, h_t)| \le \theta$ (Definition 8), Eq. (24), and the elimination threshold $\phi + \theta$. Below we prove the second claim.

For any particular level h, suppose $i_1 < \cdots < i_\tau < \cdots < i_{T_h}$ are the iteration indices with $h_t = h$, $\{t : h_t = h\}$ ordered from first to last, and $T_h = |\{t : h_t = h\}|$. For convenience define $i_0 = 0$. The goal is to prove an upper bound on T_h.

Define notations:

- $p_1, \ldots, p_{T_h}$. $p_\tau := \nu_h(f_{i_\tau})$ where $\nu_h(\cdot)$ is given in Definition 10. Recall that f_{i_τ} is the optimistic function used for exploration in iteration $t = i_\tau$.
- $\mathcal{U}(\mathcal{F}_{i_0}), \mathcal{U}(\mathcal{F}_{i_1}), \ldots, \mathcal{U}(\mathcal{F}_{i_{T_h}})$. $\mathcal{U}(\mathcal{F}_{i_\tau}) = \{\xi_h(f) : f \in \mathcal{F}_{i_\tau}\}$ where $\xi_h(f) \in \mathbb{R}^M$ is given in Definition 10.
- $\Psi = \sup_{f \in \mathcal{F}} \|\nu_h(f)\|_2$, and $\Phi = \sup_{f \in \mathcal{F}} \|\xi_h(f)\|_2$. By Definition 10, $\Psi \cdot \Phi \le \zeta$.
- $V_0, V_1, \ldots, V_{T_h}$. $V_0 := \{v : \|v\|_2 \le \Phi\}$, and $V_\tau := \{v \in V_{\tau-1} : |p_\tau^\top v| \le 2\phi + \theta + \eta\}$.
- $B_0, B_1, \ldots, B_{T_h}$. B_τ is a minimum volume enclosing ellipsoid (MVEE) of V_τ.

For every $\tau = 0, \ldots, T_h$, we first show that $\mathcal{U}(\mathcal{F}_{i_\tau}) \subseteq V_\tau$. When $\tau = 0$ this is obvious. For $\tau \ge 1$, we have $\forall f \in \mathcal{F}_{i_\tau}$,

$$|\mathcal{E}(f, \pi_{f_{i_\tau}}, h)| \le 2\phi + \theta.$$

by the elimination criterion and Eq. (24). By Definition 10, this implies that, $\forall v \in \mathcal{U}(\mathcal{F}_{i_\tau})$,

$$|p_\tau^\top v| \le 2\phi + \theta + \eta,$$

so $\mathcal{U}(\mathcal{F}_{i_\tau}) \subseteq V_\tau$.

Next we show that $\exists v \in V_{\tau-1}$ such that $|p_\tau^\top v| \ge 3\sqrt{M}(2\phi + \theta + \eta)$. In fact, Eq. (25) and the fact that f_{i_t} was chosen (implying that it survived) implies that this v can be chosen as

$$v = \xi_h(f_{i_\tau}) \in \mathcal{U}(\mathcal{F}_{i_\tau - 1}) \subseteq \mathcal{U}(\mathcal{F}_{i_{\tau-1}}) \subseteq V_{\tau-1}.$$

(The first "$\subseteq$" follows from the fact that $\mathcal{F}_t$ shrinks monotonically in Algorithm 3, since the learning steps between $t = i_{\tau-1} + 1$ and $t = i_\tau - 1$ on other levels can only eliminate functions.) We verify that this v satisfies the desired property, given by Definition 10 and Eq. (25):

$$|p_\tau^\top v| = |\langle \nu_h(f_{i_\tau}), \xi_h(f_{i_\tau}) \rangle| \ge |\mathcal{E}(f_{i_\tau}, \pi_{i_\tau}, h)| - \eta \ge 3\sqrt{M}(2\phi + \theta + \eta).$$

Observing that V_t is centrally symmetric and consequently so is B_t (Todd & Yıldırım, 2007), we apply Lemma 11 and Fact 4 with the variables set to $d := M, B := B_{\tau-1}, \kappa := 3\sqrt{M}(2\phi + \theta + \eta), \gamma := 2\phi + \theta + \eta$. We obtain that

$$\frac{vol(B_+)}{vol(B_{t-1})} \le 0.6,$$

where B_+ is the MVEE of $V'_\tau := \{v \in B_{\tau-1} : |p_\tau^\top v| \le 2\phi + \theta + \eta\}$. Note that $V_\tau = \{v \in V_{\tau-1} : |p_\tau^\top v| \le 2\phi + \theta + \eta\} \subseteq V'_\tau$ given that $V_{\tau-1} \subseteq B_{\tau-1}$. Since B_+ is *an* enclosing ellipsoid of V'_τ, and B_τ is the MVEE of V_τ, we have $vol(B_\tau) \le vol(B_+)$. Altogether we claim that

$$\frac{vol(B_\tau)}{vol(B_{\tau-1})} \le 0.6.$$

This result shows that the volume of B_τ shrinks exponentially with τ. To prove that T_h is small, it suffices to show that the volume of B_0 is not too large, and that of B_{T_h} is not too small. Let c_M be the volume of Euclidean sphere with unit radius in $\mathbb{R}^M$. By definition, $vol(B_0) = c_M(\Phi)^M$.

For $vol(B_{T_h})$, since $\|p_\tau\|_2 \leq \Psi$ always holds, we can guarantee that

$$\begin{aligned} V_T &\supseteq \left\{ q \in \mathbb{R}^M : \bigcap_{p \in \mathbb{R}^M : \|p\|_2 \leq \Psi} |\langle p, q \rangle| \leq 2\phi + \theta + \eta \right\} \\ &\supseteq \left\{ q \in \mathbb{R}^M : \|q\|_2 \leq (2\phi + \theta + \eta)/\Psi \right\} && \text{(Hölder's inequality)} \\ &\supseteq \left\{ q \in \mathbb{R}^M : \|q\|_2 \leq 2\phi/\Psi \right\}. \end{aligned}$$

Hence, $vol(B_{T_h}) \geq c_M \left(2\phi/\Psi\right)^M$, and

$$\frac{c_M (2\phi/\Psi)^M}{c_M(\Phi)^M} \leq \frac{vol(B_{T_h})}{vol(B_0)} = \prod_{t=1}^{T_h} \frac{vol(B_t)}{vol(B_{t-1})} \leq 0.6^{T_h}.$$

Algebraic manipulations give

$$M \log\left(\frac{\Psi\Phi}{2\phi}\right) \geq T_h \log \frac{5}{3}.$$

The second claim of the lemma statement follows by recalling that $\Psi\Phi \leq \zeta$. □

D.2. Lemmas for the Volumetric Argument

We adapt the work of Todd (1982) to derive lemmas that we use in D.1. The main result of this section is Lemma 11. As this section focuses on generic geometric results, we adopt notation more standard for these arguments unlike the notation used in the rest of the paper.

