Variance-Dependent Regret Bounds for Linear Bandits and Reinforcement Learning: Adaptivity and Computational Efficiency

Heyang Zhao HYZHAO@CS.UCLA.EDU
University of California, Los Angeles

Jiafan He JIAFANHE19@UCLA.EDU
University of California, Los Angeles

Dongruo Zhou DRZHOU@CS.UCLA.EDU
University of California, Los Angeles

Tong Zhang TONGZHANG@TONGZHANG-ML.ORG
Google Research and The Hong Kong University of Science and Technology

Quanquan Gu QGU@CS.UCLA.EDU
University of California, Los Angeles

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Recently, several studies (Zhou et al., 2021a; Zhang et al., 2021b; Kim et al., 2021; Zhou and Gu, 2022) have provided variance-dependent regret bounds for linear contextual bandits, which interpolates the regret for the worst-case regime and the deterministic reward regime. However, these algorithms are either computationally intractable or unable to handle unknown variance of the noise. In this paper, we present a novel solution to this open problem by proposing the *first computationally efficient* algorithm for linear bandits with heteroscedastic noise. Our algorithm is adaptive to the unknown variance of noise and achieves an $\widetilde{O}(d\sqrt{\sum_{k=1}^{K}\sigma_k^2} + d)$ regret, where σ_k^2 is the *variance* of the noise at the round k, d is the dimension of the contexts and K is the total number of rounds. Our results are based on an adaptive variance-aware confidence set enabled by a new Freedman-type concentration inequality for self-normalized martingales and a multi-layer structure to stratify the context vectors into different layers with different uniform upper bounds on the uncertainty.

Furthermore, our approach can be extended to linear mixture Markov decision processes (MDPs) in reinforcement learning. We propose a variance-adaptive algorithm for linear mixture MDPs, which achieves a problem-dependent horizon-free regret bound that can gracefully reduce to a nearly constant regret for deterministic MDPs. Unlike existing nearly minimax optimal algorithms for linear mixture MDPs, our algorithm does not require explicit variance estimation of the transitional probabilities or the use of high-order moment estimators to attain horizon-free regret. We believe the techniques developed in this paper can have independent value for general online decision making problems.

Keywords: Linear bandits, reinforcement learning, instance-dependent regret

© 2023 H. Zhao, J. He, D. Zhou, T. Zhang & Q. Gu.

1. Introduction

The Multi-Armed Bandits (MAB) problem has been persistently studied since 1933 (Thompson, 1933; Robbins, 1952; Cesa-Bianchi and Fischer, 1998; Auer et al., 2002). In the past decades, a variety of bandit algorithms have been developed under different settings, emerging their practicality in assorted real world tasks such as online advertising (Li et al., 2010), clinical experiments (Villar et al., 2015) and resource allocations (Lattimore et al., 2015), to mention a few. For a thorough review of bandit algorithms, please refer to Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2020).

To deal with a large number of arms, contextual linear bandits (Auer, 2002; Abe et al., 2003; Li et al., 2010), where each arm is associated with a context vector and the expected reward is a linear function of the context vector, have garnered a lot of attention. Numerous studies have attempted to design algorithms to achieve the optimal regret bound for linear bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011). Despite the achievement of minimax-optimal regret bounds in various settings, they only quantify the performance of a specific algorithm under the worst-case scenario. However, in the noiseless scenario (i.e., the variance of the noise equal 0), the learner only requires $\widetilde{O}(d)$ regret to recover the underlying coefficients of the linear function. This motivates a series of work on variance-dependent regret for linear bandits (Zhou et al., 2021a; Zhang et al., 2021b; Zhou and Gu, 2022; Zhao et al., 2022), which bridges the gap between the worst-case constant-variance regime (i.e., noisy case) and the deterministic regime (i.e., noiseless case). In these works, the regret bounds depend on the variance of noise at each round, i.e., $\{\sigma_i^2\}_{i=1}^K$ where K is the total number of rounds. Unfortunately, all these prior approaches are either computationally inefficient or non-adaptive, meaning the agent must possess prior knowledge of the variance to learn the reward function. As a result, none of the existing algorithms are practical enough for real-world use, despite being designed for better performance in reality. Therefore, an open question arises:

Can we design computationally tractable algorithms for linear bandits with heteroscedastic noise to obtain a variance-dependent regret bound without prior knowledge on the noise?

1.1. Our Contributions

In this paper, we answer this question affirmatively by proposing the first computationally efficient algorithm for heteroscedastic linear bandits with unknown variance and attaining a regret bound scales as $\widetilde{O}\big(d\sqrt{\sum_{k=1}^K \sigma_k^2} + d\big)$, where σ_k^2 is the *variance* of the noise at the round k, d is the dimension of the contexts and K is the total number of rounds. Our result is significant in the sense that it is minimax optimal in both the deterministic case and the worst case. When there is no noise, the above regret degenerates to $\widetilde{O}(d)$, which corresponds to the benign regime. In the worst case when the noise is R-sub-Gaussian, the above regret reduces to $\widetilde{O}(dR\sqrt{K} + d)$, which matches the minimax regret bound proved in Abbasi-Yadkori et al. (2011). Please refer to Table 1 for a comparison between our result and the previous results in linear contextual bandits.

Our algorithm and its analysis rely on the following new techniques.

- We propose a new Freedman-type concentration inequality for vector-valued self-normalized martingales, which is applicable to the heteroscedastic random variables. This strictly generalizes the previous Bernstein-type concentration inequality (Theorem 4.1, Zhou et al. 2021a) for vector-valued self-normalized martingales with homoscedastic random variables.

[1] For the deterministic-case, the variance at stage $k \in [K]$ satisfies $\sigma_k = 0$. The regret guarantee is the same as the general case for variance-unaware algorithms.

Table 1: Comparison between different algorithms for stochastic linear contextual bandits. d, K, $\{\sigma_k\}_{k\in[K]}$ are the dimension of context vectors, the number of rounds and the variance of noise at round $k \in [K]$. The last column indicates whether the corresponding algorithm requires the variance information to achieve variance-dependent regret.

Algorithm	Regret (General-Case)	Regret (Deterministic-Case)[1]	Efficiency	Variances
ConfidenceBall$_2$ (Dani et al., 2008)	$\widetilde{O}(d\sqrt{K})$	$\widetilde{O}(d\sqrt{K})$	Yes	N/A
OFUL (Abbasi-Yadkori et al., 2011)	$\widetilde{O}(d\sqrt{K})$	$\widetilde{O}(d\sqrt{K})$	Yes	N/A
Weighted OFUL (Zhou et al., 2021a)	$\widetilde{O}\Big(d\sqrt{\sum_{k=1}^K \sigma_k^2} + \sqrt{dK}\Big)$	$\widetilde{O}(\sqrt{dK})$	Yes	Known
Weighted OFUL+ (Zhou and Gu, 2022)	$\widetilde{O}\Big(d\sqrt{\sum_{k=1}^K \sigma_k^2} + d\Big)$	$O(d \cdot \text{polylog}(d, K))$	Yes	Known
VOFUL (Zhang et al., 2021b)	$\widetilde{O}\Big(d^{4.5}\sqrt{\sum_{k=1}^K \sigma_k^2} + d^5\Big)$	$O(d^5 \cdot \text{polylog}(d, K))$	No	Unknown
VOFUL2 (Kim et al., 2021)	$\widetilde{O}\Big(d^{1.5}\sqrt{\sum_{k=1}^K \sigma_k^2} + d^2\Big)$	$O(d^2 \cdot \text{polylog}(d, K)))$	No	Unknown
`SAVE` (Theorem 2.3)	$\widetilde{O}\Big(d\sqrt{\sum_{k=1}^K \sigma_k^2} + d\Big)$	$O(d \cdot \text{polylog}(d, K))$	Yes	Unknown

- We employ a multi-layer structure to partition the observed context vectors according to their elliptical norm. Different from the classic `SupLinUCB` algorithm (Chu et al., 2011), we use carefully designed weights within each layer to ensure that all the reweighted context vectors in the same layer have a uniform elliptical norm.
- Equipped with the new concentration inequality, we design a new adaptive variance-aware exploration strategy. Specifically, we adopt a self-adaptive confidence set whose radius is updated at each round based on the 'square loss' incurred by the online estimator.

