Achieving Constant Regret in Linear Markov Decision Processes

Weitong Zhang*
School of Data Science and Society
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
weitongz@unc.edu

Zhiyuan Fan*
EECS
Massachusetts Institute of Technology
Cambridge, MA 02139
fanzy@mit.edu

Jiafan He
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
jiafanhe19@ucla.edu

Quanquan Gu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
qgu@cs.ucla.edu

Abstract

We study the constant regret guarantees in reinforcement learning (RL). Our objective is to design an algorithm that incurs only finite regret over infinite episodes with high probability. We introduce an algorithm, `Cert-LSVI-UCB`, for misspecified linear Markov decision processes (MDPs) where both the transition kernel and the reward function can be approximated by some linear function up to misspecification level ζ. At the core of `Cert-LSVI-UCB` is an innovative certified estimator, which facilitates a fine-grained concentration analysis for multi-phase value-targeted regression, enabling us to establish an instance-dependent regret bound that is constant w.r.t. the number of episodes. Specifically, we demonstrate that for a linear MDP characterized by a minimal suboptimality gap Δ, `Cert-LSVI-UCB` has a cumulative regret of $\widetilde{\mathcal{O}}(d^3H^5/\Delta)$ with high probability, provided that the misspecification level ζ is below $\widetilde{\mathcal{O}}(\Delta/(\sqrt{d}H^2))$. Here d is the dimension of the feature space and H is the horizon. Remarkably, this regret bound is independent of the number of episodes K. To the best of our knowledge, `Cert-LSVI-UCB` is the first algorithm to achieve a constant, instance-dependent, high-probability regret bound in RL with linear function approximation without relying on prior distribution assumptions.

1 Introduction

Reinforcement learning (RL) has been a popular approach for teaching agents to make decisions based on feedback from the environment. RL has shown great success in a variety of applications, including robotics (Kober et al., 2013), gaming (Mnih et al., 2013), and autonomous driving. In most of these applications, there is a common expectation that RL agents will master tasks after making only a bounded number of mistakes, even over indefinite runs. However, theoretical support for this expectation is limited in RL literature: in the worst case, existing works such as Jin et al. (2020); Ayoub et al. (2020); Wang et al. (2019) only provided $\widetilde{\mathcal{O}}(\sqrt{K})$ regret upper bounds with K being the number of episodes; in the instance-dependent case, Simchowitz and Jamieson (2019); Yang et al. (2021); He et al. (2021a) achieved logarithmic high-probability regret upper bounds (e.g.,

*equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-4154

$\widetilde{\mathcal{O}}(\Delta^{-1}\log K)$) for both tabular MDPs and MDPs with linear function approximations, provided a minimal suboptimality gap Δ. However, these findings suggest that an agent's regret increases with the number of episodes K, contradicting to the expectation of finite mistakes in practice. To close this gap between theory and practice, there is a recent line of work proving constant regrets bound for RL and bandits, suggesting that an RL agent's regret may remain bounded even when it encounters an indefinite number of episodes. Papini et al. (2021a); Zhang et al. (2021) have provided instance-dependent constant regret bound under certain coverage assumptions on the data distribution. However, verifying these data distribution assumptions can be difficult or even infeasible. On the other hand, it is known that high-probability constant regret bound can be achieved unconditionally in multi-armed bandits (Abbasi-Yadkori et al., 2011) and contextual linear bandits if and only if the misspecification is sufficiently small with respect to the minimal sub-optimality gap (Zhang et al., 2023b). This raises a critical question:

Is it possible to design a reinforcement learning algorithm that incurs only constant regret under minimal assumptions?

To answer this question, we introduce a novel algorithm, which we refer to as `Cert-LSVI-UCB`, for reinforcement learning with linear function approximation. To encompass a broader range of real-world scenarios characterized by large state-action spaces and the need for function approximation, we consider the *misspecified linear MDP* (Jin et al., 2020) setting, where both the transition kernel and reward function can be approximated by a linear function with approximation error ζ. We show that, with our innovative design of certified estimator and novel analysis, `Cert-LSVI-UCB` achieves constant regret without relying on any prior assumption on data distributions. Our key contributions are summarized as follows:

- We introduce a parameter-free algorithm, referred to as `Cert-LSVI-UCB`, featuring a novel certified estimator for testing when the confidence set fails. This certified estimator enables `Cert-LSVI-UCB` to achieve a constant, instance-dependent, high probability regret bound of $\widetilde{\mathcal{O}}(d^3H^5/\Delta)$ for tasks with a suboptimality gap Δ, under the condition that the misspecification level ζ is bounded by $\zeta < \widetilde{\mathcal{O}}\big(\Delta/(\sqrt{d}H^2)\big)$. This bound is termed a *high probability constant regret bound*, indicating that it does not depend on the number of episodes K. We note that this constant regret bound matches the logarithmic expected regret lower bound of $\Omega(\Delta^{-1}\log K)$, suggesting that our result is valid and optimal in terms of the dependence on the suboptimality gap Δ.
- When restricted to a well-specified linear MDP (i.e., $\zeta = 0$), the constant high probability regret bound improves the previous logarithmic result $\widetilde{\mathcal{O}}(d^3H^5\Delta^{-1}\log K)$ in He et al. (2021a) by a $\log K$ factor. Our results suggest that the total suboptimality incurred by `Cert-LSVI-UCB` remains constantly bounded, regardless of the number of episodes K. In contrast to the previous constant regret bound achieved by Papini et al. (2021a), our regret bound does not require any prior assumption on the feature mapping, such as the UniSOFT assumption made in Papini et al. (2021a). To the best of our knowledge, `Cert-LSVI-UCB` is the first algorithm to achieve a *high probability constant regret bound* for MDPs without prior assumptions on data distributions. We further show that this constant regret high-probability bound does not violate the logarithmic expected regret bound by letting $\delta = 1/K$ [2].

Notation. Vectors are denoted by lower case boldface letters such as $\mathbf{x}$, and matrices by upper case boldface letters such as $\mathbf{A}$. We denote by $[k]$ the set $\{1, 2, \cdots, k\}$ for positive integers k. We use $\log x$ to denote the logarithm of x to base 2. For two non-negative sequence $\{a_n\}, \{b_n\}$, $a_n \le \mathcal{O}(b_n)$ means that there exists a positive constant C such that $a_n \le Cb_n$; $a_n \le \widetilde{\mathcal{O}}(b_n)$ means there exists a positive constant k such that $a_n \le \mathcal{O}(b_n \log^k b_n)$; $a_n \ge \Omega(b_n)$ means that there exists a positive constant C such that $a_n \ge Cb_n$; $a_n \ge \widetilde{\Omega}(b_n)$ means there exists a positive constant k such that $a_n \ge \Omega(b_n \log^{-k} b_n)$; $a_n \ge \omega(b_n)$ means that $\lim_{n\to\infty} b_n/a_n = 0$. For a vector $\mathbf{x} \in \mathbb{R}^d$ and a positive semi-definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we define $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^\top \mathbf{A}\mathbf{x}$. For any set $\mathcal{C}$, we use $|\mathcal{C}|$ to denote its cardinality. We denote the identity matrix by $\mathbf{I}$ and the empty set by $\varnothing$. The total variation distance of two distribution measures $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$ is denoted by $\|\mathbb{P}(\cdot) - \mathbb{Q}(\cdot)\|_{\mathrm{TV}}$.

[2] The detailed conversion is presented in Remark 5.2.

Algorithm	Misspecified MDP?	Result
LSVI-UCB (He et al., 2021a)	$\times$	$\widetilde{\mathcal{O}}(d^3H^5\Delta^{-1}\log(K))$
LSVI-UCB (Papini et al., 2021a)	$\times$	$\widetilde{\mathcal{O}}(d^3H^5\Delta^{-1}\log(1/\lambda))$
Cert-LSVI-UCB (ours, Theorem 5.1)	✓	$\widetilde{\mathcal{O}}(d^3H^5\Delta^{-1})$

Table 1: Instance-dependent regret bounds for different algorithms under the linear MDP setting. Here d is the dimension of the linear function $\boldsymbol{\phi}(s,a)$, H is the horizon length, Δ is the minimal suboptimality gap. All results in the table represent high probability regret bounds. The regret bound depends the number of episodes K in He et al. (2021a) and the minimum positive eigenvalue λ of features mapping in Papini et al. (2021b). **Misspecified MDP?** indicates if the algorithm can (✓) handle the misspecified linear MDP or not ($\times$).

2 Related Work

Instance-dependent regret bound in RL. Although most of the theoretical RL works focus on worst-case regret bounds, instance-dependent (a.k.a., problem-dependent, gap-dependent) regret bound is another important bound to understanding how the hardness of different instance can affect the sample complexity of the algorithm. For tabular MDPs, Jaksch et al. (2010) proved a $\widetilde{\mathcal{O}}(D^2S^2A\Delta^{-1}\log K)$ instance-dependent regret bound for average-reward MDP where D is the diameter of the MDP and Δ is the policy suboptimal gap. Simchowitz and Jamieson (2019) provided a lower bound for episodic MDP which suggests that the any algorithm will suffer from $\Omega(\Delta^{-1})$ regret bound. Yang et al. (2021) analyzed the optimistic Q-learning and proved a $\mathcal{O}(SAH^6\Delta^{-1}\log K)$ logarithmic instance-dependent regret bound. In the domain of linear function approximation, He et al. (2021a) provided instance-dependent regret bounds for both linear MDPs (i.e., $\widetilde{\mathcal{O}}(d^3H^5\Delta^{-1}\log K)$) and linear mixture MDPs (i.e., $\widetilde{\mathcal{O}}(d^2H^5\Delta^{-1}\log K)$). Furthermore, Dann et al. (2021) provided an improved analysis for this instance-dependent result with a redefined suboptimal gap. Zhang et al. (2023a) proved a similar logarithmic instance-dependent bound with He et al. (2021a) in misspecified linear MDPs, showing the relationship between misspecification level and suboptimality bound. Despite all these bounds are logarithmic depended on the number of episode K, many recent works are trying to remove this logarithmic dependence. Papini et al. (2021a) showed that under the linear MDP assumption, when the distribution of contexts $\boldsymbol{\phi}(s,a)$ satisfies the 'diversity assumption' (Hao et al., 2020) called 'UniSOFT', then LSVI-UCB algorithm may achieve an expected constant regret w.r.t. K. Zhang et al. (2021) showed a similar result on bilinear MDP (Yang and Wang, 2020), and extended this result to offline setting, indicating that the algorithm only need a finite offline dataset to learn the optimal policy. Table 1 summarizes the most relevant results mentioned above for the ease of comparison with our results.

RL with model misspecification. All of the aforementioned works consider the well-specified setting and ignore the approximation error in the MDP model. To better understand this misspecification issue, Du et al. (2019) showed that having a good representation is insufficient for efficient RL unless the approximation error (i.e., misspecification level) by the representation is small enough. In particular, Du et al. (2019) showed that an $\widetilde{\Omega}(\sqrt{H/d})$ misspecification will lead to $\Omega(2^H)$ sample complexity for RL to identify the optimal policy, even with a generative model. On the other hand, a series of work (Jin et al., 2020; Zanette et al., 2020b,a) provided $\widetilde{\mathcal{O}}(\sqrt{K}+\zeta K)$-type regret bound for RL in various settings, where ζ is the misspecification level[3] and we ignore the dependence on the dimension of the feature mapping d and the planing horizon H for simplicity. These algorithms, however, require the knowledge of misspecification level ζ, thus are not *parameter-free*. Another concern for these algorithms is that some of the algorithms (Jin et al., 2020) would possibly suffer from a *trivial asymptotic regret*, i.e., Regret$(k) > \omega(k\zeta \cdot \text{poly}(d, H, \log(1/\delta)))$, as suggested by Vial et al. (2022). This means the performance of the RL algorithm will possibly degenerate as the number of episodes k grows. To tackle these two issues, Vial et al. (2022) propose the Sup-LSVI-UCB algorithm which requires a parameter ε_{tol}. When $\varepsilon_{\text{tol}} = d/\sqrt{K}$, the proposed

[3]The misspecification level for these upper bounds is measured in the total variation distance between the ground truth transition kernel and approximated transition kernel, which is strictly stronger than the infinite-norm misspecification used in Du et al. (2019).

https://doi.org/10.52202/079017-4154

algorithm is *parameter-free* but will have a trivial *asymptotic regret bound*. When $\varepsilon_{\text{tol}} = \zeta$, the algorithm will have a non-trivial *asymptotic regret bound* but is not *parameter-free* since it requires knowledge of the misspecification level. Another series of works (He et al., 2022b; Lykouris et al., 2021; Wei et al., 2022) are working on the *corruption robust* setting. In particular, Lykouris et al. (2021); Wei et al. (2022) are using the *model-selection* technique to ensure the robustness of RL algorithms under adversarial MDPs.

3 Preliminaries

We consider episodic Markov Decision Processes, which are denoted by $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{r_h\}, \{\mathbb{P}_h\})$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the finite action space, H is the length of each episode, $r_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the reward function at stage h and $\mathbb{P}_h(\cdot|s, a)$ is the transition probability function at stage h. The policy $\pi = \{\pi_h\}_{h=1}^H$ denotes a set of policy functions $\pi_h : \mathcal{S} \mapsto \mathcal{A}$ for each stage h. For given policy π, we define the state-action value function $Q_h^\pi(s, a)$ and the state value function $V_h^\pi(s)$ as

$$Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}\left[\sum\nolimits_{h'=h+1}^H r_{h'}\big(s_{h'}, \pi_{h'}(s_{h'})\big) \,\Big|\, s_h = s, a_h = a\right], V_h^\pi(s) = Q_h^\pi\big(s, \pi_h(s)\big),$$

where $s_{h'+1} \sim \mathbb{P}_h(\cdot|s_{h'}, a_{h'})$. The optimal state-action value function Q_h^* and the optimal state value function V_h^* are defined by $Q_h^*(s,a) = \max_\pi Q_h^\pi(s,a), V_h^*(s) = \max_\pi V_h^\pi(s)$.

By definition, both the state-action value function $Q_h^\pi(s, a)$ and the state value function $V_h^\pi(s)$ are bounded by $[0, H]$ for any state s, action a and stage h. For any function $V : \mathcal{S} \mapsto \mathbb{R}$, we denote by $[\mathbb{P}_h V](s,a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} V(s')$ the expected value of V after transitioning from state s given action a at stage h and $[\mathbb{B}_h V](s,a) = r_h(s,a) + [\mathbb{P}_h V](s,a)$ where $\mathbb{B}$ is referred to as the *Bellman operator*. For each stage $h \in [H]$ and policy π, the Bellman equation, as well as the Bellman optimality equation, are presented as follows

$$Q_h^\pi(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^\pi](s,a) := [\mathbb{B}_h V_{h+1}^\pi](s,a),$$
$$Q_h^*(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^*](s,a) := [\mathbb{B}_h V_{h+1}^*](s,a).$$

We use regret to measure the performance of RL algorithms. It is defined as $\text{Regret}(K) = \sum_{k=1}^K \big(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\big)$, where π^k represents the agent's policy at episode k. This definition quantifies the cumulative difference between the expected rewards that could have been obtained by following the optimal policy and those achieved under the agent's policy across the first K episodes, measuring the total loss in performance due to suboptimal decisions.

We consider linear function approximation in this work, where we adopt the *misspecified linear MDP* assumption, which is firstly proposed in Jin et al. (2020).

Assumption 3.1 (ζ-Approximate Linear MDP, Jin et al. 2020)**.** For any $\zeta \le 1$, we say a MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{r_h\}, \{\mathbb{P}_h\})$ is a ζ*-approximate linear MDP* with a feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, if for any $h \in [H]$, there exist d *unknown* (signed) measures $\boldsymbol{\mu}_h = \big(\mu_h^{(1)}, \cdots, \mu_h^{(d)}\big)$ over $\mathcal{S}$ and an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$ such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\big\|\mathbb{P}_h(\cdot|s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot)\rangle\big\|_{\text{TV}} \le \zeta, \quad \big|r_h(s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h\rangle\big| \le \zeta,$$

w.l.o.g. we assume $\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \|\boldsymbol{\phi}(s,a)\| \le 1$ and $\forall h \in [H] : \|\boldsymbol{\mu}_h(\mathcal{S})\| \le \sqrt{d}, \|\boldsymbol{\theta}_h\| \le \sqrt{d}$.

The ζ*-approximate linear MDP* suggests that for any policy π, the state-action value function Q_h^π can be approximated by a linear function of the given feature mapping $\boldsymbol{\phi}$ up to some misspecification level, which is summarized in the following proposition.

Proposition 3.2 (Lemma C.1, Jin et al. 2020)**.** For a ζ*-approximate linear MDP*, for any policy π, there exist corresponding weights $\{\mathbf{w}_h^\pi\}_{h \in [H]}$ where $\mathbf{w}_h^\pi = \boldsymbol{\theta}_h + \int V_{h+1}^\pi(s') \mathrm{d}\boldsymbol{\mu}_h(s')$ such that for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $\big|Q_h^\pi(s,a) - \langle \boldsymbol{\phi}(s,a), \mathbf{w}_h^\pi\rangle\big| \le 2H\zeta$. We have $\|\mathbf{w}_h^\pi\|_2 \le 2H\sqrt{d}$.

Next, we introduce the definition of the suboptimal gap as follows.

Definition 3.3 (Minimal suboptimality gap)**.** For each $s \in \mathcal{S}, a \in \mathcal{A}$ and step $h \in [H]$, the suboptimality gap $\text{gap}_h(s,a)$ is defined by $\Delta_h(s,a) = V_h^*(s) - Q_h^*(s,a)$ and the minimal suboptimality gap Δ is defined by $\Delta = \min_{h,s,a} \big\{\Delta_h(s,a) : \Delta_h(s,a) \ne 0\big\}$.

Algorithm 1 `Cert-LSVI-UCB`

1: Set $V_{H+1}^k(s) = 0$ for all $(s,k) \in \mathcal{S} \times [K]$, $\mathcal{C}_{h,l}^k = \varnothing$ for all $(h,l) \in [H] \times \mathbb{N}^+$, $\lambda = 16$
2: **for** episode $k = 1, \cdots, K$ **do**
3: Set $L_k = \max\{\lceil \log_4(k/d) \rceil, 0\}$
4: **for** step $h = H, \cdots, 1$ **do**
5: **for** phase $l = 1, \cdots, L_k + 1$ **do**
6: $\mathbf{U}_{h,l}^k = \lambda \mathbf{I} + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau (\boldsymbol{\phi}_h^\tau)^\top$
7: $\mathbf{w}_{h,l}^k = (\mathbf{U}_{h,l}^k)^{-1} \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big(r_h^\tau + \widehat{V}_{h+1}^k(s_{h+1}^\tau) \big)$
8: $\widetilde{\mathbf{U}}_{h,l}^{k,-1} = \kappa_l \lceil (\mathbf{U}_{h,l}^k)^{-1} / \kappa_l \rfloor$, $\widetilde{\mathbf{w}}_{h,l}^k = \kappa_l \lceil \mathbf{w}_{h,l}^k / \kappa_l \rfloor$ where $\kappa_l = 0.01 \cdot 2^{-4l} d^{-1}$
9: **end for**
10: $\widehat{V}_h^k(s_h^\tau), \cdot, \cdot = \texttt{Cert-LinUCB}(s_h^\tau; \{\widetilde{\mathbf{w}}_{h,l}^k\}_l, \{\widetilde{\mathbf{U}}_{h,l}^{k,-1}\}_l, L_k)$ for all $\tau \in [k-1]$
11: **end for**
12: Observe $s_1^k \in \mathcal{S}$
13: **for** step $h = 1, \cdots, H$ **do**
14: $\cdot, \pi_h^k(s_h^k), l_h^k(s_h^k), f_h^k(s_h^k) = \texttt{Cert-LinUCB}(s_h^k; \{\widetilde{\mathbf{w}}_{h,l}^k\}_l, \{\widetilde{\mathbf{U}}_{h,l}^{k,-1}\}_l, L_k)$
15: $\mathcal{C}_{h,l_h^k(s_h^k)}^k = \mathcal{C}_{h,l_h^k(s_h^k)}^{k-1} \cup \{k\}$ **if** $f_h^k(s_h^k) = 1$ **else** $\mathcal{C}_{h,l_h^k(s_h^k)}^{k-1}$
16: $\mathcal{C}_{h,l}^k = \mathcal{C}_{h,l}^{k-1}$ for all $l \neq l_h^k(s_h^k)$
17: Play $\pi_h^k(s_h^k)$, set $\boldsymbol{\phi}_h^k = \boldsymbol{\phi}\big(s_h^k, \pi_h^k(s_h^k)\big)$, receive r_h^k and observe $s_{h+1}^k \in \mathcal{S}$
18: **end for**
19: **end for**

Notably, a task with a larger Δ means it is easier to distinguish the optimal action $\pi_h^*(s)$ from other actions $a \in \mathcal{A}$, while a task with lower gap Δ means it is more difficult to distinguish the optimal action.

4 Proposed Algorithms

4.1 Main algorithm: `Cert-LSVI-UCB`

We begin by introducing our main algorithm `Cert-LSVI-UCB`, which is a modification of the `Sup-LSVI-UCB` (Vial et al., 2022). As presented in Algorithm 1, for each episode k, our algorithm maintains a series of index sets $\mathcal{C}_{k,h}^l$ for each stage $h \in [H]$ and phase l. The algorithm design ensures that for any episode k, the maximum number of phases l is bounded by $L_k \le \max\{\lceil \log_4(k/d) \rceil, 0\}$. During the exploitation step, for each phase l associated with the index set $\mathcal{C}_{k-1,h}^l$, the algorithm constructs the estimator vector $\mathbf{w}_{h,l}^k$ by solving the following ridge regression problem in Line 6 and Line 7:

$$\mathbf{w}_{h,l}^k \leftarrow \mathrm{argmin}_{\mathbf{w} \in \mathbb{R}^d} \lambda \|\mathbf{w}\|_2^2 + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \big(\mathbf{w}^\top \boldsymbol{\phi}_h^\tau - r_h^\tau - \widehat{V}_{h+1}^k(s_{h+1}^\tau) \big)^2.$$

After calculating the estimator vector $\mathbf{w}_{h,l}^k$ in Line 8, the algorithm quantilizes $\mathbf{w}_{h,l}^k$ and $(\mathbf{U}_{h,l}^k)^{-1}$ to the precision of κ_l. Similar to `Sup-LSVI-UCB` (Vial et al., 2022), we note $\widetilde{\mathbf{U}}_{h,l}^{k,-1}$ is the quantized version of inverse covariance matrix $(\mathbf{U}_{h,l}^k)^{-1}$ rather than the inverse of quantized covariance matrix $(\widetilde{\mathbf{U}}_{h,l}^k)^{-1}$. The main difference between our implementation and that in Vial et al. (2022) is that we use a layer-dependent quantification precision κ_l instead of the global quantification precision $\kappa = 2^{-4L}/d$, which enables our algorithm get rid of the dependence on $\mathcal{O}(\log K)$ in the maximum number of phases L_k.