Theorem 5 (Theorem 2 of Todd (1982))**.** *Define $E = \{w \in \mathbb{R}^d : w^\top w \leq 1\}$ and $E_\beta = \{w \in E : |e_1^\top w| \leq \beta\}$ for $0 < \beta \leq d^{-1/2}$. The ellipsoid,*

$$E_+ = \{w \in \mathbb{R}^d \,|\, w^\top (\rho(I - \sigma e_1 e_1^\top))^{-1} w \leq 1\}, \tag{27}$$

is a minimum volume enclosing ellipsoid (MVEE) for E_β if

$$\sigma = \frac{1 - d\beta^2}{1 - \beta^2} \qquad \text{and} \qquad \rho = \frac{d(1 - \beta^2)}{d - 1}.$$

Fact 3. *With E, E_+, σ, ρ as in Theorem 5, we have*

$$\frac{Vol(E_+)}{Vol(E)} = \sqrt{d}\beta \left(\frac{d}{d-1}\right)^{(d-1)/2} \left(1 - \beta^2\right)^{(d-1)/2}. \tag{28}$$

Proof. For convenience, let us define $G = \rho(I - \sigma e_1 e_1^\top)$ so that $E_+ = \{w \in \mathbb{R}^d : w^\top G^{-1} w \leq 1\}$. Notice that E can be obtained from E_+ by the affine transformation $v = G^{-1/2} w$, which means that if $w \in E_+$ then $v = G^{-1/2} w \in E$. Via change of variables this implies that

$$\frac{\text{Vol}(E_+)}{\text{Vol}(E)} = \det(G^{1/2}).$$

The determinant is simply the product of the eigenvalues, which is easy to calculate since G is diagonal,

$$\det(G^{1/2}) = \rho^{(d-1)/2} (\rho(1 - \sigma))^{1/2}.$$

Plugging in the definitions of ρ, σ from Theorem 5 proves the statement. □

Lemma 11. *Consider a closed and bounded set $V \subset \mathbb{R}^d$ and a vector $p \in \mathbb{R}^d$. Let B be any enclosing ellipsoid of V that is centered at the origin, and we abuse the same symbol for the symmetric positive definite matrix that defines the ellipsoid, i.e., $B = \{v \in \mathbb{R}^d : v^\top B^{-1} v \leq 1\}$. Suppose there exists $v \in V$ with $|p^\top v| \geq \kappa$ and define B_+ as the minimum volume enclosing ellipsoid of $\{v \in B : |p^\top v| \leq \gamma\}$. If $\gamma/\kappa \leq 1/\sqrt{d}$, we have*

$$\frac{vol(B_+)}{vol(B)} \leq \sqrt{d}\frac{\gamma}{\kappa} \left(\frac{d}{d-1}\right)^{(d-1)/2} \left(1 - \frac{\gamma^2}{\kappa^2}\right)^{(d-1)/2}. \tag{29}$$

Proof. The first claim is to prove a bound on $p^\top Bp$.

$$\kappa \le |p^\top v| = |p^\top B^{1/2} B^{-1/2} v| \le \sqrt{p^\top Bp}\sqrt{v^\top B^{-1} v} \le \sqrt{p^\top Bp}.$$

The last inequality applies since $v \in B$ so that $v^\top B^{-1} v \le 1$. Now we proceed to work with the ellipsoids, let $L = \{v : |v^\top p| \le \gamma\}$. Set $B_+ = MVEE(B \bigcap L)$. We apply two translations of the coordinate system so that B gets mapped to the unit ball and so that p gets mapped to αe_1 (i.e. a scaled multiple of the first standard basis vector). The first translation is done by setting $w = B^{-1/2} v$ where w is in the new coordinate system and v is in the old coordinate system. Let $p_1 = B^{1/2} p$ so that we can equivalently write $L = \{w : |w^\top p_1| \le \gamma\}$. The second translation maps p_1 to αe_1 via a rotation matrix R such that $RB^{1/2} p = Rp_1 = \alpha e_1$. We also translate w to Rw but this doesn't affect the now spherically symmetric ellipsoid, so we do not change the variable names.

To summarize, after applying the scaling and the rotation, we are interested in $MVEE(I \bigcap \{w : |w^\top e_1| \le \gamma/\alpha\})$ and specifically, since volume ratios are invariant under affine transformation, we have

$$\frac{\mathrm{Vol}(B_+)}{\mathrm{Vol}(B)} = \frac{\mathrm{Vol}(MVEE(I \bigcap \{w : |w^\top e_1| \le \gamma/\alpha\}))}{\mathrm{Vol}(I)}.$$

Here I is the unit ball (i.e. the ellipsoid with identity matrix). Further applying Fact 3, we obtain

$$\frac{\mathrm{Vol}(B_+)}{\mathrm{Vol}(B)} = \sqrt{d}\frac{\gamma}{\alpha}\left(\frac{d}{d-1}\right)^{(d-1)/2}\left(1 - \frac{\gamma^2}{\alpha^2}\right)^{(d-1)/2}.$$

It remains to lower bound α, which is immediate since

$$\alpha = \|RB^{1/2} p\|_2 = \|B^{1/2} p\|_2 \ge \kappa.$$

Substituting this lower bound on α completes the proof. □

Fact 4. *When $\gamma/\kappa = \frac{1}{3\sqrt{d}}$, the RHS of Eq. (29) is less than* 0.6.

Proof. Plugging in the numbers, we have the RHS of Eq. (29) equal to

$$\frac{1}{3}\left(\frac{d}{d-1}\frac{9d-1}{9d}\right)^{(d-1)/2} = \frac{1}{3}\left(1 + \frac{8}{9(d-1)}\right)^{9(d-1)/8 \cdot 4/9} \le \frac{1}{3}\exp(4/9) \le 0.52.$$

Here we used the fact that $(1 + \frac{1}{x})^x$ is monotonically increasing towards e on $x \in [1, \infty)$. □

D.3. Deviation Bounds

In this section we prove the deviation bounds. Note that the statement of the lemmas in this section, which are for OLIVER, coincide with those stated in Appendix C for OLIVE. This is not surprising as the two algorithms draw data and estimate quantities in the same way.