Furthermore, we apply our novel techniques to episodic Markov decision processes, where the agent interacts with the environment by taking actions and observing states and rewards generated by the unknown dynamics over time. We focus on linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b) in this paper, whose transition dynamic is assumed to be a linear combination of d basic transition models. For linear mixture MDPs, both minimax optimal horizon-dependent regret (Zhou et al., 2021a) and horizon-free regret Zhang et al. (2021b); Kim et al. (2021); Zhou and Gu (2022) have been achieved. We propose an algorithm named `UCRL-AVE` and derive a tight problem-dependent regret bound that has no explicit polynomial dependency on neither the number of episodes K nor the planing horizon H. Our regret bound gracefully degrades to the nearly minimax optimal horizon-free regret bound achieved by Zhou and Gu (2022) in the worst case. See Table 2 for a comparison between our regret bound with the previous results regarding linear mixture MDPs.

1.2. Other Related Work

Here we discuss the related work on problem-dependent regret in RL. For additional related work, please refer to Appendix A. Most of the performance guarantees for episodic MDPs have been focused on worst-case regret bounds. However, some works have achieved problem-dependent regret, which holds in various scenarios, as demonstrated by studies such as Zanette and Brunskill (2019); Simchowitz and Jamieson (2019); Jin et al. (2020a); Dann et al. (2021); Xu et al. (2021);

Table 2: Comparison of our variance-dependent regret with existing regret bounds for linear mixture MDPs. H, d, K, are the horizon of the underlying MDP, the dimension of the feature vectors and the number of episodes. Var_K^* is defined in Section 3 to characterize the randomness of MDPs. It is shown later in Section 3 that our variance-dependent regret degrades to $\widetilde{O}(d\sqrt{K}+d^2)$ in the worst case, which matches the nearly minimax optimal horizon-free regret in linear mixture MDPs.

Algorithm	Regret (General-Case)	Variance-Dependent	Assumption	Efficiency
UCRL-VTR (Ayoub et al., 2020; Jia et al., 2020)	$\widetilde{O}(d\sqrt{H^3K})$	No	Homogeneous $\sum_{h=1}^H r_h \le H$	Yes
UCRL-VTR+ (Zhou et al., 2021a)	$\widetilde{O}(\sqrt{d^2H^3+dH^4}\sqrt{K}$ $+d^2H^3+d^3H^2)$	No	Inhomogeneous $\sum_{h=1}^H r_h \le H$	Yes
VarLin (Zhang et al., 2021b)	$\widetilde{O}\Big(d^{4.5}\sqrt{K}+d^9\Big)$	No	Homogeneous $\sum_{h=1}^H r_h \le 1$	No
VarLin2 (Kim et al., 2021)	$\widetilde{O}(d\sqrt{K}+d^2)$	No	Homogeneous $\sum_{h=1}^H r_h \le 1$	No
HF-UCRL-VTR+ (Zhou and Gu, 2022)	$\widetilde{O}(d\sqrt{K}+d^2)$	No	Homogeneous $\sum_{h=1}^H r_h \le 1$	Yes
`UCRL-AVE` (Theorem 2.3)	$\widetilde{O}\Big(d\sqrt{\mathrm{Var}_K^*}+d^2\Big)$	Yes	Homogeneous $\sum_{h=1}^H r_h \le 1$	Yes

Wagenmaker et al. (2022); He et al. (2021a). These results can be broadly categorized into two groups. The first group is first-order regret in RL, which was originally proposed by Zanette and Brunskill (2019) and later extended to the linear MDP setting by Wagenmaker et al. (2022). The second group is gap-dependent regret guarantees, which have been studied for both tabular MDPs (Simchowitz and Jamieson, 2019; Xu et al., 2021; Dann et al., 2021) and linear MDPs/linear mixture MDPs (He et al., 2021a). We also notice that a concurrent work by Zhou et al. (2023) considers variance-dependent bound in tabular MDPs. Our paper utilizes the same definition of *trajectory-based total variance* as Zhou et al. (2023), which characterizes the *randomness* of an episodic MDP.

Notation We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. We denote by $[n]$ the set $\{1,\ldots,n\}$. For a vector $\mathbf{x}\in\mathbb{R}^d$ and a positive semi-definite matrix $\boldsymbol{\Sigma}\in\mathbb{R}^{d\times d}$, we denote by $\|\mathbf{x}\|_2$ the vector's Euclidean norm and define $\|\mathbf{x}\|_{\boldsymbol{\Sigma}}=\sqrt{\mathbf{x}^\top\boldsymbol{\Sigma}\mathbf{x}}$. For $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, let $\mathbf{x}\odot\mathbf{y}$ be the Hadamard (componentwise) product of $\mathbf{x}$ and $\mathbf{y}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$ with $n=1,2,\ldots$, we write $a_n=O(b_n)$ if there exists an absolute constant $C>0$ such that $a_n\le Cb_n$ holds for all $n\ge 1$ and write $a_n=\Omega(b_n)$ if there exists an absolute constant $C>0$ such that $a_n\ge Cb_n$ holds for all $n\ge 1$. We use $\widetilde{O}(\cdot)$ to further hide the polylogarithmic factors. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function. For $a,b\in\mathbb{R}$ satisfying $a\le b$, we use $[x]_{[a,b]}$ to denote the truncation function $x\cdot\mathbb{1}\{a\le x\le b\}+a\cdot\mathbb{1}\{x<a\}+b\cdot\mathbb{1}\{x>b\}$.

2. Variance-Aware Learning for Heteroscedastic Linear Bandits

In this section, we propose a computationally efficient variance-aware algorithm, dubbed `SAVE` (**S**uplin + **A**daptive **V**ariance-aware **E**xploration), for stochastic linear contextual bandits and present its theoretical guarantees. `SAVE` does not require the knowledge of the noise variance (or its upper bound), making it adaptable to varying levels of noise variance.

2.1. Problem Setup

We consider a heteroscedastic variant of the classic linear contextual bandit problem (Zhou et al., 2021a; Zhang et al., 2021b). Let K be the total number of rounds. At each round $k \in [K]$, the interaction between the agent and the environment is as follows: (1) the environment generates an arbitrary decision set $\mathcal{D}_k \subseteq \mathbb{R}^d$ where each element represents a feasible action that can be selected by the agent; (2) the agent observes $\mathcal{D}_k$ and selects $\mathbf{a}_k \in \mathcal{D}_k$; and (3) the environment generates the stochastic noise ϵ_k at round k and reveal the stochastic reward $r_k = \langle \boldsymbol{\theta}^*, \mathbf{a}_k \rangle + \epsilon_k$ to the agent. We assume that there exists a uniform bound $A > 0$ for the ℓ_2 norm of the feasible actions, i.e., for all $k \in [K]$, $\mathbf{a} \in \mathcal{D}_k$, it holds that $\|\mathbf{a}\|_2 \leq A$. Following Zhou et al. (2021a), we assume the following condition on the random noise ϵ_k at each round k:

$$\mathbb{P}\left(|\epsilon_k| \leq R\right) = 1, \quad \mathbb{E}[\epsilon_k | \mathbf{a}_{1:k}, \epsilon_{1:k-1}] = 0, \quad \mathbb{E}[\epsilon_k^2 | \mathbf{a}_{1:k}, \epsilon_{1:k-1}] = \sigma_k^2. \tag{2.1}$$

Without loss of generality, we assume that the size of $\mathcal{D}_k$ is finite and is bounded by $|\mathcal{D}|$ for all $k \in [K]$. If the size of $\mathcal{D}_k$ is infinite, we can use standard covering argument to convert it to be finite.

The goal of the agent is to minimize the cumulative regret defined as follows:

$$\mathrm{Regret}(K) = \sum_{k \in [K]} \big(\langle \mathbf{a}_k^*, \boldsymbol{\theta}^* \rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}^* \rangle \big), \quad \text{where } \mathbf{a}_k^* = \operatorname*{argmax}_{\mathbf{a} \in \mathcal{D}_k} \langle \mathbf{a}, \boldsymbol{\theta}^* \rangle. \tag{2.2}$$

2.2. Technical Challenges

The key technical challenge we face is to provide a tight upper bound of $\langle \mathbf{a}_k^*, \boldsymbol{\theta}^* \rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}^* \rangle$. The classical approach is to use the *optimism-in-the-face-of-uncertainty* principle (Abbasi-Yadkori et al., 2011), and construct a confidence set $\mathcal{C}_k$ which includes $\boldsymbol{\theta}^*$ w.h.p., then upper bound $\langle \mathbf{a}_k^*, \boldsymbol{\theta}^* \rangle$ with $\langle \mathbf{a}_k, \boldsymbol{\theta}_k \rangle$, where $\boldsymbol{\theta}_k \in \mathcal{C}_k$. Starting from here, there are two main approaches to bound $\langle \mathbf{a}_k, \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle$ for heteroscedastic linear bandits.