After obtaining $\widetilde{\mathbf{w}}_{h,l}^k$ and $\widetilde{\mathbf{U}}_{h,l}^{k,-1}$, a subroutine, `Cert-LinUCB`, is called to calculate an optimistic value function $\widehat{V}_h^k(s_h^\tau)$ for all historical states s_h^τ in Line 10. Then the algorithm transits to stage $h-1$ and iteratively computes $\widetilde{\mathbf{w}}_{h,l}^k$ and $\widetilde{\mathbf{U}}_{h,l}^{k,-1}$ for all phase l and stage $h \in [H]$.

In the exploration step, the algorithm starts to do planning from the initial state s_1^k. For each observed state s_h^k, the same subroutine, `Cert-LinUCB`, will be called in Line 14 for the policy $\pi_h^k(s_h^k)$, the

https://doi.org/10.52202/079017-4154

Algorithm 2 `Cert-LinUCB` : $\big(s; \{\widetilde{\mathbf{w}}_{h,l}^k\}_l, \{\widetilde{\mathbf{U}}_{h,l}^{k,-1}\}_l, L\big) \mapsto \big(\widehat{V}_h^k(s), \pi_h^k(s), l_h^k(s), f_h^k(s)\big)$

1: **input:** $s \in \mathcal{S}, \forall l : \widetilde{\mathbf{w}}_{h,l}^k \in \mathbb{R}^d, \widetilde{\mathbf{U}}_{h,l}^{k,-1} \in \mathbb{R}^{d\times d}, L \in \mathbb{N}^+$
2: **output:** $\widehat{V}_h^k(s) \in \mathbb{R}, \pi_h^k(s) \in \mathcal{A}, l_h^k(s) \in \mathbb{N}^+, f_h^k(s) \in \{0,1\}$
3: $\mathcal{A}_{h,1}^k(s) = \mathcal{A}, \check{V}_{h,0}^k(s) = 0, \widehat{V}_{h,0}^k(s) = H$
4: **for** phase $l = 1, \cdots, L+1$ **do**
5: Set $Q_{h,l}^k(s,a) = \big\langle \boldsymbol{\phi}(s,a), \widetilde{\mathbf{w}}_{h,l}^k \big\rangle$
6: Set $\pi_{h,l}^k(s) = \mathrm{argmax}_{a \in \mathcal{A}_{h,l}^k} Q_{h,l}^k(s,a), V_{h,l}^k(s) = Q_{h,l}^k\big(s, \pi_{h,l}^k(s)\big)$
7: **if** $l > L$ **then**
8: **return** $\big(\widehat{V}_h^k(s), \pi_h^k(s), l_h^k(s), f_h^k(s)\big) = \big(\widehat{V}_{h,l-1}^k(s), \pi_{h,l-1}^k(s), l, 1\big)$
9: **else if** $\gamma_l \cdot \max_{a \in \mathcal{A}_{h,l}^k(s)} \|\boldsymbol{\phi}(s,a)\|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}} \geq 2^{-l}$ **then**
10: **return** $\big(\widehat{V}_h^k(s), \pi_h^k(s), l_h^k(s), f_h^k(s)\big) = \big(\widehat{V}_{h,l-1}^k(s), \mathrm{argmax}_{a \in \mathcal{A}_{h,l}^k(s)} \|\boldsymbol{\phi}(s,a)\|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}}, l, 1\big)$
11: **else if** $\max\big\{V_{h,l}^k(s) - 3\cdot 2^{-l}, \check{V}_{h,l-1}^k(s)\big\} > \min\big\{V_{h,l}^k(s) + 3\cdot 2^{-l}, \widehat{V}_{h,l-1}^k(s)\big\}$ **then**
12: **return** $\big(\widehat{V}_h^k(s), \pi_h^k(s), l_h^k(s), f_h^k(s)\big) = \big(\widehat{V}_{h,l-1}^k(s), \pi_{h,l-1}^k(s), l, 0\big)$
13: **else**
14: $\widehat{V}_{h,l}^k(s) = \min\big\{V_{h,l}^k(s) + 3\cdot 2^{-l}, \widehat{V}_{h,l-1}^k(s)\big\}$
15: $\check{V}_{h,l}^k(s) = \max\big\{V_{h,l}^k(s) - 3\cdot 2^{-l}, \check{V}_{h,l-1}^k(s)\big\}$
16: $\mathcal{A}_{h,l+1}^k(s) = \Big\{a \in \mathcal{A}_{h,l}^k(s) : Q_{h,l}^k(s,a) \geq V_{h,l}^k(s) - 4\cdot 2^{-l}\Big\}$
17: **end if**
18: **end for**

corresponding phase $l_h^k(s_h^k)$, and a flag $f_h^k(s_h^k)$. If the flag $f_h^k(s_h^k) = 1$, the algorithm adds the index k to the index set $\mathcal{C}_{h,l_h^k(s_h^k)}^k$ in Line 15. Otherwise, the algorithm skips the current index k and all index sets remain unchanged. Finally, the algorithm executes policy $\pi_h^k(s_h^k)$, receives reward r_h^k and observes the next state s_{h+1}^k in Line 17.

4.2 Subroutine: `Cert-LinUCB`

Next we introduce subroutine `Cert-LinUCB`, improved from `Sup-Lin-UCB-Var` (Vial et al., 2022) that computes the optimistic value function $\widehat{V}_h^k$. The algorithm is described as follows. Starting from phase $l = 1$, the algorithm first calculates the estimated state-action function $Q_{h,l}^k(s,a)$ as a linear function over the quantified parameter $\widetilde{\mathbf{w}}_{h,l}^k$ and feature mapping $\boldsymbol{\phi}(s,a)$, following Proposition 3.2. After calculating the estimated state-action value function $Q_{h,l}^k(s)$, the algorithm computes the greedy policy $\pi_{h,l}^k(s)$ and its corresponding value function $V_{h,l}^k(s)$.

Similar to `Sup-Lin-UCB-Var` (Vial et al., 2022), our algorithm has several conditions starting from Line 7 to determine whether to stop at the current phase or to eliminate the actions and proceed to the next phase $l+1$, which are listed in the following conditions.

- **Condition 1**: In Line 7, if the current phase l is greater than the maximum phase L, we directly stop at that phase and take the greedy policy on previous phase $\pi_h^k(s) = \pi_{h,l-1}^k(s)$.
- **Condition 2**: In Line 9, if there exists an action whose uncertainty $\|\boldsymbol{\phi}(s,a)\|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}}$ is greater than the threshold $2^{-l}\gamma_l^{-1}$, our algorithm will perform exploration by selecting that action.
- **Condition 3**: In Line 11, we compare the value of the pessimistic value function $\check{V}_{h,l}^k(s)$ and the optimistic value function $\widehat{V}_{h,l}^k(s)$ which will be assigned in Line 14 and Line 15, if the pessimistic estimation will be greater than the optimistic estimation, we will stop at that phase and take the greedy policy on previous phase $\pi_h^k(s) = \pi_{h,l-1}^k(s)$. Only in this case, the Algorithm 2 outputs flag $f_h^k(s) = 0$, which means this observation will not be used in Line 15 in Algorithm 1.
- **Condition 4**: In the default case in Line 16, the algorithm proceeds to the next phase after eliminating actions.

Notably, in **Condition 4**, since the expected estimation precision in the l-th phase is about $\widetilde{\mathcal{O}}(2^{-l})$, our algorithm can eliminate the actions whose state-action value is significantly less than others, i.e., less than $\widetilde{\mathcal{O}}(2^{-l})$, while retaining the remaining actions for the next phase.

Specially, our algorithm differs from that in Vial et al. (2022) in terms of **Condition 3** to certify the performance of the estimation. In particular, a well-behaved estimation should always guarantee that the optimistic estimation is greater than the pessimistic estimation. According to Line 14 and Line 15, this is equivalent to the confidence region for l-th phase has intersection of the previous confidence region $[\check{V}_{h,l-1}^k(s), \widehat{V}_{h,l-1}^k(s)]$. Otherwise, we hypothesis the estimation on l-th phase is corrupted by either misspecification or bad concentration event, thus will stop the algorithm. We will revisit the detail of this design later.

It's important to highlight that our algorithms provide unique approaches when compared with previous works. In particular, He et al. (2021b) does not eliminate actions and combines estimations from all layers by considering the minimum estimated optimistic value function. This characteristic prevents their algorithm from achieving a uniform PAC guarantee in the presence of misspecification. For a more detailed comparison with He et al. (2021b), please refer to Appendix B.1. Additionally, Lykouris et al. (2021); Wei et al. (2022) focus on a model-selection regime where a set of base learners are employed in the algorithms, whereas we adopt a multi-phase approach similar with SupLinUCB rather than conducting model selection over base learners.

5 Constant Regret Guarantee

Theorem 5.1. Under Assumption 3.1, let $\gamma_l = 5(l + 20 + \lceil \log(ld) \rceil) dH \sqrt{\log(16ldH/\delta)}$ for some fixed $0 < \delta < 1/4$. With probability at least $1 - 4\delta$, if misspecification level ζ is below $\widetilde{\mathcal{O}}\big(\Delta/(\sqrt{d}H^2)\big)$ where Δ is the minimal suboptimality gap, then for all $K \in \mathbb{N}^+$, the regret of Algorithm 1 is upper bounded by

$$\text{Regret}(K) \le \widetilde{\mathcal{O}}\big(d^3 H^5 \Delta^{-1} \log(1/\delta)\big).$$

This regret bound is constant w.r.t. the episode K.

Theorem 5.1 demonstrates a constant regret bound with respect to number of episodes K. Compared with Papini et al. (2021a), our regret bound does not require any prior assumption on the feature mapping $\boldsymbol{\phi}$, such as the *UniSOFT* assumption made in Papini et al. (2021a). In addition, compared with the previous logarithmic regret bound He et al. (2021a) in the well-specified setting, our constant regret bound removes the $\log K$ factor, indicating the cumulative regret no longer grows w.r.t. the number of episode K, with high probability.

Remark 5.2. As discussed in Zhang et al. (2023b) in the misspecified linear bandits, Our *high probability* constant regret bound does not violate the lower bound proved in Papini et al. (2021a), which says that certain diversity condition on the contexts is necessary to achieve an *expected* constant regret bound. When extending this high probability constant regret bound to the expected regret bound, we have

$$\mathbb{E}[\text{Regret}(K)] \le \widetilde{\mathcal{O}}\big(d^3 H^5 \Delta^{-1} \log(1/\delta)\big) \cdot (1 - \delta) + \delta K,$$

which depends on the number of episodes k. To obtain a sub-linear expected regret, we can choose $\delta = 1/K$, which yields a logarithmic expected regret $\widetilde{\mathcal{O}}(d^3 H^5 \Delta^{-1} \log K)$ and does not violate the lower bound in Papini et al. (2021a).

Remark 5.3. Du et al. (2019) provide a lower bound showing the interplay between the misspecification level ζ and suboptimality gap Δ in a weaker setting, which we discuss in detail in Appendix B.2. Along with the result from Du et al. (2019), our results suggests that ignoring the dependence on H, $\zeta = \widetilde{\mathcal{O}}(\Delta/\sqrt{d})$ plays an important seperation for if a misspeficied model can be efficiently learned. This result is also aligned with the positive result and negative result for linear bandits (Lattimore et al., 2020; Zhang et al., 2023b).

6 Technical Challenges and Highlight of Proof Techniques

In this section, we highlight several major challenges in obtaining the constant regret under misspecified linear MDP assumption and how our method, especially the certified estimator, tackles these challenges.

https://doi.org/10.52202/079017-4154

6.1 Challenge 1. Achieving layer-wise local estimation error.

In the analysis of the value function under misspecified linear MDPs, we follow the multi-phase estimation strategy (Vial et al., 2022) to eliminate suboptimal actions and improve the robustness of the next phase estimation. Similar approaches have been observed in Zhang et al. (2023b); Chu et al. (2011) within the framework of (misspecified) linear bandits. However, unlike linear bandits, when constructing the empirical value function $\widehat{V}_h$ for stage h in linear MDPs, Jin et al. (2020) requires a covering statement on value functions to ensure the convergence of the regression, which is written by: (see Lemma D.4 in Jin et al. (2020) for details)

$$\Big\| \sum_{\tau \in \mathcal{C}} \boldsymbol{\phi}_h^\tau \big[\widehat{V}_{h+1}^k(s^\tau) - \mathbb{E}[\widehat{V}_{h+1}^k(s^\tau)]\big] \Big\|_{\mathbf{U}_h^{-1}} \le \widetilde{\mathcal{O}}_H\Big(\sqrt{d\log(|\mathcal{C}|) + \log(|\mathcal{V}_{h+1}^k|/\delta)} + \sqrt{d}\kappa\Big), \tag{6.1}$$

where we employ notation $\widetilde{\mathcal{O}}_H$ to obscure the dependence on H to simplify the presentation. We use the notation $\mathcal{V}_{h+1}^k$ to denote as an κ-covering, (or quantification in Takemura et al. (2021); Vial et al. (2022)) for the value functions $\widehat{V}_{h+1}^k$. However, in the multi-phase algorithm, the empirical value function $\widehat{V}_{h+1}^k$ from the subsequent stage $h+1$, which is formulated using all pairs of parameters $\{\mathbf{w}_{h,\ell}^k, \mathbf{U}_{h,\ell}^k\}_\ell$ in L phases. Consequently, the covering number $\log|\mathcal{V}_{h+1}^k|$ is directly proportional to the number of phases $L = \mathcal{O}(\log K)$.

Therefore, when analyzing any single phase l, prior analysis cannot eliminate the $\log K$ term from (6.1) to achieve a *local* estimation error independent that is independent of the logarithmic number of global episodes $\log K$. Furthermore, due to the algorithm design of previous methods (Vial et al., 2022), additional $\log K$ terms may be introduced by global quantification (i.e., $\varepsilon_{tol} = d/\sqrt{K}$).

Our approach: `Cert-LinUCB`. To tackle this challenge, we introduce the certified estimator into Algorithm 2 and use a 'local quantification' to ensure the quantification error of each phase l depend on the local phase $\widetilde{\mathcal{O}}(l)$ instead of the global parameter $\log K$. The certified estimator works as follows: Considering the concentration term we need to control for each phase l:

$$\Big\| \sum_{\tau \in \mathcal{C}_{h,l}^k} \boldsymbol{\phi}_h^\tau \big[\widehat{V}_{h+1}^k(s^\tau) - \mathbb{E}[\widehat{V}_{h+1}^k(s^\tau)]\big] \Big\|_{(\mathbf{U}_{h,l}^k)^{-1}}, \tag{6.2}$$

as discussed in **Challenge 1**, the function class $\widehat{\mathcal{V}}_{h+1}^k \ni \widehat{V}_{h+1}^k$ involves $L = \mathcal{O}(\log K)$ parameters, leading to a $\log K$ dependence in the results when using traditional routines. The idea of certified estimator is to get rid of this by not directly controlling $\log|\mathcal{V}_{h+1}^k|$. Instead, certified estimator establishes a covering statement for the value function class $\mathcal{V}_{h+1,l_+}^k \ni \widehat{V}_{h+1,l_+}^k$, where $\widehat{V}_{h+1,l_+}^k$ is the value function that only incorporates the first l_+ phases of parameters $\{\mathbf{w}_{h,\ell}^k, \mathbf{U}_{h,\ell}^k\}_\ell$. Under this framework, the covering statement becomes:

Lemma 6.1 (Lemma C.4, informal)**.** Let $\widehat{V}_{h+1,l_+}^k$ be the output of Algorithm 2 terminated at phase $l_+ \in \mathbb{N}^+$, then with probability is at least $1-2\delta$,

$$\Big\| \sum_{\tau \in \mathcal{C}_{h,l}^k} \boldsymbol{\phi}_h^\tau \big[\widehat{V}_{h+1,l_+}^k(s^\tau) - \mathbb{E}[\widehat{V}_{h+1,l_+}^k(s^\tau)]\big] \Big\|_{(\mathbf{U}_{h,l}^k)^{-1}} \le \gamma_{l,l_+} = 5l_+dH\sqrt{\log(16ldH/\delta)}.$$

Lemma 6.1 suggests a concentration inequality at any phase l_+, and the following lemma suggests that this procedure will only introduce an $\widetilde{\mathcal{O}}(2^{-l_+})$ error, under some faithful extension of the $\widehat{V}_{h,l_+}^k(s)$:

Lemma 6.2 (Lemma C.2, informal)**.** For any $l_+ \in \mathbb{N}^+$, $|\widehat{V}_h^k(s) - \widehat{V}_{h,l_+}^k(s)| \le 6 \cdot 2^{-l_+}$.

Therefore, if a large enough l_+ can be reached in Algorithm 2, combining Lemma 6.1 and Lemma 6.2 allow us to bound (6.2) without introducing $\log K$ factors. The next lemma shows that the Line 11 will only never be triggered in shallow layer l.

Lemma 6.3 (Lemma C.8, informal)**.** With probability at least $1-2\delta$, for any $(k,h) \in [K] \times [H]$, Line 11 in Algorithm 2 can only be triggered on phase $l \ge \widetilde{\Omega}\big(\log(1/\zeta)\big)$.

Lemma 6.3 delivers a clear message: In the well-specified setting, Line 11 will never be triggered ($l \geq \infty$). When the misspecification level is large, then Line 11 will be more likely triggered, indicating it's harder for the algorithm to proceed to deeper layer. The contribution of the certified estimator yields the following important lemma regarding the 'local estimation error':

Lemma 6.4 (Lemma C.12, Informal)**.** With high probability, for any $\varepsilon > \widetilde{\Omega}(\sqrt{d}H^2\zeta)$ and $h \in [H]$, `Cert-LSVI-UCB` ensures $\sum_{k=1}^{\infty} \mathbb{1}\left[V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) \geq \varepsilon\right] \leq \widetilde{\mathcal{O}}\big(d^3H^4\varepsilon^{-2}\big)$.

Remark 6.5. He et al. (2021b) achieved a similar $\mathcal{O}\big(d^3H^5\varepsilon^{-2}\big)$ *uniform-PAC* bound for (well-specified) linear MDP. Comparing with Lemma 6.4 with $\zeta = 0$, one can find that our result is better than He et al. (2021b). In addition, Lemma 6.4 ensures this *uniform-PAC* result under all stage $h \in [H]$ while He et al. (2021b) only ensure the $h = 1$. This improvement is achieved by a more efficient data selection strategy which we will discuss in detail in Appendix B.1.

6.2 Challenge 2. Achieving constant regret from local estimation error

In misspecified linear bandits, Zhang et al. (2023b) concludes their proof by controlling $\sum_{k=1}^{\infty} \mathbb{1}[V_1^*(s_1^k) - V_1^{\pi}(s_1^k) \geq \Delta]$[4]. Although it is trivial showing that rounds with instantaneous regret $V_1^*(s_1^k) - V_1^{\pi}(s_1^k) < \Delta$ is optimal in bandits (i.e., $V_1^*(s_1^k) = V_1^{\pi}(s_1^k)$), previous works fail to reach a similar result for RL settings. This difficulty arises from the randomness inherent in MDPs: Consider a policy π that is optimal at the initial stage $h = 1$. After the initial state and action, the MDP may transition to a state s_2' with a small probability p where the policy π is no longer optimal, or to another state s_2 where π remains optimal until the end. In this context, the gap between $V_1^*(s_1)$ and $V_1^{\pi}(s_1)$ can be arbitrarily small, given a sufficiently small $p > 0$:

$$V_1^*(s_1) - V_1^{\pi}(s_1) = p\big(V_2^*(s_2') - V_2^{\pi}(s_2')\big) + (1-p)\big(V_2^*(s_2) - V_2^{\pi}(s_2)\big) = p\big(V_2^*(s_2') - V_2^{\pi}(s_2')\big).$$

Therefore, one cannot easily draw a constant regret conclusion simply by controlling $\sum_{k=1}^{\infty} \mathbb{1}[V_1^*(s_1^k) - V_1^{\pi}(s_1^k) \geq \Delta]$ since the gap between $V_1^*(s_1^k) - V_1^{\pi}(s_1^k)$ needs to be further fine-grained controlled. In short, the existence of Δ describing the minimal gap between $V^*(s) - Q^*(s,a)$ cannot be easily applied to controlling regret $V^*(s) - V^{\pi}(s)$.

Our approach: A fine-grained concentration analysis We address this challenge by providing a fine-grained concentration analysis in connecting the gap with the regret. Notice that the regret $V_h^*(s_h) - V_h^{\pi^k}(s_h)$ in episode k is the expectation of cumulative suboptimality gap $\mathbb{E}[\sum_{h=1}^{H} \Delta_h^k]$ taking over trajectory $\{s_h^k\}_{h=1}^{H}$. In addition, the variance of the random variable can be self-bounded according to

$$\mathrm{Var}\left[\sum_{h=1}^{H}\Delta_h^k\right] \leq \mathbb{E}\left[\left(\sum_{h=1}^{H}\Delta_h^k\right)^2\right] \leq H^2\,\mathbb{E}\left[\sum_{h=1}^{H}\Delta_h^k\right] = H^2\big(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\big).$$

Denote η_k be the difference between $V_h^*(s_h) - V_h^{\pi^k}(s_h)$ and the actual $\sum_{h=1}^{H}\Delta_h^k$. Freedman inequality (Lemma H.5) implies that $\sum_{t=1}^{T}\eta^t \geq aC$ and $\sum_{t=1}^{T}\mathrm{Var}[\eta^t] \leq vC$ happens at the same with a small probability for certain constant a and v. Using a fine-grained union bound statement over C, we can reach the following statement indicates the cumulative regret can be upper bounded using the cumulative suboptimality gap:

Lemma 6.6 (Lemma C.14, Informal)**.** The following statement holds with high probability:

$$\sum_{k=1}^{K}\big(V_h^*(s_h) - V_h^{\pi^k}(s_h)\big) \leq \widetilde{\mathcal{O}}\Big(\sum_{k=1}^{K}\sum_{h=1}^{H}\Delta_h^k + H^2\Big).$$

Comparing with Lemma 6.1 in He et al. (2021a), Lemma 6.6 eliminates the $\log K$ dependence, which is achieved by the aforementioned fine-grained union bound. As a result, together with Lemma 6.4, we reach the desired statement that `Cert-LSVI-UCB` achieves constant regret bound when the misspecification is sufficiently small against the minimal suboptimality gap.

7 Conclusions and Limitations

In this work, we proposed a new algorithm, called certified estimator, for reinforcement learning with a misspecified linear function approximation. Our algorithm is parameter-free and does not

[4]We employ the RL notations and set $h = 1$ for the ease of comparison.

 https://doi.org/10.52202/079017-4154

require prior knowledge of misspecification level ζ or the suboptimality Δ. Our algorithm is based on a novel certified estimator and provides the first constant regret guarantee for misspecified linear MDPs and (well-specified) linear MDPs.

Limitations. Despite these advancements, several aspects of our algorithm and analysis warrant further investigation. One significant open question is whether the dependency on the planning horizon and dimension d, H can achieve optimal instance-dependent regret bounds. For the gap-independent regret bounds, the regret lower bound is $\Omega(d\sqrt{H^3K})$ as shown by Zhou et al. (2021a), and this benchmark has recently been met by works such as He et al. (2022a); Agarwal et al. (2022). Additionally, our analysis assumes uniform misspecification across all actions. Investigating other types of misspecifications could lead to more sophisticated results, enhancing the algorithm's robustness and applicability to diverse real-world scenarios. This exploration remains an important direction for future research.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their helpful comments. This work was done while WZ was a PhD student at UCLA. WZ is partially supported by UCLA dissertation year fellowship and the research fund from UCLA-Amazon Science Hub. JH and QG are partially supported by the research fund from UCLA-Amazon Science Hub. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24** 2312–2320.

