Nearly Minimax Optimal Reinforcement Learning for Linear Markov Decision Processes

Jiafan He [1] **Heyang Zhao** [1] **Dongruo Zhou** [1] **Quanquan Gu** [1]

Abstract

We study reinforcement learning (RL) with linear function approximation. For episodic time-inhomogeneous linear Markov decision processes (linear MDPs) whose transition probability can be parameterized as a linear function of a given feature mapping, we propose the first computationally efficient algorithm that achieves the nearly minimax optimal regret $\widetilde{O}(d\sqrt{H^3K})$, where d is the dimension of the feature mapping, H is the planning horizon, and K is the number of episodes. Our algorithm is based on a weighted linear regression scheme with a carefully designed weight, which depends on a new variance estimator that (1) directly estimates the variance of the *optimal* value function, (2) monotonically decreases with respect to the number of episodes to ensure a better estimation accuracy, and (3) uses a rare-switching policy to update the value function estimator to control the complexity of the estimated value function class. Our work provides a complete answer to optimal RL with linear MDPs, and the developed algorithm and theoretical tools may be of independent interest.

1 Introduction

How to make reinforcement learning (RL) efficient with large state and action spaces has been a central research problem in the RL community. A widely used approach is *function approximation*, which approximates the value function in RL with a predefined function class for efficient exploration and exploitation. Although the intuition is simple, some basic questions about the function approximation approach still remain open. For instance, what is the optimal sample complexity (or regret) for RL algorithms with function approximation to find the optimal policy? Such optimal sample complexity results have been widely-studied and established for tabular RL methods (e.g., Azar et al. 2017; Zhang and Ji 2019; Zhang et al. 2020), but are still understudied for RL with function approximation.

Some recent works have studied the optimal regret results for a special class of MDPs called *linear mixture Markov decision processes (linear mixture MDPs)* (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021a; Zhou and Gu, 2022), which assume that the transition probability of the MDP is a *linear* combination of several base models. More specifically, Zhou et al. (2021a) proposed the near-optimal algorithm for time-inhomogeneous linear mixture MDPs. Zhou and Gu (2022) further proposed near-optimal horizon-free algorithm for time-homogeneous linear mixture MDPs under the assumption that the total reward is bounded by 1. However, the computational efficiency of their algorithms highly depends on the *value-targeted regression* procedure (Jia et al., 2020; Ayoub et al., 2020), which relies on an integration or sampling oracle of the individual base model. Such an integration or sampling oracle exists for some special linear mixture MDPs but can be computationally expensive or even intractable in the general case.

Another line of works studies the *linear Markov decision processes* (linear MDPs) (Yang and Wang, 2019; Jin et al., 2020), which assumes that the transition probability and the reward of the environment enjoys a compact low-rank representation. The most appealing feature of linear MDPs is that they can induce a linear structure of the value function for any policy, which makes sample-efficient RL possible. Meanwhile, the algorithms for linear MDPs directly approximate the value function itself, which is computationally more efficient than the algorithms for linear mixture MDPs. In particular, Yang and Wang (2019) first proposed a near-optimal RL algorithm with the access to a generative model, which can generate any number of samples for any given state-action pairs. Without accessing the generative model, Jin et al. (2020) proposed an LSVI-UCB algorithm based on the principle of optimism in the face of uncertainty and achieved $\widetilde{O}(\sqrt{d^3H^4K})$ regret, where d is the dimension of a linear MDP, H is the planning horizon and K is the number of episodes. Nevertheless, their algorithm is not optimal since there exists an $O(\sqrt{dH})$ gap between

[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

their regret upper bound, and the lower bound $O(d\sqrt{H^3K})$ proved in Zhou et al. (2021a). Zanette et al. (2020b) studied a more general MDP class called *low Bellman error class*, which contains linear MDPs as a special case, and they proposed a computationally inefficient algorithm with a near-optimal regret.

Therefore, a natural question arises[1]:

Can we design a computationally efficient algorithm that achieves the minimax optimality for linear MDPs?

We give an affirmative answer to the above question in this work. Our contributions are listed as follows.

- We propose an algorithm LSVI-UCB++ which attains a near-optimal regret $\widetilde{O}(d\sqrt{H^3K})$ when K is large, which matches the lower bound (Zhou et al., 2021a) up to logarithmic factors. To the best of our knowledge, this is the first computationally efficient RL algorithm that is nearly minimax-optimal for linear MDPs.
- The first key component of our algorithm is a variance-aware weighted ridge regression scheme, which is firstly introduced to acheive nearly minimax optimal regret for linear mixture MDPs in Zhou et al. (2021a) and later improved in Zhou and Gu (2022) to achieve horizon-free regret. Such a component reduces the variance of the estimators in our algorithm, which leads to a $\sqrt{H}$ improvement in the regret over Jin et al. (2020).
- To improve the d dependence, inspired by previous works for tabular RL (Azar et al., 2017), our algorithm utilizes a new strategy to estimate the variance of the estimated value function. Unlike the previous approach for linear mixture MDPs (Zhou et al., 2021a), our new estimator directly estimates the variance of the *true* value function and computes the difference between the variances of the true value function and the estimated one. Such a strategy allows the variance estimator to focus on a simpler function class that only includes the true value function, and therefore gives a tighter confidence set than that in Jin et al. (2020).
- To obtain a uniform variance upper bound, we construct our value function estimator as a monotonically decreasing estimator with a "rare-switching" update strategy, which makes the estimated value function decrease with respect to the episodes and being updated rarely. Together with our new variance estimator, we can remove the additional $\sqrt{d}$ dependency from the previous regret, which makes our algorithm nearly minimax optimal. Notably, our algorithm only needs to update the policy $O(\log K)$ times instead of K times, and therefore enjoys a low-switching cost.

[1]We are aware of a recently published work (Hu et al., 2022), which claims to achieve the nearly minimax optimal regret for linear MDPs. However, a closer examination of their proof can find a technical error, which makes their result invalid. We will discuss it in more detail and show why our algorithm and proof can get around the issue in Appendix A. Using the techniques proposed by our paper, Hu et al. (2022) recently fixed the technical flaw by using the "rare-switching" update strategy and also abandoning the over-optimistic estimator. This is acknowledged in the updated arXiv version of Hu et al. (2022).

Model	Algorithm	Regret
Linear MDP	LSVI-UCB (Jin et al., 2020)	$\widetilde{O}\Big(\sqrt{d^3H^4K}\Big)$
	LSVI-UCB++ (**Our work**)	$\widetilde{O}\Big(d\sqrt{H^3K}\Big)$
Lower bound	Zhou et al. (2021a)	$\Omega(d\sqrt{H^3K})$

Table 1. Comparison of RL with linear function approximation in terms of regret guarantee.

For the ease of comparison, we summarize the regret bounds of our algorithm and previous algorithms for linear MDPs in Table 1.

Recently, an independent concurrent work (Agarwal et al., 2022) proposed a different algorithm that can also achieve near-optimal regret for linear MDPs. Their algorithm follows the algorithm design in Hu et al. (2022), which introduces an additional over-optimistic value function to construct a monotonic variance estimator, and a non-Markovian policy to fix the technical flaw in Hu et al. (2022). In contrast, our algorithm takes a neat approach and constructs the monotonic variance estimator with a simple "rare-switching" update strategy, which enjoys low-switching cost. Agarwal et al. (2022) also studied RL with nonlinear function approximation, which is beyond the scope of this work.

Notation In this work, we use lowercase letters to denote scalars and use lower and uppercase boldface letters to denote vectors and matrices respectively. For a vector $\mathbf{x} \in \mathbb{R}^d$ and matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, we denote by $\|\mathbf{x}\|_2$ the Euclidean norm and $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma}\mathbf{x}}$. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an absolute constant C such that $a_n \le Cb_n$, and we write $a_n = \Omega(b_n)$ if there exists an absolute constant C such that $a_n \ge Cb_n$. We use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to further hide the logarithmic factors. For any $a \le b \in \mathbb{R}$, $x \in \mathbb{R}$, let $[x]_{[a,b]}$ denote the truncate function $a\cdot\mathbb{1}(x \le a)+x\cdot\mathbb{1}(a \le x \le b)+b\cdot\mathbb{1}(b \le x)$, where $\mathbb{1}(\cdot)$ is the indicator function. For a positive integer n, we use $[n] = \{1, 2, .., n\}$ to denote the set of integers from 1 to n.

2 Related Work

Near-optimal tabular reinforcement learning There is a voluminous amount of works developing nearly minimax optimal algorithms for tabular MDPs under different settings (Azar et al., 2017; Zanette and Brunskill, 2019; Zhang and Ji, 2019; Simchowitz and Jamieson, 2019; Zhang et al., 2020; 2021a; He et al., 2021b). A key idea behind these works is to exploit the $O(H^2)$ total variance of the value functions for each episode (Azar et al., 2017; Jin et al., 2018). Azar et al. (2017) first proposed this idea to design Bernstein-type bonuses in tabular MDPs and provided an $\widetilde{O}(\sqrt{H^2SAK})$ regret upper bound, which matches the lower bound for the tabular setting. In their analysis, Azar et al. (2017) also introduced a new value function decomposition scheme which mainly focuses on the variances of the optimal value function rather than the estimated value function. Zhang et al. (2021a) further improved the dependence on H for the constant terms and achieved a nearly minimax optimal horizon-free regret (nearly independent of H) under the assumption that the total reward is bounded by 1. Our algorithm extends the idea of Bernstein-type bonuses and value function decomposition in Azar et al. (2017) to RL with linear function approximation.

Reinforcement Learning with Linear Function Approximation. There exists a large body of literature on RL with linear function approximation (Jiang et al., 2017; Dann et al., 2018; Yang and Wang, 2019; Jin et al., 2020; Wang et al., 2020; Du et al., 2019; Sun et al., 2019; Zanette et al., 2020a; Yang and Wang, 2020; Modi et al., 2020; Ayoub et al., 2020; Zhou et al., 2021a; He et al., 2021a; Zhou and Gu, 2022). All these works assume certain linear structures of the underlying MDP. The most related work to ours is initiated by Yang and Wang (2019), which assumes that the reward function and the transition probability are linear in the feature mapping $\phi(s,a)$ for each state-action pair (s,a). Jin et al. (2020) further considered *Linear MDPs* and proposed LSVI-UCB which achieves an $\widetilde{O}(\sqrt{d^3H^4K})$ regret bound. Zanette et al. (2020a) proposed a Thompson sampling based algorithm for linear MDPs, which attains a regret upper bound of order $\widetilde{O}(\sqrt{d^4H^5K})$. Another popular MDP model for RL with linear function approximation is *linear mixture Markov Decision Processes* (Modi et al., 2020; Yang and Wang, 2020; Jia et al., 2020; Ayoub et al., 2020), or *Linear Kernel MDPs* (Zhou et al., 2021b), where the transition probability is a linear combination of several base models. For linear mixture MDPs, Zhou et al. (2021a) is the first to achieve a nearly minimax optimal regret bound. There are also works achieving horizon-free regret bounds for time-homogeneous linear mixture MDPs (Zhang et al., 2021b; Zhou and Gu, 2022). Compared with Zhou et al. (2021a), our algorithm is the first to achieve the near-optimality for linear MDPs.

3 Preliminaries

In this work, we consider the episodic Markov Decision Processes (MDP), where the MDP can be denoted by a tuple of $M(\mathcal{S},\mathcal{A},H,\{r_h\}_{h=1}^H,\{\mathbb{P}_h\}_{h=1}^H)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the finite action space, H is the length of each episode (i.e., planning horizon), $r_h:\mathcal{S}\times\mathcal{A}\to[0,1]$[2] is the reward function at stage h and $\mathbb{P}_h(s'|s,a)$ is the transition probability function at stage h which denotes the probability for state s to transfer to next state s' with current action a. Following Jin et al. (2020), we assume that $\mathcal{S}$ is a measurable space with possibly infinite number of states and $\mathcal{A}$ is a finite set. A policy $\pi:\mathcal{S}\times[H]\to\mathcal{A}$ is a function that maps a state s and the stage number h to an action a. For any stage $h\in[H]$ and policy π, we define the value function $V_h^\pi(s)$ and the action-value function $Q_h^\pi(s,a)$ as follows:

$$\begin{aligned}Q_h^\pi(s,a) &= r_h(s,a)\\ &\quad+\mathbb{E}\bigg[\sum_{h'=h+1}^{H} r_{h'}\big(s_{h'},a_{h'}\big)\big|s_h=s,a_h=a\bigg],\\ V_h^\pi(s) &= Q_h^\pi\big(s,\pi(s,h)\big),\end{aligned}$$

where $s_{h'+1}\sim\mathbb{P}_h(\cdot|s_{h'},a_{h'})$ denotes the state at stage $h'+1$ and $a_{h'}=\pi(s_{h'},h')$ denotes the action at stage h' . Furthermore, we can define the optimal value function V_h^* and the optimal action-value function Q_h^* as $V_h^*(s)=\max_\pi V_h^\pi(s)$ and $Q_h^*(s,a)=\max_\pi Q_h^\pi(s,a)$. By this definition, the value function $V_h^\pi(s)$ and action-value function $Q_h^\pi(s,a)$ are bounded in $[0,H]$. For any function $V:\mathcal{S}\to\mathbb{R}$, we denote $[\mathbb{P}_hV](s,a)=\mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}V(s')$ and $[\mathbb{V}_hV](s,a)=[\mathbb{P}_hV^2](s,a)-\big([\mathbb{P}_hV](s,a)\big)^2$ for simplicity. Thus, for every stage $h\in[H]$ and policy π, we have the following Bellman equation for value functions $Q_h^\pi(s,a)$ and $V_h^\pi(s)$, as well as the Bellman optimality equation for optimal value functions $Q_h^*(s,a)$ and $V_h^*(s)$:

$$\begin{aligned}Q_h^\pi(s,a) &= r_h(s,a)+[\mathbb{P}_hV_{h+1}^\pi](s,a),\\ Q_h^*(s,a) &= r_h(s,a)+[\mathbb{P}_hV_{h+1}^*](s,a),\end{aligned}$$

where $V_{H+1}^\pi(s)=V_{H+1}^*(s)=0$. At the beginning of each episode $k\in[K]$, the agent selects a policy π_k to be followed in this episode. At each stage $h\in[H]$, the agent first observes the current state s_h^k, chooses an action following the policy π_k and then observes the next state with $s_{h+1}^k\sim\mathbb{P}_h(\cdot|s_h^k,a_h^k)$. Based on these definitions, we further define the regret in the first K episodes as follows:

Definition 3.1. For any algorithm `Alg`, we define its regret on learning an MDP $M(\mathcal{S},\mathcal{A},H,r,\mathbb{P})$ in the first K

[2] In this work, we study the deterministic and known reward functions for simplicity, and it is not difficult to generalize our results to stochastic and unknown linear reward functions in (Jin et al., 2019), where $r_h(s,a)=\langle\boldsymbol{\phi}(s,a),\boldsymbol{\mu}_h\rangle$.

episodes as the sum of the sub-optimality gaps for episode $k = 1, \ldots, K$, i.e.,

$$\text{Regret}(K) = \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k),$$

where π_k is the agent's policy in the k-th episode.

Linear Markov Decision Process In this work, we focus on the linear Markov decision Process (Jin et al., 2020; Yang and Wang, 2019), which is formally defined as follows:

Definition 3.2. An MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$ is a linear MDP if for any stage $h \in [H]$, there exists an unknown measure $\boldsymbol{\theta}_h(\cdot) : \mathcal{S} \to \mathbb{R}^d$ and a known feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, such that for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and state $s' \in \mathcal{S}$, we have

$$\mathbb{P}_h(s'|s, a) = \big\langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h(s') \big\rangle. \tag{3.1}$$

For simplicity, we assume that the norms of $\boldsymbol{\theta}_n(\cdot)$ and $\boldsymbol{\phi}(\cdot, \cdot)$ are upper bounded by $\|\boldsymbol{\phi}(s, a)\|_2 \leq 1$ and $\big\|\boldsymbol{\theta}_h(\mathcal{S})\big\|_2 \leq \sqrt{d}$. For linear MDPs, we have the following property:

Proposition 3.3 (Proposition 2.3, Jin et al. 2020)**.** *For any policy π, there exist weights $\{\mathbf{w}_h^\pi\}_{h=1}^H$ such that for any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and stage $h \in [H]$, we have $[\mathbb{P}V_{h+1}^\pi](s, a) = \langle \boldsymbol{\phi}(s, a), \mathbf{w}_h^\pi \rangle$.*

4 The Proposed Algorithm

In this section, we propose a new algorithm LSVI-UCB++ to learn the linear MDPs (See Definition 3.2). The main algorithm is illustrated in Algorithm 1. In the sequel, we introduce the key ideas of the proposed algorithm one by one.

4.1 Weighted Ridge Regression

The basic framework of our algorithm follows the LSVI-UCB algorithm proposed by Jin et al. (2020). Based on Proposition 3.3 that the expected value function $[\mathbb{P}_h V_{h+1}^\pi](s, a) = \langle \boldsymbol{\phi}(s, a), \mathbf{w}_h^\pi \rangle$, Algorithm 1 reduces the learning of the optimal action-value function into a series of linear regression problems. In order to have a good estimation for the vector $\mathbf{w}_h^\pi$ and achieve the minimax-optimal regret result, Algorithm 1 adapts the weighted ridge regression method (Henderson, 1975), which was used in heteroscedastic linear bandits (Lattimore et al., 2015; Kirschner and Krause, 2018) and more recently RL with linear function approximation (Zhou et al., 2021a) for linear mixture MDPs. In detail, for each stage $h \in [H]$ and episode $k \in [K]$, we construct the estimator $\widehat{\mathbf{w}}_{k,h}$ by solving the following weighted ridge regression

$$\widehat{\mathbf{w}}_{k,h} \leftarrow \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^d} \lambda \|\mathbf{w}\|_2^2$$

Algorithm 1 LSVI-UCB++

Require: Regularization parameter $\lambda > 0$, confidence radius $\beta, \bar{\beta}, \widetilde{\beta}$

1: Initialize $k_{\text{last}} = 0$ and for each stage $h \in [H]$ set $\boldsymbol{\Sigma}_{0,h}, \boldsymbol{\Sigma}_{1,h} \leftarrow \lambda \mathbf{I}$
2: For each stage $h \in [H]$ and state-action $(s, a) \in \mathcal{S} \times \mathcal{A}$, set $Q_{0,h}(s, a) \leftarrow H, \check{Q}_{0,h}(s, a) \leftarrow 0$
3: **for** episodes $k = 1, \ldots, K$ **do**
4: Received the initial state s_1^k.
5: **for** stage $h = H, \ldots, 1$ **do**
6: $\widehat{\mathbf{w}}_{k,h} = \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) V_{k,h+1}(s_{h+1}^i)$
7: $\check{\mathbf{w}}_{k,h} = \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \check{V}_{k,h+1}(s_{h+1}^i)$
8: **if** there exists a stage $h' \in [H]$ such that $\det(\boldsymbol{\Sigma}_{k,h'}) \geq 2 \det(\boldsymbol{\Sigma}_{k_{\text{last}},h'})$ **then**
9: $Q_{k,h}(s, a) = \min \Big\{ r_h(s, a) + \widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s, a) + \beta \sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s, a)}, Q_{k-1,h}(s, a), H \Big\}$
10: $\check{Q}_{k,h}(s, a) = \max \Big\{ r_h(s, a) + \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s, a) - \bar{\beta} \sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s, a)}, \check{Q}_{k-1,h}(s, a), 0 \Big\}$
11: Set the last updating episode $k_{\text{last}} = k$
12: **else**
13: $Q_{k,h}(s, a) = Q_{k-1,h}(s, a)$
14: $\check{Q}_{k,h}(s, a) = \check{Q}_{k-1,h}(s, a)$
15: **end if**
16: $V_{k,h}(s) = \max_a Q_{k,h}(s, a)$
17: $\check{V}_{k,h}(s) = \max_a \check{Q}_{k,h}(s, a)$
18: **end for**
19: **for** stage $h = 1, \ldots, H$ **do**
20: Take action $a_h^k \leftarrow \operatorname{argmax}_a Q_{k,h}(s_h^k, a)$
21: Set the estimated variance $\sigma_{k,h}$ as in (4.1)
22: $\bar{\sigma}_{k,h} \leftarrow \max \big\{ \sigma_{k,h}, H, 2d^3 H^2 \|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}^{1/2} \big\}$
23: $\boldsymbol{\Sigma}_{k+1,h} = \boldsymbol{\Sigma}_{k,h} + \bar{\sigma}_{k,h}^{-2} \boldsymbol{\phi}(s_h^k, a_h^k) \boldsymbol{\phi}(s_h^k, a_h^k)^\top$
24: Receive next state s_{h+1}^k
25: **end for**
26: **end for**

$$+ \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \big(\mathbf{w}^\top \boldsymbol{\phi}(s_h^i, a_h^i) - V_{k,h+1}(s_{h+1}^i) \big)^2.$$

Here, we take the inverse of the estimated variances $\sigma_{k,h}^2$ as the weights for the regression problem and set $\sigma_{k,h}$ as

$$\bar{\sigma}_{k,h} = \max \big\{ \sigma_{k,h}, H, 2d^3 H^2 \|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}^{1/2} \big\}$$

in Line 22 of Algorithm 1, which depends on the uncertainty term $\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}$. Note that the uncertainty-dependent weight has also been used in He et al. (2022) to defend the adversarial corruption in the linear bandits problem. The reason why we want to use an uncertainty-dependent weight can be explained by the following lemma.