The first approach bounds $\langle \mathbf{a}_k, \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle$ with $\|\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_k^{-1}} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_k}$ by Cauchy-Schwarz inequality. Zhou et al. (2021a); Zhou and Gu (2022) constructed $\mathcal{C}_k$ as an ellipsoid centering at $\widehat{\boldsymbol{\theta}}_k$, which is the solution to a *weighted linear regression* problem over the past contexts $\mathbf{a}_i$, and their weight is based on the upper bound of the variance of heteroscedastic noise σ_k^2. Then they bound $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_k}$ by $\|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_k}$ and $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_k}$ separately, each of which can be bounded properly by using the self-normalized concentration inequalities proposed in Zhou et al. (2021a); Zhou and Gu (2022). However, as we have mentioned before, their approach is limited to the case where an upper bound of σ_k^2 is known.

The second approach (Zhang et al., 2021b; Kim et al., 2021) follows the *test-based* framework. Instead of constructing $\mathcal{C}_k$ as an ellipsoid centering at a least square estimator $\widehat{\boldsymbol{\theta}}_k$ for each round k, Zhang et al. (2021b); Kim et al. (2021) constructed $\mathcal{C}_k$ as the intersection of a series of sub-confidence sets denoted by different tests, where each test represents a constraint over a potential direction of the action $\mathbf{a}_k$. The limitation of their approach is that, in order to have a uniform upper bound on $\langle \mathbf{a}_k, \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle$, they have to cover all possible directions of $\mathbf{a}_k$, which leads to an $\exp(d)$ number of test sets by the standard covering argument. This makes the computational time of their test-based algorithms depend on d exponentially, which is computationally inefficient.

2.3. A New Freedman-Type Concentration Inequality for Vector-Valued Martingales

To tackle the above technical challenges posed by both weighted linear regression and test-based approach, we seek to develop a new Freedman-type concentration inequality for vector-valued self-normalized martingales with heteroscedastic noise (i.e., non-uniform variance).

Theorem 2.1 *Let $\{\mathcal{G}_k\}_{k=1}^\infty$ be a filtration, and $\{\mathbf{x}_k, \eta_k\}_{k\ge 1}$ be a stochastic process such that $\mathbf{x}_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma, \lambda, \epsilon > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $k \ge 1$, let $y_k = \langle \boldsymbol{\mu}^*, \mathbf{x}_k\rangle + \eta_k$, where $\eta_k, \mathbf{x}_k$ satisfy*

$$\mathbb{E}[\eta_k|\mathcal{G}_k] = 0,\ |\eta_k| \le R,\ \sum_{i=1}^k \mathbb{E}[\eta_i^2|\mathcal{G}_i] \le v_k,\ \textit{for}\ \forall\ k \ge 1$$

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

$$\beta_k = 16\rho\sqrt{v_k \log(4k^2/\delta)} + 6\rho R \log(4k^2/\delta),$$

where $\rho \ge \sup_{k\ge 1} \|\mathbf{x}_k\|_{\mathbf{Z}_{k-1}^{-1}}$. Then, for any $0 < \delta < 1$, we have with probability at least $1-\delta$ that,

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

Theorem 2.1 can be viewed as an extension of Freedman's inequality (Freedman, 1975) from scalar-valued martingales to vector-valued self-normalized martingales. Though Zhou et al. (2021a) proposed Bernstein-type concentration inequalities for vector-valued martingales (Theorem 4.1, Zhou et al. 2021a), their inequality relies on a uniform upper bound on the variance of random variables, i.e., $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_k} \le \widetilde{O}(\sigma\sqrt{d} + R)$, where $\sigma^2 \ge \sup_{k\ge 1} \mathbb{E}[\eta_i^2|\mathcal{G}_i]$. In contrast, the upper bound provided by Theorem 2.1 depends on the maximum of $\|\mathbf{x}_k\|_{\mathbf{Z}_{k-1}^{-1}}$, which is of the order $\widetilde{O}(\sqrt{d/k})$ under certain conditions (Carpentier et al., 2020). For these cases, our upper bound for $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_k}$ scales as $\widetilde{O}(\sqrt{d \cdot v_k/k} + R \cdot \sqrt{d/k})$, which is more fine-grained and strictly tighter than the previous upper bounds when $k \ge d$.

2.4. The Proposed Algorithm

Equipped with the new Freedman-type concentration inequality, we will design a new algorithm that is adaptive to the unknown variance of noise.

2.4.1. SupLin with Adaptive Variance-Aware Exploration

As discussed in the last subsection, in order to exploit Theorem 2.1 effectively in the heteroscedastic linear bandits setting, we need the uncertainty/bonus term $\|\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_{k-1}^{-1}}$ to be small, where $\widehat{\boldsymbol{\Sigma}}_{k-1}$ is the covariance matrix of $\mathbf{a}_k$. However, such a term is in the order of $O(1)$ in the worst case. Our algorithm partition the observed contexts into different layers such that the uncertainty of the contexts within each layer is small. Our algorithm is displayed in Algorithm 1, namely `SAVE`.

Overall algorithm structure In general, Algorithm 1 shares a similar multi-layer structure as `SupLinUCB` in Chu et al. (2011). Algorithm 1 maintains L context sets $\Psi_{k,\ell}, \ell \in [L]$ at the k-th round. The goal of Algorithm 1 at the k-th round is to select $\mathbf{a}_k$ which maximizes $\langle \mathbf{a}, \boldsymbol{\theta}^*\rangle$. Since $\boldsymbol{\theta}^*$ is unknown, the selection process is based on L number of estimates of $\boldsymbol{\theta}^*$, which we denote them by $\widehat{\boldsymbol{\theta}}_{k,\ell}, \ell \in [L]$. $\widehat{\boldsymbol{\theta}}_{k,\ell}$ is the solution to some regression problem over contexts in $\Psi_{k,\ell}$ and