AGARWAL, A., JIN, Y. and ZHANG, T. (2022). Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. *arXiv preprint arXiv:2212.06069* .

AYOUB, A., JIA, Z., SZEPESVARI, C., WANG, M. and YANG, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*. PMLR.

CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, learning, and games*. Cambridge university press.

CHU, W., LI, L., REYZIN, L. and SCHAPIRE, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings.

DANN, C., MARINOV, T. V., MOHRI, M. and ZIMMERT, J. (2021). Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems* **34**.

DU, S. S., KAKADE, S. M., WANG, R. and YANG, L. F. (2019). Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*.

HAO, B., LATTIMORE, T. and SZEPESVARI, C. (2020). Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

HE, J., ZHAO, H., ZHOU, D. and GU, Q. (2022a). Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132* .

HE, J., ZHOU, D. and GU, Q. (2021a). Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*. PMLR.

HE, J., ZHOU, D. and GU, Q. (2021b). Uniform-PAC bounds for reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems*.

HE, J., ZHOU, D., ZHANG, T. and GU, Q. (2022b). Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Advances in Neural Information Processing Systems*.

HOEFFDING, W. (1963). Probability inequalities for sum of bounded random variables.

ISHFAQ, H., CUI, Q., NGUYEN, V., AYOUB, A., YANG, Z., WANG, Z., PRECUP, D. and YANG, L. (2021). Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*. PMLR.

JAKSCH, T., ORTNER, R. and AUER, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* **11**.

JIA, Z., YANG, L., SZEPESVARI, C. and WANG, M. (2020). Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*. PMLR.

JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR.

KOBER, J., BAGNELL, J. A. and PETERS, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* **32** 1238–1274.

LATTIMORE, T., SZEPESVARI, C. and WEISZ, G. (2020). Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*. PMLR.

LYKOURIS, T., SIMCHOWITZ, M., SLIVKINS, A. and SUN, W. (2021). Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*. PMLR.

MNIH, V., KAVUKCUOGLU, K., SILVER, D., GRAVES, A., ANTONOGLOU, I., WIERSTRA, D. and RIEDMILLER, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* .

MODI, A., JIANG, N., TEWARI, A. and SINGH, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*.

PAPINI, M., TIRINZONI, A., PACCHIANO, A., RESTELLI, M., LAZARIC, A. and PIROTTA, M. (2021a). Reinforcement learning in linear mdps: Constant regret and representation selection. *Advances in Neural Information Processing Systems* **34** 16371–16383.

PAPINI, M., TIRINZONI, A., RESTELLI, M., LAZARIC, A. and PIROTTA, M. (2021b). Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*. PMLR.

SIMCHOWITZ, M. and JAMIESON, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems* **32** 1153–1162.

TAKEMURA, K., ITO, S., HATANO, D., SUMITA, H., FUKUNAGA, T., KAKIMURA, N. and KAWARABAYASHI, K.-I. (2021). A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

VIAL, D., PARULEKAR, A., SHAKKOTTAI, S. and SRIKANT, R. (2022). Improved algorithms for misspecified linear markov decision processes. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

WANG, Y., WANG, R., DU, S. S. and KRISHNAMURTHY, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*.

WEI, C.-Y., DANN, C. and ZIMMERT, J. (2022). A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*. PMLR.

YANG, K., YANG, L. and DU, S. (2021). Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

 https://doi.org/10.52202/079017-4154

YANG, L. and WANG, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*. PMLR.

ZANETTE, A., BRANDFONBRENER, D., BRUNSKILL, E., PIROTTA, M. and LAZARIC, A. (2020a). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*.

ZANETTE, A., LAZARIC, A., KOCHENDERFER, M. and BRUNSKILL, E. (2020b). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*. PMLR.

ZHANG, W., HE, J., FAN, Z. and GU, Q. (2023a). On the interplay between misspecification and sub-optimality gap: From linear contextual bandits to linear MDPs.

ZHANG, W., HE, J., FAN, Z. and GU, Q. (2023b). On the interplay between misspecification and sub-optimality gap in linear contextual bandits. *arXiv preprint arXiv:2303.09390* .

ZHANG, W., HE, J., ZHOU, D., ZHANG, A. and GU, Q. (2021). Provably efficient representation learning in low-rank markov decision processes. *arXiv preprint arXiv:2106.11935* .

ZHOU, D., GU, Q. and SZEPESVARI, C. (2021a). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

ZHOU, D., HE, J. and GU, Q. (2021b). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR.

A Additional Related Work

RL with linear function approximation. Recent years have witnessed a line of work focusing on RL with linear function approximation to tackle RL tasks in large state space. A widely studied MDP model is linear MDP (Jin et al., 2020), where both the transition kernel and the reward function are linear functions of a given feature mapping of the state-action pairs $\boldsymbol{\phi}(s, a)$. Several works have developed RL algorithms with polynomial sample complexity or sublinear regret bound in this setting. For example, LSVI-UCB (Jin et al., 2020) has an $\widetilde{\mathcal{O}}(\sqrt{d^3H^4K})$ regret bound, randomized LSVI (Zanette et al., 2020a) has an $\widetilde{\mathcal{O}}(\sqrt{d^4H^5K})$ regret bound and Ishfaq et al. (2021) achieved an $\widetilde{\mathcal{O}}(\sqrt{d^3H^4K})$. He et al. (2022a) then improves this regret bound to a nearly minimax-optimal result $\widetilde{\mathcal{O}}(d\sqrt{H^3K})$ while Agarwal et al. (2022) provides a general function approximation extension given the above result. Linear mixture/kernel MDPs (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b) have also emerged as another model that enables model-based RL with linear function approximation. In this setting, the transition kernel is a linear function of a feature mapping on the triplet of state, action, and next state $\boldsymbol{\phi}(s, a, s')$. Nearly minimax optimal regrets can be achieved for both finite-horizon episodic MDPs (Ayoub et al., 2020; Zhou et al., 2021a) and infinite-horizon discounted MDPs (Zhou et al., 2021b) under this assumption.

B Additional Discussions on Algorithm Design and Result

B.1 Comparison with He et al. (2021b)

It is worth comparing our algorithm with He et al. (2021b), which also provides a uniform PAC bound for linear MDPs. Both our algorithm and theirs utilize a multi-phase structure that maintains multiple regression-based value function estimators at different phases. Despite this similarity, there are several major differences between our algorithm and that in He et al. (2021b), which are highlighted as follows:

(1) In Line 7 of Algorithm 1, when calculating the regression-based estimator, for different phase l, we use the same regression target $\widehat{V}_{h+1}^k$, while their algorithm uses different $V_{h+1,l}^k$ for different phase l.
(2) When aggregating the regression estimators over all different L_k phases, we follow the arm elimination method as in Chu et al. (2011), while He et al. (2021b) simply take the point-wise minimum of all estimated state-action functions, i.e., $Q(s, a) = \min_{l\in[L_k]} Q_{k,h}^l(s, a)$.

(3) When calculating the phase $l_h^k(s_h^k)$ for a trajectory $s_1^k, s_2^k, \cdots, s_H^k$, He et al. (2021b) require that the phase $l_h^k(s_h^k)$ to be monotonically decreasing with respect to the stage h, i.e., $l_h^k(s_h^k) \le l_{h-1}^k(s_{h-1}^k)$ (see line 19 in Algorithm 2 in He et al. (2021b)). Such a requirement will lead to a poor estimation for later stages and thus increase the sample complexity. In contrast, we do not have this requirement or any other requirements related to $l_h^k(s_h^k)$ and $l_{h-1}^k(s_{h-1}^k)$.

As a result, by (3), He et al. (2021b) have to sacrifice some sample complexity to make their algorithm work for different target value functions $V_{h+1,l}^k$. As a comparison, since we use the same regression target for different phase l, we do not have to make such a sacrifice in (3). Moreover, by (2), He et al. (2021b) cannot deal with linear MDPs with misspecification, while our algorithm can handle misspecification as in Vial et al. (2022).

B.2 Discussion on Lower Bounds of Sample Complexity

We present a lower bound from Du et al. (2019) to better illustrate the interplay between the misspecification level ζ and the suboptimality gap Δ.

Assumption B.1 (Assumption 4.3, Du et al. 2019, ζ-Approximate Linear MDP)**.** There exists $\zeta > 0$, $\boldsymbol{\theta}_h \in \mathbb{R}^d$ and $\boldsymbol{\mu}_h : \mathcal{S} \mapsto \mathbb{R}^d$ for each stage $h \in [H]$ such that for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have $\left|\mathbb{P}_h(s'|s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(s')\rangle\right| \le \zeta$ and $\left|r(s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h\rangle\right| \le \zeta$.

Theorem B.2 (Theorem 4.2, Du et al. 2019)**.** There exists a family of hard-to-learn linear MDPs with action space $|\mathcal{A}| = 2$ and a feature mapping $\boldsymbol{\phi}(s,a)$ satisfying Assumption B.1, such that for any algorithm that returns a $1/2$-optimal policy with probability 0.9 needs to sample at least $\Omega(\min\{|\mathcal{S}|, 2^H, \exp(d\zeta^2/16)\})$ episodes.

Remark B.3. As claimed in Du et al. (2019), Theorem B.2 suggests that when misspecification in the ℓ_∞ norm satisfies $\zeta = \Omega(\Delta\sqrt{H/d})$, the agent needs an exponential number of episodes to find a near-optimal policy, where $\Delta = 1/2$ in their setting. It is worth noting that Assumption B.1 is a ℓ_∞ approximation for the transition matrix. Such a ℓ_∞ guarantee ($\|\cdot\|_\infty \le \zeta$) is weaker than the ℓ_1 guarantee ($\|\cdot\|_1 \le \zeta$) provided in Assumption 3.1. So it's natural to observe a positive result when making a stronger assumption and a negative result when making a weaker assumption. In addition, despite of this difference, one could find that $\zeta \sim \Delta/\sqrt{d}$ plays a vital role in determining if the task can be efficiently learned. Similar positive and negative results are also provided in Lattimore et al. (2020); Zhang et al. (2023b) in the linear contextual bandit setting (a special case of linear MDP with $H = 1$).

C Constant Regret Guarantees for `Cert-LSVI-UCB`

In this section, we present the proof of Theorem 5.1. To begin with, we recap the notations used in the algorithm and introduce several shorthand notations that would be employed for the simplicity of latter proof. The notation table is presented in Table 2.Any proofs not included in this section are deferred to Appendix D.

C.1 Quantized State Value Function set $\mathcal{V}_{h,l}^k$.

To begin our proof, we first extend the definition of $\widehat{V}_{h,l}^k$ to arbitrary l and give a formal definition of the state value function class $\mathcal{V}_{h,l}^k$ as we skip the detail of this definition in Section 6.

Definition C.1. We extend the definition of state value function $\widehat{V}_{h,l}^k$ to any tuple $(k,h,l) \in [K] \times [H] \times \mathbb{N}^+$ by

$$\widehat{V}_{h,l}^k, \cdot, \cdot, \cdot = \texttt{Cert-LinUCB}\big(s; \{\widetilde{\mathbf{w}}_{h,\ell}^k\}_{\ell=1}^l, \{\widetilde{\mathbf{U}}_{h,\ell}^{k,-1}\}_{\ell=1}^l, l\big)$$

We also define the state value function family $\mathcal{V}_{h,l}^k$ be the set of all possible $\widehat{V}_{h,l}^k$.

$$\mathcal{V}_{h,l}^k = \left\{\widehat{V}_{h,l}^k \;\middle|\; \widehat{V}_{h,l}^k, \cdot, \cdot, \cdot = \texttt{Cert-LinUCB}\big(s; \{\widetilde{\mathbf{w}}_{\cdot,\ell}^{\cdot}\}_{\ell=1}^l, \{\widetilde{\mathbf{U}}_{\cdot,\ell}^{\cdot,-1}\}_{\ell=1}^l, l\big)\right\}$$

where $\{\widetilde{\mathbf{w}}_{\cdot,\ell}^{\cdot}\}_{\ell=1}^l$ and $\{\widetilde{\mathbf{U}}_{\cdot,\ell}^{\cdot,-1}\}_{\ell=1}^l$ are referring to *any* possible parameters generated by Line 8 in Algorithm 1.

https://doi.org/10.52202/079017-4154

Notation	Meaning
ζ	Misspecification level of feature map ϕ_h. (see Definiton 3.1)
Δ	Minimal suboptimality gap among Δ_h. (see Definition 3.3)
s_h^k, a_h^k	States and actions introduced in the episode k by the policy π_k.
$Q_h^\pi(s,a), V_h^\pi(s)$	Ground-truth state-action value function and state value function of policy π.
$Q_h^*(s,a), V_h^*(s)$	The optimal ground-truth state-action value function and state value function.
$\Delta_h(s,a)$	Suboptimal gap with respect to the optimal policy π^*. (see Definition 3.3)
$\mathbb{P}_h, \mathbb{B}_h$	The ground-truth transition kernel and the Bellman operator.
κ_l	The quantification precision in the phase l. (see Algorithm 1)
γ_l	The confidence radius in the phase l. (see Theorem 5.1)
$\mathcal{C}_{h,l}^k$	Index sets during phase l in the episode k. (see Algorithm 1)
$\mathbf{w}_{h,l}^k, \mathbf{U}_{h,l}^k$	Empirical weights and covariance matrix in the phase l. (see Algorithm 1)
$\widetilde{\mathbf{w}}_{h,l}^k, \widetilde{\mathbf{U}}_{h,l}^k$	Quantified version of $\mathbf{w}_{h,l}^k$ and $\mathbf{U}_{h,l}^k$. (see Algorithm 1)
$\widehat{V}_h^k(s)$	The overall optimistic state value function. (see Algorithm 2)
$Q_{h,l}^k(s,a)$	Empirical state-action value function in phase l. (see Algorithm 2)
$V_{h,l}^k(s)$	Empirical state value function in phase l. (see Algorithm 2)
$\widehat{V}_{h,l}^k(s)$	Optimistic state value function in phase l. (see Definition C.1)
$\check{V}_{h,l}^k(s)$	Pessimistic state value function in phase l. (see Algorithm 2)
π_h^k	Policy played in the episode k. (see Algorithm 2)
$\pi_{h,l}^k$	Policy induced at state s during phase l of episode k. (see Algorithm 2)
$l_h^k(s)$	The index of the phase at which state s stops in episode k. (see Algorithm 2)
ϕ_h^k	The feature vector observed in the episode k. (see Algorithm 1)
$\mathcal{V}_{h,l}^k$	Function family of all optimistic state function $\widehat{V}_{h,l}^k$. (see Definition C.1)
γ_{l,l_+}	The confidence radius with covering on phase l_+. (see Definition C.3)
l_+	The phase offsets for the covering statement. (see Lemma C.5)
χ	The inflation on misspecification. (see Lemma C.6)
L_ζ	The deepest phase that tolerance ζ misspecification. (see Lemma C.8).
L_ε	The shallowest phase that guarantees ε accuracy. (see Lemma C.9).
Δ_h^k	The suboptimaility gap of played policy π_h^k at state s_h^k. (see Lemma C.13)
$\mathcal{G}_1$	The event defined in Definition C.3.
$\mathcal{G}_2$	The event defined in Definition D.13.
$\mathcal{G}_\varepsilon$	The condition defined in Definition C.10.

Table 2: Notations used in algorithm and proof

It is worth noting that one can check the definition of $\widehat{V}_{h,l}^k$ here is consistent with those computed in Algorithm 2 with $l < l_h^k(s)$. Therefore, we will not distinguish between the notations in the remainder of the proof.

The following lemma controls the distance between $\widehat{V}_h^k(s)$ and $\widehat{V}_{h,l}^k(s)$ for any phase l.

Lemma C.2. For any $(k,h,s) \in [K] \times [H] \times \mathcal{S}, l \in [l_h^k(s) - 1]$, it holds that

$$\check{V}_{h,l}^k(s) \le \widehat{V}_h^k(s) \le \widehat{V}_{h,l}^k(s), \quad |\widehat{V}_h^k(s) - \widehat{V}_{h,l}^k(s)| \le 6 \cdot 2^{-l}.$$

Moreover, for any tuple $(k,h,s,l_+) \in [K] \times [H] \times \mathcal{S} \times \mathbb{N}^+$, the difference $|\widehat{V}_h^k(s) - \widehat{V}_{h,l_+}^k(s)|$ is bounded by $6 \cdot 2^{-l_+}$, following the extension of the definition scope of $\widehat{V}_{h,l_+}^k$ as outlined in Definition C.1.

Lemma C.2 suggests that given any phase l_+, $\widehat{V}_{h,l}^k$ is close to $\widehat{V}_h^k$. This enables us to construct covering on $\widehat{V}_h^k$ using the covering on $\widehat{V}_{h,l}^k$.

C.2 Concentration of State Value Function $\widehat{V}_h^k(s)$

In this subsection, we provide a new analysis for bounding the self-normalized concentration of $\left\| \sum_\tau \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h \widehat{V}_h^k](s_h^\tau, a_h^\tau) - \widehat{V}_h^k(s_{h+1}^\tau) \big) \right\|_{\mathbf{U}^{-1}}$ to get rid of the $\log k$ factor in Vial et al. (2022).

To facilitate our proof, we define the filtration list $\mathcal{F}_h^k = \left\{\{s_i^j, a_i^j\}_{i=1,j=1}^{H,k-1}, \{s_i^k, a_i^k\}_{i=1}^{h}\right\}$. It is easy to verify that s_h^k, a_h^k are both $\mathcal{F}_h^k$-measurable. Also, for any function V built on $\mathcal{F}_h^k$, $[\mathbb{P}_h V](s_h^k, a_h^k) - V(s_{h+1}^k)$ is $\mathcal{F}_{h+1}^k$-measurable and it is also a zero-mean random variable conditioned on $\mathcal{F}_h^k$.

The first lemma we provide is similar with Vial et al. (2022), which shows the self-normalized concentration property for each phase l and any function $V \in \mathcal{V}_{h,l}^k$.

Definition C.3. For some fixed mapping $l \mapsto l_+ = l_+(l)$ that $l_+ \geq l$, we define the bad event as

$$\mathcal{B}_1(k,h,l,V) = \left\{ \left\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h V](s_h^\tau, a_h^\tau) - V(s_{h+1}^\tau)\big) \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} > \gamma_{l,l_+} \right\}.$$

The good event is defined by $\mathcal{G}_1 = \bigcap_{k=1}^K \bigcap_{h=1}^H \bigcap_{l\geq 1} \bigcap_{V \in \mathcal{V}_{h,l_+}^k} \mathcal{B}_1^{\complement}(k,h,l,V)$ where we define $\gamma_{l,l_+} = 5 l_+ dH \sqrt{\log(16ldH/\delta)} = \widetilde{\mathcal{O}}(ldH \log(\delta^{-1}))$.

Lemma C.4. The good event $\mathcal{G}_1$ defined in Definition C.3 happens with probability at least $1 - 2\delta$.

Lemma C.4 establishes the concentration bounds for any given phase l. However, the total number of phases for the state value function $V_h^k(s)$ can be bounded only trivially by$l = \mathcal{O}(\log K)$, resulting in $\log K$ dependence. To address this issue, the following lemma proposes a method to eliminate this logarithmic factor:

Lemma C.5. Under event $\mathcal{G}_1$, for any $(k,h,l) \in [K] \times [H] \times \mathbb{N}^+$,

$$\left\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h \widehat{V}_{h+1}^k](s_h^\tau, a_h^\tau) - \widehat{V}_{h+1}^k(s_{h+1}^\tau)\big) \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} \leq 1.1\gamma_l. \tag{C.1}$$

where we set $\gamma_l = \gamma_{l,l_+}$ with $l_+ = l + 20 + \lceil \log(ld) \rceil$.

Then Lemma C.5 immediately yields the following lemma regarding the estimation error of the state-action value function $Q_{h,l}^k$:

Lemma C.6. Under event $\mathcal{G}_1$, for any $(k,h,s) \in [K] \times [H] \times \mathcal{S}$, $l \in [l_h^k(s) - f_h^k(s)]$, $a_l \in \mathcal{A}_{h,l}^k(s)$,

$$\big| Q_{h,l}^k(s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) \big| \leq 2 \cdot 2^{-l} + \chi \sqrt{l} \zeta \tag{C.2}$$

where we define $\chi = 12\sqrt{d}H$.

Lemma C.6 build an estimation error for any $l \in [l_h^k(s) - 1]$. As we mentioned in the algorithm design, a larger l here will lead to more precise estimation (a smaller 2^{-l} term in (C.2)) but will suffer from a larger covering number (a larger γ_l term in (C.2)). Following a similar proof sketch from Vial et al. (2022), the next lemma shows that any action that is not eliminated has a low regret,

Lemma C.7. Fix some arbitrary $L_0 \geq 1$ and let $\chi = 12\sqrt{d}H$. Under event $\mathcal{G}_1$, for any $(k,h,s) \in [K] \times [H] \times \mathcal{S}$, $l \in [\min\{L_0, l_h^k(s) - f_h^k(s)\}]$, $a_{l+1} \in \mathcal{A}_{h,l+1}^k(s)$,

$$\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_{l+1}) \leq 8 \cdot 2^{-l} + 2l \cdot \chi \sqrt{L_0} \zeta.$$

C.3 The Impact of Misspecification Level ζ

Next, we are ready to show the criteria where Line 11 in Algorithm 2 will be triggered, which shows the impact of misspecification on this multi-phased estimation.

Lemma C.8. Under event $\mathcal{G}_1$, for any $(k,h) \in [K] \times [H]$ such that $f_h^k(s_h^k) = 0$, we have $l_h^k(s_h^k) > L_\zeta$ where L_ζ is the maximal integer satisfying $2^{-L_\zeta} \geq \chi L_\zeta^{1.5} \zeta$ for $\chi = 12\sqrt{d}H$, i.e., $L_\zeta = \Omega(\log(1/\zeta))$.

Equipped with Lemma C.8, the following lemma suggests that how much estimation precision ε can be achieved by accumulating the error $2^{-l_h^k(s_h^k)}$ that occurred in Lemma C.6.

https://doi.org/10.52202/079017-4154

Lemma C.9. Under event $\mathcal{G}_1$ and for all $\varepsilon > 0$, define L_ε to be the minimal integer satisfying $2^{-L_\varepsilon} \le 0.01\varepsilon/H$, i.e., $L_\varepsilon = \lceil -\log(0.01\varepsilon/H)\rceil$. When $L_\varepsilon \le L_\zeta$, then for any $\mathcal{K} \subseteq [K], h \in [H]$,

$$\sum_{k\in\mathcal{K}} 2^{-l_h^k(s_h^k)} \le 0.01|\mathcal{K}| \cdot \varepsilon/H + 2^{12} L_\varepsilon dH\gamma_{L_\varepsilon}^2 \cdot \varepsilon^{-1}.$$

The relationship between $L_\varepsilon \le L_\zeta$ can be translated to the relationship between ε and ζ. We characterize this condition as follows:

Definition C.10. Condition $\mathcal{G}_\varepsilon$ is defined for a given ε, and is satisfied if $L_\zeta \ge L_\varepsilon$ where L_ε is the minimal integer satisfying $2^{-L_\varepsilon} \le 0.01\varepsilon/H$ and L_ζ is the maximal integer satisfying $2^{-L_\zeta} \ge \chi L_\zeta^{1.5}\zeta$.

