Provably Efficient Reinforcement Learning via Surprise Bound

Hanlin Zhu
Electrical Engineering
and Computer Sciences,
UC Berkeley

Ruosong Wang
Paul G. Allen School of
Computer Science & Engineering,
University of Washington

Jason D. Lee
Electrical and
Computer Engineering,
Princeton University

Abstract

Value function approximation is important in modern reinforcement learning (RL) problems especially when the state space is (infinitely) large. Despite the importance and wide applicability of value function approximation, its theoretical understanding is still not as sophisticated as its empirical success, especially in the context of general function approximation. In this paper, we propose a provably efficient RL algorithm (both computationally and statistically) with general value function approximations. We show that if the value functions can be approximated by a function class $\mathcal{F}$ which satisfies the Bellman-completeness assumption, our algorithm achieves an $\widetilde{O}(\text{poly}(\iota H)\sqrt{T})$ regret bound where ι is the product of the surprise bound and log-covering numbers, H is the planning horizon, K is the number of episodes and $T = HK$ is the total number of steps the agent interacts with the environment. Our algorithm achieves reasonable regret bounds when applied to both the linear setting and the sparse high-dimensional linear setting. Moreover, our algorithm only needs to solve $O(H \log K)$ empirical risk minimization (ERM) problems, which is far more efficient than previous algorithms that need to solve ERM problems for $\Omega(HK)$ times.

1 INTRODUCTION

Modern Reinforcement Learning (RL) problems are often challenging due to the huge state spaces, and in practice, value function approximation schemes are usually employed to tackle this issue. Empirically, combining various reinforcement learning algorithms with function approximation schemes has led to tremendous success on various tasks (Mnih et al., 2013, 2015; Silver et al., 2017). However, despite the great empirical success, our theoretical understanding of RL with function approximation is still not as sophisticated as its empirical counterpart. Until recently, most existing theoretical work in RL has been focusing on the tabular setting or the linear setting (Azar et al., 2017; Jin et al., 2018; Yang and Wang, 2019; Wang et al., 2019; Du et al., 2019b,a; Agarwal et al., 2020; Wang et al., 2020a; Du et al., 2020; Jin et al., 2020; Zanette et al., 2020; Li et al., 2020), while in practice, complex function approximators like neural networks are usually employed. Over the years, understanding conditions on the function class that permit sample-efficient RL has evolved into an important open research problem in machine learning theory.

Existing provably efficient RL algorithms that can handle general function approximation (Jiang et al., 2017; Sun et al., 2019; Ayoub et al., 2020; Jin et al., 2021; Du et al., 2021) usually require solving computationally intractable optimization problems and are therefore computationally inefficient. Recently, Wang et al. (2020b) proposed a provably efficient RL algorithm with general function approximation for function classes with bounded eluder dimensions. The algorithm by Wang et al. (2020b) is based on Least Squares Value Iteration (LSVI) and the principle of "optimism in the face of uncertainty". There are two shortcomings in the work of Wang et al. (2020b). First, in order to calculate the exploration bonus, their algorithm applies sensitivity sampling (Langberg and Schulman, 2010; Feldman and Langberg, 2011; Feldman et al., 2013) to reduce the size of the replay buffer. Using a replay buffer with bounded complexity to calculate the exploration bonus is crucial for the correctness of their algorithm. On the other hand, such a step is complicated in nature and could be hard to implement in practice. Therefore, to make the algorithm practical, it is much more desirable to use simpler dimensionality reduction techniques (like uniform sampling) without sacrificing the theoretical guarantee. Second, as mentioned in Foster et al. (2018), showing examples with a small eluder dimension beyond linearly parameterized functions is challenging. In addition, taking the worst-case

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

over all histories, as in the definition of the eluder dimension, is usually overly pessimistic in practice. In contextual bandits, it is known that provable efficiency can be established by assuming distributional conditions on the problem. For example, Foster et al. (2018) establishes regret bound for an optimism-based contextual bandits algorithm by assuming bounded surprise bound. It is natural to ask whether similar conditions can be used to establish provable efficiencies of RL algorithms.

Recently, Foster et al. (2020) established instance-dependent regret bounds for contextual bandits and reinforcement learning problems by assuming a bounded disagreement coefficient, which is a distribution-dependent assumption. Foster et al. (2020) show that the disagreement coefficient is always upper bounded by the eluder dimension of the function class. The RL algorithm in Foster et al. (2020), which is also based on Least Squares Value Iteration (LSVI) and the principle of "optimism in the face of uncertainty", has two drawbacks. First, their algorithm achieves provable guarantees only in the block MDP setting which might not be realistic in practice. Second, when calculating the exploration bonus, their algorithm uses the *star hull* to reduce the complexity of the replay buffer, which is also complicated in nature and therefore difficult to implement in practice.

In this paper, we develop a novel provably efficient RL algorithm with general function approximation. Similar to previous algorithms (Wang et al., 2020b; Foster et al., 2020), our algorithm is an optimistic version of LSVI. Compared to previous ones, our algorithm has the following advantages:

- The regret bound of our algorithm is based on a variant of surprise bound proposed in (Foster et al., 2018), which is a distribution-dependent quantity and could therefore be smaller than the eluder dimension which considers the worst-case over all histories. Moreover, our theory does not rely on the block MDP assumption. Furthermore, the surprise bound can be upper bounded in the tabular setting, the linear setting and the high dimensional sparse linear setting, which implies our algorithm achieves reasonable regret bound in all these three settings.

- The dimensionality reduction technique for reducing the complexity of the replay buffer is based on uniform sampling. This is much simpler than the sensitivity sampling framework in Wang et al. (2020b) and the method based on star hull in Foster et al. (2020).

- Our algorithm requires solving only $O(H \log K)$ empirical risk minimization (ERM) problems, while previous algorithms (Wang et al., 2020b; Foster et al., 2020) require solving $\Omega(HK)$ ERM problems.

1.1 Related Work

Tabular reinforcement learning. Tabular RL is well studied in the context of sample complexity and regret bound in numerous literature (Kearns and Singh, 2002; Kakade, 2003; Strehl et al., 2006, 2009; Jaksch et al., 2010; Azar et al., 2013; Lattimore and Hutter, 2014; Dann and Brunskill, 2015; Agrawal and Jia, 2017; Azar et al., 2017; Jin et al., 2018; Dann et al., 2019; Zanette and Brunskill, 2019; Zhang et al., 2020; Wang et al., 2020a; Yang et al., 2021). In particular, for episodic MDP without further assumptions, the best regret bound is $\widetilde{O}(\sqrt{H^2SAT})$ for both model-based (Azar et al., 2017) and model-free (Zhang et al., 2020) algorithms, which matches the lower bound $\Omega\left(\sqrt{H^2SAT}\right)$ proved by Jin et al. (2018). Recently, Yang et al. (2021) propose an RL algorithm with a regret bound of $O\left(\frac{SA\text{poly}(H)}{\Delta_{\min}}\log(SAT)\right)$ assuming the existence of a positive sub-optimality gap. However, all algorithms mentioned above cannot be applied to RL problems with huge or infinite state spaces due to the polynomial dependence on $\sqrt{S}$ in the regret bound. Therefore, in this paper, we assume the value function lies in a function class with bounded complexity and design a provably efficient algorithm whose regret bound depends polynomially on the complexity of the function class instead of the size of the state space.

Bandits. There is also rich literature studying stochastic (contextual) bandits, which can be viewed as a special case of MDP without state transitions (Auer, 2002; Dani et al., 2008; Li et al., 2010; Rusmevichientong and Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Foster et al., 2018, 2020; Li et al., 2019). In particular, Foster et al. (2018) study contextual bandit problems with general value function approximation, and prove their algorithms could achieve a regret bound depending polynomially on the surprise bound and the implicit exploration coefficient (IEC). In this paper, we study RL with general value function approximation, and prove that the regret bound of our algorithm also depends on the (slightly modified) surprise bound as well as the log-covering numbers. However, we note that the RL setting is much more complicated than the contextual bandits setting since there is no state transition in bandit problems.

Reinforcement learning with function approximation. In the setting of linear function approximation, there has been great interest recently in the theoretical analysis of the sample complexity of RL algorithms (Yang and Wang, 2019, 2020; Jin et al., 2020; Cai et al., 2020; Du et al., 2019b, 2020; Wang et al., 2019; Zanette et al., 2020; Zhou et al., 2021). Compared to linear function approximation, however, many current provably efficient algorithms for general value function approximation are relatively impractical. For example, algorithms in Jiang et al. (2017); Sun et al. (2019); Dong et al. (2020) achieve regret bound in

terms of the witness rank or the Bellman rank, but they are not computationally efficient. Foster et al. (2020) devise REGRL algorithm which is both computationally and statistically efficient. However, it requires the block MDP assumption which greatly alleviates the difficulty of (infinitely) huge state space and might not be realistic in practice. Ayoub et al. (2020) propose a model-based algorithm and Wang et al. (2020b) propose a model-free algorithm for general value function approximation, and the regret bound of both algorithms depend on the eluder dimension. Kong et al. (2021) propose an efficient algorithm both computationally and statistically for general value function approximation, of which the regret bound also depends on the eluder dimension. However, the eluder dimension considers the worst-case over all histories and is thus often overly pessimistic. Instead, the regret bound of our algorithm depends polynomially on the surprise bound which is a distribution-dependent quantity and thus could be smaller than the eluder dimension for practical scenarios.

2 PRELIMINARIES

In this paper, we study episodic *Markov Decision Process* (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r, \mu)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the finite action space, $H \in \mathbb{N}_+$ is the planning horizon, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel which maps a state-action pair to a distribution over the state space, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function and $\mu \in \Delta(\mathcal{S})$ is the initial state distribution [1].

A (stochastic) policy

$$\pi = \{\pi_h\}_{h=1}^H : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$$

maps any state s to a distribution over the action space at each step h, where we use $[N]$ to denote the set $\{1, 2, \ldots, N\}$ for any positive integer N. A trajectory

$$(s_1, a_1, r_1), (s_2, a_2, r_2), \ldots, (s_H, a_H, r_H)$$

is induced by a policy π if $s_1 \sim \mu$, $a_h \sim \pi_h(s_h), r_h = r(s_h, a_h), \forall h \in [H]$ and $s_{h+1} \sim P(s_h, a_h), \forall h \in [H-1]$. Furthermore, a policy $\pi = \{\pi_h\}_{h=1}^H$ is deterministic if for each step $h \in [H]$, $\pi_h : \mathcal{S} \to \mathcal{A}$ maps a state to only one action.

For any policy π, the expected cumulative reward starting from state s at step h is defined as the value function

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^H r_{h'} | s_h = s \right],$$

where we use superscript π to denote that the trajectory is induced by π. Similarly, the expected cumulative reward starting from state-action pair (s, a) at step h is defined as the Q-function

$$Q_h^\pi(s, a) = \mathbb{E}_\pi [\sum_{h'=h}^H r_{h'} | s_h = s, a_h = a].$$

Let π^* denote the optimal policy which maximizes $\mathbb{E}_{s_1 \sim \mu}[V_1^\pi(s_1)]$. Also, let $V_h^*(s) = V_h^{\pi^*}(s)$ and $Q_h^*(s, a) = Q_h^{\pi^*}(s, a)$.

The agent interacts with the environment for K episodes. At the beginning of each episode $k \in [K]$, the agent specifies a policy π^k based on previous trajectories and interacts with the environment using π^k for H steps. We assume the agent knows the number of episodes K, and we define $T = KH$ to be the total number of steps that the agent interacts with the environment. The *regret* of an algorithm after K episodes is defined as

$$\mathrm{Reg}(K) = \sum_{k=1}^K \left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right),$$

which compares the accumulated rewards between the agent's policy and the optimal policy. The goal of the agent is to minimize the regret. In this paper, we consider the typical regime that H is fixed while K grows to infinity.

Width function and norms. For notation convenience, we define the width function for any function class $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ and several norms for any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The width function is defined as

$$w(\mathcal{F}, s, a) = \max_{f, f' \in \mathcal{F}} \left(f(s, a) - f'(s, a) \right),$$

$\forall (s, a) \in \mathcal{S} \times \mathcal{A}$. For any dataset $\mathcal{Z} \subseteq \mathcal{S} \times \mathcal{A}$ and $\mathcal{D} \subseteq \mathcal{S} \times \mathcal{A} \times \mathbb{R}$, define $\mathcal{Z}$-norm

$$\|f\|_{\mathcal{Z}} = \sqrt{\sum_{(s,a) \in \mathcal{Z}} f^2(s, a)},$$

$\mathcal{D}$-norm

$$\|f\|_{\mathcal{D}} = \sqrt{\sum_{(s,a,r) \in \mathcal{D}} (f(s, a) - r)^2},$$

and infinite norm

$$\|f\|_\infty = \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |f(s, a)|$$

respectively. In addition, define $\|v\|_\infty = \max_{s \in \mathcal{S}} |v(s)|$ for any $v : \mathcal{S} \to \mathbb{R}$.

Additional notations for algorithms. For any finite multiset $\mathcal{X}$, let $\texttt{Unif}(\mathcal{X})$ denote the uniform distribution over $\mathcal{X}$ and $\mathrm{Card_d}(\mathcal{X})$ denote the number of distinct elements in $\mathcal{X}$. For any $x \in \mathbb{R}_+$, let $\lfloor x \rfloor$ denote the integer part of x and define $\lceil x \rceil = \lfloor x \rfloor + 1$ if x is not an integer and otherwise $\lceil x \rceil = x$. We use the standard $O(\cdot), \Omega(\cdot)$ notations to hide constants and use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ to suppress log factors. Also, we use $x \lesssim y$ to denote that there exists a constant $c > 0$ s.t. $x \le cy$, and use $x \gtrsim y$ if $y \lesssim x$.

[1] Our analysis can be naturally extended to the time-inhomogeneous settings where the reward function and the transition kernel are different for each $h \in [H]$.

3 ALGORITHM

In this section, we first introduce the assumptions for the algorithm and then present our main algorithm (Algorithm 1). The theoretical guarantee of our algorithm is presented in Section 4.

3.1 Assumptions

Assume our algorithm (Algorithm 1) receives a function class $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, H+1]\}$ as part of the input. Since the complexity of $\mathcal{F}$ determines the efficiency of the algorithm, it is natural and necessary to require bounded complexities of the function class under appropriate measures. We make the following assumptions on the function class $\mathcal{F}$.