Algorithm 1 SupLin + Adaptive Variance-aware Exploration (SAVE)

```
Require: $\alpha > 0$, and the upper bound on the $\ell_2$-norm of $\mathbf{a}$ in $\mathcal{D}_k (k \geq 1)$, i.e., $A$.
1: Initialize $L \leftarrow \lceil \log_2(1/\alpha) \rceil$.
2: Initialize the estimators for all layers: $\widehat{\mathbf{\Sigma}}_{1,\ell} \leftarrow 2^{-2\ell} \cdot \mathbf{I}$, $\widehat{\mathbf{b}}_{1,\ell} \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_{1,\ell} \leftarrow \mathbf{0}$, $\Psi_{k,\ell} \leftarrow \varnothing$, $\widehat{\beta}_{1,\ell} \leftarrow 2^{-\ell+1}$ for all $\ell \in [L]$.
3: for $k = 1, \ldots, K$ do
4:     Observe $\mathcal{D}_k$.
5:     Let $\mathcal{A}_{k,1} \leftarrow \mathcal{D}_k$, $\ell \leftarrow 1$.
6:     while $\mathbf{a}_k$ is not specified do
7:         if $\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}} \leq \alpha$ for all $\mathbf{a} \in \mathcal{A}_{k,\ell}$ then
8:             Choose $\mathbf{a}_k \leftarrow \mathrm{argmax}_{\mathbf{a} \in \mathcal{A}_{k,\ell}} \langle \mathbf{a}, \widehat{\boldsymbol{\theta}}_{k,\ell} \rangle + \widehat{\beta}_{k,\ell} \|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}}$ and observe $r_k$.
9:             Keep the same index sets at all layers: $\Psi_{k+1,\ell'} \leftarrow \Psi_{k,\ell'}$ for all $\ell' \in [L]$.
10:        else if $\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}} \leq 2^{-\ell}$ for all $\mathbf{a} \in \mathcal{A}_{k,\ell}$ then
11:            $\mathcal{A}_{k,\ell+1} \leftarrow \big\{\mathbf{a} \in \mathcal{A}_{k,\ell} \big| \langle \mathbf{a}, \widehat{\boldsymbol{\theta}}_{k,\ell} \rangle \geq \max_{\mathbf{a}' \in \mathcal{A}_{k,\ell}} \langle \mathbf{a}', \widehat{\boldsymbol{\theta}}_{k,\ell} \rangle - 2 \cdot 2^{-\ell} \widehat{\beta}_{k,\ell} \big\}$.
12:        else
13:            Choose $\mathbf{a}_k$ such that $\|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}} > 2^{-\ell}$ and observe $r_k$.
14:            Compute the weight: $w_k \leftarrow 2^{-\ell} / \|\mathbf{a}_k\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}}$.
15:            Update the index sets: $\Psi_{k+1,\ell} \leftarrow \Psi_{k,\ell} \cup \{k\}$ and $\Psi_{k+1,\ell'} \leftarrow \Psi_{k,\ell'}$ for $\ell' \in [L] \backslash \{\ell\}$.
16:        end if
17:        $\ell \leftarrow \ell + 1$.
18:    end while
19:    For $\ell \in [L]$ such that $\Psi_{k+1,\ell} \neq \Psi_{k,\ell}$, update the estimators as follows:

           $\widehat{\mathbf{\Sigma}}_{k+1,\ell} \leftarrow \widehat{\mathbf{\Sigma}}_{k,\ell} + w_k^2 \mathbf{a}_k \mathbf{a}_k^\top, \widehat{\mathbf{b}}_{k+1,\ell} \leftarrow \widehat{\mathbf{b}}_{k,\ell} + w_k^2 \cdot r_k \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_{k+1,\ell} \leftarrow \widehat{\mathbf{\Sigma}}_{k+1,\ell}^{-1} \widehat{\mathbf{b}}_{k+1,\ell}.$

       Compute the adaptive confidence radius $\widehat{\beta}_{k+1,l}$for the next round according to (2.3).
20:    For $\ell \in [L]$ such that $\Psi_{k+1,\ell} = \Psi_{k,\ell}$, let $\widehat{\mathbf{\Sigma}}_{k+1,\ell} \leftarrow \widehat{\mathbf{\Sigma}}_{k,\ell}, \widehat{\mathbf{b}}_{k+1,\ell} \leftarrow \widehat{\mathbf{b}}_{k,\ell}, \widehat{\boldsymbol{\theta}}_{k+1,\ell} \leftarrow \widehat{\boldsymbol{\theta}}_{k,\ell}, \widehat{\beta}_{k+1,\ell} \leftarrow \widehat{\beta}_{k,\ell}$.
21: end for
```

their corresponding rewards. Starting from $\ell = 1$, the decision set $\mathcal{A}_{k,\ell}$ will keep 'shrinking' by eliminating all $\mathbf{a} \in \mathcal{A}_{k,\ell}$ which are unlikely to be the maximizer of $\langle \mathbf{a}, \boldsymbol{\theta}^* \rangle$ (notably, since $\boldsymbol{\theta}^*$ is unknown, here $\boldsymbol{\theta}^*$ needs to be replaced by $\widehat{\boldsymbol{\theta}}_{k,\ell}$, as displayed in Line 11). The elimination process will not stop until some action $\mathbf{a}$ with large uncertainty $\|\mathbf{a}\|_{\widehat{\mathbf{\Sigma}}_{k,\ell}^{-1}}$ emerges. Then Algorithm 1 will either select the action $\mathbf{a}$ with a large uncertainty (Line 13 to Line 15), or the action which maximizes the upper confidence bound of estimated reward if there is no action with a large uncertainty (Line 8 to Line 9). The context set $\Psi_{k,\ell}$ will be updated by appending k to it only when $\mathbf{a}_k$ enjoys a large uncertainty.

Construction of the estimate $\widehat{\boldsymbol{\theta}}_{k,\ell}$ The first difference between our algorithm and `SupLinUCB` is the construction of the estimate $\widehat{\boldsymbol{\theta}}_{k,\ell}$. Unlike the *unweighted* ridge regression estimator applied in

SupLinUCB, we employ a *weighted ridge-regression* estimator as follows

$$\forall k \in [K] \text{ and } \ell \in [L], \quad \widehat{\boldsymbol{\theta}}_{k,\ell} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i \in \Psi_{k,\ell}} w_i^2 \big(r_i - \langle \boldsymbol{\theta}, \mathbf{a}_i \rangle\big)^2 + 2^{-2\ell} \lambda \|\boldsymbol{\theta}\|_2^2,$$

where the weight w_i is chosen such that for $\forall \ell \in [L]$ and $i \in \Psi_{k,\ell}$, $\|w_i \mathbf{a}_i\|_{\widehat{\boldsymbol{\Sigma}}_{i,\ell}^{-1}} = 2^{-\ell}$. We explain here why we want to adopt such a weighted regression scheme. In particular, the estimate $\widehat{\boldsymbol{\theta}}_{k,\ell}$ can be regarded as $\boldsymbol{\mu}_k$ in Theorem 2.1. By our construction of w_i, we can ensure that the context $\mathbf{x}_i$ in Theorem 2.1, which is $w_i \mathbf{a}_i$ here, enjoys a uniform upper bound on the uncertainty, i.e., $\|\mathbf{x}_i\|_{\mathbf{Z}_i^{-1}} \le 2^{-\ell}$. Such a result can further imply that $\|\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}}$ is in the order of $O(2^{-\ell})$, which is tighter than the vanilla bound $O(1)$ deduced by previous works.

Remark 2.2 *It is worth noting that weighted ridge-regression technique has been used in heteroscedastic bandit setting (Kirschner and Krause, 2018; Zhou et al., 2021a; Zhou and Gu, 2022) for the known variance case. The most related work to ours is Zhou et al. (2021a), which applies the following weighted ridge-regression estimator* $\boldsymbol{\theta}_k = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^{k-1} \frac{1}{\sigma_i^2} \big(r_i - \langle \boldsymbol{\theta}, \mathbf{a}_i \rangle\big)^2 + \lambda \|\boldsymbol{\theta}\|_2^2$, *where the* $1/\sigma_i^2$ *weight is introduced to normalize the variance of noise. Our weight* w_i*, in contrast, is set to reweight the feature vectors such that they have the same elliptical norm* $\|\mathbf{a}_i\|_{\widehat{\boldsymbol{\Sigma}}_{i,\ell}^{-1}}$. *Weighted ridge-regression technique has also been applied to other bandit settings such as linear multi-resource allocation (Lattimore et al., 2015) and corruption-robust linear bandits (He et al., 2022b). In particular, He et al. (2022b) adopts a similar weight* $w_i = O\big(1/\|\mathbf{a}_i\|_{\widehat{\boldsymbol{\Sigma}}_i^{-1}}^{1/2}\big)$ *to balance the effect of adversarial corruption and stochastic noise. Nevertheless, the specific bandit problems they are solving are quite different from ours.*

Adaptive variance-aware exploration According to previous discussion, we can bound the estimation error of $\widehat{\boldsymbol{\theta}}_{k,\ell}$ following Theorem 2.1, which leads to a confidence bound of $\boldsymbol{\theta}^*$, i.e., $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,\ell}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}} \le \widetilde{O}(2^{-\ell} \cdot \sqrt{\sum_{i \in \Psi_{k,\ell}} w_i^2 \sigma_i^2} + 2^{-\ell} R)\}$. This can be used in the arm selection step (Line 8 and Line 11). However, such a confidence set requires the knowledge of variances σ_i^2 apriori. To address this issue, we need to replace σ_i^2 with their *empirical estimator*. In detail, since $\sigma_i^2 = \mathbb{E}[(r_i - \langle \boldsymbol{\theta}^*, \mathbf{a}_i \rangle)^2 | \mathbf{a}_{1:i-1}, \epsilon_{1:i-1}]$, we simply use an *one-point plug-in estimator* $(r_i - \langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \mathbf{a}_i \rangle)^2$. With such an estimator, we define the confidence radius $\beta_{k+1,\ell}$ at round $k+1$ and layer ℓ as

$$\begin{aligned} \widehat{\beta}_{k+1,\ell} := 16 \cdot 2^{-\ell} \sqrt{\Big(8 \widehat{\mathrm{Var}}_{k+1,\ell} + 6R^2 \log(4(k+1)^2 L/\delta) + 2^{-2\ell+4}\Big) \log(4k^2 L/\delta)} \\ + 6 \cdot 2^{-\ell} R \log(4k^2 L/\delta) + 2^{-\ell+1}, \end{aligned} \tag{2.3}$$

where $\widehat{\mathrm{Var}}_{k+1,\ell} := \begin{cases} \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \big(r_i - \langle \widehat{\boldsymbol{\theta}}_{k+1,\ell}, \mathbf{a}_i \rangle\big)^2, & 2^\ell \ge 64 \sqrt{\log\big(4(k+1)^2 L/\delta\big)} \\ R^2 \, |\Psi_{k+1,\ell}|, & \text{otherwise.} \end{cases}$