Lemma C.11. If $\varepsilon \ge \Omega\big(\sqrt{d}H^2\zeta\log^2(1/\zeta)\big)$, then $\mathcal{G}_\varepsilon$ is satisfied.

Proof. If $\varepsilon \ge \Omega\big(\sqrt{d}H^2\zeta\log^2(1/\zeta)\big)$, we have

$$2^{-L_\varepsilon} \ge 0.005\varepsilon/H \ge 2\chi L_\zeta^{1.5}\zeta \ge 2^{-L_\zeta}.$$

where the first inequality is given by the definition of L_ε, the last inequality is given by the definition of L_ζ, and the second inequality holds since $H\chi L_\zeta^{1.5} \le \mathcal{O}\big(\sqrt{d}H^2\log^2(1/\zeta)\big)$, and the last inequality is given by the definition of L_ε and L_ζ, respectively. Since 2^{-l} decreases as l increases, we can conclude that $L_\varepsilon \le L_\zeta$. □

The above analysis of the interplay between misspecification level ζ and precision ε yields the following important lemma in our proof, showing a local decision error across all $h \in [H]$:

Lemma C.12. Under Assumption 3.1, let $\gamma_l = 5(l+20+\lceil\log(ld)\rceil)dH\sqrt{\log(16ldH/\delta)}$, for some fixed $0 < \delta < 1/3$. With probability at least $1-3\delta$, for any $\varepsilon > \Omega\big(\sqrt{d}H^2\zeta\log^2(1/\zeta)\big)$ and $h \in [H]$, we have

$$\sum_{k=1}^{\infty} \mathbb{1}\left[V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) \ge \varepsilon\right] \le \mathcal{O}\big(d^3H^4\varepsilon^{-2}\log^4(dH\varepsilon^{-1})\log(\delta^{-1})\iota\big),$$

where ι refers to some polynomial of $\log\log(dH\varepsilon^{-1}\delta^{-1})$. This can also be written as

$$\Pr\left[\exists\varepsilon > \varepsilon_0, h \in [H], \sum_{k=1}^{\infty} \mathbb{1}\left[V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) > \varepsilon\right] > f(\varepsilon,\delta)\right] \le \delta.$$

with $\varepsilon_0 = \widetilde{\Omega}(\sqrt{d}H^2\zeta)$ and $f(\varepsilon,\delta) = \widetilde{\mathcal{O}}(d^3H^4\varepsilon^{-2}\log(\delta^{-1}))$.

C.4 From Local Step-wise Decision Error to Constant Regret

The next lemma shows that the total incurred suboptimality gap is constant if the minimal suboptimality gap Δ satisfies $\Delta > \varepsilon_0$.

Lemma C.13. Suppose an RL algorithm `Alg.` satisfies

$$\Pr\left[\exists\varepsilon > \varepsilon_0, h \in [H], \sum_{k=1}^{\infty} \mathbb{1}\left[V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) > \varepsilon\right] > f(\varepsilon,\delta)\right] \le \delta,$$

such that $f(\varepsilon,\delta) = \widetilde{\mathcal{O}}(C_1/\varepsilon + C_2/\varepsilon^2)$ where $C_1, C_2 > 0$ are constant in ε, but may depend on other quantities such as $d, H, \log(\delta^{-1})$. If the minimal suboptimality gap Δ satisfies $\Delta > \varepsilon_0$, then

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \Delta_h^k \le \widetilde{\mathcal{O}}(C_2H/\Delta + C_1H)$$

where $\Delta_h^k = \Delta_h\big(s_h^k, \pi_h^k(s_h^k)\big) = V_h^*(s_h^k) - Q_h^*\big(s_h^k, \pi_h^k(s_h^k)\big)$ is the suboptimality gap suffered in stage h of episode k.

The following Lemma is a refined version of Lemma 6.1 in He et al. (2021a) that removes the dependence between regret and number of episodes K.

Lemma C.14. For each MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{r_h\}, \{\mathbb{P}_h\})$ and any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\text{Regret}(K) < \widetilde{\mathcal{O}}\bigg(\sum_{k=1}^{K}\sum_{h=1}^{H}\Delta_h^k + H^2\log(1/\delta)\bigg).$$

We are now ready to prove Theorem 5.1:

Proof of Theorem 5.1. By plugging in Lemma C.12 and Lemma C.13 into Lemma C.14, we can reach the desired statement. □

D Proof of Lemmas in Appendix C

In this section, we prove lemmas outlined in Appendix C. Any proofs not included in this section are deferred to Appendix E.

D.1 Proof of Lemma C.2

Proof of Lemma C.2. According to the criteria for Line 11, we have $\check{V}_{h,l}^k(s) \le \widehat{V}_{h,l}^k(s)$ for any $l \in [l_h^k(s) - 1]$. From the definition of $\check{V}_{h,l}^k(s)$ and $\widehat{V}_{h,l}^k(s)$, they are monotonic in l that $\check{V}_{h,l-1}^k(s) \le \check{V}_{h,l}^k(s)$ and $\widehat{V}_{h,l}^k(s) \le \widehat{V}_{h,l-1}^k(s)$ hold. Combining with $\widehat{V}_{h+1}^k(s) = \widehat{V}_{h,l_h^k(s)-1}^k$, we have

$$\forall l \in [l_h^k(s) - 1], \check{V}_{h,l}^k(s) \le \widehat{V}_h^k(s) \le \widehat{V}_{h,l}^k(s) \quad \text{(D.1)}$$

From the definition of $\widehat{V}_{h,l}^k(s)$ and $\check{V}_{h,l}^k(s)$, we have

$$0 \le \widehat{V}_{h,l}^k(s) - \check{V}_{h,l}^k(s) \le \big(\widehat{V}_{h,l}^k(s) - V_{h,l}^k(s)\big) + \big(V_{h,l}^k(s) - \check{V}_{h,l}^k(s)\big) \le 6 \cdot 2^{-l}. \quad \text{(D.2)}$$

Plugging (D.1) into (D.2), we conclude that for any phase $l \in [l_h^k(s) - 1]$, it holds that $|\widehat{V}_h^k(s) - \widehat{V}_{h,l}^k(s)| \le 6 \cdot 2^{-l}$.

Now consider the extended state value function $\widehat{V}_{h,l_+}^k$ with an arbitrary $l_+ \in \mathbb{N}^+$. For every s where $l_+ \le l_h^k(s) - 1$, we have $|\widehat{V}_h^k(s) - V_{h,l_+}^k(s)| \le 6 \cdot 2^{-l_+}$ as reasoned above. For the other $s \in \mathcal{S}$ where $l_+ \ge l_h^k(s)$, we have $\widehat{V}_{h,l}^k(s) = \widehat{V}_h^k(s)$ following the procedure of Algorithm 2. This suggest that $|\widehat{V}_h^k(s) - \widehat{V}_{h,l_+}^k(s)| \le 6 \cdot 2^{-l_+}$ always holds. □

D.2 Proof of Lemma C.4

The following Lemma shows the rounding only cast bounded effects on the recovered parameters.

Lemma D.1. For any $(k, h, s) \in [K] \times [H] \times \mathcal{S}, l \in [l_h^k(s) - f_h^k(s)], a \in \mathcal{A}_{h,l}^k(s)$, it holds that

$$\big|\langle\boldsymbol{\phi}(s,a), \mathbf{w}_{h,l}^k\rangle - \langle\boldsymbol{\phi}(s,a), \widetilde{\mathbf{w}}_{h,l}^k\rangle\big| \le 0.01 \cdot 2^{-4l}, \big|\|\boldsymbol{\phi}(s,a)\|_{(\mathbf{U}_{h,l}^k)^{-1}} - \|\boldsymbol{\phi}(s,a)\|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}}\big| \le 0.1 \cdot 2^{-2l}.$$

The following lemma shows the number of episodes that are taken into regression $|\mathcal{C}_{h,l}^k|$ is bounded independently from the number of episodes k.

Lemma D.2. For any tuple $(k, h, l) \in [K] \times [H] \times \mathbb{N}^+$, we have $|\mathcal{C}_{h,l}^k| \le 16l \cdot 4^l \gamma_l^2 d$.

The following lemma shows the number of possible state value functions $|\mathcal{V}_{h,l}^k|$ is bounded independently from the number of episodes k.

Lemma D.3. For any tuple $(k, h, l) \in [K] \times [H] \times \mathbb{N}^+$, we have $|\mathcal{V}_{h,l}^k| \le (2^{22}d^6H^4)^{l^2d^2}$.

Now we are ready to prove Lemma C.4.

 https://doi.org/10.52202/079017-4154

Proof of Lemma C.4. Recall in Definition C.3, the good event defined by the union of each single bad event:

$$\mathcal{G}_1 = \bigcap_{k=1}^{K} \bigcap_{h=1}^{H} \bigcap_{l\geq 1} \bigcap_{V\in\mathcal{V}_{h,l_+}^k} \mathcal{B}_1^{\complement}(k,h,l,V),$$

where each single bad event is given by

$$\mathcal{B}_1(k,h,l,V) = \left\{ \left\| \sum_{\tau\in\mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h V](s_h^\tau, a_h^\tau) - V(s_{h+1}^\tau)\big) \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} > \gamma_l \right\},$$

in which $[\mathbb{P}_h V](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V(s)$.

Consider some fixed $(h,l)\in[H]\times\mathbb{N}^+$, $V\in\mathcal{V}_{h,l_+}^K$. Arrange elements of $\mathcal{C}_{h,l}^k$ in ascending order as $\{\tau_i\}_i$. Since the environment sample $s_{h+1}^{\tau_i}$ according to $\mathbb{P}_h(\cdot|s_h^{\tau_i}, a_h^{\tau_i})$, we have $[\mathbb{P}_h V](s_h^{\tau_i}, a_h^{\tau_i}) - V(s_{h+1}^{\tau_i})$ is $\mathcal{F}_h^{\tau_i}$-measurable with $\mathbb{E}\left[[\mathbb{P}_h V](s_h^{\tau_i}, a_h^{\tau_i}) - V(s_{h+1}^{\tau_i})\middle|\mathcal{F}_h^{\tau_i}\right] = 0$. Since $0\leq V(s_{h+1}^{\tau_i})\leq H$, we have $|[\mathbb{P}_h V](s_h^{\tau_i}, a_h^{\tau_i}) - V(s_{h+1}^{\tau_i})|\leq H$. This further leads to

$$\begin{aligned}
&\left\| \sum_{\tau\in\mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h V](s_h^\tau, a_h^\tau) - V(s_{h+1}^\tau)\big) \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} \\
&= \left\| \sum_{i=1}^{|\mathcal{C}_{h,l}^{k-1}|} \boldsymbol{\phi}_h^{\tau_i} \big([\mathbb{P}_h V](s_h^{\tau_i}, a_h^{\tau_i}) - V(s_{h+1}^{\tau_i})\big) \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} \\
&\leq H\sqrt{2d\ln\big(1+|\mathcal{C}_{h,l}^k|/(d\lambda)\big) + 2\ln(l^2 H|\mathcal{V}_{h,l_+}^K|/\delta)} \\
&\leq H\sqrt{2d\ln(1+l\cdot 4^l\gamma_l^2) + 2\ln(l^2 H(2^{22}d^6H^4)^{l_+^2 d^2}/\delta)} \\
&\leq \gamma_{l,l_+},
\end{aligned}$$

where the first inequality holds following from the good event of probability $1-\delta/(l^2 H|\mathcal{V}_{h,l_+}^K|)$ defined in Lemma H.2 over filtration $\{\mathcal{F}_h^{\tau_i}\}_i$, the second inequality is derived from combining Lemma D.2 and Lemma D.3, and the last inequality is given by Lemma G.3. According to Lemma H.2, we have the bad event $\bigcup_{k=1}^K \mathcal{B}_1(k,h,l,V)$ happens with probability at most $\delta/(l^2H|\mathcal{V}_{h,l_+}^K|)$. Taking union bound over all $(h,l)\in[H]\times\mathbb{N}^+$, $V\in\mathcal{V}_{h,l_+}^K$, we have the bad event happens with probability at most

$$\Pr[\mathcal{G}_1^{\complement}] \leq \sum_{h=1}^{H}\sum_{l=1}^{\infty}\sum_{V\in\mathcal{V}_{h,l_+}^K} \Pr\Big[\bigcup_{k=1}^K \mathcal{B}_1(k,h,l,V)\Big] \leq \sum_{h=1}^{H}\sum_{l=1}^{\infty}\sum_{V\in\mathcal{V}_{h,l_+}^K} \frac{\delta}{l^2 H|\mathcal{V}_{h,l_+}^K|} \leq 2\delta,$$

where the last inequality holds due to $\sum_{n\geq 1} n^{-2} = \pi^2/6$. This completes our proof. □

D.3 Proof of Lemma C.5

Proof of Lemma C.5. Denote the martingale difference between $\widehat{V}_{h,l_+}^k - \widehat{V}_h^k$ as:

$$\mu_{h,l}^k = [\mathbb{P}_h(\widehat{V}_{h,l_+}^k - \widehat{V}_{h+1}^k)](s_h^k, \pi_h^k(s_h^k)) - \big(\widehat{V}_{h,l_+}^k(s_{h+1}^k) - \widehat{V}_{h+1}^k(s_{h+1}^k)\big).$$

By triangle inequality:

$$\begin{aligned}
&\left\| \sum_{\tau\in\mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h \widehat{V}_{h+1}^k](s_h^\tau, a_h^\tau) - \widehat{V}_{h+1}^k(s_{h+1}^\tau)\big) \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} \\
&\leq \left\| \sum_{\tau\in\mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h V_{h,l_+}^k](s_h^\tau, a_h^\tau) - \widehat{V}_{h,l_+}^k(s_{h+1}^\tau)\big) \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} + \left\| \sum_{\tau\in\mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \mu_{h,l_+}^\tau \right\|_{(\mathbf{U}_{h,l}^k)^{-1}}. \quad \text{(D.3)}
\end{aligned}$$

According to the definition of event $\mathcal{G}_1$, we can upper bound the first term by

$$\bigg\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h V_{h,l_+}^k](s_h^\tau, a_h^\tau) - \widehat{V}_{h,l_+}^k(s_{h+1}^\tau) \big) \bigg\|_{(\mathbf{U}_{h,l}^k)^{-1}} \leq \gamma_{l,l_+} = \gamma_l. \tag{D.4}$$

According to Lemma C.2, we have $|\widehat{V}_{h,l_+}^k(s) - \widehat{V}_{h+1}^k(s)| \leq 6 \cdot 2^{-l_+}$ for any $s \in \mathcal{S}$. Thus, the difference can be bounded by $|\mu_{h,l_+}^\tau| \leq 6 \cdot 2^{-l_+}$. Consequently, we can bound the second term by

$$\begin{aligned} \bigg\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \mu_{h,l_+}^\tau \bigg\|_{(\mathbf{U}_{h,l}^k)^{-1}} &\leq 6 \cdot 2^{-l_+} \sqrt{|\mathcal{C}_{h,l}^k|} \\ &\leq 6 \cdot 2^{-l_+} \sqrt{16 l \cdot 4^l \gamma_l^2 d} \\ &= 24 \cdot 2^{l-l_+} \gamma_l \sqrt{ld}, \end{aligned} \tag{D.5}$$

where the first inequality is provided by Lemma H.3, utilizing the condition $|\mu_{h,l_+}^\tau| \leq 6 \cdot 2^{-l_+}$, the second inequality is from Lemma D.2. By plugging in the definition of l_+, we can further bound the final term of (D.5) by

$$\bigg\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \mu_{h,l_+}^\tau \bigg\|_{(\mathbf{U}_{h,l}^k)^{-1}} \leq 24 \cdot 2^{l-l_+} \gamma_l \sqrt{ld} \leq 24 \cdot 2^{-20} \gamma_l \leq 0.1 \gamma_l. \tag{D.6}$$

Plugging (D.4) and (D.6) into (D.3) yields the desired statement such that

$$\bigg\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \big([\mathbb{P}_h \widehat{V}_{h+1}^k](s_h^\tau, a_h^\tau) - \widehat{V}_{h+1}^k(s_{h+1}^\tau) \big) \bigg\|_{(\mathbf{U}_{h,l}^k)^{-1}} \leq 1.1 \gamma_l,$$

which concludes our proof. □

D.4 Proof of Lemma C.6

The following lemma shows the state-action value function $Q_{h,l}^k(s,a)$ is always well estimated.

Lemma D.4. Under event $\mathcal{G}_1$, for any $(k,h,l,s,a) \in [K] \times [H] \times \mathbb{N}^+ \times \mathcal{S} \times \mathcal{A}$,

$$\big| Q_{h,l}^k(s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) \big| \leq \big(1.2 + 8\sqrt{ld} H \cdot 2^l \zeta \big) \gamma_l \| \boldsymbol{\phi}(s,a) \|_{(\mathbf{U}_{h,l}^k)^{-1}} + 0.01 \cdot 2^{-4l} + 2H\zeta.$$

Equipped with Lemma D.1 and Lemma D.4, we are ready to prove Lemma C.6.

Proof of Lemma C.6. In case that $l \leq l_h^k(s) - f_h^k(s)$, for any $a \in \mathcal{A}_{h,l}^k(s)$, we have that

$$\begin{aligned} \| \boldsymbol{\phi}(s,a) \|_{(\mathbf{U}_{h,l}^k)^{-1}} &\leq \| \boldsymbol{\phi}(s,a) \|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}} + \big| \| \boldsymbol{\phi}(s,a) \|_{(\mathbf{U}_{h,l}^k)^{-1}} - \| \boldsymbol{\phi}(s,a) \|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}} \big| \\ &\leq 2^{-l} \gamma_l^{-1} + 0.1 \cdot 2^{-2l} \leq 1.1 \cdot 2^{-l} \gamma_l^{-1}, \end{aligned} \tag{D.7}$$

where the first inequality holds due to triangle inequality, and in the second inequality, the first term is satisfied since state s passes the criterion in Line 9 in phase l and the second term follows from Lemma D.1, and the last inequality is given by Lemma G.2 which implies $2^l > \gamma_l$. Plugging (D.7) into Lemma D.4 gives

$$\begin{aligned} \big| Q_{h,l}^k(s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) \big| &\leq 0.01 \cdot 2^{-4l} + 1.32 \cdot 2^{-l} + 8.8\sqrt{ld} H\zeta + 2H\zeta \\ &\leq 2 \cdot 2^{-l} + 12\sqrt{ld} H\zeta, \end{aligned}$$

which proves the desired statement. □

 https://doi.org/10.52202/079017-4154

D.5 Proof of Lemma C.7

Equipped with Lemma C.6, we are able to show several properties of the state value function $V_{h,l}^k$ throught the arm-elimination process. The first lemma suggests that for any action $a_l \in \mathcal{A}_{h,l}^k(s)$, there is at least one action $a_{l+1} \in \mathcal{A}_{h,l+1}^k(s)$ close to a_l in terms of the Bellman operator $[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)$ after the elimination.

Lemma D.5. Under event $\mathcal{G}_1$, for any $(k,h,s) \in [K] \times [H] \times \mathcal{S}, l \in [\min\{L_0, l_h^k(s) - f_h^k(s)\}], a_l \in \mathcal{A}_{h,l}^k(s)$, there exists $a_{l+1} \in \mathcal{A}_{h,l+1}^k(s)$ that

$$[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a_l) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a_{l+1}) \leq 2\chi\sqrt{L_0}\zeta$$

where $\chi = 12\sqrt{d}H$ for arbitrary $L_0 \geq 1$.

Then the following lemma shows that by induction on stage $h \in [H]$, we can show the elimination process keep at least one near-optimal action $a_{l+1} \in \mathcal{A}_{h,l+1}^k(s)$.

Lemma D.6. Under event $\mathcal{G}_1$, for any $(k,h,s) \in [K] \times [H] \times \mathcal{S}, l \in [\min\{L_0, l_h^k(s) - f_h^k(s)\}]$, there exists $a_{l+1} \in \mathcal{A}_{h,l+1}^k(s)$ that,

$$\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a_{l+1}) \leq 2l \cdot \chi\sqrt{L_0}\zeta.$$

where $\chi = 12\sqrt{d}H$ for arbitrary $L_0 \geq 1$.

The following two lemmas indicate that the state value function $V_{h,l}^k(s)$ on stage h is a good estimation for the state value function given by Bellman operator $V(s) = \max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)$.

Lemma D.7. Under event $\mathcal{G}_1$, for any $(k,h,s) \in [K] \times [H] \times \mathcal{S}, l \in [\min\{L_0, l_h^k(s) - f_h^k(s)\}]$,

$$\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - V_{h,l}^k(s) \leq 2 \cdot 2^{-l} + (2l-1)\chi\sqrt{L_0}\zeta.$$

where $\chi = 12\sqrt{d}H$ for arbitrary $L_0 \geq 1$.

Lemma D.8. Under event $\mathcal{G}_1$, for any $(k,h,s) \in [K] \times [H] \times \mathcal{S}, l \in [\min\{L_0, l_h^k(s) - f_h^k(s)\}]$,

$$V_{h,l}^k(s) - \max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) \leq 2 \cdot 2^{-l} + \chi\sqrt{L_0}\zeta,$$

where $\chi = 12\sqrt{d}H$ for arbitrary $L_0 \geq 1$.

Now we are ready to show any actions remaining in the elimination process are near-optimal.

Proof of Lemma C.7. First, according to Lemma D.7, we can write

$$\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - V_{h,l}^k(s) \leq 2 \cdot 2^{-l} + (2l-1)\chi\sqrt{L_0}\zeta. \tag{D.8}$$

Any action $a_{l+1} \in \mathcal{A}_{h,l+1}^k(s)$ passes the elimination process will satisfy:

$$Q_{h,l}^k(s,a_{l+1}) \geq V_{h,l}^k(s) - 4 \cdot 2^{-l}. \tag{D.9}$$

According to Lemma C.6 with the condition that $l \leq L_0$, we have that the empirical state-action value function $Q_{h,l}^k(s,\cdot)$ is a good estimation for $[\mathbb{B}_h \widehat{V}_{h+1}^k](s,\cdot)$ among every $a_{l+1} \in \mathcal{A}_l^k(s)$ under event $\mathcal{G}_1$:

$$\left|[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a_{l+1}) - Q_{h,l}^k(s,a_{l+1})\right| \leq 2 \cdot 2^{-l} + \chi\sqrt{L_0}\zeta. \tag{D.10}$$

Combining (D.8), (D.9), and (D.10) gives

$$\begin{aligned}
&\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a_{l+1}) \\
&= \big(\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - V_{h,l}^k(s)\big) + \big(V_{h,l}^k(s) - Q_{h,l}^k(s,a_{l+1})\big) + \big(Q_{h,l}^k(s,a_{l+1}) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a_{l+1})\big) \\
&\leq \big(2 \cdot 2^{-l} + (2l-1)\chi\sqrt{L_0}\zeta\big) + 4 \cdot 2^{-l} + \big(2 \cdot 2^{-l} + \chi\sqrt{L_0}\zeta\big) \\
&= 8 \cdot 2^{-l} + 2l \cdot \chi\sqrt{L_0}\zeta,
\end{aligned}$$

which proves the desired statement. □

D.6 Proof of Lemma C.8

The following two lemmas demonstrate that, at stage h, both the optimistic state value function, $\widehat{V}_{h,l}^k(s)$, and the pessimistic state value function, $\check{V}_{h,l}^k(s)$, exhibit a gap relative to the state value function determined by the Bellman operator, given as $V(s) = \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)$.

