Nearly Minimax Optimal Reinforcement Learning with Linear Function Approximation

Pihe Hu [1]   **Yu Chen** [2]   **Longbo Huang** [1]

Abstract

We study reinforcement learning with linear function approximation where the transition probability and reward functions are linear with respect to a feature mapping $\phi(s, a)$. Specifically, we consider the episodic inhomogeneous linear Markov Decision Process (MDP), and propose a novel computation-efficient algorithm, LSVI-UCB$^+$, which achieves an $\widetilde{O}(Hd\sqrt{T})$ regret bound where H is the episode length, d is the feature dimension, and T is the number of steps. LSVI-UCB$^+$ builds on weighted ridge regression and upper confidence value iteration with a Bernstein-type exploration bonus. Our statistical results are obtained with novel analytical tools, including a new Bernstein self-normalized bound with conservatism on elliptical potentials, and refined analysis of the correction term. To the best of our knowledge, this is the first minimax optimal algorithm[1] for linear MDPs up to logarithmic factors, which closes the $\sqrt{Hd}$ gap between the best known upper bound of $\widetilde{O}(\sqrt{H^3d^3T})$ in (Jin et al., 2020) and lower bound of $\Omega(Hd\sqrt{T})$ for linear MDPs.

1. Introduction

Reinforcement Learning (RL) has demonstrated phenomenal empirical success in many areas, including games, robotic control, etc., where improving sample complexity is always an important topic. When the state space and action space are finite, the Markov decision process (MDP)

has been proven to achieve nearly minimax optimal sample-complexity with the generative model in (Azar et al., 2013). For harder RL settings, nearly minimax optimal sample-complexities are obtained in (Azar et al., 2017) for finite horizon episodic MDPs, and (He et al., 2021; Tossou et al., 2019) for infinite horizon MDPs[2]. However, MDPs are known to suffer from the curse-of-dimensionality due to large and possibly infinite state and action space.

Function approximation is an essential approach for handling large MDPs, which assumes that the problem structure has a compact representation concerning state or state-action pairs and enables the development of nearly minimax optimal theoretical guarantees for RL problems. Linear function approximation is one of the most fundamental function approximations. It has a significant impact since many problems can be linearly-parameterized structurally or combined with embedding, where linear MDPs and linear mixture MDPs are two of the most popular models. Representative works for these two settings are presented in Table 1, from which we find that the best regret upper bound for linear MDPs and linear mixture MDPs are $\widetilde{O}(\sqrt{H^3d^3T})$ by LSVI-UCB in (Jin et al., 2020) and $\widetilde{O}(\sqrt{H^2d^2T + H^3dT})$ by UCRL-VTR$^+$ in (Zhou et al., 2021), respectively. However, the best-known lower bound for these two settings are both $\Omega(Hd\sqrt{T})$ according to (Zhou et al., 2021), which means the following fundamental question remains open:

Does there exist a computation-efficient and minimax optimal algorithm for RL with linear function approximation?

We answer this question affirmatively for linear MDPs by designing the LSVI-UCB$^+$ algorithm, which reaches minimax optimal regret up to logarithmic factors. LSVI-UCB$^+$ overcomes barriers to nearly minimax optimality in existing works (Jin et al., 2020; Wang et al., 2020b; Zanette et al., 2020b; Wang et al., 2020a) for linear MDPs and their variations, including overly aggressive exploration and extra cost for building a uniform convergence argument by covering net. It constructs a Bernstein-type bonus to perform efficient exploration, which enables a $\sqrt{H}$ factor reduction in regret. Besides, the extra $\sqrt{d}$ dependency from building the uni-

---

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China [2]Department of Mathematical Sciences, Tsinghua University, Beijing, China. Correspondence to: Longbo Huang <longbohuang@tsinghua.edu.cn>.

[1]In this version, we fix the issue in building the over-optimistic value function $\breve{V}_{k,h}(\cdot)$ by adopting the "rare-switching" value function in (He et al., 2022). Our result still achieves minimax optimal regret for linear MDPs, with minor modifications of constant terms, compared to our last version.

[2]An algorithm is nearly minimax optimal if its sample complexity matches the minimax lower bound up to logarithmic factors.

Table 1. Theoretical results on RL with linear function approximation, where † denotes that rewards are adversarial, and the lower bound holds for both settings.

Algorithm	Regret
Linear MDP	
OPT-RLSVI (Zanette et al., 2020a)	$\widetilde{O}(H^2 d^2 \sqrt{T})$
LSVI-UCB (Jin et al., 2020)	$\widetilde{O}(\sqrt{H^3 d^3 T})$
LSVI-UCB$^+$ (**this paper**)	$\widetilde{O}(H d \sqrt{T})$
Linear Mixture MDP	
OPPO† (Cai et al., 2020)	$\widetilde{O}(\sqrt{H^3 d^2 T})$
UCRL-VTR (Ayoub et al., 2020)	$\widetilde{O}(\sqrt{H^3 d^2 T})$
UCRL-VTR$^+$ (Zhou et al., 2021)	$\widetilde{O}(\sqrt{H^2 d^2 T + H^3 dT})$
Lower Bound (Zhou et al., 2021)	$\Omega(H d \sqrt{T})$

form convergence argument can be removed by our novel technique of bounding the correction term $(\widehat{\mathbb{P}} - \mathbb{P})(\widehat{V} - V^*)$. Notably, minimax optimal algorithms (Azar et al., 2017; Zanette & Brunskill, 2019) for tabular MDPs also utilize the Bernstein-type bonus for exploration with refined consideration of the correction term in analysis. However, results for tabular MDPs cannot be applied directly to our settings due to the need for building the Bernstein inequality for vector-valued martingales in linear settings. It is worth mentioning that the above Bernstein inequality has been studied in (Zhou et al., 2021), where the UCRL-VTR$^+$ is nearly minimax optimal when $d \geq H$ for linear mixture MDPs. By contrast, our proposed LSVI-UCB$^+$ algorithm achieves nearly minimax optimal regret without requiring this assumption for linear MDPs and can be further generalized to linear mixture MDPs such that the nearly minimax optimal regret can also be obtained without $d \geq H$. This is because our proposed Bernstein inequality (Theorem 7.1) is sharper by considering the conservatism on elliptical potentials. Our contributions are summarized below:

- We develop a novel Bernstein bound of $\widetilde{O}(\sigma\sqrt{d} + R)$ for self-normalized martingales, which is sharper than the analog inequalities in (Zhou et al., 2021). By utilizing the conservatism on elliptical potentials, the bound can be further improved to $\widetilde{O}(\sigma\sqrt{d})$, which serves as a new analytical tool for RL.

- We propose the LSVI-UCB$^+$ algorithm based on a Bernstein-type exploration bonus and weighted ridge regression, with weights determined by value function variances and exploration uncertainty. LSVI-UCB$^+$ achieves an $\widetilde{O}\left(H d \sqrt{T}\right)$ regret, and is minimax optimal up to logarithmic factors in large-sample regime.

- We improve the analytical framework of statistical complexity for linear MDPs by bounding the correction term $(\widehat{\mathbb{P}} - \mathbb{P})(\widehat{V} - V)$. Combined with the Bernstein

self-normalized bound, this new analytical framework can remove the extra dependencies on H and d, which is very different from the traditional Hoeffding bound used in (Jin et al., 2020; Wang et al., 2020b;a).

Notations   Scalars are denoted in lower case letters, and vectors/matrices are denoted in boldface letters. Denote $\|\boldsymbol{x}\|_{\boldsymbol{\Lambda}}^2 = \boldsymbol{x}^\top \boldsymbol{\Lambda} \boldsymbol{x}$ for vector $\boldsymbol{x}$ and positive definite matrix $\boldsymbol{\Lambda}$. Denote $\{1, ..., n\}$ as $[n]$ and the truncated value of x in $[a, b]$ interval as $[x]_{[a,b]}$ for $a \leq b$. Define $a_n = O(b_n)$ if there exists an absolute constant $c > 0$ such that $a_n \leq cb_n$ holds for all $n \geq 1$ and define $a_n = \Omega(b_n)$ for inverse direction. $\widetilde{O}(\cdot)$ further suppresses the polylogarithmic factors in $O(\cdot)$.

2. Related Work

Linear Bandits   Linear stochastic bandits can be regarded as a special case of linearly-parameterized MDPs with episode length 1. (Dani et al., 2008) proposes an algorithm with $O(d\sqrt{T \log^3 T})$ regret by building confidence ball with Freedman inequality (Freedman, 1975). (Abbasi-Yadkori et al., 2011) improves the regret to $O(d\sqrt{T \log^2 T})$ with a self-normalized tail inequality, derived by the method of mixture (Victor et al., 2009). (Li et al., 2021) further proposes an algorithm with $O(d\sqrt{T \log T} \, \text{poly}(\log\log T))$ regret by bounding the supremum of self-normalized processes, which matches the lower bound up to a $\text{poly}(\log\log T)$ factor. The self-normalized tail inequalities for linear bandits in these works are all Hoeffding-type, i.e., only consider sub-Gaussian noises. However, for linear RL, Bernstein-type inequalities considering the sub-exponential noise, are necessary for sharper statistical results.

RL with Linear Function Approximation   Recent works have focused on designing statistically and/or computationally efficient algorithms for RL with linear function approximation. The first sample efficient algorithm is introduced by (Jiang et al., 2017), where low Bellman rank is considered. Subsequent works on this setting include (Dann et al., 2018; Sun et al., 2019). (Yang & Wang, 2019) develops the first statistically and computationally efficient algorithm for linear MDPs with a simulator, where the transition probability and reward functions are linear concerning a feature mapping $\phi(s, a)$. Subsequently, (Jin et al., 2020) considers RL settings for linear MDPs and propose LSVI-UCB algorithm reaching $\widetilde{O}(\sqrt{H^3 d^3 T})$ regret. Concurrently, (Zanette et al., 2020a) provides a Thompson sampling based algorithm with regret bound of $\widetilde{O}(d^2 H^2 \sqrt{T})$. More works generalize linear MDPs includes (Zanette et al., 2020b) for low inherent bellman error, (Wang et al., 2020b) for linear Q function, and (Wang et al., 2020a) for bounded Eluder dimension.

Another popular linearly-parameterized MDP is the linear mixture MDP, where transition probability is linear to the feature function over (state, action, next state) triples.

(Modi et al., 2020) firstly considers the statistical complexity of this setting and (Yang & Wang, 2020) provides an $\widetilde{O}(H^2 d \log T \sqrt{T})$ regret with special case of low-dimensional representation of the transition matrix. Subsequently, (Jia et al., 2020; Ayoub et al., 2020) proposes UCRL-VTR algorithm with $\widetilde{O}(\sqrt{H^3 d^2 T})$ regret, and (Cai et al., 2020) considers adversarial rewards setting, giving same regret. Notably, the nearly minimax optimal regret for linear mixture MDP is first obtained by UCRL-VTR$^+$ in (Zhou et al., 2021) under $d \geqslant H$ case.

3. Preliminaries

We consider episodic finite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_h, \{r_h\}_h\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $H \in \mathbb{Z}^+$ is the length of each episode, $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ and $r_h : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ are time-dependent transition probability and deterministic reward function. We assume that $\mathcal{S}$ is a measurable space with possibly infinite number of elements and $\mathcal{A}$ is a finite set.

For a time-inhomogeneous MDP, the policy is time-dependent, which is denoted as $\pi = \{\pi_1, ..., \pi_H\}$. Here $\pi_h(s)$ is the action that agent takes at state s at the h-th step. The value function $V_h^\pi : \mathcal{S} \rightarrow \mathbb{R}$ is the expected value of cumulative rewards received under policy π when starting from a state at h-th step, given as

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

for any $s \in \mathcal{S}, h \in [H]$. The state-action function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ gives the expected value of cumulative rewards starting from a state-action pair at h-th step, defined as

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a, \pi\right],$$

for any $(s, a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$. For any function $V : \mathcal{S} \rightarrow \mathbb{R}$, we denote $\mathbb{P}_h V(s, a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} V(s')$ and $[\mathbb{V}_h V](s, a) = \mathbb{P}_h V^2(s, a) - [\mathbb{P}_h V(s, a)]^2$, where V^2 stands for the function whose value at s is $V^2(s)$. The Bellman equation associated with a policy π is

$$Q_h^\pi(s, a) = r_h(s, a) + \mathbb{P}_h V_{h+1}^\pi(s, a)$$
$$V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$$

for any $(s, a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$. Since the action space and the episode length are both finite, there always exists an optimal policy π^* such that $V_h^\star(s) = \sup_\pi V_h^\pi(s)$ for any $s \in \mathcal{S}, h \in [H]$, with Bellman optimality equation as

$$Q_h^\star(s, a) = r_h(s, a) + \mathbb{P}_h V_{h+1}^\star(s, a)$$
$$V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s, a)$$

for any $(s, a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$.

The structural assumption we make in this paper is a linear structure in both transition and reward, which has been considered in (Yang & Wang, 2019; Jin et al., 2020; Zanette et al., 2020a). The formal definition is as follows.

Definition 3.1 (Linear MDP). A MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_h, \{r_h\}_h\}$ is a linear MDP with a known feature mapping $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$, if for any $h \in [H]$, there exist $|\mathcal{S}|$ unknown d-dimensional measures $\boldsymbol{\mu}_h = (\mu_h(1), ..., \mu_h(|\mathcal{S}|)) \in \mathbb{R}^{d \times |\mathcal{S}|}$ and an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$, such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\mathbb{P}_h(\cdot \mid s, a) = \langle \phi(s, a), \boldsymbol{\mu}_h(\cdot) \rangle, \; r_h(s, a) = \langle \phi(s, a), \boldsymbol{\theta}_h \rangle.$$

We make the following assumptions, similar to existing literature (Jin et al., 2020; Agarwal et al., 2019). Specifically, for any $h \in [H]$, (i) $\sup_{s,a} \|\phi(x, a)\|_2 \leqslant 1$, (ii) $\|\boldsymbol{\mu}_h \boldsymbol{v}\|_2 \leqslant \sqrt{d}$ for any vector $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ with $\|\boldsymbol{v}\|_\infty \leqslant 1$, (iii) $\|\boldsymbol{\theta}_h\|_2 \leqslant W$, and (iv) $r_h(s, a) \in [0, 1]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

In this paper, we focus on the setting where the reward function $\{r_h\}_{h \in [H]}$, i.e., $\{\boldsymbol{\theta}_h\}_{h \in [H]}$ is known, but our algorithm can readily be extended to handle unknown rewards.

Learning Protocol In every episode k, the learner first proposes a policy π^k based on all history information up to the end of episode $k - 1$. The learner then executes π^k to generate a single trajectory $\tau^k = \{s_h^k, a_h^k\}_{h=1}^H$ with $a_h^k = \pi_h^k(s_h^k)$ and $s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k, a_h^k)$. The goal of the learner is to learn the optimal policy by interacting with the environment during K episodes. For the k-th episode, the initial state s_1^k is picked by the adversary and the optimal policy will minimize the cumulative regret over K episodes:

$$\text{Regret}(K) = \sum_{k=1}^K [V_1^\star(s_1^k) - V_1^{\pi_k}(s_1^k)].$$

4. Strategic Exploration in Linear MDP

Section 4.1 illustrates the standard ways of strategic exploration in linear MDPs in existing works, i.e., optimistic value iteration with parameters estimated by linear ridge regression. Next, we point out in Section 4.2 barriers to minimax optimality in existing algorithms, which also helps explain our algorithm design Section 5.

4.1. Optimistic Learning in Linear MDPs

Optimistic learning evolves in an episodic fashion. In episode k, the agent first estimates unknown parameters of the linear MDP by historical data up to episodes $k - 1$. One standard approach is estimating the parameter $\boldsymbol{\omega}_h^* = \boldsymbol{\theta}_h + \boldsymbol{\mu}_h V_{h+1}^*$ by linear ridge regression, as LSVI-UCB in (Jin et al., 2020) and its variants in (Wang et al., 2020b;a),

for any $(s, a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$.

since the optimal Q function $Q_h^*(s,a) = \langle \boldsymbol{\omega}_h^*, \boldsymbol{\phi}(s,a) \rangle$ according to Proposition 2.3 in (Jin et al., 2020). Subsequently, an optimistic Q function $Q_{k,h}$ in Eq. (2) is constructed with the learned parameter $\boldsymbol{\omega}_{k,h}$ and the exploration bonus $\beta \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Lambda}_{k,h}^{-1}}$. The agent then follows a greedy policy π^k of $Q_{k,h}$ to interact with the environment and repeat the above procedure in the next episode.

We illustrate two major steps of linear ridge regression and the construction of the optimistic Q function below.

Linear Ridge Regression　To estimate the optimal $\boldsymbol{\omega}_h^*$, the following regularized least-squares problem is proposed:

$$\min_{\boldsymbol{\omega} \in \mathbb{R}^d} \sum_{i=1}^{k-1} [V_{k,h+1}(s_{h+1}^i) - \boldsymbol{\omega}^\top \boldsymbol{\phi}(s_h^i, a_h^i)]^2 + \lambda \|\boldsymbol{\omega}\|_2^2 \quad (1)$$

The closed-form solution to Eq. (1) is

$$\boldsymbol{\omega}_{k,h} = \boldsymbol{\Lambda}_{k,h}^{-1} \sum_{i=1}^{k-1} \boldsymbol{\phi}(s_h^i, a_h^i)^\top V_{k,h+1}(s_{h+1}^i)$$

where $\boldsymbol{\Lambda}_{k,h} = \sum_{i=1}^{k-1} \boldsymbol{\phi}(s_h^i, a_h^i)\boldsymbol{\phi}(s_h^i, a_h^i)^\top + \lambda \mathbf{I}$, and $V_{k,h}(\cdot)$ is given in Eq. (3).

Optimistic Estimator　$\boldsymbol{\omega}_{k,h}$ is then used to build an optimistic state-action function in Eq. (2) with exploration bonus $\beta \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Lambda}_{k,h}^{-1}}$ to encourage exploration, and optimistic value function is given in Eq. (3) as well. Notice that these two functions are built in a backwards fashion from stage H to 1, such that named as optimistic value iteration.

$$Q_{k,h}(\cdot,\cdot) = \langle \boldsymbol{\omega}_{k,h}, \boldsymbol{\phi}(\cdot,\cdot) \rangle + \beta \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Lambda}_{k,h}^{-1}} \quad (2)$$

$$V_{k,h}(\cdot) = \max_{a \in \mathcal{A}} Q_{k,h}(\cdot,a) \quad (3)$$

In particular, denote the optimistic confidence set $\mathcal{C}_{k,h} := \{\boldsymbol{\omega} : \|\boldsymbol{\omega} - \boldsymbol{\omega}_{k,h}\|_{\boldsymbol{\Lambda}_{k,h}} \leqslant \beta\}$ such that $Q_{k,h}(\cdot,\cdot) = \max_{\boldsymbol{\omega} \in \mathcal{C}_{k,h}} \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\cdot,\cdot) \rangle$. Notably, confidence set $\mathcal{C}_{k,h}$ is an ellipsoid centered at $\boldsymbol{\omega}_{k,h}$, with shape parameter $\boldsymbol{\Lambda}_{k,h}$ and radius β (usually named as the exploration radius). It can be proved that with high probability, $\boldsymbol{\omega}_h^* \in \mathcal{C}_{k,h}$ by using self-normalized tail inequalities for vector-valued martingales, e.g., Theorem 1 in (Abbasi-Yadkori et al., 2011), used broadly in the analysis of linear bandits or RL with linear function approximation. Consequently, functions in Eq. (2), (3) obtains optimism in high probability.

4.2. Barriers to Minimax Optimality

The above optimistic learning based value iteration is a commonly adopted paradigm of RL with linear function approximation in existing works, e.g., (Jin et al., 2020; Wang et al., 2020b;a). However, the best-known regret upper bound for linear MDPs is $\widetilde{O}(\sqrt{H^3 d^3 T})$ by LSVI-UCB algorithm

in (Jin et al., 2020), while the best known lower bound is $\Omega(Hd\sqrt{T})$ according to (Zhou et al., 2021). As shown in Section 6, the lower bound is tight. We analyze where the $\sqrt{Hd}$ gap comes from and then propose corresponding solutions, which immediately sheds light on designing the efficient LSVI-UCB$^+$ algorithm in the next section.

4.2.1. OVERLY AGGRESSIVE EXPLORATION

The tradeoff between exploitation and exploration is a central task for RL algorithms, implemented by designing exploration bonuses in optimistic learning. The current $\sqrt{H}$ gap stems from the overly aggressive exploration, which means that the current exploration radius $\beta = \widetilde{O}(Hd)$ in existing works, e.g., (Jin et al., 2020; Wang et al., 2020b; Ayoub et al., 2020) is too large and leads to insufficient exploitation. The underlying reason remains that a bonus with $\widetilde{O}(Hd)$ radius is intrinsically Hoeffding-type since it has the order of the magnitude of the considered martingale difference sequence (MDS). We prove that a Bernstein-type bonus, based on the variance of the MDS, combined with the Law of Total Variance (LTV) (Lattimore & Hutter, 2012), can reduce one $\sqrt{H}$ factor of regrets in linear MDPs. The motivation for this improvement comes from prior works (Azar et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019) for tabular MDPs, which succeeded in achieving $\sqrt{H}$ regret reduction by introducing a Bernstein-type bonus. For linear mixtures MDPs, UCRL-VTR$^+$ in (Zhou et al., 2021) firstly introduces a Bernstein-type bonus and also achieves a $\sqrt{H}$ regret reduction. However, a direct adaption of UCRL-VTR$^+$ in linear MDPs will not improve the regret due to the extra cost of building a uniform convergence argument.

4.2.2. EXTRA UNIFORM CONVERGENCE COST

Introducing a ε-covering net is a common approach to build a uniform convergence argument over a function class. Many algorithms for RL with linear function approximation achieve polynomial sample complexity with this approach. However, this brings extra dependency on d in the regret, as presented in prior analysis, e.g., LSVI-UCB in (Jin et al., 2020) and its variants in (Wang et al., 2020b;a). Specifically, when bounding the deviation term $(\widehat{\mathbb{P}} - \mathbb{P})\widehat{V}$, the self-normalized tail inequality cannot be applied directly since $\widehat{V}$ is not well-measurable. Prior works fix a value function $V(\cdot) \in \widehat{\mathcal{V}}$, where $\widehat{\mathcal{V}}$ is the function class contains all possible $\widehat{V}$, and build a uniform convergence argument by taking uniform bound over all functions in the ε-covering net $\widehat{\mathcal{N}}_\varepsilon$ of $\widehat{\mathcal{V}}$. In this way, a self-normalized bound concerning $\widehat{V}$ can be established (refer to proof of Lemma C.8 in Appendix for details). However, the covering number of $\widehat{\mathcal{N}}_\varepsilon$ highly depends on the feature space dimension, resulting in extra dependency on d in the regret. We propose a novel technique of bounding the deviation term by dominant term $(\widehat{\mathbb{P}} - \mathbb{P})V^*$ and the correction term $(\widehat{\mathbb{P}} - \mathbb{P})(\widehat{V} - V^*)$ sep-

arately, to remove the extra dependency on d. Note that bounding the correction term is also required for RL algorithms in tabular MDPs to achieve minimax optimality. However, adopting this idea to linear MDPs is nontrivial since we need to build the self-normalized bound for vector-valued martingale other than the well-studied scalar bound in tabular MDPs.

5. Optimal Exploration for linear MDPs

Algorithm 1 LSVI-UCB$^+$ for Linear MDPs

Require: Regularization parameter λ, $\widehat{\beta}$, $\widecheck{\beta}$.
1: **for** step $h = H, ..., 1$ **do**
2:     $\widehat{\boldsymbol{\Lambda}}_{1,h}, \widetilde{\boldsymbol{\Lambda}}_{1,h} \leftarrow \lambda \mathbf{I}$; $\widehat{\boldsymbol{\mu}}_{1,h} \leftarrow \mathbf{0}$
3: **end for**
4: $k_0 \leftarrow 0$
5: **for** episode $k = 1, ..., K$ **do**
6:     $\widehat{V}_{k,H+1}(\cdot), \widecheck{V}_{k,H+1}(\cdot) \leftarrow 0$
7:     **for** step $h = H, ..., 1$ // Optimistic value iteration **do**
8:         **if** there exists a stage $h' \in [H]$ such that $\det(\widehat{\Lambda}_{k,h'}) \geqslant 2 \det(\widehat{\Lambda}_{k_0,h'})$ **then**
9:             $\widehat{Q}_{k,h}(\cdot, \cdot) \leftarrow \min\{r_h(\cdot, \cdot) + \langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1},$
            $\boldsymbol{\phi}(\cdot, \cdot)\rangle + \widehat{\beta}_k \|\boldsymbol{\phi}(\cdot, \cdot)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}, \widehat{Q}_{k-1,h}(\cdot, \cdot), H\}$
10:             $k_0 \leftarrow k$ // Last updating episode
11:         **else**
12:             $\widehat{Q}_{k,h}(\cdot, \cdot) \leftarrow \widehat{Q}_{k-1,h}(\cdot, \cdot)$
13:         **end if**
14:         $\widecheck{Q}_{k,h}(\cdot, \cdot) = r_h(\cdot, \cdot) + \langle \widehat{\boldsymbol{\mu}}_{k,h} \widecheck{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(\cdot, \cdot)\rangle -$
        $\widecheck{\beta}_k \|\boldsymbol{\phi}(\cdot, \cdot)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$ // Pessimistic Q function
15:         $\widehat{V}_{k,h}(\cdot) \leftarrow \max_{a \in \mathcal{A}} \widehat{Q}_{k,h}(\cdot, a)$
16:         $\widecheck{V}_{k,h}(\cdot) \leftarrow \max\{\max_{a \in \mathcal{A}} \widecheck{Q}_{k,h}(\cdot, a), 0\}$
17:         $\pi_h^k(\cdot) \leftarrow \arg\max_{a \in \mathcal{A}} \widehat{Q}(\cdot, a)$
18:     **end for**
19:     Receive the initial state s_1^k.
20:     **for** step $h = 1, ..., H$ **do**
21:         $a_h^k \leftarrow \pi_h^k(s_h^k)$, and observe $s_{h+1}^k \sim \mathbb{P}_h(\cdot | s_h^k, a_h^k)$.
22:         $\widetilde{\sigma}_{k,h} \leftarrow$
        $\sqrt{\max\{H, Hd^3 E_{k,h}, [\widehat{\mathbb{V}}_{k,h} \widehat{V}_{k,h+1}](s_h^k, a_h^k) + U_{k,h}\}}$
23:         $\widetilde{\boldsymbol{\Lambda}}_{k+1,h} \leftarrow \widetilde{\boldsymbol{\Lambda}}_{k,h} + \widetilde{\sigma}_{k,h}^{-2} \boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top$
24:         **if** $\|\widetilde{\sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widetilde{\boldsymbol{\Lambda}}_{k,h}^{-1}} \leqslant 1/(H^3 d^5)$ **then**
25:             $\varsigma_{k,h} \leftarrow \sqrt{H}$
26:         **else**
27:             $\varsigma_{k,h} \leftarrow H^2 \sqrt{d^5}$ // Enlarge ς_k if necessary
28:         **end if**
29:         $\widehat{\sigma}_{k,h} \leftarrow$
        $\sqrt{\max\{\varsigma_{k,h}^2, d^3 H E_{k,h}, [\widehat{\mathbb{V}}_{k,h} \widehat{V}_{k,h+1}](s_h^k, a_h^k) + U_{k,h}\}}$
30:         $\widehat{\boldsymbol{\Lambda}}_{k+1,h} \leftarrow \widehat{\boldsymbol{\Lambda}}_{k,h} + \widehat{\sigma}_{k,h}^{-2} \boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top$
31:         $\widehat{\boldsymbol{\mu}}_{k+1,h} \leftarrow \widehat{\boldsymbol{\Lambda}}_{k+1,h}^{-1} \sum_{i=1}^k \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)\boldsymbol{\delta}(s_{h+1}^i)^\top$
32:     **end for**
33: **end for**

In this section, we present the proposed LSVI-UCB$^+$ algo-rithm (Algorithm 1), where the optimistic value iteration is performed in Lines 5-18, and the learned policy is executed in Line 21. The remaining parts of Algorithm 1 are responsible for estimating parameter $\boldsymbol{\mu}_h$ by linear weighted ridge regression. Specifically, the estimated variance is given in Line 29, whose lower bound is controlled in Lines 24-28, and the solution to the regression is given in Line 31.

LSVI-UCB$^+$ is an optimistic algorithm similar to existing works (Yang & Wang, 2019; Jin et al., 2020; Ayoub et al., 2020), but upgrading the Hoeffding-type bonus to a carefully designed Bernstein-type one. The exploration radius in LSVI-UCB$^+$ is proportional to the standard deviation of the optimal value function conditioned on some state-action pair, which accounts for two key novelties of LSVI-UCB$^+$:

(i) We replace the linear ridge regression in prior works (Yang & Wang, 2019; Jin et al., 2020; Ayoub et al., 2020) with a carefully designed weighted version such that LTV can be applied. Note that the linear weighted ridge regression estimator was originally built for linear bandits with heteroscedastic noises, e.g., (Lattimore et al., 2015; Kirschner & Krause, 2018). Besides, the regression is performed to estimate $\boldsymbol{\mu}_h$, i.e., transition matrix, instead of estimating indirect variables, e.g., $\boldsymbol{\omega}_h^*$ of LSVI-UCB in (Jin et al., 2020).

(ii) A variance estimator, based on the estimated parameter $\widehat{\boldsymbol{\mu}}_{k,h}$, is built for the optimal value function to determine the weights in regression. UCRL-VTR$^+$ in (Zhou et al., 2021) also introduces weighted ridge regression for linear mixture MDPs, and weights are determined by variances of the constructed optimistic value function. However, the weights in LSVI-UCB$^+$ are very different from those in UCRL-VTR$^+$, since our variances are estimated with respect to the optimal value function, not the constructed value function.