Lemma 4.1 (Theorem 4.3, Zhou and Gu 2022)**.** *Let $\{\mathcal{G}_k\}_{k=1}^{\infty}$ be a filtration, and $\{\mathbf{x}_k, \eta_k\}_{k\geq 1}$ be a stochastic process such that $\mathbf{x}_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $k \geq 1$, let $y_k = \langle \boldsymbol{\mu}^*, \mathbf{x}_k \rangle + \eta_k$ and suppose that $\eta_k, \mathbf{x}_k$ also satisfy*

$$\mathbb{E}[\eta_k|\mathcal{G}_k] = 0,\ \mathbb{E}[\eta_k^2|\mathcal{G}_k] \leq \sigma^2,\ |\eta_k| \leq R,\ \|\mathbf{x}_k\|_2 \leq L.$$

For $k \geq 1$, let $\mathbf{Z}_k = \lambda\mathbf{I} + \sum_{i=1}^{k} \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_k = \sum_{i=1}^{k} y_i\mathbf{x}_i$, $\boldsymbol{\mu}_k = \mathbf{Z}_k^{-1}\mathbf{b}_k$, and $\beta_k = \widetilde{O}\big(\sigma\sqrt{d} + \max_{1\leq i\leq k} |\eta_i| \min\{1, \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\}\big)$. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$, for all $k \in [K]$, we have

$$\big\|\textstyle\sum_{i=1}^{k} \mathbf{x}_i\eta_i\big\|_{\mathbf{Z}_k^{-1}} \leq \beta_k,\ \|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_k} \leq \beta_k + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2.$$

By Lemma 4.1, one can easily verify that $|\langle \widehat{\mathbf{w}}_{k,h}, \boldsymbol{\phi}(s,a)\rangle - \mathbb{P}_h V_{k,h+1}(s,a)| = O\Big(\beta\big\|\boldsymbol{\Sigma}_{k,h}^{-1/2}\boldsymbol{\phi}(s,a)\big\|_2\Big)$, where $\beta = \widetilde{O}(\sqrt{d})$. Such an $\widetilde{O}(\sqrt{d})$ dependence is similar to that in Zhou and Gu (2022), which allows our algorithm to use a tighter confidence set than Jin et al. (2020). Therefore, we can construct the optimistic value function $Q_{k,h}$ with the linear function and an additional exploration bonus term (Line 7 in Algorithm 1), i.e.,

$$Q_{k,h}(s,a) \approx r_h(s,a) + \widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s,a) + \beta\big\|\boldsymbol{\Sigma}_{k,h}^{-1/2}\boldsymbol{\phi}(s,a)\big\|_2.$$

With the help of the exploration bonus, we can show that the optimistic value function $Q_{k,h}(s,a)$ is an upper bound of the optimal value function $Q_h^*(s,a)$ and the summation of the sub-optimality gaps can be upper bounded by the summation of exploration bonus $\sum_{h=1}^{H}\sum_{k=1}^{K} \beta\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)}$. By adapting the weighted ridge regression, Zhou and Gu (2022) proposed HF-UCRL-VTR+, which is able to achieve a nearly minimax optimal regret for linear mixture MDPs. However, their algorithm and approach cannot be directly applied to linear MDPs, and we need to construct a pessimistic value function $\check{V}_{k,h}$ for the optimal value function $Q_h^*(s,a)$ to estimate the gap between $V_{k,h}(s)$ and $V_h^*(s)$, where we have $V_{k,h}(s) - V_h^*(s) \leq V_{k,h}(s) - \check{V}_{k,h}(s)$. Similar to the optimistic value function, we construct the vector $\check{\mathbf{w}}_{k,h}$ by solving the following weighted ridge regression,

$$\check{\mathbf{w}}_{k,h} \leftarrow \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \lambda\|\mathbf{w}\|_2^2 + \textstyle\sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2}\big(\mathbf{w}^\top\boldsymbol{\phi}(s_h^i, a_h^i) - \check{V}_{k,h+1}(s_{h+1}^i)\big)^2,$$

and compute the pessimistic value function $\check{Q}_{k,h}$ as:

$$\check{Q}_{k,h}(s,a) \approx r_h(s,a) + \check{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s,a) - \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)},$$

where $\bar{\beta} = \widetilde{O}\Big(\sqrt{d^3H^2}\Big)$. We can show that the pessimistic value function $\check{V}_{k,h}(s)$ is a lower bound for the optimal value function $V_h^*(s)$.

4.2 Variance Estimator

We compare our variance estimator and its counterparts in Zhou et al. (2021a). Zhou et al. (2021a) first introduced variance estimators into RL with linear function approximation. They studied linear mixture MDPs, and their algorithm estimates the variance of the *optimistic* value function $V_{k,h+1}(s)$ directly. In comparison, for linear MDPs, estimating the variance of the *optimistic* value function $V_{k,h+1}(s)$ will encounter the dependence issue, which is discussed in Jin et al. (2020) and will introduce an additional $\sqrt{d}$ factor in the regret due to the covering-based decoupling argument. Inspired by the previous works (Azar et al., 2017; Hu et al., 2022), we decompose the *optimistic* value function $V_{k,h+1}(s)$ into the *optimal* value function $V_{h+1}^*(s)$ and the sub-optimality gap $V_{k,h+1}(s) - V_{h+1}^*(s)$, then estimate their variances $[\mathbb{V}_h V_{h+1}^*](s,a)$ and $\big[\mathbb{V}_h(V_{k,h+1} - V_{h+1}^*)\big](s,a)$ separately.

For the variance of *optimal* value function $[\mathbb{V}_h V_{h+1}^*](s,a)$, since neither the variance operator $\mathbb{V}_h$ nor the optimal value function V_{h+1}^* is observable, Algorithm 1 takes several steps to estimate these two quantities. In detail, Algorithm 1 uses the optimistic value function $V_{k,h+1}$ to estimate the optimal value function V_{h+1}^* and introduce an error term $D_{k,h}$ to bound the difference between $\mathbb{V}_h V_{k,h+1}$ and $\mathbb{V}_h V_{h+1}^*$. For the variance operator, Algorithm 1 follows Zhou et al. (2021a) to write the variance as the difference between the second-order moment and the square of the first-order moment of $V_{k,h}$, which is upper bounded by the bonus term $E_{k,h}$. More specifically, the variance of function $V_{k,h}$ can be denoted by

$$[\mathbb{V}_h V_{k,h}](s,a) = [\mathbb{P}_h V_{k,h}^2](s,a) - \big([\mathbb{P}_h V_{k,h}](s,a)\big)^2.$$

According to the Proposition 3.3, the expectation $\mathbb{P}_h V_{k,h}(s,a)$ and $\mathbb{P}_h V_{k,h}^2(s,a)$ are linear in the feature mapping $\boldsymbol{\phi}(s,a)$ and can be approximated as follows,

$$\begin{aligned}[\mathbb{V}_h V_{k,h}](s,a) &\approx \bar{\mathbb{V}}_h V_{k,h+1}(s_h^k, a_h^k)\\ &:= \big[\widetilde{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k, a_h^k)\big]_{[0,H^2]} - \big[\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k, a_h^k)\big]_{[0,H]}^2,\end{aligned}$$

where

$$\widetilde{\mathbf{w}}_{k,h} := \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \lambda\|\mathbf{w}\|_2^2 + \textstyle\sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2}\big(\mathbf{w}^\top\boldsymbol{\phi}(s_h^i, a_h^i) - V_{k,h+1}^2(s_{h+1}^i)\big)^2$$

is the solution to the weighted ridge regression problem for the squared value function. To summarize, LSVI-UCB++

constructs the estimated variance $\sigma_{k,h}$ as follows:

$$\sigma_{k,h} = \sqrt{[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h} + D_{k,h} + H}, \tag{4.1}$$

where $E_{k,h}$ and $D_{k,h}$ are defined as follows

$$\begin{aligned}
E_{k,h} &= \min\Big\{\widetilde{\beta}\big\|\boldsymbol{\Sigma}_{k,h}^{-1/2}\boldsymbol{\phi}(s_h^k, a_h^k)\big\|_2, H^2\Big\} \\
&\quad + \min\Big\{2H\bar{\beta}\big\|\boldsymbol{\Sigma}_{k,h}^{-1/2}\boldsymbol{\phi}(s_h^k, a_h^k)\big\|_2, H^2\Big\}, \\
D_{k,h} &= \min\Big\{4d^3H^2\Big(\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k, a_h^k) - \check{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k, a_h^k) \\
&\quad + 2\bar{\beta}\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k, a_h^k)}\Big), d^3H^3\Big\}.
\end{aligned}$$

Here $E_{k,h}$ is the error between the estimated variance and the true variance of $V_{k,h+1}$, and $D_{k,h}$ is the error between the variance of $V_{k,h+1}$ and the variance of the optimal value function V_h^*. For term $D_{k,h}$, we use the difference between the optimistic value function $V_{k,h}$ and the pessimistic value function $\check{V}_{k,h}$ to bound the difference between $V_{k,h}$ and V_h^*. More discussions on the decomposition and the variance estimator can be found in Section 6.

5 Main Results

In this section, we provide the regret bound for our algorithm LSVI-UCB++.

Theorem 5.1. *For any linear MDP M, if we set the parameters $\lambda = 1/H^2$ and confidence radius $\beta, \bar{\beta}, \widetilde{\beta}$ as*

$$\begin{aligned}
\beta &= O\Big(H\sqrt{d\lambda} + \sqrt{d\log^2\big(1 + dKH/(\delta\lambda)\big)}\Big), \\
\bar{\beta} &= O\Big(H\sqrt{d\lambda} + \sqrt{d^3H^2\log^2\big(dHK/(\delta\lambda)\big)}\Big), \\
\widetilde{\beta} &= O\Big(H^2\sqrt{d\lambda} + \sqrt{d^3H^4\log^2\big(dHK/(\delta\lambda)\big)}\Big),
\end{aligned}$$

then with high probability of at least $1 - 7\delta$, the regret of LSVI-UCB++ is upper bounded as follows:

$$\textit{Regret}(K) \le \widetilde{O}\Big(d\sqrt{H^3K} + d^7H^8\Big).$$

In addition, the number of updates for $Q_{k,h}, \check{Q}_{k,h}$ is upper bounded by $O(dH\log(1 + K/\lambda))$.

Remark 5.2. When the number of episodic K satisfies that K is large, the regret can be simplified as $\widetilde{O}(d\sqrt{H^3K})$. Compared with the lower bound $\Omega(d\sqrt{H^3K})$ proved in Zhou et al. (2021a), our regret bound matches the lower bound up to logarithmic factors, which suggests that LSVI-UCB++ is near-optimal for linear MDPs.

Remark 5.3. For LSVI-UCB++, based on the optimistic property $Q_{k,h}(s,a) \ge V_h^*(s)$ and the pessimistic property $V_{k,h}^{\pi^k}(s) \ge \check{V}_{k,h}(s)$, the sub-optimality at episode k is upper bounded by $V_{k,1}(s_1) - \check{V}_{k,1}(s_1)$. Thus, the total regret for the first K episodes can be roughly upper bounded by $\textbf{Regret}(K) \le \sum_{k=1}^K \big(V_{k,1}(s_1) - \check{V}_{k,1}(s_1)\big) = O(d^2\sqrt{H^5/K} + d^8H^9/K)$. When the initial state s_1^k is fixed across all episodes $k \in [K]$, according to the monotinoic property of the optimistic value function $V_{k,1}$ and the pessimistic value function $\check{V}_{k,1}$, the sub-optimality gap $\big(V_{k,1}(s_1) - \check{V}_{k,1}(s_1)\big)$ is decreasing. As a result, the cumulative regret up to episode K satisfies $V_{K,1}(s_1) - \check{V}_{K,1}(s_1) \le 1/K \times \sum_{k=1}^K \big(V_{k,1}(s_1) - \check{V}_{k,1}(s_1)\big) = O(1/\sqrt{K})$, which implies a Probably Approximately Correct (PAC) guarantee. Therefore, LSVI-UCB++ will converge to the optimal policy and enjoys an (ϵ, δ)-PAC guarantee with sample complexity $\widetilde{O}(1/\epsilon^2)$ without any modification of the algorithm. In contrast, to obtain the (ϵ, δ)-PAC guarantee, the LSVI-UCB algorithm in Jin et al. (2020) needs to randomly select a policy uniformly from the previous K policies.

Computational Complexity As shown in Jin et al. (2020), the computational complexity of the original LSVI-UCB is $O(d^2|\mathcal{A}|HK^2)$, where $\mathcal{A}$ is a finite action space and $|\mathcal{A}|$ is the size of the action set. Compared with the LSVI-UCB algorithm, Algorithm 1 uses the "rare-switching" technique, where the algorithm only updates the estimated value functions if the determinant of the covariance matrix $\det(\boldsymbol{\Sigma}_{h,k})$ doubles (Line 8). According to Lemma F.1, the number of episodes that triggers the updating criterion is at most $dH\log(1 + K/\lambda)$ and the action-value function $Q_{k,h}(s,a)$ can be represented as a minimum over $dH\log(1 + K/\lambda)$ quadratic functions. Therefore, given all previous optimistic weight vectors $\mathbf{w}_{i,h}$ and covariance matrices $\boldsymbol{\Sigma}_{i,h}$, computing the optimistic value function $Q_{k,h}(s,a)$ needs $\widetilde{O}(d^3H)$ computational complexity. Thus, for each episode $k \in [K]$, calculating the value function $Q_{k,h}(s_h^k, a)$, choosing the action $a_h^k \leftarrow \arg\max_a Q_{k,h}(s_h^k, a)$ and estimating the variance $\bar{\sigma}_{k,h}$ will only lead to $\widetilde{O}(d^3H^2|\mathcal{A}|)$ computational complexity.

For computing the linear regression weight vectors (Line 6 to Line 7), if the updating criterion is not triggered in episode k, then LSVI-UCB++ only needs to update the weight vectors $\widehat{\mathbf{w}}_{k,h}$ and $\check{\mathbf{w}}_{k,h}$. Since the value functions $V_{k,h+1}$ and $\check{V}_{k,h+1}$ remain unchanged, we only need to compute the new terms $\sigma_{k-1,h}^{-2}\boldsymbol{\phi}(s_h^{k-1}, a_h^{k-1})V_{k,h+1}(s_{h+1}^{k-1})$ and $\sigma_{k-1,h}^{-2}\boldsymbol{\phi}(s_h^{k-1}, a_h^{k-1})\check{V}_{k,h+1}(s_{h+1}^{k-1})$, which has an $O(d^3H|\mathcal{A}|)$ computational complexity. On the other hand, if the updating criterion is triggered in episode k, then LSVI-UCB++ needs to update the value function and recalculate the weight vectors $\widehat{\mathbf{w}}_{k,h}, \check{\mathbf{w}}_{k,h}$, which incurs an $\widetilde{O}(d^4H^2|\mathcal{A}|K)$ computational complexity. Combining the computational complexity for all episodes and noticing that the number of episodes that trigger the updating criterion

is at most $\widetilde{O}(dH)$, the total computational complexity of LSVI-UCB++ is $\widetilde{O}(d^4H^3|\mathcal{A}|K)$, which improves the original LSVI-UCB algorithm by a factor of K.

6 Overview of Key Techniques

In this section, we provide an overview of the key techniques in our algorithm design and analysis.

6.1 Decompose $V_{k,h+1}$ to V^*_{h+1} and $V_{k,h+1} - V^*_{h+1}$

We start with estimating the Bellman backup $[\mathbb{P}_h V_{k,h+1}](s,a)$, which is the main difficulty in almost all existing analyses of algorithms for linear MDPs. According to Proposition 3.3, for any value function V and state-action pair (s,a), the Bellman backup $[\mathbb{P}_h V](s,a)$ is always a linear function of the feature mapping $\phi(s,a)$ can be approximated as follows

$$\begin{aligned}&[\widehat{\mathbb{P}}_{k,h} V_{k,h+1}](s,a)\\ &\approx \boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \sigma_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) V_{k,h+1}(s_{h+1}^i),\end{aligned}$$

which utilizes all the past observations $\phi(s_h^i, a_h^i)$ and the associated values $V(s_{h+1}^k)$. In addition, the estimation error for this estimator can be measured by

$$\begin{aligned}&[\widehat{\mathbb{P}}_{k,h} V_{k,h+1}](s,a) - [\mathbb{P}_h V_{k,h+1}](s,a)\\ &\approx \boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \sigma_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \eta_{i,h}(V_{k,h+1}), \quad (6.1)\end{aligned}$$

where $\eta_{i,h}(V) = V(s_{h+1}^i) - [\mathbb{P}_h V](s_{h+1}^i)$ denotes the stochastic transition noise at episode i with value function V. According to the Bernstein-type self-normalized martingale inequality (Lemma 4.1), the summation of stochastic noise can be bounded by a small value (e.g., $\big|[\widehat{\mathbb{P}}_{k,h} V](s,a) - [\mathbb{P}_h V](s,a)\big| \le \beta\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)}$). However, Jin et al. (2020) noticed that the estimation of the optimistic value function $V_{k,h+1}$ *depends* on the past observations $(s_h^k, a_h^k, s_{h+1}^k)$, which violates the conditional independence required by the martingale concentration inequality, i.e., $\mathbb{E}\big[\eta_{i,h}(V_{k,h+1})\big] \neq 0$.

To deal with this problem, Jin et al. (2020) applied the uniform convergence argument based on covering number for all possible value functions and introduce a fixed covering set to replace $V_{k,h}$ in their analysis. In detail, the function class considered in Jin et al. (2020) is denoted by

$$\begin{aligned}\mathcal{V} = \Bigg\{ V \Bigg| V(\cdot) = \max_a \min \Bigg(H, \mathbf{w}_i^\top \boldsymbol{\phi}(\cdot, a)\\ + \beta\sqrt{\boldsymbol{\phi}(\cdot,a)^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\phi}(\cdot,a)} \Bigg), \|\mathbf{w}_i\| \le L, \boldsymbol{\Sigma}_i \succeq \lambda \mathbf{I} \Bigg\}.\end{aligned}$$

We can cover $\mathcal{V}$ by an ϵ-net denoted by $\mathcal{N}_\epsilon$, and its covering entropy $\log \mathcal{N}_\epsilon$ satisfies $\log \mathcal{N}_\epsilon = \widetilde{O}(d^2)$. Such an approach, although fixing the dependency issue in (6.1), introduces an additional $\sqrt{d}$ factor to their final regret since each V belongs to a quadratic function class by their optimistic construction, which prevents them from achieving the optimal d dependency in the regret.