Assumption 3.1 (Bellman-completeness)**.** *For any function $V : \mathcal{S} \to [0, H]$, there exists a function $f_V \in \mathcal{F}$, s.t.*

$$f_V(\cdot,\cdot) = r(\cdot,\cdot) + \sum_{s' \in \mathcal{S}} P(s'|\cdot,\cdot)V(s').$$

Assumption 3.1 indicates the closedness under Bellman equations. This is a general assumption that summarizes many previous assumptions in special settings and is commonly adopted in previous literature for general value function approximation (Wang et al., 2020b; Foster et al., 2020; Kong et al., 2021). For tabular RL, $\mathcal{F}$ can be chosen as the set of all functions mapping from $\mathcal{S} \times \mathcal{A}$ to $[0, H+1]$. In the linear MDP setting (Bradtke and Barto, 1996; Jin et al., 2020; Yang and Wang, 2019, 2020; Wang et al., 2019) where the transition kernel and the reward function are both linear in a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, $\mathcal{F}$ can be the set of all linear functions with respect to ϕ. In sparse high-dimensional linear MDP settings where the transition kernel and the reward function are both s-sparse linear functions in ϕ, $\mathcal{F}$ can be the set of all $(2s)$-sparse linear functions with respect to ϕ. Furthermore, Assumption 3.1 approximately holds in practice as long as $\mathcal{F}$ is rich enough (e.g., deep neural networks) and we show in Section 5 that our algorithm is robust to model misspecification.

Assumption 3.2 (Bounded covering number)**.** *Given any $\varepsilon > 0$, there exist covering sets $\mathcal{C}(\mathcal{F}, \varepsilon) \subseteq \mathcal{F}$ and $\mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon) \subseteq \mathcal{S} \times \mathcal{A}$ with bounded size $\mathcal{N}(\mathcal{F}, \varepsilon)$ and $\mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon)$ respectively, where*

- $\forall f \in \mathcal{F}$, $\exists f' \in \mathcal{C}(\mathcal{F}, \varepsilon)$, *s.t.* $\|f - f'\|_\infty \le \varepsilon$.
- $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, $\exists (s',a') \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, \varepsilon)$, *s.t.* $\max_{f \in \mathcal{F}} |f(s,a) - f(s',a')| \le \varepsilon$.

Assumption 3.2 requires bounded covering numbers $\mathcal{N}(\cdot, \varepsilon)$ for both $\mathcal{F}$ and $\mathcal{S} \times \mathcal{A}$, and the regret bound of our algorithm depends only logarithmically on the covering numbers (Theorem 4.1). In the tabular RL setting, $\ln \mathcal{N}(\mathcal{F}, \varepsilon) = \widetilde{O}(|\mathcal{S}||\mathcal{A}|)$ and $\ln \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon) = O(\ln(|\mathcal{S}||\mathcal{A}|))$. In d-dimensional linear MDP settings, $\ln \mathcal{N}(\mathcal{F}, \varepsilon) = \widetilde{O}(d)$ and $\ln \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon) = \widetilde{O}(d)$. In s-sparse high-dimensional linear MDP settings, $\ln \mathcal{N}(\mathcal{F}, \varepsilon) = \widetilde{O}(s)$. If we further assume that $\phi(s,a)$ is s-sparse for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, then $\ln \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon) = \widetilde{O}(s)$.

Surprise bound. Another important complexity measure in this paper is *surprise bound*, which was first introduced in Foster et al. (2018) to characterize the complexity of the function class in the contextual bandit setting.

Definition 3.3 (Surprise bound)**.** *The surprise bound is the smallest positive constant L_1 s.t.*

$$\begin{aligned}&(f(s,a) - f'(s,a))^2 \\ \le& L_1 \mathbb{E}_{s' \sim \mathcal{D}_h(\pi)} \mathbb{E}_{a' \sim \pi_h(s')} \left[(f(s',a') - f'(s',a'))^2 \right]\end{aligned}$$

for all $f, f' \in \mathcal{F}, s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$ and any policy π, where $\mathcal{D}_h(\pi)$ is the distribution of s_h when the policy is π.

Intuitively, the surprise bound is small if all pairs of functions with a small expected squared error with respect to any policy, do not encounter a much larger squared error on any state-action pair. The following proposition gives upper bounds of the surprise bound for linear and sparse linear settings (see Appendix C for the proof).

Proposition 3.4. *In the (sparse) linear MDP setting with a fixed feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, consider the function class $\mathcal{F} = \{(s,a) \mapsto w^\mathsf{T} \phi(s,a) | w \in \mathcal{W}\}$ for some $\mathcal{W} \subseteq \mathbb{R}^d$.*

- *If $\|\phi(s,a)\|_2 \le 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ and $\|w\|_2 \le 2H\sqrt{d}, \forall w \in \mathcal{W}$, then L_1 is upper bounded by*

$$\sup_{\pi, h \in [H]} \frac{1}{\lambda_{\min}\left(\mathbb{E}_{s \sim \mathcal{D}_h(\pi), a \sim \pi_h(s)} \left[\phi(s,a)\phi(s,a)^\mathsf{T}\right]\right)}.$$

- *If $\|\phi(s,a)\|_\infty \le 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ and $\|w\|_\infty \le 2H\sqrt{d}, \|w\|_0 \le 2s, \forall w \in \mathcal{W}$, then L_1 is upper bounded by*

$$\sup_{\pi, h \in [H]} \frac{4s}{\psi_{\min}\left(\mathbb{E}_{s \sim \mathcal{D}_h(\pi), a \sim \pi_h(s)} \left[\phi(s,a)\phi(s,a)^\mathsf{T}\right]\right)},$$

  where $\psi_{\min}(A) = \min_{w \neq 0 : \|w\|_0 \le 4s} w^\mathsf{T} A w / w^\mathsf{T} w$ is the minimum restricted eigenvalue for $(4s)$-sparse predictors (Raskutti et al., 2010).

3.2 Algorithm

In this section, we present our main algorithm (Algorithm 1) and discuss in detail several important components of our algorithm.

Algorithm 1 Optimistic LSVI with doubling epoch schedule

Input: number of epochs M, number of warm-start epochs M_0, failure probability $\delta \in (0,1)$
for episode $k = 1, 2, \ldots, \tau_{M_0} - 1$ **do**
  Receive initial state $s_1^k \sim \mu$
  for $h = 1, 2, \ldots, H$ **do**
    Take action $a_h^k \sim \texttt{Unif}(\mathcal{A})$, observe $s_{h+1}^k \sim P(\cdot|s_h^k, a_h^k)$ and receive $r_h^k = r(s_h^k, a_h^k)$
  end for
end for
for epoch $m = M_0, M_0 + 1, \ldots, M$ **do**
  $Q_{H+1}^m(\cdot,\cdot) \leftarrow 0$ and $V_{H+1}^m(\cdot) \leftarrow 0$
  $\mathcal{Z}^m \leftarrow \big\{(s_h^k, a_h^k)\big\}_{(h,k)\in[H]\times[\tau_m - 1]}$
  for $h = H, H-1, \ldots, 1$ **do**
    $\mathcal{D}_h^m \leftarrow \big\{\big(s_{h'}^k, a_{h'}^k, r_{h'}^k + V_{h+1}^m(s_{h'+1}^k)\big)\big\}$, $\forall (h', k) \in [H] \times [\tau_m]$
    $f_h^m \leftarrow \arg\min_{f\in\mathcal{F}} \|f\|_{\mathcal{D}_h^m}^2$
    $b_h^m(\cdot,\cdot) \leftarrow \texttt{Bonus}(\mathcal{F}, f_h^m, \mathcal{Z}^m, \delta)$ (Algorithm 3)
    $Q_h^m(\cdot,\cdot) \leftarrow \min\{f_h^m(\cdot,\cdot) + b_h^m(\cdot,\cdot), H\}$
    $V_h^m(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q_h^m(\cdot, a)$
    $\pi_h^m(\cdot) \leftarrow \arg\max_{a\in\mathcal{A}} Q_h^m(\cdot, a)$
  end for
  for episode $k = \tau_m, \tau_m + 1, \ldots, \tau_{m+1} - 1$ **do**
    Receive initial state $s_1^k \sim \mu$
    for $h = 1, 2, \ldots, H$ **do**
      Take action $a_h^k \leftarrow \pi_h^m(s_h^k)$, observe $s_{h+1}^k \sim P(\cdot|s_h^k, a_h^k)$ and receive $r_h^k = r(s_h^k, a_h^k)$
    end for
  end for
end for

3.2.1 Doubling Epoch Schedule

Our algorithm consists of M epochs where each epoch $m \in [M]$ starts at the beginning of episode $\tau_m = 2^{m-1}$ and consists of $T_m = 2^{m-1}$ episodes. Thus, the total number of episodes $K = 2^M - 1$ and $M = O(\log K)$. At the beginning of epoch m, the algorithm fixes a policy $\pi^m = \{\pi_h^m\}_{h=1}^H$ and the agent executes π^m for all episodes $k \in [\tau_m, \tau_m + T_m - 1]$. The M epochs can be divided into two phases.

- **Phase 1: Warm-up epochs.** For the first $(M_0 - 1)$ epochs, the agent plays a uniformly random policy. These warm-up epochs are designed to encourage exploration at the initial episodes.
- **Phase 2: Optimistic LSVI.** Starting from epoch M_0, we use an optimistic version of Least Squares Value Iteration (LSVI) similar to Jin et al. (2020); Wang et al. (2019, 2020b); Foster et al. (2020). At the beginning of each epoch $m \geq M_0$, we maintain all previous trajectories as a replay buffer, and find the best fit $f^m = \{f_h^m\}_{h=1}^H \in \mathcal{F}^H$ with respect to the replay buffer in the sense of mean squared error (MSE), i.e.,

$$f_h^m \leftarrow \arg\min_{f\in\mathcal{F}} \|f\|_{\mathcal{D}_h^m}^2$$

  where $\mathcal{D}_h^m$ is the replay buffer (see definition in Algorithm 1). To avoid overfitting and encourage exploration, we design a bonus function $b_h^m(\cdot,\cdot)$ which we will discuss later in Section 3.2.2, and approximate the optimal Q function $Q_h^*(\cdot,\cdot)$ by

$$Q_h^m(\cdot,\cdot) = \min\left\{f_h^m(\cdot,\cdot) + b_h^m(\cdot,\cdot), H\right\}.$$

  Our design of the bonus function ensures that Q_h^m is an optimistic estimator of Q_h^* with high probability (Lemma B.3). Finally, for each episode $k \in [\tau_m, \tau_{m+1} - 1]$ in epoch m, the agent plays the greedy policy with respect to Q_h^m and collect the trajectory in episode k.

The advantages of the doubling epoch schedule are two folded:

- **Computationally efficient.** Since our algorithm only conducts large amount of computation at the beginning of each epoch (computing f_h^m by empirical risk minimization and b_h^m by the width function as in Section 3.2.2, which can often be solved efficiently by appropriate optimization methods or assuming access to appropriate regression oracles (Wang et al., 2020b; Foster et al., 2018)) and there are only $O(\log K)$ epochs, our algorithm is much more computationally efficient than previous methods (Wang et al., 2020b; Foster et al., 2020) which require to solve $\Omega(HK)$ equivalent optimization problems.

Recently, Kong et al. (2021) proposes an online subsampling technique which improves the computational complexity of Wang et al. (2020b). However, our algorithm is still much more computationally efficient than Kong et al. (2021). The algorithm of Kong et al. (2021) adopts sensitivity sampling, which requires computing sensitivities for each state action pair (s_h^k, a_h^k). Since the calculation of sensitivity requires solving a regression oracle for $\Omega(\log(TH))$ times (see Section 4.4. in Kong et al. (2021)), and there are $T = KH$ such state-action pairs, their algorithm needs to solve $\Omega(KH\log(TH))$ regression oracles to calculate sensitivities and subsample the dataset. While in our algorithm, we use uniform sampling to avoid the complex and time-consuming sensitivity calculation and thus does not need any oracle to perform the subsampling procedure.

- **Stabilizing adjacent trajectories.** The doubling epoch schedule together with the warm-up epochs stabilizes the adjacent trajectories by ensuring that at the beginning of each epoch, at least half of the historical trajectories in the replay buffer are induced by the same policy. This property enables us to adopt uniform sampling (Algorithm 2) to reduce the complexity of the replay buffer.

3.2.2 Uniform Sampling

An important technical novelty of our algorithm is the design of the bonus function via uniform sampling. To ensure optimism of our estimator Q_h^m, we can choose b_h^m as the upper bound of the difference between Q_h^* and f_h^m. If we are able to obtain a confidence region $\mathcal{F}_h^m$ which contains both f_h^m and Q_h^*, it suffices to define the bonus function as the width function of $\mathcal{F}_h^m$.

A naive way to choose the confidence region is $\mathcal{F}_h^m = \big\{f \in \mathcal{F} \,\big|\, \|f - f_h^m\|_{\mathcal{Z}^m}^2 \le \beta\big\}$ with a carefully selected β. However, since the confidence region depends on the whole replay buffer with size at most T, the confidence region and thus the bonus function would suffer extremely high complexity. This implies that β needs to be set extremely large to ensure the accuracy of the confidence region. To obtain a bonus function with low complexity, we reduce the complexity of the replay buffer by uniform sampling, which is formally stated in Algorithm 2.

Comparison to previous methods. Actually, the algorithms in Wang et al. (2020b); Foster et al. (2020) also suffer the high complexity of the bonus function and address the issue by sensitivity sampling and star hull respectively. However, sensitivity sampling requires estimating the sensitivity of each state-action pair, which is time-consuming; the star hull is complicated in nature and thus is hard to implement in practice. In contrast, our uniform sampling is conceptually simple and easy to implement.

Algorithm 2 `Uniform-Sampling`$(\mathcal{F}, \mathcal{Z}, \lambda, \varepsilon, \delta)$

Input: function class $\mathcal{F}$, dataset $\mathcal{Z}$, parameters $\lambda, \varepsilon > 0$ and failure probability $\delta \in (0,1)$
Set $\varepsilon_0 \leftarrow \varepsilon/72 \cdot \sqrt{\lambda\delta/|\mathcal{Z}|}$
Set $p^{-1} \leftarrow \max\Big\{1, \Big\lfloor \frac{1}{384L_1 \cdot \ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/(\varepsilon^2\cdot|\mathcal{Z}|)} \Big\rfloor\Big\}$
Initialize $\mathcal{Z}' \leftarrow \{\}$
for $z \in \mathcal{Z}$ **do**
  Add $1/p$ copies of z to $\mathcal{Z}'$ with probability p
end for
Output: $\mathcal{Z}'$

Note that there is only one single parameter p to be determined in Algorithm 2. When the surprise bound L_1 is known in advance, we can directly calculate the value of p. When L_1 is unknown, we can perform a grid-search in a log-space of L_1. Specifically, we can set a small value $L_{\min}$ as the lower bound of L_1 and a large value $L_{\max}$ as the upper bound, and perform Algorithm 1 for $L_1 \in \mathcal{L} \triangleq \{L_{\min}, 2L_{\min}, 2^2L_{\min}, \ldots, L_{\max}\}$. Then we can pick the policy with the best performance under different choices of L_1.

Theorem 4.1 shows that the regret of our main algorithm (Algorithm 1) is $\tilde{O}(\sqrt{T})$ in T dependence. We also emphasize that the above grid-search procedure won't result in higher total regret, since one can first try each possible $L_1 \in \mathcal{L}$ for $O(\sqrt{T})$ times, and then exploit the best L_1 for the remaining $O(T - \sqrt{T}\log(L_{\max}/L_{\min})) = O(T)$ steps. The resulting total regret is still $\tilde{O}(\sqrt{T})$.