5.1. Linear Weighted Ridge Regression

Denote $\boldsymbol{\delta}(s) \in \mathbb{R}^{|\mathcal{S}|}$ as a one-hot vector that is zero everywhere except that the entry corresponding to state s is one, and define $\boldsymbol{\epsilon}_h^k := \mathbb{P}_h(\cdot | s_h^k, a_h^k) - \boldsymbol{\delta}(s_{h+1}^k)$. Since $\mathbb{E}[\boldsymbol{\epsilon}_h^k | \mathcal{F}_{k,h}] = 0$, $\boldsymbol{\delta}(s_{h+1}^k)$ is an unbiased estimate of $\mathbb{P}_h(\cdot | s_h^k, a_h^k) = \boldsymbol{\mu}_h^\top \boldsymbol{\phi}(s_h^k, a_h^k)$. Thus, $\boldsymbol{\mu}_h$ can be learned via regression from $\boldsymbol{\phi}(s_h^k, a_h^k)$ to $\boldsymbol{\delta}(s_{h+1}^k)$. In addition, samples are normalized by the estimated standard deviation $\widehat{\sigma}_{k,h}$. Thus, the estimated parameter $\widehat{\boldsymbol{\mu}}_{k,h}$ in Line 31 of Algorithm 1 is the solution to the following weighted ridge regression problem:

$$\min_{\boldsymbol{\mu} \in \mathbb{R}^{d \times |\mathcal{S}|}} \sum_{i=1}^{k-1} \left\| [\boldsymbol{\mu}_h^\top \boldsymbol{\phi}(s_h^k, a_h^k) - \boldsymbol{\delta}(s_{h+1}^i)] \widehat{\sigma}_{i,h}^{-1} \right\|_2^2 + \lambda \|\boldsymbol{\mu}\|_F^2,$$

where $\| \cdot \|_F$ denotes Frobenius norm. The solution is

$$\widehat{\boldsymbol{\mu}}_{k,h} = \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)\boldsymbol{\delta}(s_{h+1}^i)^\top, \qquad (4)$$

where $\widehat{\boldsymbol{\Lambda}}_{k,h} = \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda\mathbf{I}$. Thus, the estimated transition probability is denoted as

$$\widehat{\mathbb{P}}_{k,h}(\cdot \mid s, a) = \widehat{\boldsymbol{\mu}}_{k,h}^\top \phi(s, a)$$

for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. After estimating the transition matrix, Lines 8-13 in Algorithm 1 constructs an optimistic state-action function, which is equivalent to

$$\widehat{Q}_{k,h}(\cdot, \cdot) = \min\{ \max_{\boldsymbol{\mu}\in\widehat{\mathcal{C}}_{k_0,h}} r_h(\cdot, \cdot) + \langle\boldsymbol{\mu}\widehat{\boldsymbol{V}}_{k_0,h+1}, \phi(\cdot, \cdot)\rangle, H\},$$

where k_0 is the updating episode and the optimistic confidence set is given by

$$\widehat{\mathcal{C}}_{k_0,h} := \{\boldsymbol{\mu} : \|(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k_0,h})\widehat{\boldsymbol{V}}_{k_0,h+1}\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}} \leqslant \widehat{\beta}\},$$

and $\widehat{\beta}$ is the exploration radius. The updating condition in Line 8 of Algorithm 1 is inspired by (He et al., 2022), which utilizes a "rare switching" mechanism (detailed in Lemma F.8 and F.11 in Appendix) to ensure a small covering number of considered optimistic value function classes. In addition, the pessimistic state-action function in Line 14 is equivalent to

$$\widecheck{Q}_{k,h}(\cdot, \cdot) = \min_{\boldsymbol{\mu}\in\widecheck{\mathcal{C}}_{k,h}} r_h(\cdot, \cdot) + \langle\boldsymbol{\mu}\widecheck{\boldsymbol{V}}_{k,h+1}, \phi(\cdot, \cdot)\rangle,$$

where the pessimistic confidence set is

$$\widecheck{\mathcal{C}}_{k,h} := \{\boldsymbol{\mu} : \|(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h})\widecheck{\boldsymbol{V}}_{k,h+1}\|_{\widecheck{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widecheck{\beta}\}.$$

Subsequently, optimistic value function $\widehat{V}_{k,h}(\cdot)$ and pessimistic value function $\widecheck{V}_{k,h}(\cdot)$ can be defined. Note that $\widehat{V}_{k,h}(\cdot)$ in Algorithm 1 is strictly decreasing in k, which ensures that the optimistic value function approaches the optimal value function $V_h^*(\cdot)$ almost surely. Besides, the pessimistic value function $\widecheck{V}_{k,h}(\cdot)$ is required for estimating the variance upper bound later.

5.2. Variance Estimation

After estimating the transition matrix in Eq. (4), LSVI-UCB$^+$ estimates the variance of the optimal value function $[\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)$ and the variance of sub-optimality gap $[\mathbb{V}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)$. This is a major difference with prior UCRL-VTR$^+$ algorithm in (Zhou et al., 2021) for linear mixture MDPs, which only estimates the variance of the constructed optimistic value function $[\mathbb{V}_h \widehat{V}_{k,h+1}](s_h^k, a_h^k)$. The purpose to estimate these two variances remains that we utilize Bernstein self-normalized tail inequality in Theorem 7.1 to bound the dominant term $[(\widehat{\mathbb{P}}_{k,h} - \mathbb{P}_h)V_{h+1}^*](s_h^k, a_h^k)$ and the correction term $[(\widehat{\mathbb{P}}_{k,h} - \mathbb{P}_h)(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)$ separately to remove the extra dependency of regrets, such that we need to estimate these two variance, which are illusated below.

Variance of Optimal Value Function   We first consider the case where the transition matrix and optimal value function $V_{h+1}^*(\cdot)$ were given. In this case, the variance of the optimal value function is given by

$$[\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k) = \mathbb{P}_h V_{h+1}^{*}{}^2(s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^*(s_h^k, a_h^k)]^2$$

However, only empirical estimation $\widehat{\boldsymbol{\mu}}_{k,h}$ and optimistic value function $\widehat{V}_{k,h}$ are obtainable, which means we only have the empirical variance of the optimistic value function:

$$[\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k) = \widehat{\mathbb{P}}_{k,h}\widehat{V}_{k,h+1}^2(s_h^k, a_h^k)_{[0,H^2]} \\ - [\widehat{\mathbb{P}}_{k,h}\widehat{V}_{k,h+1}(s_h^k, a_h^k)_{[0,H]}]^2. \quad (5)$$

To ensure the accuracy of the estimation, we introduce an offset term $U_{k,h}$ to guarantee that $|[\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k) - [\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k)| \leqslant U_{k,h}$ with high probability. Moreover, the exact form of offset term $U_{k,h}$ is specified in Lemma 7.5, which requires accessing the pessimistic value functions as detailed in Lemma C.12 in Appnedix.

Variance of Sub-optimality Gap   In particular, we try to build a upper bound for the variance of the sub-optimality gap, which is given as

$$[\mathbb{V}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)$$
$$= [\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)^2](s_h^k, a_h^k) - \left[[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)\right]^2$$
$$\leqslant [\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)^2](s_h^k, a_h^k) \leqslant H[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)$$
$$\leqslant H[\mathbb{P}_h(\widehat{V}_{k,h+1} - \widecheck{V}_{k,h+1})](s_h^k, a_h^k)$$

where the second and last inequalities holds by the optimism and pessimism of $\widehat{V}_{k,h+1}$ and $\widecheck{V}_{k,h+1}$, respectively. Thus, it suffices to upper bound the deviation $[\mathbb{P}_h(\widehat{V}_{k,h+1} - \widehat{V}_{k,h+1})](s_h^k, a_h^k)$. In addition, the upper bound of the variance of the sub-optimality gap is denoted as $E_{k,h}$, specified in Lemma 7.5.

Putting two variances together, the weight $\widehat{\sigma}_{k,h}$ in Algorithm 1 is given by

$$\widehat{\sigma}_{k,h} = \sqrt{\max\{\varsigma_{k,h}^2, dE_{k,h}, [\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k) + U_{k,h}\}},$$

which is the maximum over the weight lower bound $\varsigma_{k,h}$, the variance upper bound of the optimal value function, the variance of the sub-optimality gap with a factor d scaling. Here $\varsigma_{k,h}$ controls the lower bound of $\widehat{\sigma}_{k,h}$ and is dynamically determined in Lines 24-28. In particular, we try to keep the magnitude of the considered MDS to be small by adaptively enlarging $\varsigma_{k,h}$, which is detailed in Remark 7.4.

6. Main Results

This section presents the results of the statistical, space, and computational complexities of the LSVI-UCB$^+$ algorithm.

In particular, LSVI-UCB$^+$ firstly reaches nearly minimax optimal regret in linear MDPs, while the space and computational complexities are no worse than prior works.

6.1. Statistical Complexity

We first present the regret upper bound of LSVI-UCB$^+$ in Theorem 6.1.

Theorem 6.1 (Regret Upper Bound). *Set $\lambda = 1/(H^2\sqrt{d})$. Then, with probability at least $1 - 10\delta$, the regret of LSVI-UCB$^+$ is upper bounded by*

$$\text{Regret}(K) = \tilde{O}\left(Hd\sqrt{T} + H^6 d^9\right), \qquad (6)$$

where $T = KH$.

Proof Sketch. We prove the result conditioning on the conclusion of Lemma 7.5. Initially, with the standard regret decomposition, we can show that the total regret is bounded by the summation of the exploration bonus, i.e.,

$$\text{Regret}(K) \leqslant \sum_{k=1}^{K}\sum_{h=1}^{H} \hat{\beta}\|\phi(s_h^k, a_h^k)\|_{\hat{\Lambda}_{k,h}^{-1}}$$

$$= \sum_{k=1}^{K}\sum_{h=1}^{H} \hat{\beta}\hat{\sigma}_{k,h}\|\hat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\hat{\Lambda}_{k,h}^{-1}}$$

$$\leqslant \hat{\beta}\underbrace{\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\hat{\sigma}_{k,h}^2}}_{\tilde{O}\left(\sqrt{HT+c\sqrt{T}}\right)}\underbrace{\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\|\hat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\hat{\Lambda}_{k,h}^{-1}}^2}}_{\tilde{O}\left(\sqrt{Hd}\right)} \qquad (7)$$

where the second inequality holds by Cauchy-Schwarz inequality and c is a constant. The summation of $\|\hat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\hat{\Lambda}_{k,h}^{-1}}^2$ can be addressed by Elliptical Potential Lemma (Lemma F.5 in Appendix), and the summation of $\hat{\sigma}_{k,h}^2$ can be bounded by

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\hat{\sigma}_{k,h}^2 \leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}\varsigma_{k,h}^2 + \sum_{k=1}^{K}\sum_{h=1}^{H}[dE_{k,h} + U_{k,h}]$$

$$+ \sum_{k=1}^{K}\sum_{h=1}^{H}[\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1}](s_h^k, a_h^k) \leqslant \tilde{O}\left(HT + c\sqrt{T}\right),$$

where the first inequality holds by definition of $\hat{\sigma}_{k,h}$, and the second inequality holds by $\sum_{k=1}^{K}\sum_{h=1}^{H}\varsigma_{k,h}^2 \leqslant \tilde{O}(HT)$ due to the conservatism of elliptical potentials, $\sum_{k=1}^{K}\sum_{h=1}^{H}[dE_{k,h}+U_{k,h}] \leqslant \tilde{O}(c\sqrt{T})$ due to the Elliptical Potential Lemma, and $\sum_{k=1}^{K}\sum_{h=1}^{H}[\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1}](s_h^k, a_h^k) \leqslant \tilde{O}(HT)$ due to the LTV. Besides, the exploration radius $\hat{\beta} = \tilde{O}(\sqrt{d})$, which determined by the upper bound of $(\hat{\mathbb{P}}_{k,h} - \mathbb{P}_h)\hat{V}_{k,h+1}(s_h^k, a_h^k)$, detailed in Section 7.3. The full proof is given in Appendix D. □

Theorem 6.1 is proved under the event that the optimistic confidence set $\hat{\mathcal{C}}_{k,h}$ holds, which is built in Lemma 7.5.

In addition, we find that the exploration radius $\hat{\beta}$ of the optimistic confidence set $\hat{\mathcal{C}}_{k,h}$ determines the sharpness of the final regret, as shown in Eq. (7) .

Remark 6.2. When[3] $T \geqslant H^{10}d^{16}$, the regret in Eq. (6) can be simplified to $\tilde{O}(Hd\sqrt{T})$, which improves the $\tilde{O}\left(\sqrt{H^3 d^3 T}\right)$ regret of LSVI-UCB (Jin et al., 2020) by a factor of $\sqrt{Hd}$. Moreover, our algorithm design an analytical tools including Theorem 7.1 and Lemma 7.3 in next sections can further improve the regret bound of UCRL-VTR$^+$ in (Zhou et al., 2021) for linear mixture MDPs to $\tilde{O}(Hd\sqrt{T})$ from existing $\tilde{O}(\sqrt{H^2 d^2 T + H^3 dT})$, such that it is minimax optimal up to logarithmic factor without large dimension assumption that $d \geqslant H$ in (Zhou et al., 2021).

Lower Bound We formalize a linear MDP instance in Appendix E to establish an $\Omega(Hd\sqrt{T})$ regret lower bound of linear MDPs. This linear MDP instance is firstly proposed in Remark 23 in (Zhou et al., 2021), which shares the same regret lower bound of a linear mixture MDP instance. This class of MDP is hard due to the intrinsical sparsity of reward and indistinguishability of large action space, which can be regarded as an extension of hard instances in linear bandits literature (Dani et al., 2008; Lattimore & Szepesvári, 2020). According to Theorem 8 in (Zhou et al., 2021), linear mixture MDPs have regret lower bound of $\Omega(Hd\sqrt{T})$. Thus linear MDPs have the same regret lower bound. The lower bound, together with the upper bound of LSVI-UCB$^+$ in Theorem 6.1 show that LSVI-UCB$^+$ is minimax optimal up to logarithmic factors when $T \geqslant \max\{H^4 d^{10}, H^5 d^6\}$.

6.2. Space and Computational Complexities

As stated above, LSVI-UCB$^+$ reaches minimax optimal regret up to logarithmic factors, which is also computationally efficient. In particular, the space and computational complexities of LSVI-UCB$^+$ are briefly stated below, which are both the same as LSVI-UCB in (Jin et al., 2020).

Space Complexity Although $\hat{\mu}_{k,h} \in \mathbb{R}^{d \times |\mathcal{S}|}$ and $|\mathcal{S}|$ can be infinitely large, we do not store it explicitly, as Algorithm 1 only utilizes indirect variables $\hat{\mu}_{k,h}V = \hat{\Lambda}_{k,h}^{-1}\sum_{i=1}^{k-1}\hat{\sigma}_{i,h}^{-2}\phi(s_h^i, a_h^i)V(s_{h+1}^i)$ where $V \in \{\hat{V}_{k,h+1}, \hat{V}_{k,h+1}^2, \check{V}_{k,h+1}\}$. In episode $k \in [K]$, Algorithm 1 only stores $\hat{\mu}_{k,h}V, \hat{\Lambda}_{k,h}, \tilde{\Lambda}_{k,h}, \hat{\sigma}_{k,h}, \tilde{\sigma}_{k,h}$ for all $h \in [H]$, and $\{\phi(s_h^{k'}, a)\}_{a \in \mathcal{A}}$ for all $(k', h) \in [k] \times [H]$, which means LSVI-UCB$^+$ requires $O(d^2 H + d|\mathcal{A}|T)$ space.

Computational Complexity Assume $\hat{\mu}_{k,h}V$ is given for some $V \in \{\hat{V}_{k,h+1}, \hat{V}_{k,h+1}^2, \check{V}_{k,h+1}\}$, then each evaluation of $V(s)$ takes $O(d^2|\mathcal{A}|)$ operations. Thus, calculating

---

[3]Large-sample regime conditions are required in many RL algorithms to obtain satisfactory statistical complexities, e.g. UCRL-VTR$^+$ in (Zhou et al., 2021) requires $T \geqslant d^4 H^2 + d^3 H^3$.

$\widehat{\boldsymbol{\mu}}_{k,h}\boldsymbol{V}$ takes $O(d^2|\mathcal{A}|)K$ operations. Besides, $\widehat{\boldsymbol{\Lambda}}_{k,h}, \widetilde{\boldsymbol{\Lambda}}_{k,h}$ can be computed by Sherman-Morrison formula (Hager, 1989) with $O(d^2)$ operations and other steps take less operations. Thus, LSVI-UCB$^+$ has a running time of $O(d^2|\mathcal{A}|KT)$, which is computationally efficient since its running time is polynomial on $d, K, H, |\mathcal{A}|$, and does not depend on $|\mathcal{S}|$, which can be possibly infinite.

7. Mechanism Towards Minimax Optimality

In this section, we highlight our technical contributions in building the sharp optimistic confidence set $\widehat{\mathcal{C}}_{k,h}$. We first present two novel analytical tools, a sharp Bernstein self-normalized tail inequality for vector-valued martingales in Section 7.1, the conservatism of elliptical potentials in Section 7.2. Together, these two analytical tools remove the additional dependency of regret on H in the regret of the LSVI-UCB$^+$ algorithm. In addition, we also upper bounds the correction term of the form $(\widehat{\mathbb{P}} - \mathbb{P})(\widehat{V} - V^*)$ to avoid extra cost from the covering net such that the additional dependency of regret on d is removed as well. Consequently, the sharp confidence set $\widehat{\mathcal{C}}_{k,h}$ is built in Lemma 7.5 in Section 7.3. These technical contributions together enable LSVI-UCB$^+$ to achieve nearly minimax optimal regret and have the potential to improve other statistical results of algorithms for RL with linear function approximation.

7.1. Sharp Bernstein Self-normalized Bound

Most existing self-normalized concentrations used in prior works for RL with linear function approximation (Jin et al., 2020; Wang et al., 2020b;a; Ayoub et al., 2020) are all Hoeffding-type, i.e., they consider sub-Gaussian noises. Our self-normalized bound below considers sub-exponential noises, which is a Bernstein-type one.

Theorem 7.1 (Bernstein self-normalized bound). *Let* $\{\mathcal{G}_t\}_{t=1}^\infty$ *be a filtration, and* $\{\boldsymbol{x}_t, \eta_t\}_{t \geqslant 1}$ *be a stochastic process such that* $\boldsymbol{x}_t \in \mathbb{R}^d$ *is* $\mathcal{G}_t$*-measurable and* $\eta_t \in \mathbb{R}$ *is* $\mathcal{G}_{t+1}$*-measurable. Define* $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \boldsymbol{x}_i\boldsymbol{x}_i^\top$ *for* $t \geqslant 1$ *and* $\mathbf{Z}_0 = \lambda\mathbf{I}$. *If* $\|\boldsymbol{x}_t\|_2 \leqslant L$, *and* η_t *satisfies* $\mathbb{E}[\eta_t \mid \mathcal{G}_t] = 0$, $\mathbb{E}[\eta_t^2 \mid \mathcal{G}_t] \leqslant \sigma^2$, *and* $|\eta_t \cdot \min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\}| \leqslant R$ *for all* $t \geqslant 1$. *Then, for any* $0 < \delta < 1$, *with probability at least* $1 - \delta$, *we have:*

$$\forall t > 0, \left\|\sum_{i=1}^t \boldsymbol{x}_i\eta_i\right\|_{\mathbf{Z}_t^{-1}} \leqslant \widetilde{O}(\sigma\sqrt{d} + R).$$

Proof. Please refer to Appendix B. $\square$

Remark 7.2. The proof of Theorem 7.1 in Appendix B shows that bounding the self-normalized vector-valued martingales is equivalent to bounding a scalar-valued MDS $\{\eta_t \cdot \min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\}, \mathcal{G}_{t+1}\}$, where η_t is scaled by the factor of $\min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\}$. In particular, $\|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}} =$

$\sqrt{\boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1}\boldsymbol{x}_t}$ is denoted as the elliptical potential, which is common in online learning literature (Cesa-Bianchi & Lugosi, 2006). Notice that Elliptical Potential Lemma shows that $\min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\}$ can be roughly regarded as an attenuated sequence. Theorem 7.1 looks similar to but is sharper than Theorem 2 in (Zhou et al., 2021), because it pay extra attentions on elliptical potentials $\|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}$. However, the scaling factor $\min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\}$ is crudely deflated to 1 in Theorem 2 in (Zhou et al., 2021), such that the attenuation of the MDS is neglected, which is highlighted in Lemma D.7.

7.2. Conservatism of Elliptical Potentials

Notice the self-normalized bound in Theorem 7.1 will determine the order of exploration radius $\widehat{\beta}$. We try to keep the second term R, the magnitude of the MDS, in Theorem 7.1 smaller than the first $\sigma\sqrt{d}$ by utilizing the conservatism of elliptical potentials. Specifically, the following lemma characterizes the conservatism of elliptical potentials, i.e., elliptical potentials are usually small. This lemma is firstly proposed at Exercise 19.3 in (Lattimore & Szepesvári, 2020) for case $c = 1$, and we generalize it to case $c > 0$.

Lemma 7.3 (Elliptical Potentials are Usually Small). *Given* $\lambda > 0$ *and sequence* $\{\boldsymbol{x}_t\}_{t=1}^T \subset \mathbb{R}^d$ *with* $\|\boldsymbol{x}_t\|_2 \leqslant L$ *for all* $t \in [T]$, *define* $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \boldsymbol{x}_i\boldsymbol{x}_i^\top$ *for* $t \geqslant 1$ *and* $\mathbf{Z}_0 = \lambda\mathbf{I}$. *The number of times* $\|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}} \geqslant c$ *is at most*

$$\frac{3d}{\log(1+c^2)}\log\left(1 + \frac{L^2}{\lambda\log(1+c^2)}\right)$$

for any $t \in [T]$, *where* $c > 0$ *is a constant.*

Proof. Please refer to Lemma D.7 in Appendix. $\square$

On the one hand, for some stage $h \in [H]$, the noise $\eta_k = \widehat{\sigma}_{k,h}^{-1}\boldsymbol{V}^\top\boldsymbol{\epsilon}_h^k$ for some value function V as detailed in Appendix C. On the other hand, in Theorem 7.1, R is the absolute bound of $|\eta_k \cdot \min\{1, \|\widehat{\sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}\}|$. Lemma 7.3 reveals that R is intrinsically small since the elliptical potential $\|\widehat{\sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}^2$ is small in most episodes. In addition, we only need to enlarge the lower bound of $\widehat{\sigma}_{k,h}$, i.e., $\varsigma_{k,h}$, when $\|\widehat{\sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}^2$ is large such that R can remain small uniformly, which is detailed in the following remark.

Remark 7.4. Lines 22-29 of Algorithm 1 ensure the following facts for any $(k, h) \in [K] \times [H]$ by introducing indicator variable $\|\widetilde{\sigma}_{i,h}^{-1}\boldsymbol{\phi}(s_h^i, a_h^i)\|_{\widetilde{\boldsymbol{\Lambda}}_{i,h}^{-1}}$:

(i) In most cases, we have $\|\widetilde{\sigma}_{i,h}^{-1}\boldsymbol{\phi}(s_h^i, a_h^i)\|_{\widetilde{\boldsymbol{\Lambda}}_{i,h}^{-1}} \leqslant 1/(Hd^3)$, then $\varsigma_{k,h} = \sqrt{H}$ such that $\widehat{\sigma}_{i,h} = \widetilde{\sigma}_{i,h}$. We can prove that the elliptical potential $\|\widehat{\sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$ is small;

(ii) Otherwise, $\varsigma_{k,h} = H\sqrt{d^3}$, such that $\hat{\sigma}_{k,h} \geqslant H\sqrt{d^3}$. In this case, the R is still small since $\hat{\sigma}_{k,h}$ is large.

Notice that always enlarging $\varsigma_{k,h}$ for any $k \in [K]$ is a simple method to keep R small, but contributing to final regret linearly since it is an additive term in $\sum_{k=1}^{K}\sum_{h=1}^{H}\hat{\sigma}_{k,h}$. Nevertheless, the enlarging operation in case (ii) only contributes an additive constant term to the regret, since the elliptical potential $\|\tilde{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}}^2$ is small in most episodes such that the enlarging operation in case (ii) happens rarely.

As a consequence, the R in Theorem 7.1 can be controlled to be smaller than the $\sigma\sqrt{d}$ in the LSVI-UCB$^+$ algorithm, such that the exploration radius $\hat{\beta}$ in LSVI-UCB$^+$ is $\tilde{O}(\sqrt{d})$. However, analog Bernstein self-normalized bounds, such as Theorem 2 in (Zhou et al., 2021) and Theorem 1 in (Faury et al., 2020), cannot lead to such exploration radius in the SVI-UCB$^+$ algorithm, while that of Theorem 2 in (Zhou et al., 2021) is $\tilde{O}(\sqrt{d} + \sqrt{H}d^2)$, and Theorem 1 in (Faury et al., 2020) is $\tilde{O}(\sqrt{H\sqrt{d^5}})$.

7.3. Building Confidence Set with Correction Term

This subsection explains critical steps of building a sharp optimistic confidence set $\hat{\mathcal{C}}_{k,h}$ with the correction term. Specifically, the exploration bonus $\hat{\beta}\|\phi(s_h^k, a_h^k)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$ is the upper bound of the deviation term $[(\hat{\mathbb{P}}_{k,h} - \mathbb{P}_h)\hat{V}_{k,h+1}](s_h^k, a_h^k)$, which is the basis of optimistic learning. Specifically, the deviation term can be decomposed by triangle inequality as the sum of dominant term and correction term:

$$\underbrace{[(\hat{\mathbb{P}}_{k,h} - \mathbb{P}_h)V_{h+1}^*](s_h^k, a_h^k)}_{\text{Dominant term with respect to }\hat{V}_{h+1}^*}$$

$$+ \underbrace{[(\hat{\mathbb{P}}_{k,h} - \mathbb{P}_h)(\hat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)}_{\text{Correction Term}}$$

$$\leqslant \|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})V_{h+1}^*\|_{\hat{\boldsymbol{\Lambda}}_{k,h}}\|\phi(s_h^k, a_h^k)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$+ \|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})(\hat{V}_{k,h+1} - V_{h+1}^*)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}}\|\phi(s_h^k, a_h^k)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$\leqslant \hat{\beta}^{(1)}\|\phi(s_h^k, a_h^k)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \hat{\beta}^{(2)}\|\phi(s_h^k, a_h^k)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$= \hat{\beta}\|\phi(s_h^k, a_h^k)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}},$$

where the first inequality holds by Cauchy-Schwarz inequality, the second inequality holds since

$$\hat{\mathcal{C}}_{k,h}^{(1)} = \{\boldsymbol{\mu} : \|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})V_{h+1}^*\|_{\hat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \hat{\beta}^{(1)}\}$$

$$\hat{\mathcal{C}}_{k,h}^{(2)} = \{\boldsymbol{\mu} : \|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})(\hat{V}_{k,h+1} - V_{h+1}^*)\|_{\hat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \hat{\beta}^{(2)}\},$$

and the last equality holds since $\hat{\beta} = \hat{\beta}^{(1)} + \hat{\beta}^{(2)}$. In the following, we briefly illustrate how to use Theorem 7.1 to build confidence sets $\hat{\mathcal{C}}_{k,h}^{(1)}$ and $\hat{\mathcal{C}}_{k,h}^{(2)}$. Initially, by

$$\|(\hat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h)\boldsymbol{V}\|_{\hat{\boldsymbol{\Lambda}}_{k,h}} \lesssim \|\sum_{i=1}^{k-1}\hat{\sigma}_{i,h}^{-2}\phi(s_h^i, a_h^i)\boldsymbol{\epsilon}_h^{i\top}\boldsymbol{V}\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$(8)$$

for some fixed function V in Lemma F.9, building a confidence set with respect to V is equivalent to building a self-normalized bound for $\|\sum_{i=1}^{k-1}\hat{\sigma}_{i,h}^{-2}\phi(s_h^i, a_h^i)\boldsymbol{\epsilon}_h^{i\top}\boldsymbol{V}\|_{\hat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$.

Building $\hat{\mathcal{C}}_{k,h}^{(1)}$: We build confidence set $\hat{\mathcal{C}}_{k,h}^{(1)}$ by applying the Bernstein self-normalized inequality in Theorem 7.1 with dynamic control of MDS magnitude, highlighted in Remark 7.2. Thus, $U_{k,h}$ is specified in Lemma 7.5 to guarantee that $\hat{\sigma}_{k,h}$ upper bounds $\mathbb{V}_h V_{h+1}^*(s_h^k, a_h^k)$, and $\varsigma_{k,h}$ is set dynamically. Besides, the uniform convergence argument by covering net in (Jin et al., 2020; Wang et al., 2020b;a) is not required, since now $V = V_{h+1}^*$ in Eq. (8) is a fixed function and there is no measurability issue. Consequently, we get $\hat{\beta}_k^{(1)} = \tilde{O}(\sqrt{d})$, which is detailed in Lemma C.15.

Building $\hat{\mathcal{C}}_{k,h}^{(2)}$: We apply also Theorem 7.1 with dynamic control of MDS magnitude to build $\hat{\mathcal{C}}_{k,h}^{(2)}$ as well. Similarly, $E_{k,h}$ is specified in Lemma 7.5 to guarantee that $\hat{\sigma}_{k,h}$ upper bounds $[\mathbb{V}_h(\hat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)$, and $\varsigma_{k,h}$ is set dynamically to keep the MDS magnitude R in Theorem 7.1 small. Since now $V = \hat{V}_{k,h+1} - V_{h+1}^*$ in Eq. (8) suffers from the measurability issue, a uniform convergence argument by covering net is still required, which bring extra dependency on d in the exploration radius $\hat{\beta}^{(2)}$. That is why we enlarge $E_{k,h}$ with a d factor in $\hat{\sigma}_{k,h}$.