In comparison, our approach gets around the covering number issue by decomposing the value function $V_{k,h}$ into the optimal value function V^*_{h+1} and the sub-optimality gap $V_{k,h+1} - V^*_{h+1}$. Such an analysis approach has been firstly considered in the tabular MDPs Azar et al. (2017); Zhang et al. (2021a) and later in the linear MDP by Hu et al. (2022). More specifically, we have

$$\begin{aligned}&[\widehat{\mathbb{P}}_{k,h} V_{k,h+1}](s,a) - [\mathbb{P}_h V_{k,h+1}](s,a)\\ &\approx \boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \sigma_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \eta_{i,h}(V_{k,h+1})\\ &= \underbrace{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \sigma_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \eta_{i,h}(V^*_{h+1})}_{I_1}\\ &+ \underbrace{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \sigma_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \eta_{i,h}(\Delta V_{k,h+1})}_{I_2}, \quad (6.2)\end{aligned}$$

where $\Delta V_{k,h+1} = V_{k,h+1} - V^*_{h+1}$ denotes the estimation error for value function $V_{k,h+1}$. For the first term I_1, as discussed in Section 4.2, we can use $V_{k,h+1}$ to approximate the optimal value function V^*_{h+1} and the estimation error for the variance can be bounded by:

$$\big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V^*_{h+1}](s_h^k, a_h^k)\big| \le E_{k,h} + D_{k,h}.$$

Since the optimal value function V^*_{h+1} is fixed across all episodes $k \in [K]$ and does not depend on the past observations, such an approach can prevent the covering number argument and save a $\sqrt{d}$ factor in the regret compared with Jin et al. (2020). For the second term I_2, the sub-optimality gap $\Delta V_{k,h+1} = V_{k,h+1} - V^*_{h+1}$ depends on the past observations and we still need to use the covering number argument. However, the magnitude of the sub-optimality gap $\Delta V_{k,h+1}$ is small provided that $V_{k,h+1}$ is an accurate estimate for V^*_{h+1}. In this case, term I_2 will be dominated by term I_1 even with the extra factors from the covering number argument. With the help of the decomposition, we have the following Bernstein-type error bound between the estimated $\widehat{\mathbb{P}}_{k,h} V_{k,h+1}$ and its true value:

$$|\widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \mathbb{P}_h V_{k,h+1}(s,a)| \le \beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}.$$

This analysis also explains why Algorithm 1 needs to estimate the variance of V^*_{h+1} and $\Delta V_{k,h+1}$ instead of $V_{k,h+1}$.

6.2 Monotonic Variance Estimator

Here we provide more details about the variance estimator $\sigma_{k,h}$. According to our previous discussion, we decompose the value function $V_{k,h}$, and only need to control the estimation errors I_1 and I_2 in (6.2) separately. In order to derive a Bernstein-type error bound, we use Lemma 4.1 for both the optimal value function V^*_{h+1} and $\Delta V_{k,h+1}$, which require an estimation for the variance $\mathbb{V}_h V^*_{h+1}$ and $\mathbb{V}_h[V_{k,h+1} - V^*_{h+1}]$. For the variance $\mathbb{V}_h V^*_{h+1}$, as we discussed in Section 4.2, we approximate it with the following empirical variance:

$$\begin{aligned}
&\bar{\mathbb{V}}_h V_{k,h+1}(s_h^k, a_h^k) - [\mathbb{V}_h V^*_{h+1}](s_h^k, a_h^k) \\
&= \underbrace{\bar{\mathbb{V}}_h V_{k,h+1}(s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)}_{J_1} \\
&\quad + \underbrace{[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V^*_{h+1}](s_h^k, a_h^k)}_{J_2},
\end{aligned}$$

where the estimation error J_1 can be controlled by a Hoeffding-type bound (term $E_{k,h}$) and the estimation error J_2 can be upper bounded by

$$\begin{aligned}
&\big|[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V^*_{h+1}](s_h^k, a_h^k)\big| \\
&\le 4H[\mathbb{P}_h(V_{k,h+1} - V^*_{h+1})](s_h^k, a_h^k).
\end{aligned}$$

For this error bound, the optimal value function V^*_{h+1} is not observable and we replace it by the *pessimistic* value function $\check{V}_{k,h}$, which gives an upper bound

$$\begin{aligned}
&4H[\mathbb{P}_h(V_{k,h+1} - V^*_{h+1})](s_h^k, a_h^k) \\
&\le 4H[\mathbb{P}_h(V_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k).
\end{aligned}$$

The above term is further dominated by the term $D_{k,h}$. With a similar approach, for the variance $\mathbb{V}_h[V_{k,h+1} - V^*_{h+1}]$, it can be upper bound by

$$\begin{aligned}
&[\mathbb{V}_h(V_{k,h+1} - V^*_{h+1})](s_h^k, a_h^k) \\
&\le 2H[\mathbb{P}_h(V_{k,h+1} - V^*_{h+1})](s_h^k, a_h^k) \\
&\le 2H[\mathbb{P}_h(V_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) \approx D_{k,h}/(d^3H),
\end{aligned} \tag{6.3}$$

where we approximate the optimal value function V^*_{h+1} by the *pessimistic* value function $\check{V}_{k,h}$ and introduce an extra d^3H-factor in the $D_{k,h}$ to offset the error caused by the covering number argument.

However, there exists another difficulty in our algorithm and analysis when we extend the result in (6.3) to future episode $i > k$. In detail, while the value function $V_{i,h+1}(s^k_{h+1})$ and corresponding variance $[\mathbb{V}_h(V_{i,h+1} - V^*_{h+1})](s_h^k, a_h^k)$ will change across different episodes, the estimated variance $\sigma_{k,h}$ is chosen at episode k and cannot be changed in the subsequent episode. Therefore, $\sigma_{k,h}$ should be a uniform variance upper bound for all subsequent episodes. To achieve such a uniform variance upper bound, it suffices to have the sub-optimality gap $V_{k,h+1} - V^*_{h+1}$ to be monotonically decreasing. Our solution is to set $V_{k,h+1}(s)$ to be a monotonically decreasing sequence in k given any state s, by setting it as the minimum between its current estimate and its predecessor $V_{k-1,h+1}$ (Line 9 of Algorithm 1). A similar approach is also applied to $\check{V}_{k,h+1}$ to guarantee the estimate sequence is monotonically increasing in k. Then, the following property shows that the estimated variance of the sub-optimality at episode k $D_{k,h}$ holds for all the subsequent episodes.

$$[\mathbb{V}_h(V_{i,h+1} - V^*_{h+1})](s_h^k, a_h^k) \le D_{k,h}/(d^3H), \forall i > k.$$

This idea was firstly introduced by Azar et al. (2017) for tabular MDPs. Hu et al. (2022) adopted a similar idea to guarantee the monotonicity for linear MDPs, while their approach is to construct another sequence of "over-optimistic" value functions, which turns out to be flawed as we will discuss in Appendix A.

6.3 Rare-Switching Value Function Update

As we discussed in Section 6.2, we ensure the monotonicity and construct the variance estimation, by taking minimization with its predecessor $V_{k-1,h+1}$. However, this approach will introduce an extra issue for the augmented value function class. In detail, the optimistic value function $V_{k,h}$ can be denoted by the minimum over several quadratic functions and belongs to the following function class,

$$\begin{aligned}
\mathcal{V}_h = \Bigg\{ V \Bigg| V(\cdot) = \max_a \min_{1\le i\le l} \min \Big(H, r_h(\cdot, a) + \mathbf{w}_i^\top \boldsymbol{\phi}(\cdot, a) \\
+ \beta\sqrt{\boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\phi}(\cdot, a)} \Big), \|\mathbf{w}_i\| \le L, \boldsymbol{\Sigma}_i \succeq \lambda \mathbf{I} \Bigg\},
\end{aligned}$$

where l is the number of quadratic functions and equals to the number of policy updates in Algorithm 1. Here, we denote the covering number of that function class by $\mathcal{N}$, and the covering number of a quadratic function class by $\mathcal{N}_q$. We are specifically interested in the covering entropy $\log \mathcal{N}$, which is a standard complexity measure of the function class, and it will directly affect the regret of our algorithm. The standard approach to computing the covering entropy suggests that $\log \mathcal{N} = l \log \mathcal{N}_q$. In this case, if we update the value function at each episode $k \in [K]$ and minimize with its predecessor, then there will be an extra K factor in the covering number, which is unacceptable. Inspired by the "rare-switching" technique (Abbasi-Yadkori et al., 2011; Wang et al., 2021), it is not necessary or efficient to update the value functions $V_{k,h+1}$ and $\check{V}_{k,h+1}$ at each episode. Instead, we only need to update the value function when the determinant of the covariance matrix grows much larger than before (Line 8 in Algorithm 1), which requires at most $\widetilde{O}(dH)$

updates. Such an update strategy reduces the covering entropy from $\log \mathcal{N} = K \log \mathcal{N}_q$ to $\log \mathcal{N} = \widetilde{O}(dH) \log \mathcal{N}_q$, which makes the regret of Algorithm 1 tight. In addition, the "rare-switching" nature also reduces the computational complexity from $\Omega(K^2)$ to $\Omega(K)$, which makes LSVI-UCB++ more efficient.

7 Conclusions and Future Work

In this paper, we propose a near-optimal algorithm LSVI-UCB++ for linear MDPs. LSVI-UCB++ is based on weighted ridge regression, where the weights are constructed from a novel variance estimator that comes from a direct estimation of the variance of the true value function, and a "rare-switching" updating rule to update the value function estimator. We prove that with high probability, LSVI-UCB++ obtains an $\widetilde{O}(d\sqrt{H^3K})$ regret, which matches the lower bound in Zhou et al. (2021a) up to logarithmic factors. Our algorithm is also computationally efficient. In the future, we will study how to design computationally efficient near-optimal RL algorithms with general nonlinear function approximation with misspecification.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. JH, HZ, DZ and QG are supported in part by the National Science Foundation CAREER Award 1906169 and the Sloan Research Fellowship. 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. In *NIPS*, vol. 11.

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.

AZAR, M. G., OSBAND, I. and MUNOS, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*. PMLR.

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

DANN, C., JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2018). On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems* **31**.

DU, S. S., KAKADE, S. M., WANG, R. and YANG, L. F. (2019). Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016* .

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). Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems* **34** 22288–22300.

HE, J., ZHOU, D., ZHANG, T. and GU, Q. (2022). Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2205.06811* .

HENDERSON, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 423–447.

HU, P., CHEN, Y. and HUANG, L. (2022). Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*. PMLR.

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.

JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*. PMLR.

JIN, C., ALLEN-ZHU, Z., BUBECK, S. and JORDAN, M. I. (2018). Is q-learning provably efficient? *Advances in neural information processing systems* **31**.

JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388* .

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.

KIRSCHNER, J. and KRAUSE, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*. PMLR.

LATTIMORE, T., CRAMMER, K. and SZEPESVÁRI, C. (2015). Linear multi-resource allocation with semi-bandit feedback. *Advances in Neural Information Processing Systems* **28**.

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*. PMLR.

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

SUN, W., JIANG, N., KRISHNAMURTHY, A., AGARWAL, A. and LANGFORD, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*. PMLR.

WANG, T., ZHOU, D. and GU, Q. (2021). Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems* **34** 13524–13536.

WANG, Y., WANG, R., DU, S. S. and KRISHNAMURTHY, A. (2020). Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*.

YANG, L. and WANG, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*.

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*. PMLR.

ZANETTE, A. and BRUNSKILL, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*. PMLR.

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, Z. and JI, X. (2019). Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems* **32**.

ZHANG, Z., JI, X. and DU, S. (2021a). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*. PMLR.

ZHANG, Z., YANG, J., JI, X. and DU, S. S. (2021b). Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems* **34** 4342–4355.

ZHANG, Z., ZHOU, Y. and JI, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems* **33** 15198–15207.

ZHOU, D. and GU, Q. (2022). Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *arXiv preprint arXiv:2205.11507* .

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 Comparison with Hu et al. (2022)

In this section, we give a detailed comparison with Hu et al. (2022). We first elaborate on the importance of the monotonic property, then discuss the issue on the over-optimistic value function $\dot{\widehat{V}}_{k,h}(s)$ proposed in Hu et al. (2022) and finally illustrate the difference in the algorithm design between the algorithm in Hu et al. (2022) and our algorithm.

As we discuss in the Section 6, both our LSVI-UCB++ algorithm and LSVI-UCB+ algorithm in Hu et al. (2022) get rid of the covering number issue by decomposing the value function $V_{k,h+1}(s)$ to $V^*_{h+1}(s)$ and $V_{k,h+1}(s) - V^*_{h+1}(s)$. In detail, from the proof of Lemma B.5, we have shown in (D.18) that the estimation error can be decomposed as

$$
\begin{aligned}
&\Big\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(V_{k,h+1}(s_{h+1}^i)-\mathbb{P}_hV_{k,h+1}(s_h^i,a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}}\\
&=\Big\|\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(V_{k,h+1}(s_{h+1}^i)-\mathbb{P}_hV_{k,h+1}(s_h^i,a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}\\
&\le\underbrace{\Big\|\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(V^*_{h+1}(s_{h+1}^i)-\mathbb{P}_hV^*_{h+1}(s_h^i,a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}}_{J_1}\\
&\quad+\underbrace{\Big\|\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(\Delta V_{k,h+1}(s_{h+1}^i)-[\mathbb{P}_h(\Delta V_{k,h+1})](s_h^i,a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}}_{J_2}.
\end{aligned}
$$

To control the concentration error on the term J_2, we use Lemma 4.1 and only need to estimate the variance $\mathbb{V}_h[\Delta V_{k,h+1}](s,a)=\mathbb{V}_h\big[(V_{k,h+1}-V^*_{h+1})\big](s,a)$, which is trivially upper bounded by $2H\cdot[\mathbb{P}_h\Delta V_{k,h+1}](s,a)$. In order to make the upper bound of variance at episode i hold for all subsequent episode $k>i$, we need to guarantee that the trivial upper bound $2H\cdot[\mathbb{P}_h\Delta V_{k,h+1}](s,a)$ is decreasing in k, which requires the optimistic value function $V_{k,h+1}(s)$ to be monotonically decreasing.

To satisfy this requirement, Hu et al. (2022) constructs an over-optimistic value function $\dot{\widehat{V}}_{k,h}(s)$, which has the following monotonicity property.

Lemma A.1 (Lemma D.2, Hu et al. 2022). *For any stage $h\in[H]$ and episodes $i<j$, the over-optimistic value function $\dot{\widehat{V}}_{j,h}(s)$ satisfies:*

$$\dot{\widehat{V}}_{i,h}(s)\ge\widehat{V}_{j,h}(s),$$

where $\widehat{V}_{j,h}(s)$ is the optimistic value function.

Based on this monotonically decreasing property, the estimation error $2H\cdot\big[\mathbb{P}_h(\dot{\widehat{V}}_{i,h+1}-V^*_{h+1})\big](s,a)$ is a uniform variance upper bound for all subsequent episodes. Unfortunately, the last inequality in the proof of Lemma A.1 claims that $[\mathbb{P}_hV_{i,h+1}](s,a)-[\mathbb{P}_hV_{j,h+1}](s,a)$ holds due to $\dot{\widehat{V}}_{i,h}(s)\ge\widehat{V}_{j,h}(s)$, which is not true. Thus, the monotonic property of over-optimistic value function $\dot{\widehat{V}}_{k,h}(s)$ does not hold and the estimated variance $\sigma_{i,h}$ may no longer be a variance upper bound for the subsequent episodes.

In comparison, our LSVI-UCB++ ensures the monotonic property by choosing the minimum of the optimistic value functions in the first k episodes (See Line 9). As the cost of ensuring monotonic property, the resulting value function class can be regarded as a minimum over K quadratic function classes, and the covering number grows exponentially in K. To overcome this problem, we utilize the 'rare-switching' technique from previous works (Abbasi-Yadkori et al., 2011; Wang et al., 2021), which reduces the number of updates to $\widetilde{O}(dH)$ and thus controls the complexity growth of the resulting value function class.

B Proof Sketch

This section is devoted to provide a proof sketch of Theorem 5.1.

B.1 High-Probability Events

We first introduce the following high-probability events:

1. We define $\mathcal{E}$ as the event that the following inequalities hold for all $s, a, k, h \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$.

$$\begin{aligned}
\big|\widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - [\mathbb{P}_h V_{k,h+1}](s,a)\big| &\le \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)},\\
\big|\widetilde{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - [\mathbb{P}_h V_{k,h+1}^2](s,a)\big| &\le \widetilde{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)},\\
\big|\check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - [\mathbb{P}_h \check{V}_{k,h+1}](s,a)\big| &\le \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)},
\end{aligned}$$

   where $\widetilde{\beta} = O\Big(H^2\sqrt{d\lambda} + \sqrt{d^3H^4\log^2\big(dHK/(\delta\lambda)\big)}\Big)$ and $\bar{\beta} = O\Big(H\sqrt{d\lambda} + \sqrt{d^3H^2\log^2\big(dHK/(\delta\lambda)\big)}\Big)$.

2. We define $\widetilde{\mathcal{E}}_h$ as the event such that for all episode $k \in [K]$, stage $h \le h' \le H$ and state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, the weight vector $\widehat{\mathbf{w}}_{k,h}$ satisfies that

$$\big|\widehat{\mathbf{w}}_{k,h'}^\top \boldsymbol{\phi}(s,a) - [\mathbb{P}_h V_{k,h'+1}](s,a)\big| \le \beta\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h'}^{-1}\boldsymbol{\phi}(s,a)}, \tag{B.1}$$

   where $\beta = O\Big(H\sqrt{d\lambda} + \sqrt{d\log^2\big(1 + dKH/(\delta\lambda)\big)}\Big)$. For simplicity, we further define events $\widetilde{\mathcal{E}} = \widetilde{\mathcal{E}}_1$ that (B.1) holds for all stage $h \in [H]$.

Our ultimate goal is to show that $\widetilde{\mathcal{E}}$ holds with high probability. Intuitively speaking, $\mathcal{E}$ serves as a 'coarse' event where the concentration results hold with a larger confidence radius $\widetilde{\beta}$ and $\bar{\beta}$, and $\widetilde{\mathcal{E}}$ serves as a 'refined' event where the confidence radius β is tighter than $\widetilde{\beta}$ and $\bar{\beta}$. To start with, the following lemma shows that $\mathcal{E}$ holds with high probability.

Lemma B.1. *Event $\mathcal{E}$ holds with probability at least $1 - 7\delta$.*

Next, we prove $\widetilde{\mathcal{E}} = \widetilde{\mathcal{E}}_1$ holds with high probability. Since $\widehat{\mathbf{w}}_{k,h}$'s are obtained from weighted linear regression whose weights depend on the variances of $V_{k,h+1}$, the key technical challenge is to show that our adapted weights $\sigma_{k,h}$'s are indeed upper bounds of these variances for all $h \in [H]$. We use backward induction to prove such a statement. In detail, the following two lemmas provide estimation error bounds at stage h conditioned on $\widetilde{\mathcal{E}}_{h+1}$.

Lemma B.2. *On the event $\mathcal{E}$ and $\widetilde{\mathcal{E}}_{h+1}$, for each episode $k \in [K]$ and stage h, the estimated variance satisfies*

$$\begin{aligned}
\big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\big| &\le E_{k,h},\\
\big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\big| &\le E_{k,h} + D_{k,h}.
\end{aligned}$$

Lemma B.3. *On the event $\mathcal{E}$ and $\widetilde{\mathcal{E}}_{h+1}$, for any episode k and $i > k$, we have*

$$[\mathbb{V}_h(V_{i,h+1} - V_{h+1}^*)](s_h^k, a_h^k) \le D_{k,h}/(d^3H).$$

We also have the following lemma, which shows that our constructed value functions Q, V, and $\check{Q}, \check{V}$ are optimistic and pessimistic estimators of the true value functions under the events we defined before.