Algorithm 3 `Bonus`$(\mathcal{F}, \bar{f}, \mathcal{Z}, \delta)$

Input: function class $\mathcal{F}$, reference function $\bar{f}$, dataset $\mathcal{Z}$ and failure probability $\delta \in (0,1)$
$\mathcal{Z}' \leftarrow$ `Uniform-Sampling`$(\mathcal{F}, \mathcal{Z}, \frac{\delta}{(16T)^2}, \frac{1}{2}, \frac{\delta}{16T})$ (Algorithm 2)
if $|\mathcal{Z}'| > 64T^2/\delta$ **or** $\mathrm{Card}_{\mathrm{d}}(\mathcal{Z}') \ge 9216L_1 \cdot \ln(64T\mathcal{N}(\mathcal{F}, \delta/(9216T^2))/\delta)$ **then**
  $\mathcal{Z}' \leftarrow \{\}$
end if
Let $\hat{f} \in \mathcal{C}(\mathcal{F}, 1/(8\sqrt{64T^2/\delta}))$ such that $\|\hat{f} - \bar{f}\|_\infty \le 1/(8\sqrt{64T^2/\delta})$
$\widehat{\mathcal{Z}} \leftarrow \{\}$
for $z \in \mathcal{Z}'$ **do**
  Let $\hat{z} \in \mathcal{C}(\mathcal{S}\times\mathcal{A}, 1/(8\sqrt{64T^2/\delta}))$ such that $\sup_{f\in\mathcal{F}} |f(z) - f(\hat{z})| \le 1/(8\sqrt{64T^2/\delta})$
  $\widehat{\mathcal{Z}} \leftarrow \widehat{\mathcal{Z}} \cup \{\hat{z}\}$
end for
$\beta \triangleq \beta(\mathcal{F}, \delta) \leftarrow c' \cdot L_1 H^2 \ln^3(T/\delta)\ln(\mathcal{N}(\mathcal{F}, \delta/T^3)) \times \ln(\mathcal{N}(\mathcal{S}\times\mathcal{A}, \delta/T^2))$ for some constant $c' > 0$
$\widehat{\mathcal{F}} \leftarrow \Big\{f \in \mathcal{F} \mid \|f - \hat{f}\|_{\widehat{\mathcal{Z}}}^2 \le 3\beta + 2\Big\}$
Output: $\hat{w}(\cdot,\cdot) \leftarrow w(\widehat{\mathcal{F}}, \cdot, \cdot)$

Design of the bonus function via uniform sampling. Now we are able to design a bonus function with low complexity as in Algorithm 3 via uniform sampling. After obtaining the reduced dataset $\mathcal{Z}'$, we round each data in $\mathcal{Z}'$ and the reference function $\bar{f}$ to their nearest neighbors in covering sets. The confidence region and the bonus function is then defined by the rounded reference function and the rounded dataset. Note that in Algorithm 3, the rounding operation does not need to be performed explicitly since all the data are stored in computers with bounded precision, and thus all the data will be implicitly rounded. For the choice of β, we can use the same grid-search method of L_1 since β is also determined by L_1.

Efficient computation of the bonus function. The computation of the bonus function is equivalent to an optimization problem of the following form:

$$\max_{f_1, f_2 \in \mathcal{F}} f_1(s,a) - f_2(s,a)$$
$$\text{s.t.} \quad \|f_1 - f_2\|_{\mathcal{Z}} \le \varepsilon.$$

This problem can be solved efficiently by either assuming access to an optimization oracle, or assuming access to only a regression oracle (which is a milder assumption than optimization oracles) as mentioned in Section 4.4 of Kong et al. (2021).

4 THEORETICAL RESULTS

In this section, we formally present our main theorem of the regret bound and defer the proof to Appendix B.

Theorem 4.1 (Main theorem)**.** *Under Assumptions 3.1 and 3.2, let* $M_0 = \left\lceil \ln\left(16L_1^2 \ln \frac{128T\mathcal{N}(\mathcal{F},\delta/(9216T^2))^2}{\delta}\right)\right\rceil$ *where the number of total steps* $T = H \cdot (2^M - 1)$ *is sufficiently large. With probability at least* $1-\delta$*, the regret of* $Algorithm\ 1$ *is at most*

$$O(\iota \cdot H^{3/2} \cdot \sqrt{T}),$$

where $\iota = L_1 \cdot \ln^2(T/\delta) \cdot \max(\ln(\mathcal{N}(\mathcal{F}, \delta/T^3)), \ln(\mathcal{N}(\mathcal{S}\times\mathcal{A}, \delta/T^2)))$.

Proof sketch. In this proof sketch, we ignore the rounding operation in Algorithm 3 for convenience. The proof can be decomposed into three main steps.

- **Step 1: Bounding the complexity of the bonus function.** First, we show that our bonus function has low complexity (Proposition A.5). Note that the bonus function is defined as the width function of the confidence region

$$\hat{\mathcal{F}}_h^m = \left\{ f \in \mathcal{F} \,\middle|\, \|f - \hat{f}_h^m\|_{\hat{\mathcal{Z}}^m}^2 \le \beta \right\}.$$

Since the reduced dataset $\hat{\mathcal{Z}}^m$ has bounded size (Lemma A.1) and bounded number of distinct elements (Lemma A.3), our bonus function which is defined by $\hat{\mathcal{Z}}^m$ also has low complexity. Now it remains to show that the bonus function defined over the reduced dataset $\hat{\mathcal{Z}}^m$ is (almost) the same as the bonus function defined over the original dataset $\mathcal{Z}^m$. It is equivalent to show that the confidence region remains (almost) unchanged after uniform sampling. This can be proved by showing that for any function pairs $f, f' \in \mathcal{F}$, the $\mathcal{Z}'$-norm of $f - f'$ approximates well the $\mathcal{Z}$-norm of $f - f'$ (Lemma A.2). For a fixed function pair (f, f'), $\|f - f'\|_{\mathcal{Z}'}^2$ is an unbiased estimator of $\|f - f'\|_{\mathcal{Z}}^2$ and its variance can be controlled, since the trajectories in the replay buffer are stabilized by the doubling epoch and thus $\mathcal{Z}'$ has low complexity after uniform sampling. Then we can apply the Bernstein inequality to a fixed function pair (f, f') to show that $\|f - f'\|_{\mathcal{Z}'}^2$ is close to $\|f - f'\|_{\mathcal{Z}}^2$ with high probability. Applying a union bound over all function pairs in the covering set of $\mathcal{F}$, we can obtain the desired result.

- **Step 2: Optimism of the estimated Q-function.** The next step is to show that the estimated Q-function is an optimistic version of the true Q-function of the optimal policy (Lemma B.3). To achieve this, we need to show that the best fit f_h^m is close to $r(\cdot,\cdot) + \sum_{s' \in \mathcal{S}} P(s'|\cdot,\cdot) V_{h+1}^m(s')$. If f_h^m and V_{h+1}^m are independent, a standard concentration argument concludes the result. However, V_{h+1}^m and f_h^m are subtly dependent since they are both determined by the previous dataset. To address the difficulty, we first apply the standard concentration result on a fixed V (Lemma B.1), and then apply a union bound over all V in a covering set (Lemma B.2) to obtain the result. This method is similar to Wang et al. (2020b).

- **Step 3: Regret decomposition.** Finally, we decompose the regret by the summation of the bonus functions (Lemma B.4). Then, we use similar arguments as in Foster et al. (2018) to bound each bonus term by the surprise bound separately since the bonus function is defined as the (approximate) width function of the confidence region.

□

Remark 4.2. *Recently, Foster et al. (2021) proposes a high-level algorithm E2D. When applying E2D algorithm to our settings, one can show that it also achieves a similar regret bound* $\tilde{O}(poly(L_1)\sqrt{T})$ *(other parameters omitted). However, we want to emphasize that E2D algorithm is too high-level to implement in practice. The implementation of E2D algorithm requires an online estimation oracle (see Algorithm 1 in Foster et al. (2021)), which is a very strong assumption in RL settings. While in our algorithm, we only*

require a ERM oracle and a regression oracle, which are mild and common assumptions in machine learning problems.

While our algorithm works for general value function class, it also achieves reasonable regret in special cases.

Tabular settings. In the tabular RL setting, it holds that $\ln \mathcal{N}(\mathcal{F}, \varepsilon) = \widetilde{O}(|\mathcal{S}||\mathcal{A}|)$ and $\ln \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon) = O(\ln(|\mathcal{S}||\mathcal{A}|))$. When $\mu(s) \geq \varepsilon$ and $P(s'|s,a) \geq \varepsilon$ for all $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$ for a (not too) small positive value ε, $L_1 = O(\text{poly}(|\mathcal{S}||\mathcal{A}|))$, which implies that the regret bound is $\tilde{O}(\text{poly}(|\mathcal{S}||\mathcal{A}|)H^{3/2}\sqrt{T})$. This is a reasonable regret bound since it is optimal in terms of T, the most important term in the regret bound, and has polynomial dependency in other parameters.

Linear settings. When $\mathcal{F}$ is a d-dimensional linear function class, we have $\ln \mathcal{N}(\mathcal{F}, \varepsilon) = \ln \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon) = \widetilde{O}(d)$. When

$$\lambda_{\min}\left(\mathbb{E}_{s' \sim \mathcal{D}_h(\pi)} \mathbb{E}_{a' \sim \pi_h(s')}\left[\phi(s', a')\phi(s', a')^{\mathsf{T}}\right]\right)$$

is lower bounded (of order $\Omega(1/d)$) and thus $L_1 = O(d)$ by Proposition 3.4, the regret bound is $\widetilde{O}(d^2 \cdot H^{3/2} \cdot \sqrt{T})$, which is optimal in T-dependency and matches the result of Wang et al. (2020b) in d-dependency.

Sparse linear settings. Furthermore, when $\mathcal{F}$ is an s-sparse high-dimensional linear function class where typically $d \geq T \gg s$, we have $\ln \mathcal{N}(\mathcal{F}, \varepsilon) = \widetilde{O}(s)$. When

$$\psi_{\min}\left(\mathbb{E}_{s' \sim \mathcal{D}_h(\pi)} \mathbb{E}_{a' \sim \pi_h(s')}\left[\phi(s', a')\phi(s', a')^{\mathsf{T}}\right]\right)$$

is lower bounded (of order $\Omega(1)$) and thus L_1 is $O(s)$ by Proposition 3.4, the regret bound is $\widetilde{O}(s \cdot \max(s, \ln(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2))) \cdot H^{3/2} \cdot \sqrt{T})$. If we further assume that $\phi(s', a')$ is s-sparse for all $(s', a') \in \mathcal{S} \times \mathcal{A}$, we have $\ln \mathcal{N}(\mathcal{S} \times \mathcal{A}, \varepsilon) = \widetilde{O}(s)$ and thus obtain an $\widetilde{O}(s^2 \cdot H^{3/2} \cdot \sqrt{T})$ regret bound. However, directly applying the result in linear settings of Wang et al. (2020b) can only obtain a linear regret when $d \geq T$. This shows the superiority of our algorithm since we can provide theoretical guarantee for more general function classes, and thus it is an important step toward studying general value function approximation beyond the tabular and linear settings.

We also emphasize a subtle difference between linear and sparse linear settings. In linear settings, when $\lambda_{\min}\left(\mathbb{E}_{s' \sim \mathcal{D}_h(\pi)} \mathbb{E}_{a' \sim \pi_h(s')}\left[\phi(s', a')\phi(s', a')^{\mathsf{T}}\right]\right)$ is lower bounded, we typically expect it to be of order $\Omega(1/d)$ since we assume the *2-norm* $\|\phi\|_2 \leq 1$. While for sparse linear settings, when $\psi_{\min}\left(\mathbb{E}_{s' \sim \mathcal{D}_h(\pi)} \mathbb{E}_{a' \sim \pi_h(s')}\left[\phi(s', a')\phi(s', a')^{\mathsf{T}}\right]\right)$ is lower bounded, we typically expect it to be of order $\Omega(1)$ since we assume the *infinity norm* $\|\phi\|_\infty \leq 1$ in this setting.

5 MODEL MISSPECIFICATION

Our main theorem (Theorem 4.1) requires Bellman-completeness assumption (Assumption 3.1). Although the Bellman-completeness assumption is fairly common in theoretical analysis, especially in the presence of general value function approximation, the ground truth model together with the function class might slightly violate this assumption in real-world scenario. This phenomenon is known as model misspecification (Jin et al., 2020; Wang et al., 2020b).

In this section, we show that as long as the violation of the Bellman-completeness assumption is small, the regret of our algorithm is still bounded. To state the result formally, we first introduce the following assumption, which can be viewed as a model misspecification version of the Bellman-completeness assumption.

Assumption 5.1 (Model misspecification)**.** *There exists a constant $\zeta > 0$ satisfying that for any function $V : \mathcal{S} \to [0, H]$, there exists a function $f_V \in \mathcal{F}$, s.t.*

$$\left\| f_V(\cdot, \cdot) - r(\cdot, \cdot) + \sum_{s' \in \mathcal{S}} P(s'|\cdot, \cdot)V(s') \right\|_\infty \leq \zeta.$$

Under Assumption 5.1, one can directly apply Algorithm 1 to the model misspecification setting with only a different choice of the parameter β in Algorithm 3. Specifically, for some constant $c' > 0$ we set

$$\begin{aligned} \beta =& c'(L_1 H^2 \ln^3(T/\delta) \ln(\mathcal{N}(\mathcal{F}, \delta/T^3)) \\ & \cdot \ln(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2)) + HT\zeta). \end{aligned} \tag{1}$$

Note that when Assumption 3.1 holds, it is equivalent to Assumption 5.1 with $\zeta = 0$, and thus the parameter β is exactly the same as the one in our original algorithm. The following theorem provides theoretical guarantees of our algorithm for model misspecification, and the proof is attached in Appendix D, which is very similar to the proof of Theorem 4.1.

Theorem 5.2 (Theoretical guarantee for model misspecification)**.** *Under Assumptions 3.2 and 5.1, let $M_0 = \left\lceil \ln\left(16 L_1^2 \ln \frac{128T\mathcal{N}(\mathcal{F}, \delta/(9216T^2))^2}{\delta}\right)\right\rceil$ and the number of total steps $T = H \cdot (2^M - 1)$. With probability at least $1 - \delta$, the regret of Algorithm 1 (where the parameter β is defined as in* (1)*) is at most*

$$O(\iota \cdot H^{3/2} \cdot \sqrt{T} + \sqrt{L_1 \cdot H^2 \cdot \zeta \cdot \log T} \cdot T),$$

where $\iota = L_1 \cdot \ln^2(T/\delta) \cdot \max(\ln(\mathcal{N}(\mathcal{F}, \delta/T^3)), \ln(\mathcal{N}(\mathcal{S} \times \mathcal{A}, \delta/T^2)))$.