Putting everything together gives the following key technical lemma that builds the sharp optimistic confidence set $\hat{\mathcal{C}}_{k,h}$.

Lemma 7.5. *Set $\hat{\beta} = \hat{\beta}^{(1)} + \hat{\beta}^{(2)}$, then there exists an absolute constant $c > 0$ such that for any $\delta \in (0, 1)$, with probability at least $1 - 7\delta$, we have that simultaneously for any $k \in [K]$ and any $h \in [H]$,*

$$\boldsymbol{\mu}_h \in \hat{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h},$$

and $\left|[\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\right| \leqslant U_{k,h}$, $[\mathbb{V}_h(\hat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k) \leqslant E_{k,h}$ where $\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$, $U_{k,h}$, $E_{k,h}$ are specified in Lemma C.17 in Appendix.

Proof. Please refer to Appendix C.4. $\qquad\square$

8. Conclusion

This paper presents a computationally and statistically efficient algorithm, LSVI-UCB$^+$, which builds on linear weighted ridge regression and upper confidence value iteration with a Bernstein-type exploration bonus. LSVI-UCB$^+$ is the first algorithm to reach minimax optimal regret bound up to logarithmic factors for linear MDPs. Our sharp result builds on a novel Bernstein self-normalized bound with the conservatism of elliptical potentials, and refined analysis of the correction term, which serve as new analytical tools for RL with linear function approximation.

Acknowledgements

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

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

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory-colt*, 2008.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient pac rl with rich observations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1429–1439, 2018.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Hager, W. W. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989.

He, J., Zhou, D., and Gu, Q. Minimax optimal reinforcement learning for discounted mdps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2021.

He, J., Zhao, H., Zhou, D., and Gu, Q. Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132*, 2022.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 4868–4878, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Kirschner, J. and Krause, A. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pp. 358–384. PMLR, 2018.

Lattimore, T. and Hutter, M. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pp. 320–334. Springer, 2012.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lattimore, T., Crammer, K., and Szepesvári, C. Linear multi-resource allocation with semi-bandit feedback. In *NIPS*, pp. 964–972, 2015.

Li, Y., Wang, Y., Chen, X., and Zhou, Y. Tight regret bounds for infinite-armed linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 370–378. PMLR, 2021.

Modi, A., Jiang, N., Tewari, A., and Singh, S. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Tossou, A., Basu, D., and Dimitrakakis, C. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *arXiv preprint arXiv:1905.12425*, 2019.

Victor, H., la Peña, D., Lai, T. L., and Shao, Q.-M. *Self-normalized processes: Limit theory and Statistical Applications*, volume 204. Springer, 2009.

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

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020b.

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

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

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

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964. PMLR, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020b.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021.

Appendices

In the appendix, we present some additional results and supporting materials to supplement the statements, theorems and proofs in the main papers. There are 6 sections in appendix:

- Appendix A presents additional comparisons with related works.

- Appendix B presents the proof of our proposed sharp Bernstein tail inequality for self-normalized vector-valued martingales.

- Appendix C presents the construction of several high probability confidences sets.

- Appendix D presents the former proof of the regret upper bound of LSVI-UCB$^+$ algorithm, with conclusion given in Theorem 6.1.

- Appendix E constructs a hard-to-learn MDP to build a regret lower bound for linear MDPs.

- Appendix F presents auxiliary lemmas necessary for proofs in above sections and important properties that will be helpful in algorithm design.

A. Additional Comparisons of Related Works

Table 2 serves as a more complete table compared to Table 1 in the main paper, which lists some representative works in RL with linear function approximation. The top part of Table 2 lists representative works for the linear MDP and its generalizations and the buttom part is for the linear mixture MDP and its generalizations.

Table 2. Theoretical results on RL with linear function approximation, where † denotes that rewards are adversarial, and d_h is the dimension of the feature mapping at the h-th stage within the episodes and K is the number of episodes.

Setting	Algorithm	Technique	Regret
Linear MDP	OPT-RLSVI (Zanette et al., 2020a)	Hoeffding+Covering	$\widetilde{O}(H^2 d^2 \sqrt{T})$
Linear MDP	LSVI-UCB (Jin et al., 2020)	Hoeffding+Covering	$\widetilde{O}(\sqrt{H^3 d^3 T})$
Linear MDP	LSVI-UCB$^+$ (**this paper**)	Bernstein+*Covering*	$\widetilde{O}(Hd\sqrt{T})$
Linear Q Function	LSVI-UCB* (Wang et al., 2020b)	Hoeffding+Covering	$\widetilde{O}(\sqrt{d^3 T})$
Low Bellman Error	ELEANOR (Zanette et al., 2020b)	Hoeffding+Covering	$\widetilde{O}(\sum_{h=1}^{H} d_h \sqrt{K})$
Bounded Eluder Dimension	$\mathcal{F}$-LSVI(δ) (Wang et al., 2020a)	Hoeffding+Covering	$\widetilde{O}(\mathrm{poly}(dH)\sqrt{T})$
Linear Mixture MDP	UCRL-VTR (Jia et al., 2020; Ayoub et al., 2020)	Hoeffding	$\widetilde{O}(\sqrt{H^3 d^2 T})$
Linear Mixture MDP	UCRL-VTR$^+$ (Zhou et al., 2021)	Bernstein	$\widetilde{O}(\sqrt{H^2 d^2 T + H^3 dT})$
Feature Space	MatrixRL (Yang & Wang, 2020)	Hoeffding	$\widetilde{O}(H^2 d \log T \sqrt{T})$
Linear Mixture MDP†	OPPO (Cai et al., 2020)	Hoeffding	$\widetilde{O}(\sqrt{d^2 H^3 T})$

From Table 2, we can find that existing algorithms for the linear MDP and its generalizations all use a classical Hoeffding self-normalized bound such as Theorem 1 in (Abbasi-Yadkori et al., 2011) with the covering net argument, while our work is the first to introduce a Bernstein self-normalized bound with a covering net argument. Moreover, building a covering net argument in our work does not brings extra dependency on feature space dimension d since we only consider covering net argument in bounding the correction term which can be made small.

As for the linear mixture MDP and its generalizations, the covering net argument is not required due to the structure of the linear mixture MDP. In addition, prior works (Yang & Wang, 2020; Jia et al., 2020; Ayoub et al., 2020; Cai et al., 2020) utilize Hoeffding self-normalized bound to build confidence sets, while (Zhou et al., 2021) considers the Bernstein self-normalized bound for the first time in the setting of linear mixture MDP. Compared with regret bound of $\widetilde{O}(\sqrt{H^3 d^2 T})$ obtained in (Jia et al., 2020; Ayoub et al., 2020) for linear mixture MDPs, the regret bound $\widetilde{O}(\sqrt{H^2 d^2 T + H^3 dT})$ in (Zhou et al., 2021) is better and a $\sqrt{H}$ factor is further saved if $d \geqslant H$.

B. Sharp Bernstein Self-Normalized Bound

In this section, we prove the proposed sharp Bernstein tail inequality for self-normalized vector-valued martingales. Our proof diagram is based on the proof of Theorem 1 in (Zhou et al., 2021), which is firstly proposed in the proof of Lemma 14 in (Dani et al., 2008). However, our Bernstein self-normalized bound is sharper than Theorem 1 in (Zhou et al., 2021) with critical changes of the attenuation of the martingale difference sequence.

Specifically, $\{\min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\}\}_{t \in [T]}$ can be roughly considered as an attenuated sequence since we can prove $\sum_{t=1}^{T} \min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\} = O(d \log T)$ by Elliptical Potential Lemma. On the contrary, $\min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\}$ is deflated to 1 in Theorem 1, (Zhou et al., 2021) such that the bound is looser than ours. In the following proof, we do not deflate $\min\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\}$ to 1, and we take into account the elliptical potential $\|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}$ in our algorithm design, which is one of the major contributions in this paper.

Firstly, we give the following definitions to simplifying notations during the proof.

Definition B.1.

$$\boldsymbol{d}_t := \sum_{i=1}^{t} \boldsymbol{x}_i \eta_i,$$

$$Z_t := \|\boldsymbol{d}_t\|_{\mathbf{Z}_t^{-1}},$$

$$w_t := \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}},$$
$$\beta_t := 8\sigma\sqrt{d \log\left(1 + tL^2/(d\lambda)\right) \log\left(4t^2/\delta\right)} + 4R \log\left(4t^2/\delta\right),$$
$$\mathcal{E}_t := \mathbb{1}\left\{0 \leqslant s \leqslant t, Z_s \leqslant \beta_s\right\},$$

for $t \geqslant 1$ and $\boldsymbol{d}_0 = 0$, $Z_0 = 0$, $\beta_0 = 0$ for $t = 0$.

Measurability With the assumptions in Theorem 7.1, x_t is $\mathcal{G}_t$-measurable and η_t is $\mathcal{G}_{t+1}$-measurable. Thus, w_t is $\mathcal{G}_t$-measurable, and d_t, Z_t and $\mathcal{E}_t$ are $\mathcal{G}_{t+1}$-measurable.

Our goal is to upper bound Z_t. By definition of Z_t, we have

$$
\begin{aligned}
Z_t^2 &= (\boldsymbol{d}_{t-1} + \boldsymbol{x}_t\eta_t)^\top \mathbf{Z}_t^{-1} (\boldsymbol{d}_{t-1} + \boldsymbol{x}_t\eta_t) \\
&= \boldsymbol{d}_{t-1}^\top \mathbf{Z}_t^{-1} \boldsymbol{d}_{t-1} + 2\eta_t \boldsymbol{x}_t^\top \mathbf{Z}_t^{-1} \boldsymbol{d}_{t-1} + \eta_t^2 \boldsymbol{x}_t^\top \mathbf{Z}_t^{-1} \boldsymbol{x}_t \\
&\leqslant Z_{t-1}^2 + \underbrace{2\eta_t \boldsymbol{x}_t^\top \mathbf{Z}_t^{-1} \boldsymbol{d}_{t-1}}_{I_1} + \underbrace{\eta_t^2 \boldsymbol{x}_t^\top \mathbf{Z}_t^{-1} \boldsymbol{x}_t}_{I_2}
\end{aligned}
$$

where the inequality holds since $\mathbf{Z}_t \succeq \mathbf{Z}_{t-1}$.

Since $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \boldsymbol{x}_t \boldsymbol{x}_t^\top$, by the Sherman–Morrison formula (Hager, 1989), we obtain

$$\mathbf{Z}_t^{-1} = \mathbf{Z}_{t-1}^{-1} - \frac{\mathbf{Z}_{t-1}^{-1} \boldsymbol{x}_t \boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1}}{1 + w_t^2}, \quad \text{for } t \geqslant 1.$$

Subsequently,

$$I_1 = 2\eta_t \left(\boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{d}_{t-1} - \frac{\boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{x}_t \boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{d}_{t-1}}{1 + w_t^2} \right) = 2\eta_t \left(\boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{d}_{t-1} - \frac{w_t^2 \boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{d}_{t-1}}{1 + w_t^2} \right) = \frac{2\eta_t \boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{d}_{t-1}}{1 + w_t^2}$$

$$I_2 = \eta_t^2 \left(\boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{x}_t^\top - \frac{\boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{x}_t \boldsymbol{x}_t^\top \mathbf{Z}_{t-1}^{-1} \boldsymbol{x}_t}{1 + w_t^2} \right) = \eta_t^2 \left(w_t^2 - \frac{w_t^4}{1 + w_t^2} \right) = \frac{\eta_t^2 w_t^2}{1 + w_t^2}$$

Therefore, we have

$$Z_t^2 \leqslant \sum_{i=1}^t \frac{2\eta_i \boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} + \sum_{i=1}^t \frac{\eta_i^2 w_i^2}{1 + w_i^2} \tag{9}$$

Now we try to bound the two summation terms on the r.h.s. of Eq. (9) in Lemma B.3 and Lemma B.4, respectively. Before that, we present a uniform Bernstein bound required for proving Lemma B.3 and Lemma B.4.

Lemma B.2 (Uniform Bernstein Bound). *Let* $\{x_i, \mathcal{F}_i\}$ *be a martingale difference sequence with* $\forall i \geqslant 1$, $\mathbb{E}(x_i \mid \mathcal{F}_{i-1}) = 0$, $\mathbb{E}\left(x_i^2 \mid \mathcal{F}_{i-1}\right) = \sigma_i^2$, $V_i^2 = \sum_{j=1}^i \sigma_j^2$. *Furthermore, assume that* $\mathbb{P}\left(|x_i| \leqslant c \mid \mathcal{F}_{i-1}\right) = 1$ *for any* $0 < c < \infty$.

Then, for any $\delta > 0$, *with probability at least* $1 - \delta$, *simultaneously for any* $t \geqslant 1$, *it holds that*

$$\sum_{i=1}^t d_i \leqslant \sqrt{2V_t^2 \log(2t^2/\delta)} + \frac{2c \log(2t^2/\delta)}{3}.$$

Proof. By Freedman's inequality in Lemma F.2, for any $\delta > 0$ and some $t \geqslant 1$, with probability at least $1 - \delta/(2t^2)$, we have:

$$\sum_{i=1}^t d_i \leqslant \sqrt{2V_t^2 \log(1/\delta)} + \frac{2}{3}c \log(1/\delta). \tag{10}$$

Taking a union bound for Eq. (10) from $t = 1$ to ∞ and using the fact that $\sum_{t=1}^\infty t^{-2} < 2$ complete the proof. $\qquad\square$

Next, we bound the first summation term on the r.h.s. of Eq. (9) in Lemma B.3.

Lemma B.3. *Under assumptions in Theorem 7.1 and Definition B.1, with probability at least $1 - \delta/2$, simultaneously for all $t \geqslant 1$ it holds that*

$$\sum_{i=1}^{t} \frac{2\eta_i \boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \leqslant 3\beta_t^2/4$$

Proof. Define

$$\ell_i = \frac{2\eta_i \boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1},$$

and we will give a uniform upper bound of $\sum_{i=1}^{t} l_i$ by Lemma B.2 below.

Firstly, for any $1 \leqslant i \leqslant t$,

$$\mathbb{E}[\ell_i \mid \mathcal{G}_i] = \frac{2\boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \mathbb{E}[\eta_i \mid \mathcal{G}_i] = 0.$$

Besides, we have

$$\left| \frac{2\boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \right| \leqslant \frac{2 \|\boldsymbol{x}_i\|_{\mathbf{Z}_{i-1}^{-1}} \left[\|\boldsymbol{d}_{i-1}\|_{\mathbf{Z}_{i-1}^{-1}} \mathcal{E}_{i-1} \right]}{1 + w_i^2} \leqslant \frac{2w_i \beta_{i-1}}{1 + w_i^2} \leqslant 2\min\{1, w_i\} \beta_{i-1} \tag{11}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to the definition of $\mathcal{E}_{i-1}$, and the last inequality holds by algebra. Thus,

$$|\ell_i| = |2\eta_i \cdot \min\{1, w_i\} \beta_i| \leqslant 2R\beta_i \leqslant 2R\beta_t$$

where the last inequality holds since $\{\beta_i\}_{i \in [t]}$ is an increasing sequence.

Secondly, we also have

$$\sum_{i=1}^{t} \mathbb{E}\left[\ell_i^2 \mid \mathcal{G}_i\right] \leqslant \sigma^2 \sum_{i=1}^{t} \left(\frac{2\boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \right)^2$$

$$\leqslant \sigma^2 \sum_{i=1}^{t} \left[2\min\{1, w_i\} \beta_{i-1} \right]^2$$

$$\leqslant 4\sigma^2 \beta_t^2 \sum_{i=1}^{t} \min\{1, w_i^2\}$$

$$\leqslant 8\sigma^2 \beta_t^2 d \log\left(1 + tL^2/(d\lambda)\right)$$

where the first inequality holds since $\mathbb{E}\left[\eta_i^2 \mid \mathcal{G}_i\right] \leqslant \sigma^2$, the second inequality holds due to Eq. (11), the third inequality holds since $\{\beta_i\}_{i \in [t]}$ is an increasing sequence, and the last inequality holds due to Lemma F.5.

Therefore, using Lemma B.2, simultaneously for all $t \geqslant 1$, with probability at least $1 - \delta/\left(4t^2\right)$, it holds that

$$\sum_{i=1}^{t} \ell_i \leqslant \sqrt{16\sigma^2 \beta_t^2 d \log\left(1 + tL^2/(d\lambda)\right) \log\left(4t^2/\delta\right)} + 2/3 \cdot 2R\beta_t \log\left(4t^2/\delta\right)$$

$$\leqslant \frac{\beta_t^2}{4} + 16\sigma^2 d \log\left(1 + tL^2/(d\lambda)\right) \log\left(4t^2/\delta\right) + \frac{\beta_t^2}{4} + 4R^2 \log^2\left(4t^2/\delta\right)$$

$$\leqslant \beta_t^2/2 + \frac{1}{4} \left(8\sigma\sqrt{d \log\left(1 + tL^2/(d\lambda)\right) \log\left(4t^2/\delta\right)} + 4R\log\left(4t^2/\delta\right) \right)^2$$

$$= 3\beta_t^2/4$$

where the first inequality holds due to Lemma B.2, the second inequality holds due to $2\sqrt{|ab|} \leqslant |a| + |b|$, and the last equality follows from the definition of β_t. □

Accordingly, the second summation terms on the r.h.s. of Eq. (9) is bounded in Lemma B.4.

Lemma B.4. *Under assumptions in Theorem 7.1 and Definition B.1, with probability at least $1 - \delta/2$, simultaneously for all $t \geqslant 1$ it holds that*

$$\sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1 + w_i^2} \leqslant \beta_t^2/4$$

Proof. Define

$$\ell_i = \frac{\eta_i^2 w_i^2}{1 + w_i^2} - \mathbb{E}\left[\frac{\eta_i^2 w_i^2}{1 + w_i^2} \mid \mathcal{G}_i\right].$$

We will give a uniform upper bound of $\sum_{i=1}^{t} l_i$ by Lemma B.2, similar to Lemma B.3.

Clearly, for any $1 \leqslant i \leqslant t$, we have $\mathbb{E}\left[\ell_i \mid \mathcal{G}_i\right] = 0$. We further have that

$$\sum_{i=1}^{t} \mathbb{E}\left[\ell_i^2 \mid \mathcal{G}_i\right] \leqslant \sum_{i=1}^{t} \mathbb{E}\left[\frac{\eta_i^4 w_i^4}{\left(1 + w_i^2\right)^2} \mid \mathcal{G}_i\right]$$

$$\leqslant R^2 \sum_{i=1}^{t} \mathbb{E}\left[\frac{\eta_i^2 w_i^2}{1 + w_i^2} \mid \mathcal{G}_i\right]$$

$$\leqslant R^2 \sigma^2 \sum_{i=1}^{t} \frac{w_i^2}{1 + w_i^2}$$

$$\leqslant 2R^2 \sigma^2 d \log\left(1 + tL^2/(d\lambda)\right)$$

where the first inequality holds due to the fact $\mathbb{E}(X - \mathbb{E}X)^2 \leqslant \mathbb{E}X^2$, the second inequality holds since $|\eta_i \cdot \min\{1, w_i\}| \leqslant R$, the third inequality holds since $\mathbb{E}\left[\eta_i^2 \mid \mathcal{G}_i\right] \leqslant \sigma^2$, and the fourth inequality holds due to the fact $w_i^2/\left(1 + w_i^2\right) \leqslant \min\{1, w_i^2\}$ and Lemma F.5.

Furthermore, using the fact that $|\eta_i \cdot \min\{1, w_i\}| \leqslant R$ holds almost surely under filtration $\mathcal{G}_i$, we obtain

$$\left|\frac{\eta_i^2 w_i^2}{1 + w_i^2}\right| \leqslant \left|\eta_i^2 \cdot \min\{1, w_i^2\}\right| = |\eta_i \cdot \min\{1, w_i\}|^2 = R^2,$$

and

$$|\ell_i| \leqslant \left|\frac{\eta_i^2 w_i^2}{1 + w_i^2}\right| + \left|\mathbb{E}\left[\frac{\eta_i^2 w_i^2}{1 + w_i^2} \mid \mathcal{G}_i\right]\right| \leqslant 2R^2.$$

Therefore, by Lemma B.2, simultaneously for all $t \geqslant 1$, with probability at least $1 - \delta/\left(4t^2\right)$, it holds that

$$\sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1 + w_i^2} \leqslant \sum_{i=1}^{t} \mathbb{E}\left[\frac{\eta_i^2 w_i^2}{1 + w_i^2} \mid \mathcal{G}_i\right] + \sqrt{4R^2\sigma^2 d \log\left(1 + tL^2/(d\lambda)\right)\log\left(4t^2/\delta\right)} + 4/3 \cdot R^2 \log\left(4t^2/\delta\right)$$

$$\leqslant \sigma^2 \sum_{i=1}^{t} \frac{w_i^2}{1 + w_i^2} + 2R\sigma\sqrt{d \log\left(1 + tL^2/(d\lambda)\right)\log\left(4t^2/\delta\right)} + 2R^2 \log\left(4t^2/\delta\right)$$

$$\leqslant 2\sigma^2 d \log\left(1 + tL^2/(d\lambda)\right) + 2R\sigma\sqrt{d \log\left(1 + tL^2/(d\lambda)\right)\log\left(4t^2/\delta\right)} + 2R^2 \log\left(4t^2/\delta\right)$$

$$\leqslant 1/4 \cdot \left(8\sigma\sqrt{d}\sqrt{\log\left(1 + tL^2/(d\lambda)\right)\log\left(4t^2/\delta\right)} + 4R\log\left(4t^2/\delta\right)\right)^2$$

$$= \beta_t^2/4$$

where the first inequality holds due to Lemma B.2, the second inequality holds due to $\mathbb{E}\left[\eta_i^2 \mid \mathcal{G}_i\right] \leqslant \sigma^2$, the third inequality holds due to the fact $w_i^2/\left(1 + w_i^2\right) \leqslant \min\{1, w_i^2\}$ and Lemma 12, and the last inequality holds due to the definition of β_t. $\square$

B.1. Proof of Theorem 7.1

Proof of Theorem 7.1. Consider the case when conclusions of Lemma B.3 and Lemma B.4 hold. Conditioning on this event, we claim $Z_t \leqslant \beta_t$ for any $t \geqslant 0$.

We prove this by induction on t. Initially, the base case of $t = 0$ holds since $\beta_0 = 0 = Z_0$ by definition. Now fix some $t \geqslant 1$ and assume that for all $0 \leqslant i \leqslant t-1$, we have $Z_i \leqslant \beta_i$. This implies that $\mathcal{E}_1 = \mathcal{E}_2 = \cdots = \mathcal{E}_{t-1} = 1$. Then, by Eq. (9), we have

$$Z_t^2 \leqslant \sum_{i=1}^{t} \frac{2\eta_i \boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} + \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1 + w_i^2} = \sum_{i=1}^{t} \frac{2\eta_i \boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} + \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1 + w_i^2}. \tag{12}$$

Since the conclusions of Lemma B.3 and Lemma B.4 hold, we have

$$\sum_{i=1}^{t} \frac{2\eta_i \boldsymbol{x}_i^\top \mathbf{Z}_{i-1}^{-1} \boldsymbol{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \leqslant 3\beta_t^2/4, \tag{13}$$

$$\sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1 + w_i^2} \leqslant \beta_t^2/4. \tag{14}$$

Therefore, substituting Eq. (13) and (14) into Eq. (12), we have $Z_t \leqslant \beta_t$, which ends the induction. Taking the union bound of the events in Lemma B.3 and Lemma B.4 implies that with probability at least $1 - \delta$, $Z_t \leqslant \beta_t$ holds for any $t \geqslant 1$. $\square$

C. High Probability Events

In this section, we define some high probability events, i.e., confidence sets concerning the parameter $\boldsymbol{\mu}_h$, and show how to build them. The goal of this section is to build the sharp optimistic confidence set $\widehat{\mathcal{C}}_{k,h}$ in Lemma 7.5 for all $(k, h) \in [K] \times [H]$.

We lists all confidence sets encountered during the proof in the following. Confidence sets $\bar{\mathcal{C}}_{k,h}, \widetilde{\mathcal{C}}_{k,h}, \check{\mathcal{C}}_{k,h}$ are called independent confidence sets, since they can be built by applying self-normalized concentration inequality directly without conditioning on other events. Instead, confidence sets $\widehat{\mathcal{C}}_{k,h}^{(1)}, \widehat{\mathcal{C}}_{k,h}^{(2)}, \widehat{\mathcal{C}}_{k,h}$ are called dependent confidence sets since they can only be built by conditioning on other confidence sets, apart from self-normalized concentration inequalities.

Definition C.1 (Confidence Set). • Independent Confidence Sets:

$$\bar{\mathcal{C}}_{k,h} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \, \widehat{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \bar{\beta} \right\}$$

$$\widetilde{\mathcal{C}}_{k,h} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \, \widehat{\boldsymbol{V}}_{k,h+1}^2 \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widetilde{\beta} \right\}$$

$$\check{\mathcal{C}}_{k,h} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \, \check{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \check{\beta} \right\}$$

• Dependent Confidence Sets:

$$\widehat{\mathcal{C}}_{k,h}^{(1)} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \, \boldsymbol{V}_{h+1}^* \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widehat{\beta}^{(1)} \right\}$$

$$\widehat{\mathcal{C}}_{k,h}^{(2)} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \left(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}_{h+1}^* \right) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widehat{\beta}^{(2)} \right\}$$

$$\widehat{\mathcal{C}}_{k,h} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \, \widehat{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widehat{\beta}^{(1)} + \widehat{\beta}^{(2)} = \widehat{\beta} \right\}$$

To simplify notations during the proof, we further define the following events that optimistic and pessimistic confidence sets hold in multiple stages under some episode $k \in [K]$ or all episodes.

Definition C.2 (Optimism Event).

$$\widehat{\Psi}_{k,h} := \left\{ \forall h \leqslant h' \leqslant H : \boldsymbol{\mu}_{h'} \in \widehat{\mathcal{C}}_{k,h'} \right\}$$

$$\widehat{\Psi}_h := \left\{ \forall i \in [K], \forall h \leqslant h' \leqslant H : \boldsymbol{\mu}_{h'} \in \widehat{\mathcal{C}}_{i,h'} \right\}$$

(15)

Definition C.3 (Pessimism Event).

$$\breve{\Psi}_{k,h} := \left\{ \forall h \leqslant h' \leqslant H : \boldsymbol{\mu}_{h'} \in \breve{\mathcal{C}}_{k,h'} \right\}$$

$$\breve{\Psi}_h := \left\{ \forall i \in [K], \forall h \leqslant h' \leqslant H : \boldsymbol{\mu}_{h'} \in \breve{\mathcal{C}}_{i,h'} \right\}$$

(16)

In this section, independent confidence sets $\bar{\mathcal{C}}_{k,h}, \widetilde{\mathcal{C}}_{k,h}, \breve{\mathcal{C}}_{k,h}$ are built in Lemma C.8, C.9, and C.10 respectively in Appendix C.1. These independent confidence sets are built to upper bound the variance of the considered value function. Specifically, the difference between the estimated variance of the constructed optimistic value function and the real variance of the optimal value function, i.e., $\left| [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k) - [\widehat{\mathbb{V}}_{k,h} \widehat{V}_{k,h+1}](s_h^k, a_h^k) \right|$, is upper bounded in high probability in Lemma C.12. In addition, the variance $[\mathbb{V}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)$ is also upper bounded in Lemma C.13. Subsequently, dependent confidence sets $\widehat{\mathcal{C}}_{k,h}^{(1)}, \widehat{\mathcal{C}}_{k,h}^{(2)}$ can be built based on the independent confidence sets in Lemma C.15, C.16 respectively in Appendix C.3. Thus, the confidence set $\widehat{\mathcal{C}}_{k,h}$, the goal of this section, holds trivially if $\widehat{\mathcal{C}}_{k,h}^{(1)}, \widehat{\mathcal{C}}_{k,h}^{(2)}$ both hold. Finally, Lemma 7.5 in the main paper is proved in Appendix C.4.

Before the formal proof begins, we give some necessary definitions. We first give definitions about-measurable space and filtration required for our proofs.

Measurable Space   Note that the stochasticity in the transition probability of the MDP are the only source of randomness. Denote $\mathbb{P}$ as the gather of the distributions over state-action pair sequence $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$, induced by the interconnection of policy obtained from LSVI-UCB$^+$ algorithm and the episodic linear MDP $\mathcal{M}$. Denote $\mathbb{E}$ as the corresponding expectation operator. Hence, all random variables can be defined over the sample space $\Omega = (\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$. Thus, we work with the probability space given by the triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is the product σ-algebra generated by the discrete σ-algebras underlying $\mathcal{S}$ and $\mathcal{A}$.

Definition C.4 (Filtration).   For any $k \in [K]$ and any $h \in [H]$, let $\mathcal{F}_{k,h}$ be the σ-algebra generated by the random variables representing the state-action pairs up to and including that appears in stage h of episode k.