Lemma D.9. Under event $\mathcal{G}_1$, for any $(k,h,s) \in [K]\times[H]\times\mathcal{S}, l \in [\min\{L_0, l_h^k(s) - f_h^k(s)\}]$,

$$\min\left\{V_{h,l}^k(s) + 3\cdot 2^{-l}, \widehat{V}_{h,l-1}^k(s)\right\} - \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) \ge 2^{-l} - (2l-1)\chi\sqrt{L_0}\zeta,$$

where $\chi = 12\sqrt{d}H$ for arbitrary $L_0 \ge 1$. In case that $l \le l_h^k(s) - 1$, the inequality is equivalent to

$$\widehat{V}_{h,l}^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) \ge 2^{-l} - (2l-1)\chi\sqrt{L_0}\zeta.$$

Lemma D.10. Under event $\mathcal{G}_1$, for any $(k,h,s) \in [K]\times[H]\times\mathcal{S}, l \in [\min\{L_0, l_h^k(s) - f_h^k(s)\}]$,

$$\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - \max\left\{V_{h,l}^k(s) - 3\cdot 2^{-l}, \check{V}_{h,l-1}^k(s)\right\} \ge 2^{-l} - \chi\sqrt{L_0}\zeta,$$

where $\chi = 12\sqrt{d}H$ for arbitrary $L_0 \ge 1$. In case that $l \le l_h^k(s) - 1$, the inequality is equivalent to

$$\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - \check{V}_{h,l}^k(s) \ge 2^{-l} - \chi\sqrt{L_0}\zeta.$$

Proof of Lemma C.8. Set $L_0 = L_\zeta$ be the maximal integer satisfying $2^{-L_\zeta} - \chi L_\zeta^{1.5}\zeta \ge 0$. Combining Lemma D.10 and Lemma D.9, for any $l \in [\min\{L_0, l_h^k(s) - f_h^k(s)\}]$, we have that

$$\begin{aligned}
&\min\left\{V_{h,l}^k(s) + 3\cdot 2^{-l}, \widehat{V}_{h,l-1}^k(s)\right\} - \max\left\{V_{h,l}^k(s) - 3\cdot 2^{-l}, \check{V}_{h,l-1}^k(s)\right\} \\
&= \left(\widehat{V}_{h,l}^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)\right) + \left(\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - \check{V}_{h,l}^k(s)\right) \\
&\ge \left(2^{-l} - (2l-1)\chi\sqrt{L_0}\zeta\right) + \left(2^{-l} - \chi\sqrt{L_0}\zeta\right) \\
&= 2\cdot 2^{-l} - 2l\cdot\chi\sqrt{L_0}\zeta \\
&\ge 2\cdot 2^{-L_0} - 2\chi L_0^{1.5}\zeta \ge 0.
\end{aligned}$$

where the second inequality holds since 2^{-l} decreases as l increases and the last inequality holds according to the selection of L_0.

When $f_h^k(s) = 0$, consider $l = l_h^k(s)$. The above reasoning indicates the criterion in Line 11 can never satisfied. Thus $f_h^k(s) = 0$ can only happen if $l_h^k(s) > L_0 = L_\zeta$. □

D.7 Proof of Lemma C.9

By partitioning $[K]$ based on whether Algorithm 2 stops before phase L_ε, we can prove Lemma C.9. Specifically, Lemma D.2 bounds the number of episodes in which Algorithm 2 stops before phase L_ε. This allows us to establish an upper bound for the desired summation over these episodes. Furthermore, for episodes that stop after phase L_ε, the contribution of $2^{-l_h^k(s_h^k)}\gamma_{l_h^k(s_h^k)}$ is small according to the definition of L_ε.

Proof of Lemma C.9. Denote $\mathcal{C}_{h,+}^K = [K] - \bigcup_{l=1}^{L_\varepsilon-1} \mathcal{C}_{h,l}^K$. In this sense, we have

$$\sum_{k\in\mathcal{K}} 2^{-l_h^k(s_h^k)} = \sum_{k\in\mathcal{K}\cap\mathcal{C}_{h,+}^K} 2^{-l_h^k(s_h^k)} + \sum_{l=1}^{L_\varepsilon-1}\sum_{k\in\mathcal{K}\cap\mathcal{C}_{h,l}^K} 2^{-l_h^k(s_h^k)}. \tag{D.11}$$

From the construction of $\mathcal{C}_{h,l}^K$, we have $l_h^k(s_h^k) = l$ for any $k \in \mathcal{C}_{h,l}^K$. Fix some $k \in \mathcal{C}_{h,+}^K$. If $f_h^k(s_h^k) = 0$, we have $l_h^k(s_h^k) \ge L_\zeta \ge L_\varepsilon$ where the first inequality is given by Lemma C.8 and the second inequality is given by the assignment of L_ε. Otherwise, we have $l_h^k(s_h^k) \ge L_\varepsilon$ according to

 https://doi.org/10.52202/079017-4154

the definition of $\mathcal{C}_{h,l}^K$. This indicates $l_h^k(s_h^k) \geq L_\varepsilon$ holds for any $k \in \mathcal{C}_{h,+}^K$. This allow is to bound the first term by

$$\sum_{k \in \mathcal{K} \cap \mathcal{C}_{h,+}^K} 2^{-l_h^k(s_h^k)} \leq \sum_{k \in \mathcal{K} \cap \mathcal{C}_{h,+}^K} 2^{-L_\varepsilon} \leq 0.01|\mathcal{K}| \cdot \varepsilon/H, \tag{D.12}$$

where the first inequality holds since $l_h^k(s_h^k) > L_\varepsilon$ and the second inequality holds from both $2^{-L_\varepsilon} \leq 0.01\varepsilon/H$ and $|\mathcal{K} \cap \mathcal{C}_{h,+}^K| \leq |\mathcal{K}|$.

Furthermore, we can bound the second term by

$$\begin{aligned}
\sum_{l=1}^{L_\varepsilon - 1} \sum_{k \in \mathcal{K} \cap \mathcal{C}_{h,l}^K} 2^{-l_h^k(s_h^k)} &\leq \sum_{l=1}^{L_\varepsilon - 1} |\mathcal{K} \cap \mathcal{C}_{h,l}^K| \cdot 2^{-l} \\
&\leq \sum_{l=1}^{L_\varepsilon - 1} 16l \cdot 4^l \gamma_l^2 d \cdot 2^{-l} \\
&\leq 16 L_\varepsilon d \cdot 2^{L_\varepsilon} \gamma_{L_\varepsilon}^2 \leq 2^{12} L_\varepsilon d H \gamma_{L_\varepsilon}^2 \varepsilon^{-1}.
\end{aligned} \tag{D.13}$$

where the second inequality is given by Lemma D.2, and the last inequality holds due to $0.005\varepsilon/H \leq 2^{-L_\varepsilon}$ which is because L_ε is a minimal integer such that $2^{-L_\varepsilon} \leq 0.01\varepsilon/H$.

Finally, plugging (D.12) and (D.13) into (D.11) gives

$$\sum_{k \in \mathcal{K}} 2^{-l_h^k(s_h^k)} \leq 0.01|\mathcal{K}| \cdot \varepsilon/H + 2^{12} L_\varepsilon d H \gamma_{L_\varepsilon}^2 \varepsilon^{-1}.$$

□

D.8 Proof of Lemma C.12

The following lemma provides an upper bound for the underestimation error of the empirical state value function $\widehat{V}_h^k$ with respect to the optimal state value function V_h^*.

Lemma D.11. Under event $\mathcal{G}_1$ and for all $\varepsilon > 0$ that $\mathcal{G}_\varepsilon$ is satisfied, for any $(k, h, s) \in [K] \times [H] \times \mathcal{S}$,

$$V_h^*(s) - \widehat{V}_h^k(s) \leq 0.07\varepsilon.$$

As $\widehat{V}_h^k$ represents an empirical state value function with potentially optimal policy $\pi_h^k(s)$, the following lemma provides an upper bound for the overestimation error of $\widehat{V}_h^k$ with respect to deploying the policy $\pi_h^k(s)$ on the ground-truth transition kernel.

Lemma D.12. Under event $\mathcal{G}_1$ and for all $\varepsilon > 0$ that $\mathcal{G}_\varepsilon$ is satisfied, for any $(k, h, s) \in [K] \times [H] \times \mathcal{S}$,

$$\widehat{V}_h^k(s) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, \pi_h^k(s)) \leq 20 \cdot 2^{-l_h^k(s)} + 0.16\varepsilon/H.$$

To start with, we define a good event according to:

Definition D.13. For some $\varepsilon > 0$, let $\mathcal{K}_h^\varepsilon = \{k \in [K] : V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) \geq \varepsilon\}$. We define the bad event as

$$\mathcal{B}_2(h, \varepsilon) = \left\{ \sum_{k \in \mathcal{K}_h^\varepsilon} \sum_{h'=h}^{H} \eta_{h'}^k > 4\sqrt{H^3 |\mathcal{K}_h^\varepsilon| \log(4H|\mathcal{K}_h^\varepsilon| \log(\varepsilon^{-1})/\delta)} \right\}.$$

where $\eta_h^k = [\mathbb{P}_h(\widehat{V}_{h+1}^k - V_{h+1}^{\pi^k})](s_h^k, \pi_h^k(s_h^k)) - \big(\widehat{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big)$. The good event is defined as $\mathcal{G}_2 = \bigcap_{h=1}^H \bigcap_{l \geq 1} \mathcal{B}_2^{\complement}(h, 2^{-l})$.

The following lemma provides the concentration property such that the cumulative sample error is small with high probability.

Lemma D.14. Event $\mathcal{G}_2$ happens with probability at least $1 - \delta$.

Using the above results, we can bound the instantaneous regret of any subsets once the misspecification level is appropriately controlled,

Lemma D.15. Under event $\mathcal{G}_1, \mathcal{G}_2$ and for all $\varepsilon > 0$ that $\mathcal{G}_\varepsilon$ is satisfied, for any $\mathcal{K} \subseteq [K]$ and $h \in [H]$, it satisfies that

$$\sum_{k \in \mathcal{K}} \left(V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k)\right) \le 0.49|\mathcal{K}|\varepsilon + 2^{17} L_\varepsilon d H^2 \gamma_{L_\varepsilon}^2 \varepsilon^{-1} + 4\sqrt{H^3|\mathcal{K}| \log(4H|\mathcal{K}| \log(\varepsilon^{-1})/\delta)}.$$

With all lemmas stated above, we can show Cert-LSVI-UCBachieves constant step-wise decision error. The following lemma gives a sufficient condition that $\mathcal{G}_\varepsilon$ defined in Definition C.10 is satisfied.

Now, we are ready to prove Lemma C.12.

Proof of Lemma C.12. We focus on the case where the good event $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_\varepsilon$ occurs. By the union bound statement over Lemma C.4 and Lemma D.14, and Lemma C.11, this good event happens with a probability of at least $1 - 3\delta$ and requires $\varepsilon \ge \Omega(\zeta \sqrt{d} H^2 \log^2(dH\zeta^{-1}))$. W.l.o.g, consider $\mathcal{K}_h^\varepsilon$ for some $h \in [H]$ and $\varepsilon = 2^{-l}$ where $l > 0$ is an integer. On the one hand, we have

$$\sum_{k \in \mathcal{K}_h^\varepsilon} \left(V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k)\right) \ge |\mathcal{K}_h^\varepsilon|\varepsilon. \tag{D.14}$$

On the other hand, Lemma D.15 gives

$$\begin{aligned}\sum_{k \in \mathcal{K}_h^\varepsilon} \left(V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k)\right) &\le 0.49|\mathcal{K}_h^\varepsilon|\varepsilon + 2^{17} L_\varepsilon d H^2 \gamma_{L_\varepsilon}^2 \varepsilon^{-1} \\ &\quad + 4\sqrt{H^3|\mathcal{K}_h^\varepsilon| \log(4H|\mathcal{K}_h^\varepsilon| \log(\varepsilon^{-1})/\delta)}.\end{aligned} \tag{D.15}$$

Combining (D.14) and (D.15) gives

$$0.51|\mathcal{K}_h^\varepsilon|\varepsilon \le 2^{17} L_\varepsilon d H^2 \gamma_{L_\varepsilon}^2 \varepsilon^{-1} + 4\sqrt{H^3|\mathcal{K}_h^\varepsilon| \log(4H|\mathcal{K}_h^\varepsilon| \log(\varepsilon^{-1})/\delta)}.$$

Plugging the value of γ_{L_ε}, we have

$$\begin{aligned}0.51|\mathcal{K}_h^\varepsilon|\varepsilon &\le 2^{22} L_\varepsilon (L_\varepsilon + \log(2^{20} dH))^2 d^3 H^4 \varepsilon^{-1} \log(16 L_\varepsilon d/\delta) \\ &\quad + 4\sqrt{H^3|\mathcal{K}_h^\varepsilon| \log(4H|\mathcal{K}_h^\varepsilon| \log(\varepsilon^{-1})/\delta)}.\end{aligned} \tag{D.16}$$

According to Lemma G.4, (D.16) implies

$$|\mathcal{K}_h^\varepsilon| \le \mathcal{O}(L_\varepsilon (L_\varepsilon + \log(dH))^2 d^3 H^4 \varepsilon^{-2} \log(L_\varepsilon d) \log(\delta^{-1})\iota),$$

where ι refers to a polynomial of $\log\log(dH\varepsilon^{-1}\delta^{-1})$. With the definition of L_ε, we conclude that

$$|\mathcal{K}_h^\varepsilon| \le \mathcal{O}(d^3 H^4 \varepsilon^{-2} \log^4(dH\varepsilon^{-1}) \log(\delta^{-1})\iota).$$

□

D.9 Proof of Lemma C.13

Proof of Lemma C.13. From the definition of suboptimality gap, we have

$$\begin{aligned}\Delta_h^k &= V_h^*(s_h^k) - [\mathbb{B}_h V_{h+1}^*](s_h^k, \pi_h^k(s_h^k)) \\ &\le V_h^*(s_h^k) - [\mathbb{B}_h V_{h+1}^{\pi^k}](s_h^k, \pi_h^k(s_h^k)) \\ &= V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k).\end{aligned} \tag{D.17}$$

According to the assumption,

$$\sum_{k=1}^K \mathbb{1}\left[V_h^*(s_1^k) - V_h^{\pi^k}(s_h^k) \ge \varepsilon\right] \le \left(\frac{C_1}{\varepsilon} + \frac{C_2}{\varepsilon^2}\right) \log^a\left(\frac{C_1}{\varepsilon} + \frac{C_2}{\varepsilon^2}\right)$$

https://doi.org/10.52202/079017-4154

holds for every $\varepsilon > \varepsilon_0$ with probability at least $1-\delta$, replacing the $V_h^*(s_1^k) - V_h^{\pi^k}(s_h^k)$ with its lower bound Δ_h^k yields for every $\varepsilon > \varepsilon_0$,

$$\sum_{k=1}^K \mathbb{1}\left[\Delta_h^k \ge \varepsilon\right] \le \Big(\frac{C_1}{\varepsilon} + \frac{C_2}{\varepsilon^2}\Big) \log^a \Big(\frac{C_1}{\varepsilon} + \frac{C_2}{\varepsilon^2}\Big).$$

In addition, according to the definition of minimal suboptimality gap Δ in Definition 3.3, we have Δ_h^k is either 0 or no less than Δ. Since for any $x \in \{0\} \cup [\Delta, H]$, it holds that $x \le \Delta \cdot \mathbb{1}[x \ge \Delta] + \int_\Delta^H \mathbb{1}[x \ge \varepsilon]\,\mathrm{d}\varepsilon$, we decompose the total suboptimality incurred in stage h by

$$\begin{aligned}\sum_{k=1}^K \Delta_h^k &\le \sum_{k=1}^K \left(\Delta \cdot \mathbb{1}\left[\Delta_h^k \ge \Delta\right] + \int_\Delta^H \mathbb{1}\left[\Delta_h^k \ge \varepsilon\right] \mathrm{d}\varepsilon\right)\\ &= \Delta \sum_{k=1}^K \mathbb{1}\left[\Delta_h^k \ge \Delta\right] + \int_\Delta^H \sum_{k=1}^K \mathbb{1}\left[\Delta_h^k \ge \varepsilon\right] \mathrm{d}\varepsilon. \end{aligned} \tag{D.18}$$

In case that $\varepsilon_0 \le \Delta$, the first term in (D.18) can be bounded by

$$\Delta \sum_{k=1}^K \mathbb{1}\left[\Delta_h^k \ge \Delta\right] \le \Delta\Big(\frac{C_1}{\Delta} + \frac{C_2}{\Delta^2}\Big) \log^a \Big(\frac{C_1}{\Delta} + \frac{C_2}{\Delta^2}\Big). \tag{D.19}$$

We can further bound the second term by

$$\begin{aligned}\int_\Delta^H \sum_{k=1}^K \mathbb{1}\left[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \ge \varepsilon\right] \mathrm{d}\varepsilon &\le \int_\Delta^H \Big(\frac{C_1}{\varepsilon} + \frac{C_2}{\varepsilon^2}\Big) \log^a \Big(\frac{C_1}{\varepsilon} + \frac{C_2}{\varepsilon^2}\Big) \mathrm{d}\varepsilon\\ &\le \log^a \Big(\frac{C_1}{\Delta} + \frac{C_2}{\Delta^2}\Big) \cdot \Big(C_1 \ln \frac{H}{\Delta} + \frac{C_2}{\Delta}\Big)\\ &\le (C_1 \log H + C_2/\Delta) \cdot \mathrm{polylog}(C_1, C_2, \Delta^{-1}). \end{aligned} \tag{D.20}$$

Plugging (D.19) and (D.20) into (D.18) with summation over $h \in [H]$, we conclude that the total suboptimality incurred in stage h is bounded by

$$\begin{aligned}\sum_{k=1}^K \sum_{h=1}^H \Delta_h^k &\le H \cdot (C_1 + C_2/\Delta + C_1 \log H + C_2/\Delta) \cdot \mathrm{polylog}(C_1, C_2, \Delta^{-1})\\ &\le \widetilde{\mathcal{O}}(C_2 H/\Delta + C_1 H).\end{aligned}$$

□

D.10 Proof of Lemma C.14

Proof of Lemma C.14. For a given policy π and any state $s_h \in \mathcal{S}$, we have

$$\begin{aligned}&V_h^*(s_h) - V_h^\pi(s_h)\\ &= \big(V_h^*(s_h) - [\mathbb{B}_h V_{h+1}^*](s_h, \pi_h(s_h))\big) + \big([\mathbb{B}_h V_{h+1}^*](s_h, \pi_h(s_h)) - [\mathbb{B}_h V_{h+1}^\pi](s_h, \pi_h(s_h))\big)\\ &= \Delta_h(s_h, \pi_h(s_h)) + [\mathbb{P}_h(V_{h+1}^* - V_{h+1}^\pi)](s_h, \pi_h(s_h)),\end{aligned}$$

where the second equality is given by the definition of suboptimality gap $\Delta_h(\cdot,\cdot)$ in Definition 3.3. Taking expectation on both sides with respect to the randomness of state-transition and taking telescoping sum over all $h \in [H]$ gives

$$V_1^*(s_1) - V_h^\pi(s_1) = \mathbb{E}\left[\sum_{h=1}^H \Delta_h(s_h, \pi_h(s_h))\right],$$

where $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, \pi_h(s_h))$. Let the filtration list be $\mathcal{F}^k = \Big\{\{s_i^j, a_i^j\}_{i=1,j=1}^{H,k-1}\Big\}$, we have

$$\mathbb{E}\Big[\sum_{h=1}^H \Delta_h^k \Big| \mathcal{F}^k\Big] = V_1^*(s_1^k) - V_h^{\pi^k}(s_1^k).$$

Denote random variable $\eta^k = \big(V_1^*(s_1^k) - V_h^{\pi^k}(s_1^k)\big) - \sum_{h=1}^H \Delta_h^k$. We can see η^k is $\mathcal{F}_{k+1}$-measurable with $|\mathbb{E}[\eta^k|\mathcal{F}^k]| = 0$. Furthermore, for the variance of η^k, we have

$$\begin{aligned}\mathrm{Var}[\eta^k|\mathcal{F}^k] &\le \mathbb{E}\Big[\Big(\sum_{h=1}^H \Delta_h^k\Big)^2\Big|\mathcal{F}^k\Big]\\ &\le H^2\,\mathbb{E}\Big[\sum_{h=1}^H \Delta_h^k\Big|\mathcal{F}^k\Big]\\ &= H^2\big(V_1^*(s_1^k) - V_h^{\pi^k}(s_1^k)\big),\end{aligned}$$

where the first inequality holds due to $\mathrm{Var}[X] \le \mathbb{E}[(X-t)^2]$ for any fixed t, the second inequality follows $0 \le \Delta_h^k \le H$. As a result, the total variance of the random variables $\{\eta^k\}$ can be bounded by

$$\sum_{k=1}^K \mathrm{Var}[\eta^k|\mathcal{F}^k] \le \sum_{k=1}^K H^2\big(V_1^*(s_1^k) - V_h^{\pi^k}(s_1^k)\big) = H^2\mathrm{Regret}(K).$$

Let $F(x) = \sqrt{2xH^2\log(x/\delta)} + H^2\log(x/\delta)$, using peeling technique, we can write

$$\begin{aligned}&\Pr\Big[\Big(\sum_{k=1}^K \eta^k\Big) \ge F(\mathrm{Regret}(K)), 1 < \mathrm{Regret}(K)\Big]\\ &= \Pr\Big[\Big(\sum_{k=1}^K \eta^k\Big) \ge F(\mathrm{Regret}(K)), 1 < \mathrm{Regret}(K), \sum_{k=1}^K \mathrm{Var}[\eta^k|\mathcal{F}^k] \le H^2\mathrm{Regret}(K)\Big]\\ &\le \sum_{i=1}^\infty \Pr\Big[\Big(\sum_{k=1}^K \eta^k\Big) \ge F(\mathrm{Regret}(K)), 2^{i-1} < \mathrm{Regret}(K) \le 2^i, \sum_{k=1}^K \mathrm{Var}[\eta^k|\mathcal{F}^k] \le H^2\mathrm{Regret}(K)\Big]\\ &\le \sum_{i=1}^\infty \Pr\Big[\Big(\sum_{k=1}^K \eta^k\Big) \ge F(2^i), \sum_{k=1}^K \mathrm{Var}[\eta^k|\mathcal{F}^k] \le 2^iH^2\Big]\\ &\le \sum_{i=1}^\infty \exp\Big(\frac{-F(2^i)^2}{2^{i+1}H^2 + 2F(2^i)H^2/3}\Big), \qquad \text{(D.21)}\end{aligned}$$

where the last inequality follows Lemma H.5. Plugging $F(x) = \sqrt{2xH^2\log(x/\delta)} + H^2\log(x/\delta)$ back into (D.21) yields

$$\begin{aligned}&\Pr\Big[\Big(\sum_{k=1}^K \eta^k\Big) \ge \sqrt{2\mathrm{Regret}(K)H^2\log(\mathrm{Regret}(K)/\delta)} + H^2\log(\mathrm{Regret}(K)/\delta), 1 < \mathrm{Regret}(K)\Big]\\ &\le \sum_{i=1}^\infty \exp(-\log(2^i/\delta)) = \sum_{i=1}^\infty \delta/2^i = \delta.\end{aligned}$$