6 CONCLUSION

In this paper, we propose a provably efficient RL algorithm (both computationally and statistically) with general value

function approximation. The regret bound of our algorithm depends on the surprise bound, which is a distribution-dependent quantity and could therefore be smaller than the eluder dimension considered in previous work. Our algorithm achieves reasonable regret bound when instantiating to special function classes.

As a future direction, it would be interesting to see if it is possible to establish the provable efficiency of RL algorithms using other distribution-dependent complexity measures. For example, it would be interesting to study whether it is possible to design a provably efficient RL algorithm by assuming a bounded disagreement coefficient (as in Foster et al. (2020)) but without the block MDP assumption.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *arXiv preprint arXiv:1510.08906*, 2015.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory*, pages 1554–1557. PMLR, 2020.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019a.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle. *arXiv preprint arXiv:1906.06321*, 2019b.

Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity. *arXiv preprint arXiv:2002.07125*, 2020.

Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578, 2011.

Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1434–1453, 2013.

Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learn-

ing: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

Dingwen Kong, Ruslan Salakhutdinov, Ruosong Wang, and Lin F Yang. Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

Michael Langberg and Leonard J Schulman. Universal ε-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.

Tor Lattimore and Marcus Hutter. Near-optimal pac bounds for discounted mdps. *Theoretical Computer Science*, 558:125–143, 2014.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11: 2241–2259, 2010.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, 2006.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(11), 2009.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019.

Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020a.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020b.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR, 2021.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

A ANALYSIS OF THE BONUS FUNCTION

In this section, we analyze our bonus function, and the main proposition is presented in Proposition A.5.

A.1 Analysis of Algorithm 2

Note that the notation δ in Algorithm 3 and Algorithm 2 are different. In this subsection, all the notation δ refer to δ in Algorithm 2, and therefore, $\lambda = \delta/(16T)$. Also, let $\varepsilon_0 = \varepsilon/72 \cdot \sqrt{\lambda\delta/|\mathcal{Z}|}$ throughout this subsection.

We assume that the input dataset of Algorithm 2 is $\mathcal{Z} = \{(s_h^k, a_h^k)\}_{(h,k)\in[H]\times[t]}$ where more than half of the trajectories are induced by the same policy and the number of trajectories

$$t \geq 4L_1^2 \ln \frac{8\mathcal{N}(\mathcal{F}, \varepsilon_0)^2}{\delta}$$

which is satisfied if $t \geq \tau_{M_0}$ and M_0 is chosen as in Theorem 4.1.

The first lemma gives an upper bound on the size of the dataset produced by uniform sampling.

Lemma A.1. *With probability at least* $1 - \delta/4$, $|\mathcal{Z}'| \leq 4|\mathcal{Z}|/\delta$.

Proof. We define random variable

$$X_z = \begin{cases} 1/p & z \text{ is added into } \mathcal{Z}' \text{ for } 1/p \text{ times} \\ 0 & \text{otherwise} \end{cases}.$$

Since $|\mathcal{Z}'| = \sum_{z\in\mathcal{Z}} X_z$ and $\mathbb{E}[X_z] = 1$, we can obtain

$$\Pr\{|\mathcal{Z}'| > 4|\mathcal{Z}|/\delta\} \leq \delta/4$$

by Markov inequality. □

The next lemma proves that after uniform sampling, the norms of difference of any function pairs are approximately preserved with high probability.

Lemma A.2. *With probability at least* $1 - \delta/2$, *for any* $f, f' \in \mathcal{F}$,

$$(1-\varepsilon)\|f - f'\|_{\mathcal{Z}}^2 - 2\lambda \leq \|f - f'\|_{\mathcal{Z}'}^2 \leq (1+\varepsilon)\|f - f'\|_{\mathcal{Z}}^2 + 8|\mathcal{Z}|\lambda/\delta.$$

Proof. When $p = 1$, $\mathcal{Z} = \mathcal{Z}'$, the result directly holds. So we only consider the case when $p < 1$, which means

$$p \geq 384L_1 \cdot \ln(4\mathcal{N}(\mathcal{F}, \varepsilon_0)/\delta)/(\varepsilon^2 \cdot |\mathcal{Z}|).$$

We separately consider the cases when $\|f - f'\|_{\mathcal{Z}}^2 < 2\lambda$ and $\|f - f'\|_{\mathcal{Z}}^2 \geq 2\lambda$.

For any function pair $f, f' \in \mathcal{F}$ where $\|f - f'\|_{\mathcal{Z}}^2 < 2\lambda$, conditioned on the event in Lemma A.1 which holds with probability at least $1 - \delta/4$, we can obtain that $\|f - f'\|_{\mathcal{Z}'}^2 \leq |\mathcal{Z}'|\|f - f'\|_{\mathcal{Z}}^2 \leq 4|\mathcal{Z}|/\delta \cdot \|f - f'\|_{\mathcal{Z}}^2 \leq 8|\mathcal{Z}|\lambda/\delta$. Also, by the fact that $\|f - f'\|_{\mathcal{Z}}^2 < 2\lambda$ and $\|f - f'\|_{\mathcal{Z}'}^2 \geq 0$, we can conclude that

$$(1-\varepsilon)\|f - f'\|_{\mathcal{Z}}^2 - 2\lambda \leq \|f - f'\|_{\mathcal{Z}'}^2 \leq (1+\varepsilon)\|f - f'\|_{\mathcal{Z}}^2 + 8|\mathcal{Z}|\lambda/\delta.$$

In the remaining part of the proof, we consider the case that $\|f - f'\|_{\mathcal{Z}}^2 \geq 2\lambda$.

We first fix any pair of distinct functions $f, f' \in \mathcal{C}(\mathcal{F}, \varepsilon_0)$. Assume the first $u = \lfloor (t+1)/2 \rfloor$ trajectories are all induced by the same policy π. Also, for any $1 \leq k \leq u$, let

$$g_k = \sum_{h=1}^{H} (f(s_h^k, a_h^k) - f'(s_h^k, a_h^k))^2.$$

Therefore,

$$\mathbb{E}\left[g_k\right] = \sum_{h=1}^{H} \mathbb{E}_{s\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a\sim\pi_h(s)}\left[(f(s,a)-f'(s,a))^2\right].$$

Note that

$$0 \le g_k \le H \times \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} (f(s,a)-f'(s,a))^2 = H\|f-f'\|_\infty^2.$$

Also, by Definition 3.3,

$$\mathbb{E}\left[g_k\right] \ge \frac{H}{L_1} \max_{s\in\mathcal{S},a\in\mathcal{A}} (f(s,a)-f'(s,a))^2 = \frac{H}{L_1}\|f-f'\|_\infty^2.$$

Therefore, by Hoeffding's inequality,

$$\begin{aligned}
&\Pr\left\{\frac{1}{u}\sum_{k=1}^{u}\left(g_k - \mathbb{E}\left[g_k\right]\right) \le -v\mathbb{E}\left[g_1\right]\right\} \le \exp\left(-\frac{2u^2v^2\mathbb{E}\left[g_1\right]^2}{uH^2\|f-f'\|_\infty^4}\right)\\
\le&\exp\left(-\frac{2uv^2}{H^2\|f-f'\|_\infty^4}\cdot\frac{H^2\|f-f'\|_\infty^4}{L_1^2}\right) \le \exp\left(-\frac{tv^2}{L_1^2}\right)\\
\le&\exp\left(-\frac{v^2}{L_1^2}\cdot 4L_1^2\ln\frac{8\mathcal{N}(\mathcal{F},\varepsilon_0)^2}{\delta}\right) \le \exp\left(-4v^2\ln\frac{8\mathcal{N}(\mathcal{F},\varepsilon_0)^2}{\delta}\right).
\end{aligned}$$

Setting $v=\frac{1}{2}$, we can obtain

$$\Pr\left\{\frac{1}{u}\sum_{k=1}^{u} g_k \le \frac{1}{2}\mathbb{E}\left[g_1\right]\right\} \le \frac{\delta}{8\mathcal{N}(\mathcal{F},\varepsilon_0)^2}.$$

Let $\mathcal{E}_1$ denote the event that

$$\frac{1}{u}\sum_{k=1}^{u} g_k \ge \frac{1}{2}\mathbb{E}\left[g_1\right],$$

then $\Pr\{\mathcal{E}_1\} \ge 1-\frac{\delta}{8\mathcal{N}(\mathcal{F},\varepsilon_0)^2}$.

Now, we condition on $\mathcal{E}_1$ for the following analysis. For each $z\in\mathcal{Z}$, define

$$X_z = \begin{cases} \frac{1}{p}(f(z)-f'(z))^2 & z \text{ is added into } \mathcal{Z}' \text{ for } 1/p \text{ times}\\ 0 & \text{otherwise}\end{cases}.$$

Obviously, $\|f-f'\|_{\mathcal{Z}'}^2 = \sum_{z\in\mathcal{Z}} X_z$, and $\mathbb{E}[X_z] = \left(f(z)-f'(z)\right)^2$. Also,

$$\begin{aligned}
&\sum_{z\in\mathcal{Z}}\mathrm{Var}[X_z] \le \sum_{z\in\mathcal{Z}}\mathbb{E}[X_z^2] \le \max_{z\in\mathcal{Z}}(f(z)-f'(z))^2/p\cdot\sum_{z\in\mathcal{Z}}(f(z)-f'(z))^2\\
=&\frac{\|f-f'\|_{\mathcal{Z}}^4}{p}\cdot\frac{\max_{z\in\mathcal{Z}}(f(z)-f'(z))^2}{\sum_{z\in\mathcal{Z}}(f(z)-f'(z))^2}\\
\le&\frac{\|f-f'\|_{\mathcal{Z}}^4}{p}\cdot\frac{\frac{1}{H}\sum_{h=1}^{H}L_1\mathbb{E}_{s\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a\sim\pi_h(s)}\left[(f(s,a)-f'(s,a))^2\right]}{\sum_{k=1}^{u}\sum_{h=1}^{H}(f(s_h^k,a_h^k)-f'(s_h^k,a_h^k))^2}\\
\le&\frac{\|f-f'\|_{\mathcal{Z}}^4}{p\cdot uH}\cdot\frac{L_1\mathbb{E}[g_1]}{\frac{1}{u}\sum_{k=1}^{u}g_k}\\
\le&\frac{2L_1\|f-f'\|_{\mathcal{Z}}^4}{pu\cdot H} \le \frac{\|f-f'\|_{\mathcal{Z}}^4\cdot\varepsilon^2}{96\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)}.
\end{aligned}$$

Moreover,

$$\begin{aligned}\max_{z\in\mathcal{Z}} X_z =& \max_{z\in\mathcal{Z}} \frac{(f(z)-f'(z))^2}{p}\\ \leq& \frac{\|f-f'\|_{\mathcal{Z}}^2}{p}\cdot\frac{\max_{z\in\mathcal{Z}}(f(z)-f'(z))^2}{\sum_{z\in\mathcal{Z}}(f(z)-f'(z))^2}\\ =& \frac{\varepsilon^2\|f-f'\|_{\mathcal{Z}}^2}{96\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)}\end{aligned}$$

Then, by Azuma-Bernstein's Inequality,

$$\begin{aligned}&\Pr\left\{|\|f-f'\|_{\mathcal{Z}}^2-\|f-f'\|_{\mathcal{Z}'}^2|\geq\varepsilon/4\cdot\|f-f'\|_{\mathcal{Z}}^2\,|\mathcal{E}_1\right\}\\ =&\Pr\left\{\left|\sum_{z\in\mathcal{Z}}\mathbb{E}[X_z]-\sum_{z\in\mathcal{Z}}X_z\right|\geq\varepsilon/4\cdot\|f-f'\|_{\mathcal{Z}}^2\,\middle|\,\mathcal{E}_1\right\}\\ \leq&2\exp\left(-\frac{\varepsilon^2/16\cdot\|f-f'\|_{\mathcal{Z}}^4}{2\sum_{z\in\mathcal{Z}}\mathrm{Var}[X_z]+2/3\max_{z\in\mathcal{Z}}X_z\cdot\varepsilon/4\cdot\|f-f'\|_{\mathcal{Z}}^2}\right)\\ \leq&2\exp\left(-\frac{\varepsilon^2/16\cdot\|f-f'\|_{\mathcal{Z}}^4\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)}{\|f-f'\|_{\mathcal{Z}}^4\cdot\varepsilon^2/48+\|f-f'\|_{\mathcal{Z}}^4\cdot\varepsilon^2/576}\right)\\ \leq&2\exp\left(-2\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)\right)\\ \leq&(\delta/8)/\left(\mathcal{N}(\mathcal{F},\varepsilon_0)\right)^2.\end{aligned}$$

Since the above inequality holds conditioned on $\mathcal{E}_1$, if we do not condition on $\mathcal{E}_1$,

$$\Pr\left\{|\|f-f'\|_{\mathcal{Z}}^2-\|f-f'\|_{\mathcal{Z}'}^2|\geq\varepsilon/4\cdot\|f-f'\|_{\mathcal{Z}}^2\right\}\leq(\delta/4)/\left(\mathcal{N}(\mathcal{F},\varepsilon_0)\right)^2.$$

By union bound, the inequality above implies that with probability at least $1-\delta/4$, for any $f,f'\in\mathcal{C}(\mathcal{F},\varepsilon_0)$,

$$(1-\varepsilon/4)\|f-f'\|_{\mathcal{Z}}^2\leq\|f-f'\|_{\mathcal{Z}'}^2\leq(1+\varepsilon/4)\|f-f'\|_{\mathcal{Z}}^2.$$

Denote the event above and the event in Lemma A.1 by $\mathcal{E}_2$, where

$$\begin{aligned}&\mathcal{E}_2=\{|\mathcal{Z}'|\leq4|\mathcal{Z}|/\delta\}\\ &\cap\left\{(1-\varepsilon/4)\|f-f'\|_{\mathcal{Z}}^2\leq\|f-f'\|_{\mathcal{Z}'}^2\leq(1+\varepsilon/4)\|f-f'\|_{\mathcal{Z}}^2,\forall f,f'\in\mathcal{C}(\mathcal{F},\varepsilon_0)\right\}.\end{aligned}$$

Now we condition on $\mathcal{E}_2$ where $\Pr\{\mathcal{E}_2\}\geq1-\delta/2$. For any function pair $f,f'\in\mathcal{F}$ where $\|f-f'\|_{\mathcal{Z}}^2\geq2\lambda$, there exists $\hat{f},\hat{f}'\in\mathcal{C}(\mathcal{F},\varepsilon_0)$, s.t.

$$\|f-\hat{f}\|_\infty\leq\varepsilon_0=\varepsilon/72\cdot\sqrt{\lambda\delta/|\mathcal{Z}|}\leq\sqrt{\lambda/(25|\mathcal{Z}|)},\|f'-\hat{f}'\|_\infty\leq\sqrt{\lambda/(25|\mathcal{Z}|)}.$$