Measurability   Thus, $[\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k), U_{k,h}, E_{k,h}, \varsigma_{k,h}, \widetilde{\sigma}_{k,h}, \widehat{\sigma}_{k,h}, \widehat{\boldsymbol{\Lambda}}_{k+1,h}$ are $\mathcal{F}_{k,h}$-measurable, $\widehat{\boldsymbol{\mu}}_{k+1,h}$ is $\mathcal{F}_{k,h+1}$-measurable, $\widehat{Q}_{k,h}, \breve{Q}_{k,h}, \widehat{V}_{k,h}, \breve{V}_{k,h}, \pi_h^k$ are $\mathcal{F}_{k-1,H}$-measurable, but not $\mathcal{F}_{k-1,h}$-measurable due to their backwards construction.

C.1. Independent Confidence Sets

In this subsection, independent confidence sets $\bar{\mathcal{C}}_{k,h}, \widetilde{\mathcal{C}}_{k,h}, \breve{\mathcal{C}}_{k,h}$ are built in Lemma C.8, C.9, C.10, respectively. During the construction of these confidence sets, it is unavoidable to build a uniform convergence argument by covering net of the encountered function class. Thus, we also present the definition of possibly encountered function classes in the following.

Definition C.5 (Optimistic Value Function Class).   For fixed J updating episode, let $\widehat{\mathcal{V}}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form

$$\widehat{V}(\cdot) = \max_a \min_{1 \leqslant i \leqslant J} \min \left\{ \boldsymbol{w}_i^\top \boldsymbol{\phi}(\cdot, a) + \beta \sqrt{\boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{\Lambda}_i^{-1} \boldsymbol{\phi}(\cdot, a)}, H \right\},$$

where the parameters $(\boldsymbol{w}_i, \beta, \boldsymbol{\Lambda}_i)$ satisfy $\|\boldsymbol{w}_i\|_2 \leqslant L, \beta \in [0, B]$, the minimum eigenvalue satisfies $\lambda_{\min}(\boldsymbol{\Lambda}_i) \geqslant \lambda$, and $\sup_{s,a} \|\boldsymbol{\phi}(s, a)\|_2 \leqslant 1$.

Definition C.6 (Squared Optimistic Value Function Class).   For fixed J updating episode, let $\widehat{\mathcal{V}}^2$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form

$$\widehat{V}^2(\cdot) = \max_a \min_{1 \leqslant i \leqslant J} \left[\min \left\{ \boldsymbol{w}_i^\top \boldsymbol{\phi}(\cdot, a) + \beta \sqrt{\boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{\Lambda}_i^{-1} \boldsymbol{\phi}(\cdot, a)}, H \right\} \right]^2,$$

where the parameters $(\boldsymbol{w}_i, \beta, \boldsymbol{\Lambda}_i)$ satisfy $\|\boldsymbol{w}_i\|_2 \leqslant L, \beta \in [0, B]$, the minimum eigenvalue satisfies $\lambda_{\min}(\boldsymbol{\Lambda}_i) \geqslant \lambda$, and $\sup_{s,a} \|\boldsymbol{\phi}(s, a)\|_2 \leqslant 1$.

Definition C.7 (Pessimistic Value Function Class). Let $\check{\mathcal{V}}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form

$$\check{V}(\cdot) = \max\left\{\max_a \boldsymbol{w}^\top \boldsymbol{\phi}(\cdot, a) - \beta\sqrt{\boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{\phi}(\cdot, a)}, 0\right\},$$

where the parameters $(\boldsymbol{w}, \beta, \boldsymbol{\Lambda})$ satisfy $\|\boldsymbol{w}\|_2 \leqslant L, \beta \in [0, B]$, the minimum eigenvalue satisfies $\lambda_{\min}(\boldsymbol{\Lambda}) \geqslant \lambda$, and $\sup_{s,a} \|\boldsymbol{\phi}(s, a)\|_2 \leqslant 1$.

Now we are ready to build four independent confidence sets $\bar{\mathcal{C}}_{k,h}, \widetilde{\mathcal{C}}_{k,h}, \check{\mathcal{C}}_{k,h}$. Since radius of independent confidence sets, i.e., $\bar{\beta}, \widetilde{\beta}, \check{\beta}$, will not become dominant terms in the final regret bound, we build these four confidence sets with traditional Hoeffding inequality (Lemma F.3) with covering net arguments.

Lemma C.8. *In Algorithm 1, for any $\delta \in (0, 1)$, any $k \in [K]$ fixed $h \in [H]$, with probability at least $1 - \delta/H$:*

$$\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{k,h} = \left\{\boldsymbol{\mu} : \left\|(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \widehat{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \bar{\beta}\right\},$$

where

$$\bar{\beta} = \sqrt{H}\sqrt{d\log\left(1 + \frac{K}{Hd\lambda}\right) + \log\left(\frac{H}{\delta}\right) + dJ\log\left(1 + \frac{4KL}{H\sqrt{\lambda}}\right) + d^2 J\log\left(1 + \frac{8K^2\widehat{B}^2\sqrt{d}}{H^2\lambda^2}\right)} + H\sqrt{\lambda d} + 2.$$

Here $J = dH\log(1 + K), L = W + K/\lambda$ and $\widehat{B}$ is a constant satisfying $\widehat{\beta} \leqslant \widehat{B}$ with $\widehat{\beta}$ given in Lemma C.17.

Proof. Initially, note that we have

$$\left\|(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}}$$

$$= \left\|\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}\left[-\lambda\boldsymbol{\mu}_h + \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^\top\right] \widehat{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}}$$

$$= \left\|-\lambda\boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1} + \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^\top \widehat{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \qquad (17)$$

$$\leqslant \left\|-\lambda\boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^\top \widehat{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$\leqslant \frac{1}{\sqrt{\lambda}} \cdot \lambda H\sqrt{d} + \left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^\top \widehat{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}},$$

where the first equality is due to Eq. (55) in Lemma F.9, the first inequality is due to triangle inequality, and the second inequality holds since $\|\boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1}\|_2 \leqslant H\sqrt{d}$ and the minimum eigenvalue of $\widehat{\boldsymbol{\Lambda}}_{k,h}$ is no less than λ.

Thus, we bound $\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^\top \widehat{\boldsymbol{V}}_{k,h+1}\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$ in the following. However, $\widehat{V}_{k,h+1}$ is $\mathcal{F}_{k,h}$-measurable, which brings obstacles in directly applying self-normalized bound for martingales. We need to build a uniform convergence argument for $\widehat{V}_{k,h+1}$.

For any $(k, h) \in \times[K] \times [H]$, $\widehat{V}_{k,h}(\cdot) = \min\{\max_a \langle \boldsymbol{\theta}_h + \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(\cdot, a)\rangle + \widehat{\beta}\|\boldsymbol{\phi}(\cdot, a)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}, H\}$ in Algorithm 1. Moreover, we have

$$\left\|\boldsymbol{\theta}_h + \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}\right\|_2 = \left\|\boldsymbol{\theta}_h + \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)\widehat{V}_{k,h+1}(s_{h+1}^i)\right\|_2$$

$$\leqslant W + \left\| \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \widehat{V}_{k,h+1}(s_{h+1}^i) \right\|_2$$

$$\leqslant W + \frac{H}{(\sqrt{H})^2} \left\| \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \boldsymbol{\phi}(s_h^i, a_h^i) \right\|_2 \leqslant W + K/\lambda,$$

where the first inequality holds due to triangle inequality, the second inequality holds since $\widehat{V}_{k,h+1}(\cdot) \leqslant H$ and $\widehat{\sigma}_{i,h} \geqslant \sqrt{H}$ for any $i \in [k]$, and the last inequality holds since $\lambda_{\min}(\boldsymbol{\Lambda}_{k,h}) \geqslant \lambda$ and $\sup_{s,a} \|\boldsymbol{\phi}(s,a)\|_2 \leqslant 1$. Subsequently, we claim $\widehat{V}_{k,h} \in \widehat{\mathcal{V}}$, where $\widehat{\mathcal{V}}$ is defined in Definition C.5, with $L = W + K/\lambda$ and $B = \widehat{B}$. Here, $\widehat{B}$ is a constant satisfying $\widehat{\beta} \leqslant \widehat{B}$ with $\widehat{\beta}$ specified in Lemma C.17.

Then, we fix a function $V(\cdot) \in \widehat{\mathcal{V}} : \mathcal{S} \mapsto [0, H]$. Let $\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{x}_i = \widehat{\sigma}_{i,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i)$ and $\eta_i = \widehat{\sigma}_{i,h}^{-1} {\boldsymbol{\epsilon}_h^i}^\top V = \widehat{\sigma}_{i,h}^{-1} \langle \boldsymbol{\mu}_h V, \boldsymbol{\phi}(s_h^i, a_h^i) \rangle - \widehat{\sigma}_{i,h}^{-1} V(s_{h+1}^i)$. It is clear that $\boldsymbol{x}_i$ is $\mathcal{G}_i$-measurable and η_i is $\mathcal{G}_{i+1}$-measurable. Since $\varsigma_i \geqslant \sqrt{H}$, we have $\widehat{\sigma}_{i,h} \geqslant \sqrt{H}$. Besides, we have $\|\boldsymbol{x}_i\|_2 \leqslant 1/\sqrt{H}$, $\mathbb{E}[\eta_i \mid \mathcal{G}_i] = 0$, $|\eta_i| \leqslant \sqrt{H}$ and $\mathbb{E}[\eta_i^2 \mid \mathcal{G}_i] \leqslant H$. By Lemma F.3, we obtain that, with probability at least $1 - \delta/H$, for any $k \in [K]$ and fixed $h \in [H]$,

$$\left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top V \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \leqslant \sqrt{H} \sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) + \log\left(\frac{H}{\delta}\right)}.$$

Denote the ε-cover of function class $\widehat{\mathcal{V}}$ as $\widehat{\mathcal{N}}_\varepsilon$. Consider an arbitrary $f \in \widehat{\mathcal{V}}$. From the definition of ε-cover, we know that for f, there exists a $V \in \widehat{\mathcal{N}}_\varepsilon$, such that $\|\boldsymbol{f} - \boldsymbol{V}\|_\infty \leqslant \varepsilon$. Since $\|{\boldsymbol{\epsilon}_h^i}^\top (\boldsymbol{f} - \boldsymbol{V})\|_2 \leqslant \|\boldsymbol{\epsilon}_h^i\|_1 \|\boldsymbol{f} - \boldsymbol{V}\|_\infty \leqslant 2\varepsilon$ and $\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \leqslant K/(H\sqrt{\lambda})$, we have

$$\left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top (\boldsymbol{f} - \boldsymbol{V}) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \leqslant \frac{2\varepsilon K}{H\sqrt{\lambda}}. \tag{18}$$

This further implies the following inequality holds with probability at least $1 - \delta/H$:

$$
\begin{aligned}
&\left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top \boldsymbol{f} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \\
&\leqslant \left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top \boldsymbol{V} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top (\boldsymbol{f} - \boldsymbol{V}) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \\
&\leqslant \left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top \boldsymbol{V} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \frac{2\varepsilon K}{H\sqrt{\lambda}} \\
&\leqslant \sqrt{H} \sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) + \log\left(\frac{H}{\delta}\right) + \log\left|\widehat{\mathcal{N}}_\varepsilon\right|} + \frac{2\varepsilon K}{H\sqrt{\lambda}}.
\end{aligned}
\tag{19}
$$

where the first inequality is due to the triangle inequality, the second one holds by Eq. (18), and the third inequality holds by a union bound over all functions in $\widehat{\mathcal{N}}_\varepsilon$ with

$$\log\left|\widehat{\mathcal{N}}_\varepsilon\right| \leqslant dJ \log\left(1 + \frac{4L}{\varepsilon}\right) + d^2 J \log\left(1 + \frac{8\widehat{B}^2 \sqrt{d}}{\lambda \varepsilon^2}\right)$$

according to Lemma F.11 with $J = dH \log(1 + K)$ by Lemma F.8.

Note that Eq. (17) holds and $\widehat{V}_{k,h+1}(\cdot) \in \widehat{\mathcal{V}}$. We have with probability at least $1 - \delta/H$, for any $k \in [K]$ and fixed $h \in [H]$:

$$\left\| \left(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h\right) \widehat{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top \widehat{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + H\sqrt{\lambda d}$$

$$\leqslant \sqrt{H} \sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) + \log\left(\frac{H}{\delta}\right) + dJ \log\left(1 + \frac{4KL}{H\sqrt{\lambda}}\right) + d^2 J \log\left(1 + \frac{8K^2 \widehat{B}^2 \sqrt{d}}{H^2 \lambda^2}\right)} + H\sqrt{\lambda d} + 2$$

$$= \bar{\beta},$$

where the last inequality holds by Eq. (19) and setting $\varepsilon = H\sqrt{\lambda}/K$. $\qquad\square$

After building the confidence set $\bar{\mathcal{C}}_{k,h}$ in Lemma C.8, confidence sets $\widetilde{\mathcal{C}}_{k,h}$ and $\check{\mathcal{C}}_{k,h}$ can be built similarly.

Lemma C.9. *In Algorithm 1, for any $\delta \in (0,1)$, any $k \in [K]$ and fixed $h \in [H]$, with probability at least $1 - \delta/H$:*

$$\boldsymbol{\mu}_h \in \widetilde{\mathcal{C}}_{k,h} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \, \widehat{\boldsymbol{V}}^2_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widetilde{\beta} \right\},$$

where

$$\widetilde{\beta} = \sqrt{H^3} \sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) + \log\left(\frac{H}{\delta}\right) + dJ \log\left(1 + \frac{8KL}{\sqrt{\lambda}}\right) + d^2 J \log\left(1 + \frac{32K^2 \widehat{B}^2 \sqrt{d}}{\lambda^2}\right)} + H^2 \sqrt{\lambda d} + 2.$$

Here $J = dH \log(1 + K)$, $L = W + K/\lambda$ and and $\widehat{B}$ is a constant satisfying $\widehat{\beta} \leqslant \widehat{B}$ with $\widehat{\beta}$ given in Lemma C.17.

Proof. The proof of this lemma is almost the same as that of Lemma C.8, except for replacing $\widehat{V}_{k,h+1}$, $\widehat{\mathcal{V}}$, and $\widehat{\mathcal{N}}_\varepsilon$ by $\widehat{V}^2_{k,h+1}$, $\widehat{\mathcal{V}}^2$, and $\widehat{\mathcal{N}}^2_\varepsilon$, respectively. Here

$$\log \left| \widehat{\mathcal{N}}^2_\varepsilon \right| \leqslant dJ \log\left(1 + \frac{8LH}{\varepsilon}\right) + d^2 J \log\left(1 + \frac{32 \widehat{B}^2 H^2 \sqrt{d}}{\lambda \varepsilon^2}\right),$$

where $L = W + k/\lambda$ and $\widehat{B}$ is a constant satisfying $\widehat{\beta} \leqslant \widehat{B}$ with $\widehat{\beta}$ given in Lemma C.17.

After setting $\varepsilon = H\sqrt{\lambda}/K$, it can be proved that with probability at least $1 - \delta/H$, for any $k \in [K]$ and fixed $h \in [H]$:

$$\left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \, \widehat{\boldsymbol{V}}^2_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}}$$

$$\leqslant \sqrt{H^3} \sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) + \log\left(\frac{H}{\delta}\right) + dJ \log\left(1 + \frac{8KL}{\sqrt{\lambda}}\right) + d^2 J \log\left(1 + \frac{32K^2 \widehat{B}^2 \sqrt{d}}{\lambda^2}\right)} + H^2 \sqrt{\lambda d} + 2$$

$$= \widetilde{\beta}.$$

$\qquad\square$

Lemma C.10. *In Algorithm 1, for any $\delta \in (0,1)$, any $k \in [K]$ and fixed $h \in [H]$, with probability at least $1 - \delta/H$:*

$$\boldsymbol{\mu}_h \in \check{\mathcal{C}}_{k,h} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \, \check{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \check{\beta} \right\},$$

where

$$\check{\beta} = \sqrt{H} \sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) + \log\left(\frac{H}{\delta}\right) + d \log\left(1 + \frac{4KL}{H\sqrt{\lambda}}\right) + d^2 \log\left(1 + \frac{8K^2 \check{B}^2 \sqrt{d}}{H^2 \lambda^2}\right)} + H\sqrt{\lambda d} + 2. \quad (20)$$

Here $L = W + K/\lambda$ and $\check{B}$ is a constant satisfying $\check{\beta} \leqslant \check{B}$.

Proof. The proof of this lemma is similar to that of Lemma C.8 except for replacing $\widehat{V}_{k,h+1}$, $\widehat{\mathcal{V}}$, and $\widehat{\mathcal{N}}_\varepsilon$ by $\widecheck{V}_{k,h+1}$, $\widecheck{\mathcal{V}}$, and $\widecheck{\mathcal{N}}_\varepsilon$, respectively, where

$$\log\left|\widecheck{\mathcal{N}}_\varepsilon\right| \leqslant d\log\left(1 + \frac{4L}{\varepsilon}\right) + d^2\log\left(1 + \frac{8\widecheck{B}^2\sqrt{d}}{\lambda\varepsilon^2}\right).$$

Here $L = W + K/\lambda$ and $\widecheck{B}$ is a constant satisfying $\widecheck{\beta} \leqslant \widecheck{B}$ with $\widecheck{\beta}$ specified in Eq. (20).

After setting $\varepsilon = H\sqrt{\lambda}/K$, it can be proved that with probability at least $1 - \delta/H$, for any $k \in [K]$ and fixed $h \in [H]$:

$$\left\|\left(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h\right)\widecheck{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}}$$

$$\leqslant \sqrt{H}\sqrt{d\log\left(1 + \frac{K}{Hd\lambda}\right) + \log\left(\frac{H}{\delta}\right) + d\log\left(1 + \frac{4KL}{H\sqrt{\lambda}}\right) + d^2\log\left(1 + \frac{8K^2\widecheck{B}^2\sqrt{d}}{H^2\lambda^2}\right)} + H\sqrt{\lambda d} + 2$$

$$= \widecheck{\beta}.$$

$\square$

C.2. Variance Upper Bound

In this section, we prove some necessary lemmas to build upper bounds of value function variances, including variances of $V_{h+1}^*(\cdot)$ and $[\widehat{V}_{k,h+1} - V_{h+1}^*](\cdot)$. Specifically, we present Lemma C.12 in Appendix C.2.1 to bound the difference between the estimated variance of the constructed optimistic value function and the real variance of the optimal value function, i.e., $\left|[\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k) - [\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k)\right|$, with high probability. In addition, we also present Lemma C.13 in Appendix C.2.2 to upper bound the variance $[\mathbb{V}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)$.

C.2.1. VARIANCE OF $V_{h+1}^*(\cdot)$

Before proving Lemma C.12, we first present Lemma C.11, which upper bounds $\left|[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)\right|$ under optimism and pessimism events $\widehat{\Psi}_{k,h+1} \cap \widecheck{\Psi}_{k,h+1}$, and serves as the building block for Lemma C.12.

Lemma C.11. *In Algorithm 1, for any $k \in [K]$ and any $h \in [H]$, under $\widehat{\Psi}_{k,h+1} \cap \widecheck{\Psi}_{k,h+1}$, we have*

$$\left|\langle\boldsymbol{\mu}_h\widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle - \langle\boldsymbol{\mu}_h\boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right| \leqslant \left|\langle(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h)\widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right|$$

$$+ \left|\langle\widehat{\boldsymbol{\mu}}_{k,h}\widehat{\boldsymbol{V}}_{k,h+1} - \widehat{\boldsymbol{\mu}}_{k,h}\widecheck{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right| + \left|\langle(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h)\widecheck{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right|.$$

Proof. By definition, we have

$$\left|\langle\boldsymbol{\mu}_h\widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle - \langle\boldsymbol{\mu}_h\boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right|$$

$$= \left|\mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h V_{h+1}^*(s_h^k, a_h^k)\right| = \mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h V_{h+1}^*(s_h^k, a_h^k)$$

$$\leqslant \mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h\widecheck{V}_{k,h+1}(s_h^k, a_h^k) = \left|\mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h\widecheck{V}_{k,h+1}(s_h^k, a_h^k)\right|,$$

where the second equality holds since $\mathbb{P}_h$ is a valid distribution and $\widehat{V}_{k,h+1}(\cdot) \geqslant V_{h+1}^*(\cdot)$ under $\widehat{\Psi}_{k,h+1}$ by Lemma D.1, the first inequality holds since $\mathbb{P}_h$ is a valid distribution plus $V_{h+1}^*(\cdot) \geqslant \widecheck{V}_{k,h+1}(\cdot)$ under $\widecheck{\Psi}_{k,h+1}$ by Lemma D.1, and the last equality holds since $\mathbb{P}_h$ is a valid distribution plus $\widehat{V}_{k,h+1}(\cdot) \geqslant \widecheck{V}_{k,h+1}(\cdot)$ under $\widehat{\Psi}_{k,h+1} \cap \widecheck{\Psi}_{k,h+1}$ by Lemma D.1.

Therefore,

$$\left|\mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h V_{h+1}^*(s_h^k, a_h^k)\right| \leqslant \left|\mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h\widecheck{V}_{k,h+1}(s_h^k, a_h^k)\right|$$

$$= \left|\mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h}\widehat{V}_{k,h+1}(s_h^k, a_h^k) + \widehat{\mathbb{P}}_{k,h}\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h}\widecheck{V}_{k,h+1}(s_h^k, a_h^k)\right.$$

$$+ \widehat{\mathbb{P}}_{k,h} \check{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h \check{V}_{k,h+1}(s_h^k, a_h^k) \Big|$$

$$\leqslant \Big| \mathbb{P}_h \widehat{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h} \widehat{V}_{k,h+1}(s_h^k, a_h^k) \Big| + \Big| \widehat{\mathbb{P}}_{k,h} \Big[\widehat{V}_{k,h+1} - \check{V}_{k,h+1} \Big] (s_h^k, a_h^k) \Big|$$

$$+ \Big| \widehat{\mathbb{P}}_{k,h} \check{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h \check{V}_{k,h+1}(s_h^k, a_h^k) \Big|$$

$$= \Big| \langle (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big| + \Big| \langle \widehat{\boldsymbol{\mu}}_{k,h} (\widehat{\boldsymbol{V}}_{k,h+1} - \check{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|$$

$$+ \Big| \langle (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|.$$

$\square$

Based on Lemma C.11, we are ready to present Lemma C.12.

Lemma C.12. *In Algorithm 1, for any $k \in [K]$ and any $h \in [H]$, under $\widehat{\Psi}_{k,h+1} \cap \check{\Psi}_{k,h+1}$, we have*

$$\Big| \big[\mathbb{V}_h V_{h+1}^* \big] (s_h^k, a_h^k) - \big[\widehat{\mathbb{V}}_{k,h} \widehat{V}_{k,h+1} \big] (s_h^k, a_h^k) \Big| \leqslant \min \left\{ \Big\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}^2 \Big\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \| \boldsymbol{\phi}(s_h^k, a_h^k) \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + 4H\Delta_{k,h}, 2H^2 \right\},$$

where

$$\Delta_{k,h} = \Big\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1} \Big\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \| \boldsymbol{\phi}(s_h^k, a_h^k) \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \Big| \langle \widehat{\boldsymbol{\mu}}_{k,h} (\widehat{\boldsymbol{V}}_{k,h+1} - \check{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|$$

$$+ \Big\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{k,h+1} \Big\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \| \boldsymbol{\phi}(s_h^k, a_h^k) \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}.$$

Proof. By definition, we have

$$\Big| \big[\mathbb{V}_h V_{h+1}^* \big] (s_h^k, a_h^k) - \big[\widehat{\mathbb{V}}_{k,h} \widehat{V}_{k,h+1} \big] (s_h^k, a_h^k) \Big|$$

$$= \Big| \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^{*\,2}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle - \big[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]_{[0,H^2]}$$

$$+ \left\{ \big[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]_{[0,H]} \right\}^2 - \big[\langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]^2 \Big|$$

$$\leqslant \underbrace{\Big| \big[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]_{[0,H^2]} - \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^{*\,2}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|}_{I_1}$$

$$+ \underbrace{\Big| \left\{ \big[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]_{[0,H]} \right\}^2 - \big[\langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]^2 \Big|}_{I_2}$$

where the inequality holds due to the triangle inequality. We bound I_1 first.

$$I_1 = \Big| \big[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]_{[0,H^2]} - \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^{*\,2}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|$$

$$= \Big| \big[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]_{[0,H^2]} - \langle \boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle$$

$$+ \langle \boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle - \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^{*\,2}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|$$

$$\leqslant \Big| \big[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \big]_{[0,H^2]} - \langle \boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|$$

$$+ \Big| \langle \boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle - \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^{*\,2}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|$$

$$\leqslant \Big| \langle (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big| + 2H \Big| \langle \boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle - \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \Big|$$

$$\leqslant \left| \langle (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}^2, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right| + 2H \Big[\left| \langle (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right|$$

$$+ \left| \langle \widehat{\boldsymbol{\mu}}_{k,h} (\widehat{\boldsymbol{V}}_{k,h+1} - \check{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right| + \left| \langle (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right| \Big]$$

$$\leqslant \left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}^2 \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + 2H \Big[\left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$+ \left| \langle \widehat{\boldsymbol{\mu}}_{k,h} (\widehat{\boldsymbol{V}}_{k,h+1} - \check{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right| + \left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \Big]$$

where the first inequality holds due to the triangle inequality, the second inequality holds since $\mathbb{P}_h(\cdot \mid s_h^k, a_h^k) = \langle \boldsymbol{\mu}_h(\cdot), \boldsymbol{\phi}(s_h^k, a_h^k) \rangle$ is valid distribution and $\widehat{V}_{k,h+1}(\cdot), V_{h+1}^*(\cdot) \in [0, H]$, the third inequality holds due to Lemma C.11 under $\widehat{\Psi}_{k,h+1} \cap \check{\Psi}_{k,h+1}$, and the last inequality holds due to the Cauchy-Schwarz inequality.

For I_2, we have

$$I_2 = \left| \left\{ \left[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right]_{[0,H]} \right\}^2 - \left[\langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right]^2 \right|$$

$$= \left| \left[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right]_{[0,H]} + \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right|$$

$$\cdot \left| \left[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right]_{[0,H]} - \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right|$$

$$\leqslant 2H \left| \left[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right]_{[0,H]} - \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right|$$

$$\leqslant 2H \Big[\left| \langle (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right| + \left| \langle \widehat{\boldsymbol{\mu}}_{k,h} (\widehat{\boldsymbol{V}}_{k,h+1} - \check{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right|$$

$$+ \left| \langle (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right| \Big]$$

$$\leqslant 2H \Big[\left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \left| \langle \widehat{\boldsymbol{\mu}}_{k,h} (\widehat{\boldsymbol{V}}_{k,h+1} - \check{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \right|$$

$$+ \left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \Big]$$

where the first inequality holds since $0 \leqslant [\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle]_{[0,H]} \leqslant H$, $0 \leqslant \mathbb{P}_h V_{h+1}^*(s_h^k, a_h^k) = \langle \boldsymbol{\mu}_h \boldsymbol{V}_{h+1}^*, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle \leqslant H$, the second inequality holds due to Lemma C.11 under $\widehat{\Psi}_{k,h+1} \cap \check{\Psi}_{k,h+1}$, and the third inequality holds due to the Cauchy-Schwarz inequality. Combining the upper bound of I_1, I_2 in above two inequalities and using the fact that I_1 and I_2 are both bounded by H^2 give the final result. $\qquad \square$

C.2.2. VARIANCE OF $[\widehat{V}_{k,h+1} - V_{h+1}^*](\cdot)$

Lemma C.13. *In Algorithm 1, for any $k \in [K]$ and any $h \in [H]$, under $\widehat{\Psi}_{k,h+1} \cap \check{\Psi}_{k,h+1}$, we have $V_{h+1}^*(\cdot) \leqslant \widehat{V}_{k,h+1}(\cdot)$. Moreover, for any function $V : \mathcal{S} \mapsto [0, H]$ satisfying $V_{h+1}^*(\cdot) - \zeta \leqslant V(\cdot) \leqslant \widehat{V}_{k,h+1}(\cdot) + \zeta$, where $0 < \zeta \leqslant H$ is a constant we have*

$$\left[\mathbb{V}_h(V - V_{h+1}^*) \right](s_h^k, a_h^k) \leqslant \min \Big\{ 2H \Big[\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle - \langle \widehat{\boldsymbol{\mu}}_{k,h} \check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle$$

$$+ \left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$+ \left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \zeta \Big], H^2 \Big\}.$$

Proof. Initially, we have $V_{h+1}^*(\cdot) \leqslant \widehat{V}_{k,h+1}(\cdot)$ by Lemma D.1 under $\widehat{\Psi}_{k,h+1}$. Denote $\widetilde{V}(\cdot) = V(\cdot) - V_{h+1}^*(\cdot)$. By definition

of the variance, we have

$$
\begin{aligned}
[\mathbb{V}_h(V - V_{h+1}^*)](s_h^k, a_h^k) &= [\mathbb{V}_h \widetilde{V}](s_h^k, a_h^k) = \mathbb{P}_h \widetilde{V}^2(s_h^k, a_h^k) - [\mathbb{P}_h \widetilde{V}(s_h^k, a_h^k)]^2 \leqslant \mathbb{P}_h \widetilde{V}^2(s_h^k, a_h^k) \\
&\leqslant 2H \cdot \mathbb{P}_h |\widetilde{V}|(s_h^k, a_h^k) \leqslant 2H \cdot \left[\mathbb{P}_h \widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h V_{h+1}^*(s_h^k, a_h^k) + \zeta \right] \\
&\leqslant 2H \cdot \left[\mathbb{P}_h \widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h \check{V}_{k,h+1}(s_h^k, a_h^k) + \zeta \right],
\end{aligned}
\tag{21}
$$

where the first inequality holds since $[\mathbb{P}_h \widetilde{V}(s_h^k, a_h^k)]^2 \geqslant 0$, the second and third inequalities both holds since $-H \leqslant -\zeta \leqslant V(\cdot) - V_{h+1}^*(\cdot) \leqslant \widehat{V}_{k,h+1}(\cdot) - V_{h+1}^* + \zeta \leqslant 2H$, and the last inequality holds since $\check{V}_{k,h+1}(\cdot) \leqslant V_{h+1}^*(\cdot)$ under $\check{\Psi}_{k,h+1}$ by Lemma D.1.