We would like to emphasize that although our used one-point estimator $(r_i - \langle \widehat{\boldsymbol{\theta}}_{k,l}, \mathbf{a}_i \rangle)^2$ might be an inaccurate estimator of the target σ_i^2 for some round i, the *weighted summation* of the one-point estimators $\sum w_i^2 (r_i - \langle \boldsymbol{\theta}_{k,l}, \mathbf{a}_i \rangle)^2$ actually serves as a sufficiently accurate estimator of the total variance $\sum w_i^2 \sigma_i^2$. That is because our employed weight can effectively 'calibrate' the term $(r_i - \langle \boldsymbol{\theta}_{k,l}, \mathbf{a}_i \rangle)^2$ and reduce its error, leading to an accurate estimate when these terms are summed.

2.4.2. COMPUTATIONAL COMPLEXITY

At each round $k \in [K]$, the learner executes the arm elimination step (Line 11 in Algorithm 1) for $O(L)$ times, and then applies Sherman-Morrison formula (Golub and Van Loan, 2013) and matrix multiplication to update the estimator in $O(d^2)$ time (Line 19). Note that we need to compute the confidence radius at each round in Line 19, which will take $O(k)$ time if we compute it directly. However, we can compute $\widehat{\mathrm{Var}}_{k+1,\ell}$ by $\widehat{\mathrm{Var}}_{k+1,\ell} = \sum_{i\in\Psi_{k+1,\ell}} w_i^2 r_i^2 - 2\widehat{\boldsymbol{\theta}}_{k+1,\ell}^\top \cdot \sum_{i\in\Psi_{k+1,\ell}} w_i r_i \mathbf{a}_i + \widehat{\boldsymbol{\theta}}_{k+1,\ell}^\top \Big(\sum_{i\in\Psi_{k+1,\ell}} w_i^2 \mathbf{a}_i \mathbf{a}_i^\top \Big) \widehat{\boldsymbol{\theta}}_{k+1,\ell}$, where the first term can be computed in $O(1)$ time at each round, the second term can be computed in $O(d)$ time by maintaining the prefix sum of $w_i r_i \cdot \mathbf{a}_i$ and the third term can be computed in $O(d^2)$ time by maintaining the value of the weighted covariance matrix. By adding these steps together, we can conclude that the time complexity of Algorithm 1 is $O(K|\mathcal{D}|Ld^2)$, where L is actually a logarithmic term according to the choice of α in Theorem 2.3.

2.5. Regret Bounds

We provide the regret guarantee of Algorithm 1 in the following theorem.

Theorem 2.3 *Suppose that for all $k \geq 1$ and all $\mathbf{a} \in \mathcal{D}_k$, $\|\mathbf{a}\|_2 \leq A$, $\|\boldsymbol{\theta}^*\|_2 \leq 1$, $\langle \mathbf{a}, \boldsymbol{\theta}^* \rangle \in [-1, 1]$. If $\{\beta_{k,\ell}\}_{k\geq 1, \ell\in[L]}$ is defined in (2.3) and $\alpha = 1/(R \cdot K^{3/2})$, then the cumulative regret of Algorithm 1 is bounded as follows with probability at least $1 - 3\delta$:*

$$\mathrm{Regret}(K) = \widetilde{O}\Bigg(d\sqrt{\sum_{k=1}^K \sigma_k^2} + dR + d \Bigg).$$

Remark 2.4 *If we treat R as a constant, the regret can be simplified as $\widetilde{O}\Big(d\sqrt{\sum_{k=1}^K \sigma_k^2} + d\Big)$. Compared with Weighted OFUL+ (Zhou and Gu, 2022), our algorithm achieves the same order of regret guarantee and does not require any prior knowledge about the variance σ_k. Compared with VOFUL2 (Kim et al., 2021), our `SAVE` algorithm improves the regret from $\widetilde{O}\Big(d^{1.5}\sqrt{\sum_{k=1}^K \sigma_k^2} + d^2\Big)$ to $\widetilde{O}\Big(d\sqrt{\sum_{k=1}^K \sigma_k^2} + d\Big)$. Furthermore, `VOFUL2` needs to perform the arm elimination for each possible direction $\boldsymbol{\mu}$ in the d-dimension unit ball, which requires an exponential computational time (See the discussion in Section 2.2).*

Remark 2.5 *Consider the deterministic reward setting where $\sigma_k = 0$ holds for all round $k \in [K]$. If we treat R as a constant, then Theorem 2.3 suggests an $\widetilde{O}(d)$ regret guarantee, which matches the $\Omega(d)$ lower bound up to logarithmic factors (Chu et al., 2011).*

3. Variance-Aware Learning for Linear Mixture MDPs

In this section, we apply the techniques developed in Section 2 to reinforcement learning, and propose a variance-aware algorithm for linear mixture MDPs.

3.1. Problem Setup

Episodic MDPs. A time-homogenous episodic MDP (Puterman, 2014) is denoted by a tuple $M = M(\mathcal{S}, \mathcal{A}, H, r, \mathbb{P})$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is a finite action space, H is the planning horizon (i.e., length of each episode), $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ is a deterministic reward function, $\mathbb{P}(s'|s, a)$ is the transition probability function denoting the probability of transition from state s to state s' under

Algorithm 2 UCRL-AVE

```
Require: Regularization parameter $\lambda > 0$, $\alpha > 0$, $B$, an upper bound on the $\ell_2$-norm of $\boldsymbol{\theta}^*$.
Set $L = \lceil \log_2(1/\alpha) \rceil$.
Initialize: $\widehat{\boldsymbol{\Sigma}}_{0,H+1,\ell} \leftarrow 2^{-2\ell}\lambda \cdot \mathbf{I}$, $\widehat{\mathbf{b}}_{0,H+1,\ell} \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_{1,\ell} \leftarrow \mathbf{0}$ for all $\ell \in [L]$.
for $k = 1, \ldots, K$ do
    $V_{k,H+1}(\cdot) \leftarrow 0$.
    Update the current estimators: $\widehat{\boldsymbol{\Sigma}}_{k,1,\ell} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k-1,H+1,\ell}$, $\widehat{\mathbf{b}}_{k,1,\ell} \leftarrow \widehat{\mathbf{b}}_{k-1,H+1,\ell}$, $\widehat{\boldsymbol{\theta}}_{k,\ell} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}\widehat{\mathbf{b}}_{k,1,\ell}$ for all $\ell \in [L]$.
    Compute $\widehat{\beta}_{k,\ell}$ according to (3.4).
    for $h = H, \ldots, 1$ do
        $Q_{k,h}(\cdot,\cdot) \leftarrow \min\Big\{1, \min_{\ell\in[L]}\Big[r(\cdot,\cdot) + \big\langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\rangle + \widehat{\beta}_{k,\ell}\big\|\boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}}\Big]\Big\}$.
        $\pi_k(\cdot,h) \leftarrow \mathrm{argmax}_{a\in\mathcal{A}} Q_{k,h}(\cdot,a)$,   $V_{k,h}(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q_{k,h}(\cdot,a)$.
    end for
    Observe $s_1^k$.
    for $h = 1, \ldots H$ do
        Take action $a_h^k \leftarrow \pi_k(s_h^k, h)$ and observe $s_{h+1}^k$.
        $\mathcal{L}_{k,h} \leftarrow \Big\{\ell \in [L] \Big| \big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\big\|_{\boldsymbol{\Sigma}_{k,h,\ell}^{-1}} \geq 2^{-\ell}\Big\}$.
        Set $\ell_{k,h} \leftarrow \begin{cases} L+1, & \mathcal{L}_{k,h} = \varnothing \\ \min(\mathcal{L}_{k,h}), & \text{otherwise} \end{cases}$.
        if $\ell_{k,h} \neq L+1$ then
            $w_{k,h} \leftarrow 2^{-\ell_{k,h}} / \big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\big\|_{\boldsymbol{\Sigma}_{k,h,\ell_{k,h}}^{-1}}$.
            $\widehat{\boldsymbol{\Sigma}}_{k,h+1,\ell_{k,h}} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,h,\ell_{k,h}} + w_{k,h}^2 \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)^\top$.
            $\widehat{\mathbf{b}}_{k,h+1,\ell_{k,h}} \leftarrow \widehat{\mathbf{b}}_{k,h,\ell_{k,h}} + w_{k,h}^2 V_{k,h+1}(s_{h+1}^k)\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$.
        end if
        $\widehat{\boldsymbol{\Sigma}}_{k,h+1,\ell} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,h,\ell}$, $\widehat{\mathbf{b}}_{k,h+1,\ell} \leftarrow \widehat{\mathbf{b}}_{k,h,\ell}$ for all $\ell \in [L]$ and $\ell \neq \ell_{k,h}$.
    end for
end for
```