Therefore,

$$(1-\varepsilon/4)\|\hat{f}-\hat{f}'\|_{\mathcal{Z}}^2\leq\|\hat{f}-\hat{f}'\|_{\mathcal{Z}'}^2\leq(1+\varepsilon/4)\|\hat{f}-\hat{f}'\|_{\mathcal{Z}}^2$$

by $\mathcal{E}_2$. Then we can obtain that

$$\begin{aligned}
\|f-f'\|_{\mathcal{Z}'}^2 &\le \Big(\|f-\hat{f}\|_{\mathcal{Z}'}+\|\hat{f}-\hat{f}'\|_{\mathcal{Z}'}+\|\hat{f}'-f'\|_{\mathcal{Z}'}\Big)^2 \\
&\le \Big((1+\varepsilon/8)\|\hat{f}-\hat{f}'\|_{\mathcal{Z}}+2\sqrt{|\mathcal{Z}'|}\cdot\varepsilon_0\Big)^2 \\
&= \Big((1+\varepsilon/8)\|\hat{f}-\hat{f}'\|_{\mathcal{Z}}+2\sqrt{|\mathcal{Z}'|}\cdot\varepsilon/72\cdot\sqrt{\lambda\delta/|\mathcal{Z}|}\Big)^2 \\
&\overset{|\mathcal{Z}'|\le 4|\mathcal{Z}|/\delta}{\le} \Big((1+\varepsilon/8)\|\hat{f}-\hat{f}'\|_{\mathcal{Z}}+\sqrt{\lambda}\cdot\varepsilon/18\Big)^2 \\
&\le \Big((1+\varepsilon/8)\|f-f'\|_{\mathcal{Z}}+\sqrt{\lambda}\cdot\varepsilon/18+2\|\hat{f}-f\|_{\mathcal{Z}}+2\|\hat{f}'-f'\|_{\mathcal{Z}}\Big)^2 \\
&\le \Big((1+\varepsilon/8)\|f-f'\|_{\mathcal{Z}}+\sqrt{\lambda}\cdot\varepsilon/18+4\sqrt{|\mathcal{Z}|}\cdot\varepsilon/72\cdot\sqrt{\lambda\delta/|\mathcal{Z}|}\Big)^2 \\
&\le \Big((1+\varepsilon/8)\|f-f'\|_{\mathcal{Z}}+\sqrt{\lambda}\cdot\varepsilon/9\Big)^2 \\
&\overset{\|f-f'\|_{\mathcal{Z}}\ge\sqrt{\lambda}}{\le} (1+\varepsilon)\|f-f'\|_{\mathcal{Z}}^2.
\end{aligned}$$

By similar methods, we can also obtain that

$$\begin{aligned}
\|f-f'\|_{\mathcal{Z}'}^2 &\ge \Big(\|\hat{f}-\hat{f}'\|_{\mathcal{Z}'}-\|f-\hat{f}\|_{\mathcal{Z}'}-\|\hat{f}'-f'\|_{\mathcal{Z}'}\Big)^2 \\
&\ge \Big((1-\varepsilon/6)\|\hat{f}-\hat{f}'\|_{\mathcal{Z}}-2\sqrt{|\mathcal{Z}'|}\cdot\varepsilon/72\cdot\sqrt{\lambda\delta/|\mathcal{Z}|}\Big)^2 \\
&\overset{|\mathcal{Z}'|\le 4|\mathcal{Z}|/\delta}{\ge} \Big((1-\varepsilon/6)\|\hat{f}-\hat{f}'\|_{\mathcal{Z}}-\sqrt{\lambda}\cdot\varepsilon/18\Big)^2 \\
&\ge \Big((1-\varepsilon/6)\|f-f'\|_{\mathcal{Z}}-\sqrt{\lambda}\cdot\varepsilon/18-\|\hat{f}-f\|_{\mathcal{Z}}-\|\hat{f}'-f'\|_{\mathcal{Z}}\Big)^2 \\
&\ge \Big((1-\varepsilon/6)\|f-f'\|_{\mathcal{Z}}-\sqrt{\lambda}\cdot\varepsilon/18-2\sqrt{|\mathcal{Z}|}\cdot\varepsilon/72\cdot\sqrt{\lambda\delta/|\mathcal{Z}|}\Big)^2 \\
&\ge \Big((1-\varepsilon/6)\|f-f'\|_{\mathcal{Z}}-\sqrt{\lambda}\cdot\varepsilon/12\Big)^2 \\
&\overset{\|f-f'\|_{\mathcal{Z}}\ge\sqrt{\lambda}}{\ge} (1-\varepsilon)\|f-f'\|_{\mathcal{Z}}^2.
\end{aligned}$$

□

We also give the bound of the number of distinct elements in $\mathcal{Z}'$.

Lemma A.3. *With probability at least* $1-\delta/4$, $\mathit{Card}_d(\mathcal{Z}') \le 2304L_1\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/\varepsilon^2$.

Proof. First, note that

$$p \le 768L_1\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/(\varepsilon^2\cdot|\mathcal{Z}|)$$

since for any $0<x<1$, there must exists $\hat{x}\in[x,2x]$ s.t. $1/\hat{x}$ is an integer.

When $p=1$, which means $\mathcal{Z}=\mathcal{Z}'$ and

$$768L_1\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/(\varepsilon^2\cdot|\mathcal{Z}|)\ge 1,$$

we have

$$|\mathcal{Z}'|=|\mathcal{Z}|\le 768L_1\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/\varepsilon^2.$$

When $p<1$, we have $p\ge 384L_1\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/(\varepsilon^2\cdot|\mathcal{Z}|)$. Now, For each $z\in\mathcal{Z}$, define

$$X_z=\begin{cases}1 & z \text{ is added into } \mathcal{Z}' \text{ for } 1/p \text{ times}\\ 0 & \text{otherwise}\end{cases}.$$

Then the number of distinct elements in $\mathcal{Z}'$ is upper bounded by $\sum_{z\in\mathcal{Z}} X_z$. Since $\mathbb{E}[X_z] = p$,

$$\sum_{z\in\mathcal{Z}} \mathbb{E}[X_z] = p\cdot|\mathcal{Z}| \le 768L_1\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/\varepsilon^2.$$

By Chernoff bound,

$$\begin{aligned}&\Pr\left\{\sum_{z\in\mathcal{Z}} X_z \ge 3\times 768L_1\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/\varepsilon^2\right\} \le \Pr\left\{\sum_{z\in\mathcal{Z}} X_z \ge 3\sum_{z\in\mathcal{Z}}\mathbb{E}[X_z]\right\}\\ \le&\exp\left\{-p\cdot|\mathcal{Z}|\right\} \le \exp\left\{-384L_1\cdot\ln(4\mathcal{N}(\mathcal{F},\varepsilon_0)/\delta)/\varepsilon^2\right\} \le \exp\left\{-\ln(4/\delta)\right\} = \delta/4.\end{aligned}$$

□

A.2 Analysis of Algorithm 3

In this subsection, all the notation δ refer to δ in Algorithm 3. In other words, we replace all the δ in Appendix A.1 by $\delta/(16T)$. Also, we still assume that the input dataset of Algorithm 3 is $\mathcal{Z} = \{(s_h^k, a_h^k)\}_{(h,k)\in[H]\times[t]}$ where more than half of the trajectories are induced by the same policy and the number of trajectories t satisfies

$$4L_1^2\ln\frac{128T\mathcal{N}(\mathcal{F},\delta/(9216T^2))^2}{\delta} \le t \le K = T/H,$$

which is satisfied if $t \ge \tau_{M_0}$ and M_0 is chosen as in Theorem 4.1.

Combining the three lemmas in Appendix A.1 with a union bound, we can obtain the following proposition.

Proposition A.4. *Let $\mathcal{Z}'$ denote the dataset returned by Algorithm 2. With probability at least $1-\delta/(16T)$, $|Z'| \le 64T^2/\delta$, the number of distinct elements in $\mathcal{Z}'$ does not exceed*

$$9216L_1\cdot\ln(64T\mathcal{N}(\mathcal{F},\delta/(9216T^2))/\delta),$$

and for any $f, f' \in \mathcal{F}$,

$$\|f-f'\|_{\mathcal{Z}}^2/2 - 1/2 \le \|f-f'\|_{\mathcal{Z}'}^2 \le 3\|f-f'\|_{\mathcal{Z}}^2/2 + 1/2.$$

By Proposition A.4, we can deduce the following proposition.

Proposition A.5. *For Algorithm 3, the following holds.*

1. *With probability at least $1-\delta/(16T)$,*

$$w(\underline{\mathcal{F}}, s, a) \le \hat{w}(s,a) \le w(\overline{\mathcal{F}}, s, a),$$

where $\underline{\mathcal{F}} = \left\{f\in\mathcal{F} \mid \|f-\bar{f}\|_{\mathcal{Z}}^2 \le \beta(\mathcal{F},\delta)\right\}$ and $\overline{\mathcal{F}} = \left\{f\in\mathcal{F} \mid \|f-\bar{f}\|_{\mathcal{Z}}^2 \le 12\beta(\mathcal{F},\delta)+12\right\}$.

2. *There exists a function set $\mathcal{W}$ s.t. $\hat{w}(\cdot,\cdot)\in\mathcal{W}$ and*

$$\begin{aligned}\ln|\mathcal{W}| \le& 9216L_1\cdot\ln\left(64T\mathcal{N}(\mathcal{F},\delta/(9216T^2))/\delta\right)\ln\left(\mathcal{N}(\mathcal{S}\times\mathcal{A},1/(8\sqrt{64T^2/\delta}))\times 64T^2/\delta\right)\\ &+\ln\left(\mathcal{N}(\mathcal{F},1/(8\sqrt{64T^2/\delta}))\right)+1\\ \le& C\cdot L_1\cdot\ln\left(\mathcal{N}(\mathcal{F},\delta/T^3)\times T/\delta\right)\ln\left(\mathcal{N}(\mathcal{S}\times\mathcal{A},\delta/T^2)\times T/\delta\right)\end{aligned}$$

for some absolute constant $C>0$ when T is sufficiently large.

Proof. For the first part, we condition on the event defined in Proposition A.4. We only need to prove that $\underline{\mathcal{F}} \subseteq \widehat{\mathcal{F}} \subseteq \overline{\mathcal{F}}$, where $\widehat{\mathcal{F}}$ is defined in Algorithm 3. For any $f\in\mathcal{F}$, we have

$$\|f-\bar{f}\|_{\mathcal{Z}}^2/2 - 1/2 \le \|f-\bar{f}\|_{\mathcal{Z}'}^2 \le 3\|f-\bar{f}\|_{\mathcal{Z}}^2/2 + 1/2.$$

Therefore,

$$\begin{aligned}\|f-\hat{f}\|_{\widehat{\mathcal{Z}}}^2 \le& \left(\|f-\hat{f}\|_{\mathcal{Z}'}+\sqrt{64T^2/\delta}/(4\sqrt{64T^2/\delta})\right)^2\\ \le& \left(\|f-\bar{f}\|_{\mathcal{Z}'}+\sqrt{64T^2/\delta}/(8\sqrt{64T^2/\delta})+\sqrt{64T^2/\delta}/(4\sqrt{64T^2/\delta})\right)^2\\ \le& 2\|f-\bar{f}\|_{\mathcal{Z}'}^2+1/2\le 3\|f-\bar{f}\|_{\mathcal{Z}}^2+2.\end{aligned}$$

This means for any $f\in\underline{\mathcal{F}}$, we have $\|f-\bar{f}\|_{\mathcal{Z}}^2\le\beta(\mathcal{F},\delta)$, which implies $\|f-\hat{f}\|_{\widehat{\mathcal{Z}}}^2\le 3\beta(\mathcal{F},\delta)+2$, i.e., $f\in\widehat{\mathcal{F}}$. Similarly,

$$\begin{aligned}\|f-\hat{f}\|_{\widehat{\mathcal{Z}}}^2 \ge& \left(\|f-\hat{f}\|_{\mathcal{Z}'}-\sqrt{64T^2/\delta}/(4\sqrt{64T^2/\delta})\right)^2\\ \ge& \left(\|f-\bar{f}\|_{\mathcal{Z}'}-\sqrt{64T^2/\delta}/(8\sqrt{64T^2/\delta})-\sqrt{64T^2/\delta}/(4\sqrt{64T^2/\delta})\right)^2\\ \ge& \|f-\bar{f}\|_{\mathcal{Z}'}^2/2-1/4\ge\|f-\bar{f}\|_{\mathcal{Z}}^2/4-1.\end{aligned}$$

So for any $f\in\widehat{\mathcal{F}}$, we have $\|f-\hat{f}\|_{\widehat{\mathcal{Z}}}^2\le 3\beta(\mathcal{F},\delta)+2$, which implies $\|f-\bar{f}\|_{\mathcal{Z}}^2\le 12\beta(\mathcal{F},\delta)+12$, i.e., $f\in\overline{\mathcal{F}}$.

For the second part, since function $\hat{w}(\cdot,\cdot)$ is uniquely defined by $\widehat{\mathcal{F}}$, we only need to analyze the maximal number of different possible function classes $\widehat{\mathcal{F}}$. When $|\mathcal{Z}'|>64T^2/\delta$ or the number of distinct elements in $\mathcal{Z}'$ is larger than

$$9216L_1\cdot\ln(64T\mathcal{N}(\mathcal{F},\delta/(9216T^2))/\delta),$$

$|\mathcal{Z}'|=0$ and thus $\widehat{\mathcal{F}}=\mathcal{F}$. Otherwise, $\widehat{\mathcal{F}}$ is determined by $\widehat{\mathcal{Z}}$ and $\hat{f}$. Since $\hat{f}\in\mathcal{C}(\mathcal{F},1/(8\sqrt{64T^2/\delta}))$, the number of different $\hat{f}$ does not exceed $\mathcal{N}(\mathcal{F},1/(8\sqrt{64T^2/\delta}))$. Moreover, since there are at most

$$9216L_1\cdot\ln(64T\mathcal{N}(\mathcal{F},\delta/(9216T^2))/\delta)$$

distinct elements in $\widehat{\mathcal{Z}}$, where $|\widehat{\mathcal{Z}}|\le 64T^2/\delta$ and each element belongs to $\mathcal{C}(\mathcal{S}\times\mathcal{A},1/(8\sqrt{64T^2/\delta}))$, the number of different $\widehat{\mathcal{Z}}$ is upper bounded by

$$\left(\mathcal{N}(\mathcal{S}\times\mathcal{A},1/(8\sqrt{64T^2/\delta}))\times 64T^2/\delta\right)^{9216L_1\cdot\ln(64T\mathcal{N}(\mathcal{F},\delta/(9216T^2))/\delta)}.$$

□

B ANALYSIS OF THE MAIN ALGORITHM

Now we start to prove the regret bound of Algorithm 1. The following lemma provides a bound on the estimation of a single backup.