On the other hand, by Cauchy-Schwarz inequality, we have

$$
\left| \mathbb{P}_h \widehat{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h} \widehat{V}_{k,h+1}(s_h^k, a_h^k) \right| \leqslant \left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}
\tag{22}
$$

$$
\left| \mathbb{P}_h \check{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h} \check{V}_{k,h+1}(s_h^k, a_h^k) \right| \leqslant \left\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{k,h+1} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}
\tag{23}
$$

Combining Eq (21), Eq (22), Eq (23) and using the fact that $[\mathbb{V}_h(V - V_{h+1}^*)](s_h^k, a_h^k) \leqslant H^2$ give the final result. $\quad\square$

C.3. Dependent Confidence Sets

Based on independent confidence sets $\bar{\mathcal{C}}_{k,h}, \widetilde{\mathcal{C}}_{k,h}, \check{\mathcal{C}}_{k,h}$ built above and Lemma C.12, C.13, dependent confidence sets $\widehat{\mathcal{C}}_{k,h}^{(1)}, \widehat{\mathcal{C}}_{k,h}^{(2)}$ are built in Lemma C.15, C.16, respectively. As a results, the confidence set $\widehat{\mathcal{C}}_{k,h}$, the goal of this section, holds trivially if $\widehat{\mathcal{C}}_{k,h}^{(1)}, \widehat{\mathcal{C}}_{k,h}^{(2)}$ both hold. We build confidence sets $\widehat{\mathcal{C}}_{k,h}^{(1)}$ and $\widehat{\mathcal{C}}_{k,h}^{(2)}$ elegantly because the radius of the confidence set $\widehat{\mathcal{C}}_{k,h}$ will exactly determine the sharpness of the regret obtained by LSVI-UCB$^+$ algorithm. In particular, we utilize the conservatism of elliptical potentials, which is detailed in Remark 7.4 in the main paper. To formally utilize this property, we first present Lemma C.14 to keep the magnitude of the considered MDS small with the conservatism of elliptical potentials.

Lemma C.14. *In Algorithm 1, for any $k \in [K]$ and any $h \in [H]$, we have*

$$
\widehat{\sigma}_{k,h}^{-1} \cdot \min \left\{ 1, \left\| \widehat{\sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \right\} \leqslant \frac{1}{H^2 \sqrt{d^5}}.
$$

Proof. In Algorithm 1, for any $k \in [K]$ and any $h \in [H]$, we have following two cases:

- If $\left\| \widetilde{\sigma}_{i,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i) \right\|_{\widetilde{\boldsymbol{\Lambda}}_{i,h}^{-1}} \leqslant 1/(H^3 d^5)$, then $\varsigma_{k,h} = \sqrt{H}$ such that $\widehat{\sigma}_{i,h} = \widetilde{\sigma}_{i,h}$. In this case, we have

$$
\left\| \widehat{\sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} = \left\| \widetilde{\sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \leqslant \sqrt{H^3 d^5} \left\| \widetilde{\sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i) \right\|_{\widetilde{\boldsymbol{\Lambda}}_{k,h}^{-1}} \leqslant \sqrt{H^3 d^5}/(H^3 d^5) \leqslant 1/\sqrt{H^3 d^5}.
$$

where the inequality holds since $H^3 d^5 \cdot \widehat{\boldsymbol{\Lambda}}_{k,h} \succeq \widetilde{\boldsymbol{\Lambda}}_{k,h}$ by following facts:

$$
H^3 d^5 \cdot \widehat{\boldsymbol{\Lambda}}_{k,h} = H^3 d^5 \cdot \left(\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \boldsymbol{\phi}(s_h^j, a_h^j)^\top + \lambda \mathbf{I} \right) \succeq \sum_{i=1}^{k-1} \widetilde{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \boldsymbol{\phi}(s_h^i, a_h^i)^\top + \lambda \mathbf{I} = \widetilde{\boldsymbol{\Lambda}}_{k,h},
$$

where the inequality holds since $\widehat{\sigma}_{i,h} \leqslant \sqrt{H^3 d^5} \widetilde{\sigma}_{i,h}$ in Algorithm 1, which implies $H^3 d^5 \cdot \widehat{\boldsymbol{\Lambda}}_{k,h} - \widetilde{\boldsymbol{\Lambda}}_{k,h}$ is a semi-positive definite matrix.

Therefore, the conclusion holds in this case since $\widehat{\sigma}_{k,h} \geqslant \sqrt{H}$.

- Otherwise, $\varsigma_{k,h} = H^2 \sqrt{d^5}$, such that $\widehat{\sigma}_{k,h} \geqslant H^2 \sqrt{d^5}$. In this case, the conclusion still holds since $\min\{1, \|\widehat{\sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}\} \leqslant 1$.

$\quad\square$

Now we are ready to prove Lemma C.15 which builds the dependent confidence set $\widehat{\mathcal{C}}_{k,h}^{(1)}$, based on independent confidence sets $\bar{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}$ and Lemma C.12. Indeed, the confidence set $\widehat{\mathcal{C}}_{k,h}^{(1)}$ corresponds to the deviation term of the form $[(\widehat{\mathbb{P}}_{k,h} - \mathbb{P}_h)V_{h+1}^*](s_h^k, a_h^k)$ in the main paper.

Lemma C.15. *In Algorithm 1, for any $\delta \in (0,1)$, any $k \in [K]$ and fixed $h \in [H]$, under $\widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$, with probability at least $1 - 4\delta/H$:*

$$\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h}^{(1)} \cap \bar{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h},$$

where

$$\widehat{\mathcal{C}}_{k,h}^{(1)} = \left\{ \boldsymbol{\mu} : \left\| (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}) \boldsymbol{V}_{h+1}^* \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widehat{\beta}^{(1)} \right\},$$

$$\widehat{\beta}^{(1)} = 8\sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) \log\left(\frac{4K^2H}{\delta}\right)} + 4\log\left(\frac{4K^2H}{\delta}\right) + H\sqrt{\lambda d}. \tag{24}$$

Proof. Let $\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{x}_i = \widehat{\sigma}_{i,h}^{-1} \phi(s_h^i, a_h^i)$, $\mathbf{Z}_i = \lambda \mathbf{I} + \sum_{j=1}^i \boldsymbol{x}_i \boldsymbol{x}_i^\top$, and $\eta_i = \widehat{\sigma}_{i,h}^{-1} \boldsymbol{\epsilon}_h^{i\top} \boldsymbol{V}_{h+1}^* \cdot \mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\}$. Now that V_{h+1}^* is a fixed function and $\mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\}$ is $\mathcal{G}_i$-measurable, it is clear that $\boldsymbol{x}_i$ are $\mathcal{G}_i$-measurable and η_i is $\mathcal{G}_{i+1}$-measurable.

Besides, we have $\mathbb{E}[\eta_i \mid \mathcal{G}_i] = 0$. Since $\widehat{\sigma}_{i,h} \geqslant \varsigma_{i,h} \geqslant \sqrt{H}$, $|\eta_i| \leqslant \sqrt{H}$ and $\|\boldsymbol{x}_i\|_2 \leqslant 1/\sqrt{H}$. In particular, we claim $|\eta_i \cdot \min\{1, \|\boldsymbol{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\}| \leqslant 1$ because of the following three facts: (i) $|\boldsymbol{\epsilon}_h^{i\top} \boldsymbol{V}_{h+1}^* \cdot \mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\}| \leqslant H$ holds by $|V_{h+1}*(\cdot)| \leqslant H$; (ii) $\widehat{\sigma}_{i,h}^{-1} \cdot \min\{1, \|\boldsymbol{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\} \leqslant 1/(H^2\sqrt{d^5})$ holds by Lemma C.14; and (iii) $|\mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\}| \leqslant 1$.

Furthermore, it holds that

$$\begin{aligned}
\mathbb{E}[\eta_i^2 \mid \mathcal{G}_i] =& \widehat{\sigma}_{i,h}^{-2} \cdot \mathbb{1}\left\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\right\} \cdot \mathbb{1}\left\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\right\} [\mathbb{V}_h V_{h+1}^*](s_h^i, a_h^i) \\
\leqslant& \widehat{\sigma}_{i,h}^{-2} \cdot \mathbb{1}\left\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\right\} \bigg[\left[\widehat{\mathbb{V}}_{i,h} \widehat{V}_{i,h+1}\right](s_h^i, a_h^i) \\
&+ \min\left\{ \left\|(\widehat{\boldsymbol{\mu}}_{i,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{i,h+1}^2\right\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}} \|\phi(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}^{-1}} + 4H\Delta_{i,h}, 2H^2\right\} \bigg] \\
\leqslant& \widehat{\sigma}_{i,h}^{-2} \bigg[\left[\widehat{\mathbb{V}}_{i,h} \widehat{V}_{i,h+1}\right](s_h^i, a_h^i) + \min\left\{\widetilde{\beta} \|\phi(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}^{-1}} + 4H\delta_{i,h}, 2H^2\right\} \bigg] \\
\leqslant& 1,
\end{aligned}$$

where the first inequality holds due to Lemma C.12 under $\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}$, the second inequality holds due to the definition of indicator function, and the last inequality holds due to the definition of $\widehat{\sigma}_{i,h}$. Here $\Delta_{i,h}$ and $\delta_{i,h}$ are given by

$$\begin{aligned}
\Delta_{i,h} =& \left\|(\widehat{\boldsymbol{\mu}}_{i,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{i,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}} \|\phi(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}^{-1}} + \left|\left\langle \widehat{\boldsymbol{\mu}}_{i,h}\left(\widehat{\boldsymbol{V}}_{i,h+1} - \check{\boldsymbol{V}}_{i,h+1}\right), \phi(s_h^i, a_h^i)\right\rangle\right| \\
&+ \left\|(\widehat{\boldsymbol{\mu}}_{i,h} - \boldsymbol{\mu}_h) \check{\boldsymbol{V}}_{i,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}} \|\phi(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}^{-1}}, \\
\delta_{i,h} =& \bar{\beta} \|\phi(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}^{-1}} + \left|\left\langle \widehat{\boldsymbol{\mu}}_{i,h}\left(\widehat{\boldsymbol{V}}_{i,h+1} - \check{\boldsymbol{V}}_{i,h+1}\right), \phi(s_h^i, a_h^i)\right\rangle\right| + \check{\beta} \|\phi(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{i,h}^{-1}}.
\end{aligned}$$

Then, by Lemma F.4, with probability at least $1 - \delta/H$, for all $k \in [K]$ and fixed $h \in [H]$,

$$\begin{aligned}
&\left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i) \boldsymbol{\epsilon}_h^{i\top} \boldsymbol{V}_{h+1}^* \mathbb{1}\left\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\right\} \mathbb{1}\left\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\right\} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \\
&\leqslant 8\sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) \log\left(\frac{4K^2H}{\delta}\right)} + 4\log\left(\frac{4K^2H}{\delta}\right).
\end{aligned} \tag{25}$$

Denote $\mathcal{E}_h^{(1)}$ as the event that $\boldsymbol{\mu}_h \in \bigcap_{k \in [K]} \bar{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}$ and Eq. (25) hold, which happens with probability at least $1 - 4\delta/H$ by taking a union bound. In addition, we claim that with probability at least $1 - 4\delta/H$, for all $k \in [K]$ and fixed $h \in [H]$,

$$\left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top \boldsymbol{V}_{h+1}^* \cdot \mathbb{1}\left\{ \boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h} \right\} \cdot \mathbb{1}\left\{ \widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1} \right\} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$= \left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top \boldsymbol{V}_{h+1}^* \cdot \mathbb{1}\left\{ \widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1} \right\} \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$\leqslant 8\sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) \log\left(\frac{4K^2 H}{\delta}\right)} + 4\log\left(\frac{4K^2 H}{\delta}\right),$$

where the equality holds since under event $\mathcal{E}_h^{(1)}$, for any $i \in [K]$, $\mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} = 1$. Moreover, if we further assume $\widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$ holds, which means $\mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\} = 1$ for any $i \in [k]$, then with probability at least $1 - 4\delta/H$, for any $k \in [K]$ and fixed $h \in [H]$:

$$\left\| \left(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h\right) \boldsymbol{V}_{h+1}^* \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant 8\sqrt{d \log\left(1 + \frac{K}{Hd\lambda}\right) \log\left(\frac{4K^2 H}{\delta}\right)} + 4\log\left(\frac{4K^2 H}{\delta}\right) + H\sqrt{\lambda d} = \widehat{\beta}^{(1)}.$$

since

$$\left\| \left(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h\right) \boldsymbol{V}_{h+1}^* \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant H\sqrt{\lambda d} + \left\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) {\boldsymbol{\epsilon}_h^i}^\top \boldsymbol{V}_{h+1}^* \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

with a similar argument as in Eq. (17). Thus, we conclude that for any $k \in [K]$ and fixed $h \in [H]$, under $\widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$, with probability at least $1 - 4\delta/H$:

$$\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h}^{(1)} \cap \bar{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}.$$

$\square$

Subsequently, we prove Lemma C.16 which builds the dependent confidence set $\widehat{\mathcal{C}}_{k,h}^{(2)}$, based on independent confidence sets $\bar{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}$ and Lemma C.13. The confidence set $\widehat{\mathcal{C}}_{k,h}^{(2)}$ corresponds to the deviation term of the form $[(\widehat{\mathbb{P}}_{k,h} - \mathbb{P}_h)(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)$ in main paper, which is controlled to be small in LSVI-UCB$^+$.

Lemma C.16. *In Algorithm 1, for any $\delta \in (0,1)$, any $k \in [K]$ and fixed $h \in [H]$, under $\widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$, with probability at least $1 - 3\delta/H$:*

$$\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h}^{(2)} \cap \bar{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}.$$

where

$$\widehat{\mathcal{C}}_{k,h}^{(2)} = \left\{ \boldsymbol{\mu} : \left\| \left(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h}\right)\left(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}_{h+1}^*\right) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widehat{\beta}^{(2)} \right\},$$

$$\widehat{\beta}^{(2)} = 8\sqrt{\frac{2}{Hd^2} \log\left(1 + \frac{K}{Hd\lambda}\right) \left[\log\left(\frac{4K^2 H}{\delta}\right) + dJ\log\left(1 + \frac{4KL}{H\sqrt{\lambda}}\right) + d^2 J\log\left(1 + \frac{8K^2 \widehat{B}^2 \sqrt{d}}{H^2 \lambda^2}\right)\right]}$$

$$+ \frac{4}{H\sqrt{d^5}}\left[\log\left(\frac{4K^2 H}{\delta}\right) + dJ\log\left(1 + \frac{4KL}{H\sqrt{\lambda}}\right) + d^2 J\log\left(1 + \frac{8K^2 \widehat{B}^2 \sqrt{d}}{H^2 \lambda^2}\right)\right] + H\sqrt{\lambda d} + 2. \quad (26)$$

Here $J = dH\log(1 + K), L = W + K/\lambda$ and $\widehat{B}$ is a constant satisfying $\widehat{\beta} \leqslant \widehat{B}$ with $\widehat{\beta}$ given in Lemma C.17.

Proof. It suffices to upper bound $\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i) \epsilon_h^{i\top} (\widehat{V}_{k,h+1} - V_{h+1}^*)\|_{\widehat{\Lambda}_{k,h}^{-1}}$ with a similar argument as Eq. (17). Besides, we need to build a uniform convergence argument by covering net since $\widehat{V}_{k,h+1}(\cdot)$ is $\mathcal{F}_{k-1,h}$-measurable. As stated in the proof of Lemma C.8, $\widehat{V}_{k,h} \in \widehat{\mathcal{V}}$, where $\widehat{\mathcal{V}}$ is defined in Definition C.5, with $J = dH \log(1 + K)$, $L = W + K/\lambda$ and $B = \widehat{B}$. Here, $\widehat{B}$ is a constant satisfying $\check{\beta} \leqslant \widehat{B}$ with $\check{\beta}$ specified in Lemma C.17.

Then, for a fixed function $V(\cdot) \in \widehat{\mathcal{V}} : \mathcal{S} \mapsto [0, H]$ and a constant $\zeta = H\sqrt{\lambda}/K$, let $\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{x}_i = \widehat{\sigma}_{i,h}^{-1} \phi(s_h^i, a_h^i)$, $\mathbf{Z}_i = \lambda \mathbf{I} + \sum_{j=1}^i \boldsymbol{x}_i \boldsymbol{x}_i^\top$, and $\eta_i = \widehat{\sigma}_{i,h}^{-1} \epsilon_h^{i\top} (V - V_{h+1}^*) \cdot \mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} \cdot \mathbb{1}\{V_{h+1}^* - \zeta \leqslant V \leqslant \widehat{V}_{i,h+1} + \zeta\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\}$.

Since $V(\cdot)$ and $V(\cdot)_{h+1}^*$ are fixed functions, and $\widehat{V}_{i,h+1}$ and $\mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\}$ are $\mathcal{G}_i$-measurable, it is clear that $\boldsymbol{x}_i$ is $\mathcal{G}_i$-measurable and η_i is $\mathcal{G}_{i+1}$-measurable. Besides, we have $\mathbb{E}[\eta_i \mid \mathcal{G}_i] = 0$. Since $\widehat{\sigma}_{i,h} \geqslant \varsigma_{i,h} \geqslant \sqrt{H}$, $|\eta_i| \leqslant \sqrt{H}$ and $\|\boldsymbol{x}_i\|_2 \leqslant 1/\sqrt{H}$.

Similar to the proof in Lemma C.15, we claim $|\eta_i \cdot \min\{1, \|\boldsymbol{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\}| \leqslant 1/H\sqrt{d^5}$ because of the following three facts: (i) $|\epsilon_h^{i\top} (V - V_{h+1}^*) \cdot \mathbb{1}\{V_{h+1}^* - \zeta \leqslant V \leqslant \widehat{V}_{i,h+1} + \zeta\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\}| \leqslant H$ holds by $|(V(\cdot) - V_{h+1}^*(\cdot)) \cdot \mathbb{1}\{V_{h+1}^* - \zeta \leqslant V \leqslant \widehat{V}_{i,h+1} + \zeta\}| \leqslant H$; (ii) $\widehat{\sigma}_{i,h}^{-1} \cdot \min\{1, \|\boldsymbol{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\} \leqslant 1/(H^2\sqrt{d^5})$ holds by Lemma C.14; and (iii) $|\mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\}| \leqslant 1$. Furthermore, it holds that

$$
\begin{aligned}
\mathbb{E}[\eta_i^2 \mid \mathcal{G}_i] =& \widehat{\sigma}_{i,h}^{-2} \cdot \mathbb{1}\left\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\right\} \cdot \mathbb{1}\left\{V_{h+1}^* - \zeta \leqslant V \leqslant \widehat{V}_{i,h+1} + \zeta\right\} \cdot \mathbb{1}\left\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\right\} \\
& \cdot [\mathbb{V}_h(V - V_{h+1}^*)](s_h^i, a_h^i) \\
\leqslant& \widehat{\sigma}_{i,h}^{-2} \cdot \mathbb{1}\left\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\right\} \cdot 2H\Big[\widehat{\mathbb{P}}_{k,h} \widehat{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h} \check{V}_{k,h+1}(s_h^k, a_h^k) \\
& + \left\|(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{V}_{k,h+1}\right\|_{\widehat{\Lambda}_{k,h}} \|\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}} \\
& + \left\|(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \check{V}_{k,h+1}\right\|_{\widehat{\Lambda}_{k,h}} \|\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}} + \zeta\Big] \\
\leqslant& \widehat{\sigma}_{i,h}^{-2} \cdot 2H\Big[\widehat{\mathbb{P}}_{k,h} \widehat{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h} \check{V}_{k,h+1}(s_h^k, a_h^k) + \bar{\beta} \|\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}} + \check{\beta} \|\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}} + \zeta\Big] \\
\leqslant& \frac{2}{Hd^3},
\end{aligned}
$$

where the first inequality holds due to Lemma C.13 under $\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}$, the second inequality holds due to the definition of indicator function, and the last inequality holds due to the definition of $\widehat{\sigma}_{i,h}$.

Then, by Lemma F.4, for all $k \in [K]$ and fixed $h \in [H]$, with probability at least $1 - \delta/H$:

$$
\begin{aligned}
& \Big\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i) \epsilon_h^{i\top} (V - V_{h+1}^*) \\
& \cdot \mathbb{1}\left\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\right\} \cdot \mathbb{1}\left\{V_{h+1}^* - \zeta \leqslant V \leqslant \widehat{V}_{i,h+1} + \zeta\right\} \cdot \mathbb{1}\left\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\right\} \Big\|_{\widehat{\Lambda}_{k,h}^{-1}} \\
& \leqslant 8\sqrt{\frac{2}{Hd^2} \log\left(1 + \frac{K}{Hd\lambda}\right) \log\left(\frac{4K^2 H}{\delta}\right)} + \frac{4}{H\sqrt{d^5}} \log\left(\frac{4K^2 H}{\delta}\right).
\end{aligned}
$$

We further proceed our proof under the event that $\widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$ holds, which implies $\mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\} = 1$ for any $i \in [k]$. Denote $\mathcal{E}_h^{(2)}$ as the event that $\boldsymbol{\mu}_h \in \bigcap_{k \in [K]} \bar{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}$ and the above inequality holds, which happens with probability at least $1 - 3\delta/H$ by taking a union bound. In addition, we claim that under $\widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$, with probability at least $1 - 3\delta/H$, for all $k \in [K]$ and fixed $h \in [H]$,

$$
\Big\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i) \epsilon_h^{i\top} (V - V_{h+1}^*)
$$

$$\cdot \, \mathbb{1}\left\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\right\} \cdot \mathbb{1}\left\{\boldsymbol{V}_{h+1}^* - \zeta \leq \boldsymbol{V} \leq \widehat{\boldsymbol{V}}_{i,h+1} + \zeta\right\} \cdot \mathbb{1}\left\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\right\}\Big\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$= \left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\epsilon_h^i}^\top \left(\boldsymbol{V} - \boldsymbol{V}_{h+1}^*\right) \cdot \mathbb{1}\left\{\boldsymbol{V}_{h+1}^* - \zeta \leq \boldsymbol{V} \leq \widehat{\boldsymbol{V}}_{i,h+1} + \zeta\right\}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$\leq 8\sqrt{\frac{2}{Hd^2}\log\left(1 + \frac{K}{Hd\lambda}\right)\log\left(\frac{4K^2 H}{\delta}\right)} + \frac{4}{H\sqrt{d^5}}\log\left(\frac{4K^2 H}{\delta}\right),$$

where the equality holds since under event $\mathcal{E}_h^{(2)} \cap \widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$, for any $i \in [K]$, $\mathbb{1}\{\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{i,h} \cap \check{\mathcal{C}}_{i,h}\} \cdot \mathbb{1}\{\widehat{\Psi}_{i,h+1} \cap \check{\Psi}_{i,h+1}\} = 1$.

Denote the ε-cover of function class $\widehat{\mathcal{V}}$ as $\widehat{\mathcal{N}}_\varepsilon$. Since $\widehat{V}_{k,h+1}(\cdot) \in \widehat{\mathcal{V}}$, for any $\widehat{V}_{k,h+1}$, there exists a $V' \in \widehat{\mathcal{N}}_\varepsilon$, such that $\|\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}'\|_\infty \leq \varepsilon$. This implies $\boldsymbol{V}_{h+1}^* - \varepsilon \leq \widehat{\boldsymbol{V}}_{k,h+1} - \varepsilon \leq \boldsymbol{V}' \leq \widehat{\boldsymbol{V}}_{k,h+1} + \varepsilon \leq \widehat{\boldsymbol{V}}_{i,h+1} + \varepsilon$ for any $i \in [k]$, where the first inequality holds by Lemma D.1 under $\widehat{\Psi}_{h+1}$, and the last inequality holds by definition of optimistic value function in Algorithm 1.

In addition, setting $\varepsilon = \zeta = H\sqrt{\lambda}/K$ makes $\mathbb{1}\{\boldsymbol{V}_{h+1}^* - \zeta \leq \boldsymbol{V}' \leq \widehat{\boldsymbol{V}}_{i,h+1} + \zeta\} = 1$ for any $i \in [k]$. Moreover, since $\|{\epsilon_h^i}^\top(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}')\|_2 \leq \|\epsilon_h^i\|_1 \|\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}'\|_\infty = 2\varepsilon$ and $\|\sum_{i=1}^{k-1}\widehat{\sigma}_{i,h}^{-2}\boldsymbol{\phi}(s_h^i, a_h^i)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \leq K/(H\sqrt{\lambda})$, we have

$$\left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\epsilon_h^i}^\top \left(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}'\right) \cdot \mathbb{1}\left\{\boldsymbol{V}_{h+1}^* - \zeta \leq \boldsymbol{V}' \leq \widehat{\boldsymbol{V}}_{i,h+1} + \zeta\right\}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \leq \frac{2\varepsilon K}{H\sqrt{\lambda}}. \tag{27}$$

This further implies that the following inequality holds with probability at least $1 - 3\delta/H$:

$$\left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\epsilon_h^i}^\top \left(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}_{h+1}^*\right)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$= \left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\epsilon_h^i}^\top \left(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}_{h+1}^*\right) \cdot \mathbb{1}\left\{\boldsymbol{V}_{h+1}^* - \zeta \leq \boldsymbol{V}' \leq \widehat{\boldsymbol{V}}_{i,h+1} + \zeta\right\}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$\leq \left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\epsilon_h^i}^\top \left(\boldsymbol{V}' - \boldsymbol{V}_{h+1}^*\right) \cdot \mathbb{1}\left\{\boldsymbol{V}_{h+1}^* - \zeta \leq \boldsymbol{V}' \leq \widehat{\boldsymbol{V}}_{i,h+1} + \zeta\right\}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$+ \left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\epsilon_h^i}^\top \left(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}'\right) \cdot \mathbb{1}\left\{\boldsymbol{V}_{h+1}^* - \zeta \leq \boldsymbol{V}' \leq \widehat{\boldsymbol{V}}_{i,h+1} + \zeta\right\}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$\leq \left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\epsilon_h^i}^\top \left(\boldsymbol{V}' - \boldsymbol{V}_{h+1}^*\right) \cdot \mathbb{1}\left\{\boldsymbol{V}_{h+1}^* - \zeta \leq \boldsymbol{V}' \leq \widehat{\boldsymbol{V}}_{i,h+1} + \zeta\right\}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \frac{2\varepsilon K}{H\sqrt{\lambda}}$$

$$\leq 8\sqrt{\frac{2}{Hd^2}\log\left(1 + \frac{K}{Hd\lambda}\right)\left[\log\left(\frac{4K^2 H}{\delta}\right) + \log\left|\widehat{\mathcal{N}}_\varepsilon\right|\right]} + \frac{4}{H\sqrt{d^5}}\left[\log\left(\frac{4K^2 H}{\delta}\right) + \log\left|\widehat{\mathcal{N}}_\varepsilon\right|\right] + \frac{2\varepsilon K}{H\sqrt{\lambda}}$$

where the first inequality is due to triangle inequality, the second inequality holds by Eq. (27), and the third inequality holds by a union bound over all functions in $\widehat{\mathcal{N}}_\varepsilon$ with

$$\log\left|\widehat{\mathcal{N}}_\varepsilon\right| \leq dJ\log\left(1 + \frac{4L}{\varepsilon}\right) + d^2 J\log\left(1 + \frac{8\widehat{B}^2\sqrt{d}}{\lambda\varepsilon^2}\right)$$

according to Lemma F.11 with $J = dH\log(1 + K)$ by Lemma F.8.

Similar to Eq. (17), for any $k \in [K]$ and fixed $h \in [H]$, under $\widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$, we have that, with probability at least

$1 - 3\delta/H$:

$$
\left\|\left(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h\right)\left(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}_{h+1}^*\right)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}}
$$

$$
\leqslant \frac{1}{\sqrt{\lambda}} \cdot \lambda H \sqrt{d} + \left\|\sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \boldsymbol{\epsilon}_h^{i\,\top}\left(\widehat{\boldsymbol{V}}_{k,h+1} - \boldsymbol{V}_{h+1}^*\right)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}
$$

$$
\leqslant 8\sqrt{\frac{2}{Hd^2}\log\left(1 + \frac{K}{Hd\lambda}\right)\left[\log\left(\frac{4K^2 H}{\delta}\right) + dJ\log\left(1 + \frac{4KL}{H\sqrt{\lambda}}\right) + d^2 J\log\left(1 + \frac{8K^2\widehat{B}^2\sqrt{d}}{H^2\lambda^2}\right)\right]} \tag{28}
$$

$$
+ \frac{4}{H\sqrt{d^5}}\left[\log\left(\frac{4K^2 H}{\delta}\right) + dJ\log\left(1 + \frac{4KL}{H\sqrt{\lambda}}\right) + d^2 J\log\left(1 + \frac{8K^2\widehat{B}^2\sqrt{d}}{H^2\lambda^2}\right)\right] + H\sqrt{\lambda d} + 2
$$

$$
= \widehat{\beta}^{(2)},
$$

where the last inequality holds by the above proved self-normalized bound and $\varepsilon = \zeta = H\sqrt{\lambda}/K$. Thus, we conclude that for any $k \in [K]$ and fixed $h \in [H]$, under $\widehat{\Psi}_{h+1} \cap \widecheck{\Psi}_{h+1}$, with probability at least $1 - 3\delta/H$:

$$
\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h}^{(2)} \cap \overline{\mathcal{C}}_{k,h} \cap \widecheck{\mathcal{C}}_{k,h}.
$$

$\square$

C.4. Proof of Lemma 7.5

Now we are ready to prove Lemma 7.5, i.e., building the sharp confidence set $\widehat{\mathcal{C}}_{k,h}$, in the main paper, based on above building blocks including confidence sets $\overline{\mathcal{C}}_{k,h}, \widetilde{\mathcal{C}}_{k,h}, \widecheck{\mathcal{C}}_{k,h}, \widehat{\mathcal{C}}_{k,h}^{(1)}, \widehat{\mathcal{C}}_{k,h}^{(2)}$, and Lemma C.12 and C.13 for upper bounding variances of value functions. In the following, we present Lemma C.17, which is the full version of Lemma 7.5 in the main paper.