Lemma B.4. *On the event $\mathcal{E}$ and $\widetilde{\mathcal{E}}_h$, for all episode $k \in [K]$ and stage $h \le h' \le H$, we have $Q_{k,h}(s,a) \ge Q_h^*(s,a) \ge \check{Q}_{k,h}(s,a)$. In addition, we have $V_{k,h}(s) \ge V_h^*(s) \ge \check{V}_{k,h}(s)$.*

Equipped with Lemmas B.2, B.3 and B.4, one can easily prove $\sigma_{k,h}$ indeed serves as an upper bound of the true variance of $V_{k,h+1}$ at stage h. Therefore, by the backward induction, we can prove the following lemma.

Lemma B.5. *On the events $\mathcal{E}$, event $\widetilde{\mathcal{E}}$ holds with probability at least $1 - \delta$.*

B.2 Regret Decomposition

Now, we prove the regret bound based on the high-probability events defined before. Based on Lemma B.4, for all stage $h \in [H]$ and episode $k \in [K]$, we have $Q_{k,h}(s_h^k, a_h^k) = V_{k,h}(s_h^k) \ge V_h^*(s_h^k)$. Thus, we can bound the regret as follows,

$$\textbf{Regret}(K) = \sum_{k=1}^{K}\big(V_1^*(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big)$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K} \big(V_{k,1}(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \Big\{ \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k, a_h^k) - \big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big) \Big\} \\
&\quad + O\left(\sum_{k=1}^{K}\sum_{h=1}^{H} \min\Big(\beta\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, H\Big)\right),
\end{aligned}
$$

where the last inequality holds due to the decomposition of the difference of value functions and the high-probability events defined in Section B.1. Using standard regret analysis, we can bound the first term as the sum of a martingale difference sequence. Then it remains to bound the summation of bonus terms, $\sum_{k=1}^{K}\sum_{h=1}^{H} \beta\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}$. By Cauchy-Schwartz inequality, this summation can be bounded by

$$
\sum_{k=1}^{K}\sum_{h=1}^{H} \beta\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \leq \widetilde{O}\left(d^4H^8 + \beta d^7 H^5 + \beta\sqrt{dHK + dH\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2}\right),
$$

where the calculation details are deferred to Lemma E.1. According to the definition of $\sigma_{k,h}^2$, we have $\sigma_{k,h}^2 \leq O\big([\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h} + D_{k,h} + H\big)$. By carefully bounding the summation of $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k)$ by relating them to the summation of $[\mathbb{V}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k)$ and using the total variance lemma (Lemma C.5, Jin et al. 2018), we have

$$
\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2 \leq \widetilde{O}\left(H^2K + d^{10.5}H^{16}\right).
$$

Putting all pieces together, we can obtain the high-probability regret bound

$$
\mathbf{Regret}(K) \leq \widetilde{O}\left(d\sqrt{H^3K} + d^7H^8\right).
$$

C Detailed Proof of Theorem 5.1

In this section, we provide the proof of Theorem 5.1. Firstly, for the stochastic transition noises, we define the following high-probability events:

$$
\begin{aligned}
\mathcal{E}_1 = \Bigg\{ &\forall h \in [H], \sum_{k=1}^{K}\sum_{h'=h}^{H} \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k, a_h^k) \\
&\qquad - \sum_{k=1}^{K}\sum_{h'=h}^{H} \big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big) \leq 2\sqrt{2H^3K\log(H/\delta)} \Bigg\}, \\
\mathcal{E}_2 = \Bigg\{ &\forall h \in [H], \sum_{k=1}^{K}\sum_{h'=h}^{H} \big[\mathbb{P}_h(V_{k,h+1} - \check{V}_{k,h+1})\big](s_h^k, a_h^k) \\
&\qquad - \sum_{k=1}^{K}\sum_{h'=h}^{H} \big(V_{k,h+1}(s_{h+1}^k) - \check{V}_{k,h+1}(s_{h+1}^k)\big) \leq 2\sqrt{2H^3K\log(H/\delta)} \Bigg\}.
\end{aligned}
$$

Then according to the Azuma–Hoeffding inequality (Lemma G.2), we have $\Pr(\mathcal{E}_1) \geq 1-\delta$ and $\Pr(\mathcal{E}_2) \geq 1-\delta$. Based on the definition of events $\mathcal{E}_1, \mathcal{E}_2$ and events $\mathcal{E}, \widetilde{\mathcal{E}}$ in Section B.1, the regret in the first K episodes can be upper bounded by the summation of estimated variance $\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2$ and we have the following lemma.

Lemma C.1. *On the events $\widetilde{\mathcal{E}}$, $\mathcal{E}$ and $\mathcal{E}_1$, for all stage $h \in [H]$, the regret in the first K episodes is upper bounded by:*

$$
\sum_{k=1}^{K}\big(V_{k,h}(s_h^k) - V_{k,h}^{\pi^k}(s_h^k)\big) \leq 16d^4H^8\iota + 40\beta d^7H^5\iota + 8\beta\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)} + 4\sqrt{H^3K\log(H/\delta)},
$$

and for all stage $h \in [H]$, we further have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k, a_h^k)$$
$$\leq 16d^4H^9\iota + 40\beta d^7H^6\iota + 8H\beta\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)} + 4\sqrt{H^5K\log(H/\delta)},$$

where $\iota = \log\big(1 + K/(d\lambda)\big)$.

In addition, for the sub-optimality gap between the optimistic value function $V_{k,h}(s)$ and pessimistic value function $\check{V}_{k,h}(s)$, we have the following lemma.

Lemma C.2. *On the events $\widetilde{\mathcal{E}}$, $\mathcal{E}$ and $\mathcal{E}_2$, the difference between the optimistic value function $V_{k,h}$ and the pessimistic value function $\check{V}_{k,h}$ is upper bounded by:*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\big[\mathbb{P}_h(V_{k,h+1} - \check{V}_{k,h+1})\big](s_h^k, a_h^k)$$
$$\leq 32d^4H^9\iota + 40(\beta+\bar{\beta})d^7H^6\iota + 8H(\beta+\bar{\beta})\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)} + 4\sqrt{H^5K\log(H/\delta)},$$

where $\iota = \log\big(1 + K/(d\lambda)\big)$.

For the summation of variance $\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k)$, we denote the following high probability events $\mathcal{E}_3$:

$$\mathcal{E}_3 = \bigg\{\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k) \leq 3H^2K + 3H^3\log(1/\delta)\bigg\}.$$

Then Lemma C.5 in Jin et al. (2018) shows that the probability of events $\mathcal{E}_3$ is lower bounded by $\Pr(\mathcal{E}_3) \geq 1 - \delta$. Furthermore, on the event $\mathcal{E} \cap \widetilde{\mathcal{E}} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, the following lemma gives an upper bound of the total estimated variance $\sum_{h=1}^{H}\sum_{k=1}^{K}\sigma_{k,h}^2$.

Lemma C.3. *On the event $\mathcal{E} \cap \widetilde{\mathcal{E}} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, the total estimated variance is upper bounded by:*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2 \leq O\big(H^2K + d^{10.5}H^{16}\log^{1.5}(1 + dKH/\delta)\big).$$

With all previous lemma, we start to prove our main Theorem 5.1.

Proof of Theorem 5.1. On the event $\mathcal{E} \cap \widetilde{\mathcal{E}} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,, the regret is upper bounded by:

$$\begin{aligned}
\textbf{Regret}(K) &= \sum_{k=1}^{K}\big(V_1^*(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big)\\
&\leq \sum_{k=1}^{K}\big(V_{k,1}(s_1^k) - V_{k,1}^{\pi^k}(s_1^k)\big)\\
&\leq 16d^4H^8\iota + 40\beta d^7H^5\iota + 8\beta\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)} + 4\sqrt{H^3K\log(H/\delta)}\\
&= \widetilde{O}\Big(d^7H^8 + d\sqrt{H^3K\log^2(1 + dKH/\delta)}\Big), \qquad \text{(C.1)}
\end{aligned}$$

where $\iota = \log\big(1 + K/(d\lambda)\big)$, the first inequality holds due to Lemma B.4, the second inequality holds due to Lemma C.2 and the last inequality holds due to Lemma C.3. Since the event $\mathcal{E} \cap \widetilde{\mathcal{E}} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ holds with probability at least $1 - 7\delta$, (C.1) holds. In addition, according to Lemma F.1, the number of updates for $Q_{k,h}, \check{Q}_{k,h}$ is upper bounded by $O(dH\log(1 + K/\lambda))$. Thus, we complete the proof of Theorem 5.1. updates for $Q_{k,h}, \check{Q}_{k,h}$ is upper bounded by $O(dH\log(1 + K/\lambda))$. □

D Proof of Lemmas in Section B

In this section, we provide the proof of Lemmas in Section B and we need the following lemma, which extends Lemma D.4 in Jin et al. (2020) to weighted ridge regression.

Lemma D.1. *(Lemma D.4, Jin et al. 2020 with weighted linear regression) Let $\{x_k\}_{k=1}^{\infty}$ be a real-valued stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$. Let $\{\boldsymbol{\phi}_k\}_{k=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process, where $\boldsymbol{\phi}_k \in \mathcal{F}_{k-1}$ and $\|\boldsymbol{\phi}_k\|_2 \le 1$. Let $\{w_k\}_{k=1}^{\infty}$ be an real-valued stochastic process where $w_k \in \mathcal{F}_{k-1}$ and $0 \le w_k \le C$. For any $k \ge 0$, we define $\boldsymbol{\Sigma}_k = \lambda\mathbf{I} + \sum_{i=1}^{k} w_k^2 \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$. Then with probability at least $1 - \delta$, for all $k \in \mathbb{N}$ and all function $V \in \mathcal{V}$ with $\max_s |V(x)| \le H$, we have*

$$\bigg\| \sum_{i=1}^{k} w_i^2 \boldsymbol{\phi}_i \Big\{ V(x_i) - \mathbb{E}\big[V(x_i)|\mathcal{F}_{i-1}\big] \Big\} \bigg\|_{\boldsymbol{\Sigma}_k^{-1}}^2 \le 4C^2H^2 \bigg[\frac{d}{2}\log(1 + kC^2/\lambda) + \log\frac{\mathcal{N}_\epsilon}{\delta} \bigg] + 8k^2C^4\epsilon^2/\lambda,$$

where $\mathcal{N}_\epsilon$ is the ϵ-covering number of the function class $\mathcal{V}$ with respect to the distance function dist$(V_1, V_2) = \max_s |V_1(s) - V_2(s)|$.

Proof of Lemma D.1. For any function $V \in \mathcal{V}$, based on the definition of ϵ-covering number, there exists a function $\widetilde{V}$ in the ϵ-net, such that

$$\text{dist}(V, \widetilde{V}) \le \epsilon. \tag{D.1}$$

Therefore, the concentration error for the value function can be decomposed as

$$\begin{aligned}
&\bigg\| \sum_{i=1}^{k} w_i^2 \boldsymbol{\phi}_i \Big\{ V(x_i) - \mathbb{E}\big[V(x_i)|\mathcal{F}_{i-1}\big] \Big\} \bigg\|_{\boldsymbol{\Sigma}_k^{-1}}^2 \\
&\le \underbrace{2\bigg\| \sum_{i=1}^{k} w_i^2 \boldsymbol{\phi}_i \Big\{ \widetilde{V}(x_i) - \mathbb{E}\big[\widetilde{V}(x_i)|\mathcal{F}_{i-1}\big] \Big\} \bigg\|_{\boldsymbol{\Sigma}_k^{-1}}^2}_{I_1} + \underbrace{2\bigg\| \sum_{i=1}^{k} w_i^2 \boldsymbol{\phi}_i \Big\{ \Delta_V(x_i) - \mathbb{E}\big[\Delta_V(x_i)|\mathcal{F}_{i-1}\big] \Big\} \bigg\|_{\boldsymbol{\Sigma}_k^{-1}}^2}_{I_2},
\end{aligned} \tag{D.2}$$

where $\Delta_V = V - \widetilde{V}$ and the inequality holds due to $\|\mathbf{a} + \mathbf{b}\|_{\boldsymbol{\Sigma}}^2 \le 2\|\mathbf{a}\|_{\boldsymbol{\Sigma}}^2 + 2\|\mathbf{b}\|_{\boldsymbol{\Sigma}}^2$. For any fixed value function $\widetilde{V}$, we apply Lemma G.5 with $\mathbf{x}_i = w_i\boldsymbol{\phi}_i, \eta_t = w_i\widetilde{V}(x_i) - w_i\mathbb{E}[\widetilde{V}(x_i)]$. According to the definition of $\mathbf{x}_i, \eta_i$, we have following property

$$\begin{aligned}
\|\mathbf{x}_i\|_2 &= w_i\boldsymbol{\phi}_i \le C, \\
\mathbb{E}[\eta_i|\mathcal{F}_i] = 0, |\eta_i| &= \big|w_i\widetilde{V}(x_i) - w_i\mathbb{E}[\widetilde{V}(x_i)]\big| \le HC.
\end{aligned}$$

Therefore, according to Lemma G.5, after taking an union bound over the ϵ-net of the function class $\mathcal{V}$, with probability at least $1 - \delta/H$, the first term I_1 is upper bounded by:

$$\begin{aligned}
I_1 &= \bigg\| 2\sum_{i=1}^{k} w_i^2 \boldsymbol{\phi}_i \Big\{ \widetilde{V}(x_i) - \mathbb{E}\big[\widetilde{V}(x_i)|\mathcal{F}_{i-1}\big] \Big\} \bigg\|_{\boldsymbol{\Sigma}_k^{-1}}^2 \\
&\le 4C^2H^2 \bigg[\frac{d}{2}\log(1 + kC^2/\lambda) + \log\frac{\mathcal{N}_\epsilon}{\delta} \bigg].
\end{aligned} \tag{D.3}$$

For the second term, it can be upper bounded by

$$I_2 = 2\bigg\| \sum_{i=1}^{k} w_i^2 \boldsymbol{\phi}_i \Big\{ \Delta_V(x_i) - \mathbb{E}\big[\Delta_V(x_i)|\mathcal{F}_{i-1}\big] \Big\} \bigg\|_{\boldsymbol{\Sigma}_k^{-1}}^2$$

$$\leq 2k\sum_{i=1}^{k}\Big\|w_i^2\boldsymbol{\phi}_i\Big\{\Delta_V(x_i)-\mathbb{E}\big[\Delta_V(x_i)|\mathcal{F}_{i-1}\big]\Big\}\Big\|^2_{\boldsymbol{\Sigma}_k^{-1}}$$
$$\leq 8k^2C^4\epsilon^2/\lambda, \tag{D.4}$$

where the first inequality holds due to the Cauchy-Schwartz inequality and the last inequality holds due to the facts that $|\Delta V(x_i)| \leq \epsilon, 0 \leq w_i \leq C, \|\boldsymbol{\phi}_i\|_2 \leq 1, \boldsymbol{\Sigma}_k \succeq \lambda\mathbf{I}$. Substituting the results in (D.3) and (D.4) into (D.2), we finish the proof of Lemma D.1. □

D.1 Proof of Lemma B.1

In this subsection, we provide the proof of Lemma B.1, which suggests a Hoeffding-type upper bound for the estimation error.

Proof of Lemma B.1. Firstly, for any fixed stage $h \in [H]$ and the optimistic value function $V_{k,h+1}$, according to Lemma G.1, there exists a vector $\mathbf{w}_{k,h}$ such that $\mathbb{P}_hV_{k,h+1}(s,a)$ can be represented by $\mathbf{w}_{k,h}^\top\boldsymbol{\phi}(s,a)$ and $\|\mathbf{w}_{k,h}\|_2 \leq H\sqrt{d}$. Therefore, the estimation error can be decomposed as

$$\begin{aligned}
&\|\widehat{\mathbf{w}}_{k,h}-\mathbf{w}_{k,h}\|_{\boldsymbol{\Sigma}_{k,h}}\\
&=\Big\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)V_{k,h+1}(s_{h+1}^i)-\boldsymbol{\Sigma}_{k,h}^{-1}\Big(\lambda\mathbf{I}+\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\boldsymbol{\phi}(s_h^i,a_h^i)^\top\Big)\mathbf{w}_{k,h}\Big\|_{\boldsymbol{\Sigma}_{k,h}}\\
&=\Big\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(V_{k,h+1}(s_{h+1}^i)-[\mathbb{P}_hV_{k,h+1}](s_h^i,a_h^i)\big)-\lambda\boldsymbol{\Sigma}_{k,h}^{-1}\mathbf{w}_{k,h}\Big\|_{\boldsymbol{\Sigma}_{k,h}}\\
&\leq\underbrace{\|\lambda\boldsymbol{\Sigma}_{k,h}^{-1}\mathbf{w}_{k,h}\|_{\boldsymbol{\Sigma}_{k,h}}}_{I_1}+\underbrace{\Big\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(V_{k,h+1}(s_{h+1}^i)-[\mathbb{P}_hV_{k,h+1}](s_h^i,a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}}}_{I_2},
\end{aligned}\tag{D.5}$$

where the first inequality holds due to the fact that $\|\mathbf{a}+\mathbf{b}\|_{\boldsymbol{\Sigma}} \leq \|\mathbf{a}\|_{\boldsymbol{\Sigma}}+\|\mathbf{b}\|_{\boldsymbol{\Sigma}}$. For the first term I_1, since $\boldsymbol{\Sigma}_{k,h} \succeq \lambda\mathbf{I}$ and $\|\mathbf{w}_{k,h}\|_2 \leq H\sqrt{d}$, it is upper bounded by

$$I_1=\|\lambda\mathbf{w}_{k,h}\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \leq \sqrt{\lambda}\cdot\|\mathbf{w}_{k,h}\|_2 \leq H\sqrt{d\lambda}. \tag{D.6}$$

For the second term I_2, we apply Lemma D.1 with the optimistic value function class $\mathcal{V}_h$ and $\epsilon = H\sqrt{\lambda}/K$, then for any fixed stage $h \in [H]$, with probability at least $1-\delta/H$, for all episode $k \in [K]$, we have

$$\begin{aligned}
I_2&=\Big\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(V_{k,h+1}(s_{h+1}^i)-[\mathbb{P}_hV_{k,h+1}](s_h^i,a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}}\\
&\leq\sqrt{4C^2H^2\bigg[\frac{d}{2}\log(1+kC^2/\lambda)+\log\frac{H\mathcal{N}_\epsilon}{\delta}\bigg]+8k^2C^4\epsilon^2/\lambda}\\
&\leq\sqrt{4H\bigg[\frac{d}{2}\log\big(1+k/(\lambda H)\big)+\log\frac{H\mathcal{N}_\epsilon}{\delta}\bigg]+8k^2\epsilon^2/(\lambda H^2)}\\
&\leq\sqrt{4H\bigg[\frac{d}{2}\log\big(1+k/(\lambda H)\big)+\log\frac{H\mathcal{N}_\epsilon}{\delta}\bigg]+8}\\
&=O\bigg(\sqrt{d^3H^2\log^2\big(dHK/(\delta\lambda)\big)}\bigg),
\end{aligned}\tag{D.7}$$

where the first inequality holds due to Lemma D.1, the second inequality holds due to $0 \leq \bar{\sigma}_{i,h}^{-1} \leq 1/\sqrt{H}$, the third inequality holds due to Lemma F.6 and $\epsilon = H\sqrt{\lambda}/K$. Substituting (D.6) and (D.7) into (D.5), we have

$$\|\widehat{\mathbf{w}}_{k,h}-\mathbf{w}_{k,h}\|_{\boldsymbol{\Sigma}_{k,h}} \leq I_1+I_2=O\Big(H\sqrt{d\lambda}+\sqrt{d^3H^2\log^2\big(dHK/(\delta\lambda)\big)}\Big)=\bar{\beta}. \tag{D.8}$$

Therefore, the estimation error is upper bounded by

$$\begin{aligned}\big|\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s,a)-[\mathbb{P}_hV_{k,h+1}](s,a)\big| &= |\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s,a)-\mathbf{w}_{k,h}^\top\boldsymbol{\phi}(s,a)|\\ &\le \|\widehat{\mathbf{w}}_{k,h}-\mathbf{w}_{k,h}\|_{\boldsymbol{\Sigma}_{k,h}}\cdot\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}\\ &\le \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)},\end{aligned}\tag{D.9}$$

where the first inequality holds due to Cauchy-Schwartz inequality and the last inequality holds due to (D.8). Replacing the value function class by the pessimistic value function class $\check{\mathcal{V}}_h$ (or squared value function class $\mathcal{V}_h^2$) and following the same proof of (D.9), we can derive the following upper bound for the estimation errors:

$$\begin{aligned}\big|\widetilde{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s,a)-[\mathbb{P}_hV_{k,h+1}^2](s,a)\big| &\le \widetilde{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)},\\ \big|\check{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s,a)-[\mathbb{P}_h\check{V}_{k,h+1}](s,a)\big| &\le \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)},\end{aligned}$$

where $\widetilde{\beta}=O\Big(H\sqrt{d\lambda}+\sqrt{d^3H^4\log^2\big(dHK/(\delta\lambda)\big)}\Big)$ and $\bar{\beta}=O\Big(H\sqrt{d\lambda}+\sqrt{d^3H^2\log^2\big(dHK/(\delta\lambda)\big)}\Big)$ Thus, we finish the proof of Lemma B.1. □

D.2 Proof of Lemma B.2

In this subsection, we provide the proof of Lemma B.2 for the variance estimator.