action a. A policy $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ is a function which maps a state s and the stage number h to an action a. For any policy π and stage $h \in [H]$, we define the following action-value function $Q_h^\pi(s,a)$ and value function $V_h^\pi(s)$ as follows

$$Q_h^\pi(s,a) = r(s,a) + \mathbb{E}\bigg[\sum_{h'=h+1}^{H} r\big(s_{h'}, \pi(s_{h'}, h')\big)\bigg| s_h = s, a_h = a\bigg], \quad V_h^\pi(s) = Q_h^\pi(s, \pi(s,h)),$$

where $s_{h'+1} \sim \mathbb{P}(\cdot|s_{h'}, a_{h'})$. We further define the optimal value function V_h^* and the optimal action-value function Q_h^* as $V_h^*(s) = \max_\pi V_h^\pi(s)$ and $Q_h^*(s,a) = \max_\pi Q_h^\pi(s,a)$. In addition, for any function $V : \mathcal{S} \to \mathbb{R}$, we denote $[\mathbb{P}V](s,a) = \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}V(s')$. Therefore, for each stage $h \in [H]$ and policy π, we have the following Bellman equation, as well as the Bellman optimality equation:

$$Q_h^\pi(s,a) = r(s,a) + [\mathbb{P}V_{h+1}^\pi](s,a), \quad Q^*(s,a) = r(s,a) + [\mathbb{P}V_{h+1}^*](s,a),$$

where $V_{H+1}^{\pi}(\cdot) = V_{H+1}^{*}(\cdot) = 0$. At the beginning of episode k, the agent chooses a policy π to guide its actions throughout the episode. At each stage $h \in [H]$, the agent observes the state s_h^k, chooses an action by the policy π and observes the next state with $s_{h+1}^k \sim \mathbb{P}(\cdot|s_h^k, a_h^k)$.

Following previous work on horizon-free regret in linear mixture MDPs (Zhang et al., 2021b; Kim et al., 2021; Zhou and Gu, 2022), we consider the setting where the total reward (i.e., return of an episode) is bounded by 1.

Assumption 3.1 *For any policy π, let $(s_h, a_h)_{h=1}^H$ be one trajectory following π, then $\sum_{h\in[H]} r(s_h, a_h) \le 1$ almost surely.*

For simplicity, let $[\mathbb{V}V](s,a) = \left[\mathbb{P}V^2\right](s,a) - ([\mathbb{P}V](s,a))^2$ denote the conditional variance of V conditioned on (s,a). We define the following instance-dependent quantity:

$$\mathrm{Var}_K^* = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}V_{h+1}^*](s_h^k, a_h^k). \tag{3.1}$$

The quantity (3.1) characterizes the stochasticity of the MDP under the optimal policy. For a deterministic MDP where the transition function is deterministic , we have $\mathrm{Var}_K^* = 0$. Similar quantities have been considered in Maillard et al. (2014); Zanette and Brunskill (2019), and the same quantity has been proposed by a concurrent work (Zhou et al., 2023) on tabular RL.

Linear Mixture MDPs. We consider a special MDP class called *linear mixture MDPs*.

Definition 3.2 (Episodic linear mixture MDPs, Jia et al. 2020; Ayoub et al. 2020) *An episodic MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, r, \mathbb{P})$ is a homogeneous, episodic B-bounded linear mixture MDP if there exists vectors $\boldsymbol{\theta}^* \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}^*\|_2 \le B$ and $\boldsymbol{\phi}(\cdot|\cdot,\cdot)$ satisfying (3.2), such that for each $(s,a) \in \mathcal{S}\times\mathcal{A}$, $s' \in \mathcal{S}$ and stage $h \in [H]$, $\mathbb{P}(s'|s,a) = \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta}^*\rangle$. Moreover, $\boldsymbol{\phi}$ satisfies that for any bounded function $V : \mathcal{S} \to [0,1]$ and any tuple $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\|\boldsymbol{\phi}_V(s,a)\|_2 \le 1, \textit{where } \boldsymbol{\phi}_V(s,a) := \sum_{s'\in\mathcal{S}} \boldsymbol{\phi}(s'|s,a)V(s'). \tag{3.2}$$

The goal of the agent is to minimize the following cumulative regret at the first K rounds:

$$\mathrm{Regret}(K) = \sum_{k\in[K]} \left[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right].$$

3.2. The Proposed Algorithm

We present an adaptive variance-aware algorithm named UCRL with **A**daptive **V**ariance-Aware **E**xploration (`UCRL-AVE`) in Algorithm 2. The backbone of our algorithm is the *value-targeted-regression* scheme proposed by `UCRL-VTR` (Jia et al., 2020; Ayoub et al., 2020). In detail, Algorithm 2 aims to estimate the optimal value function Q_h^* by $Q_{k,h}$, utilizing the Bellman optimal equation. Since $\mathbb{P}V_{k,h+1}$ is not tractable ($\mathbb{P}$ is unknown), Algorithm 2 uses the fact that $\mathbb{P}V_{k,h+1}(s,a) = \langle \boldsymbol{\phi}_{V_{k,h+1}}(s,a), \boldsymbol{\theta}^*\rangle$ is a linear function of the feature $\boldsymbol{\phi}_{V_{k,h+1}}(s,a)$, and estimates $\mathbb{P}V_{k,h+1}(s,a)$ by a plug-in estimator $\langle \boldsymbol{\phi}_{V_{k,h+1}}(s,a), \widehat{\boldsymbol{\theta}}_k\rangle$, where $\widehat{\boldsymbol{\theta}}_k$ is the estimate of $\boldsymbol{\theta}^*$. Then `UCRL-VTR` computes $Q_{k,h}$ by the *upper confidence bound* of the empirical estimator with truncation (Line 8).

The main difference between Algorithm 2 and `UCRL-VTR` is the construction of $\widehat{\boldsymbol{\theta}}_k$: instead of using a single estimate, Algorithm 2 maintains L estimates $\widehat{\boldsymbol{\theta}}_{k,\ell}$, constructed on a multi-layer structure of feature vectors. We highlight several important technical innovations here.

Multi-layer structure of feature vectors. We first demonstrate how Algorithm 2 utilizes L number of estimates $\widehat{\boldsymbol{\theta}}_{k,\ell}$ to build the value function estimate $Q_{k,h}$, then we show how Algorithm 2 updates $\widehat{\boldsymbol{\theta}}_{k,\ell}$ accordingly. Algorithm 2 constructs $Q_{k,h}$ as the minimum of L optimistic estimates computed by $\langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{k,h+1}} \rangle$. The minimum step makes the estimate $Q_{k,h}$ tighter than that in `UCRL-VTR`.