Lemma C.17. *Set $\widehat{\beta} = \widehat{\beta}^{(1)} + \widehat{\beta}^{(2)}$ with $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ given in Lemma C.15 and C.16, respectively. Then for any $\delta \in (0,1)$, with probability at least $1 - 7\delta$, we have that simultaneously for any $k \in [K]$ and any $h \in [H]$,*

$$
\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h} \cap \widehat{\mathcal{C}}_{k,h}^{(1)} \cap \widehat{\mathcal{C}}_{k,h}^{(2)} \cap \overline{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \cap \widecheck{\mathcal{C}}_{k,h}
$$

and

$$
\left|[\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\right| \leqslant U_{k,h}
$$

$$
\left[\mathbb{V}_h(V - V_{h+1}^*)\right](s_h^k, a_h^k) \leqslant E_{k,h}
$$

where

$$
U_{k,h} = \min\left\{\widetilde{\beta}\left\|\boldsymbol{\phi}(s_h^k, a_h^k)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + 4H\left[\left|\langle\widehat{\boldsymbol{\mu}}_{k,h}(\widehat{\boldsymbol{V}}_{k,h+1} - \widecheck{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right| + \overline{\beta}\left\|\boldsymbol{\phi}(s_h^k, a_h^k)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}\right.\right.
$$

$$
\left.\left. + \widecheck{\beta}\left\|\boldsymbol{\phi}(s_h^k, a_h^k)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}\right], 2H^2\right\} \tag{29}
$$

$$
E_{k,h} = \min\left\{H\left[\langle\widehat{\boldsymbol{\mu}}_{k,h}\widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle - \langle\widehat{\boldsymbol{\mu}}_{k,h}\widecheck{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle + \overline{\beta}\left\|\boldsymbol{\phi}(s_h^k, a_h^k)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \widecheck{\beta}\left\|\boldsymbol{\phi}(s_h^k, a_h^k)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}\right.\right.
$$

$$
\left.\left. + H\sqrt{\lambda}/K\right], H^2\right\}. \tag{30}
$$

Here $\overline{\beta}, \widetilde{\beta}, \widecheck{\beta}$ are specified in Lemma C.8, C.9, C.10, respectively, and $\zeta = H\sqrt{\lambda}/K$.

Proof. We first prove the following claim:

For any $\delta \in (0,1)$, any $k \in [K]$ and fixed $h \in [H]$, with probability at least $1 - 7(H-h)\delta/H$, for any h' such that $h \leqslant h' \leqslant H$:

$$
\boldsymbol{\mu}_{h'} \in \widehat{\mathcal{C}}_{k,h'} \cap \widehat{\mathcal{C}}_{k,h'}^{(1)} \cap \widehat{\mathcal{C}}_{k,h'}^{(2)} \cap \overline{\mathcal{C}}_{k,h'} \cap \widetilde{\mathcal{C}}_{k,h'} \cap \widecheck{\mathcal{C}}_{k,h'},
$$

and

$$\left|[\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\right| \leqslant U_{k,h}$$
$$\left[\mathbb{V}_h(V - V_{h+1}^*)\right](s_h^k, a_h^k) \leqslant E_{k,h}$$

hold simultaneously.

We prove this claim by introduction.

- We first prove the claim for $h = H$. Since $\check{V}_{k,H+1}(\cdot) = V_{H+1}^*(\cdot) = \widehat{V}_{k,H+1}(\cdot) = 0$ in Algorithm 1 for any $k \in [K]$, the conclusion holds for sure.

- Assume the claim holds for $h + 1 \leqslant H$. Then, for any $k \in [K]$ and fixed $h + 1 \in [H]$, with probability at least $1 - 7(H - h - 1)\delta/H$, $\boldsymbol{\mu}_{h'} \in \widehat{\mathcal{C}}_{k,h'} \cap \check{\mathcal{C}}_{k,h'}$ for any $h + 1 \leqslant h' \leqslant H$, which implies $\check{\Psi}_{h+1} \cap \widehat{\Psi}_{h+1}$ holds.

  Combined with conclusions from Lemma C.15 and Lemma C.16, for any $k \in [K]$, the following events holds with probability at least $1 - 7\delta/H$:

$$\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h}^{(1)} \cap \widehat{\mathcal{C}}_{k,h}^{(2)} \cap \bar{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}.$$

  Moreover, we have

$$\left|[\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k)\right| \leqslant U_{k,h}$$
$$\left[\mathbb{V}_h(V - V_{h+1}^*)\right](s_h^k, a_h^k) \leqslant E_{k,h}$$

  under the event $\widehat{\Psi}_{h+1} \cap \check{\Psi}_{h+1}$ and $\bar{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}$ by Lemma C.12 and C.13 with $\zeta = H\sqrt{\lambda}/K$.

  Considering $\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h}^{(1)} \cap \widehat{\mathcal{C}}_{k,h}^{(2)}$, we have

$$\left\|\left(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h}\right)\widehat{V}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \left\|\left(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h}\right)V_{h+1}^*\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} + \left\|\left(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h}\right)\left(\widehat{V}_{k,h+1} - V_{h+1}^*\right)\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}}$$
$$\leqslant \widehat{\beta}^{(1)} + \widehat{\beta}^{(2)} = \widehat{\beta},$$

  which implies $\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h}$ since $\widehat{\mathcal{C}}_{k,h} = \{\boldsymbol{\mu} : \|(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{k,h})\widehat{V}_{k,h+1}\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \leqslant \widehat{\beta}\}$. In other words, $\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h} \cap \widehat{\mathcal{C}}_{k,h}^{(1)} \cap \widehat{\mathcal{C}}_{k,h}^{(2)} \cap \bar{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}$.

  Thus, by taking a union bound over these two events, we claim that with probability at least $1 - 7(H - h)\delta/H$, the claim holds for h.

Therefore, the claim is proved by induction and setting $h = 1$ gives the desired results in Lemma C.17. $\qquad\square$

D. Regret Upper Bound

In this section, we upper bound the final regret, where we show that the total regret is roughly bounded by the summation of the exploration bonus, i.e.,

$$\text{Regret}(K) \leqslant \sum_{k=1}^{K}\left[\widehat{V}_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)\right] \lesssim \sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$\lesssim \widehat{\beta}\underbrace{\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}}_{\tilde{O}(\sqrt{HT})}\underbrace{\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\|\widehat{\sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}^2}}_{\tilde{O}(\sqrt{Hd})} \leqslant \tilde{O}(Hd\sqrt{T}).$$

- The first inequality holds by the optimism of constructed value functions, which is built in Lemma D.1.

- The third inequality is proved in Lemma D.4, which bounds the cumulative difference between the optimistic value function $\widehat{V}_{k,h}$ and the value function associated with policy π^k value function $V_h^{\pi^k}$ in Lemma D.4. We further bound the cumulative difference between the optimistic value function $\widehat{V}_{k,h}$ and the pessimistic value function $\widecheck{V}_{k,h}$ in Lemma D.5.

- The fourth inequality holds by Cauchy-Schwarz inequality, where first summation of estimated variance is bounded in Lemma D.9 in Appendix D.3. The summation of estimated variance $\sum_{k=1}^{K} \sum_{h=1}^{H} \widehat{\sigma}_{k,h}^2 = \widetilde{O}(HT)$, utilizing the the Law of Total Variance in (Lattimore & Hutter, 2012), detailed in Lemma D.6. The second summation can be bounded by classical Elliptical Potential Lemma, presented in Lemma F.5 in Appendix F.

Putting these building blocks together, we are finally ready to upper bound the regret in Appendix D.4 as $\widetilde{O}(Hd\sqrt{T} + H^4 d^4 + H^3 d^6)$. Before the formal proof begins, denote the event when the conclusion of Lemma 7.5 holds as Υ. Also denote the event that the conclusion of Lemma D.2, D.3 and D.6 holds as Ξ_1, Ξ_2 and Ξ_3, respectively. The final regret bound is a high probability bound builds under event $\Upsilon \cap \Xi_1 \cap \Xi_2 \cap \Xi_3$.

D.1. Monotonicity

In this subsection, we build the optimism of the constructed optimistic value function $\widehat{V}_{k,h}$, and the pessimism of the constructed pessimistic value function $\widecheck{V}_{k,h}$ over the optimal value function V_h^* in Lemma D.1. These are the preliminaries for our later proofs.

Lemma D.1 (Optimism and Pessimism). *In Algorithm 1, if $\widecheck{\Psi}_{k,h} \cap \widehat{\Psi}_{k,h}$ holds, then for any $k \in [K]$ and any $h \in [H]$, we have*

$$\widecheck{V}_{k,h}(s) \overset{(a)}{\leqslant} V_h^*(s) \overset{(b)}{\leqslant} \widehat{V}_{k,h}(s), \quad \forall s \in \mathcal{S}. \tag{31}$$

Proof. We prove two inequalities by induction on respective hypotheses.

(a) Pessimism: For any $k \in [K]$, the statement holds for $h = H + 1$ since $\widecheck{V}_{k,H+1}(\cdot) = V_{H+1}^*(\cdot) = 0$.

Assume the statement holds for $h + 1$, which means $\widecheck{V}_{k,h+1}(\cdot) \leqslant V_{h+1}^*(\cdot)$ under $\widecheck{\Psi}_{k,h+1}$. Since $\widecheck{Q}_{k,h}(\cdot, \cdot) = r_h(\cdot, \cdot) + \langle \widehat{\boldsymbol{\mu}}_{k,h} \widecheck{V}_{k,h+1}, \boldsymbol{\phi}(\cdot, \cdot) \rangle - \widecheck{\beta} \| \boldsymbol{\phi}(\cdot, \cdot) \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$, for $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, we have:

$$
\begin{aligned}
&Q_h^*(s,a) - \widecheck{Q}_{k,h}(s,a) \\
=&r_h(s,a) + \mathbb{P}_h V_{h+1}^*(s,a) - \left[r_h(s,a) + \langle \widehat{\boldsymbol{\mu}}_{k,h} \widecheck{V}_{k,h+1}, \boldsymbol{\phi}(s,a) \rangle - \widecheck{\beta} \| \boldsymbol{\phi}(s,a) \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \right] \\
=&\langle \boldsymbol{\mu}_h \widecheck{V}_{k,h+1}, \boldsymbol{\phi}(s,a) \rangle - \langle \widehat{\boldsymbol{\mu}}_{k,h} \widecheck{V}_{k,h+1}, \boldsymbol{\phi}(s,a) \rangle + \widecheck{\beta} \| \boldsymbol{\phi}(s,a) \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \mathbb{P}_h V_{h+1}^*(s,a) - \mathbb{P}_h \widecheck{V}_{k,h+1}(s,a) \\
\geqslant& -\| (\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widecheck{V}_{k,h+1} \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \| \boldsymbol{\phi}(s,a) \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \widecheck{\beta} \| \boldsymbol{\phi}(s,a) \|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \mathbb{P}_h V_{h+1}^*(s,a) - \mathbb{P}_h \widecheck{V}_{k,h+1}(s,a) \\
\geqslant& \mathbb{P}_h V_{h+1}^*(s,a) - \mathbb{P}_h \widecheck{V}_{k,h+1}(s,a) \\
\geqslant& 0,
\end{aligned}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds by the assumption that $\boldsymbol{\mu}_h \in \widecheck{\mathcal{C}}_{k,h}$ under $\widecheck{\Psi}_{k,h}$, the third inequality holds since the induction assumption $\widecheck{V}_{k,h+1}(\cdot) \leqslant V_{h+1}^*(\cdot)$ under $\widecheck{\Psi}_{k,h+1}$ and $\mathbb{P}_h$ is a valid distribution. Therefore, we have $\widecheck{Q}_{k,h}(s,a) \leqslant Q_h^*(s,a)$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Since $\widecheck{V}_{k,h}(\cdot) = \max \{ \max_{a \in \mathcal{A}} \widecheck{Q}_{k,h}(\cdot, a), 0 \}$, for any $s \in \mathcal{S}$, we have the following two cases:

- If $\max_{a \in \mathcal{A}} \widecheck{Q}_{k,h}(s,a) \leqslant 0$, we have $0 = \widecheck{V}_{k,h}(s) \leqslant V_h^*(s)$.

- Otherwise, $\widecheck{V}_{k,h}(s) = \max_{a \in \mathcal{A}} \widecheck{Q}_{k,h}(s,a) \leqslant \max_{a \in \mathcal{A}} Q_h^*(s,a) = V_h^*(s)$.

Therefore, we have $\widecheck{V}_{k,h}(\cdot) \leqslant V_h^*(\cdot)$ under $\widecheck{\Psi}_{k,h}$ for any $(k,h) \in [K] \times [H]$.

(b): Optimism: We first prove the optimism for some fixed episode $k \in [K]$ by induction.

For any $k \in [K]$, the statement holds for $h = H + 1$ since $\widehat{V}_{k,H+1}(\cdot) = V^*_{H+1}(\cdot) = 0$. Assume the statement holds for $h + 1$, which means $\widehat{V}_{k,h+1}(\cdot) \geqslant V^*_{h+1}(\cdot)$ under $\widehat{\Psi}_{k,h+1}$ for any $k \in [K]$. For any $k \in [K]$ and any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and, we have

$$
\begin{aligned}
&r_h(\cdot, \cdot) + \langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(\cdot, \cdot) \rangle + \widehat{\beta} \|\boldsymbol{\phi}(\cdot, \cdot)\|_{\widehat{\boldsymbol{\Lambda}}^{-1}_{k,h}} - Q^*_h(s, a) \\
={}& r_h(s, a) + \langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s, a) \rangle + \widehat{\beta} \|\boldsymbol{\phi}(s, a)\|_{\widehat{\boldsymbol{\Lambda}}^{-1}_{k,h}} - \left[r_h(s, a) + \mathbb{P}_h V^*_{h+1}(s, a) \right] \\
={}& \langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s, a) \rangle - \langle \boldsymbol{\mu}_h \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s, a) \rangle + \widehat{\beta} \|\boldsymbol{\phi}(s, a)\|_{\widehat{\boldsymbol{\Lambda}}^{-1}_{k,h}} + \mathbb{P}_h \widehat{V}_{k,h+1}(s, a) - \mathbb{P}_h V^*_{h+1}(s, a) \\
\geqslant{}& -\|(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h) \widehat{\boldsymbol{V}}_{k,h+1}\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \|\boldsymbol{\phi}(s, a)\|_{\widehat{\boldsymbol{\Lambda}}^{-1}_{k,h}} + \widehat{\beta} \|\boldsymbol{\phi}(s, a)\|_{\widehat{\boldsymbol{\Lambda}}^{-1}_{k,h}} + \mathbb{P}_h \widehat{V}_{k,h+1}(s, a) - \mathbb{P}_h V^*_{h+1}(s, a) \\
\geqslant{}& \mathbb{P}_h \widehat{V}_{k,h+1}(s, a) - \mathbb{P}_h V^*_{h+1}(s, a) \\
\geqslant{}& 0,
\end{aligned}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds by the assumption that $\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k,h}$ under $\widehat{\Psi}_{k,h}$, the third inequality holds by the induction assumption $\widehat{V}_{k,h+1}(\cdot) \geqslant V^*_{h+1}(\cdot)$ under $\widehat{\Psi}_{k,h+1}$ and $\mathbb{P}_h$ is a valid distribution. Then we have

$$
Q^*_h(s, a) \leqslant \min_{1 \leqslant i \leqslant k} \min \left\{ r_h(s, a) + \langle \widehat{\boldsymbol{\mu}}_{i,h} \widehat{\boldsymbol{V}}_{i,h+1}, \boldsymbol{\phi}(s, a) \rangle + \widehat{\beta} \|\boldsymbol{\phi}(s, a)\|_{\widehat{\boldsymbol{\Lambda}}^{-1}_{i,h}}, H \right\} \leqslant \widehat{Q}_{k,h}(s, a)
$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, which further implies $\widehat{V}_{k,h}(s) = \max_{a \in \mathcal{A}} \widehat{Q}_{k,h} \geqslant \max_{a \in \mathcal{A}} Q^*_h(s, a)$ for any $s \in \mathcal{S}$.

Therefore, we have $\widehat{V}_{k,h}(\cdot) \geqslant V^*_h(\cdot)$ under $\widehat{\Psi}_{k,h}$ for any $(k, h) \in [K] \times [H]$. $\qquad \square$

D.2. Suboptimality Gap

In this subsection, we establish Lemma D.4 and Lemma D.5 that bound the distance of the optimistic value function $\widehat{V}_{k,h}$ to the value function $V^{\pi^k}_h$ associated with policy π_k and the pessimistic value function $\widecheck{V}_{k,h}$, respectively. Before that, we present two high probability events Ξ_1 and Ξ_2 in Lemma D.2 and D.3, respectively.

Lemma D.2. *In Algorithm 1, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, simultaneously for all $h' \in [H]$, we have*

$$
\sum_{k=1}^K \sum_{h=h'}^H \left[[\mathbb{P}_h (\widehat{V}_{k,h+1} - V^{\pi^k}_{h+1})](s^k_h, a^k_h) - [\widehat{V}_{k,h+1} - V^{\pi^k}_{h+1}](s^k_{h+1}) \right] \leqslant 2H \sqrt{2T \log \left(\frac{H}{\delta} \right)}
$$

Proof. Denote $\Delta_{k,h} = [\mathbb{P}_h (\widehat{V}_{k,h+1} - V^{\pi^k}_{h+1})](s^k_h, a^k_h) - [\widehat{V}_{k,h+1} - V^{\pi^k}_{h+1}](s^k_{h+1})$. Since s^k_{h+1} is $\mathcal{F}_{k,h+1}$-measurable, $\Delta_{k,h}$ is $\mathcal{F}_{k,h+1}$-measurable and $\mathbb{E}[\Delta_{k,h} \mid \mathcal{F}_{k,h}] = 0$. Thus, for some $h' \in [H]$, $\{\Delta_{k,h'}, \Delta_{k,h'+1}, ..., \Delta_{k,H}\}_{k \in [K]}$ is a martingale difference sequence. Since $|\Delta_{k,h}| \leqslant 2H$ by $-H \leqslant \widehat{V}_{k,h+1}(\cdot) - V^{\pi^k}_{h+1}(\cdot) \leqslant H$, we can apply Azuma-Hoeffding inequality (Lemma F.1) to this martingale difference sequence and obtain

$$
\sum_{k=1}^K \sum_{h=h'}^H \Delta_{k,h} \leqslant 2H \sqrt{2K(H - h' + 1)) \log(1/\delta)} \leqslant 2H \sqrt{2T \log(1/\delta)}, \tag{32}
$$

for some $h' \in [H]$ with probability at least $1 - \delta$. Taking a union bound over all $h' \in [H]$ gives the final conclusion. $\qquad \square$

Lemma D.3. *In Algorithm 1, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, simultaneously for all $h' \in [H]$, we have*

$$
\sum_{k=1}^K \sum_{h=h'}^H \left[[\mathbb{P}_h (\widehat{V}_{k,h+1} - \widecheck{V}_{k,h+1})](s^k_h, a^k_h) - [\widehat{V}_{k,h+1} - \widecheck{V}_{k,h+1}](s^k_{h+1}) \right] \leqslant 2H \sqrt{2T \log \left(\frac{H}{\delta} \right)}
$$

Proof. The proof is almost the same as that of Lemma D.2, except for replacing $V^{\pi^k}_{k,h+1}(\cdot)$ by $\widecheck{V}_{k,h+1}(\cdot)$. $\qquad \square$

Lemma D.4. *In Algorithm 1, under $\widehat{\Psi}_1 \cap \Xi_1$, we have*

$$
\sum_{k=1}^{K} \widehat{V}_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k) \leqslant 4\widehat{\beta}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{\sigma}_{k,h}^2}\sqrt{H \cdot 2d \log\left(1 + \frac{K}{Hd\lambda}\right)}
$$
$$
+ \frac{3H^2 d}{\log(2)} \log\left(1 + \frac{1}{\lambda H \log(2)}\right) + 2H\sqrt{2T \log\left(\frac{H}{\delta}\right)} \tag{33}
$$

$$
\sum_{k=1}^{K}\sum_{h=1}^{H} [\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) \leqslant 4\widehat{\beta}H\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{\sigma}_{k,h}^2}\sqrt{H \cdot 2d \log\left(1 + \frac{K}{Hd\lambda}\right)}
$$
$$
+ \frac{3H^3 d}{\log(2)} \log\left(1 + \frac{1}{\lambda \log(2)}\right) + 4H^2\sqrt{2T \log\left(\frac{H}{\delta}\right)} \tag{34}
$$

Proof. By Algorithm 1, for any $k \in [K], h \in [H]$, we have

$$
\widehat{Q}_{k,1}(s,a) \leqslant r_1(s,a) + \langle \widehat{\boldsymbol{\mu}}_{k_0,1}\widehat{\boldsymbol{V}}_{k_0,2}, \boldsymbol{\phi}(s,a)\rangle + \widehat{\beta}\|\boldsymbol{\phi}(s,a)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}^{-1}}
$$
$$
= r_1(s,a) + \widehat{\mathbb{P}}_{k_0,1}\widehat{V}_{k_0,2}(s,a) + \widehat{\beta}\|\boldsymbol{\phi}(s,a)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}^{-1}}, \tag{35}
$$
$$
Q_1^{\pi^k}(s,a) = r_1(s,a) + \mathbb{P}_1 V_2^{\pi^k}(s,a).
$$

Then,

$$
\widehat{V}_{k,1}(s_1^k) - V_h^{\pi^k}(s_1^k) = \widehat{Q}_{k,1}(s_1^k, a_1^k) - Q_1^{\pi^k}(s_1^k, a_1^k)
$$
$$
\leqslant r_1(s_1^k, a_1^k) + \widehat{\mathbb{P}}_{k_0,1}\widehat{V}_{k_0,2}(s_1^k, a_1^k) + \widehat{\beta}\|\boldsymbol{\phi}(s_1^k, a_1^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}^{-1}} - r_1(s_1^k, a_1^k) - \mathbb{P}_1 V_2^{\pi^k}(s_1^k, a_1^k)
$$
$$
= \widehat{\beta}\|\boldsymbol{\phi}(s_1^k, a_1^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}^{-1}} + [\widehat{\mathbb{P}}_{k_0,1}\widehat{V}_{k_0,2}(s_1^k, a_1^k) - \mathbb{P}_1 \widehat{V}_{k_0,2}(s_1^k, a_1^k)] + [\mathbb{P}_1(\widehat{V}_{k,2} - V_2^{\pi^k})](s_1^k, a_1^k)
$$
$$
= \widehat{\beta}\|\boldsymbol{\phi}(s_1^k, a_1^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}^{-1}} + \langle (\widehat{\boldsymbol{\mu}}_{k_0,1} - \boldsymbol{\mu}_1)\widehat{\boldsymbol{V}}_{k_0,2}, \boldsymbol{\phi}(s_1^k, a_1^k)\rangle + [\mathbb{P}_1(\widehat{V}_{k,2} - V_2^{\pi^k})](s_1^k, a_1^k)
$$
$$
\leqslant \widehat{\beta}\|\boldsymbol{\phi}(s_1^k, a_1^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}^{-1}} + \|(\widehat{\boldsymbol{\mu}}_{k_0,1} - \boldsymbol{\mu}_h)\widehat{\boldsymbol{V}}_{k_0,2}\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}}\|\boldsymbol{\phi}(s_1^k, a_1^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}^{-1}} + [\mathbb{P}_1(\widehat{V}_{k,2} - V_2^{\pi^k})](s_1^k, a_1^k) \tag{36}
$$
$$
\leqslant 2\widehat{\beta}\|\boldsymbol{\phi}(s_1^k, a_1^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,1}^{-1}} + [\mathbb{P}_1(\widehat{V}_{k,2} - V_2^{\pi^k})](s_1^k, a_1^k)
$$
$$
\leqslant 4\widehat{\beta}\|\boldsymbol{\phi}(s_1^k, a_1^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,1}^{-1}} + [\mathbb{P}_1(\widehat{V}_{k,2} - V_2^{\pi^k})](s_1^k, a_1^k)
$$
$$
= 4\widehat{\beta}\|\boldsymbol{\phi}(s_1^k, a_1^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,1}^{-1}} + [\widehat{V}_{k,2} - V_2^{\pi^k}](s_2^k) + [\mathbb{P}_1(\widehat{V}_{k,2} - V_2^{\pi^k})](s_1^k, a_1^k) - [\widehat{V}_{k,2} - V_2^{\pi^k}](s_2^k)
$$
$$
\cdots
$$
$$
\leqslant \sum_{h=1}^{H} 4\widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \sum_{h=1}^{H}\left[[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\right]
$$

where the first inequality is due to Eq. (35), the first equality holds since $\widehat{V}_{k,2} = \widehat{V}_{k_0,2}$ in Algorithm 1, the second inequality holds by the Cauchy-Schwarz inequality, third inequality holds since $\boldsymbol{\mu}_1 \in \widehat{\mathcal{C}}_{k_0,1}$ under $\widehat{\Psi}_1$, the fourth inequality holds due to Lemma F.7 with the updating rule in Line 8 of Algorithm 1, and the last inequality holds since we can expand $\widehat{V}_{k,h+1}(s_h^k) - \check{V}_{k,h+1}(s_h^k)$ in a recursive way until stage H.

Summing up Eq.(36) for $k \in [K]$ gives

$$
\sum_{k=1}^{K} \widehat{V}_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)
$$
$$
\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H} 4\widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,1}^{-1}} + \sum_{k=1}^{K}\sum_{h=1}^{H}\left[[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\right]
$$

$$\leqslant 4\widehat{\beta} \sum_{k=1}^{K} \sum_{h=1}^{H} \underbrace{\widehat{\sigma}_{k,h} \|\widehat{\sigma}_{k,h}^{-1} \phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}}_{I} + 2H\sqrt{2T \log\left(\frac{H}{\delta}\right)},$$

where the second inequality holds under Ξ_1 by Lemma D.2. Subsequently, we try to bound I. For fixed $h \in [H]$, denote $x_k = \widehat{\sigma}_{k,h}^{-1} \phi(s_h^k, a_h^k)$ in Lemma D.7. Then, there are at most $3d \log[1 + d/(\lambda H \log(2))]/\log(2)$ episodes that $\|\widehat{\sigma}_{k,h}^{-1} \phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}} \geqslant 1$, which further implies that there are at most $H \cdot 3d \log[1 + d/(\lambda H \log(2))]/\log(2)$ episodes that there exists $h' \in [H]$ such that $\|\widehat{\sigma}_{k,h'}^{-1} \phi(s_{h'}^k, a_{h'}^k)\|_{\widehat{\Lambda}_{k,h'}^{-1}} \geqslant 1$. Moreover, we can bound $\widehat{V}_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)$ in these episodes by H since $\widehat{V}_{k,1}(\cdot) - V_1^{\pi^k}(\cdot) \leqslant H$ for any $k \in [K]$. Thus, we have

$$\sum_{k=1}^{K} \widehat{V}_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)$$

$$\leqslant 4\widehat{\beta} \sum_{k=1}^{K} \sum_{h=1}^{H} \widehat{\sigma}_{k,h} \min\left\{\|\widehat{\sigma}_{k,h}^{-1} \phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}, 1\right\} + H \cdot H \cdot \frac{3d}{\log(2)} \log\left(1 + \frac{1}{\lambda H \log(2)}\right) + 2H\sqrt{2T \log\left(\frac{H}{\delta}\right)}$$

$$\leqslant 4\widehat{\beta} \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \widehat{\sigma}_{k,h}^2} \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \min\left\{\|\widehat{\sigma}_{k,h}^{-1} \phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}^2, 1\right\}} + \frac{3H^2 d}{\log(2)} \log\left(1 + \frac{1}{\lambda H \log(2)}\right) + 2H\sqrt{2T \log\left(\frac{H}{\delta}\right)}$$

$$\leqslant 4\widehat{\beta} \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \widehat{\sigma}_{k,h}^2} \sqrt{H \cdot 2d \log\left(1 + \frac{K}{Hd\lambda}\right)} + \frac{3H^2 d}{\log(2)} \log\left(1 + \frac{L^2}{\lambda H \log(2)}\right) + 2H\sqrt{2T \log\left(\frac{H}{\delta}\right)},$$

(37)

where the second inequality holds by Cauchy-Schwarz inequality, and the last inequality hols by Lemma F.5 with the fact that $\|\widehat{\sigma}_{k,h}^{-1} \phi(s_h^k, a_h^k)\|_2 \leqslant 1/\sqrt{H}$. Thus, Eq. (33) is obtained. Besides, by similar argument in Eq. (36), we obtain

$$\widehat{V}_{k,h'}(s_{h'}^k) - V_{h'}^{\pi^k}(s_{h'}^k) \leqslant \sum_{h=h'}^{H} 4\widehat{\beta} \|\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}} + \sum_{h=h'}^{H} \left[[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\right]$$

which further gives

$$\sum_{k=1}^{K} \widehat{V}_{k,h'}(s_{h'}^k) - V_{h'}^{\pi^k}(s_{h'}^k)$$

$$\leqslant \sum_{k=1}^{K} \sum_{h=h'}^{H} 4\widehat{\beta} \|\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}} + \sum_{k=1}^{K} \sum_{h=h'}^{H} \left[[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\right]$$

$$\leqslant 4\widehat{\beta} \sum_{k=1}^{K} \sum_{h=1}^{H} \widehat{\sigma}_{k,h} \|\widehat{\sigma}_{k,h}^{-1} \phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}} + \sum_{k=1}^{K} \sum_{h=h'}^{H} \left[[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\right]$$

(38)

$$\leqslant 4\widehat{\beta} \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \widehat{\sigma}_{k,h}^2} \sqrt{H \cdot 2d \log\left(1 + \frac{K}{Hd\lambda}\right)} + \frac{3H^2 d}{\log(2)} \log\left(1 + \frac{1}{\lambda \log(2)}\right) + 2H\sqrt{2T \log\left(\frac{H}{\delta}\right)},$$

where the last inequality holds by similar argument of bounding I in Eq. (37) and Lemma D.2 under Ξ_1.