Proof of Lemma B.2. Firstly, according to Lemma B.1, we have

$$\begin{aligned}&\big|[\bar{\mathbb{V}}_hV_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)\big|\\ &=\Big|\big[\widetilde{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k,a_h^k)\big]_{[0,H^2]}-\big[\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k,a_h^k)\big]_{[0,H]}^2-[\mathbb{P}_hV_{k,h+1}^2](s_h^k,a_h^k)+\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\big)^2\Big|\\ &\le\Big|\big[\widetilde{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k,a_h^k)\big]_{[0,H^2]}-[\mathbb{P}_hV_{k,h+1}^2](s_h^k,a_h^k)\Big|+\Big|\big[\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k,a_h^k)\big]_{[0,H]}^2-\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\big)^2\Big|\\ &=\Big|\big[\widetilde{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k,a_h^k)\big]_{[0,H^2]}-[\mathbb{P}_hV_{k,h+1}^2](s_h^k,a_h^k)\Big|\\ &\quad+\Big|\big[\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k,a_h^k)\big]_{[0,H]}+[\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\Big|\cdot\Big|\big[\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k,a_h^k)\big]_{[0,H]}-[\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\Big|\\ &\le\min\Big\{\widetilde{\beta}_k\big\|\boldsymbol{\Sigma}_{k,h}^{-1/2}\boldsymbol{\phi}(s_h^k,a_h^k)\big\|_2,H^2\Big\}+\min\Big\{2H\bar{\beta}_k\big\|\boldsymbol{\Sigma}_{k,h}^{-1/2}\boldsymbol{\phi}(s_h^k,a_h^k)\big\|_2,H^2\Big\}\\ &=E_{k,h},\end{aligned}\tag{D.10}$$

where the first inequality holds due to $|a+b|\le|a|+|b|$ and the last inequality holds due to Lemma B.1 with the fact that $0\le\big[\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s_h^k,a_h^k)\big]_{[0,H]}+[\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\le 2H$. In addition, for the variance $\big[\mathbb{V}_hV_{h+1}^*\big](s_h^k,a_h^k)$, we have

$$\begin{aligned}&\big|[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{h+1}^*](s_h^k,a_h^k)\big|\\ &=\Big|[\mathbb{P}_hV_{k,h+1}^2](s_h^k,a_h^k)-\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\big)^2-[\mathbb{P}_h(V_{h+1}^*)^2](s_h^k,a_h^k)+\big([\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\big)^2\Big|\\ &\le\big|[\mathbb{P}_hV_{k,h+1}^2](s_h^k,a_h^k)-[\mathbb{P}_h(V_{h+1}^*)^2](s_h^k,a_h^k)\big|+\Big|\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\big)^2-\big([\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\big)^2\Big|\\ &=\big|[\mathbb{P}_h(V_{k,h+1}-V_{h+1}^*)(V_{k,h+1}+V_{h+1}^*)](s_h^k,a_h^k)\big|\\ &\quad+\Big|\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)-[\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\big)\cdot\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)+[\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\big)\Big|\\ &\le 4H\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)-[\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\big),\end{aligned}\tag{D.11}$$

where the first inequality holds due to $|a+b|\le|a|+|b|$ and the last inequality holds due to Lemma B.4 ($V_{k,h+1}(s')\ge V_{h+1}^*(s')$) with the fact that $0\le V_{h+1}^*(s'),V_{k,h+1}(s')\le H$. Based on the event $\mathcal{E}$ and $\widetilde{\mathcal{E}}_{h+1}$, (D.11) can be further bounded by

$$\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)-[\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\big)$$

$$
\begin{aligned}
&\leq \big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h \check{V}_{k,h+1}](s_h^k, a_h^k)\big) \\
&\leq \widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) + \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)} - \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) + \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)}, \qquad \text{(D.12)}
\end{aligned}
$$

where the first inequality holds due to Lemma B.4 ($V_{h+1}^*(s') \geq \check{V}_{k,h+1}(s')$) and the last inequality holds due to the definition of events $\mathcal{E}$. Combining the results in (D.10), (D.11) and (D.10), we have

$$
\begin{aligned}
&\big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\big| \\
&\leq \big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\big| + \big|[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\big| \\
&\leq E_{k,h} + 4H\Big(\widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) + \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)} - \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) + \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s,a)}\Big).
\end{aligned}
$$

In addition, since the value functions $V_{k,h}(s)$ and $V_{h+1}^*(s)$ is upper bounded by H, we have $\big|[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\big|$ which implies that

$$
\begin{aligned}
&\big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\big| \\
&\leq \big|[\bar{\mathbb{V}}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\big| + \big|[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\big| \\
&\leq E_{k,h} + H^2.
\end{aligned}
$$

Thus, we finish the proof of Lemma B.2. □

D.3 Proof of Lemma B.3

In this subsection, we provide the proof of Lemma B.3 for the variance estimator.

Proof of Lemma B.3. On the event $\mathcal{E}$ and $\widetilde{\mathcal{E}}_{h+1}$, we have

$$
\begin{aligned}
&\big[\mathbb{V}_h(V_{i,h+1} - V_{h+1}^*)\big](s_h^k, a_h^k) \\
&\leq [\mathbb{P}_h(V_{i,h+1} - V_{h+1}^*)^2](s_h^k, a_h^k) \\
&\leq 2H\big[\mathbb{P}_h(V_{i,h+1} - V_{h+1}^*)\big](s_h^k, a_h^k) \\
&\leq 2H\big([\mathbb{P}_h V_{i,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h \check{V}_{k,h+1}](s_h^k, a_h^k)\big) \\
&\leq 2H\big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h \check{V}_{k,h+1}](s_h^k, a_h^k)\big) \\
&\leq 2H\Big(\widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s_h^k, a_h^k) + \bar{\beta}\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)} - \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s_h^k, a_h^k) + \bar{\beta}\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}\Big),
\end{aligned}
$$

where the first inequality holds due to $\mathrm{Var}(x) \leq \mathbb{E}[x^2]$, the second and third inequality holds due to Lemma B.4 with the fact that $0 \leq V_{i,h+1}(s'), V_{h+1}^*(s') \leq H$, the fourth inequality the fact $V_{k,h+1} \geq V_{i,h+1}$ from the update-rule in Algorithm 1 and the fifth inequality holds due to Lemma B.1. On the other hand, since value function $0 \leq V_{i,h+1}(s'), V_{h+1}^*(s') \leq H$, we have

$$
\big[\mathbb{V}_h(V_{i,h+1} - V_{h+1}^*)\big](s_h^k, a_h^k) \leq H^2 = (d^3H^3)/(d^3H).
$$

Thus, we finish the proof of Lemma B.3. □

D.4 Proof of Lemma B.4

In this subsection, we provide proof of optimistic property.

Proof of Lemma B.4. We prove this lemma by induction. First, we prove the base case for the last stage $H+1$. Under this situation, for all state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, we have $Q_{k,H+1}(s,a) = Q_h^*(s,a) = \check{Q}_{k,h}(s,a) = 0$ and $V_{k,h}(s) \geq V_h^*(s) \geq \check{V}_{k,h}(s) = 0$. Thus, the results in Lemma B.4 holds for stage $H+1$.

Now, we focus on stage $h+1$. Since events $\widetilde{\mathcal{E}}_h$ directly implies the events $\widetilde{\mathcal{E}}_{h+1}$, according to the reduction assumption, we have

$$
V_{k,h+1}(s) \geq V_{h+1}^*(s) \geq \check{V}_{k,h}(s). \qquad \text{(D.13)}
$$

Thus, for all episode $k \in [K]$, we have

$$r_h(s,a) + \widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) + \beta\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)} - Q_h^*(s,a) \geq \big[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^*)\big](s,a) \geq 0,$$

where the first inequality holds due to the definition of events $\widetilde{\mathcal{E}}_h$ and the second inequality holds due to (D.13). Furthermore, the optimal value function is upper bounded by $Q_h^*(s,a) \leq H$ and it implies that

$$Q_h^*(s,a) \leq \min\Big\{\min_{1\leq i\leq k} r_h(s,a) + \widehat{\mathbf{w}}_{i,h}^\top \boldsymbol{\phi}(s,a) + \beta\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{i,h}^{-1}\boldsymbol{\phi}(s,a)}, H\Big\} \leq Q_{k,h}(s,a). \tag{D.14}$$

With a similar argument, for the pessimistic action-value function $\check{Q}_{k,h}(s,a)$, we have

$$r_h(s,a) + \check{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)} - Q_h^*(s,a) \leq \big[\mathbb{P}_h(\check{V}_{k,h+1} - V_{h+1}^*)\big](s,a) \leq 0.$$

Since the optimal value function is lower bounded by $Q_h^*(s,a) \geq 0$, the result further implies that

$$Q_h^*(s,a) \geq \max\Big\{\max_{1\leq i\leq k} r_h(s,a) + \widehat{\mathbf{w}}_{k_{\text{last}},h}^\top \boldsymbol{\phi}(s,a) + \beta\sqrt{\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Sigma}_{k_{\text{last}},h}^{-1}\boldsymbol{\phi}(s,a)}, 0\Big\} \geq \check{Q}_{k,h}(s,a). \tag{D.15}$$

In addition, for the value function V, we have

$$V_{k,h}(s) = \max_a Q_{i,h}(s,a) \geq \min_{1\leq i\leq k}\max_a Q_h^*(s,a) = V_h^*(s),$$
$$\check{V}_{k,h}(s) = \max_a Q_{i,h}(s,a) \leq \max_{1\leq i\leq k}\max_a Q_h^*(s,a) = V_h^*(s),$$

where the first inequality holds due to (D.14) and the second inequality holds due to (D.15). Thus, by induction, we finish the proof of Lemma B.4. □

D.5 Proof of Lemma B.5

In this subsection, we provide the proof of Lemma B.5, which suggests a Bernstein-type upper bound for the estimation error.

Proof of Lemma B.5. We prove Lemma B.5 by induction. First, we prove the base case for the last stage H. Under this situation, the weight vector $\widehat{\mathbf{w}}_{k,h} = 0$ and $V_{k,h+1}(s,a) = 0$. Thus, the result in Lemma B.5 holds for stage H.

For stage $h \in [H]$ and $k \in [K]$, according to Lemma G.1, there exists a vector $\mathbf{w}_{k,h}$ such that $\mathbb{P}_h V_{k,h+1}(s,a)$ can be represented by $\mathbf{w}_{k,h}^\top \boldsymbol{\phi}(s,a)$ and $\|\mathbf{w}_{k,h}\|_2 \leq H\sqrt{d}$. Conditioned on the event $\widetilde{\mathcal{E}}_{h+1}$, the estimation error can be decomposed as

$$\begin{aligned}
&\|\widehat{\mathbf{w}}_{k,h} - \mathbf{w}_{k,h}\|_{\boldsymbol{\Sigma}_{k,h}} \\
&= \bigg\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)V_{k,h+1}(s_{h+1}^i) - \boldsymbol{\Sigma}_{k,h}^{-1}\Big(\lambda\mathbf{I} + \sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\boldsymbol{\phi}(s_h^i,a_h^i)^\top\Big)\mathbf{w}_{k,h}\bigg\|_{\boldsymbol{\Sigma}_{k,h}} \\
&= \bigg\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(V_{k,h+1}(s_{h+1}^i) - [\mathbb{P}_h V_{k,h+1}](s_h^i,a_h^i)\big) - \lambda\boldsymbol{\Sigma}_{k,h}^{-1}\mathbf{w}_{k,h}\bigg\|_{\boldsymbol{\Sigma}_{k,h}} \\
&\leq \underbrace{\|\lambda\boldsymbol{\Sigma}_{k,h}^{-1}\mathbf{w}_{k,h}\|_{\boldsymbol{\Sigma}_{k,h}}}_{I_1} + \underbrace{\bigg\|\boldsymbol{\Sigma}_{k,h}^{-1}\sum_{i=1}^{k-1}\bar{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i,a_h^i)\big(V_{k,h+1}(s_{h+1}^i) - [\mathbb{P}_h V_{k,h+1}](s_h^i,a_h^i)\big)\bigg\|_{\boldsymbol{\Sigma}_{k,h}}}_{I_2},
\end{aligned} \tag{D.16}$$

where the first inequality holds due to the fact that $\|\mathbf{a}+\mathbf{b}\|_{\boldsymbol{\Sigma}} \leq \|\mathbf{a}\|_{\boldsymbol{\Sigma}} + \|\mathbf{b}\|_{\boldsymbol{\Sigma}}$. For the first term I_1, since $\boldsymbol{\Sigma}_{k,h} \succeq \lambda\mathbf{I}$ and $\|\mathbf{w}_{k,h}\|_2 \leq H\sqrt{d}$, it is upper bounded by

$$I_1 = \|\lambda\mathbf{w}_{k,h}\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \leq \sqrt{\lambda}\cdot\|\mathbf{w}_{k,h}\|_2 \leq H\sqrt{d\lambda}. \tag{D.17}$$

For the second term I_2, we have

$$
\begin{aligned}
I_2 &= \Big\| \boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \big(V_{k,h+1}(s_{h+1}^i) - [\mathbb{P}_h V_{k,h+1}](s_h^i, a_h^i) \big) \Big\|_{\boldsymbol{\Sigma}_{k,h}} \\
&= \Big\| \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \big(V_{k,h+1}(s_{h+1}^i) - [\mathbb{P}_h V_{k,h+1}](s_h^i, a_h^i) \big) \Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&\le \underbrace{\Big\| \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \big(V_{h+1}^*(s_{h+1}^i) - [\mathbb{P}_h V_{h+1}^*](s_h^i, a_h^i) \big) \Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}}_{J_1} \\
&\quad + \underbrace{\Big\| \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \Big(\Delta V_{k,h+1}(s_{h+1}^i) - \big[\mathbb{P}_h(\Delta V_{k,h+1})\big](s_h^i, a_h^i) \Big) \Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}}}_{J_2},
\end{aligned} \tag{D.18}
$$

where $\Delta V_{k,h+1} = V_{k,h+1} - V_{h+1}^*$.

For the term J_1, we apply Lemma 4.1 with $\mathbf{x}_i = \bar{\sigma}_{i,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i)$ and $\eta_i = \mathbb{1}\big\{ [\mathbb{V}_h V_{h+1}^*](s_h^i, a_h^i) \le \bar{\sigma}_{i,h}^2 \big\} \cdot \bar{\sigma}_{i,h}^{-1} \big(V_{h+1}^*(s_{h+1}^i) - [\mathbb{P}_h V_{h+1}^*](s_h^i, a_h^i) \big)$. For $\mathbf{x}_t, \eta_t$, we have the following property:

$$
\begin{aligned}
\|\mathbf{x}_i\|_2 &= \big\| \bar{\sigma}_{i,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i) \big\|_2 \le \big\| \boldsymbol{\phi}(s_h^i, a_h^i) \big\|_2 / \sqrt{H} \le 1/\sqrt{H}, \\
\mathbb{E}[\eta_i | \mathcal{F}_i] &= 0, |\eta_t| \le \Big| \bar{\sigma}_{i,h}^{-1} \big(V_{h+1}^*(s_{h+1}^i) - [\mathbb{P}_h V_{h+1}^*](s_h^i, a_h^i) \big) \Big| \le \sqrt{H}, \\
\mathbb{E}[\eta_i^2 | \mathcal{F}_i] &= \mathbb{E}\Big[\mathbb{1}\big\{ [\mathbb{V}_h V_{h+1}^*](s_h^i, a_h^i) \le \bar{\sigma}_{i,h}^2 \big\} \cdot \bar{\sigma}_{i,h}^{-2} [\mathbb{V}_h V_{h+1}^*](s_h^i, a_h^i) \Big] \le 1, \\
\max_i \big\{ |\eta_i| \cdot \min\{1, \|\mathbf{x}_i\|_{\boldsymbol{\Sigma}_{i,h}^{-1}}\} \big\} &\le 2H \bar{\sigma}_{i,h}^{-1} \|\mathbf{x}_i\|_{\boldsymbol{\Sigma}_{i,h}^{-1}} \le \sqrt{d}.
\end{aligned}
$$

Thus, with probability at least $1 - \delta/H$, for all $k \in [K]$, we have

$$
\Big\| \sum_{i=1}^{k-1} \mathbf{x}_i \eta_i \Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \le O\Big(\sqrt{d \log^2 \big(1 + dKH/(\delta\lambda)\big)} \Big).
$$

In addition, on the event $\widetilde{\mathcal{E}}_{h+1}$ and $\mathcal{E}$, according to Lemma B.2, we have

$$
\bar{\sigma}_{k,h}^2 \ge [\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) + E_{k,h} + D_{k,h} \ge [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k),
$$

which further implies that

$$
\begin{aligned}
J_1 &= \Big\| \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \big(V_{h+1}^*(s_{h+1}^i) - [\mathbb{P}_h V_{h+1}^*](s_h^i, a_h^i) \big) \Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&= \Big\| \sum_{i=1}^{k-1} \mathbf{x}_i \eta_i \Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&\le O\Big(\sqrt{d \log^2 \big(1 + dKH/(\delta\lambda)\big)} \Big).
\end{aligned} \tag{D.19}
$$

For the term J_2, we can not directly use Lemma 4.1, Since the stochastic noise $\big(\Delta V_{k,h+1}(s_{h+1}^i) - [\mathbb{P}_h(\Delta V_{k,h+1})](s_h^i, a_h^i) \big)$ is not $\mathcal{F}_{i+1}$ measurable. Thus, we need to use the ϵ-net covering argument. In detail, for each episode, i, the value function $V_{i,h}$ belongs to the optimistic value function class $\mathcal{V}$. If we set $\epsilon = \sqrt{\lambda}/(4H^2 d^2 K)$, then according to Lemma F.6, the covering entropy for function class $\mathcal{V} - V_{h+1}^*$ is upper bounded by

$$
\log \mathcal{N}_\epsilon \le O\big(d^3 H^2 \log^2(dHK/\lambda) \big). \tag{D.20}
$$