Lemma B.1 (Single step optimization error)**.** *Consider a fixed epoch $m\in[M]\backslash[M_0]$. We define*

$$\mathcal{Z}^m=\left\{(s_h^k,a_h^k)\right\}_{(h,k)\in[H]\times[\tau_m-1]}$$

as in Algorithm 1. Also, for any function $V:\mathcal{S}\to[0,H]$, we define

$$\mathcal{D}_V^m=\left\{\left(s_h^k,a_h^k,r_h^k+V(s_{h+1}^k)\right)\right\}_{(h,k)\in[H]\times[\tau_m-1]}$$

and

$$\hat{f}_V=\arg\min_{f\in\mathcal{F}}\|f\|_{\mathcal{D}_V^m}^2.$$

Then, for any function $V:\mathcal{S}\to[0,H]$ and $\delta\in(0,1)$, there exists an event $\mathcal{E}_{V,\delta}$ where $\Pr\{\mathcal{E}_{V,\delta}\}\ge 1-\delta$, s.t. conditioned on $\mathcal{E}_{V,\delta}$, for any $V':\mathcal{S}\to[0,H]$ with $\|V-V'\|_\infty\le 1/T$, we have

$$\left\|\hat{f}_{V'}(\cdot,\cdot)-r(\cdot,\cdot)-\sum_{s'\in\mathcal{S}}P(s'|\cdot,\cdot)V'(s')\right\|_{\mathcal{Z}^m}\le c'H\sqrt{\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T)}.$$

for some constant $c'>0$.

Proof. For any $V : \mathcal{S} \to [0, H]$, we define

$$f_V(\cdot,\cdot) = r(\cdot,\cdot) + \sum_{s' \in \mathcal{S}} P(s'|\cdot,\cdot)V(s'),$$

and now we consider a fixed V. For any $f \in \mathcal{F}$, define

$$\xi_h^k(f) = 2(f(s_h^k, a_h^k) - f_V(s_h^k, a_h^k)) \cdot (f_V(s_h^k, a_h^k) - r_h^k - V(s_{h+1}^k)), \forall (h,k) \in [H] \times [\tau_m - 1].$$

Also, for any $(h,k) \in [H] \times [\tau_m - 1]$, define $\mathbb{F}_h^k$ as the filtration induced by

$$\{(s_{h'}^{k'}, a_{h'}^{k'}, r_{h'}^{k'})\}_{(h',k') \in [H] \times [k-1]} \cup \{(s_{h'}^k, a_{h'}^k, r_{h'}^k)\}_{h' \in [h]}.$$

Then we have $\mathbb{E}[\xi_h^k(f)|\mathbb{F}_h^k] = 0$ and $\mathbb{E}[(\xi_h^k(f))^2|\mathbb{F}_h^k] \le 4(H+1)^2(f(s_h^k, a_h^k) - f_V(s_h^k, a_h^k))^2$. Applying Lemma 10 of Kirschner and Krause (2018) by setting $\{X_t\} = \{\xi_h^k(f)\}$, we can obtain that with probability at least $1 - \delta$,

$$\sum_{(h,k) \in [H] \times [\tau_m - 1]} \xi_h^k(f) \le 8(H+1)^2 \log \frac{2T+2}{\delta} + 4(H+1)\|f - f_V\|_{\mathcal{Z}^m} \sqrt{\log \frac{2T+2}{\delta}}.$$

Applying a union bound of $\xi_h^k(f), -\xi_h^k(f)$ over all $f \in \mathcal{C}(\mathcal{F}, 1/T)$, we can further obtain that with probability at least $1 - \delta$,

$$\left| \sum_{(h,k) \in [H] \times [\tau_m - 1]} \xi_h^k(f) \right| \le O\left(H^2(\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)) + H\|f - f_V\|_{\mathcal{Z}^m} \sqrt{\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)} \right)$$

holds for all $f \in \mathcal{C}(\mathcal{F}, 1/T)$.

Let $\mathcal{E}_{V,\delta}$ denote the above event, and for the rest of the proof, we condition on $\mathcal{E}_{V,\delta}$.

Now, for any $f \in \mathcal{F}$, there exists a function $g \in \mathcal{C}(\mathcal{F}, 1/T)$, s.t. $\|f - g\|_\infty \le 1/T$. Therefore,

$$\begin{aligned}
\left| \sum_{(h,k) \in [H] \times [\tau_m - 1]} \xi_h^k(f) \right| \le & \left| \sum_{(h,k) \in [H] \times [\tau_m - 1]} \xi_h^k(g) \right| + 2(H+1)\|f - g\|_\infty |\mathcal{Z}^m| \\
\lesssim & H^2(\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)) + H\|g - f_V\|_{\mathcal{Z}^m} \sqrt{\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)} \\
\lesssim & H^2(\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)) + H\|f - f_V\|_{\mathcal{Z}^m} \sqrt{\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)}.
\end{aligned}$$

For any $V' : \mathcal{S} \to [0, H]$ with $\|V' - V\|_\infty \le 1/T$, we can obtain that

$$\|f_{V'} - f_V\|_\infty = \left\| \sum_{s' \in \mathcal{S}} P(s'|\cdot,\cdot)(V'(s') - V(s')) \right\|_\infty \le \|V' - V\|_\infty \le 1/T.$$

Furthermore, for any $f \in \mathcal{F}$,

$$\begin{aligned}
& \|f\|_{\mathcal{D}_{V'}^m}^2 - \|f_{V'}\|_{\mathcal{D}_{V'}^m}^2 - \|f - f_{V'}\|_{\mathcal{Z}^m}^2 \\
= & 2 \sum_{(s_h^k, a_h^k) \in \mathcal{Z}^m} (f(s_h^k, a_h^k) - f_{V'}(s_h^k, a_h^k)) \cdot (f_{V'}(s_h^k, a_h^k) - r_h^k - V'(s_{h+1}^k)) \\
\ge & 2 \sum_{(s_h^k, a_h^k) \in \mathcal{Z}^m} (f(s_h^k, a_h^k) - f_V(s_h^k, a_h^k)) \cdot (f_V(s_h^k, a_h^k) - r_h^k - V(s_{h+1}^k)) - 6(H+1) \\
= & \sum_{(h,k) \in [H] \times [\tau_m - 1]} \xi_h^k(f) - 6(H+1) \\
\gtrsim & -H^2(\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)) - H\|f - f_V\|_{\mathcal{Z}^m} \sqrt{\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)} \\
\gtrsim & -H^2(\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)) - H\|f - f_{V'}\|_{\mathcal{Z}^m} \sqrt{\ln(T/\delta) + \ln \mathcal{N}(\mathcal{F}, 1/T)}.
\end{aligned}$$

If we let $f = \hat{f}_{V'}$, since $\hat{f}_{V'} = \arg\min_{f\in\mathcal{F}} \|f\|_{\mathcal{D}^m_{V'}}$, we have

$$\begin{aligned}0 \geq& \|\hat{f}_{V'}\|^2_{\mathcal{D}^m_{V'}} - \|f_{V'}\|^2_{\mathcal{D}^m_{V'}} \\ \gtrsim& \|\hat{f}_{V'} - f_{V'}\|^2_{\mathcal{Z}^m} - H^2(\ln(T/\delta) + \ln\mathcal{N}(\mathcal{F}, 1/T)) - H\|\hat{f}_{V'} - f_{V'}\|_{\mathcal{Z}^m}\sqrt{\ln(T/\delta) + \ln\mathcal{N}(\mathcal{F}, 1/T)},\end{aligned}$$

which implies

$$\|\hat{f}_{V'} - f_{V'}\|_{\mathcal{Z}^m} \leq c' H\sqrt{\ln(T/\delta) + \ln\mathcal{N}(\mathcal{F}, 1/T)}.$$

for some constant $c' > 0$. □

Lemma B.2 (Confidence region)**.** *In Algorithm 1, for $m > M_0$, define confidence region*

$$\mathcal{F}^m_h = \left\{ f \in \mathcal{F} \,\middle|\, \|f - f^m_h\|^2_{\mathcal{Z}^m} \leq \beta(\mathcal{F}, \delta) \right\}.$$

Then with probability at least $1 - \delta/16$, for all $(h, m) \in [H] \times ([M]\backslash[M_0])$,

$$r(\cdot,\cdot) + \sum_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V^m_{h+1}(s') \in \mathcal{F}^m_h,$$

given

$$\beta(\mathcal{F}, \delta) \geq c' H^2(\ln(T/\delta) + \ln\mathcal{N}(\mathcal{F}, 1/T) + \ln|\mathcal{W}|).$$

for some constant $c' > 0$. Here, $\mathcal{W}$ is given in Proposition A.5.

Proof. By Proposition A.5, $b^m_h(\cdot,\cdot) \in \mathcal{W}, \forall (h, m) \in [H] \times ([M]\backslash[M_0])$. Note that

$$\mathcal{Q} = \{\min\{f(\cdot,\cdot) + w(\cdot,\cdot), H\} | f \in \mathcal{C}(\mathcal{F}, 1/T), w \in \mathcal{W}\} \cup \{0\}$$

is a $(1/T)$-cover of

$$Q^m_{h+1}(\cdot,\cdot) = \begin{cases} \min\{f^m_{h+1}(\cdot,\cdot) + b^m_{h+1}(\cdot,\cdot), H\}, & h < H \\ 0, & h = H \end{cases},$$

i.e., there exists $q \in \mathcal{Q}$, s.t. $\|q - Q^m_{h+1}\|_\infty \leq 1/T$. Therefore,

$$\mathcal{V} = \left\{ \max_{a\in A} q(\cdot, a) | q \in \mathcal{Q} \right\}$$

is a $(1/T)$-cover of V^m_{h+1} with $\ln|\mathcal{V}| \leq \ln|\mathcal{W}| + \ln\mathcal{N}(\mathcal{F}, 1/T) + 1$.

Now, for each $V \in \mathcal{V}$, let $\mathcal{E}_{V,\delta/(16|\mathcal{V}|T)}$ denote the event defined in Lemma B.1. By union bound, $\Pr\{\bigcap_{V\in\mathcal{V}} \mathcal{E}_{V,\delta/(16|\mathcal{V}|T)}\} \geq 1 - \delta/(16T)$. In the rest of the proof, we condition on the event $\bigcap_{V\in\mathcal{V}} \mathcal{E}_{V,\delta/(16|\mathcal{V}|T)}$.

Since $f^m_h = \arg\min_{f\in\mathcal{F}} \|f\|^2_{\mathcal{D}^m_h}$, and there exists $V \in \mathcal{V}$ s.t. $\|V - V^m_{h+1}\|_\infty \leq 1/T$, by Lemma B.1, we have

$$\left\| f^m_h(\cdot,\cdot) - r(\cdot,\cdot) - \sum_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V^m_{h+1}(s') \right\|_{\mathcal{Z}^m} \leq c' H\sqrt{\ln(T/\delta) + \ln\mathcal{N}(\mathcal{F}, 1/T) + \ln|\mathcal{W}|}$$

for some constant $c' > 0$. Applying a union bound over all $(h, m) \in [H] \times ([M]\backslash[M_0])$, we have that with probability at least $1 - \delta/16$,

$$r(\cdot,\cdot) + \sum_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V^m_{h+1}(s') \in \mathcal{F}^m_h, \forall (h, m) \in [H] \times ([M]\backslash[M_0]).$$

□

The above lemma proves that the confidence region contains $r(\cdot,\cdot) + \sum_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V^m_{h+1}(s')$ with high probability, which implies that all the estimated Q-function Q^m_h are optimistic with high probability as well. We formally state the conclusion in the next lemma.

Lemma B.3 (Optimistic Q-function)**.** *With probability at least* $1-\delta/8$*,*

$$Q_h^*(s,a) \le Q_h^m(s,a) \le r(s,a) + \sum_{s'\in\mathcal{S}} P(s'|s,a)V_{h+1}^m(s') + 2b_h^m(s,a)$$

for all $(h,m)\in[H]\times([M]\backslash[M_0])$ *and* $(s,a)\in\mathcal{S}\times\mathcal{A}$*.*

Proof. Let $\mathcal{F}_h^m$ be the confidence region as defined in Lemma B.2. Let $\mathcal{E}_1$ denote the event that

$$r(\cdot,\cdot) + \sum_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V_{h+1}^m(s') \in \mathcal{F}_h^m, \forall (h,m)\in[H]\times([M]\backslash[M_0]).$$

By Lemma B.2, $\Pr\{\mathcal{E}_1\}\ge 1-\delta/16$. Let $\mathcal{E}_2$ denote the event that

$$b_h^m(s,a) \ge w(\mathcal{F}_h^m, s, a), \forall (h,m)\in[H]\times([M]\backslash[M_0]), (s,a)\in\mathcal{S}\times\mathcal{A}.$$

By Proposition A.5 and union bound over all $(h,m)\in[H]\times([M]\backslash[M_0])$, $\Pr\{\mathcal{E}_2\}\ge 1-\delta/16$. We condition on $\mathcal{E}_1\cap\mathcal{E}_2$ in the rest of the proof, which holds with failure probability at most $\delta/8$.

By the definition of width function,

$$\max_{f\in\mathcal{F}_h^m} |f(s,a) - f_h^m(s,a)| \le w(\mathcal{F}_h^m, s, a) \le b_h^m(s,a), \forall (s,a)\in\mathcal{S}\times\mathcal{A}.$$

Since $r(\cdot,\cdot) + \sum_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V_{h+1}^m(s') \in \mathcal{F}_h^m$, we have

$$\left| r(s,a) + \sum_{s'\in\mathcal{S}} P(s'|s,a)V_{h+1}^m(s') - f_h^m(s,a) \right| \le b_h^m(s,a), \forall (s,a)\in\mathcal{S}\times\mathcal{A}. \tag{2}$$

Therefore, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$Q_h^m(s,a) \le f_h^m(s,a) + b_h^m(s,a) \le r(s,a) + \sum_{s'\in\mathcal{S}} P(s'|s,a)V_{h+1}^m(s') + 2b_h^m(s,a).$$

Next, we start to prove $Q_h^*(\cdot,\cdot)\le Q_h^m(\cdot,\cdot)$ by induction on h. When $h=H+1$, the inequality directly holds since $Q_{H+1}^*(\cdot,\cdot) = Q_{H+1}^m(\cdot,\cdot) = 0$. Now for any $h\in[H]$, assume $Q_{h+1}^*(\cdot,\cdot)\le Q_{h+1}^m(\cdot,\cdot)$. This also implies $V_{h+1}^*(\cdot)\le V_{h+1}^m(\cdot)$. Therefore, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
Q_h^*(s,a) =& r(s,a) + \sum_{s'\in\mathcal{S}} P(s'|s,a)V_{h+1}^*(s') \\
\le& \min\left\{ H, r(s,a) + \sum_{s'\in\mathcal{S}} P(s'|s,a)V_{h+1}^m(s') \right\} \\
\overset{(2)}{\le}& \min\left\{ H, f_h^m(s,a) + b_h^m(s,a) \right\} = Q_h^m(s,a),
\end{aligned}$$

which completes the proof. □

Now, we can decompose the regret and bound it by the summation of bonus functions.