Therefore, whenever $\mathrm{Regret}(K) > 1$, with probability at least $1-\delta$, we have

$$\sum_{k=1}^K \eta^k < \sqrt{2\mathrm{Regret}(K)H^2\log(\mathrm{Regret}(K)/\delta)} + H^2\log(\mathrm{Regret}(K)/\delta).$$

Combining with the fact that $\mathrm{Regret}(K) = \sum_{k=1}^K \eta^k + \sum_{k=1}^K\sum_{h=1}^H \Delta_h^k$, we have

$$\mathrm{Regret}(K) < \sum_{k=1}^K\sum_{h=1}^H \Delta_h^k + \sqrt{2\mathrm{Regret}(K)H^2\log(\mathrm{Regret}(K)/\delta)} + H^2\log(\mathrm{Regret}(K)/\delta),$$

https://doi.org/10.52202/079017-4154

whenever Regret$(K) > 1$. Taking $x = \text{Regret}(K)$, $a = \sum_{k=1}^K \sum_{h=1}^H \Delta_h^k + H^2 \log(\text{Regret}(K)/\delta)$, and $b = 2H^2 \log(\text{Regret}(K)/\delta)$, inequality (1) yields

$$\begin{aligned}\text{Regret}(K) &\leq 2\sum_{k=1}^K \sum_{h=1}^H \Delta_h^k + 2H^2 \log(\text{Regret}(K)/\delta) + 4H^2 \log(\text{Regret}(K)/\delta) \\ &= 2\sum_{k=1}^K \sum_{h=1}^H \Delta_h^k + 6H^2 \log(1/\delta) + 6H^2 \log(\text{Regret}(K)) \qquad \text{(D.22)}\end{aligned}$$

According to the fact that $x \leq a \log x + b \Rightarrow x \leq 4a \log(2a) + 2b$, letting $x = \text{Regret}(K), a = 2\sum_{k=1}^K \sum_{h=1}^H \Delta_h^k + 6H^2 \log(1/\delta)$ and $b = 6H^2$, (D.22) becomes

$$\text{Regret}(K) \leq \left(8\sum_{k=1}^K \sum_{h=1}^H \Delta_h^k + 24H^2 \log(1/\delta)\right) \log\left(4\sum_{k=1}^K \sum_{h=1}^H \Delta_h^k + 12H^2 \log(1/\delta)\right) + 12H^2.$$

Hiding the logarithmic factors within the $\widetilde{O}$ notation yields

$$\text{Regret}(K) < \widetilde{\mathcal{O}}\Big(\sum_{k=1}^K \sum_{h=1}^H \Delta_h^k + H^2 \log(1/\delta)\Big).$$

□

E Proof of Lemmas in Appendix D

In this section, we prove lemmas outlined in Appendix D. Any proofs not included in this section are deferred to Appendix F.

E.1 Proof of Lemma D.1

We first introduce the claim from Vial et al. (2022) controlling the rounding error:

Lemma E.1 (Claim 1, Vial et al. 2022, restate)**.** For any $(k, h, l, s, a) \in [K] \times [H] \times \mathbb{N}^+ \times \mathcal{S} \times \mathcal{A}$, we have

$$\boldsymbol{\phi}(s,a)^\top (\mathbf{w}_{h,l}^k - \widetilde{\mathbf{w}}_{h,l}^k) \leq \sqrt{d}\kappa_l, \big|\|\boldsymbol{\phi}(s,a)_{(\mathbf{U}_{h,l}^k)^{-1}} - \|\boldsymbol{\phi}(s,a)\|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}}\big| \leq \sqrt{d\kappa_l},$$

where κ_l is used to quantify the vector $\mathbf{w}_{h,l}^k$ and inverse matrix $(\mathbf{U}_{h,l}^l)^{-1}$.

Proof of Lemma D.1. From Lemma E.1 we have

$$\big|\big\langle \boldsymbol{\phi}(s,a), \mathbf{w}_{h,l}^k \big\rangle - \big\langle \boldsymbol{\phi}(s,a), \widetilde{\mathbf{w}}_{h,l}^k \big\rangle\big| \leq \sqrt{d}\kappa_l \leq 0.01 \cdot 2^{-4l}$$

where the first inequality is due to Lemma E.1, and the second inequality is valid due to $\kappa_l = 0.01 \cdot 2^{-4l}$. Similarly, we have

$$\big|\|\boldsymbol{\phi}(s,a)\|_{(\mathbf{U}_{h,l}^k)^{-1}} - \|\boldsymbol{\phi}(s,a)\|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}}\big| \leq \sqrt{d\kappa_l} \leq 0.1 \cdot 2^{-2l}.$$

□

E.2 Proof of Lemma D.2

Proof of Lemma D.2. First, both $l_h^\tau(s_h^\tau) = l$ and $f_h^\tau(s_h^\tau) = 1$ held for any $\tau \in \mathcal{C}_{h,l}^k$. This implies that the criteria for either Line 7 or Line 9 holds as $l = l_h^\tau(s_h^\tau)$. For τ that satisfies the first criterion, we have $l_h^\tau(s_h^\tau) > L_\tau$. Note that $L_\tau = \max\{\lceil \log_4(\tau/d) \rceil, 0\}$, so this only happens for $\tau < 4^l d$. For other τ that satisfies the second criterion, we have that

$$\|\boldsymbol{\phi}_h^\tau\|_{(\mathbf{U}_{h,l}^\tau)^{-1}} \geq \|\boldsymbol{\phi}_h^\tau\|_{\widetilde{\mathbf{U}}_{h,l}^{\tau,-1}} - \big|\|\boldsymbol{\phi}_h^\tau\|_{\widetilde{\mathbf{U}}_{h,l}^{\tau,-1}} - \|\boldsymbol{\phi}_h^\tau\|_{(\mathbf{U}_{h,l}^\tau)^{-1}}\big| \geq 2^{-l}\gamma_l^{-1} - 0.1 \cdot 2^{-l}\gamma_l^{-1} = 0.9 \cdot 2^{-l}\gamma_l^{-1},$$

where the first inequality holds due to the triangle inequality. In the second inequality, the first term $\|\boldsymbol{\phi}_h^\tau\|_{\widetilde{\mathbf{U}}_{h,l}^{\tau,-1}}$ is bounded by criterion in Line 9 while the second term $\big|\|\boldsymbol{\phi}_h^\tau\|_{\widetilde{\mathbf{U}}_{h,l}^{\tau,-1}} - \|\boldsymbol{\phi}_h^\tau\|_{(\mathbf{U}_{h,l}^\tau)^{-1}}\big|$ follows from Lemma D.1.

Arrange elements of $\mathcal{C}_{h,l}^k$ in ascending order as $\{\tau_i\}_i$. According to the above reasoning, the number of elements $\tau \in \mathcal{C}_{h,l}^k$ that $\|\boldsymbol{\phi}_h^\tau\|_{(\mathbf{U}_{h,l}^\tau)^{-1}} \geq 0.9 \cdot 2^{-l}\gamma_l^{-1}$ is at least $|\mathcal{C}_{h,l}^k| - 4^l d$. This gives

$$\sum_{i=1}^{|\mathcal{C}_{h,l}^k|} \min\{1, \|\boldsymbol{\phi}_h^\tau\|^2_{(\mathbf{U}_{h,l}^\tau)^{-1}}\} \geq (0.9 \cdot 2^{-l}\gamma_l^{-1})^2 \cdot (|\mathcal{C}_{h,l}^k| - 4^l d). \tag{E.1}$$

On the other hand, Lemma H.1 upper bounds the LHS of (E.1) by

$$\sum_{i=1}^{|\mathcal{C}_{h,l}^k|} \min\{1, \|\boldsymbol{\phi}_h^\tau\|^2_{(\mathbf{U}_{h,l}^\tau)^{-1}}\} \leq 2d \ln\big(1 + |\mathcal{C}_{h,l}^k|/(d\lambda)\big). \tag{E.2}$$

Combining (E.1) and (E.2) gives

$$0.81 \cdot 4^{-l}\gamma_l^{-2}(|\mathcal{C}_{h,l}^k| - 4^l d) \leq 2d \ln\big(1 + |\mathcal{C}_{h,l}^k|/(16d)\big). \tag{E.3}$$

From algebra analysis in Lemma G.1, a necessary condition for (E.3) is $|\mathcal{C}_{h,l}^k| \leq 16l \cdot 4^l \gamma_l^2 d$. □

E.3 Proof of Lemma D.3

We first present a claim from Vial et al. (2022) controlling the infinite norm of coefficient $\mathbf{w}$.

Lemma E.2 (Claim 10, Vial et al. 2022)**.** For any $(k, h, l) \in [K] \times [H] \times \mathbb{N}^+$, we have $\|\mathbf{w}_{h,l}^k\|_\infty \leq \|\mathbf{w}_{h,l}^k\|_2 \leq (2^l dH)^4$.

Proof of Lemma D.3. Denote $\mathcal{X}_\ell$ as the set of all $\widetilde{\mathbf{w}}_{h,\ell}^k$ and $\mathcal{Y}_\ell$ as the set of all $\widetilde{\mathbf{U}}_{h,\ell}^{k,-1}$. From the definition of $\mathcal{V}_{h,l}^k$, we have that $|\mathcal{V}_{h,l}^k| \leq \prod_{\ell=1}^l \big(|\mathcal{X}_\ell| \cdot |\mathcal{Y}_\ell|\big)$. From Lemma E.2, we have $\|\mathbf{w}_{h,\ell}^k\|_\infty \leq (2^\ell dH)^4$. Note that $\mathbf{w}_{h,\ell}^k \in \mathbb{R}^d$, we have the number of different $\widetilde{\mathbf{w}}_{h,\ell}^k$ controlled by

$$|\mathcal{X}_\ell| \leq (1 + 2 \cdot (2^\ell dH)^4/\kappa_\ell)^d \leq (2 \cdot (2^\ell dH)^4 \cdot 2^{6+4\ell} d)^d \leq 2^{(7+8\ell)d} d^{5d} H^{4d}.$$

In addition, we have $\|(\mathbf{U}_{h,l}^k)^{-1}\|_\infty \leq 1/\lambda = 1/16$. So we can bound the number of $\widetilde{\mathbf{U}}_{h,\ell}^{k,-1}$ by

$$|\mathcal{Y}_\ell| \leq (1 + 2 \cdot 1/(16\kappa_\ell))^{d^2} \leq (2 \cdot 2^{2+4\ell} d)^{d^2} \leq 2^{(3+4\ell)d^2} d^{d^2}.$$

As a result, we can conclude that

$$|\mathcal{V}_{h,l}^k| \leq \prod_{\ell=1}^l \big(|\mathcal{X}_\ell| \cdot |\mathcal{Y}_\ell|\big) \leq \prod_{\ell=1}^l \big(2^{(7+8\ell)d} d^{5d} H^{4d} \cdot 2^{(3+4\ell)d^2} d^{d^2}\big) \leq (2^{22} d^5 H^4)^{l^2 d^2}.$$

□

E.4 Proof of Lemma D.4

Proof of Lemma D.4. According to Proposition 3.2, there exists a parameter $\mathbf{w}_h$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, it holds that $\big|\langle \boldsymbol{\phi}(s, a), \mathbf{w}_h\rangle - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a)\big| \leq 2H\zeta$. Denoting $\eta_h^\tau = \langle \boldsymbol{\phi}_h^\tau, \mathbf{w}_h\rangle - [\mathbb{B}_h \widehat{V}_{h+1}^k](s_h^\tau, a_h^\tau)$ and $\varepsilon_h^\tau = \big(\widehat{V}_{h+1}^k(s_{h+1}^\tau) - [\mathbb{P}_h \widehat{V}_{h+1}^k](s_h^\tau, a_h^\tau)\big)$, we have

$$\begin{aligned}
\mathbf{U}_{h,l}^k(\mathbf{w}_{h,l}^k - \mathbf{w}_h) &= \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \Big(r_h^\tau + \widehat{V}_{h+1}^k(s_{h+1}^\tau)\Big) - \Big(\lambda \mathbf{I} + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau (\boldsymbol{\phi}_h^\tau)^\top\Big)\mathbf{w}_h \\
&= -\lambda \mathbf{w}_h + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \Big(r_h^\tau + \widehat{V}_{h+1}^k(s_{h+1}^\tau) - \langle \boldsymbol{\phi}_h^\tau, \mathbf{w}_h\rangle\Big) \\
&= -\lambda \mathbf{w}_h + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \Big(r_h^\tau + \widehat{V}_{h+1}^k(s_{h+1}^\tau) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s_h^\tau, a_h^\tau)\Big) + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \eta_h^\tau \\
&= -\lambda \mathbf{w}_h + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \varepsilon_h^\tau + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \eta_h^\tau,
\end{aligned} \tag{E.4}$$

https://doi.org/10.52202/079017-4154

where the first equality holds due to the definition of $\mathbf{U}_{h,l}^k, \mathbf{w}_{h,l}^k$, the second equality holds by rearranging the terms, the third equality holds according the definition of η_h^τ, and the last equality holds from the relationship that $[\mathbb{B}_h \widehat{V}_{h+1}^k](s_h^\tau, a_h^\tau) = r_h^\tau + [\mathbb{P}_h \widehat{V}_{h+1}^k](s_h^\tau, a_h^\tau)$. Therefore, for any vector $\boldsymbol{\phi} \in \mathbb{R}^d$, it holds that

$$\begin{aligned}
\big|\big\langle \boldsymbol{\phi}, \mathbf{w}_{h,l}^k - \mathbf{w}_h \big\rangle\big| &= \big|\boldsymbol{\phi}^\top \big(\mathbf{U}_{h,l}^k\big)^{-1} \mathbf{U}_{h,l}^k (\mathbf{w}_{h,l}^k - \mathbf{w}_h)\big| \\
&= \left|\boldsymbol{\phi}^\top \big(\mathbf{U}_{h,l}^k\big)^{-1} \cdot \left(- \lambda \mathbf{w}_h + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \varepsilon_h^\tau + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \eta_h^\tau \right)\right| \\
&\le \|\boldsymbol{\phi}\|_{(\mathbf{U}_{h,l}^k)^{-1}} \left\| - \lambda \mathbf{w}_h + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \varepsilon_h^\tau + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \eta_h^\tau \right\|_{(\mathbf{U}_{h,l}^k)^{-1}}, \qquad \text{(E.5)}
\end{aligned}$$

where the second equality follows (E.4) and the inequality holds from Cauchy–Schwarz inequality (i.e., $|\mathbf{x}^\top \mathbf{U} \mathbf{y}| \le \|\mathbf{x}\|_{\mathbf{U}} \|\mathbf{y}\|_{\mathbf{U}}$). From the triangle inequality, we have

$$\begin{aligned}
&\left\| - \lambda \mathbf{w}_h + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \varepsilon_h^\tau + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \eta_h^\tau \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} \\
&\le \lambda \|\mathbf{w}_h\|_{(\mathbf{U}_{h,l}^k)^{-1}} + \left\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \varepsilon_h^\tau \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} + \left\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \eta_h^\tau \right\|_{(\mathbf{U}_{h,l}^k)^{-1}}. \qquad \text{(E.6)}
\end{aligned}$$

There are three terms which we will bound respectively. For the first term, we have

$$\lambda \|\mathbf{w}_h\|_{(\mathbf{U}_{h,l}^k)^{-1}} \le 2\sqrt{d\lambda} H \le 0.1 \gamma_l, \qquad \text{(E.7)}$$

where the first inequality holds due to the fact that $\|\mathbf{w}_h\|_2 \le 2H\sqrt{d}$ as of Proposition 3.2 and the fact that $\mathbf{U}_{h,l}^k \succeq \lambda \mathbf{I}$. Under the good event $\mathcal{G}_1$ and Lemma C.5, the second term can be bounded by the following:

$$\left\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \varepsilon_h^\tau \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} \le 1.1 \gamma_l. \qquad \text{(E.8)}$$

And the last term can be bounded by:

$$\left\| \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \eta_h^\tau \right\|_{(\mathbf{U}_{h,l}^k)^{-1}} \le 2H\zeta \sqrt{|\mathcal{C}_{h,l}^k|} \le 2H\zeta \sqrt{16 l \cdot 4^l \gamma_l^2 d} = 8\sqrt{ld} H \cdot 2^l \gamma_l \zeta, \qquad \text{(E.9)}$$

where the first inequality is due to Lemma H.3, and the second inequality follows from Lemma D.2. Plugging (E.6), (E.7), (E.8), and (E.9) into (E.5) gives

$$\big|\big\langle \boldsymbol{\phi}, \mathbf{w}_{h,l}^k - \mathbf{w}_h \big\rangle\big| \le \big(1.2\gamma_l + 8\sqrt{ld} H \cdot 2^l \gamma_l \zeta\big) \|\boldsymbol{\phi}\|_{(\mathbf{U}_{h,l}^k)^{-1}}. \qquad \text{(E.10)}$$

So for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
&\big|Q_{h,l}^k(s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)\big| = \big|\big\langle \boldsymbol{\phi}(s,a), \widetilde{\mathbf{w}}_{h,l}^k \big\rangle - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)\big| \\
&\le \big|\big\langle \boldsymbol{\phi}(s,a), \widetilde{\mathbf{w}}_{h,l}^k - \mathbf{w}_{h,l}^k \big\rangle\big| + \big|\big\langle \boldsymbol{\phi}(s,a), \mathbf{w}_{h,l}^k - \mathbf{w}_h \big\rangle\big| + \big|\big\langle \boldsymbol{\phi}(s,a), \mathbf{w}_h \big\rangle - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)\big| \\
&\le 0.01 \cdot 2^{-4l} + \big(1.2 + 8\sqrt{ld} H \cdot 2^l \zeta\big) \gamma_l \|\boldsymbol{\phi}(s,a)\|_{(\mathbf{U}_{h,l}^k)^{-1}} + 2H\zeta. \qquad \text{(E.11)}
\end{aligned}$$

where the first inequality holds from the triangle inequality, and there are three terms in the second inequality which we will bound them respectively: the first term is given by Lemma D.1, the second term follows (E.10), and the third term holds from the definition of $\mathbf{w}_h$. □

E.5 Proof of Lemma D.5

Proof of Lemma D.5. We prove by doing case analysis. In case that action $a_l \in \mathcal{A}_{h,l+1}^k(s)$, we can assign $a_{l+1} = a_l \in \mathcal{A}_{h,l+1}^k(s)$ so that

$$[\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_{l+1}) = 0. \tag{E.12}$$

On the other hand, in the case that $a_l \notin \mathcal{A}_{h,l+1}^k(s)$, the action a_l is eliminated with $Q_{h,l}^k(s, a_l) < V_{h,l}^k(s) - 4 \cdot 2^{-l}$. Note in this case, there exists $a_{l+1} = \pi_{h,l}^k(s) \in \mathcal{A}_{h,l+1}^k(s)$ such that

$$Q_{h,l}^k(s, a_l) + 4 \cdot 2^{-l} < V_{h,l}^k(s) = Q_{h,l}^k(s, a_{l+1}). \tag{E.13}$$

According to Lemma C.6 and the condition that $l \le L_0$, we have that empirical state-value function $Q_{h,l}^k(s, \cdot)$ is a good estimation for $[\mathbb{B}_h \widehat{V}_{h+1}^k](s, \cdot)$ on actions $a_l, a_{l+1} \in \mathcal{A}_l^k(s)$ under event $\mathcal{G}_1$:

$$\big|[\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) - Q_{h,l}^k(s, a_l)\big| \le 2 \cdot 2^{-l} + \chi\sqrt{L_0}\zeta \tag{E.14}$$

$$\big|[\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_{l+1}) - Q_{h,l}^k(s, a_{l+1})\big| \le 2 \cdot 2^{-l} + \chi\sqrt{L_0}\zeta. \tag{E.15}$$

Moreover,

$$\begin{aligned}
&[\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_{l+1}) \\
&= \big([\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) - Q_{h,l}^k(s, a_l)\big) \\
&\quad + \big(Q_{h,l}^k(s, a_l) - Q_{h,l}^k(s, a_{l+1})\big) + \big(Q_{h,l}^k(s, a_{l+1}) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_{l+1})\big) \\
&\le 2 \cdot \big(2 \cdot 2^{-l} + \chi\sqrt{L_0}\zeta\big) - 4 \cdot 2^{-l} \\
&= 2\chi\sqrt{L_0}\zeta.
\end{aligned} \tag{E.16}$$

where the first inequality is derived from combining (E.13), (E.14), and (E.15). So from (E.12) and (E.16), we have that $[\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_{l+1}) \le 2\chi\sqrt{L_0}\zeta$ holds in both cases. □

E.6 Proof of Lemma D.6

Proof of Lemma D.6. We prove by induction on l. The induction basis holds at $l = 0$ by selecting $a_1 = \operatorname{argmax}_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a) \in \mathcal{A}$ which ensures $\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_1) = 0$. Additionally, if the induction hypothesis holds for $l - 1$, we have that

$$\begin{aligned}
&\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_{l+1}) \\
&= \big(\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l)\big) + \big([\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_{l+1})\big) \\
&\le 2(l-1)\chi\sqrt{L_0}\zeta + 2\chi\sqrt{L_0}\zeta \\
&= 2l \cdot \chi\sqrt{L_0}\zeta,
\end{aligned}$$

where the first inequality term is derived from combining induction hypothesis with Lemma D.5. We can then reach desired statement holds for all l in the range by induction. □

E.7 Proof of Lemma D.7

Proof of Lemma D.7. According to Lemma D.6, there exists some action $a_l \in \mathcal{A}_{h,l}^k(s)$ that

$$\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) \le 2(l-1)\chi\sqrt{L_0}\zeta. \tag{E.17}$$

Moreover, we have

$$[\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) - V_{h,l}^k(s) \le [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a_l) - Q_{h,l}^k(s, a_l) \le 2 \cdot 2^{-l} + \chi\sqrt{L_0}\zeta, \tag{E.18}$$

where the first inequality comes from the definition $V_{h,l}^k(s) = \max_{a \in \mathcal{A}_{h,l}^k} Q_{h,l}^k(s, a)$ and the second inequality holds according to Lemma C.6 with $l \le L_0$. Adding up (E.17) and (E.18) leads to

$$\max_{a \in \mathcal{A}} [\mathbb{B}_h \widehat{V}_{h+1}^k](s, a) - V_{h,l}^k \le 2 \cdot 2^{-l} + (2l-1)\chi\sqrt{L_0}\zeta.$$