Then for function $V_{k,h}$, there must exists a function $\widetilde{V}$ in the ϵ-net, such that

$$\text{dist}(\Delta V_{k,h}, \widetilde{V}) \leq \epsilon. \tag{D.21}$$

Therefore, the variance of function $\widetilde{V}$ is upper bounded by

$$\begin{aligned}
&[\mathbb{V}_h \widetilde{V}](s_h^i, a_h^i) - \big[\mathbb{V}_h(\Delta V_{k,h+1})\big](s_h^i, a_h^i) \\
&= [\mathbb{P}_h \widetilde{V}^2](s_h^i, a_h^i) - \big[\mathbb{P}_h(\Delta V_{k,h+1})^2\big](s_h^i, a_h^i) + \Big(\big[\mathbb{P}_h(\Delta V_{k,h+1})\big](s_h^i, a_h^i)\Big)^2 - \big(\mathbb{P}_h \widetilde{V}(s_h^i, a_h^i)\big)^2 \\
&\leq 2\text{dist}(\Delta V_{k,h}, \widetilde{V}) \cdot \max_{s'} |\Delta V_{k,h+1} + \widetilde{V}|(s') \\
&\leq 4H \cdot \text{dist}(\Delta V_{k,h}, \widetilde{V}) \\
&\leq 1/d^2,
\end{aligned} \tag{D.22}$$

where the first inequality holds due to the definition of distance between different functions, the third inequality holds since $|\Delta V_{k,h+1}(s') + \widetilde{V}(s')| \leq 2H$ and the last inequality holds due to the definition of ϵ-net. Thus, we apply Lemma 4.1 with $\mathbf{x}_i = \bar{\sigma}_{i,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i)$ and $\eta_i = \mathbb{1}\big\{[\mathbb{V}_h \widetilde{V}](s_h^i, a_h^i) \leq \bar{\sigma}_{i,h}^2/(d^3H)\big\} \cdot \bar{\sigma}_{i,h}^{-1}\big(\widetilde{V}(s_{h+1}^i) - [\mathbb{P}_h \widetilde{V}](s_h^i, a_h^i)\big)$. Therefore, For $\mathbf{x}_t, \eta_t$, we have the following property:

$$\begin{aligned}
\|\mathbf{x}_i\|_2 &= \big\|\bar{\sigma}_{i,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i)\big\|_2 \leq \big\|\boldsymbol{\phi}(s_h^i, a_h^i)\big\|_2 / \sqrt{H} \leq 1/\sqrt{H}, \\
\mathbb{E}[\eta_i | \mathcal{F}_i] &= 0, |\eta_t| \leq \Big|\bar{\sigma}_{i,h}^{-1}\big(V_{h+1}^*(s_{h+1}^i) - [\mathbb{P}_h \widetilde{V}_{h+1}](s_h^i, a_h^i)\big)\Big| \leq \sqrt{H}, \\
\mathbb{E}[\eta_i^2 | \mathcal{F}_i] &= \mathbb{E}\Big[\mathbb{1}\big\{[\mathbb{V}_h \widetilde{V}](s_h^i, a_h^i) \leq \bar{\sigma}_{i,h}^2/(d^3H)\big\} \cdot \bar{\sigma}_{i,h}^{-2} [\mathbb{V}_h \widetilde{V}](s_h^i, a_h^i)\Big] \leq 1/(d^3H), \\
\max_i \big\{|\eta_i| \cdot \min\{1, \|\mathbf{x}_i\|_{\boldsymbol{\Sigma}_{i,h}^{-1}}\}\big\} &\leq 2H\bar{\sigma}_{i,h}^{-1} \|\mathbf{x}_i\|_{\boldsymbol{\Sigma}_{i,h}^{-1}} \leq 1/(d^3H).
\end{aligned}$$

After taking a union bound over the ϵ-net, with probability at least $1 - \delta$, we have

$$\Big\|\sum_{i=1}^{k-1} \mathbf{x}_i \eta_i\Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \leq O\Big(\sqrt{d \log^2\big(1 + dKH/(\delta\lambda)\big)}\Big). \tag{D.23}$$

In addition, on the event $\widetilde{\mathcal{E}}_{h+1}$ and $\mathcal{E}$, according to Lemmas B.2 and B.3, we have

$$\begin{aligned}
\bar{\sigma}_{i,h}^2 &\geq [\bar{\mathbb{V}}_{i,h} V_{i,h+1}](s_h^k, a_h^k) + E_{i,h} + D_{i,h} + H \\
&\geq D_{i,h} + H \\
&\geq d^3 H\big[\mathbb{V}_h(\Delta V_{k,h+1})\big](s_h^i, a_h^i) + H \\
&\geq d^3 H [\mathbb{V}_h \widetilde{V}](s_h^i, a_h^i),
\end{aligned}$$

For simplicity, we denote $\bar{V} = \Delta V_{k,h+1} - \widetilde{V}$ and it further implies that

$$\begin{aligned}
J_2 &= \Big\|\sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)\big(\Delta V(s_{h+1}^i) - [\mathbb{P}_h(\Delta V_{k,h+1})](s_h^i, a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&\leq 2\Big\|\sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)\big(\widetilde{V}(s_{h+1}^i) - [\mathbb{P}_h \widetilde{V}](s_h^i, a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&\quad + 2\Big\|\sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)\big(\bar{V}(s_{h+1}^i) - [\mathbb{P}_h \bar{V}](s_h^i, a_h^i)\big)\Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} \\
&\leq 2\Big\|\sum_{i=1}^{k-1} \mathbf{x}_i \eta_i\Big\|_{\boldsymbol{\Sigma}_{k,h}^{-1}} + 8\epsilon^2 k^2/\lambda \\
&\leq O\Big(\sqrt{d \log^2\big(1 + dKH/(\delta\lambda)\big)}\Big).
\end{aligned} \tag{D.24}$$

where the first inequality holds due to $\|\mathbf{a}+\mathbf{b}\|_{\mathbf{\Sigma}}^2 \leq 2\|\mathbf{a}\|_{\mathbf{\Sigma}}^2 + 2\|\mathbf{b}\|_{\mathbf{\Sigma}}^2$, the second inequality holds due to the fact that $|\bar{V}(s')| \leq \epsilon, \|\boldsymbol{\phi}(s,a)\|_2 \leq 1, \mathbf{\Sigma}_{k,h} \succeq \lambda \mathbf{I}, \bar{\mathbf{\Sigma}}_{k,h}^{-1} \leq 1$ and the last inequality holds due to (D.23) with $\epsilon = \sqrt{\lambda}/(4H^2d^2K)$. Substituting the results in (D.17), (D.18), (D.19) and (D.23) into (D.16), we obtain

$$\|\widehat{\mathbf{w}}_{k,h} - \mathbf{w}_{k,h}\|_{\mathbf{\Sigma}_{k,h}} \leq I_1 + J_1 + J_2 \leq O\Big(H\sqrt{d\lambda} + \sqrt{d\log^2\big(1 + dKH/(\delta\lambda)\big)}\Big) = \beta, \tag{D.25}$$

Therefore, the estimation error is upper bounded by

$$\begin{aligned}
|\widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \mathbb{P}_h V_{k,h+1}(s,a)| &= |\widehat{\mathbf{w}}_{k,h}^\top \boldsymbol{\phi}(s,a) - \mathbf{w}_{k,h}^\top \boldsymbol{\phi}(s,a)| \\
&\leq \|\widehat{\mathbf{w}}_{k,h} - \mathbf{w}_{k,h}\|_{\mathbf{\Sigma}_{k,h}} \cdot \|\boldsymbol{\phi}(s,a)\|_{\mathbf{\Sigma}_{k,h}^{-1}} \\
&\leq \beta\sqrt{\boldsymbol{\phi}(s,a)^\top (\mathbf{\Sigma}_{k,h})^{-1} \boldsymbol{\phi}(s,a)},
\end{aligned}$$

where the first inequality holds due to Cauchy-Schwartz inequality and the last inequality holds due to (D.25), which implies the results in Lemma B.5 holds for stage h. Therefore, by induction, we finish the proof of Lemma B.5. □

E Proof of Lemmas in Appendix C

In this section, we provide the proof of Lemmas in Appendix C and we need the following auxiliary Lemma, which is modified from Lemma 4.4 in Zhou and Gu (2022)

Lemma E.1. *For any parameters $\beta' \geq 1$ and $C \geq 1$, the summation of bonuses is upper bounded by*

$$\sum_{k=1}^K \min\Big(\beta'\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \mathbf{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, C\Big) \leq 4d^4H^6C\iota + 10\beta' d^5 H^4 \iota + 2\beta'\sqrt{2d\iota \sum_{k=1}^K (\sigma_{k,h}^2 + H)},$$

where $\iota = \log\big(1 + K/(d\lambda)\big)$.

Proof of Lemma E.1. For each stage $h \in [H]$, the summation of bonuses is upper bounded by

$$\begin{aligned}
&\sum_{k=1}^K \min\Big(\beta'\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \mathbf{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, C\Big) \\
&\leq \sum_{k=1}^K \beta' \min\Big(\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \mathbf{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)}, 1\Big) + C\sum_{k=1}^K \mathbb{1}\Big\{\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \mathbf{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)} \geq 1\Big\} \\
&\leq C\sum_{k=1}^K \mathbb{1}\Big\{\sqrt{\boldsymbol{\phi}(s_h^k, a_h^k)^\top \mathbf{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)} \geq 1\Big\} + 10\beta' d^5 H^4 \iota + 2\beta'\sqrt{2d\iota \sum_{k=1}^K (\sigma_{k,h}^2 + H)},
\end{aligned} \tag{E.1}$$

where $\iota = \log\big(1 + K/(d\lambda)\big)$ and the last inequality holds due to Lemma G.6. Now, we only need to estimate the number of episodes where the bonus is larger than 1 and we denote these episodes as $\{k_1, .., k_m\}$. For simplicity, we denote

$$\mathbf{\Sigma}_i' = \lambda\mathbf{I} + \sum_{j=1}^i \bar{\sigma}_{k_j,h}^2 \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j}) \boldsymbol{\phi}(s_h^{k_j}, a_h^{k_j})^\top,$$

and we have

$$\sum_{i=1}^m \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \mathbf{\Sigma}_{i-1}' \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i}) \geq \sum_{i=1}^m \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \mathbf{\Sigma}_{k_i,h}^{-1} \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i}) \geq m. \tag{E.2}$$

On the other hand, notice that the estimated variance $\bar{\sigma}_{k,h}^2$ is upper bounded by $4d^4H^4/\lambda$, we have

$$\sum_{i=1}^m \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \mathbf{\Sigma}_{i-1}' \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i}) \leq 4d^4H^4/\lambda \cdot \sum_{i=1}^m \bar{\sigma}_{k,h}^{-2} \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i})^\top \mathbf{\Sigma}_{i-1}' \boldsymbol{\phi}(s_h^{k_i}, a_h^{k_i}) \leq 4d^4H^6\iota, \tag{E.3}$$

where $\iota = \log\big(1 + K/(d\lambda)\big)$ and the last inequality holds due to Lemma G.3. Combining the results in (E.2) and (E.3), we have $m \le 4d^4H^6\beta^2\iota$, and it further implies that

$$\begin{aligned}
&\sum_{k=1}^{K} \min\Big(\beta'\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)}, C\Big)\\
&\le C\sum_{k=1}^{K} \mathbb{1}\Big\{\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top \boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)} \ge 1\Big\} + 10\beta' d^5H^4\iota + 2\beta'\sqrt{2d\iota\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)}\\
&\le 4d^4H^6C\iota + 10\beta' d^5H^4\iota + 2\beta'\sqrt{2d\iota\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)}.
\end{aligned}$$

Thus, we finish the proof of Lemma E.1. □

E.1 Proof of Lemma C.1

Proof of Lemma C.1. For all stage $h \in [H]$ and episode $k \in [K]$, we have

$$\begin{aligned}
&V_{k,h}(s_h^k) - V_{k,h}^{\pi^k}(s_h^k)\\
&= Q_{k,h}(s_h^k,a_h^k) - Q_{k,h}^{\pi^k}(s_h^k,a_h^k)\\
&\le \min\Big(\widehat{\mathbf{w}}_{k_{\text{last}},h}^\top\boldsymbol{\phi}(s,a) + \beta\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top\boldsymbol{\Sigma}_{k_{\text{last}},h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)}, H\Big) - [\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\\
&\quad + \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k)\\
&\le \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) + 2\min\Big(\beta\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top\boldsymbol{\Sigma}_{k_{\text{last}},h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)}, H\Big)\\
&\le \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) + 4\min\Big(\beta\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)}, H\Big)\\
&= V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k) + \big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) - \big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big)\\
&\quad + 4\min\Big(\beta\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)}, H\Big),
\end{aligned} \tag{E.4}$$

where the first inequality holds due to the definition of value function $Q_{k,h}(s_h^k,a_h^k)$, the second inequality holds due to Lemma B.5 and the last inequality holds due to Lemma G.4 with the updating rule (Line 8). Furthermore, for all stage $h \in [H]$, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\big(V_{k,h}(s_h^k) - V_{k,h}^{\pi^k}(s_h^k)\big)\\
&\le \sum_{k=1}^{K}\sum_{h'=h}^{H} 4\min\Big(\beta\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)}, H\Big)\\
&\quad + \sum_{k=1}^{K}\sum_{h'=h}^{H}\Big(\big[\mathbb{P}_h(V_{k,h+1} - V_{k,h+1}^{\pi^k})\big](s_h^k,a_h^k) - \big(V_{k,h+1}(s_{h+1}^k) - V_{k,h+1}^{\pi^k}(s_{h+1}^k)\big)\Big)\\
&\le \sum_{k=1}^{K}\sum_{h'=h}^{H} 4\min\Big(\beta\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)}, H\Big) + 4\sqrt{H^3K\log(H/\delta)}\\
&\le 16d^4H^8\iota + 40\beta d^7H^5\iota + 8\beta\sum_{h'=h}^{H}\sqrt{2d\iota\sum_{k=1}^{K}(\sigma_{k,h'}^2 + H)} + 4\sqrt{H^3K\log(H/\delta)}\\
&\le 16d^4H^8\iota + 40\beta d^7H^5\iota + 8\beta\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)} + 4\sqrt{H^3K\log(H/\delta)},
\end{aligned} \tag{E.5}$$

where the first inequality holds by taking the summation of (E.4) for $k \in [K]$ and $h \le h' \le H$, the second inequality holds due to the definition of events $\mathcal{E}_1$, the third inequality holds due to Lemma E.1 and the last inequality holds due to Cauchy-Schwartz inequality. Furthermore, taking the summation of (E.5), we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\big[\mathbb{P}_h(V_{k,h+1}-V^{\pi^k}_{k,h+1})\big](s^k_h,a^k_h)\\
&=\sum_{k=1}^{K}\sum_{h=1}^{H}\big(V_{k,h+1}(s^k_{h+1})-V^{\pi^k}_{k,h+1}(s^k_{h+1})\big)\\
&\quad+\sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\big[\mathbb{P}_h(V_{k,h+1}-V^{\pi^k}_{k,h+1})\big](s^k_h,a^k_h)-\big(V_{k,h+1}(s^k_{h+1})-V^{\pi^k}_{k,h+1}(s^k_{h+1})\big)\Big)\\
&\le\sum_{k=1}^{K}\sum_{h=1}^{H}\big(V_{k,h+1}(s^k_{h+1})-V^{\pi^k}_{k,h+1}(s^k_{h+1})\big)+2\sqrt{2H^3K\log(H/\delta)}\\
&\le 16d^4H^9\iota+40\beta d^7H^6\iota+8H\beta\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma^2_{k,h}+H)}+4\sqrt{H^5K\log(H/\delta)},
\end{aligned}
$$

where the first inequality holds due to Lemma G.2 and the last inequality holds due (E.5). Therefore, we finish the proof of Lemma C.1. □

E.2 Proof of Lemma C.2

Proof of Lemma C.2. For each stage $h \in [H]$ and episode $k \in [K]$, we have

$$
\begin{aligned}
&V_{k,h}(s^k_h)-\check{V}_{k,h}(s^k_h)\\
&\le Q_{k,h}(s^k_h,a^k_h)-\check{Q}_{k,h}(s^k_h,a^k_h)\\
&\le \min\Big(\check{\mathbf{w}}^\top_{k_{\text{last}},h}\boldsymbol{\phi}(s,a)+\beta\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k_{\text{last}},h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big)-[\mathbb{P}_hV_{k,h+1}](s^k_h,a^k_h)\\
&\quad-\max\Big(\widehat{\mathbf{w}}^\top_{k_{\text{last}},h}\boldsymbol{\phi}(s,a)-\bar{\beta}\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k_{\text{last}},h}\boldsymbol{\phi}(s^k_h,a^k_h)},0\Big)+[\mathbb{P}_h\check{V}_{k,h+1}](s^k_h,a^k_h)\\
&\quad+\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s^k_h,a^k_h)\\
&\le\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s^k_h,a^k_h)+2\min\Big(\beta\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k_{\text{last}},h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big)\\
&\quad+2\min\Big(\bar{\beta}\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k_{\text{last}},h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big)\\
&\le\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s^k_h,a^k_h)+4\min\Big(\beta\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k,h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big)\\
&\quad+4\min\Big(\bar{\beta}\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k,h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big)\\
&=V_{k,h+1}(s^k_{h+1})-\check{V}_{k,h+1}(s^k_{h+1})+\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s^k_h,a^k_h)-\big(V_{k,h+1}(s^k_{h+1})-\check{V}_{k,h+1}(s^k_{h+1})\big)\\
&\quad+4\min\Big(\beta\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k,h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big)+4\min\Big(\bar{\beta}\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k,h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big),
\end{aligned}
\tag{E.6}
$$

where the first inequality holds due to the fact that $\check{V}_{k,h}(s^k_h)=\max_a\check{Q}_{k,h}(s^k_h,a)\ge\check{Q}_{k,h}(s^k_h,a^k_h)$, the second inequality holds due to the definition of value functions $Q_{k,h}$ and $\check{Q}_{k,h}$, the third inequality holds due to Lemma B.5 and Lemma B.1, and the last inequality holds due to Lemma G.4 with the updating rule (Line 8). Furthermore, for all stage $h \in [H]$, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\big(V_{k,h}(s^k_h)-\check{V}_{k,h}(s^k_h)\big)\\
&\le\sum_{k=1}^{K}\sum_{h'=h}^{H}4\min\Big(\beta\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k,h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big)+4\min\Big(\bar{\beta}\sqrt{\boldsymbol{\phi}(s^k_h,a^k_h)^\top\boldsymbol{\Sigma}^{-1}_{k,h}\boldsymbol{\phi}(s^k_h,a^k_h)},H\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{k=1}^{K}\sum_{h'=h}^{H}\Big(\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s_h^k,a_h^k)-\big(V_{k,h+1}(s_{h+1}^k)-\check{V}_{k,h+1}(s_{h+1}^k)\big)\Big)\\
&\le\sum_{k=1}^{K}\sum_{h'=h}^{H}4\min\Big(\beta\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)},H\Big)+4\min\Big(\bar{\beta}\sqrt{\boldsymbol{\phi}(s_h^k,a_h^k)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)},H\Big)\\
&\quad+4\sqrt{H^3K\log(H/\delta)}\\
&\le 32d^4H^8\iota+40(\beta+\bar{\beta})d^7H^5\iota+8(\beta+\bar{\beta})\sum_{h'=h}^{H}\sqrt{2d\iota\sum_{k=1}^{K}(\sigma_{k,h'}^2+H)}+4\sqrt{H^3K\log(H/\delta)}\\
&\le 32d^4H^8\iota+40(\beta+\bar{\beta})d^7H^5\iota+8(\beta+\bar{\beta})\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2+H)}+4\sqrt{H^3K\log(H/\delta)},
\end{aligned}
\tag{E.7}
$$

where the first inequality holds by taking the summation of (E.6) for $k\in[K]$ and $h\le h'\le H$, the second inequality holds due to the definition of event $\mathcal{E}_2$, the third inequality holds due to Lemma E.1 and the last inequality holds due to Cauchy-Schwartz inequality. Furthermore, taking the summation of (E.7), we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s_h^k,a_h^k)\\
&=\sum_{k=1}^{K}\sum_{h=1}^{H}\big(V_{k,h+1}(s_{h+1}^k)-\check{V}_{k,h+1}(s_{h+1}^k)\big)\\
&\quad+\sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s_h^k,a_h^k)-\big(V_{k,h+1}(s_{h+1}^k)-\check{V}_{k,h+1}(s_{h+1}^k)\big)\Big)\\
&\le\sum_{k=1}^{K}\sum_{h=1}^{H}\big(V_{k,h+1}(s_{h+1}^k)-\check{V}_{k,h+1}(s_{h+1}^k)\big)+2\sqrt{2H^3K\log(H/\delta)}\\
&\le 32d^4H^9\iota+40(\beta+\bar{\beta})d^7H^6\iota+8H(\beta+\bar{\beta})\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2+H)}+4\sqrt{H^5K\log(H/\delta)},
\end{aligned}
$$

where the first inequality holds due to Lemma G.2 and the last inequality holds due (E.7). Therefore, we finish the proof of Lemma C.2. □

E.3 Proof of Lemma C.3

Proof of Lemma C.3. According to the definition of estimated variance $\sigma_{k,h}$, we have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2&=\sum_{k=1}^{K}\sum_{h=1}^{H}[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)+E_{k,h}+D_{k,h}+H\\
&=H^2K+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\big([\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)\big)}_{I_1}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}E_{k,h}}_{I_2}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}D_{k,h}}_{I_3}\\
&\quad+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\big([\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}^{\pi^k}](s_h^k,a_h^k)\big)}_{I_4}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_hV_{k,h+1}^{\pi^k}](s_h^k,a_h^k)}_{I_5}.
\end{aligned}
\tag{E.8}
$$

For the term I_1, according to Lemma B.2, it is upper bounded by:

$$
I_1=\sum_{k=1}^{K}\sum_{h=1}^{H}\big([\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)\big)\le\sum_{k=1}^{K}\sum_{h=1}^{H}E_{k,h}=I_2.
\tag{E.9}
$$

For the term I_2, it is upper bounded by

$$\begin{aligned} I_2 &= \sum_{k=1}^{K}\sum_{h=1}^{H} E_{k,h} \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H} \min\Big\{\widetilde{\beta}\big\|\boldsymbol{\Sigma}_{k,h}^{-1/2}\boldsymbol{\phi}(s_h^k,a_h^k)\big\|_2, H^2\Big\} + \min\Big\{2H\bar{\beta}\big\|\boldsymbol{\Sigma}_{k,h}^{-1/2}\boldsymbol{\phi}(s_h^k,a_h^k)\big\|_2, H^2\Big\} \\ &\le 8d^4H^9\iota + (10\widetilde{\beta}+20\bar{\beta})d^5H^5\iota + (2\widetilde{\beta}+4\bar{\beta})H\sqrt{2d\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2+H)}, \end{aligned} \tag{E.10}$$

where $\iota = \log\big(1+K/(d\lambda)\big)$ and the inequality holds due to Lemma E.1.