Lemma 12 (Deviation Bound for $\hat{V}_f$)**.** *With probability at least $1 - \delta$,*

$$|\hat{V}_f - V_f| \le \sqrt{\frac{1}{2n_{est}} \log \frac{2N}{\delta}}$$

holds for all $f \in \mathcal{F}$ simultaneously. Hence, we can set $n_{est} \ge \frac{32}{\epsilon^2} \log \frac{2N}{\delta}$ to guarantee that $|\hat{V}_f - V_f| \le \epsilon/8$.

Proof. The bound follows from a straight-forward application of Hoeffding's inequality and the union bound, and we only need to verify that the V_f is the expected value of the $\hat{V}_f$, and the range of the random variables is $[0, 1]$. □

Lemma 13 (Deviation Bound for $\tilde{\mathcal{E}}(f_t, \pi_t, h)$)**.** *For any fixed f_t, with probability at least $1 - \delta$,*

$$|\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le 3\sqrt{\frac{1}{2n_{eval}} \log \frac{2H}{\delta}}$$

holds for all $h \in [H]$ simultaneously. Hence, for any $n_{eval} \ge \frac{288H^2}{\epsilon^2} \log \frac{2H}{\delta}$, with probability at least $1 - \delta$ we have $|\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le \frac{\epsilon}{8H}$.

Proof. This bound is another straight-forward application of Hoeffding's inequality and the union bound, except that the random variables that go into the average have range $[-1, 2]$, and we have to realize that $\tilde{\mathcal{E}}(f_t, \pi_t, h)$ is an unbiased estimate of $\mathcal{E}(f_t, \pi_t, h)$. □

Lemma 14 (Deviation Bound for $\widehat{\mathcal{E}}(f, \pi_t, h_t)$)**.** *For any fixed π_t and h_t, with probability at least $1-\delta$,*

$$|\widehat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| \leq \sqrt{\frac{8K \log \frac{2N}{\delta}}{n}} + \frac{2K \log \frac{2N}{\delta}}{n}$$

holds for all $f \in \mathcal{F}$ simultaneously. Hence, for any $n \geq \frac{32K}{\phi^2} \log \frac{2N}{\delta}$ and $\phi \leq 4$, with probability at least $1-\delta$ we have $|\widehat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| \leq \phi$.

Proof. We first show that $\widehat{\mathcal{E}}(f, \pi_t, h_t)$ is an average of i.i.d. random variables with mean $\mathcal{E}(f, \pi_t, h_t)$. We use μ as a shorthand for the distribution over trajectories induced by $a_1, \ldots, a_{h_t-1} \sim \pi_t, a_h \sim \text{unif}(\mathcal{A})$, which is the distribution of data used to estimate $\widehat{\mathcal{E}}(f, \pi_t, h_t)$. On the other hand, let μ' denote the distribution over trajectories induced by $a_1, \ldots, a_{h_t-1} \sim \pi_t, a_h \sim \pi_f$. The importance weight used in Eq. (10) essentially converts the distribution from μ to μ', hence the expected value of $\widehat{\mathcal{E}}(f, \pi_t, h_t)$ can be written as

$$\begin{aligned}
&\mathbb{E}_\mu \left[K\mathbf{1}[a_h = \pi_f(x_h)] \left(f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1}))\right)\right] \\
&= \mathbb{E}_{\mu'} \left[f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1}))\right] = \mathcal{E}(f, \pi_t, h_t).
\end{aligned}$$

Now, we apply Bernstein's inequality. We first analyze the 2nd-moment of the random variable. Defining $y(x_h, a_h, r_h, x_{h+1}) = f(x_h, a_h) - r_h - f(x_{h+1}, \pi_f(x_{h+1})) \in [-2, 1]$, the 2nd-moment is

$$\begin{aligned}
&\mathbb{E}_\mu \left[\left(K\mathbf{1}[a_h = \pi_f(x_h)] y(x_h, a_h, r_h, x_{h+1})\right)^2\right] \\
&= \Pr_\mu[a_h = \pi_f(x_h)] \cdot \mathbb{E}_\mu \left[\left(Ky(x_h, a_h, r_h, x_{h+1})\right)^2 \,\middle|\, a_h = \pi_f(x_h)\right] + \Pr_\mu[a_h \neq \pi_f(x_h)] \cdot 0 \\
&\leq \frac{1}{K} \mathbb{E}_\mu \left[K^2 \cdot 4 \,\middle|\, a_h = \pi_f(x_h)\right] = 4K.
\end{aligned}$$

Next we check the range of the centered random variable. The uncentered variable lies in $[-2K, K]$, and the expected value is in $[-2, 1]$, so the centered variable lies in $[-2K-1, K+2] \subseteq [-3K, 3K]$. Applying Bernstein's inequality, we have with probability at least $1-\delta$,

$$\begin{aligned}
|\widehat{\mathcal{E}}(f, \pi_t, h_t) - \mathcal{E}(f, \pi_t, h_t)| &\leq \sqrt{\frac{2 \operatorname{Var}\left[K\mathbf{1}[a_h = \pi_f(x_h)] y(x_h, a_h, r_h, x_{h+1})\right] \log \frac{2N}{\delta}}{n}} + \frac{6K \log \frac{2N}{\delta}}{3n} \\
&\leq \sqrt{\frac{8K \log \frac{2N}{\delta}}{n}} + \frac{2K \log \frac{2N}{\delta}}{n}. \qquad \text{(variance is bounded by 2nd-moment)}
\end{aligned}$$

As long as $\frac{2K \log \frac{2N}{\delta}}{n} \leq 1$, the above is bounded by $2\sqrt{\frac{8K \log \frac{2N}{\delta}}{n}}$. The choice of n follows from solving $2\sqrt{\frac{8K \log \frac{2N}{\delta}}{n}} = \phi$ for n, which indeed guarantees that $\frac{2K \log \frac{2N}{\delta}}{n} \leq 1$ as $\phi \leq 4$. □

E. Proofs of Extensions

E.1. Proof for Unknown Bellman Rank (Theorem 2)

Since we assign $\delta' = \frac{\delta}{i(i+1)}$ failure probability to the i-th call of Algorithm 2, the total failure probability is at most

$$\sum_{i=1}^{\infty} \frac{\delta}{i(i+1)} = \delta \sum_{i=1}^{\infty} \left(\frac{1}{i} - \frac{1}{i+1}\right) = \delta.$$

So with probability at least $1-\delta$, all high probability events in the analysis of OLIVE occur for every $i = 1, 2, \ldots$. Note that regardless of whether $M' < M$, we never eliminate $f^\star$ according to Lemma 5. Hence Lemma 4 holds and whenever the algorithm returns a policy it is near-optimal.