Similar to Algorithm 1, $\widehat{\boldsymbol{\theta}}_{k,l}$ is the solution to some regression problem over the features $\phi_V(s,a)$ and their corresponding target values. For simplicity, we define the following subsets of $[K] \times [H]$:

$$\Psi_{k,\ell} = \{(i,h) \in [k-1] \times [H] | \ell_{i,h} = \ell\}, \qquad \text{for } k \in [K+1], \ell \in [L+1], \tag{3.3}$$

which represents the indices of feature vectors in layer ℓ at the beginning of round k. Note that $\widehat{\boldsymbol{\theta}}_{k,\ell}$ will be updated if the feature $\phi_{V^k_{h+1}}(s^k_h, a^k_h)$ is added to the feature set $\Psi_{k,\ell}$. The rule that whether to add such a feature or not is based on the uncertainty of $\phi_{V^k_{h+1}}(s^k_h, a^k_h)$ within the feature set $\Psi_{k,\ell}$, which is similar to the multi-layer structure adopted by He et al. (2021b) for uniform-PAC bounds in linear MDPs. Finally, $\widehat{\boldsymbol{\theta}}_{k,\ell}$ is computed as the solution to the weighted regression problem over the feature set $\Psi_{k,\ell}$, where the weight $w_{k,h}$ is selected to guarantee that $\|w_{k,h}\boldsymbol{\phi}_{V_{k,h+1}}(s^k_h, a^k_h)\|_{\boldsymbol{\Sigma}^{-1}_{k,h,\ell_{k,h}}} = 2^{-\ell_{k,h}}$, similar to that in Algorithm 1.

Adaptive variance-aware exploration. Similar to Algorithm 1, we will also face the problem to construct a confidence set of $\boldsymbol{\theta}^*$ *without* knowing the variance of value functions $V_{k,h+1}$. Here we take the same approach: to replace the variance of $V_{k,h+1}$, $\mathbb{P}[V_{k,h+1} - \mathbb{P}V_{k,h+1}]^2$ with its one-point empirical estimate $(V_{i,h+1}(s^i_{h+1}) - \langle \widehat{\boldsymbol{\theta}}_{k,l}, \boldsymbol{\phi}_{V_{i,h+1}}(s^i_h, a^i_h)\rangle)^2$. In detail, the confidence radius is:

$$\begin{aligned}\widehat{\beta}_{k,\ell} := 16 \cdot 2^{-\ell} &\sqrt{\big(8\widehat{\mathrm{Var}}_{k,\ell} + 8\log(4k^2H^2L/\delta) + 2^{-2\ell+5} \cdot \lambda B^2\big)\log(4k^2H^2L/\delta)} \\ &+ 6 \cdot 2^{-\ell}\log(4k^2H^2L/\delta) + 2^{-\ell}\sqrt{\lambda} \cdot B,\end{aligned} \tag{3.4}$$

where $\widehat{\mathrm{Var}}_{k,\ell} = \begin{cases} 8\sum_{(i,h)\in\Psi_{k,\ell}} w^2_{i,h}\big(V_{i,h+1}(s^i_{h+1}) - \langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{i,h+1}}(s^i_h, a^i_h)\rangle\big)^2, & 2^\ell \geq 64\sqrt{\log(4k^2H^2L/\delta)}, \\ |\Psi_{k,\ell}|, & \text{otherwise.} \end{cases}$

We can adopt the method discussed in Subsection 2.4.2 to compute $\widehat{\mathrm{Var}}_{k,\ell}$ in an efficient way. We call the construction of the confidence set along with its radius $\widehat{\beta}_{k,\ell}$ as *adaptive variance-aware exploration*.

Compared with `UCRL-VTR+` (Zhou et al., 2021a) and `HF-UCRL-VTR+` (Zhou and Gu, 2022), our algorithm does not need to estimate the conditional variance using another ridge regression estimator on the second-order moment of value functions. Furthermore, in contrast to `HF-UCRL-VTR+` (Zhou and Gu, 2022), our algorithm does not explicitly estimate the high-order moments of value functions. Thus, our algorithm is much simpler. It is also worth noting that the multi-layer structure in Algorithm 2 is an alternative of the `SupLinUCB`-type design in Algorithm 1. Since linear bandits can be seen as a special case of linear mixture MDPs, Algorithm 2 implies another algorithm for heteroscedastic linear bandits, which enjoys the same regret guarantee as Algorithm 1.

3.3. Regret Bounds

We provide the regret guarantee of Algorithm 2 in the following theorem.

Theorem 3.3 *Set $\widehat{\beta}_{k,\ell}$ as in (3.4), $\alpha = 1/(KH)^{3/2}$ and $\lambda = 1/B^2$ in Algorithm 2. Then with probability at least $1 - (4\lceil \log_2 2HK \rceil + 9)\delta$, the regret of Algorithm 2 is bounded by:*

$$\mathrm{Regret}(K) = \widetilde{O}\Big(d\sqrt{\mathrm{Var}^*_K} + d^2\Big).$$

Corollary 3.4 *Under the same conditions as Theorem 3.3, with probability at least* $1-(4\lceil\log_2 2HK\rceil+10)\delta$*, the regret of Algorithm 2 is bounded by:* $\mathrm{Regret}(K)=\widetilde{O}\big(d\sqrt{K}+d^2\big)$.

Remark 3.5 *Our regret given by Theorem 3.3 is variance-dependent, which means that the regret of* `UCRL-AVE` *is smaller when the conditional variance of the optimistic value function is smaller. In the deterministic case where all the transitions in the MDP is deterministic, our regret reduces to* $\widetilde{O}(d^2)$*, with only a logarithmic dependence on* K*. Additionally, the regret in Corollary 3.4 matches the regret of* `HF-UCRL-VTR+` *proposed by Zhou and Gu (2022), which is the worst-case regret and matches the minimax lower bound (Zhou and Gu, 2022).*

4. Conclusion and Future Work

In this paper, we consider variance-aware learning in linear bandits and linear mixture MDPs. We propose a computationally efficient algorithm `SAVE` for heteroscedastic linear bandits, which achieves a variance-dependent regret, matching the minimax regret bounds in both the worst case and the deterministic reward case. For linear mixture MDPs, we further extend our techniques and propose an algorithm dubbed `UCRL-AVE`, attaining a tighter problem-dependent horizon-free regret bound. We leave for future work the generalization of our work to RL with nonlinear function approximation.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. HZ, DZ, JH and QG are supported in part by the National Science Foundation CAREER Award 1906169 and the Sloan Research Fellowship. TZ is supported in part by the GRF 16201320. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

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

Naoki Abe, Alan W Biermann, and Philip M Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. *arXiv preprint arXiv:2212.06069*, 2022.

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

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

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182*, 2020.

Nicolo Cesa-Bianchi and Paul Fischer. Finite-time regret bounds for the multiarmed bandit problem. In *ICML*, volume 98, pages 100–108. Citeseer, 1998.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

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

Yan Dai, Ruosong Wang, and Simon S Du. Variance-aware sparse linear bandits. *arXiv preprint arXiv:2205.13450*, 2022.

Varsha Dani, Thomas P. Hayes, and S. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, 2008.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. In *Neural Information Processing Systems*, 2021.

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

David A Freedman. On tail probabilities for martingales. pages 100–118. JSTOR, 1975.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021a.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Uniform-pac bounds for reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:14188–14199, 2021b.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132*, 2022a.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2205.06811*, 2022b.

Pihe Hu, Yu Chen, and Longbo Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 8971–9019. PMLR, 2022.

Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR, 2020.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.

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

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020a.

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

Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *arXiv preprint arXiv:2111.03289*, 2021.

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Linear multi-resource allocation with semi-bandit feedback. *Advances in Neural Information Processing Systems*, 28, 2015.

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

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

Yuanzhi Li, Ruosong Wang, and Lin F Yang. Settling the horizon-dependence of sample complexity in reinforcement learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 965–976. IEEE, 2022.

Odalric-Ambrym Maillard, Timothy A Mann, and Shie Mannor. How hard is my mdp?" the distribution-norm to the rescue". *Advances in Neural Information Processing Systems*, 27, 2014.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58 (6):527–535, 1952.

Max Simchowitz and Kevin G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Neural Information Processing Systems*, 2019.

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

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429. PMLR, 2022.

Ruosong Wang, Simon S Du, Lin Yang, and Sham Kakade. Is long horizon rl more difficult than short horizon rl? *Advances in Neural Information Processing Systems*, 33:9075–9085, 2020a.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020b.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pages 4438–4472. PMLR, 2021.