Also note that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k)$$

$$= \sum_{k=1}^{K} \sum_{h=2}^{H} [\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k) + \sum_{k=1}^{K} \sum_{h=1}^{H} \left[[\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\right]$$

$$\leqslant 4\widehat{\beta}H\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^{2}}\sqrt{H \cdot 2d\log\left(1+\frac{K}{Hd\lambda}\right)} + \frac{3H^{3}d}{\log(2)}\log\left(1+\frac{1}{\lambda\log(2)}\right)$$

$$+ 2H^{2}\sqrt{2T\log\left(\frac{H}{\delta}\right)} + 2H\sqrt{2T\log\left(\frac{H}{\delta}\right)}$$

$$\leqslant 4\widehat{\beta}H\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^{2}}\sqrt{H \cdot 2d\log\left(1+\frac{K}{Hd\lambda}\right)} + \frac{3H^{3}d}{\log(2)}\log\left(1+\frac{1}{\lambda\log(2)}\right) + 4H^{2}\sqrt{2T\log\left(\frac{H}{\delta}\right)},$$

where the first inequality holds by sum up Eq. (38) for $h' = 2, ..., H$ and Lemma D.2 under Ξ_1. Thus, Eq. (34) is also obtained. $\qquad\square$

Lemma D.5 (Gap between Optimism and Pessimism). *In Algorithm 1, under $\widehat{\Psi}_1 \cap \check{\Psi}_1 \cap \Xi_2$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) \leqslant 2(2\widehat{\beta} + \check{\beta})H\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^{2}}\sqrt{H \cdot 2d\log\left(1+\frac{K}{Hd\lambda}\right)} + 4H^{2}\sqrt{2T\log\left(\frac{H}{\delta}\right)}$$

Proof. By Algorithm 1, for any $k \in [K], h \in [H]$, we have

$$\begin{aligned}
\widehat{Q}_{k,h}(s,a) &\leqslant r_h(s,a) + \langle\widehat{\boldsymbol{\mu}}_{k_0,h}\widehat{\boldsymbol{V}}_{k_0,h+1}, \boldsymbol{\phi}(s,a)\rangle + \widehat{\beta}\|\boldsymbol{\phi}(s,a)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}^{-1}} \\
&= r_h(s,a) + \widehat{\mathbb{P}}_{k_0,h}\widehat{V}_{k_0,h+1}(s,a) + \widehat{\beta}\|\boldsymbol{\phi}(s,a)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}^{-1}}, \\
\check{Q}_{k,h}(s,a) &\geqslant r_h(s,a) + \langle\widehat{\boldsymbol{\mu}}_{k,h}\check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s,a)\rangle - \check{\beta}\|\boldsymbol{\phi}(s,a)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \\
&= r_h(s,a) + \widehat{\mathbb{P}}_{k,h}\check{V}_{k,h+1}(s,a) - \check{\beta}\|\boldsymbol{\phi}(s,a)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}},
\end{aligned} \tag{39}$$

Then,

$$\begin{aligned}
&\widehat{V}_{k,h}(s_h^k) - \check{V}_{k,h}(s_h^k) \leqslant \widehat{Q}_{k,h}(s_h^k, a_h^k) - \check{Q}_{k,h}(s_h^k, a_h^k) \\
&\leqslant r_h(s_h^k, a_h^k) + \widehat{\mathbb{P}}_{k_0,h}\widehat{V}_{k_0,h+1}(s_h^k, a_h^k) + \widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}^{-1}} - [r_h(s_h^k, a_h^k) + \widehat{\mathbb{P}}_{k,h}\check{V}_{k,h+1}(s_h^k, a_h^k) - \check{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}] \\
&= \widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}^{-1}} + \check{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + [\widehat{\mathbb{P}}_{k_0,h}\widehat{V}_{k_0,h+1}(s_h^k, a_h^k) - \mathbb{P}_h\widehat{V}_{k_0,h+1}(s_h^k, a_h^k)] \\
&\quad + [\mathbb{P}_h\check{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h}\check{V}_{k,h+1}(s_h^k, a_h^k)] + [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) \\
&= \widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}^{-1}} + \check{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \langle(\widehat{\boldsymbol{\mu}}_{k_0,h+1} - \boldsymbol{\mu}_h)\widehat{\boldsymbol{V}}_{k_0,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle \\
&\quad + \langle(\widehat{\boldsymbol{\mu}}_{k,h+1} - \boldsymbol{\mu}_h)\check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle + [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) \\
&\leqslant \widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}^{-1}} + \check{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \|(\widehat{\boldsymbol{\mu}}_{k_0,h} - \boldsymbol{\mu}_h)\widehat{\boldsymbol{V}}_{k_0,h+1}\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}^{-1}} \\
&\quad + \|(\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h)\check{\boldsymbol{V}}_{k,h+1}\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,1}^{-1}} + [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) \\
&\leqslant 2\widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k_0,h}^{-1}} + 2\check{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) \\
&\leqslant 4\widehat{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + 2\check{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) \\
&= (4\widehat{\beta} + 2\check{\beta})\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + [\widehat{V}_{k,h+1} - \check{V}_{k,h+1}](s_{h+1}^k) + [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - \check{V}_{k,h+1}](s_{h+1}^k)
\end{aligned} \tag{40}$$

where the first inequality holds since $\check{V}_{k,h+1}(s_h^k) \geqslant \check{Q}_{k,h}(s_h^k, a_h^k)$, the second inequality holds due to Eq. (39), the first equality holds since $\widehat{V}_{k,h} = \widehat{V}_{k_0,h}$, the third inequality holds by the Cauchy-Schwarz inequality, the fourth inequality holds since $\boldsymbol{\mu}_h \in \widehat{\mathcal{C}}_{k_0,h} \cap \check{\mathcal{C}}_{k,h}$ under $\widehat{\Psi}_1 \cap \check{\Psi}_1$, and the last inequality holds due to Lemma F.7 with the updating rule in Line 8 of Algorithm 1.

Since $\widehat{V}_{k,h'}(s_{h'}^k) - \check{V}_{k,h'}(s_{h'}^k) \leqslant H$, we further obtains

$$\widehat{V}_{k,h}(s_h^k) - \check{V}_{k,h}(s_h^k) \leqslant \min\left\{\widehat{Q}_{k,h}(s_h^k, a_h^k) - \check{Q}_{k,h}(s_h^k, a_h^k), H\right\}$$

$$\leqslant \min\left\{(4\widehat{\beta} + 2\check{\beta})\|\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}, H\right\} + [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k)$$

$$\leqslant (4\widehat{\beta} + 2\check{\beta})\widehat{\sigma}_{k,h}\min\left\{\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}, 1\right\} + [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) \tag{41}$$

$$= (4\widehat{\beta} + 2\check{\beta})\widehat{\sigma}_{k,h}\min\left\{\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}, 1\right\} + [\widehat{V}_{k,h+1} - \check{V}_{k,h+1}](s_{h+1}^k)$$

$$+ [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - \check{V}_{k,h+1}](s_{h+1}^k),$$

where the second inequality holds since $\check{\beta}\widehat{\sigma}_{k,h} \geqslant \sqrt{Hd^2}\sqrt{H} \geqslant H$. Summing up Eq. (41) for $k \in [K]$ and $h = h', ..., H$ gives

$$\sum_{k=1}^{K} \widehat{V}_{k,h'}(s_{h'}^k) - \check{V}_{k,h'}(s_{h'}^k)$$

$$\leqslant \sum_{k=1}^{K}\sum_{h=h'}^{H}(4\widehat{\beta} + 2\check{\beta})\widehat{\sigma}_{k,h}\min\left\{\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}, 1\right\}$$

$$+ \sum_{k=1}^{K}\sum_{h=h'}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - \check{V}_{k,h+1}](s_{h+1}^k)$$

$$\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}(4\widehat{\beta} + 2\check{\beta})\widehat{\sigma}_{k,h}\min\left\{\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}, 1\right\} + 2H\sqrt{2T\log\left(\frac{H}{\delta}\right)} \tag{42}$$

$$\leqslant (4\widehat{\beta} + 2\check{\beta})\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_{\widehat{\Lambda}_{k,h}^{-1}}^2, 1\right\}} + 2H\sqrt{2T\log\left(\frac{H}{\delta}\right)}$$

$$\leqslant (4\widehat{\beta} + 2\check{\beta})\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{H \cdot 2d\log\left(1 + \frac{K}{Hd\lambda}\right)} + 2H\sqrt{2T\log\left(\frac{H}{\delta}\right)}$$

where the second inequality holds by Lemma D.3 under Ξ_2, the third inequality holds due to Cauchy-Schwarz inequality, and the last inequality hols due to Lemma F.5 with the fact that $\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k, a_h^k)\|_2 \leqslant 1/\sqrt{H}$.

Also note that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k)$$

$$= \sum_{k=1}^{K}\sum_{h=2}^{H}[\widehat{V}_{k,h+1} - \check{V}_{k,h+1}](s_{h+1}^k) + \sum_{k=1}^{K}\sum_{h=1}^{H}\left[[\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) - [\widehat{V}_{k,h+1} - \check{V}_{k,h+1}](s_{h+1}^k)\right]$$

$$\leqslant (4\widehat{\beta} + 2\check{\beta})H\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{H \cdot 2d\log\left(1 + \frac{K}{Hd\lambda}\right)} + 2H^2\sqrt{2T\log\left(\frac{H}{\delta}\right)} + 2H\sqrt{2T\log\left(\frac{H}{\delta}\right)}$$

$$\leqslant 2(2\widehat{\beta} + \check{\beta})H\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{H \cdot 2d\log\left(1 + \frac{K}{Hd\lambda}\right)} + 4H^2\sqrt{2T\log\left(\frac{H}{\delta}\right)},$$

where the first inequality holds by sum up Eq. (42) for $h' = 2, ..., H$ and Lemma D.3 under Ξ_2. $\qquad\square$

D.3. Summation of Estimated Variances

In this subsection, we try to bound the summation of estimated variance in Lemma D.9. As shown in Lemma D.9, the summation of estimated variance $\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2 = \widetilde{O}(HT)$, which utilizes the Law of Total Variance in (Lattimore & Hutter, 2012), detailed in Lemma D.6.

Lemma D.6 (Total variance lemma, Lemma C.5 in (Jin et al., 2018)). *With probability at least $1 - \delta$, we have*

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left[\mathbb{V}_h V_{h+1}^{\pi^k} \right] (s_h^k, a_h^k) \leqslant 3 \left(HT + H^3 \log(1/\delta) \right).$$

By the definition of $\hat{\sigma}_{k,h}$, the summation of $\varsigma_{k,h}^2$ will influence the summation of $\hat{\sigma}_{k,h}^2$. We need to make $\sum_{k=1}^{K} \sum_{h=1}^{H} \varsigma_{k,h}^2$ small such that it will not become dominant term in the upper bound. However, enlarging $\varsigma_{k,h}$ is required in some stages of some episodes, as stated in Remark 7.4 in the main paper. To address this dilemma, we build the following critical lemma which characterizes the conservatism of the elliptical potential, i.e., $\|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}$ is small in most episodes, as detailed in Lemma D.7.

Lemma D.7 (Elliptical Potentials: You cannot have many big intervals). *Given $\lambda > 0$ and sequence $\{\boldsymbol{x}_t\}_{t=1}^{T} \subset \mathbb{R}^d$ with $\|\boldsymbol{x}_t\|_2 \leqslant L$ for all $t \in [T]$, define $\mathbf{Z}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ for $t \geqslant 1$ and $\mathbf{Z}_0 = \lambda \mathbf{I}$. During $[T]$, the number of times $\|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}} \geqslant c$ is at most*

$$\frac{3d}{\log(1 + c^2)} \log \left(1 + \frac{L^2}{\lambda \log(1 + c^2)} \right),$$

where $c > 0$ is a constant.

Proof. The proof of this lemma is firstly proposed at Exercise 19.3 in (Lattimore & Szepesvári, 2020) for the case of $C = 1$, i.e., Lemma F.6, we generalize it to the case with any positive constant C.

Let $\mathcal{T}$ be the set of rounds t when $\|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}^2 \geqslant C$ for $t \in [T]$ and $\mathbf{Y}_t = \mathbf{Z}_0 + \sum_{i=1}^{t} \mathbb{1}\{i \in \mathcal{T}\} \boldsymbol{x}_i \boldsymbol{x}_i^\top$. Then

$$\left(\frac{d\lambda + |\mathcal{T}|L^2}{d} \right)^d \geqslant \left(\frac{\text{trace}\,(\mathbf{Y}_T)}{d} \right)^d$$

$$\geqslant \det\,(\mathbf{Y}_T)$$

$$= \det\,(\mathbf{Z}_0) \prod_{t \in \mathcal{T}} \left(1 + \|\boldsymbol{x}_t\|_{\mathbf{Y}_{t-1}^{-1}}^2 \right)$$

$$\geqslant \det\,(\mathbf{Z}_0) \prod_{t \in \mathcal{T}} \left(1 + \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}^2 \right)$$

$$\geqslant \lambda^d (1 + c^2)^{|\mathcal{T}|}$$

Rearranging and taking the logarithm show that

$$|\mathcal{T}| \leqslant \frac{d}{\log(1 + c^2)} \log \left(1 + \frac{|\mathcal{T}|L^2}{d\lambda} \right).$$

Abbreviate $x = d/\log(1 + c^2)$ and $y = L^2/d\lambda$, which are both positive. Then

$$x \log(1 + y(3x \log(1 + xy))) \leqslant x \log \left(1 + 3x^2 y^2 \right) \leqslant x \log (1 + xy)^3 = 3x \log(1 + xy).$$

Define $f(z) = z - x \log(1 + yz)$ for $z \geqslant 0$, we have $f'(z) = [1 + y(z - x)]/(1 + yz)$, which implies $f(z)$ is increasing if $1 - xy \geqslant 0$, or $f(z)$ is first decreasing then increasing, otherwise. Since $f(0) = 0$ and $f(3x \log(1 + xy)) \geqslant 0$, if $f(z) \leqslant 0$, we must have $z \leqslant 3x \log(1 + xy)$. In other words, $f(z)$ is increasing for $z \geqslant 3x \log(1 + xy)$. It then follows that

$$|\mathcal{T}| \leqslant 3x \log(1 + xy) = \frac{3d}{\log(1 + c^2)} \log \left(1 + \frac{L^2}{\lambda \log(1 + c^2)} \right).$$

$\square$

The following Lemma is required to upper bounds the summation of offset term $U_{k,h}$.

Lemma D.8. *In Algorithm 1, under Υ, for any $k \in [K]$ and any $h \in [H]$, we have*

$$\left|\langle\widehat{\boldsymbol{\mu}}_{k,h}(\widehat{\boldsymbol{V}}_{k,h+1} - \check{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right| \leqslant \quad \bar{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})(s_h^k, a_h^k) + \check{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

Proof.

$$\left|\langle\widehat{\boldsymbol{\mu}}_{k,h}(\widehat{\boldsymbol{V}}_{k,h+1} - \check{\boldsymbol{V}}_{k,h+1}), \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right| = \left|\widehat{\mathbb{P}}_{k,h}\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h}\check{V}_{k,h+1}(s_h^k, a_h^k)\right|$$

$$\leqslant \left|\widehat{\mathbb{P}}_{k,h}\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k)\right| + \left|\mathbb{P}_h\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \mathbb{P}_h\check{V}_{k,h+1}(s_h^k, a_h^k)\right|$$

$$+ \left|\mathbb{P}_h\check{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{\mathbb{P}}_{k,h}\check{V}_{k,h+1}(s_h^k, a_h^k)\right|$$

$$= \left|\langle(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h})\widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right| + \left|\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})(s_h^k, a_h^k)\right| + \left|\langle(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h})\check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right|$$

$$= \left|\langle(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h})\widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right| + \mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})(s_h^k, a_h^k) + \left|\langle(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h})\check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k)\rangle\right|$$

$$\leqslant \left\|(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h})\widehat{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})(s_h^k, a_h^k)$$

$$+ \left\|(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_{k,h})\check{\boldsymbol{V}}_{k,h+1}\right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}} \|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}$$

$$\leqslant \bar{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} + \mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})(s_h^k, a_h^k) + \check{\beta}\|\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}},$$

where the first inequality holds due to triangle inequality, the second equality holds since $\widehat{V}_{k,h+1}(\cdot) \geqslant \check{V}_{k,h+1}(\cdot)$ under Υ by Lemma D.1, the second inequality holds due to Cauchy-Schwarz inequality, and the last inequality holds since under Υ we have $\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{k,h} \cap \tilde{\mathcal{C}}_{k,h}$. □

Now we are ready to upper bounds $\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2$ in Lemma D.9 under the high probability event $\Upsilon \cap \Xi_2 \cap \Xi_3$.

Lemma D.9. *In Algorithm 1, under event $\Upsilon \cap \Xi_2 \cap \Xi_3$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2 \leqslant 8HT + \frac{6H^5d^6}{\log(1+c^2)}\log\left(1 + \frac{d}{\lambda H\log(1+c^2)}\right) + 6H^3\log\left(\frac{1}{\delta}\right)$$

$$+ 8H^3(6 + Hd^3)\sqrt{2T\log\left(\frac{H}{\delta}\right)} + \frac{12H^4d}{\log(2)}\log\left(1 + \frac{1}{\lambda\log(2)}\right) + 2H^4d^3\sqrt{\lambda} +$$

$$8Hd\left[H\left(2H(6 + Hd^3)\widehat{\beta} + (4 + Hd^3)\bar{\beta} + (H+1)(4 + Hd^3)\check{\beta}\right) + \tilde{\beta}\right]^2\log\left(1 + \frac{K}{Hd\lambda}\right)$$

where $c = 1/(H^3d^5)$.

Proof. Initially, by definition of $\widehat{\sigma}_{k,h}$ in Algorithm 1, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2 \leqslant \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\varsigma_{k,h}^2}_{I_1} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}Hd^3E_{k,h}}_{I_2} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left[[\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k) + U_{k,h}\right]}_{I_3}, \tag{43}$$

Bounding I_1: Denote $c = 1/(H^3d^5)$. For fixed $h \in [H]$, set $\boldsymbol{x}_k$ as $\widetilde{\sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k, a_h^k)$ in Lemma D.7. Then, there are at most $3d\log[1 + d/(\lambda H\log(1+c^2))]/\log(1+c^2)$ episodes that $\|\widetilde{\sigma}_{k,h}^{-1}\boldsymbol{\phi}(s_h^k, a_h^k)\|_{\widetilde{\boldsymbol{\Lambda}}_{k,h}^{-1}} \geqslant c$ such that $\varsigma_k = H^2\sqrt{d^5}$. Thus, we obtain

$$I_1 \leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}H + H \cdot H^4d^5\frac{3d}{\log(1+c^2)}\log\left(1 + \frac{d}{\lambda H\log(1+c^2)}\right). \tag{44}$$

Bounding I_2: Since $\boldsymbol{\mu}_h \in \bar{\mathcal{C}}_{k,h} \cap \check{\mathcal{C}}_{k,h}$ under Υ, we have

$$
\begin{aligned}
\langle \widehat{\boldsymbol{\mu}}_{k,h} \widehat{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle &= \widehat{\mathbb{P}}_{k,h} \widehat{V}_{k,h+1}(s_h^k, a_h^k) \leqslant \mathbb{P}_h \widehat{V}_{k,h+1}(s_h^k, a_h^k) + \bar{\beta} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}} \\
\langle \widehat{\boldsymbol{\mu}}_{k,h} \check{\boldsymbol{V}}_{k,h+1}, \boldsymbol{\phi}(s_h^k, a_h^k) \rangle &= \widehat{\mathbb{P}}_{k,h} \check{V}_{k,h+1}(s_h^k, a_h^k) \geqslant \mathbb{P}_h \check{V}_{k,h+1}(s_h^k, a_h^k) - \check{\beta} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}
\end{aligned}
\tag{45}
$$

Combing Eq. (45) and the definition of $E_{k,h}$ in Eq. (30) gives

$$
\begin{aligned}
I_2 &\leqslant H^2 d^3 \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) + H^2 d^3 \sum_{k=1}^{K} \sum_{h=1}^{H} \min \left\{ 2(\bar{\beta} + \check{\beta}) \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}, H \right\} + H^4 d^3 \sqrt{\lambda} \\
&\leqslant H^2 d^3 \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{P}_h(\widehat{V}_{k,h+1} - \check{V}_{k,h+1})](s_h^k, a_h^k) + 2(\bar{\beta} + \check{\beta}) H^2 d^3 \sum_{k=1}^{K} \sum_{h=1}^{H} \widehat{\sigma}_{k,h} \min \left\{ \left\| \widehat{\sigma}_{k,h}^{-1} \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}}, 1 \right\} \\
&\quad + H^4 d^3 \sqrt{\lambda}
\end{aligned}
\tag{46}
$$

where the second inequality holds since $\check{\beta} \widehat{\sigma}_{k,h} \geqslant \sqrt{Hd^2}\sqrt{H} \geqslant H$.

Bounding I_3: I_3 can be bounded by

$$
\begin{aligned}
I_3 &\leqslant \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \left[\mathbb{V}_h V_{h+1}^* \right](s_h^k, a_h^k) - \left[\mathbb{V}_h V_{h+1}^{\pi^k} \right](s_h^k, a_h^k)}_{J_1} + 2 \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} U_{k,h}}_{J_2} + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \left[\mathbb{V}_h V_{h+1}^{\pi^k} \right](s_h^k, a_h^k)}_{J_3} \\
&\quad + \underbrace{\sum_{k=1}^{K} \sum_{k=1}^{H} \left[\left[\widehat{\mathbb{V}}_{k,h} \widehat{V}_{k,h+1} \right](s_h^k, a_h^k) - \left[\mathbb{V}_h V_{h+1}^* \right](s_h^k, a_h^k) - U_{k,h} \right]}_{J_4}
\end{aligned}
$$

Bounding J_1

$$
\begin{aligned}
J_1 &= \sum_{k=1}^{K} \sum_{h=1}^{H} \left[\mathbb{P}_h V_{h+1}^{*}{}^2(s_h^k, a_h^k) - \left[\mathbb{P}_h V_{h+1}^*(s_h^k, a_h^k) \right]^2 \right] - \left[\mathbb{P}_h V_{h+1}^{\pi^k}{}^2(s_h^k, a_h^k) - \left[\mathbb{P}_h V_{h+1}^{\pi^k}(s_h^k, a_h^k) \right]^2 \right] \\
&\leqslant \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{P}_h V_{h+1}^{*}{}^2(s_h^k, a_h^k) - \mathbb{P}_h V_{h+1}^{\pi^k}{}^2(s_h^k, a_h^k) \\
&\leqslant 2H \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{P}_h(V_{h+1}^* - V_{h+1}^{\pi^k})](s_h^k, a_h^k) \\
&\leqslant 2H \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{P}_h(\widehat{V}_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k),
\end{aligned}
\tag{47}
$$

where the first inequality holds since $V_{h+1}^*(\cdot) \geqslant V_{h+1}^{\pi^k}$, the second inequality holds since $V_{h+1}^{\pi^k} \leqslant V_{h+1}^*(\cdot) \leqslant H$, and the last inequality holds since $\widehat{V}_{k,h+1}(\cdot) \geqslant V_{h+1}^*(\cdot)$ under Υ by Lemma D.1.

Bounding J_2 By the definition of $U_{k,h}$ in Eq. (29) and Lemma D.8, we have

$$
\begin{aligned}
J_2 &\leqslant \sum_{k=1}^{K}\sum_{h=1}^{H}\left\{2\widetilde{\beta}\left\|\phi(s_h^k,a_h^k)\right\|_{\widehat{\mathbf{\Lambda}}_{k,h}^{-1}},2H^2\right\}+4H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1}-\check{V}_{k,h+1})](s_h^k,a_h^k)\\
&\quad+\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{8H(\bar{\beta}+\check{\beta})\left\|\phi(s_h^k,a_h^k)\right\|_{\widehat{\mathbf{\Lambda}}_{k,h}^{-1}},2H^2\right\}\\
&\leqslant[2\widetilde{\beta}+8H(\bar{\beta}+\check{\beta})]\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}\min\left\{\left\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k,a_h^k)\right\|_{\widehat{\mathbf{\Lambda}}_{k,h}^{-1}},1\right\}+4H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1}-\check{V}_{k,h+1})](s_h^k,a_h^k),
\end{aligned}
$$

$$(48)$$

where the second inequality holds since $\widetilde{\beta}\widehat{\sigma}_{k,h}\geqslant\sqrt{H^3}\cdot\sqrt{H}\geqslant H^2, 8H(\bar{\beta}+\check{\beta})\widehat{\sigma}_{k,h}\geqslant 8H\sqrt{H}\cdot\sqrt{H}\geqslant 2H^2$.

Bounding J_3   Since Ξ_3 holds, we have

$$
J_3\leqslant 3\left[HT+H^3\log\left(\frac{1}{\delta}\right)\right]
$$

$$(49)$$

Bounding J_4   Due to Lemma 7.5, we have

$$
J_4\leqslant 0
$$

$$(50)$$

Putting Together   Initially, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=h'}^{H}\widehat{\sigma}_{k,h}\min\left\{\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k,a_h^k)\|_{\widehat{\mathbf{\Lambda}}_{k,h}^{-1}},1\right\}\\
&\leqslant\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k,a_h^k)\|_{\widehat{\mathbf{\Lambda}}_{k,h}^{-1}}^2,1\right\}}\\
&\leqslant\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{H\cdot 2d\log\left(1+\frac{K}{Hd\lambda}\right)},
\end{aligned}
$$

$$(51)$$

where the first inequality holds due to Cauchy-Schwarz inequality, and the second inequality holds due to Lemma F.5 with the fact that $\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k,a_h^k)\|_2\leqslant 1/\sqrt{H}$.

Subsequently, combining Eq. (46), (47), (48) gives

$$
\begin{aligned}
&I_2+J_1+J_2\\
&\leqslant 2H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1}-V_{h+1}^{\pi^k})](s_h^k,a_h^k)+H(4+Hd^3)\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1}-\check{V}_{k,h+1})](s_h^k,a_h^k)+H^4d^3\sqrt{\lambda}\\
&\quad+2\left[H\left((4+Hd^3)\bar{\beta}+(4+Hd^3)\check{\beta}\right)+\widetilde{\beta}\right]\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}\min\left\{\left\|\widehat{\sigma}_{k,h}^{-1}\phi(s_h^k,a_h^k)\right\|_{\widehat{\mathbf{\Lambda}}_{k,h}^{-1}},1\right\}\\
&\leqslant 2H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1}-V_{h+1}^{\pi^k})](s_h^k,a_h^k)+H(4+Hd^3)\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(\widehat{V}_{k,h+1}-\check{V}_{k,h+1})](s_h^k,a_h^k)+H^4d^3\sqrt{\lambda}\\
&\quad+2\left[H\left((4+Hd^3)\bar{\beta}+(4+Hd^3)\check{\beta}\right)+\widetilde{\beta}\right]\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{H\cdot 2d\log\left(1+\frac{K}{Hd\lambda}\right)}\\
&\leqslant 4H^3(6+Hd^3)\sqrt{2T\log\left(\frac{H}{\delta}\right)}+\frac{6H^4d}{\log(2)}\log\left(1+\frac{1}{\lambda\log(2)}\right)+H^4d^3\sqrt{\lambda}+
\end{aligned}
$$

$$2\left[H\left(2H(6+Hd^3)\widehat{\beta}+(4+Hd^3)\bar{\beta}+(H+1)(4+Hd^3)\breve{\beta}\right)+\widetilde{\beta}\right]\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{H\cdot 2d\log\left(1+\frac{K}{Hd\lambda}\right)}$$

where the second inequality holds due to Eq. 51, and the third inequality holds by Lemma D.4 under $\Upsilon\cap\Xi_1$ and Lemma D.5 under $\Upsilon\cap\Xi_2$. Further considering Eq. (44), (49), (50) gives

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2$$

$$\leqslant 4HT+\frac{3H^5d^6}{\log(1+c^2)}\log\left(1+\frac{d}{\lambda H\log(1+c^2)}\right)+3H^3\log\left(\frac{1}{\delta}\right)$$

$$+4H^3(6+Hd^3)\sqrt{2T\log\left(\frac{H}{\delta}\right)}+\frac{6H^4d}{\log(2)}\log\left(1+\frac{1}{\lambda\log(2)}\right)+H^4d^3\sqrt{\lambda}+$$

$$2\left[H\left(2H(6+Hd^3)\widehat{\beta}+(4+Hd^3)\bar{\beta}+(H+1)(4+Hd^3)\breve{\beta}\right)+\widetilde{\beta}\right]\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{H\cdot 2d\log\left(1+\frac{K}{Hd\lambda}\right)}$$

$$(52)$$

Besides, for any $x,a,b\geqslant 0$, if $x\leqslant a\sqrt{x}+b$, then we have $\sqrt{x}\leqslant\sqrt{b+a^2/4}+\sqrt{a^2/4}\leqslant\sqrt{2(b+a^2/4+a^2/4)}$, i.e., $x\leqslant 2b+a^2$. Thus, Eq. (52) implies the final conclusion. $\qquad\square$

D.4. Proof of Theorem 6.1

Putting these building blocks together, we are finally ready to give high probability upper bound on the regret in this subsection, which is based on high probability event $\Upsilon\cap\Xi_1\cap\Xi_2\cap\Xi_3$.