While the algorithm returns a near-optimal policy if it terminates, we still must prove that the algorithm terminates. Since when $M' < M$ Eq. (21) and Lemma 10 do not apply, we cannot naively use arguments from the analysis of OLIVE. However, we monitor the number of iterations that have passed in each execution to OLIVE and stop the subroutine when the actual number of iterations exceeds the iteration complexity bound (Lemma 5) to prevent wasting more samples on the wrong M'.

OLIVE is guaranteed to terminate within the sample complexity bound and output near-optimal policy when $M \leq M'$. Since M' grows on a doubling schedule, for the first M' that satisfies $M \leq M'$, we have $M' \leq 2M$ and $i \leq \log_2 M + 1$. Hence, the total number of calls is bounded by $\log_2 M + 1$.

Finally, since the sample complexity bound in Theorem 1 is monotonically increasing in M and $1/\delta$ and the schedule for δ' is decreasing, we can bound the total sample complexity by that of the last call to OLIVE multiplied by the number of calls. The last call to OLIVE has $M' \leq 2M$, and $\delta' = \frac{i(i+1)}{\delta} \leq \frac{(\log_2 M+2)(\log_2 M+1)}{\delta}$, so the sample complexity bound is only affected by factors that are at most logarithmic in the relevant parameters.

E.2. Proofs for Infinite Hypothesis Classes

In this section we prove sample complexity guarantee for using infinite hypothesis classes in Appendix A.3. Recall that we are working with separated policy class Π and V-value function class $\mathcal{G}$, and when running OLIVE any occurrence of $f \in \mathcal{F}$ is replaced appropriately by $(\pi, g) \in \Pi \times \mathcal{G}$. For clarity, we use (π, g) instead of f in the derivations in this section. We assume that the two function classes have finite Natarajan dimension and pseudo dimension respectively.

The key technical step for the sample complexity guarantee is to establish the necessary deviation bounds for infinite classes. Among these deviation bounds, the bound on $\tilde{\mathcal{E}}((\pi_t, g_t), \pi_t, h)$ (Lemma 7) does not involve union bound over $\mathcal{F}$, so it can be reused without modification. The other two bounds need to be replaced by Lemma 15 and 16, stated below. With these lemmas, Theorem 3 immediately follows simply by replacing the deviation bounds.

Definition 11. *Define* $d_\Pi = \max(\textit{Ndim}(\Pi), 6), d_\mathcal{G} = \max(\textit{Pdim}(\mathcal{G}), 6)$*, and* $d = d_\Pi + d_\mathcal{G}$.

Lemma 15. *If*

$$n_{\textit{est}} \geq \frac{8192}{\epsilon^2}\left(d_\mathcal{G} \log \frac{128e}{\epsilon} + \log(8e(d_\mathcal{G}+1)) + \log \frac{1}{\delta}\right), \tag{30}$$

then with probability at least $1-\delta$, $|\hat{V}_{(\pi,g)} - V_{(\pi,g)}| \leq \epsilon/8,\ \forall (\pi, g) \in \Pi \times \mathcal{G}$.

We remark that both the estimate $\hat{V}_{(\pi,g)}$ and population quantity $V_{(\pi,g)}$ are independent of π in the separable case, and hence the sample complexity is independent of d_Π.

Lemma 16. *If*

$$n \geq \frac{1152K^2}{\phi^2}\left(6d \log(2eKd) \log \frac{48eK}{\phi} + \log\left(8e(6d\log(2eKd)+1)\right) + \log \frac{3}{\delta}\right), \tag{31}$$

then for any fixed π_t *and* h_t*, with probability at least* $1-\delta$,

$$|\widehat{\mathcal{E}}((\pi, g), \pi_t, h_t) - \mathcal{E}((\pi, g), \pi_t, h_t)| \leq \phi,\ \forall (\pi, g) \in \Pi \times \mathcal{G}.$$

Proof of Theorem 3. Set the algorithm parameters to:

$$\phi = \frac{\epsilon}{12H\sqrt{M}}, \qquad n_{\text{est}} = \frac{8192}{\epsilon^2}\left(d_{\mathcal{G}} \log \frac{128e}{\epsilon} + \log(8e(d_{\mathcal{G}}+1)) + \log\frac{3}{\delta}\right),$$
$$n_{\text{eval}} = \frac{288H^2}{\epsilon^2}\log\left(\frac{12H^2M\log(6H\sqrt{M}\zeta/\epsilon)}{\delta}\right),$$
$$n = \frac{1152K^2}{\phi^2}\Big(6d\log\Big(2eK\Big)\log\frac{48eK}{\phi} + \log\Big(8e(6d\log(2eKd)+1)\Big) + \log\frac{18HM\log(6H\sqrt{M}\zeta/\epsilon)}{\delta}\Big).$$

The rest of the proof is essentially the same as the proof of Theorem 1, and the sample complexity follows by noticing that $n_{\text{est}} = \tilde{\mathcal{O}}(\frac{d_{\mathcal{G}}+\log(1/\delta)}{\epsilon^2})$ and $n = \tilde{\mathcal{O}}(K^2(d_\Pi + d_{\mathcal{G}} + \log(1/\delta))/\phi^2)$. □

Lemma 15 is a straight-forward application of Corollary 2 introduced in E.2.1 and are not proved separately. In the remainder of this section we prove Lemma 16. Before that, we review some standard definitions and results from statistical learning theory.

E.2.1. Definitions and Basic Lemmas

Notations $\mathcal{X}, x, n, d, \xi$ in this section are used according to conventions in the literature and may not share semantics with the same symbols used elsewhere in this paper.