This completes the proof. □

https://doi.org/10.52202/079017-4154

E.8 Proof of Lemma D.8

Proof of Lemma D.8. The statement holds by simply checking:

$$\begin{aligned}V_{h,l}^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) &\le V_{h,l}^k(s) - [\mathbb{B}_h\widehat{V}_{h+1}^k](s,\pi_{h,l}^k(s))\\ &= Q_{h,l}^k(s,\pi_{h,l}^k(s)) - [\mathbb{B}_h\widehat{V}_{h+1}^k](s,\pi_{h,l}^k(s))\\ &\le 2\cdot 2^{-l} + \chi\sqrt{L_0}\zeta,\end{aligned}$$

where the first inequality holds from $\max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) \ge [\mathbb{B}_h\widehat{V}_{h+1}^k](s,\pi_{h,l}^k(s))$, the equality is from the definition $V_{h,l}^k(s) = Q_{h,l}^k(s,\pi_{h,l}^k(s))$, and the last inequality holds according to Lemma C.6 with the condition $l \le L_0$. □

E.9 Proof of Lemma D.9

Proof of Lemma D.9. The statement holds by checking

$$\begin{aligned}&\min\left\{V_{h,l}^k(s) + 3\cdot 2^{-l}, \widehat{V}_{h,l-1}^k(s)\right\} - \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a)\\ &= \min_{\ell=1}^{l}\{V_{h,\ell}^k(s) + 3\cdot 2^{-\ell}\} - \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a)\\ &\ge \min_{\ell=1}^{l}\{3\cdot 2^{-\ell} - (2\cdot 2^{-l} + (2\ell-1)\chi\sqrt{L_0}\zeta)\}\\ &= 2^{-l} - (2l-1)\chi\sqrt{L_0}\zeta,\end{aligned}$$

where the first equality holds due to $\widehat{V}_{h,l}^k(s) = \min_{\ell=1}^{l}\{V_{h,\ell}^k(s) + 3\cdot 2^{-\ell}\}$, the inequality holds according to Lemma D.7, and the last equality holds since 2^{-l} decreases as l increases. □

E.10 Proof of Lemma D.10

Proof of Lemma D.10. The statement holds by checking

$$\begin{aligned}&\max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) - \max\left\{V_{h,l}^k(s) - 3\cdot 2^{-l}, \check{V}_{h,l-1}^k(s)\right\}\\ &= \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) - \max_{\ell=1}^{l}\{V_{h,\ell}^k(s) - 3\cdot 2^{-\ell}\}\\ &= \min_{\ell=1}^{l}\left\{\max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) - V_{h,\ell}^k(s) + 3\cdot 2^{-\ell}\right\}\\ &\ge \min_{\ell=1}^{l}\{-(2\cdot 2^{-l} + \chi\sqrt{L_0}\zeta) + 3\cdot 2^{-\ell}\}\\ &= 2^{-l} - \chi\sqrt{L_0}\zeta,\end{aligned}$$

where the first equality holds due to the design of Algorithm 2 such that $\check{V}_{h,l}^k(s) = \max_{\ell=1}^{l}\{V_{h,\ell}^k(s) - 3\cdot 2^{-\ell}\}$, the inequality holds according to Lemma D.8, and the last equality holds since 2^{-l} decreases as l increases. □

E.11 Proof of Lemma D.11

We prove Lemma D.11 in this subsection. The first lemma which we introduce establishes an upper bound on the underestimation of the state value function $\widehat{V}_h^k$ for every action and every state through a categorised discussion based on whether Algorithm 2 reaches phase L_ε for state s. Specifically, if the process does not reach phase L_ε, we can substantiate the statement by applying Lemma D.9 to phase $l_h^k(s) - 1$. Conversely, if the process reaches phase L_ε, the statement can be proven by applying Lemma D.7 to phase L_ε.

Lemma E.3. Under event $\mathcal{G}_1$ and for all $\varepsilon > 0$ that $\mathcal{G}_\varepsilon$ is satisfied, for any $(k,h,s) \in [K]\times[H]\times\mathcal{S}$,

$$\max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) - \widehat{V}_h^k(s) \le 0.07\varepsilon/H.$$

Now we are ready to prove Lemma D.11 by induction.

Proof of Lemma D.11. We prove by induction on stage $h \in [H]$. It is sufficient to show for any $h \in [H], s \in \mathcal{S}$,

$$V_h^*(s) - \widehat{V}_h^k(s) \leq 0.07\varepsilon \cdot (H + 1 - h)/H. \tag{E.19}$$

We use induction on h from $H+1$ to 1 to prove the statement. The induction basis holds from the definition that $V_{H+1}^*(s) = \widehat{V}_{H+1}^k(s) = 0$. Assume the induction hypothesis (E.19) holds for $h+1$, we have

$$\begin{aligned}\max_{a\in\mathcal{A}}[\mathbb{B}_h V_{h+1}^*](s,a) - \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) &\leq \max_{a\in\mathcal{A}}[\mathbb{B}_h(V_{h+1}^* - \widehat{V}_{h+1}^k)](s,a) \\ &\leq \max_{s'\in\mathcal{S}}\big(V_{h+1}^*(s') - \widehat{V}_{h+1}^k(s')\big) \\ &\leq 0.07\varepsilon \cdot (H-h)/H. \end{aligned} \tag{E.20}$$

So for level h, it holds that

$$\begin{aligned}&V_h^*(s) - \widehat{V}_h^k(s) \\ &= \big(\max_{a\in\mathcal{A}}[\mathbb{B}_h V_{h+1}^*](s,a) - \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)\big) + \big(\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - \widehat{V}_h^k(s)\big) \\ &\leq 0.07\varepsilon \cdot (H-h)/H + 0.07\varepsilon/H \leq 0.07\varepsilon \cdot (H+1-h)/H,\end{aligned}$$

where the first inequality holds by combining (E.20) with Lemma E.3. This proves the induction statement (E.19) for h, which leads to the desired statement. ☐

E.12 Proof of Lemma D.12

We prove Lemma D.12 in this subsection, the first lemma we use establishes an upper bound on the overestimation of the state value function $\widehat{V}_h^k$ for the executed policy $\pi_h^k(s)$ across all states.

Lemma E.4. Under event $\mathcal{G}_1$ and for all $\varepsilon > 0$ that $\mathcal{G}_\varepsilon$ is satisfied, for any $(k,h,s) \in [K] \times [H] \times \mathcal{S}$,

$$\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, \pi_h^k(s)) \leq 16 \cdot 2^{-l_h^k(s)} + 0.10\varepsilon/H.$$

Then the following lemma establishes an upper bound on the decision error induced by the arm-elimination process with respect to the state-action value function given by the ground-truth transform.

Lemma E.5. Under event $\mathcal{G}_1$ and for all $\varepsilon > 0$ that $\mathcal{G}_\varepsilon$ is satisfied, for any $(k,h,s) \in [K] \times [H] \times \mathcal{S}$,

$$\widehat{V}_h^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) \leq 10 \cdot 2^{-l_h^k(s)} + 0.06\varepsilon/H.$$

Proof of Lemma D.12. We can directly reach the desired result by taking summation on Lemma E.4 and Lemma E.5:

$$\begin{aligned}&\widehat{V}_h^k(s) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s, \pi_h^k(s)) \\ &\leq \big(\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,\pi_h^k(s))\big) + \big(\widehat{V}_h^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a)\big) \\ &\leq \big(16 \cdot 2^{-l_h^k(s)} + 0.10\varepsilon/H\big) + \big(10 \cdot 2^{-l_h^k(s)} + 0.06\varepsilon/H\big) \\ &= 26 \cdot 2^{-l_h^k(s)} + 0.16\varepsilon/H.\end{aligned}$$

☐

E.13 Proof of Lemma D.14

We can prove the statement by applying a union bound to the concentration event, as given by the Azuma-Hoeffding inequality.

https://doi.org/10.52202/079017-4154

Proof of Lemma D.14. Consider some fixed $h \in [H]$ and $\varepsilon = 2^{-l} > 0$. List the episodes index k such that $V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) > \varepsilon$ holds in ascending order as $\{\tau_i\}_i$. Recall that

$$\eta_h^{\tau_i} = [\mathbb{P}_h(\widehat{V}_{h+1}^{\tau_i} - V_{h+1}^{\pi^{\tau_i}})](s_h^{\tau_i}, \pi_h^{\tau_i}(s_h^{\tau_i})) - \big(\widehat{V}_{h+1}^{\tau_i}(s_{h+1}^{\tau_i}) - V_{h+1}^{\pi^{\tau_i}}(s_{h+1}^{\tau_i})\big).$$

Since the environment sample $s_{h'+1}^{\tau_i}$ according to $\mathbb{P}_{h'}(\cdot|s_{h'}^{\tau_i}, a_{h'}^{\tau_i})$, we have $\eta_{h'}^{\tau_i}$ is $\mathcal{F}_{h'+1}^{\tau_i}$-measurable with $\mathbb{E}\left[\eta_{h'}^{\tau_i}\middle|\mathcal{F}_{h'}^{\tau_i}\right] = 0$. Since both $0 \le \widehat{V}_{h'+1}^{\tau_i}(s_{h'+1}^{\tau_i}) \le H$ and $0 \le V_{h'+1}^{\pi^{\tau_i}}(s_{h'+1}^{\tau_i}) \le H$ hold, we have $|\eta_{h'}^{\tau_i}| \le 2H$. According to Lemma H.4 over filtration

$$\mathcal{F}_h^{\tau_1} \subseteq \mathcal{F}_{h+1}^{\tau_1} \subseteq \cdots \subseteq \mathcal{F}_H^{\tau_1} \subseteq \mathcal{F}_h^{\tau_2} \subseteq \mathcal{F}_{h+1}^{\tau_2} \subseteq \cdots \subseteq \mathcal{F}_H^{\tau_2} \subseteq \cdots \subseteq \mathcal{F}_{h'}^{\tau_i} \subseteq \cdots$$

for some fixed $S = |\mathcal{K}_h^\varepsilon|$, the good event that

$$\sum_{i=1}^{|\mathcal{K}_h^\varepsilon|} \sum_{h'=h}^{H} \eta_{h'}^{\tau_i} \le 2H\sqrt{2HS\log(4HS^2l^2/\delta)} = 4\sqrt{H^3|\mathcal{K}_h^\varepsilon|\log(4H|\mathcal{K}_h^\varepsilon|\log(\varepsilon^{-1})/\delta)}$$

happens with probability at least $1 - \delta/(4HS^2l^2)$. By the union bound statement over all $(h, S, l) \in [H] \times [K] \times \mathbb{N}^+$, we have the bad event happens with probability at most

$$\Pr[\mathcal{G}_2^\complement] \le \sum_{h=1}^{H}\sum_{S=1}^{K}\sum_{l=1}^{\infty} \Pr[\mathcal{B}_2(h, 2^{-l})] \le \sum_{h=1}^{H}\sum_{S=1}^{K}\sum_{l=1}^{\infty} \frac{\delta}{4HS^2l^2} \le \delta,$$

where the last inequality holds from $\sum_{n\ge1} n^{-2} = \pi^2/6$, which reach the desired statement. □

E.14 Proof of Lemma D.15

We first provide the following instantaneous regret upper bound by combining Lemma D.11 and Lemma D.12.

Lemma E.6. Under event $\mathcal{G}_1$ and for all $\varepsilon > 0$ that $\mathcal{G}_\varepsilon$ is satisfied, for any $(k, h) \in [K] \times [H]$,

$$V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) \le 0.23\varepsilon + 26\sum_{h'=h}^{H} 2^{-l_{h'}^k(s_{h'}^k)} + \sum_{h'=h}^{H} \eta_{h'}^k,$$

where $\eta_h^k = [\mathbb{P}_h(\widehat{V}_{h+1}^k - V_{h+1}^{\pi^k})](s_h^k, \pi_h^k(s_h^k)) - \big(\widehat{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big)$ is a $\mathcal{F}_{h+1}^k$-measurable random variable that $\mathbb{E}[\eta_h^k|\mathcal{F}_h^k] = 0$ and $|\eta_h^k| \le H$.

Together with Lemma C.9 and the definition of $\mathcal{G}_2$, we can provide an upper bound for arbitrary subsets.

Proof of Lemma D.15. Taking summation on result given by Lemma E.6 to all $k \in \mathcal{K}$ gives

$$\sum_{k\in\mathcal{K}} \big(V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k)\big) \le 0.23|\mathcal{K}|\varepsilon + 26\sum_{k\in\mathcal{K}}\sum_{h'=h}^{H} 2^{-l_{h'}^k(s_{h'}^k)} + \sum_{k\in\mathcal{K}}\sum_{h'=h}^{H} \eta_{h'}^k. \tag{E.21}$$

We can bound the second term according to Lemma C.9,

$$26\sum_{k\in\mathcal{K}}\sum_{h'=h}^{H} 2^{-l_{h'}^k(s_{h'}^k)} \le 0.26|\mathcal{K}|\varepsilon + 2^{17}L_\varepsilon dH^2\gamma_{L_\varepsilon}^2\varepsilon^{-1}. \tag{E.22}$$

Under event $\mathcal{G}_2$, the third term satisfies that

$$\sum_{k\in\mathcal{K}}\sum_{h'=h}^{H} \eta_{h'}^k \le 4\sqrt{H^3|\mathcal{K}|\log(4H|\mathcal{K}|\log(\varepsilon^{-1})/\delta)}. \tag{E.23}$$

Plugging (E.22) and (E.23) into (E.21) gives

$$\sum_{k\in\mathcal{K}} \big(V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k)\big) \le 0.49|\mathcal{K}|\varepsilon + 2^{17}L_\varepsilon dH^2\gamma_{L_\varepsilon}^2\varepsilon^{-1} + 4\sqrt{H^3|\mathcal{K}|\log(4H|\mathcal{K}|\log(\varepsilon^{-1})/\delta)}. \tag{E.24}$$

□

F Proof of Lemmas in Appendix E

F.1 Proof of Lemma E.3

Proof of Lemma E.3. We start the proof by discussing different cases. First, if $l_h^k(s) \le L_\varepsilon$, we have $l_h^k(s) - 1 \le \min\{L_\varepsilon, l_h^k(s) - 1\}$, according to the definition of $\widehat{V}_{h,l}^k(s)$,

$$\begin{aligned}\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - \widehat{V}_h^k(s) &= \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - \widehat{V}_{h,l_h^k(s)-1}^k(s)\\ &\le -2^{-(l_h^k(s)-1)} + 2(l_h^k(s)-1)\chi\sqrt{L_\varepsilon}\zeta\\ &\le 0 + 2\chi L_\varepsilon^{1.5}\zeta\\ &\le 0.02\varepsilon/H, \end{aligned} \tag{F.1}$$

where the first inequality holds from Lemma D.9, and the last inequality holds due to $\chi L_\varepsilon^{1.5}\zeta \le 2^{-L_\varepsilon} \le 0.01\varepsilon/H$ given by $\mathcal{G}_\varepsilon$.

On the other hand, when $l_h^k(s) > L_\varepsilon$, we have $L_\varepsilon \le \min\{L_\varepsilon, l_h^k(s) - 1\}$ and thus

$$\widehat{V}_h^k(s) \ge \check{V}_{h,L_\varepsilon}^k(s) \ge V_{h,L_\varepsilon}^k(s) - 3\cdot 2^{-L_\varepsilon} \tag{F.2}$$

where the first inequality is due to Lemma C.2 and the second inequality holds due to the definition of $\check{V}_{h,L_\varepsilon}^k(s)$. Therefore, $L_\varepsilon \le \min\{L_\varepsilon, l_h^k(s) - 1\}$ yields

$$\begin{aligned}\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - \widehat{V}_h^k(s) &\le \max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - V_{h,L_\varepsilon}^k(s) + 3\cdot 2^{-L_\varepsilon}\\ &\le 5\cdot 2^{-L_\varepsilon} + (2L_\varepsilon - 1)\chi\sqrt{L_\varepsilon}\zeta\\ &\le 0.05\varepsilon/H + 0.02\varepsilon/H = 0.07\varepsilon/H, \end{aligned} \tag{F.3}$$

where the first inequality is given by (F.2), the second inequality is given by Lemma D.7, and the last inequality holds from $\chi L_\varepsilon^{1.5}\zeta \le 2^{-L_\varepsilon} \le 0.01\varepsilon/H$ given by $\mathcal{G}_\varepsilon$. So considering both (F.1) and (F.3), we have the first statement

$$\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - \widehat{V}_h^k(s) \le 0.07\varepsilon/H$$

always holds under event $\mathcal{G}_1$. □

F.2 Proof of Lemma E.4

We prove Lemma E.4 by applying Lemma C.7 on phase $\min\{L_\varepsilon, l_h^k(s) - 1\}$, in this subsection.

Proof of Lemma E.4. Note we have $\pi_{h,l_h^k(s)-1}^k(s) \in \mathcal{A}_{h,l_h^k(s)}^k(s)$ according to the definition of $\mathcal{A}_{h,l+1}^k(s)$. This implies $\pi_h^k(s) \in \mathcal{A}_{h,l_h^k(s)}^k(s)$ during the elimination process.

If $l_h^k(s) \le L_\varepsilon$, we have $l_h^k(s) - 1 \le \min\{L_\varepsilon, l_h^k(s) - 1\}$. Thus,

$$\begin{aligned}\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,\pi_h^k(s)) &\le 8\cdot 2^{-(l_h^k(s)-1)} + 2l_h^k(s)\cdot\chi\sqrt{L_\varepsilon}\zeta\\ &\le 16\cdot 2^{-l_h^k(s)} + 2\chi L_\varepsilon^{1.5}\zeta\\ &\le 16\cdot 2^{-l_h^k(s)} + 0.02\varepsilon/H, \end{aligned} \tag{F.4}$$

where the first inequality follows from Lemma C.7 with $\pi_h^k(s) \in \mathcal{A}_{h,l_h^k(s)}^k(s)$ and the last inequality holds due to $\chi L_\varepsilon^{1.5}\zeta \le 0.01\varepsilon/H$ given by $\mathcal{G}_\varepsilon$.

Otherwise, we have $L_\varepsilon \le \min\{L_\varepsilon, l_h^k(s) - 1\}$. In this case, we have

$$\begin{aligned}\max_{a\in\mathcal{A}}[\mathbb{B}_h \widehat{V}_{h+1}^k](s,a) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s,\pi_h^k(s)) &\le 8\cdot 2^{-L_\varepsilon} + 2\chi L_\varepsilon^{1.5}\zeta\\ &\le 0.08\varepsilon/H + 0.02\varepsilon/H = 0.10\varepsilon/H, \end{aligned} \tag{F.5}$$

https://doi.org/10.52202/079017-4154

where the first inequality follows from Lemma C.7 with $\pi_h^k(s) \in \mathcal{A}_{h,l_h^k(s)}^k(s) \subseteq \mathcal{A}_{h,L_\varepsilon}^k(s)$ according to the elimination routine and the final inequality holds due to $\chi L_\varepsilon^{1.5}\zeta \le 2^{-L_\varepsilon} \le 0.01\varepsilon/H$ given by $\mathcal{G}_\varepsilon$. So by combining (F.4) and (F.5), we have the desired statement that

$$\max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) - [\mathbb{B}_h\widehat{V}_{h+1}^k](s,\pi_h^k(s)) \le 16\cdot 2^{-l_h^k(s)} + 0.10\varepsilon/H.$$

□

F.3 Proof of Lemma E.5

We prove Lemma E.5 in this section by applying Lemma D.8 on phase $\min\{L_\varepsilon, l_h^k(s)-1\}$.

Proof of Lemma E.5. If $l_h^k(s) \le L_\varepsilon$, we have $l_h^k(s) - 1 \le \min\{L_\varepsilon, l_h^k(s)-1\}$. Firstly, we have

$$\widehat{V}_h^k(s) \le \widehat{V}_{h,l_h^k(s)-1}^k(s) \le V_{h,l_h^k(s)-1}^k(s) + 3\cdot 2^{-(l_h^k(s)-1)}. \tag{F.6}$$

where the first inequality is given by Lemma C.2 and the second inequality follows from the definition of $\widehat{V}_{h,l_h^k(s)-1}^k(s)$. This leads to

$$\begin{aligned}\widehat{V}_h^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) &\le \big(\widehat{V}_h^k(s) - V_{h,l_h^k(s)-1}^k(s)\big) + \big(V_{h,l_h^k(s)-1}^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a)\big)\\ &\le 3\cdot 2^{-(l_h^k(s)-1)} + 2\cdot 2^{-(l_h^k(s)-1)} + \chi\sqrt{L_\varepsilon}\zeta\\ &\le 10\cdot 2^{-l_h^k(s)} + 0.01\varepsilon/H,\end{aligned} \tag{F.7}$$

where in the second inequality, the first term is given by (F.6) and the second term holds according to Lemma D.8, and the third inequality holds from $\chi\sqrt{L_\varepsilon}\zeta \le 0.01\varepsilon/H$ given by $\mathcal{G}_\varepsilon$.

Otherwise, we have $L_\varepsilon \le \min\{L_\varepsilon, l_h^k(s)-1\}$, this leads to

$$\begin{aligned}\widehat{V}_h^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) &\le \big(\widehat{V}_h^k(s) - V_{h,L_\varepsilon}^k(s)\big) + \big(V_{h,L_\varepsilon}^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a)\big)\\ &\le 3\cdot 2^{-L_\varepsilon} + 2\cdot 2^{-L_\varepsilon} + \chi\sqrt{L_\varepsilon}\zeta\\ &\le 0.03\varepsilon/H + 0.02\varepsilon/H + 0.01\varepsilon/H = 0.06\varepsilon/H,\end{aligned} \tag{F.8}$$

where in the second inequality, the first term is given by the definition of $\widehat{V}_h^k(s)$ and the second term holds according to Lemma D.8, and the third inequality holds from $\chi L_\varepsilon^{1.5}\zeta \le 2^{-L_\varepsilon} \le 0.01\varepsilon/H$ given by $\mathcal{G}_\varepsilon$. Combining (F.7) and (F.8) gives the desired statement

$$\widehat{V}_h^k(s) - \max_{a\in\mathcal{A}}[\mathbb{B}_h\widehat{V}_{h+1}^k](s,a) \le 10\cdot 2^{-l_h^k(s)} + 0.06\varepsilon/H.$$

□

F.4 Proof of Lemma E.6

Proof of Lemma E.6. According to the definition in which $V_h^{\pi^k}(s_h^k) = [\mathbb{B}_h V_{h+1}^{\pi^k}](s_h^k, \pi_h^k(s_h^k))$ and $\eta_h^k + [\mathbb{P}_h(\widehat{V}_{h+1}^k - V_{h+1}^{\pi^k})](s_h^k, \pi_h^k(s_h^k)) - \big(\widehat{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big)$. We can write

$$\widehat{V}_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) = \big(\widehat{V}_h^k(s_h^k) - [\mathbb{B}_h\widehat{V}_{h+1}^k](s_h^k,\pi_h^k(s_h^k))\big) + \eta_h^k + \big(\widehat{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big).$$

By a telescoping statement from h to H with the final terminal value $\widehat{V}_{H+1}^k(\cdot) = V_{H+1}^{\pi^k}(\cdot) = 0$, we reach

$$\widehat{V}_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) = \sum_{h'=h}^{H}\big(\widehat{V}_h^k(s_h^k) - [\mathbb{B}_h\widehat{V}_{h+1}^k](s_h^k,\pi_h^k(s_h^k))\big) + \sum_{h'=h}^{H}\eta_{h'}^k. \tag{F.9}$$

As a result, we can bound the desired term by

$$
\begin{aligned}
V_h^*(s_h^k) - V_h^{\pi^k}(s_h^k) &\le \widehat{V}_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) + 0.07\varepsilon \\
&= \sum_{h'=h}^{H} \big(\widehat{V}_h^k(s_h^k) - [\mathbb{B}_h \widehat{V}_{h+1}^k](s_h^k, \pi_h^k(s_h^k))\big) + \sum_{h'=h}^{H} \eta_{h'}^k + 0.07\varepsilon \\
&\le \sum_{h'=h}^{H} \big(26 \cdot 2^{-l_{h'}^k(s_{h'}^k)} + 0.16\varepsilon/H\big) + \sum_{h'=h}^{H} \eta_{h'}^k + 0.07\varepsilon \\
&= 0.23\varepsilon + 26 \sum_{h'=h}^{H} 2^{-l_{h'}^k(s_{h'}^k)} + \sum_{h'=h}^{H} \eta_{h'}^k.
\end{aligned}
$$

where the first inequality is given by Lemma D.11, the first equality is given by (F.9), and the final inequality is given by Lemma D.12. □

G Technical Numerical Lemmas

Lemma G.1. If $|\mathcal{C}_{h,l}^k| \le 4^l d + 2.5 \cdot 4^l \gamma_l^2 d \ln\big(1 + |\mathcal{C}_{h,l}^k|/(16d)\big)$, then $|\mathcal{C}_{h,l}^k| \le 16l \cdot 4^l \gamma_l^2 d$.