Lemma B.4 (Regret decomposition)**.** *With probability at least* $1-\delta/4$*,*

$$\mathit{Reg}(K) \le \tau_{M_0+1}\cdot H + 2\sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\sum_{h=1}^{H} b_h^m(s_h^k,a_h^k) + 8H\sqrt{T\ln(16/\delta)}$$

Proof. For any step $h\in[H]$, epoch $m\in[M]\backslash[M_0]$ and episode k in epoch m, define

$$\xi_h^k = \sum_{s'\in\mathcal{S}} P(s'|s_h^k,a_h^k)\left(V_{h+1}^m(s') - V_{h+1}^{\pi^m}(s')\right) - \left(V_{h+1}^m(s_{h+1}^k) - V_{h+1}^{\pi^m}(s_{h+1}^k)\right),$$

and define $\mathbb{F}_h^k$ as the filtration induced by

$$\{(s_{h'}^{k'}, a_{h'}^{k'}, r_{h'}^{k'})\}_{(h',k')\in[H]\times[k-1]} \cup \{(s_{h'}^{k}, a_{h'}^{k}, r_{h'}^{k})\}_{h'\in[h-1]}.$$

Then $\mathbb{E}[\xi_h^k|\mathbb{F}_h^k] = 0$ and $|\xi_h^k| \le 2H$. By Azuma-Hoeffding inequality, with probability at least $1-\delta/8$,

$$\sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\sum_{h=1}^{H}\xi_h^k \le 8H\sqrt{T\ln(16/\delta)}.$$

We condition on both this event and the event defined in Lemma B.3 which also holds with probability at least $1-\delta/8$ in the rest of the proof.

Let π^0 denote the uniformly random policy adopted in the first (M_0-1) epochs. By Lemma B.3,

$$\begin{aligned}\mathrm{Reg}(K) =& \sum_{k=1}^{\tau_{M_0}-1}\left(V_1^*(s_1^k) - V_1^{\pi^0}(s_1^k)\right) + \sum_{m=M_0}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\left(V_1^*(s_1^k) - V_1^{\pi^m}(s_1^k)\right)\\ \le& \tau_{M_0+1}\cdot H + \sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\left(V_1^m(s_1^k) - V_1^{\pi^m}(s_1^k)\right).\end{aligned}$$

For each k and corresponding m, we have

$$\begin{aligned}&V_1^m(s_1^k) - V_1^{\pi^m}(s_1^k)\\ =&Q_1^m(s_1^k,a_1^k) - r(s_1^k,a_1^k) - \sum_{s'\in\mathcal{S}}P(s'|s_1^k,a_1^k)V_2^{\pi^m}(s')\\ \le& r(s_1^k,a_1^k) + \sum_{s'\in\mathcal{S}}P(s'|s_1^k,a_1^k)V_2^m(s') + 2b_1^m(s_1^k,a_1^k) - r(s_1^k,a_1^k) - \sum_{s'\in\mathcal{S}}P(s'|s_1^k,a_1^k)V_2^{\pi^m}(s')\\ =&\sum_{s'\in\mathcal{S}}P(s'|s_1^k,a_1^k)(V_2^m(s') - V_2^{\pi^m}(s')) + 2b_1^m(s_1^k,a_1^k)\\ =&(V_2^m(s_2^k) - V_2^{\pi^m}(s_2^k)) + \xi_1^k + 2b_1^m(s_1^k,a_1^k)\\ \le&(V_3^m(s_3^k) - V_3^{\pi^m}(s_3^k)) + \xi_1^k + \xi_2^k + 2b_1^m(s_1^k,a_1^k) + 2b_2^m(s_2^k,a_2^k)\\ \le&\cdots\\ \le&\sum_{h=1}^{H}\left(\xi_h^k + 2b_h^m(s_h^k,a_h^k)\right).\end{aligned}$$

Therefore,

$$\mathrm{Reg}(K) \le \tau_{M_0+1}\cdot H + 2\sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\sum_{h=1}^{H}b_h^m(s_h^k,a_h^k) + 8H\sqrt{T\ln(16/\delta)}.$$

□

To prove the main theorem, we also need the next lemma.

Lemma B.5. *With probability at least $1-\delta/2$, for all $(h,m)\in[H]\times([M]\backslash[M_0])$ and any $f, f'\in\mathcal{F}$,*

$$T_{m-1}\mathbb{E}_{s\sim\mathcal{D}_h(\pi^{m-1}),a\sim\pi_h^{m-1}(s)}[(f(s,a)-f'(s,a))^2] \le 4\sum_{k=\tau_{m-1}}^{\tau_m-1}(f(s_h^k,a_h^k)-f'(s_h^k,a_h^k))^2 + 64.$$

Proof. We first fix any $(h,m)\in[H]\times([M]\backslash[M_0])$. Define dataset

$$\mathcal{Z}_h^m = \{(s_h^k,a_h^k)\}_{k\in[\tau_{m-1},\tau_m-1]}.$$

Now we fix any pair of distinct functions $f, f' \in \mathcal{C}(\mathcal{F}, 1/T)$. Also, for any episode $k \in [\tau_{m-1},\ \tau_m\ -\ 1\ ]$, let

$$\xi_h^k = (f(s_h^k, a_h^k) - f'(s_h^k, a_h^k))^2.$$

Therefore,

$$\mathbb{E}\left[\xi_h^k\right] = \mathbb{E}_{s\sim\mathcal{D}_h(\pi^{m-1}), a\sim\pi_h^{m-1}(s)}\left[(f(s,a) - f'(s,a))^2\right].$$

Note that

$$0 \le \xi_h^k \le \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} (f(s,a) - f'(s,a))^2 = \|f - f'\|_\infty^2.$$

Also, by Definition 3.3,

$$\mathbb{E}\left[\xi_h^k\right] \ge \frac{1}{L_1} \max_{s\in\mathcal{S}, a\in\mathcal{A}} (f(s,a) - f'(s,a))^2 = \frac{1}{L_1}\|f - f'\|_\infty^2.$$

Therefore, by Hoeffding's inequality,

$$\begin{aligned}
&\Pr\left\{\frac{1}{T_{m-1}}\sum_{k=\tau_{m-1}}^{\tau_m-1}\left(\xi_h^k - \mathbb{E}\left[\xi_h^k\right]\right) \le -v\mathbb{E}\left[\xi_h^{\tau_{m-1}}\right]\right\} \le \exp\left(-\frac{2T_{m-1}^2 v^2 \mathbb{E}\left[\xi_h^{\tau_{m-1}}\right]^2}{T_{m-1}\|f - f'\|_\infty^4}\right) \\
\le &\exp\left(-\frac{2T_{m-1}v^2}{\|f - f'\|_\infty^4} \cdot \frac{\|f - f'\|_\infty^4}{L_1^2}\right) \le \exp\left(-\frac{2T_{m-1}v^2}{L_1^2}\right).
\end{aligned}$$

Since

$$T_{m-1} = 2^{m-2} \ge 2^{M_0 - 1} \ge 8L_1^2 \ln\frac{128T\mathcal{N}(\mathcal{F}, \delta/(9216T^2))^2}{\delta} \ge 2L_1^2 \ln\frac{2T\mathcal{N}(\mathcal{F}, 1/T)^2}{\delta},$$

by setting $v = \frac{1}{2}$, we can obtain that

$$\begin{aligned}
&\Pr\left\{\frac{1}{T_{m-1}}\sum_{k=1}^{u}\xi_h^k \le \frac{1}{2}\mathbb{E}\left[\xi_h^{\tau_{m-1}}\right]\right\} \le \exp\left(-\frac{v^2}{L_1^2}\cdot 4L_1^2 \ln\frac{2T\mathcal{N}(\mathcal{F}, 1/T)^2}{\delta}\right) \\
\le &\exp\left(-\ln\frac{2T\mathcal{N}(\mathcal{F}, 1/T)^2}{\delta}\right) \le \frac{\delta}{2T\mathcal{N}(\mathcal{F}, 1/T)^2}.
\end{aligned}$$

By a union bound over all such function pairs (f, f'), this implies that with probaiblity at least $1 - \delta/(2T)$, for any $f, f' \in \mathcal{C}(\mathcal{F}, 1/T)$,

$$T_{m-1}\mathbb{E}_{s\sim\mathcal{D}_h(\pi^{m-1}), a\sim\pi_h^{m-1}(s)}\left[(f(s,a) - f'(s,a))^2\right] \le 2\|f - f'\|_{\mathcal{Z}_h^m}^2.$$

Now we condition on the event above in the following part of the proof.

To simplify the notation, we denote

$$\|f - f'\|_{\pi_h^{m-1}}^2 = \mathbb{E}_{s\sim\mathcal{D}_h(\pi^{m-1}), a\sim\pi_h^{m-1}(s)}\left[(f(s,a) - f'(s,a))^2\right], \forall f, f' \in \mathcal{F}.$$

For any pair of functions $f, f' \in \mathcal{F}$, there exists $\hat{f}, \hat{f}' \in \mathcal{C}(\mathcal{F}, 1/T)$, s.t. $\|f - \hat{f}\|_\infty \le 1/T$ and $\|f' - \hat{f}'\|_\infty \le 1/T$. When $\|f - f'\|_{\pi_h^{m-1}}^2 \le 64/T_{m-1}$, we can directly obtain that

$$T_{m-1}\|f - f'\|_{\pi_h^{m-1}}^2 \le 4\|f - f'\|_{\mathcal{Z}_h^m}^2 + 64.$$

So we only consider the case when $\|f - f'\|_{\pi_h^{m-1}}^2 \ge 64/T_{m-1}$. Then, we have

$$\begin{aligned}
\|f - f'\|_{\mathcal{Z}_h^m} \ge &\|\hat{f} - \hat{f}'\|_{\mathcal{Z}_h^m} - \|f - \hat{f}\|_{\mathcal{Z}_h^m} - \|f' - \hat{f}'\|_{\mathcal{Z}_h^m} \\
\ge &\sqrt{T_{m-1}/2}\|\hat{f} - \hat{f}'\|_{\pi_h^{m-1}} - 2/\sqrt{T} \\
\ge &\sqrt{T_{m-1}/2}\left(\|f - f'\|_{\pi_h^{m-1}} - \|f - \hat{f}\|_{\pi_h^{m-1}} - \|f' - \hat{f}'\|_{\pi_h^{m-1}}\right) - 2/\sqrt{T} \\
\ge &\sqrt{T_{m-1}/2}\left(\|f - f'\|_{\pi_h^{m-1}} - 2/T\right) - 2/\sqrt{T} \\
\ge &\sqrt{T_{m-1}/2}\|f - f'\|_{\pi_h^{m-1}} - 4/\sqrt{T} \ge 0.
\end{aligned}$$

Therefore,

$$\|f-f'\|_{\mathcal{Z}_h^m}^2 \geq (T_{m-1}/4)\cdot\|f-f'\|_{\pi_h^{m-1}}^2 - 16/T \geq (T_{m-1}/4)\cdot\|f-f'\|_{\pi_h^{m-1}}^2 - 16,$$

which means

$$T_{m-1}\mathbb{E}_{s\sim\mathcal{D}_h(\pi^{m-1}),a\sim\pi_h^{m-1}(s)}[(f(s,a)-f'(s,a))^2] \leq 4\sum_{k=\tau_{m-1}}^{\tau_m-1}(f(s_h^k,a_h^k)-f'(s_h^k,a_h^k))^2+64.$$

Finally, we complete the proof by directly applying a union bound over all $(h,m)\in[H]\times([M]\backslash[M_0])$.

□

Now we are ready to prove the main theorem.

Proof of Theorem 4.1. We condition on the event defined in Lemma B.2, Lemma B.3, Lemma B.4 and Lemma B.5. Also, we condition on the event in Proposition A.5 after applying a union bound over all $(h,m)\in[H]\times([M]\backslash[M_0])$. With probability at least $1-\delta$, all the above events hold.

By Lemma B.4, we have

$$\mathrm{Reg}(K) \leq \tau_{M_0+1}\cdot H + 2\sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\sum_{h=1}^{H} b_h^m(s_h^k,a_h^k) + 8H\sqrt{T\ln(16/\delta)}.$$

For any $(h,m)\in[H]\times([M]\backslash[M_0])$, we define

$$\overline{\mathcal{F}}_h^m = \left\{f\in\mathcal{F} \mid \|f-f_h^m\|_{\mathcal{Z}^m}^2 \leq 12\beta(\mathcal{F},\delta)+12\right\},$$

where

$$\mathcal{Z}^m = \left\{(s_h^k,a_h^k)\right\}_{(h,k)\in[H]\times[\tau_m-1]}$$

as defined in Algorithm 1. Let

$$\textsc{High}_{\overline{\mathcal{F}}_h^m}(s,a) = \max_{f\in\overline{\mathcal{F}}_h^m} f(s,a), \quad \textsc{Low}_{\overline{\mathcal{F}}_h^m}(s,a) = \min_{f\in\overline{\mathcal{F}}_h^m} f(s,a).$$

By Proposition A.5, $b_h^m(\cdot,\cdot)\leq w(\overline{\mathcal{F}}_h^m,\cdot,\cdot)$. Then, for any episode $k\in[\tau_m,\tau_{m+1}-1]$,

$$\begin{aligned}
&\left(b_h^m(s_h^k,a_h^k)\right)^2 \leq \left(w(\overline{\mathcal{F}}_h^m,s_h^k,a_h^k)\right)^2 \leq \left(\textsc{High}_{\overline{\mathcal{F}}_h^m}(s_h^k,a_h^k)-\textsc{Low}_{\overline{\mathcal{F}}_h^m}(s_h^k,a_h^k)\right)^2\\
\leq&\left(\textsc{High}_{\overline{\mathcal{F}}_h^m}(s_h^k,a_h^k)-f_h^m(s_h^k,a_h^k)+f_h^m(s_h^k,a_h^k)-\textsc{Low}_{\overline{\mathcal{F}}_h^m}(s_h^k,a_h^k)\right)^2\\
\leq&2\left(\textsc{High}_{\overline{\mathcal{F}}_h^m}(s_h^k,a_h^k)-f_h^m(s_h^k,a_h^k)\right)^2+2\left(f_h^m(s_h^k,a_h^k)-\textsc{Low}_{\overline{\mathcal{F}}_h^m}(s_h^k,a_h^k)\right)^2\\
\leq&4\sup_{f\in\overline{\mathcal{F}}_h^m}\left(f(s_h^k,a_h^k)-f_h^m(s_h^k,a_h^k)\right)^2\\
\leq&4L_1\sup_{f\in\overline{\mathcal{F}}_h^m}\mathbb{E}_{s\sim\mathcal{D}_h(\pi^{m-1})}\mathbb{E}_{a\sim\pi_h^{m-1}(s)}\left[(f(s,a)-f_h^m(s,a))^2\right]\\
\overset{\text{Lemma B.5}}{\leq}&\frac{4L_1}{T_{m-1}}\cdot\sup_{f\in\overline{\mathcal{F}}_h^m}\left(4\sum_{k'=\tau_{m-1}}^{\tau_m-1}(f(s_h^{k'},a_h^{k'})-f_h^m(s_h^{k'},a_h^{k'}))^2+64\right)\\
\leq&\frac{4L_1}{T_{m-1}}\cdot\sup_{f\in\overline{\mathcal{F}}_h^m}\left(4\|f-f_h^m\|_{\mathcal{Z}^m}^2+64\right)\\
\leq&\frac{4L_1}{T_{m-1}}\cdot(4\times(12\beta(\mathcal{F},\delta)+12)+64)\\
=&\frac{64L_1}{T_{m-1}}\cdot(3\beta(\mathcal{F},\delta)+7).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left(\sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\sum_{h=1}^{H} b_h^m(s_h^k,a_h^k)\right)^2\\
\leq&\left(\sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\sum_{h=1}^{H}\left(b_h^m(s_h^k,a_h^k)\right)^2\right)\cdot T\\
\leq&64TL_1\sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\sum_{h=1}^{H}\frac{3\beta(\mathcal{F},\delta)+7}{T_{m-1}}\\
\leq&128TL_1HM(3\beta(\mathcal{F},\delta)+7),
\end{aligned}$$

which implies

$$2\sum_{m=M_0+1}^{M}\sum_{k=\tau_m}^{\tau_{m+1}-1}\sum_{h=1}^{H} b_h^m(s_h^k,a_h^k)\leq 32\sqrt{L_1THM(3\beta(\mathcal{F},\delta)+7)}.$$