Definition 12 (VC-Dimension). *Given hypothesis class $\mathcal{H} \subset \mathcal{X} \to \{0,1\}$, its VC-dimension VC-dim($\mathcal{H}$) is defined as the maximal cardinality of a set $X = \{x_1, \dots, x_{|X|}\} \subset \mathcal{X}$ that satisfies $|\mathcal{H}_X| = 2^{|X|}$ (or X is* shattered *by $\mathcal{H}$), where $\mathcal{H}_X$ is the restriction of $\mathcal{H}$ to X, namely $\{(h(x_1), \dots, h(x_{|X|})) : h \in \mathcal{H}\}$.*

Lemma 17 (Sauer's Lemma). *Given hypothesis class $\mathcal{H} \subset \mathcal{X} \to \{0,1\}$ with $d =$ VC-dim($\mathcal{H}$) $< \infty$, we have $\forall X = (x_1, \dots, x_n) \in \mathcal{X}^n$,*

$$|\mathcal{H}_X| \le (n+1)^d.$$

Lemma 18 (Sauer's Lemma for Natarajan dimension (Ben-David et al., 1992; Haussler & Long, 1995)). *Given hypothesis class $\mathcal{H} \subset \mathcal{X} \to \mathcal{Y}$ with Ndim($\mathcal{H}$) $\le d$, we have $\forall X = (x_1, \dots, x_n) \in \mathcal{X}^n$,*

$$|\mathcal{H}_X| \le \left(\frac{ne(K+1)^2}{2d}\right)^d,$$

where $K = |\mathcal{Y}|$.

Definition 13 (Covering number). *Given hypothesis class $\mathcal{H} \subset \mathcal{X} \to \mathbb{R}$, $\epsilon > 0$, $X = (x_1, \dots, x_n) \in \mathcal{X}^n$, the covering number $\mathcal{N}_1(\alpha, \mathcal{H}, X)$ is defined as the minimal cardinality of a set $C \subset \mathbb{R}^n$, such that for any $h \in \mathcal{H}$ there exists $c = (c_1, \dots, c_n) \in C$ where $\frac{1}{n}\sum_{i=1}^n |h(x_i) - c_i| \le \alpha$.*

Lemma 19 (Bounding covering number by pseudo dimension (Haussler, 1995)). *Given hypothesis class $\mathcal{H} \subset \mathcal{X} \to \mathbb{R}$ with Pdim($\mathcal{H}$) $\le d$, we have for any $X \in \mathcal{X}^n$,*

$$\mathcal{N}_1(\alpha, \mathcal{H}, X) \le e(d+1)\left(\frac{2e}{\alpha}\right)^d.$$

Lemma 20 (Uniform deviation bound using covering number (Pollard, 2012); also see Devroye et al. (1996), Theorem 29.1). *Let $\mathcal{H} \subset \mathcal{X} \to [0,b]$ be a hypothesis class, and $(x_1, \dots, x_n)$ be i.i.d. samples drawn from some distribution supported on $\mathcal{X}$. For any $\alpha > 0$,*

$$\Pr\left\{\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n h(x_i) - \mathbb{E}[h(x_1)]\right| > \alpha\right\} \le 8\,\mathbb{E}\left[\mathcal{N}_1\big(\alpha/8, \mathcal{H}, (x_1, \dots, x_n)\big)\right]\exp\left(-\frac{n\alpha^2}{128b^2}\right).$$

Corollary 2 (Uniform deviation bound using pseudo dimension)**.** *Suppose Pdim$(\mathcal{H}) \le d$, then*

$$\Pr\left\{\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}h(x_i)-\mathbb{E}[h(x_1)]\right|>\alpha\right\}\le 8e(d+1)\left(\frac{16e}{\alpha}\right)^{d}\exp\left(-\frac{n\alpha^2}{128b^2}\right).$$

To guarantee that this probability is upper bounded by δ, it suffices to have

$$n\ge\frac{128}{\alpha^2}\left(d\log\frac{16e}{\alpha}+\log(8e(d+1))+\log\frac{1}{\delta}\right).$$

E.2.2. Proof of Lemma 16

The idea is to establish deviation bounds for each of the three terms in the definition of $\widehat{\mathcal{E}}((\pi,g),\pi_t,h_t)$ (Line 13). Each term takes the form of an importance weight multiplied by a real-valued function, and we first show that the function space formed by these products has bounded pseudo dimension. We state this supporting lemma in terms of an arbitrary value-function class $\mathcal{V}$ which might operate on an input space $\mathcal{X}'$ different from the context space $\mathcal{X}$. In the sequel, we instantiate $\mathcal{V}$ and $\mathcal{X}'$ in the the lemma with specific choices to prove the desired results.

Lemma 21. *Let $\mathcal{Y}$ be a label space with $|\mathcal{Y}|=K$, let $\Pi\subseteq\mathcal{X}\to\mathcal{Y}$ be a function class with Natarajan dimension at most $d_\Pi\in[6,\infty)$, and let $\mathcal{V}\subseteq\mathcal{X}'\to[0,1]$ be a class with pseudo dimension at most $d_\mathcal{V}\in[6,\infty)$. The hypothesis class $\mathcal{H}=\{(x,a,x')\mapsto\mathbf{1}[a=\pi(x)]g(x'):\pi\in\Pi,g\in\mathcal{V}\}$ has pseudo dimension Pdim$(\mathcal{H})\le 6(d_\Pi+d_\mathcal{V})\log\left(2eK(d_\Pi+d_\mathcal{V})\right)$.*

Proof. Recall that Pdim$(\mathcal{H})$ = VC-dim$(\mathcal{H}^+)$, so it suffices to show that for any $X=\{(x_1,a_1,x_1',\xi_1),\dots,(x_d,a_d,x_d',\xi_d)\}\in(\mathcal{X}\times\mathcal{A}\times\mathcal{X}'\times\mathbb{R})^d$ where $d=6(d_\Pi+d_\mathcal{V})\log\left(2eK(d_\Pi+d_\mathcal{V})\right)$, $|\mathcal{H}_X^+|<2^d$. Note that since $g(x)\in[0,1]$ for all g,x

$$\begin{aligned}\mathcal{H}^+&=\{(x,a,x',\xi)\mapsto\mathbf{1}\Big[\mathbf{1}[a=\pi(x)]g(x')>\xi\Big]\}\\&=\{(x,a,x',\xi)\mapsto\mathbf{1}[\xi<0]+\mathbf{1}[\xi\ge0]\cdot\mathbf{1}[a=\pi(x)]\cdot\mathbf{1}[g(x')>\xi]\}\end{aligned}$$

For points where $\xi_i<0$, all hypotheses in $\mathcal{H}^+$ produce label 1, so without loss of generality we can assume that $\xi_i\ge0, i=1,\dots,d$.