Proof. Denote $c = |\mathcal{C}_{h,l}^k|/(l \cdot 4^l \gamma_l^2 d)$. We have that

$$
cl \cdot 4^l \gamma_l^2 d \le 4^l d + 2.5 \cdot 4^l \gamma_l^2 d \ln(1 + cl \cdot 4^l \gamma_l^2/16).
$$

Dividing both sides by $4^l \gamma_l^2 d$, we have that

$$
\begin{aligned}
cl &\le 1/\gamma_l^2 + 2.5 \ln(1 + cl \cdot 4^l \gamma_l^2/16) \\
&\le 1/\gamma_l^2 + 2.5 \ln(4c \cdot 5^l \gamma_l^2/16) \le 1/\gamma_l^2 + 4.1l + 2.5\ln(c).
\end{aligned}
$$

Since $l \ge 1$ and $\gamma_l \ge 1$, we can further conclude that

$$
c \le 5.1 + 2.5\ln(c) \le 5.1 + 2.5(1 + c/6).
$$

The necessary condition for the above inequality is $c \le 16$, which proves the desired statement. □

Lemma G.2. For any $l \ge 1$, $\gamma_{l+1}/\gamma_l \le 1.4$.

Proof. Firstly, we have that

$$
\frac{l + 22 + \log(l+1)}{l + 20 + \log(l)} \le \frac{l + 22 + 0.2l + 2}{l + 20} = 1.2, \tag{G.1}
$$

where the first inequality holds due to $\log(x+1) \le 0.2x + 2$. In addition, we have

$$
\frac{4 + \log(l+1)}{4 + \log(l)} \le \frac{4 + \log(l) + 1}{4 + \log(l)} \le 1.25, \tag{G.2}
$$

where the first inequality holds due to $\log(x+1) \le \log(x) + 1$. As a result, we can reach the desired statement according to

$$
\begin{aligned}
\frac{\gamma_{l+1}}{\gamma_l} &= \frac{5(l + 1 + \lceil 20 + \log((l+1)d) \rceil) dH \sqrt{\log(16(l+1)dH/\delta)}}{5(l + \lceil 20 + \log(ld) \rceil) dH \sqrt{\log(16ldH/\delta)}} \\
&\le \frac{l + 22 + \log(l+1) + \log(d)}{l + 20 + \log(l) + \log(d)} \cdot \sqrt{\frac{\log(l+1) + \log(16dH/\delta)}{\log(l) + \log(16dH/\delta)}} \\
&\le \frac{l + 22 + \log(l+1)}{l + 20 + \log(l)} \cdot \sqrt{\frac{\log(l+1)}{\log(l)}} \\
&\le 1.2\sqrt{1.25} \\
&\le 1.4,
\end{aligned}
$$

where the third inequality holds from plugging both (G.1) and (G.2). □

https://doi.org/10.52202/079017-4154

Lemma G.3.

$$\sqrt{2d\ln(1+l\cdot 4^l\gamma_l^2)+2\ln(l^2H(2^{22}d^6H^4)^{l_+^2d^2}/\delta)}\le\gamma_{l,l_+}$$

Proof. By calculation, we have that

$$\begin{aligned}
&H\sqrt{2d\ln(1+l\cdot 4^l\gamma_l^2)+2\ln(l^2H(2^{22}d^6H^4)^{l_+^2d^2}/\delta)}\\
&\le H\sqrt{2d\ln(1+l\cdot 4^l\cdot 1.4^{2l}\gamma_1^2)}+H\sqrt{12l_+^2d^2\ln(2^4ldH/\delta)}\\
&\le l_+dH\sqrt{2\ln(2^4ldH/\delta)}+l_+dH\sqrt{12\ln(2^4ldH/\delta)}\\
&\le 5l_+dH\sqrt{\log(2^4\gamma_{l_+}ldH/\delta)}\\
&=\gamma_{l,l_+}.
\end{aligned}$$

□

Lemma G.4. If some constant $c_1,c_2>0$ that

$$|\mathcal{K}_h^\varepsilon|<c_1L_\varepsilon(L_\varepsilon+\log(dH))^2d^3H^4\varepsilon^{-2}\log(L_\varepsilon d/\delta)+\varepsilon^{-1}\sqrt{c_2H^3|\mathcal{K}_h^\varepsilon|\log(H|\mathcal{K}_h^\varepsilon|\log(\varepsilon^{-1})/\delta)}.$$

Then, there exists $c_3>0$ such that

$$|\mathcal{K}_h^\varepsilon|<c_3L_\varepsilon(L_\varepsilon+\log(dH))^2d^3H^4\varepsilon^{-2}\log(L_\varepsilon d)\log(\delta^{-1})\iota,$$

where ι is a polynomial of $\log\log(L_\varepsilon dH\delta^{-1})$.

Proof. Let $x=|\mathcal{K}_h^\varepsilon|/\log(|\mathcal{K}_h^\varepsilon|)$. We have that

$$x<c_1L_\varepsilon(L_\varepsilon+\log(dH))^2d^3H^4\varepsilon^{-2}\log(L_\varepsilon d/\delta)+\varepsilon^{-1}\sqrt{c_2H^3x\log(H\log(\varepsilon^{-1})/\delta)}.$$

Since $x<a+\sqrt{bx}$ implies $x<2a+2b$, so the above inequality implies

$$x<2c_1L_\varepsilon(L_\varepsilon+\log(dH))^2d^3H^4\varepsilon^{-2}\log(L_\varepsilon d/\delta)+2c_2H^3\varepsilon^{-2}\log(H\log(\varepsilon^{-1})/\delta).$$

Moreover, since $y/\log(y)<a$ implies $y<2a\log a$, we can conclude that there exists $c_3>0$ that

$$|\mathcal{K}_h^\varepsilon|<c_3L_\varepsilon(L_\varepsilon+\log(dH))^2d^3H^4\varepsilon^{-2}\log(L_\varepsilon d)\log(\delta^{-1})\iota,$$

where ι is a polynomial of $\log\log(L_\varepsilon dH\varepsilon^{-1}\delta^{-1})$. □

H Auxiliary Lemmas

This section provides some auxiliary concentration lemmas frequently used in the proof.

Lemma H.1 (Lemma 11, Abbasi-Yadkori et al. (2011))**.** Let $\{\boldsymbol{\phi}^k\}_{k=1}^\infty$ be any bounded sequence such that $\boldsymbol{\phi}^k\in\mathbb{R}^d$ and $\|\boldsymbol{\phi}^k\|_2\le B$ for some constant $B>0$. For $k\ge 1$, let $\mathbf{U}^k=\lambda\mathbf{I}+\sum_{\tau=1}^{k-1}\boldsymbol{\phi}^\tau(\boldsymbol{\phi}^\tau)^\top$. Let $\lambda>0$, then for all $k\in[K]$, we have that

$$\sum_{\tau=1}^{k}\min\left\{1,\|\boldsymbol{\phi}^\tau\|_{(\mathbf{U}^\tau)^{-1}}^2\right\}\le 2d\ln\left(1+kB^2/(d\lambda)\right).$$

Lemma H.2 (Self-Normalized Martingale, Abbasi-Yadkori et al. (2011))**.** Let $\{\mathcal{F}^k\}_{k=1}^\infty$ be a filtration, and $\{\boldsymbol{\phi}^k,\eta^k\}_{k=1}^\infty$ be a stochastic process where $\boldsymbol{\phi}^k\in\mathbb{R}^d$ is $\mathcal{G}^k$-measurable and η^k is $\mathcal{F}^{k+1}$-measurable such that

$$|\mathbb{E}[\eta^k|\mathcal{F}^k]|=0,|\eta^k|\le R,\|\boldsymbol{\phi}^k\|_2\le B$$

for some constant $B,R>0$. Let $\lambda>0$. For $k\ge1$, let $\mathbf{U}^k=\lambda\mathbf{I}+\sum_{\tau=1}^{k-1}\boldsymbol{\phi}^\tau(\boldsymbol{\phi}^\tau)^\top$. Then for any $\delta\in(0,1)$, with probability at least $1-\delta$, for all $k\ge1$, we have that

$$\left\|\sum_{\tau=1}^{k-1}\eta^\tau\boldsymbol{\phi}^\tau\right\|_{(\mathbf{U}^k)^{-1}}\le R\sqrt{2d\ln\left(1+kB^2/(d\lambda)\right)+2\ln\delta^{-1}}.$$

Lemma H.3 (Lemma 8, Zanette et al. (2020b)). Let $\{\boldsymbol{\phi}^k, \eta^k\}_{k=1}^{\infty}$ be any bounded sequence satisfying $\boldsymbol{\phi}^k \in \mathbb{R}^d$ and $|\eta^k| \le \zeta$ for some constant $\zeta > 0$. For $k \ge 1$, let $\mathbf{U}^k = \lambda \mathbf{I} + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}^\tau (\boldsymbol{\phi}^\tau)^\top$. Then, for all $k \ge 1$, we have that

$$\left\| \sum_{\tau=1}^{k-1} \eta^\tau \boldsymbol{\phi}^\tau \right\|_{(\mathbf{U}^k)^{-1}} \le \zeta\sqrt{k}.$$

Lemma H.4 (Azuma–Hoeffding inequality, Hoeffding (1963)). Let $\{\eta^k\}_{k=1}^K$ be a martingale difference sequence with respect to a filtration $\{\mathcal{F}^k\}_{k=1}^K$ satisfying $|\eta^k| \le M$ for some constant $M > 0$ and η^k is $\mathcal{F}^{k+1}$-measurable with $|\mathbb{E}[\eta^k|\mathcal{F}^k]| = 0$. Then for some fixed $k \in [K]$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

$$\sum_{\tau=1}^{k} \eta^\tau \le M\sqrt{2k \ln \delta^{-1}}.$$

Lemma H.5 (Freedman inequality, Cesa-Bianchi and Lugosi (2006)). Let $\{\eta^k\}_{k=1}^K$ be a martingale difference sequence with respect to a filtration $\{\mathcal{F}^k\}_{k=1}^K$ satisfying $|\eta^k| \le M$ for some constant $M > 0$ and η^k is $\mathcal{F}^{k+1}$-measurable with $|\mathbb{E}[\eta^k|\mathcal{F}^k]| = 0$. Then for some fixed $k \in [K]$, $a > 0$ and $v > 0$, we have

$$\Pr\Big(\sum_{\tau=1}^{k} \eta^\tau \ge a, \sum_{\tau=1}^{k} \mathrm{Var}[\eta^\tau|\mathcal{F}^\tau] \le v\Big) \le \exp\Big(\frac{-a^2}{2v + 2aM/3}\Big).$$

I Numerical Simulation

We added experiments on synthetic datasets to verify the performance of the algorithm and the contribution of each component. Specifically, we consider a linear MDP with $S = 4$, $A = 5$, $H = 2$, and $d = 8$. Each element in the feature vector $\boldsymbol{\phi}(s, a)$ and $\boldsymbol{\mu}(s')$ is generated by a uniform distribution $U(0, 1)$. Subsequently, $\boldsymbol{\phi}$ is normalized to ensure that $\mathbb{P}(s'|s, a)$ is a probability measure, i.e., $\boldsymbol{\phi}(s, a) = \boldsymbol{\phi}(s, a)/\sum_{s'} \boldsymbol{\phi}^\top(s, a)\boldsymbol{\mu}(s')$. The reward is defined by $r(s, a) = \boldsymbol{\phi}^\top(s, a)\boldsymbol{\theta}$, where $\boldsymbol{\theta} \sim N(0, I_d)$. The model misspecification is also added to the transition $\mathbb{P}$ and reward function r. For a given misspecification ζ, the ground truth reward function is defined by $r(s, a) = \phi^\top(s, a)\theta + Z(s, a)$, where $Z(s, a) \sim U(-\zeta, \zeta)$. When adding the model misspecification to the transition kernel, we first random sample a subset $\mathcal{S}_+ \subset \mathcal{S}$ such that $|\mathcal{S}_+| = |S|/2$. Then the misspecified transition kernel is then generated by

$$\mathbb{P}'(s'|s, a) = \mathbb{P}(s'|s, a) + 2\frac{\zeta}{S}\mathbb{1}[s' \in \mathcal{S}_+] - \frac{\zeta}{S},$$

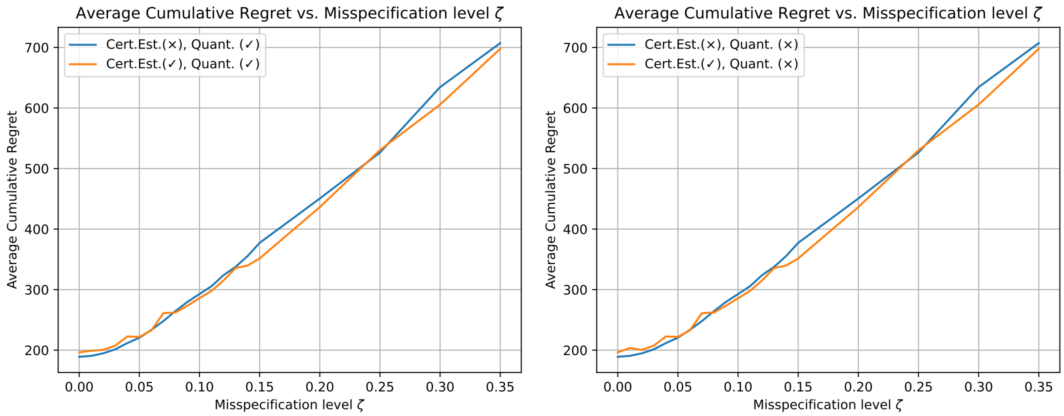
we can verify that $\|\mathbb{P}(\cdot|s, a) - \mathbb{P}'(\cdot|s, a)\|_{\mathrm{TV}} = \zeta$. We investigated the misspecification level from $\zeta = 0, 0.01, \cdots, 0.3$ in 16 randomly generated environments over 2000 episodes. We report the cumulative regret and runtime with respect to different misspecification levels. Additionally, we performed an ablation study by 1) removing the certified estimation (Algorithm 2, Line 11) and 2) removing the quantization (Algorithm 1, Line 8).

The results of these configurations are presented in the following table. The detailed regret for all misspecification level is presented in Table 3 and Figure 1, we plot the cumulative regret for 2000 episodes with respect to the misspecification level ζ. The cumulative regret curve is plotted in Figure 2.

The experimental results suggest several key findings that support our theoretical analysis:

- When the misspecification level is low, it is possible to achieve constant regret, where the instantaneous regret in the final rounds is approximately zero.
- The certified estimator and the quantization do not significantly affect the algorithm's runtime. In contrast, the certified estimator provides an 'early-stopping' condition in Algorithm 2, which slightly reduces the algorithm's runtime. In particular, our algorithm yields a computational complexity of $O(d^2AHK^2 \log K)$, which is the same as Vial et al. (2022) and only $\log K$ greater than the vanilla LSVI-UCB (Jin et al., 2020) due to the multi-phased algorithm.

 https://doi.org/10.52202/079017-4154



(a) With quantization

(b) Without quantization

Figure 1: Cumulative regret over 2000 episodes with respect to different misspecification level ζ. The result is averaged over 16 individual environments.

- The certified estimator helps the algorithm by providing robust estimation in the presence of misspecification. As shown in the table, using the certified estimator does not make a significant difference when the misspecification level ζ is low, but it becomes significant as the misspecification level increases.
- The quantization does not contribute significantly to the results, as the numerical results are intrinsically discrete and quantized. In Figure 2, the regret curve with quantization and the one without quantization are highly overlapped.

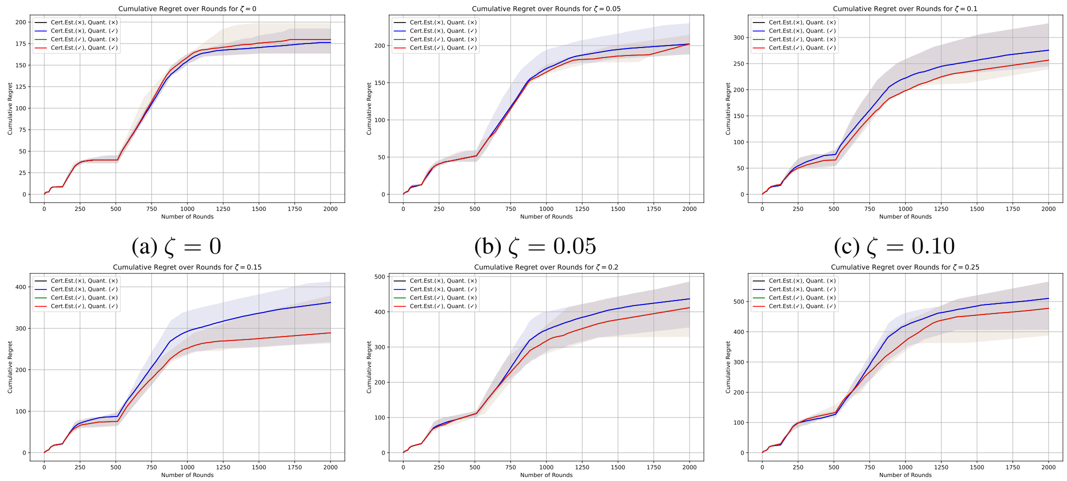
[redacted]	Q.	$\zeta = 0.0$	0.01	0.02	0.03	0.04
×	×	189.01 ± 57.21	190.57 ± 59.49	**194.74** ± 61.99	**201.36** ± 63.67	**211.67** ± 66.06
×	✓	**189.00** ± 57.18	**190.55** ± 59.49	194.75 ± 61.99	201.41 ± 63.67	211.68 ± 66.06
✓	×	196.32 ± 64.09	203.68 ± 77.24	200.41 ± 66.88	207.47 ± 70.97	222.61 ± 78.77
✓	✓	196.31 ± 64.06	199.07 ± 68.45	200.42 ± 66.88	207.52 ± 70.96	222.62 ± 78.77

C.	Q.	$\zeta = 0.05$	0.06	0.07	0.08	0.09
×	×	220.52 ± 67.85	233.57 ± 66.54	248.08 ± 68.70	264.49 ± 72.05	279.82 ± 78.28
×	✓	**220.50** ± 67.86	233.48 ± 66.50	**248.06** ± 68.69	264.56 ± 72.00	279.81 ± 78.25
✓	×	221.87 ± 73.51	232.75 ± 73.75	261.16 ± 84.44	**262.07** ± 82.05	273.56 ± 96.19
✓	✓	221.86 ± 73.51	**232.70** ± 73.81	261.14 ± 84.43	262.15 ± 82.01	**273.55** ± 96.17

C.	Q.	ζ =0.1	0.11	0.12	0.13	0.14
×	×	292.59 ± 80.52	305.73 ± 83.47	323.93 ± 90.25	337.65 ± 94.39	355.49 ± 106.22
×	✓	292.68 ± 80.55	305.71 ± 83.44	323.91 ± 90.23	337.65 ± 94.39	355.49 ± 106.22
✓	×	**285.81** ± 102.54	297.98 ± 107.21	315.22 ± 114.50	**335.69** ± 110.76	**339.81** ± 99.26
✓	✓	285.90 ± 102.57	**297.97** ± 107.18	**315.20** ± 114.49	335.69 ± 110.76	339.82 ± 99.27

C.	Q.	ζ =0.15	0.2	0.25	0.3	Time(s)
×	×	377.27 ± 127.21	450.49 ± 154.80	526.45 ± 181.90	634.46 ± 245.70	1654.76 ± 125.40
×	✓	377.23 ± 127.13	450.48 ± 154.79	**526.36** ± 181.91	634.48 ± 245.68	1654.21 ± 141.31
✓	×	351.52 ± 118.09	436.30 ± 154.89	530.15 ± 194.89	**605.64** ± 233.19	1599.66 ± 138.18
✓	✓	**351.50** ± 118.03	**436.29** ± 154.91	530.58 ± 195.24	605.67 ± 233.18	**1593.38** ± 98.11

Table 3: Average cumulative regret (± standard derivation) and execution time over 2000 episodes. The results are averaged over 16 individual runs. **C** indicates if Certified Estimator is used. **Q** indicates if Quantization is used.



(a) $\zeta = 0$ (b) $\zeta = 0.05$ (c) $\zeta = 0.10$

(d) $\zeta = 0.15$ (e) $\zeta = 0.20$ (f) $\zeta = 0.25$

Figure 2: Cumulative regret with respect to the number of episodes. We reported the median cumulative regret with the shadow area as the region from 25% percentage to 75% percentage statistics over 16 runs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We study the constant regret analysis in reinforcement learning from a theoretical perceptive. The contribution, assumptions and scope are clearly claimed in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations and future potential directions in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

https://doi.org/10.52202/079017-4154

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes] .

   Justification: We provide a detailed proof for all theorems in the appendix, starting from Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: Our submission paper does not include experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with

the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: Our paper does not include experiments requiring code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Our paper does not include experiments.

   Guidelines:

https://doi.org/10.52202/079017-4154

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work provides the theoretical understanding of reinforcement learning. Although there might be some potential social impacts on reinforcement learning applications, according to the guidelines, we believe our result does not have a direct connection with these issues.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We place this paper on the theoretical understand of reinforcement learning thus the paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

https://doi.org/10.52202/079017-4154

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: We do not use existing assets in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-4154