Then, we can obtain that

$$\begin{aligned}
&\mathrm{Reg}(K)\\
\leq&2^{M_0}\cdot H+32\sqrt{L_1THM(3\beta(\mathcal{F},\delta)+7)}+8H\sqrt{T\ln(16/\delta)}\\
\leq&64L_1^2H\ln\frac{128T\mathcal{N}(\mathcal{F},\delta/(9216T^2))^2}{\delta}+32\sqrt{L_1THM(3\beta(\mathcal{F},\delta)+7)}+8H\sqrt{T\ln(16/\delta)}\\
\leq&O(L_1^2H(\ln(T/\delta)+\ln(\mathcal{N}(\mathcal{F},\delta/T^2))))+O(L_1^2H^{3/2}\ln^2(T/\delta)\cdot\max(\ln(\mathcal{N}(\mathcal{F},\delta/T^3)),\ln(\mathcal{N}(\mathcal{S}\times\mathcal{A},\delta/T^2)))\cdot\sqrt{T})\\
\leq&O(L_1H^{3/2}\ln^2(T/\delta)\cdot\max(\ln(\mathcal{N}(\mathcal{F},\delta/T^3)),\ln(\mathcal{N}(\mathcal{S}\times\mathcal{A},\delta/T^2)))\cdot\sqrt{T})
\end{aligned}$$

□

C PROOF OF PROPOSITION 3.4

In this section, we provide the proof of Proposition 3.4.

Proof of Proposition 3.4. For linear settings, let $\mathcal{W}^*=\{w-w'|w,w'\in\mathcal{W}\}$, then by Definition 3.3,

$$\begin{aligned}
L_1\leq&\sup_{\pi}\max_{h\in[H]}\sup_{w,w'\in\mathcal{W}}\frac{\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}(w^\mathsf{T}\phi(s,a)-w'^\mathsf{T}\phi(s,a))^2}{\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[(w^\mathsf{T}\phi(s',a')-w'^\mathsf{T}\phi(s',a'))^2\right]}\\
\leq&\sup_{\pi}\max_{h\in[H]}\sup_{w\in\mathcal{W}^*}\frac{\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}(w^\mathsf{T}\phi(s,a))^2}{\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[(w^\mathsf{T}\phi(s',a'))^2\right]}\\
\leq&\sup_{\pi}\max_{h\in[H]}\sup_{w\in\mathcal{W}^*}\frac{\|w\|_2^2}{\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[(w^\mathsf{T}\phi(s',a'))^2\right]}\\
\leq&\sup_{\pi}\max_{h\in[H]}\sup_{w\in\mathcal{W}^*}\frac{\|w\|_2^2}{w^\mathsf{T}\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[\phi(s',a')\phi(s',a')^\mathsf{T}\right]w}\\
\leq&\sup_{\pi}\max_{h\in[H]}\sup_{w\in\mathcal{W}^*}\frac{\|w\|_2^2}{\|w\|_2^2\lambda_{\min}\left(\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[\phi(s',a')\phi(s',a')^\mathsf{T}\right]\right)}\\
\leq&\sup_{\pi}\max_{h\in[H]}\frac{1}{\lambda_{\min}\left(\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[\phi(s',a')\phi(s',a')^\mathsf{T}\right]\right)}
\end{aligned}$$

For sparse high-dimensional linear settings, let $\mathcal{W}^* = \{w - w' | w, w' \in \mathcal{W}\}$, then by Definition 3.3,

$$\begin{aligned}
L_1 \leq & \sup_{\pi} \max_{h\in[H]} \sup_{w,w'\in\mathcal{W}} \frac{\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}(w^\mathsf{T}\phi(s,a) - w'^\mathsf{T}\phi(s,a))^2}{\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[(w^\mathsf{T}\phi(s',a') - w'^\mathsf{T}\phi(s',a'))^2\right]} \\
\leq & \sup_{\pi} \max_{h\in[H]} \sup_{w\in\mathcal{W}^*} \frac{\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}(w^\mathsf{T}\phi(s,a))^2}{\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[(w^\mathsf{T}\phi(s',a'))^2\right]} \\
\leq & \sup_{\pi} \max_{h\in[H]} \sup_{w\in\mathcal{W}^*} \frac{4s\|w\|_2^2}{\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[(w^\mathsf{T}\phi(s',a'))^2\right]} \\
\leq & \sup_{\pi} \max_{h\in[H]} \sup_{w\in\mathcal{W}^*} \frac{4s\|w\|_2^2}{w^\mathsf{T}\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[\phi(s',a')\phi(s',a')^\mathsf{T}\right]w} \\
\leq & \sup_{\pi} \max_{h\in[H]} \sup_{w\in\mathcal{W}^*} \frac{4s\|w\|_2^2}{\|w\|_2^2\psi_{\min}\left(\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[\phi(s',a')\phi(s',a')^\mathsf{T}\right]\right)} \\
\leq & \sup_{\pi} \max_{h\in[H]} \frac{4s}{\psi_{\min}\left(\mathbb{E}_{s'\sim\mathcal{D}_h(\pi)}\mathbb{E}_{a'\sim\pi_h(s')}\left[\phi(s',a')\phi(s',a')^\mathsf{T}\right]\right)}.
\end{aligned}$$

□

D PROOF OF THEOREM 5.2

In this section, we provide the proof of Theorem 5.2 for model misspecification. First, we slightly modify Lemma B.1 and reprove it in model misspecification case.

Lemma D.1 (Single step optimization error for misspecification)**.** *Assume that our function class $\mathcal{F}$ satisfies Assumption 5.1. Consider a fixed epoch $m \in [M]\backslash[M_0]$. We define*

$$\mathcal{Z}^m = \left\{(s_h^k, a_h^k)\right\}_{(h,k)\in[H]\times[\tau_m-1]}$$

as in Algorithm 1. Also, for any function $V : \mathcal{S} \to [0, H]$, we define

$$\mathcal{D}_V^m = \left\{\left(s_h^k, a_h^k, r_h^k + V(s_{h+1}^k)\right)\right\}_{(h,k)\in[H]\times[\tau_m-1]}$$

and

$$\hat{f}_V = \arg\min_{f\in\mathcal{F}} \|f\|_{\mathcal{D}_V^m}^2.$$

Then, for any function $V : \mathcal{S} \to [0, H]$ and $\delta \in (0, 1)$, there exists an event $\mathcal{E}_{V,\delta}$ where $\Pr\{\mathcal{E}_{V,\delta}\} \geq 1 - \delta$, s.t. conditioned on $\mathcal{E}_{V,\delta}$, for any $V' : \mathcal{S} \to [0, H]$ with $\|V - V'\|_\infty \leq 1/T$, we have

$$\left\|\hat{f}_{V'}(\cdot,\cdot) - r(\cdot,\cdot) - \sum_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V'(s')\right\|_{\mathcal{Z}^m} \leq c'\sqrt{H^2(\ln(T/\delta) + \ln\mathcal{N}(\mathcal{F}, 1/T)) + HT\zeta}.$$

for some constant $c' > 0$.

Proof. For any $V : \mathcal{S} \to [0, H]$, we define

$$f_V(\cdot,\cdot) = r(\cdot,\cdot) + \sum_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V(s'),$$

and now we consider a fixed V. Note that under Assumption 5.1, it does not necessary hold that $f_V \in \mathcal{F}$, but it can be ensured that

$$\min_{f\in\mathcal{F}} \|f - f_V\|_{\mathcal{Z}^m}^2 \leq |\mathcal{Z}^m|\zeta^2 \leq T\zeta^2.$$

For any $f \in \mathcal{F}$, define

$$\xi_h^k(f) = 2(f(s_h^k, a_h^k) - f_V(s_h^k, a_h^k)) \cdot (f_V(s_h^k, a_h^k) - r_h^k - V(s_{h+1}^k)),\ \forall (h,k) \in [H] \times [\tau_m - 1].$$

By the same method as in Lemma B.1, we can prove that with probability at least $1-\delta$,

$$\left|\sum_{(h,k)\in[H]\times[\tau_m-1]}\xi_h^k(f)\right| \lesssim 8(H+1)^2\log\frac{2T+2}{\delta}+4(H+1)\|f-f_V\|_{\mathcal{Z}^m}\sqrt{\log\frac{2T+2}{\delta}}.$$

Let $\mathcal{E}_{V,\delta}$ denote the above event, and for the rest of the proof, we condition on $\mathcal{E}_{V,\delta}$.

Similarly, by the same method as in Lemma B.1, for any $f\in\mathcal{F}$, we have

$$\left|\sum_{(h,k)\in[H]\times[\tau_m-1]}\xi_h^k(f)\right| \lesssim H^2(\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T))+H\|f-f_V\|_{\mathcal{Z}^m}\sqrt{\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T)}.$$

For any $V':\mathcal{S}\to[0,H]$ with $\|V'-V\|_\infty\le 1/T$, we can obtain that

$$\|f_{V'}-f_V\|_\infty=\left\|\sum_{s'\in\mathcal{S}}P(s'|\cdot,\cdot)(V'(s')-V(s'))\right\|_\infty\le\|V'-V\|_\infty\le 1/T.$$

Furthermore, again by the same method as in Lemma B.1, we can obtain that for any $f\in\mathcal{F}$,

$$\begin{aligned}&\|f\|^2_{\mathcal{D}^m_{V'}}-\|f_{V'}\|^2_{\mathcal{D}^m_{V'}}\\ \gtrsim&\|f-f_{V'}\|^2_{\mathcal{Z}^m}-H^2(\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T))-H\|f-f_{V'}\|_{\mathcal{Z}^m}\sqrt{\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T)}.\end{aligned}$$

If we let $f=\hat{f}_{V'}=\arg\min_{f\in\mathcal{F}}\|f\|_{\mathcal{D}^m_{V'}}$, we have

$$\begin{aligned}&\|\hat{f}_{V'}\|^2_{\mathcal{D}^m_{V'}}-\|f_{V'}\|^2_{\mathcal{D}^m_{V'}}\\ \gtrsim&\|\hat{f}_{V'}-f_{V'}\|^2_{\mathcal{Z}^m}-H^2(\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T))-H\|\hat{f}_{V'}-f_{V'}\|_{\mathcal{Z}^m}\sqrt{\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T)}.\end{aligned}$$

Now let $\tilde{f}_{V'}=\arg\min_{f\in\mathcal{F}}\|f-f_{V'}\|^2_{\mathcal{Z}^m}$, then

$$\begin{aligned}&\|\hat{f}_{V'}\|_{\mathcal{D}^m_{V'}}\le\|\tilde{f}_{V'}\|_{\mathcal{D}^m_{V'}}\le\|f_{V'}\|_{\mathcal{D}^m_{V'}}+\|f_{V'}-\tilde{f}_{V'}\|_{\mathcal{Z}^m}\le\|f_{V'}\|_{\mathcal{D}^m_{V'}}+\sqrt{T}\zeta\\ \Longrightarrow&\|\hat{f}_{V'}\|_{\mathcal{D}^m_{V'}}-\|f_{V'}\|_{\mathcal{D}^m_{V'}}\le\sqrt{T}\zeta\\ \Longrightarrow&\|\hat{f}_{V'}\|^2_{\mathcal{D}^m_{V'}}-\|f_{V'}\|^2_{\mathcal{D}^m_{V'}}\le\sqrt{T}\zeta(\|\hat{f}_{V'}\|_{\mathcal{D}^m_{V'}}+\|f_{V'}\|_{\mathcal{D}^m_{V'}})\le\sqrt{T}\zeta\cdot 4\sqrt{T}H=4HT\zeta.\end{aligned}$$

Therefore,

$$\|\hat{f}_{V'}-f_{V'}\|^2_{\mathcal{Z}^m}\lesssim H^2(\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T))+H\|\hat{f}_{V'}-f_{V'}\|_{\mathcal{Z}^m}\sqrt{\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T)}+4HT\zeta.$$

which implies

$$\|\hat{f}_{V'}-f_{V'}\|_{\mathcal{Z}^m}\le c'\sqrt{H^2(\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T))+HT\zeta}.$$

for some constant $c'>0$. □

Using the above lemma, we can obtain the following lemma similar to Lemma B.2.

Lemma D.2 (Confidence region for misspecification)**.** *Assume that our function class $\mathcal{F}$ satisfies Assumption 5.1. In Algorithm 1, for $m>M_0$, define confidence region*

$$\mathcal{F}_h^m=\left\{f\in\mathcal{F}\,\middle|\,\|f-f_h^m\|^2_{\mathcal{Z}^m}\le\beta(\mathcal{F},\delta)\right\}.$$

Then with probability at least $1-\delta/16$, for all $(h,m)\in[H]\times([M]\backslash[M_0])$,

$$r(\cdot,\cdot)+\sum_{s'\in\mathcal{S}}P(s'|\cdot,\cdot)V^m_{h+1}(s')\in\mathcal{F}_h^m,$$

given

$$\beta(\mathcal{F},\delta)\ge c'(H^2(\ln(T/\delta)+\ln\mathcal{N}(\mathcal{F},1/T)+\ln|\mathcal{W}|)+HT\zeta).$$

for some constant $c'>0$. Here, $\mathcal{W}$ is given in Proposition A.5.

Proof. The proof is almost identical to that of Lemma B.2. □

Proof of Theorem 5.2. By Lemma D.2, Lemma B.3, Lemma B.4, Lemma B.5, the proof is almost the same as the proof of Theorem 4.1. □