With a slight abuse of notation, let Π_X denote the restriction of Π to the set of contexts $\{x_1,\dots,x_d\}$ (actions and future contexts $(a_1,x_1'),\dots,(a_d,x_d')$ are ignored since Π does not operate on them), and $\mathcal{V}_X^+$ denote the restriction of $\mathcal{V}^+$ to $\{(x_1',\xi_1),\dots,(x_d',\xi_d)\}$. $\mathcal{H}_X^+$ can be produced by the Cartesian product of Π_X and $\mathcal{V}_X^+$ as follows:

$$\mathcal{H}_X^+=\{(\mathbf{1}[a_1=\alpha_1]\beta_1,\dots,\mathbf{1}[a_d=\alpha_d]\beta_d):(\alpha_1,\dots,\alpha_d)\in\Pi_X,(\beta_1,\dots,\beta_d)\in\mathcal{V}_X^+\}.$$

Therefore, $|\mathcal{H}_X^+|\le|\Pi_X|\,|\mathcal{V}_X^+|$. Recall that Ndim$(\Pi)\le d_\Pi$ and VC-dim$(\mathcal{V}^+)$ = Pdim$(\mathcal{V})\le d_\mathcal{V}$. Applying Lemma 18 and 17:

$$|\mathcal{H}_X^+|\le\left(\frac{de(K+1)^2}{2d_\Pi}\right)^{d_\Pi}(d+1)^{d_\mathcal{V}}.$$

The logarithm of the RHS is

$$\begin{aligned}&d_\Pi\log\left(\frac{de(K+1)^2}{2d_\Pi}\right)+d_\mathcal{V}\log(d+1)<d_\Pi\log(de(K+1)^2)+d_\mathcal{V}\log(d+1)\\&\le d_\Pi\log d+2d_\Pi\log(2eK)+d_\mathcal{V}\log(d+1)\le2(d_\Pi+d_\mathcal{V})\log(2eK)+(d_\Pi+d_\mathcal{V})\log(2d).\end{aligned}$$

It remains to be shown that this is less than $\log(2^d)=d\log2$. Note that

$$d\log2>3(d_\Pi+d_\mathcal{V})(\log(2eK)+\log(d_\Pi+d_\mathcal{V})),$$

so we only need to show that $(d_\Pi+d_\mathcal{V})\log(2d)\le(d_\Pi+d_\mathcal{V})\log(2eK)+3(d_\Pi+d_\mathcal{V})\log(d_\Pi+d_\mathcal{V})$. Now

$$\begin{aligned}&(d_\Pi+d_\mathcal{V})\log(2d)=(d_\Pi+d_\mathcal{V})\Big(\log(12(d_\Pi+d_\mathcal{V}))+\log\log(2eK(d_\Pi+d_\mathcal{V}))\Big)\\&\le2(d_\Pi+d_\mathcal{V})\log(d_\Pi+d_\mathcal{V})+(d_\Pi+d_\mathcal{V})\log\Big(\log(2eK)+\log(d_\Pi+d_\mathcal{V})\Big)\qquad(d_\Pi+d_\mathcal{V}\ge12)\\&\le2(d_\Pi+d_\mathcal{V})\log(d_\Pi+d_\mathcal{V})+(d_\Pi+d_\mathcal{V})\big(\log(2eK)+\log(d_\Pi+d_\mathcal{V})\big).\end{aligned}$$

□

Proof of Lemma 16. Recall that when we are given a policy class Π and separate V-value function class $\mathcal{G}$, for every $\pi \in \Pi, g \in \mathcal{G}$, we instead estimate average Bellman error with

$$\widehat{\mathcal{E}}((\pi, g), \pi_t, h_t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}[a_{h_t}^{(i)} = \pi(x_{h_t}^{(i)})]}{1/K}\Big(g(x_{h_t}^{(i)}) - r_{h_t}^{(i)} - g(x_{h_t+1}^{(i)})\Big).$$

So it suffices to show that the averages of $\mathbf{1}[a_{h_t}^{(i)} = \pi(x_{h_t}^{(i)})]g(x_{h_t}^{(i)})$, $\mathbf{1}[a_{h_t}^{(i)} = \pi(x_{h_t}^{(i)})]r_{h_t}^{(i)}$, $\mathbf{1}[a_{h_t}^{(i)} = \pi(x_{h_t}^{(i)})]g(x_{h_t+1}^{(i)})$ are $\frac{\phi}{3K}$-close to their expectations with probability at least $1 - \delta/3$, respectively. It turns out that, we can use Lemma 21 for all the three terms. For the first and the third terms, we apply Lemma 21 with $\mathcal{V} = \mathcal{G}, \mathcal{X}' = \mathcal{X}$, and obtain the necessary sample size directly from Corollary 2. For the second term, we apply Lemma 21 with $\mathcal{V} = \{x \mapsto x\}, \mathcal{X}' = \mathbb{R}$. Note that in this case $\mathcal{V}$ is a singleton with the only element being the identity function over $\mathbb{R}$, so it is clear that $\mathrm{Pdim}(\mathcal{V}) < 6 \le d_{\mathcal{G}}$, hence the sample size for the other two terms is also adequate for this term. □

E.3. Proofs for OLIVER

Recall that the main lemmas for analyzing OLIVER have been proved in Appendix D.1, so below we directly prove Theorem 4.

Proof of Theorem 4. Suppose the preconditions of Lemma 9 (Eq. (22)) and Lemma 10 (Eq. (24)) hold; we show them by invoking the deviation bounds later. By Lemma 9, when the algorithm terminates, the value of the output policy is at least

$$V^{\star}_{\mathcal{F},\theta} - \epsilon' - H\theta.$$

Recall that $\epsilon' = \epsilon + 2H(3\sqrt{M}(\theta + \eta) + \eta)$ (Line 1), so the suboptimality compared to $V^{\star}_{\mathcal{F},\theta}$ is at most

$$\epsilon + 2H(3\sqrt{M}(\theta + \eta) + \eta) + H\theta \le \epsilon + 8H\sqrt{M}(\theta + \eta),$$

which establishes the suboptimality claim.