For the term I_3, it is upper bounded by

$$\begin{aligned} I_3 &= \sum_{k=1}^{K}\sum_{h=1}^{H} D_{k,h} \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H} \min\Big\{4d^3H^2\Big(\widehat{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s,a) - \check{\mathbf{w}}_{k,h}^\top\boldsymbol{\phi}(s,a) + 2\bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)}\Big), d^3H^3\Big\} \\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H} \min\Big\{4d^3H^2\Big(\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s_h^k,a_h^k) + 4\bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)}\Big), d^3H^3\Big\} \\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H} 4d^3H^2\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s_h^k,a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H}\min\Big\{16d^3H^2\bar{\beta}\sqrt{\boldsymbol{\phi}(s,a)^\top\boldsymbol{\Sigma}_{k,h}^{-1}\boldsymbol{\phi}(s,a)}, d^3H^3\Big\} \\ &\le 4d^7H^9\iota + 160\bar{\beta}d^8H^7\iota + 32d^3H^3\bar{\beta}\sqrt{2d\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2+H)} \\ &\quad + \sum_{k=1}^{K}\sum_{h=1}^{H} 4d^3H^2\big[\mathbb{P}_h(V_{k,h+1}-\check{V}_{k,h+1})\big](s_h^k,a_h^k) \\ &\le 132d^7H^{11}\iota + 320(\beta+\bar{\beta})d^{10}H^8\iota + 64d^3H^3(\beta+\bar{\beta})\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2+H)} + 4d^3\sqrt{H^9K\log(H/\delta)}, \end{aligned} \tag{E.11}$$

where $\iota = \log\big(1+K/(d\lambda)\big)$, the first inequality holds due to Lemma B.1, the second inequality holds due to the fact that $V_{k,h+1}(s) \ge V_{h+1}^*(s) \ge \check{V}_{k,h+1}(s)$, the third inequality holds due to Lemma E.1 and the last inequality holds due to Lemma C.2.

For the term I_4, it is upper bounded by

$$\begin{aligned} I_4 &= \sum_{k=1}^{K}\sum_{h=1}^{H}\big([\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k) - [\mathbb{V}_hV_{k,h+1}^{\pi^k}](s_h^k,a_h^k)\big) \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H}\Big([\mathbb{P}_hV_{k,h+1}^2](s_h^k,a_h^k) - \big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k)\big)^2 - [\mathbb{P}_h(V_{k,h+1}^{\pi^k})^2](s_h^k,a_h^k) + \big([\mathbb{P}_hV_{k,h+1}^{\pi^k}](s_h^k,a_h^k)\big)^2\Big) \\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H}\big([\mathbb{P}_hV_{k,h+1}^2](s_h^k,a_h^k) - [\mathbb{P}_h(V_{k,h+1}^{\pi^k})^2](s_h^k,a_h^k)\big) \\ &\le 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\big([\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k) - [\mathbb{P}_hV_{k,h+1}^{\pi^k}](s_h^k,a_h^k)\big) \end{aligned}$$

$$\leq 32d^4H^{10}\iota + 80\beta d^7H^7\iota + 16H^2\beta\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)} + 8\sqrt{H^7K\log(H/\delta)}, \tag{E.12}$$

where the first inequality holds due to the fact that $V_{k,h+1}^{\pi^k}(s') \leq V_{k,h+1}(s')$, the second inequality holds due to $0 \leq V_{k,h+1}(s'), V_{k,h+1}^{\pi^k}(s') \leq H$ and the last inequality holds due to Lemma C.1.

Based on the definition of events $\mathcal{E}_3$, for the term I_5, we have

$$I_5 = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{k,h+1}^{\pi^k}](s_h^k, a_h^k) \leq 3\big(H^2K + H^3\log(1/\delta)\big). \tag{E.13}$$

Substituting the results in (E.9), (E.10), (E.11), (E.12) and (E.13) into (E.8), we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2 \\
&= I_1 + I_2 + I_3 + I_4 + I_5 \\
&\leq 3H^2K + 183d^7H^{11}\iota + 460(\beta + \widetilde{\beta} + \bar{\beta})d^{10}H^8\iota + 12d^3\sqrt{H^9K\log(H/\delta)} \\
&\quad + 92d^3H^3(\beta + \widetilde{\beta} + \bar{\beta})\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}(\sigma_{k,h}^2 + H)} \\
&\leq 3H^2K + 183d^7H^{11}\iota + 460(\beta + \widetilde{\beta} + \bar{\beta})d^{10}H^8\iota + 12d^3\sqrt{H^9K\log(H/\delta)} \\
&\quad + 92d^3H^3(\beta + \widetilde{\beta} + \bar{\beta})\sqrt{2dH^2K\iota} + 92d^3H^3(\beta + \widetilde{\beta} + \bar{\beta})\sqrt{2dH\iota\sum_{h=1}^{H}\sum_{k=1}^{K}\sigma_{k,h}^2},
\end{aligned}$$

where $\iota = \log\big(1 + K/(d\lambda)\big)$,

$$\begin{aligned}
\beta &= O\Big(H\sqrt{d\lambda} + \sqrt{d\log^2\big(1 + dKH/(\delta\lambda)\big)}\Big) \\
\widetilde{\beta} &= O\Big(H^2\sqrt{d\lambda} + \sqrt{d^3H^4\log^2\big(dHK/(\delta\lambda)\big)}\Big) \\
\bar{\beta} &= O\Big(H\sqrt{d\lambda} + \sqrt{d^3H^2\log^2\big(dHK/(\delta\lambda)\big)}\Big),
\end{aligned}$$

and the last inequality holds due to the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Therefore, by the fact that $x \leq a\sqrt{x} + b$ implies $x \leq a^2 + 2b$ and $\lambda = 1/H^2$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2 \leq O\big(H^2K + d^{10.5}H^{16}\log^{1.5}(1 + dKH/\delta)\big).$$

Thus, we complete the proof of Lemma C.3. □

F Covering Number Arguments

F.1 Number of Value Function Updating

According to the determinant-based criterion in Algorithm 1 (Line 8), the number of episodes where the algorithm updates the value function is upper bounded by:

Lemma F.1. *The number of episodes where the algorithm updates the value function in Algorithm 1 is upper bounded by* $dH\log(1 + K/\lambda)$.

Proof. We denote $k_0 = 0$ and suppose that $\{k_1, .., k_m\}$ be the episodes where the algorithm updates the value function. Then according to the determinant-based criterion (Line 8), for each episodes k_i, there exists a stage $h \in [H]$ such that

$$\det(\boldsymbol{\Sigma}_{k_i,h}) \geq 2\det(\boldsymbol{\Sigma}_{k_{i-1},h}).$$

According to the update rule of $\mathbf{\Sigma}_{k,h}$ (Line 24), for other stage $h' \neq h$, we have $\mathbf{\Sigma}_{k_i,h'} \succeq \mathbf{\Sigma}_{k_i,h'}$, which implies $\det(\mathbf{\Sigma}_{k_i,h'}) \geq \det(\mathbf{\Sigma}_{k_i,h'})$. Thus, we have

$$\prod_{h=1}^{H} \det(\mathbf{\Sigma}_{k_i,h}) \geq 2 \prod_{h=1}^{H} \det(\mathbf{\Sigma}_{k_{i-1},h}). \tag{F.1}$$

Applying the result (F.1) overall episodes in $\{k_1, .., k_m\}$, we have

$$\prod_{h=1}^{H} \det(\mathbf{\Sigma}_{k_m,h}) \geq 2^m \prod_{h=1}^{H} \det(\mathbf{\Sigma}_{k_0,h}) = 2^m \prod_{h=1}^{H} \det(\lambda \mathbf{I}) = 2^m \lambda^{dH}. \tag{F.2}$$

On the other hand, the determinant $\det(\mathbf{\Sigma}_{k_m,h})$ is upper bounded by:

$$\prod_{h=1}^{H} \det(\mathbf{\Sigma}_{k_m,h}) \leq \prod_{h=1}^{H} \det(\mathbf{\Sigma}_{K,h}) \leq (\lambda + K)^{dH}, \tag{F.3}$$

where the first inequality holds due to $\mathbf{\Sigma}_{K,h} \succeq \mathbf{\Sigma}_{k_m,h}$, the last inequality holds due to $\bar{\sigma}_{k,h}^{-1} \leq 1$ and $\|\boldsymbol{\phi}(s,a)\|_2 \leq 1$. Combining the results in (F.2) and (F.3), we have

$$m \leq dH \log(1 + K/\lambda).$$

Thus, we finish the proof of Lemma F.1. □

F.2 Norm of the Weight Vectors

In this section, we provide the following upper bounds for the norm of the weight vectors.

Lemma F.2. *For all stage $h \in [H]$ and all episode $k \in \mathbb{N}$, the norm of the weight vector $\widehat{\mathbf{w}}_{k,h}$ can be upper bounded by*

$$\|\widehat{\mathbf{w}}_{k,h}\|_2 \leq H\sqrt{dK/\lambda}.$$

Proof of Lemma F.2. According to the definition of weight vector $\widehat{\mathbf{w}}_{k,h}$ in Algorithm 1, we have

$$\mathbf{\Sigma}_{k,h} = \lambda \mathbf{I} + \sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \boldsymbol{\phi}(s_h^i, a_h^i)^\top,$$
$$\widehat{\mathbf{w}}_{k,h} = \mathbf{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) V_{k,h+1}(s_{h+1}^i).$$

Then for the norm $\|\widehat{\mathbf{w}}_{k,h}\|_2$, we have the following inequality

$$\begin{aligned}
\|\widehat{\mathbf{w}}_{k,h}\|_2^2 &= \bigg\| \mathbf{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) V_{k,h+1}(s_{h+1}^i) \bigg\|_2^2 \\
&\leq k \sum_{i=1}^{k-1} \big\| \mathbf{\Sigma}_{k,h}^{-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) V_{k,h+1}(s_{h+1}^i) \big\|_2^2 \\
&\leq kH^2 \sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \big\| \mathbf{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i) \big\|_2^2 \\
&\leq \frac{kH^2}{\lambda} \sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)^\top \mathbf{\Sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i) \\
&= \frac{kH^2}{\lambda} \mathrm{tr}\bigg(\mathbf{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)^\top \boldsymbol{\phi}(s_h^i, a_h^i) \bigg),
\end{aligned} \tag{F.4}$$

where the first inequality holds due to Cauchy-Schwartz inequality, the second inequality holds due to $0 \leq V_{k,h+1}(s,a) \leq H$ and the last inequality holds due to $\boldsymbol{\Sigma}_{k,h} \succeq \lambda \mathbf{I}$. Now, we assume the eigen-decomposition of matrix $\sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)^\top \boldsymbol{\phi}(s_h^i, a_h^i)$ is $Q^\top \Lambda Q$ and we have

$$\begin{aligned}
\mathrm{tr}\bigg(\boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)^\top \boldsymbol{\phi}(s_h^i, a_h^i)\bigg) &= \mathrm{tr}\big((Q^\top \Lambda Q + \lambda \mathbf{I})^{-1} Q^\top \Lambda Q\big) \\
&= \mathrm{tr}\big((\Lambda + \lambda \mathbf{I})^{-1} \Lambda\big) \\
&= \sum_{i=1}^{d} \frac{\Lambda_i}{\Lambda_i + \lambda} \\
&\leq d. \qquad \text{(F.5)}
\end{aligned}$$

Substituting (F.5) into (F.4), we have

$$\begin{aligned}
\|\widehat{\mathbf{w}}_{k,h}\|_2^2 &\leq \frac{kH^2}{\lambda} \mathrm{tr}\bigg(\boldsymbol{\Sigma}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{k,i}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)^\top \boldsymbol{\phi}(s_h^i, a_h^i)\bigg) \\
&\leq \frac{kH^2 d}{\lambda}, \qquad \text{(F.6)}
\end{aligned}$$

where the first inequality holds due to (F.4) and the last inequality holds due to (F.5). Thus, we finish the proof of Lemma F.2 □

In addition, for the pessimistic weight vector $\check{\mathbf{w}}_{k,h}$ and weight vector $\widetilde{\mathbf{w}}_{k,h}$, we have the following lemmas:

Lemma F.3. *For all stage $h \in [H]$ and all episode $k \in \mathbb{N}$, the norm of the weight vector $\check{\mathbf{w}}_{k,h}$ can be upper bounded by*

$$\|\check{\mathbf{w}}_{k,h}\|_2 \leq H\sqrt{dK/\lambda}.$$

Proof of Lemma F.3. The proof is almost the same as Lemma F.2 and we only need to replace the optimistic value function $V_{k,h}(s,a)$ by the pessimistic value function $\check{V}_{k,h}(s,a)$. □

Lemma F.4. *For each stage $h \in [H]$ and each episode $k \in \mathbb{N}$, the norm of the weight vector $\widetilde{\mathbf{w}}_{k,h}$ can be upper bounded by*

$$\|\check{\mathbf{w}}_{k,h}\|_2 \leq H^2\sqrt{dK/\lambda}.$$

Proof of Lemma F.4. The proof is almost the same as Lemma F.2 and we only need to replace the optimistic value function $V_{k,h}(s,a)$ with the squared value function $V_{k,h}^2(s,a)$. □

F.3 Function Class and Covering Number

Combining the update rule (Line 8) with Lemma F.1 and Lemma F.2, for each episodes $k \in [K]$ and $h \in [H]$, the optimistic value function $V_{k,h} = \min_{i \leq k} \max_a Q_{i,h}(s,a)$ belong to the following function class

$$\mathcal{V}_h = \bigg\{V \bigg| V(\cdot) = \max_a \min_{1 \leq i \leq l} \min\Big(H, r_h(\cdot,a) + \mathbf{w}_i^\top \boldsymbol{\phi}(\cdot,a) + \beta\sqrt{\boldsymbol{\phi}(\cdot,a)^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\phi}(\cdot,a)}\Big), \|\mathbf{w}_i\| \leq L, \boldsymbol{\Sigma}_i \succeq \lambda \mathbf{I}\bigg\}, \quad \text{(F.7)}$$

where $l = dH\log(1 + K/\lambda)$ and $L = H\sqrt{dK/\lambda}$. Similarly, for each episodes $k \in [K]$ and $h \in [H]$, the pessimistic value function $\check{V}_{k,h} = \max_{i \leq k} \max_a \check{Q}_{i,h}(s,a)$ belongs to the following function class

$$\check{\mathcal{V}}_h = \bigg\{V \bigg| V(\cdot) = \max_a \max_{1 \leq i \leq l} \max\Big(0, r_h(\cdot,a) + \mathbf{w}_i^\top \boldsymbol{\phi}(\cdot,a) - \beta\sqrt{\boldsymbol{\phi}(\cdot,a)^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\phi}(\cdot,a)}\Big), \|\mathbf{w}_i\| \leq L, \boldsymbol{\Sigma}_i \succeq \lambda \mathbf{I}\bigg\}, \quad \text{(F.8)}$$

where $l = dH\log(1 + K/\lambda)$ and $L = H\sqrt{dK/\lambda}$. To compute the covering number of function classes $\mathcal{V}_h$, $\mathcal{V}_h^2$ and $\check{\mathcal{V}}_h$, we need the following result on the Euclidean ball.