Proof of Theorem 6.1. By construction, taking a union bound, we have that with probability $1-10\delta$, $\Upsilon\cap\Xi_1\cap\Xi_2\cap\Xi_3$ holds. In the remainder of the proof, assume that we are conditioning on this event. Initially, we have

$$\text{Regret}(K)=\sum_{k=1}^{K}\left[V_1^{\star}(s_1^k)-V_1^{\pi_k}(s_1^k)\right]\leqslant\sum_{k=1}^{K}\left[\widehat{V}_{k,1}(s_1^k)-V_1^{\pi^k}(s_1^k)\right]$$

$$\leqslant 4\widehat{\beta}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2}\sqrt{H\cdot 2d\log\left(1+\frac{K}{Hd\lambda}\right)}+\frac{3H^2d}{\log(2)}\log\left(1+\frac{1}{\lambda H\log(2)}\right)+2H\sqrt{2T\log\left(\frac{H}{\delta}\right)},$$

$$(53)$$

where the first inequality holds since $\widehat{V}_{k,1}(\cdot)\geqslant V_1^*(\cdot)$ under Υ by Lemma D.1, and the second inequality holds due to Lemma D.4.

Then, by Lemma D.9, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\sigma}_{k,h}^2\leqslant 8HT+\frac{6H^5d^6}{\log(1+c^2)}\log\left(1+\frac{d}{\lambda H\log(1+c^2)}\right)+6H^3\log\left(\frac{1}{\delta}\right)$$

$$+8H^3(6+Hd^3)\sqrt{2T\log\left(\frac{H}{\delta}\right)}+\frac{12H^4d}{\log(2)}\log\left(1+\frac{1}{\lambda\log(2)}\right)+2H^4d^3\sqrt{\lambda}+$$

$$8Hd\left[H\left(2H(6+Hd^3)\widehat{\beta}+(4+Hd^3)\bar{\beta}+(H+1)(4+Hd^3)\breve{\beta}\right)+\widetilde{\beta}\right]^2\log\left(1+\frac{K}{Hd\lambda}\right)$$

under $\Upsilon\cap\Xi_1\cap\Xi_2$, where $c=1/(H^3d^5)$.

On the one hand, since $1/\log(1 + 1/x) \leqslant 2x$ for any $x \geqslant 1$, we have

$$\frac{1}{\log(1 + c^2)} \leqslant \frac{1}{\log\left(1 + \frac{1}{H^6 d^{10}}\right)} \leqslant 2H^6 d^{10}$$

for $H^6 d^1 0 \geqslant 1$. One the other hand, setting $\lambda = 1/(H^2\sqrt{d})$ gives $\hat{\beta} = O(\sqrt{d}\log(T))$, $\bar{\beta} = O(H\sqrt{d^3}\log(T))$, $\tilde{\beta} = O(H^2\sqrt{d^3}\log(T))$ and $\breve{\beta} = O(H\sqrt{d^3}\log(T))$, where $T = KH$.

Thus, by some basic manipulations, we obtain

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\hat{\sigma}_{k,h}^2 &\leqslant O\left(HT + H^{11}d^{16} + H^4 d^3\sqrt{T\log(T)} + H^9 d^{10}\log^3(T)\right) \\
&\leqslant O\left(HT + H^{11}d^{16} + H^7 d^4\log(T) + H^9 d^{10}\log^3(T)\right) \\
&\leqslant O\left(HT + H^{11}d^{16} + H^9 d^{10}\log^3(T)\right)
\end{aligned}
\tag{54}
$$

where the second inequality holds since $H^4 d^3\sqrt{T\log(T)} \leqslant [HT + H^7 d^6\log(T)]/2$, and the third inequality holds since $H^7 d^4\log(T) \leqslant H^9 d^{10}\log^3(T)$. Substituting Eq. (54) in (53) gives

$$
\begin{aligned}
&\text{Regret}(K) \\
&\leqslant O\left(\sqrt{d}\log(T)\sqrt{HT + H^{11}d^{16} + H^9 d^{10}\log^3(T)}\sqrt{Hd\log(T)} + \frac{3H^2 d}{\log(2)}\log\left(1 + \frac{H\sqrt{d}}{\log(2)}\right) + 2H\sqrt{2T\log\left(\frac{H}{\delta}\right)}\right) \\
&\leqslant O\left(Hd\sqrt{T\log^3(T)} + H^6 d^9\sqrt{\log^3(T)} + H^5 d^6\log^2(T)\right) \\
&= \tilde{O}\left(Hd\sqrt{T} + H^6 d^9\right),
\end{aligned}
$$

where the second inequality holds by dropping lower order terms and the fact that $\sqrt{a + b + c} \leqslant \sqrt{a} + \sqrt{b} + \sqrt{c}$ for any $a, b, c > 0$. □

E. Lower Bound

Remark 23 in Appendix of (Zhou et al., 2021) constructs a hard-to-learn linear MDP instance, which shares the same order of regret lower bound as a hard-to-learn linear mixture MDP instance. According to Theorem 8 in (Zhou et al., 2021), the linear mixture MDP has a regret lower bound of $\Omega(Hd\sqrt{T})$, which means the linear MDP with known reward has the same regret lower bound as well. We present the construction of this hard-to-learn linear MDP from (Zhou et al., 2021) in this section for completeness. This hard-to-learn linear MDP can be regarded as an extension of hard instances in linear bandits literature (Dani et al., 2008; Lattimore & Szepesvári, 2020). We first illustrate the structure of this MDP and then present the specific linear parametrization.

Hard MDP Instance   This MDP instance is denoted as $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_1\}_h, \{r_h\}_h\}$. The state space $\mathcal{S}$ consists of states $s_1, \ldots s_{H+2}$ such that $|\mathcal{S}| = H + 2$. There are 2^{d-1} actions and $\mathcal{A} = \{-1, 1\}^{d-1}$ such that each action $\boldsymbol{a} \in \mathcal{A}$ is denoted in vector from.

- Reward: For any stage $h \in [H + 2]$, only transitions originating at s_{H+2} incurs a reward.

- Transition: s_{H+1} and s_{H+2} are absorbing regardless of what action is taken. For state s_i with $i \leqslant H$, the transition probability is given as

$$\mathbb{P}_h(s'|s_i, \boldsymbol{a}) = \begin{cases} \iota + \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle, & s' = s_{H+2} \\ 1 - (\iota + \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle), & s' = s_{i+1} \\ 0, & \text{Otherwise} \end{cases},$$

  where $\iota = 1/H$ and $\boldsymbol{\mu}_h \in \{-\Delta, \Delta\}^{d-1}$ with $\Delta = \sqrt{\iota/K}/(4\sqrt{2})$ to make the probabilities are well-defined. The transition of this MDP is detailed in Figure 1.
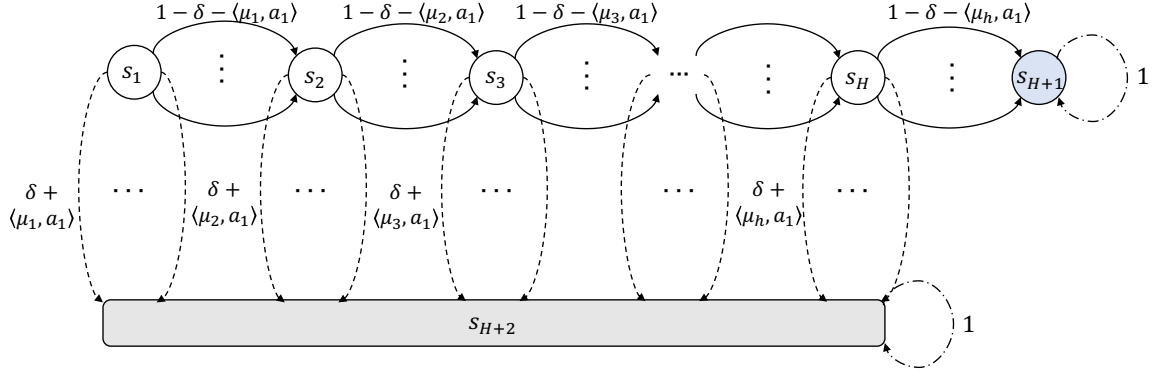
Figure 1. The transition matrix $\mathbb{P}_h$ of the hard-to-learn MDP.

Linear Parametrization Then, we specify the linear parametrization of this MDP. For each $h \in [H]$, the transition probability matrix $\mathbb{P}_h$ and the reward function r_h are defined as $\mathbb{P}_h(s' \mid s, a) = \langle \phi(s, a), \mu_h(s') \rangle$ and $r_h(s, a) = \langle \phi(s, a), \theta_h \rangle$, where $\phi(s, a) \in \mathbb{R}^{d+1}$ is the known feature mapping, $\mu_h = (\mu_h(s_1), ..., \mu_h(s_{H+2})) \in \mathbb{R}^{(d+1) \times (H+2)}$ and $\theta_h \in \mathbb{R}^{d+1}$ are unknown parameters in linear MDPs. Here, $\phi(s, a), \mu_h, \theta_h$ are specified as:

$$\phi(s, a) = \begin{cases} (\alpha, \beta a^\top, 0)^\top, & s = s_h, h \in [H+1] \\ (0, \mathbf{0}^\top, 1)^\top, & s = s_{H+2} \end{cases}$$

$$\mu_h(s') = \begin{cases} ((1-\iota)/\alpha, -\mu_h^\top/\beta, 0)^\top, & s' = s_{H+1}, \\ (\iota/\alpha, \mu_h^\top/\beta, 1)^\top, & s' = s_{H+2}, \\ \mathbf{0}, & \text{otherwise} \end{cases},$$

$$\theta_h = (\mathbf{0}^\top, 1)^\top$$

where $\alpha = \sqrt{1/(1 + \Delta(d-1))}$, $\beta = \sqrt{\Delta/(1 + \Delta(d-1))}$.

Norm Assumption We check the norm assumption of linear MDPs in the following:

1. For $s = s_h$ where $h \in [H+1]$, $\|\phi(s, a)\|_2 = \sqrt{\alpha^2 + (d-1)\beta^2} = 1$ and $\|\phi(s_{H+2}, a)\|_2 = 1$. Thus, $\|\phi(s, a)\|_2 \leqslant 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

2. For any $v = (v_1, \ldots, v_{H+2}) \in \mathbb{R}^{H+2}$ such that $\|v\|_\infty \leqslant 1$, we have

$$\|\mu_h v\|_2^2 = \left[\frac{v_1(1-\iota)}{\alpha} + \frac{v_1 \iota}{\alpha} \right]^2 + v_{H+2}^2 \leqslant \frac{1}{\alpha^2} + 1 = [1 + \Delta(d-1)]^2 + 1 = \left[1 + \frac{\sqrt{\delta/K}}{4\sqrt{2}}(d-1) \right]^2 + 1 \leqslant d+1,$$

where the last inequality holds by assuming episode number $K \geqslant (d-1)/(32H(\sqrt{d}-1))$. Thus, $\|\mu_h v\|_2 \leqslant \sqrt{d+1}$ for any $h \in [H]$.

3. In addition, $\|\theta_h\|_2 \leqslant 1 \leqslant \sqrt{d+1}$ for any $h \in [H]$.

Lower Bound The constructed linear MDP above has the same state space $\mathcal{S}$, action space $\mathcal{A}$, episode length H, reward function $\{r_h\}_{h \in [H]}$ and transition probability $\{\mathbb{P}_h\}_{h \in [H]}$ as the constructed hard-to-learn linear mixture MDP in Appendix E. of (Zhou et al., 2021), which shares the same regret lower bound $\Omega(Hd\sqrt{T})$ as shown in Theorem 8 in (Zhou et al., 2021) and is formalized in the following theorem.

Lemma E.1 (Lower bound of linear MDPs)**.** *Let $d > 1$ and suppose $K \geqslant \max\{(d-1)^2 H/2, (d-1)/(32H(d-1))\}$, $d \geqslant 4$, $H \geqslant 3$. Then for any algorithm there exists an episodic linear MDP parameterized by $\{\boldsymbol{\mu}_h\}_{h \in [H]}, \{\boldsymbol{\theta}_h\}_{h \in [H]}$ and satisfy the norm assumption given in Definition 3.1, such that the expected regret is lower bounded as follows:*

$$\mathbb{E}[\mathrm{Regret}(K)] \geqslant \Omega(Hd\sqrt{T}),$$

where $T = KH$ and the expectation is taken over the probability distribution generated by the interconnection of the algorithm and the MDP.

Proof. The proof is the same as that of Theorem 8 in (Zhou et al., 2021), except for changing B to $\sqrt{d}$ to satisfy the norm assumption of linear MDPs. $\qquad\square$

F. Auxiliary Lemmas

In this section, we give some auxiliary lemmas which serve as the preliminary for the proof above. We also include some other lemmas that are unnecessary for our theoretical analysis but can help readers be more familiar with related works. In general, these lemmas are categorized into four subsections:

- Appendix F.1 for some concentration inequalities;

- Appendix F.2 for properties related to elliptical potentials;

- Appendix F.3 presents some useful properties for linear MDPs;

- Appendix F.4 builds the covering number for covering net over some function classes of our interests.

F.1. Concentration Inequality

In this subsection, Lemma F.1 presents the Azuma-Hoeffding inequality, Lemma F.2 presents the Freedman's inequality in (Freedman, 1975), Lemma F.3 a Hoeffding-type self-normalized bound, and Lemma F.4 presents the full version of Theorem 7.1 in main paper.

Lemma F.1 (Azuma-Hoeffding Inequality)**.** *Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}_{i=1}^{n+1}$ such that $|x_i| \leqslant M$ almost surely. That is, x_i is $\mathcal{G}_{i+1}$-measurable and $\mathbb{E}[x_i \mid \mathcal{G}_i] = 0$ a.s. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$,*

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

Lemma F.2 (Freedman's Inequality, (Freedman, 1975))**.** *Let $\{x_i, \mathcal{F}_i\}$ be a martingale difference sequence with $\forall i \geqslant 1$, $\mathbb{E}(x_i \mid \mathcal{F}_{i-1}) = 0$, $\mathbb{E}(x_i^2 \mid \mathcal{F}_{i-1}) = \sigma_i^2$, $V_i^2 = \sum_{j=1}^i \sigma_j^2$. Furthermore, assume that $\mathbb{P}(|x_i| \leqslant c \mid \mathcal{F}_{i-1}) = 1$ for any $0 < c < \infty$.*

Then, for fixed $t \geqslant 1$ and any $\delta > 0$, with probability at least $1 - \delta$, we have:

$$\sum_{i=1}^t d_i \leqslant \sqrt{2V_t^2 \log(1/\delta)} + \frac{2}{3}c\log(1/\delta).$$

Lemma F.3 (Hoeffding inequality for vector-valued martingales, Theorem 1 in (Abbasi-Yadkori et al., 2011))**.** *Let $\{\mathcal{G}_t\}_{t=1}^\infty$ be a filtration, $\{\boldsymbol{x}_t, \eta_t\}_{t\geqslant 1}$ be a stochastic process so that $\boldsymbol{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable.*

Denote $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \boldsymbol{x}_i\boldsymbol{x}_i^\top$ for $t \geqslant 1$ and $\mathbf{Z}_0 = \lambda\mathbf{I}$. If $\|\boldsymbol{x}_t\|_2 \leqslant L$, and η_t satisfies

$$\mathbb{E}[\eta_t \mid \mathcal{G}_t] = 0, \quad |\eta_t| \leqslant R$$

for all $t \geqslant 1$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ we have:

$$\forall t > 0, \left\|\sum_{i=1}^t \boldsymbol{x}_i\eta_i\right\|_{\mathbf{Z}_t^{-1}} \leqslant R\sqrt{d\log(1 + tL^2/d\lambda) + \log(1/\delta)}.$$

Lemma F.4 (Bernstein inequality for vector-valued martingales, full version of Theorem 7.1). *Let $\{\mathcal{G}_t\}_{t=1}^{\infty}$ be a filtration, $\{\boldsymbol{x}_t, \eta_t\}_{t \geqslant 1}$ be a stochastic process so that $\boldsymbol{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable.*

If $\|\boldsymbol{x}_t\|_2 \leqslant L$, and η_t satisfies

$$\mathbb{E}\left[\eta_t \mid \mathcal{G}_t\right] = 0, \quad \mathbb{E}\left[\eta_t^2 \mid \mathcal{G}_t\right] \leqslant \sigma^2, \quad \left|\eta_t \cdot \min\left\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}\right\}\right| \leqslant R$$

for all $t \geqslant 1$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ we have:

$$\forall t > 0, \left\|\sum_{i=1}^{t} \boldsymbol{x}_i \eta_i\right\|_{\mathbf{Z}_t^{-1}} \leqslant 8\sigma\sqrt{d \log\left(1 + tL^2/(d\lambda)\right) \log\left(4t^2/\delta\right)} + 4R \log\left(4t^2/\delta\right)$$

where $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} \boldsymbol{x}_i\boldsymbol{x}_i^{\top}$ for $t \geqslant 1$ and $\mathbf{Z}_0 = \lambda\mathbf{I}$.

F.2. Elliptical Potentials

In this subsection, we present Lemma F.5 in (Abbasi-Yadkori et al., 2011), which is an important lemma for building $\sqrt{O}(\sqrt{T})$ worst-case regret for many algorithms for linear bandits or RL with linear function approximation. Then, we present Lemma F.6, about elliptical potentials (Exercise 19.3 in (Lattimore & Szepesvári, 2020)), which states that one cannot have more than $O(d)$ big intervals. Lemma F.6 is further generalized in Lemma D.7 in Appendix D.3. In addition, we also present Lemma F.7 (Lemma 12 in (Abbasi-Yadkori et al., 2011)) and Lemma F.8 (Lemma E.1. in (He et al., 2022)) about the "rare-switching" update strategy of the constructed value function.

Lemma F.5 (Lemma 11, (Abbasi-Yadkori et al., 2011)). *Given $\lambda > 0$ and sequence $\{\boldsymbol{x}_t\}_{t=1}^{T} \subset \mathbb{R}^d$ with $\|\boldsymbol{x}_t\|_2 \leqslant L$ for all $t \in [T]$, define $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} \boldsymbol{x}_i\boldsymbol{x}_i^{\top}$ for $t \geqslant 1$ and $\mathbf{Z}_0 = \lambda\mathbf{I}$. We have*

$$\sum_{t=1}^{T} \min\left\{1, \|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}^2\right\} \leqslant 2d \log\left(1 + \frac{TL^2}{d\lambda}\right).$$

Lemma F.6 (Elliptical Potentials: You cannot have more than $O(d)$ big intervals. Exercise 19.3 in (Lattimore & Szepesvári, 2020)). *Given $\lambda > 0$ and sequence $\{\boldsymbol{x}_t\}_{t=1}^{T} \subset \mathbb{R}^d$ with $\|\boldsymbol{x}_t\|_2 \leqslant L$ for all $t \in [T]$, define $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} \boldsymbol{x}_i\boldsymbol{x}_i^{\top}$ for $t \geqslant 1$ and $\mathbf{Z}_0 = \lambda\mathbf{I}$. The number of times $\|\boldsymbol{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}^2 \geqslant 1$ is at most*

$$\frac{3d}{\log(2)} \log\left(1 + \frac{L^2}{\lambda \log(2)}\right).$$

Lemma F.7 (Lemma 12 in (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 $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_{\mathbf{A}} \leqslant \|\boldsymbol{x}\|_{\mathbf{B}} \cdot \sqrt{\det \mathbf{A}/\det \mathbf{A}}$.*

Lemma F.8 (Lemma E.1. in (He et al., 2022)). *The number of episodes where the algorithm updates the value function in Algorithm 1 is upper bounded by $dH \log(1 + K)$.*

F.3. Linear MDP Property

This subsection gives some indirect results about the estimated parameter $\widehat{\boldsymbol{\mu}}_{k,h}$ in linear MDPs.

Lemma F.9. *In Algorithm 1, for any $k \in [K]$ and any $h \in [H]$, we have:*

$$\widehat{\boldsymbol{\mu}}_{k,h} - \boldsymbol{\mu}_h = \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1}\left[-\lambda\boldsymbol{\mu}_h + \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2}\phi(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^{\top}\right] \tag{55}$$

Proof. We start from the closed-form solution of $\widehat{\boldsymbol{\mu}}_{k,h}$:

$$
\begin{aligned}
\widehat{\boldsymbol{\mu}}_{k,h} =& \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \boldsymbol{\delta}\left(s_{h+1}^i\right)^\top = \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \left(\mathbb{P}_h(\cdot \mid s_h^i, a_h^i)^\top + {\boldsymbol{\epsilon}_h^i}^\top\right) \\
=& \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \left(\boldsymbol{\phi}(s_h^i, a_h^i)^\top \boldsymbol{\mu}_h + {\boldsymbol{\epsilon}_h^i}^\top\right) \\
=& \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i) \boldsymbol{\phi}(s_h^i, a_h^i)^\top \boldsymbol{\mu}_h + \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^\top \\
=& \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \left(\widehat{\boldsymbol{\Lambda}}_{k,h} - \lambda \mathbf{I}\right) \boldsymbol{\mu}_h + \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^\top \\
=& \boldsymbol{\mu}_h - \lambda \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \boldsymbol{\mu}_h + \widehat{\boldsymbol{\Lambda}}_{k,h}^{-1} \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}(s_h^i, a_h^i){\boldsymbol{\epsilon}_h^i}^\top .
\end{aligned}
$$

Rearranging terms gives Eq. (55). $\qquad\square$

F.4. Covering Net

This subsection presents lemmas required for bounding the covering number of considered function classes, including $\widehat{\mathcal{V}}, \widehat{\mathcal{V}}^2, \check{\mathcal{V}}$.

Lemma F.10 (Covering Number of Euclidean Ball, Lemma D.5. in (Jin et al., 2020)). *For any $\varepsilon > 0$, the ε-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\varepsilon)^d$.*

Lemma F.11 (Lemma E.6. in (He et al., 2022)). *Let $\widehat{\mathcal{N}}_\varepsilon$ be the ε-covering of $\widehat{\mathcal{V}}$ with respect to the distance $\mathrm{dist}\,(V, V') = \sup_x |V(x) - V'(x)|$, where $\widehat{\mathcal{V}}$ is defined in Definition C.5. Then*

$$
\log |\widehat{\mathcal{N}}_\varepsilon| \leqslant dJ \log(1 + 4L/\varepsilon) + d^2 J \log\left[1 + 8d^{1/2} B^2 / \left(\lambda \varepsilon^2\right)\right].
$$

Lemma F.12 (Lemma E.8. in (He et al., 2022)). *Let $\widehat{\mathcal{N}}_\varepsilon^2$ be the ε-covering of $\widehat{\mathcal{V}}^2$ with respect to the distance $\mathrm{dist}\,(V, V') = \sup_x |V(x) - V'(x)|$, where $\widehat{\mathcal{V}}^2$ is defined in Definition C.6. Then*

$$
\log |\widehat{\mathcal{N}}_\varepsilon^2| \leqslant dJ \log(1 + 8LH/\varepsilon) + d^2 J \log\left[1 + 32d^{1/2} B^2 H^2 / \left(\lambda \varepsilon^2\right)\right].
$$

Lemma F.13. *Let $\check{\mathcal{N}}_\varepsilon$ be the ε-covering of $\check{\mathcal{V}}$ with respect to the distance $\mathrm{dist}\,(V, V') = \sup_x |V(x) - V'(x)|$, where $\check{\mathcal{V}}$ is defined in Definition C.7. Then*

$$
\log |\check{\mathcal{N}}_\varepsilon| \leqslant d \log(1 + 4L/\varepsilon) + d^2 \log\left[1 + 8d^{1/2} B^2 / \left(\lambda \varepsilon^2\right)\right].
$$

Proof. Denote $\mathbf{A} = \beta^2 \Lambda^{-1}$, then for any $\check{V}(\cdot) \in \check{\mathcal{V}}$,

$$
\check{V}(\cdot) = \max\left\{\max_a \boldsymbol{w}^\top \boldsymbol{\phi}(\cdot, a) - \sqrt{\boldsymbol{\phi}(\cdot, a)^\top \mathbf{A} \boldsymbol{\phi}(\cdot, a)}, 0\right\} \tag{56}
$$

for $\|\boldsymbol{w}\| \leqslant L$ and $\|\mathbf{A}\| \leqslant B^2 \lambda^{-1}$. The proof is almost the same as that for Lemma F.11, since for any two functions $\check{V}_1, \check{V}_2 \in \check{\mathcal{V}}$, let them take the form in Eq. (56) with parameters $(\boldsymbol{w}_1, \mathbf{A}_1)$ and $(\boldsymbol{w}_2, \mathbf{A}_2)$, respectively. Since $|\min\{x, 0\} -$

$\min\{y, 0\}| \leqslant |x - y|$ for any $x, y \in \mathbb{R}$ and $\max_a$ is a contraction mapping, we have

$$
\begin{aligned}
\operatorname{dist}(\widehat{V}_1, \widehat{V}_2) &\leqslant \sup_{s,a} \left\| \left[\boldsymbol{w}_1^\top \boldsymbol{\phi}(s,a) - \sqrt{\boldsymbol{\phi}(s,a)^\top \mathbf{A}_2 \boldsymbol{\phi}(s,a)} \right] - \left[\boldsymbol{w}_2^\top \boldsymbol{\phi}(s,a) - \sqrt{\boldsymbol{\phi}(s,a)^\top \mathbf{A}_2 \boldsymbol{\phi}(s,a)} \right] \right\| \\
&\leqslant \sup_{\boldsymbol{\phi}: \|\boldsymbol{\phi}\|_2 \leqslant 1} \left\| \left[\boldsymbol{w}_1^\top \boldsymbol{\phi} - \sqrt{\boldsymbol{\phi}^\top \mathbf{A}_2 \boldsymbol{\phi}} \right] - \left[\boldsymbol{w}_2^\top \boldsymbol{\phi} - \sqrt{\boldsymbol{\phi}^\top \mathbf{A}_2 \boldsymbol{\phi}} \right] \right\| \\
&\leqslant \sup_{\boldsymbol{\phi}: \|\boldsymbol{\phi}\|_2 \leqslant 1} \left| (\boldsymbol{w}_1 - \boldsymbol{w}_2)^\top \boldsymbol{\phi} \right| + \sup_{\boldsymbol{\phi}: \|\boldsymbol{\phi}\|_2 \leqslant 1} \sqrt{\left| \boldsymbol{\phi}^\top (\mathbf{A}_2 - \mathbf{A}_1) \boldsymbol{\phi} \right|} \\
&= \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \sqrt{\|\mathbf{A}_1 - \mathbf{A}_2\|_2} \leqslant \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \sqrt{\|\mathbf{A}_1 - \mathbf{A}_2\|_F},
\end{aligned}
\tag{57}
$$

where the second last inequality follows from the fact that $|\sqrt{x} - \sqrt{y}| \leqslant \sqrt{x-y}$ holds for any $x, y \geqslant 0$. For matrices, $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the matrix operator norm and Frobenius norm, respectively.

Let $\mathcal{C}_{\boldsymbol{w}}$ be an $\varepsilon/2$-cover of $\{\boldsymbol{w} \in \mathbb{R}^d \mid \|\boldsymbol{w}\|_2 \leqslant L\}$ with respect to the 2-norm, and $\mathcal{C}_{\mathbf{A}}$ be an $\varepsilon^2/4$-cover of $\{\mathbf{A} \in \mathbb{R}^{d \times d} \mid \|\mathbf{A}\|_F \leqslant d^{1/2} B^2 \lambda^{-1}\}$ with respect to the Frobenius norm. By Lemma F.10, we have:

$$
|\mathcal{C}_{\boldsymbol{w}}| \leqslant (1 + 4L/\varepsilon)^d, \quad |\mathcal{C}_{\mathbf{A}}| \leqslant \left[1 + 8d^{1/2} B^2 / \left(\lambda \varepsilon^2 \right) \right]^{d^2}
$$

By Eq. (57), for any $\widehat{V}_1 \in \widehat{\mathcal{V}}$, there exists $\boldsymbol{w}_2 \in \mathcal{C}_{\boldsymbol{w}}$ and $\mathbf{A}_2 \in \mathcal{C}_{\mathbf{A}}$ such that $\widehat{V}_2$ parametrized by $(\boldsymbol{w}_2, \mathbf{A}_2)$ satisfies $\operatorname{dist}(\widehat{V}_1, \widehat{V}_2) \leqslant \varepsilon$. Hence, it holds that $|\mathcal{N}_\varepsilon| \leqslant |\mathcal{C}_{\boldsymbol{w}}| \cdot |\mathcal{C}_{\mathbf{A}}|$, which gives

$$
\log |\mathcal{N}_\varepsilon| \leqslant \log |\mathcal{C}_{\boldsymbol{w}}| + \log |\mathcal{C}_{\mathbf{A}}| \leqslant d \log(1 + 4L/\varepsilon) + d^2 \log \left[1 + 8d^{1/2} B^2 / \left(\lambda \varepsilon^2 \right) \right].
$$

This concludes the proof. $\qquad \square$