It remains to show the sample complexity bound. Applying Lemma 9, in every iteration t before the algorithm terminates,

$$\mathcal{E}(f_t, \pi_t, h_t) \ge \frac{\epsilon'}{2H} = \frac{\epsilon}{2H} + 3\sqrt{M}(\theta + \eta) + \eta = 3\sqrt{M}(2\phi + \theta + \eta) + \eta,$$

thanks to the choice of ϕ and ϵ'. For level $h = h_t$, Eq. (25) is satisfied. According to Lemma 10, the event $h_t = h$ can happen at most $M \log\left(\frac{\zeta}{2\phi}\right) / \log\frac{5}{3}$ times for every $h \in [H]$. Hence, the total number of iterations in the algorithm is at most

$$HM \log\left(\frac{\zeta}{2\phi}\right) / \log\frac{5}{3} = HM \log\left(\frac{6H\sqrt{M}\zeta}{\epsilon}\right) / \log\frac{5}{3}.$$

Now we are ready to apply the deviation bounds to show that Eq. (22) and 20 hold with high probability. We split the total failure probability δ among the following events:

1. Estimation of $\hat{V}_f$ (Lemma 12; only once): $\delta/3$.
2. Estimation of $\tilde{\mathcal{E}}(f_t, \pi_t, h)$ (Lemma 13; every iteration): $\delta / \left(3HM \log\left(\frac{6H\sqrt{M}\zeta}{\epsilon}\right) / \log\frac{5}{3}\right)$.
3. Estimation of $\widehat{\mathcal{E}}(f, \pi_t, h_t)$ (Lemma 14; every iteration): same as above.

Applying Lemma 12, 13, 14 with the above failure probabilities, the choices of $n_{\text{est}}, n_{\text{eval}}, n$ in the algorithm statement satisfy the preconditions of Lemmas 9 and 10. In particular, the choice of n_{est} and n_{eval} guarantee that $|\hat{V}_f - V_f| \le \epsilon/8$ and $|\tilde{\mathcal{E}}(f_t, \pi_t, h) - \mathcal{E}(f_t, \pi_t, h)| \le \epsilon/(8H)$, which are tighter than needed as $\epsilon \le \epsilon'$ (only $\epsilon'/8$ and $\epsilon'/(8H)$ are needed respectively, but tightening these bounds does not improve the sample complexity significantly, so we keep them the same as in Theorem 1 for simplicity). The remaining calculation of sample complexity is exactly the same as in the proof of Theorem 1. □

F. Lower Bounds

F.1. An Exponential Lower Bound

We include a result from Krishnamurthy et al. (2016) to formally show that, without making additional assumptions, the sample complexity of value-based RL for CDPs as introduced in Section 2 has a lower bound of order K^H.

Proposition 11 (Restatement of Proposition 2 in Krishnamurthy et al. (2016))**.** *For any $H, K \in \mathbb{N}$ with $K \geq 2$, and any $\epsilon \in (0, \sqrt{1/8})$, there exists a family of finite-horizon MDPs with horizon H and $|\mathcal{A}| = K$, and a function space $\mathcal{F}$ with $|\mathcal{F}| = K^H$ and a universal constant c, such that $Q^\star \in \mathcal{F}$ for all MDP instances in the family, yet for any algorithm and any $T \leq cK^H/\epsilon^2$, the probability that the algorithm outputs a policy $\hat{\pi}$ with $V^{\hat{\pi}} \geq V^\star - \epsilon$ after collecting T trajectories is at most $2/3$ when the problem instance is chosen from the family by an adversary.*

Proof. The proof relies on the fact that CDPs include MDPs where the state space is arbitrarily large. Each instance of the MDP family is a complete tree with branching factor K and depth H. Transition dynamics are deterministic, and only leaf nodes have non-zero rewards. All leaves give $\text{Ber}(1/2)$ rewards, except for one that gives $\text{Ber}(1/2 + \epsilon)$. Changing the position of the most rewarding leaf node yields a family of K^H MDP instances so collecting optimal Q-value functions forms the desired function class $\mathcal{F}$. Since $\mathcal{F}$ provides no information other than the fact that the true MDP lies in this family, the problem is equivalent to identifying the best arm in a multi-arm bandit with K^H arms, and the remaining analysis follows exactly as in Krishnamurthy et al. (2016). □

F.2. A Polynomial Lower Bound that Depends on Bellman Rank

In this section, we prove a new lower bound for layered episodic MDPs that meet the assumptions we make in this paper.

We first recall some definitions. A layered episodic MDP is defined by a time horizon H, a state space $\mathcal{S}$, partitioned into sets $\mathcal{S}_1, \ldots, \mathcal{S}_H$, each of size at most M, and an action space $\mathcal{A}$ of size K. The system descriptor is replaced with a transition function Γ that associates a distribution over states with each state action pair. More formally, for any $s_h \in \mathcal{S}_h$, and $a \in \mathcal{A}$, $\Gamma(s_h, a) \in \Delta(\mathcal{S}_{h+1})$. The starting state is drawn from $\Gamma_1 \in \Delta(\mathcal{S}_1)$, and all transitions from $\mathcal{S}_H$ are terminal.

There is also a reward distribution R that associates a random reward with each state-action pair. We use $r \sim R(s, a)$ to denote the random instantaneous reward for taking action a at state s. We assume that the cumulative reward $\sum_{h=1}^H r_h \in [0, 1]$, where r_h is the reward received at level h as in Assumption 1.

Observe that this process is a special case of the finite-horizon Contextual Decision Process and moreover, with the set of all value functions $\mathcal{F} = (\mathcal{S} \times \mathcal{A} \to [0, 1])$, admits a Bellman factorization with Bellman rank at most M (by Proposition 1). Thus the upper bounds for PAC learning apply directly to this setting.

We now state the lower bound.

Theorem 6. *Fix $M \geq 4, H, K \geq 2$ and $\epsilon \in (0, \frac{1}{48\sqrt{8}})$. For any algorithm and any $n \leq cMKH/\epsilon^2$, there exists a layered episodic MDP with H layers, M states per layer, and K actions, such that the probability that the algorithm outputs a policy $\hat{\pi}$ with $V(\hat{\pi}) \geq V^\star - \epsilon$ after collecting n trajectories is at most $11/12$. Here $c > 0$ is a universal constant.*

The result precludes a $o(MKH/\epsilon^2)$ PAC-learning sample complexity bound since in this case the algorithm must fail with constant probability. The result is similar in spirit to other lower bounds for PAC-learning MDPs (Dann & Brunskill, 2015; Krishnamurthy et al., 2016), but we are not aware of any lower bound that applies directly to the setting. There are two main differences between this bound and the lower bound due to Dann & Brunskill (2015) for episodic MDPs. First, that bound assumes that the total reward is in $[0, H]$, so the H^2 dependence in the sample complexity is a consequence of scaling the rewards. Second, that MDP is not layered, but instead has M total states shared across all layers. In contrast, our process is layered with M distinct states per layer and total reward bounded in $[0, 1]$. Intuitively, the additional H dependence arises simply from having MH total states.