Lemma F.5 (Lemma D.5, Jin et al. 2020)**.** *For a Euclidean ball with radius R in $\mathbb{R}^d$, the ϵ-covering number of this ball is upper bounded by $(1+2R/\epsilon)^d$.*

With the help of Lemma F.5, the covering number $\mathcal{N}_\epsilon$ of optimistic function class $\mathcal{V}$ can be upper bounded by the following lemma:

Lemma F.6. *For optimistic function class $\mathcal{V}_h$, we define the distance between two function V_1 and V_2 as $V_1, V_2 \in \mathcal{V}_h$ as $dist(V_1, V_2) = \max_s |V_1(s) - V_2(s)|$. With respect to this distance function, the ϵ-covering number $\mathcal{N}_\epsilon$ of the function class $\mathcal{V}_h$ can be upper bounded by*

$$\log \mathcal{N}_\epsilon \leq dl \log(1 + 4L/\epsilon) + d^2 l \log(1 + 8\sqrt{d}\beta^2/\epsilon^2)$$

Proof of Lemma F.6. For any two function $V_1, V_2 \in \mathcal{V}_h$, according to the definition of function class $\mathcal{V}_h$, we have

$$V_1(\cdot) = \max_a \min_{1\leq i\leq l} \min\left(H, r_h(\cdot, a) + \mathbf{w}_{1,i}^\top \boldsymbol{\phi}(\cdot, a) + \beta\sqrt{\boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{\Gamma}_{1,i} \boldsymbol{\phi}(\cdot, a)}\right),$$
$$V_2(\cdot) = \max_a \min_{1\leq i\leq l} \min\left(H, r_h(\cdot, a) + \mathbf{w}_{2,i}^\top \boldsymbol{\phi}(\cdot, a) + \beta\sqrt{\boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{\Gamma}_{2,i} \boldsymbol{\phi}(\cdot, a)}\right),$$

where $\|\mathbf{w}_{1,i}\|_2, \|\mathbf{w}_{2,i}\|_2 \leq L$ and $\boldsymbol{\Gamma}_{1,i}, \boldsymbol{\Gamma}_{2,i} \preceq \mathbf{I}$. Since all of the functions $\min_{1\leq i\leq l}$, $\max_a$ and $\min(H, \cdot)$ are contraction functions, we have

$$\begin{aligned}
\text{dist}(V_1, V_2) &= \max_{s\in\mathcal{S}} \big|V_1(s) - V_2(s)\big| \\
&\leq \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \Big|\mathbf{w}_{1,i}^\top \boldsymbol{\phi}(s, a) + \beta\sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Gamma}_{1,i} \boldsymbol{\phi}(s, a)} \\
&\qquad - \mathbf{w}_{2,i}^\top \boldsymbol{\phi}(s, a) - \beta\sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Gamma}_{2,i} \boldsymbol{\phi}(s, a)}\Big| \\
&\leq \beta \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \Big|\sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Gamma}_{1,i} \boldsymbol{\phi}(s, a)} - \sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Gamma}_{2,i} \boldsymbol{\phi}(s, a)}\Big| \\
&\qquad + \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \big|(\mathbf{w}_{1,i} - \mathbf{w}_{2,i})^\top \boldsymbol{\phi}(s, a)\big| \\
&\leq \beta \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \Big|\sqrt{\boldsymbol{\phi}(s, a)^\top (\boldsymbol{\Gamma}_{1,i} - \boldsymbol{\Gamma}_{2,i}) \boldsymbol{\phi}(s, a)}\Big| \\
&\qquad + \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \big|(\mathbf{w}_{1,i} - \mathbf{w}_{2,i})^\top \boldsymbol{\phi}(s, a)\big| \\
&\leq \beta \max_{1\leq i\leq l} \sqrt{\|\boldsymbol{\Gamma}_{1,i} - \boldsymbol{\Gamma}_{2,i}\|_F} + \max_{1\leq i\leq l} \|\mathbf{w}_{1,i} - \mathbf{w}_{2,i}\|_2,
\end{aligned} \tag{F.9}$$

where the first inequality holds due to the contraction property, the second inequality holds due to the fact that $\max_x |f(x) + g(x)| \leq \max_x |f(x)| + \max_x |g(x)|$, the third inequality holds due to $|\sqrt{x} - \sqrt{y}| \geq |\sqrt{x} - \sqrt{y}|$ and the last inequality holds due to the fact that $\|\boldsymbol{\phi}(s, a)\|_2 \leq 1$. Now, we denote $\mathcal{C}_\mathbf{w}$ as a $\epsilon/2$-cover of the set $\big\{\mathbf{w} \in \mathbb{R}^d \big| \|\mathbf{w}\|_2 \leq L\big\}$ and $\mathcal{C}_\boldsymbol{\Gamma}$ as a $\epsilon^2/(4\beta^2)$-cover of the set $\{\boldsymbol{\Gamma} \in \mathbb{R}^{d\cdot d} \big| \|\boldsymbol{\Gamma}\|_F \leq \sqrt{d}\}$ with respect to the Frobenius norms. Thus, according to Lemma F.5, we have following property:

$$|\mathcal{C}_\mathbf{w}| \leq \big(1 + 4L/\epsilon\big)^d, |\mathcal{C}_\boldsymbol{\Gamma}| \leq \big(1 + 8\sqrt{d}\beta^2/\epsilon^2\big)^{d^2}. \tag{F.10}$$

By the definition of covering number, for any function $V_1 \in \mathcal{V}$ with parameters $\mathbf{w}_{1,i}, \boldsymbol{\Gamma}_{1,i} (1 \leq i \leq l)$, s other parameters $\mathbf{w}_{2,i}, \boldsymbol{\Gamma}_{2,i} (1 \leq i \leq l)$ such that $\mathbf{w}_{2,i} \in \mathcal{C}_\mathbf{w}, \boldsymbol{\Gamma}_{2,i} \in \mathcal{C}_\boldsymbol{\Gamma}$ and $\|\mathbf{w}_{2,i} - \mathbf{w}_{1,i}\|_2 \leq \epsilon/2, \|\boldsymbol{\Gamma}_{2,i} - \boldsymbol{\Gamma}_{1,i}\|_F \leq \epsilon^2/(4\beta^2)$. Thus, we have

$$\text{dist}(V_1, V_2) \leq \beta \max_{1\leq i\leq l} \sqrt{\|\boldsymbol{\Gamma}_{1,i} - \boldsymbol{\Gamma}_{2,i}\|_F} + \max_{1\leq i\leq l} \|\mathbf{w}_{1,i} - \mathbf{w}_{2,i}\|_2 \leq \epsilon,$$

where the inequality holds due to (F.9). Therefore, the ϵ-covering number of optimistic function class $\mathcal{V}_h$ is bounded by $\mathcal{N}_\epsilon \leq |\mathcal{C}_\mathbf{w}|^l \cdot |\mathcal{C}_\boldsymbol{\Gamma}|^l$ and it implies

$$\log \mathcal{N}_\epsilon \leq dl \log(1 + 4L/\epsilon) + d^2 l \log(1 + 8\sqrt{d}\beta^2/\epsilon^2),$$

where the first inequality holds due to (F.10). Thus, we finish the proof of Lemma F.6. □

With a similar argument as Lemma F.6, the covering number $\mathcal{N}_\epsilon$ of pessimistic value function class $\check{\mathcal{V}}_h$ can be upper bounded by the following lemma:

Lemma F.7. *For pessimistic function class $\check{\mathcal{V}}_h$, we define the distance between two function V_1 and V_2 as $V_1, V_2 \in \check{\mathcal{V}}_h$ as $dist(V_1, V_2) = \max_s |V_1(s) - V_2(s)|$. With respect to this distance function, the ϵ-covering number $\mathcal{N}_\epsilon$ of the function class $\check{\mathcal{V}}_h$ can be upper bounded by*

$$\log \mathcal{N}_\epsilon \leq dl \log(1 + 4L/\epsilon) + d^2 l \log(1 + 8\sqrt{d}\beta^2/\epsilon^2)$$

Proof of Lemma F.7. For any two function $V_1, V_2 \in \check{\mathcal{V}}_h$, according to the definition of function class $\check{\mathcal{V}}_h$, we have

$$V_1(\cdot) = \max_a \max_{1\leq i\leq l} \max\left(0, r_h(\cdot, a) + \mathbf{w}_{1,i}^\top \boldsymbol{\phi}(\cdot, a) - \beta\sqrt{\boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{\Gamma}_{1,i} \boldsymbol{\phi}(\cdot, a)}\right),$$
$$V_2(\cdot) = \max_a \max_{1\leq i\leq l} \max\left(0, r_h(\cdot, a) + \mathbf{w}_{2,i}^\top \boldsymbol{\phi}(\cdot, a) - \beta\sqrt{\boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{\Gamma}_{2,i} \boldsymbol{\phi}(\cdot, a)}\right),$$

where $\|\mathbf{w}_{1,i}\|_2, \|\mathbf{w}_{2,i}\|_2 \leq L$ and $\boldsymbol{\Gamma}_{1,i}, \boldsymbol{\Gamma}_{2,i} \preceq \mathbf{I}$. Since all of the functions $\max_{1\leq i\leq l}$, $\max_a$ and $\max(0, \cdot)$ are contraction functions, we have

$$\begin{aligned}
\text{dist}(V_1, V_2) &= \max_{s\in\mathcal{S}} \big|V_1(s) - V_2(s)\big| \\
&\leq \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \Big|\mathbf{w}_{1,i}^\top \boldsymbol{\phi}(s, a) - \beta\sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Gamma}_{1,i} \boldsymbol{\phi}(s, a)} \\
&\qquad - \mathbf{w}_{2,i}^\top \boldsymbol{\phi}(s, a) + \beta\sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Gamma}_{2,i} \boldsymbol{\phi}(s, a)}\Big| \\
&\leq \beta \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \Big|\sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Gamma}_{1,i} \boldsymbol{\phi}(s, a)} - \sqrt{\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Gamma}_{2,i} \boldsymbol{\phi}(s, a)}\Big| \\
&\qquad + \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \big|(\mathbf{w}_{1,i} - \mathbf{w}_{2,i})^\top \boldsymbol{\phi}(s, a)\big| \\
&\leq \beta \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \Big|\sqrt{\boldsymbol{\phi}(s, a)^\top (\boldsymbol{\Gamma}_{1,i} - \boldsymbol{\Gamma}_{2,i}) \boldsymbol{\phi}(s, a)}\Big| \\
&\qquad + \max_{1\leq i\leq l, s\in\mathcal{S}, a\in\mathcal{A}} \big|(\mathbf{w}_{1,i} - \mathbf{w}_{2,i})^\top \boldsymbol{\phi}(s, a)\big| \\
&\leq \beta \max_{1\leq i\leq l} \sqrt{\|\boldsymbol{\Gamma}_{1,i} - \boldsymbol{\Gamma}_{2,i}\|_F} + \max_{1\leq i\leq l} \|\mathbf{w}_{1,i} - \mathbf{w}_{2,i}\|_2,
\end{aligned} \tag{F.11}$$

where the first inequality holds due to the contraction property, the second inequality holds due to the fact that $\max_x |f(x) + g(x)| \leq \max_x |f(x)| + \max_x |g(x)|$, the third inequality holds due to $|\sqrt{x} - \sqrt{y}| \geq |\sqrt{x} - \sqrt{y}|$ and the last inequality holds due to the fact that $\|\boldsymbol{\phi}(s, a)\|_2 \leq 1$. Now, we denote $\mathcal{C}_\mathbf{w}$ as a $\epsilon/2$-cover of the set $\big\{\mathbf{w} \in \mathbb{R}^d \big| \|\mathbf{w}\|_2 \leq L\big\}$ and $\mathcal{C}_\boldsymbol{\Gamma}$ as a $\epsilon^2/(4\beta^2)$-cover of the set $\{\boldsymbol{\Gamma} \in \mathbb{R}^{d\cdot d} \big| \|\boldsymbol{\Gamma}\|_F \leq \sqrt{d}\}$ with respect to the Frobenius norms. Thus, according to Lemma F.5, we have following property:

$$|\mathcal{C}_\mathbf{w}| \leq \big(1 + 4L/\epsilon\big)^d, |\mathcal{C}_\boldsymbol{\Gamma}| \leq \big(1 + 8\sqrt{d}\beta^2/\epsilon^2\big)^{d^2}. \tag{F.12}$$

By the definition of covering number, for any function $V_1 \in \check{\mathcal{V}}$ with parameters $\mathbf{w}_{1,i}, \boldsymbol{\Gamma}_{1,i} (1 \leq i \leq l)$, s other parameters $\mathbf{w}_{2,i}, \boldsymbol{\Gamma}_{2,i} (1 \leq i \leq l)$ such that $\mathbf{w}_{2,i} \in \mathcal{C}_\mathbf{w}, \boldsymbol{\Gamma}_{2,i} \in \mathcal{C}_\boldsymbol{\Gamma}$ and $\|\mathbf{w}_{2,i} - \mathbf{w}_{1,i}\|_2 \leq \epsilon/2, \|\boldsymbol{\Gamma}_{2,i} - \boldsymbol{\Gamma}_{1,i}\|_F \leq \epsilon^2/(4\beta^2)$. Thus, we have

$$\text{dist}(V_1, V_2) \leq \beta \max_{1\leq i\leq l} \sqrt{\|\boldsymbol{\Gamma}_{1,i} - \boldsymbol{\Gamma}_{2,i}\|_F} + \max_{1\leq i\leq l} \|\mathbf{w}_{1,i} - \mathbf{w}_{2,i}\|_2 \leq \epsilon,$$

where the inequality holds due to (F.11). Therefore, the ϵ-covering number of function class $\check{\mathcal{V}}_h$ is bounded by $\mathcal{N}_\epsilon \leq |\mathcal{C}_\mathbf{w}|^l \cdot |\mathcal{C}_\boldsymbol{\Gamma}|^l$ and it implies

$$\log \mathcal{N}_\epsilon \leq dl \log(1 + 4L/\epsilon) + d^2 l \log(1 + 8\sqrt{d}\beta^2/\epsilon^2),$$

where the first inequality holds due to (F.12). Thus, we finish the proof of Lemma F.7. □

In addition, according to the result in Lemma F.6, the covering number $\mathcal{N}_\epsilon$ of squared function class $\mathcal{V}_h^2$ can be upper bounded by:

Lemma F.8. *For squared function class $\mathcal{V}_h^2$, we define the distance between two function V_1^2 and V_2^2 as $V_1^2, V_2^2 \in \mathcal{V}_h^2$ as $dist(V_1^2, V_2^2) = \max_s |V_1^2(s) - V_2^2(s)|$. With respect to this distance function, the ϵ-covering number $\mathcal{N}_\epsilon$ of the function class $\mathcal{V}_h^2$ can be upper bounded by*

$$\log \mathcal{N}_\epsilon \leq dl \log(1 + 8HL/\epsilon) + d^2 l \log(1 + 32\sqrt{d}H^2\beta^2/\epsilon^2)$$

Proof of Lemma F.8. For any function $V_1^2, V_2^2 \in \mathcal{V}_h^2$, the distance can be upper bounded by:

$$\begin{aligned}
\text{dist}(V_1^2, V_2^2) &= \max_{s \in \mathcal{S}} \big|V_1^2(s) - V_2^2(s)\big| \\
&= \max_{s \in \mathcal{S}} \big|V_1(s) - V_2(s)\big| \cdot \big|V_1(s) + V_2(s)\big| \\
&\leq 2H \max_{s \in \mathcal{S}} \big|V_1(s) - V_2(s)\big| \\
&= 2H \text{dist}(V_1, V_2),
\end{aligned} \tag{F.13}$$

where the inequality holds due to the fact that $0 \leq V_1(s), V_2(s) \leq H$. Thus, any $(\epsilon/2H)$-net for optimistic function class $\mathcal{V}_h$ is also a $(\epsilon/2H)$-net for the squared function class $\mathcal{V}^2$. According to Lemma F.6, the covering number of the squared function class is upper bounded by:

$$\log \mathcal{N}_\epsilon \leq dl \log(1 + 4L/\epsilon) + d^2 l \log(1 + 8\sqrt{d}\beta^2/\epsilon^2).$$

Thus, we finish the proof of Lemma F.8. □

G Auxiliary Lemmas

Lemma G.1. *For any stage $h \in [h]$ in a linear MDP and any bounded-function $V : \mathcal{S} \to [0, B]$, there always exists a vector $\mathbf{w} \in \mathbb{R}^d$ such that for all state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$[\mathbb{P}_h V](s, a) = \mathbf{w}^\top \boldsymbol{\phi}(s, a), \text{ where } \|\mathbf{w}\|_2 \leq B\sqrt{d}.$$

Proof of Lemma G.1. According to the definition of linear MDP (Assumption 3.2), we have

$$\begin{aligned}
[\mathbb{P}_h V](s, a) &= \int \mathbb{P}_h(s'|s, a) V(s') ds' \\
&= \int \boldsymbol{\phi}(s, a)^\top V(s') d\boldsymbol{\theta}_h(s') \\
&= \boldsymbol{\phi}(s, a)^\top \int V(s') d\boldsymbol{\theta}_h(s') \\
&= \boldsymbol{\phi}(s, a)^\top \mathbf{w},
\end{aligned}$$

where we set $\mathbf{w} = \int V(s') d\boldsymbol{\theta}_h(s')$. In addition, the norm of $\mathbf{w}$ is upper bounded by:

$$\Big\| \int V(s') d\boldsymbol{\theta}_h(s') \Big\| \leq \max_{s'} V(s') \cdot \sqrt{d} = B\sqrt{d}.$$

Thus, we finish the proof of Lemma G.1. □

Lemma G.2 (Azuma–Hoeffding inequality, Cesa-Bianchi and Lugosi 2006)**.** *Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}$ satisfying $|x_i| \leq M$ for some constant M, x_i is $\mathcal{G}_{i+1}$-measurable, $\mathbb{E}[x_i|\mathcal{G}_i] = 0$. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have*

$$\sum_{i=1}^n x_i \leq M\sqrt{2n \log(1/\delta)}.$$

Lemma G.3 (Lemma 11, Abbasi-Yadkori et al. 2011)**.** *Let $\{\mathbf{x}_k\}_{k=1}^K$ be a sequence of vectors in $\mathbb{R}^d$, matrix $\mathbf{\Sigma}_0$ a $d \times d$ positive definite matrix and define $\mathbf{\Sigma}_k = \mathbf{\Sigma}_0 + \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^\top$, then we have*

$$\sum_{i=1}^k \min\Big\{1, \mathbf{x}_i^\top \mathbf{\Sigma}_{i-1}^{-1} \mathbf{x}_i\Big\} \le 2\log\bigg(\frac{\det \mathbf{\Sigma}_k}{\det \mathbf{\Sigma}_0}\bigg).$$

In addition, if $\|\mathbf{x}_i\|_2 \le L$ holds for all $i \in [K]$, then

$$\sum_{i=1}^k \min\Big\{1, \mathbf{x}_i^\top \mathbf{\Sigma}_{i-1}^{-1} \mathbf{x}_i\Big\} \le 2\log\bigg(\frac{\det \mathbf{\Sigma}_k}{\det \mathbf{\Sigma}_0}\bigg) \le 2\Big(d\log\big((\mathit{trace}(\mathbf{\Sigma}_0) + kL^2)/d\big) - \log\det \mathbf{\Sigma}_0\Big).$$

Lemma G.4 (Lemma 12, Abbasi-Yadkori et al. 2011)**.** *Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are two positive definite matrices satisfying that $\mathbf{A} \succeq \mathbf{B}$, then for any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_{\mathbf{A}} \le \|\mathbf{x}\|_{\mathbf{B}} \cdot \sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.*

Lemma G.5 (Confidence Ellipsoid, Theorem 2, Abbasi-Yadkori et al. 2011)**.** *Let $\{\mathcal{G}_k\}_{k=1}^\infty$ be a filtration, and $\{\mathbf{x}_k, \eta_k\}_{k\ge1}$ be a stochastic process such that $\mathbf{x}_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma, \mathbf{\Sigma}, \epsilon > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $k \ge 1$, let $y_k = \langle \boldsymbol{\mu}^*, \mathbf{x}_k\rangle + \eta_k$ and suppose that $\eta_k, \mathbf{x}_k$ also satisfy*

$$\mathbb{E}[\eta_k|\mathcal{G}_k] = 0,\ |\eta_k| \le R,\ \|\mathbf{x}_k\|_2 \le L. \tag{G.1}$$

For $k \ge 1$, let $\mathbf{Z}_k = \lambda\mathbf{I} + \sum_{i=1}^k \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_k = \sum_{i=1}^k y_i\mathbf{x}_i$, $\boldsymbol{\mu}_k = \mathbf{Z}_k^{-1}\mathbf{b}_k$, and

$$\beta_k = R\sqrt{d\log\bigg(\frac{1 + kL^2/\lambda}{\delta}\bigg)}.$$

Then, for any $0 < \delta < 1$, we have with probability at least $1-\delta$ that,

$$\forall k \ge 1,\ \Big\|\sum_{i=1}^k \mathbf{x}_i\eta_i\Big\|_{\mathbf{Z}_k^{-1}} \le \beta_k,\ \|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_k} \le \beta_k + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2.$$

Lemma G.6 (Lemma 4.4, Zhou and Gu 2022)**.** *Let $\{\sigma_k, \widehat{\beta}_k\}_{k\ge1}$ be a sequence of non-negative numbers, $\alpha, \gamma > 0$, $\{\mathbf{a}_k\}_{k\ge1} \subset \mathbb{R}^d$ and $\|\mathbf{a}_k\|_2 \le A$. Let $\{\bar\sigma_k\}_{k\ge1}$ and $\{\widehat{\mathbf{\Sigma}}_k\}_{k\ge1}$ be (recursively) defined as follows: $\widehat{\mathbf{\Sigma}}_1 = \lambda\mathbf{I}$,*

$$\forall k \ge 1,\ \bar\sigma_k = \max\{\sigma_k, \alpha, \gamma\|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}^{1/2}\},\ \widehat{\mathbf{\Sigma}}_{k+1} = \widehat{\mathbf{\Sigma}}_k + \mathbf{a}_k\mathbf{a}_k^\top/\bar\sigma_k^2.$$

Let $\iota = \log(1 + KA^2/(d\lambda\alpha^2))$. Then we have

$$\sum_{k=1}^K \min\Big\{1, \|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_k^{-1}}\Big\} \le 2d\iota + 2\gamma^2 d\iota + 2\sqrt{d\iota}\sqrt{\sum_{k=1}^K(\sigma_k^2 + \alpha^2)}.$$