At a high level, the proof is based on embedding $\Theta(MH)$ independent multi-arm bandit instances into a MDP and requiring that the algorithm identify the best action in $\Omega(MH)$ of them to produce a near-optimal policy. By appealing to a sample complexity lower bound for best arm identification, this implies that the algorithm requires $\Omega(MHK/\epsilon^2)$ samples to identify a near-optimal policy.

We rely on a fairly standard lower bound for best arm identification. We reproduce the formal statement from Krishnamurthy et al. (2016), although the proof is based on earlier lower bounds due to Auer et al. (2002).

Proposition 12. *For any $K \geq 2$ and $\tau \leq \sqrt{1/8}$ and any best arm identification algorithm that produces an estimate $\hat{a}$, there exists a multi-arm bandit problem for which the best arm $a^\star$ is τ better than all others, but $\mathbb{P}[\hat{a} \neq a^\star] \geq 1/3$ unless the number of samples T is at least $\frac{K}{72\tau^2}$.*

In particular, the problem instance used in this lower bound is one where the best arm $a^\star$ has reward $\mathrm{Ber}(1/2+\epsilon)$, while all other arms have reward $\mathrm{Ber}(1/2)$. Our construction embeds precisely these instances into the MDP.

Proof. We construct an MDP with M states per level, H levels, and K actions per state. At each level, we allocate three special states, w_h, g_h, and b_h, for "waiting", "good", and "bad." The remaining $M-3$ "bandit" states are denoted $s_{h,i}, i \in [M-3]$. Each bandit state has an unknown optimal action $a^\star_{h,i}$.

The dynamics are as follows.

- For waiting states w_h, all actions are equivalent and with probability $1-1/H$, they transition to the next waiting state w_{h+1}. With the remaining $1/H$ probability, they transition randomly to one of the bandit state $s_{h+1,i}$ so each subsequent bandit state is visited with probability $\frac{1}{H(M-3)}$.
- For bandit states $s_{h,i}$, the optimal action $a^\star_{h,i}$ transitions to the good state g_{h+1} with probability $1/2+\tau$ and otherwise to the bad state b_{h+1}. All other actions transition to g_{h+1} and b_{h+1} with probability $1/2$. Here τ is a parameter we set toward the end of the proof.
- Good states always transition to the next good state and bad states always transition to bad states.
- The starting state is w_1 with probability $1-1/H$ and $s_{1,i}$ with probability $\frac{1}{H(M-3)}$ for each $i \in [M-3]$.

The reward at all states except g_H is zero, and the reward at g_H is one. Clearly the optimal policy takes actions $a^\star_{h,i}$ for each bandit state, and takes arbitrary actions at the waiting, good, and bad states.

This construction embeds $H(M-3)$ best arm identification problems that are identical to the one used in Proposition 12 into the MDP. Moreover, these problems are independent in the sense that samples collected from one provides no information about any others. Appealing to Proposition 12, for each bandit state (h,i), unless $\frac{K}{72\tau^2}$ samples are collected from that state, the learning algorithm fails to identify the optimal action $a^\star_{h,i}$ with probability at least $1/3$.

After the execution of the algorithm, let B be the set of (h,i) pairs for which the algorithm identifies the correct action. Let C be the set of (h,i) pairs for which the algorithm collects fewer than $\frac{K}{72\tau^2}$ samples. For a set S, we use S^C to denote the complement.

$$\begin{aligned}
\mathbb{E}[|B|] &= \mathbb{E}\left[\sum_{(h,i)} \mathbf{1}[a_{h,i} = a^\star_{h,i}]\right] \\
&\leq ((M-3)H - |C|) + \sum_{(h,i)\in C} \mathbb{E}\mathbf{1}[a_{h,i} = a^\star_{h,i}] \\
&\leq ((M-3)H - |C|) + \frac{2}{3}|C| = (M-3)H - |C|/3
\end{aligned}$$

The second inequality is based on Proposition 12. Now, by the pigeonhole principle, if $n \leq \frac{(M-3)H}{2} \times \frac{K}{72\tau^2}$, then the algorithm can collect $\frac{K}{72\tau^2}$ samples from at most half of the bandit problems. Thus $|C| > (M-3)H/2$, which implies,

$$\mathbb{E}[|B|] < \frac{5}{6}(M-3)H$$

By Markov's inequality,

$$\mathbb{P}\left[|B| \geq \frac{11}{12}(M-3)H\right] \leq \frac{\mathbb{E}[|B|]}{\frac{11}{12}(M-3)H} < \frac{5/6}{11/12} = 10/11$$

Thus with probability at least $1/11$ we know that $|B| \leq \frac{11}{12}(M-3)H$, so the algorithm failed to identify the optimal action on $1/12$ fraction of the bandit problems. Under this event, the suboptimality of the policy produced by the algorithm is,

$$\begin{aligned} V^\star - V(\hat{\pi}) &= \mathbb{P}[\text{visit } B^C] \times \tau = \mathbb{P}[\bigcup_{(h,i)\in B^C} \text{visit } (h,i)] \times \tau = \sum_{(h,i)\in B^C} \mathbb{P}[\text{visit } (h,i)] \times \tau \\ &= \sum_{(h,i)\in B^C} \frac{1}{H(M-3)}(1-1/H)^{h-1}\tau \geq \sum_{(h,i)\in B^C} \frac{1}{H(M-3)}(1-1/H)^{H}\tau \\ &\geq \sum_{(h,i)\in B^C} \frac{1}{H(M-3)}\frac{1}{4}\tau \geq \frac{H(M-3)}{12}\frac{1}{H(M-3)}\frac{1}{4}\tau = \frac{\tau}{48}. \end{aligned}$$

Here we use the fact that the probability of visiting a bandit state is independent of the policy and that the policy can only visit one bandit state per episode, so the events are disjoint. Moreover, if we visit a bandit state for which the algorithm failed to identify the optimal action, the difference in value is τ, since the optimal action visits the good state with τ more probability than a suboptimal one. The remainder of the calculation uses the transition model, the fact that $H \geq 2$, and finally the fact that $|B| \leq \frac{11}{12}(M-3)H$. Setting $\tau = 48\epsilon$ and using the requirement on τ gives a stricter requirement on ϵ and proves the result. $\square$