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

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

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

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020a.

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

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021a.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021b.

Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pages 3858–3904. PMLR, 2022.

Heyang Zhao, Dongruo Zhou, Jiafan He, and Quanquan Gu. Bandit learning with general function classes: Heteroscedastic noise and variance-dependent regret bounds. *arXiv preprint arXiv:2202.13603*, 2022.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. In *Advances in Neural Information Processing Systems*, 2022.

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

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.

Runlong Zhou, Zihan Zhang, and Simon S Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. *arXiv preprint arXiv:2301.13446*, 2023.

Appendix A. Additional Related Work

Horizon-free regret in tabular RL. RL is considered to be more challenging than contextual bandits due to the non-trivial planning horizon and uncertain state transitions. Jiang and Agarwal (2018) conjectured that any algorithm seeking an ϵ-optimal policy for tabular RL, where the total reward is bounded by 1, would require a sample complexity with a polynomial dependence on the planning horizon H. However, this conjecture was disproven by Wang et al. (2020a), who introduced a horizon-free algorithm with a sample complexity of $\widetilde{O}(|\mathcal{S}|^5|\mathcal{A}|^4\epsilon^{-2}\text{polylog}(H))$ that only has a polylogarithmic dependence on H. Zhang et al. (2021a) then proposed a near-optimal algorithm with a regret of $O((\sqrt{|\mathcal{S}||\mathcal{A}|K} + |\mathcal{S}|^2|\mathcal{A}|)\text{polylog}(H))$ and a similar sample complexity. Later, Li et al. (2022) and Zhang et al. (2022) presented algorithms with sample complexity guarantees that are independent of H.

Heteroscedastic linear bandits. The worst-case regret of linear bandits has been extensively studied (Auer, 2002; Dani et al., 2008; Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Li et al., 2019). Recently, there is a series of works considering a heteroscedastic variant of the classic linear bandit problem where the noise distribution is assumed to vary over time. Kirschner and Krause (2018) is the first to formally propose linear bandit model with heteroscedastic noise. In their setting, the noise at round $k \in [K]$ is assumed to be σ_k-sub-Gaussian. Some recent works relaxed the sub-Gaussian assumption in the sense that the noise at the k-th round is assumed to be of variance σ_k^2 instead of σ_k-sub-Gaussian (Zhou et al., 2021a; Zhang et al., 2021b; Kim et al., 2021; Zhou and Gu, 2022; Dai et al., 2022). Among these works, Zhou et al. (2021a) and Zhou and Gu (2022) considered known-variance case where σ_k is observed by the learner after the k-th round, while Zhang et al. (2021b); Kim et al. (2021) proposed statistically efficient but computationally inefficient algorithm for the unknown-variance case. Dai et al. (2022) considered a more specific model, heteroscedastic sparse linear bandits, and proposed a general framework which converts any heteroscedastic linear bandit algorithm to an algorithm for heteroscedastic sparse linear bandits.

RL with linear function approximation. There is a huge body of literature on RL with linear function approximation (Jiang et al., 2017; Dann et al., 2018; Yang and Wang, 2019; Jin et al., 2020b; Wang et al., 2020b; Du et al., 2019; Sun et al., 2019; Zanette et al., 2020a,b; Weisz et al., 2021; Yang and Wang, 2020; Modi et al., 2020; Ayoub et al., 2020; Zhou et al., 2021a; He et al., 2021a; Zhou and Gu, 2022). Several types of assumption on the linear structure of the underlying MDPs have been made in these works, including the *linear MDP* assumption (Yang and Wang, 2019; Jin et al., 2020b; Hu et al., 2022; He et al., 2022a; Agarwal et al., 2022), the *low Bellman-rank* assumption (Jiang et al., 2017), the *low inherent Bellman error* assumption (Zanette et al., 2020b), and the *linear mixture MDP* assumption (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021a). In this paper, we focus on linear mixture MDPs, where the transition probability function is assumed to be a linear function of a known feature mapping over the state-action-next-state triplet. Recently, there is a line of works aiming for attaining horizon-free regret bounds (Zhang et al., 2021b; Kim et al., 2021; Zhou and Gu, 2022), which are most related to our work.

Appendix B. Proof of Theorem 2.1

Proof For simplicity, we introduce the following definitions:

$$\mathbf{d}_0 = 0, \mathbf{d}_k = \sum_{i=1}^{k} \mathbf{x}_i\eta_i, q_0 = 0, q_k = \|\mathbf{d}_k\|_{\mathbf{Z}_k^{-1}}, \mathcal{I}_k = \mathbb{1}\{\forall\, 0 \le s \le k, q_s \le \beta_s\},$$

where $k \geq 1$ and we further define $\beta_0 = 0, \mathcal{I}_0 = 1$. According to these definitions, the term q_k can be upper bounded by the following decomposition:

$$\begin{aligned}
q_k^2 &= (\mathbf{d}_{k-1} + \mathbf{x}_k\eta_k)^\top \mathbf{Z}_k^{-1}(\mathbf{d}_{k-1} + \mathbf{x}_k\eta_k)\\
&= \mathbf{d}_{k-1}^\top \mathbf{Z}_k^{-1}\mathbf{d}_{k-1} + \underbrace{2\eta_k\mathbf{x}_k^\top \mathbf{Z}_k^{-1}\mathbf{d}_{k-1}}_{I_{1,k}} + \underbrace{\eta_k^2\mathbf{x}_k^\top \mathbf{Z}_k^{-1}\mathbf{x}_k}_{I_{2,k}}\\
&\leq q_{k-1}^2 + I_{1,k} + I_{2,k},
\end{aligned} \tag{B.1}$$

where the inequality holds since $\mathbf{Z}_k = \mathbf{Z}_{k-1} + \mathbf{x}_k\mathbf{x}_k^\top \succeq \mathbf{Z}_{k-1}$. For the term $I_{1,k}$, by the matrix inversion lemma, we have the following equation:

$$\begin{aligned}
I_{1,k} &= 2\eta_k\left(\mathbf{x}_k^\top \mathbf{Z}_{k-1}^{-1}\mathbf{d}_{k-1} - \frac{\mathbf{x}_k\mathbf{Z}_{k-1}^{-1}\mathbf{x}_k\mathbf{x}_k^\top \mathbf{Z}_{k-1}^{-1}\mathbf{d}_{k-1}}{1 + \|\mathbf{x}_k\|_{\mathbf{Z}_{k-1}^{-1}}^2}\right)\\
&= 2\eta_k\left(\mathbf{x}_k^\top \mathbf{Z}_{k-1}^{-1}\mathbf{d}_{k-1} - \frac{\|\mathbf{x}_k\|_{\mathbf{Z}_{k-1}^{-1}}^2\mathbf{x}_k^\top \mathbf{Z}_{k-1}^{-1}\mathbf{d}_{k-1}}{1 + \|\mathbf{x}_k\|_{\mathbf{Z}_{k-1}^{-1}}^2}\right)\\
&= 2\eta_k \cdot \frac{\mathbf{x}_k^\top \mathbf{Z}_{k-1}^{-1}\mathbf{d}_{k-1}}{1 + \|\mathbf{x}_k\|_{\mathbf{Z}_{k-1}^{-1}}^2}.
\end{aligned}$$

Taking a summation over the term $I_{1,k}$ with respect to the indicator function $\mathcal{I}_{k-1}$, we have the following equation:

$$\sum_{i=1}^k I_{1,i} \cdot \mathcal{I}_{i-1} = 2\sum_{i=1}^k \eta_i \cdot \frac{\mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1 + \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}^2}\mathcal{I}_{i-1}. \tag{B.2}$$

Now, we can derive an upper bound for this summation by Freedman's inequality. In detail, for each round $i \in [k]$, we have

$$\left|\eta_i \cdot \frac{\mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1 + \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}^2}\mathcal{I}_{i-1}\right| \leq R\left|\frac{\|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\|\mathbf{d}_{i-1}\|_{\mathbf{Z}_{i-1}^{-1}}}{1 + \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}^2}\right|\mathcal{I}_{i-1} \leq R\left|\frac{\|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\beta_{i-1}}{1 + \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}^2}\right| \leq R\beta_k\rho,$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to the definition of indicator function $\mathcal{I}_{i-1}$ and the last inequality holds due to $\rho \geq \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}$. In addition, for each round $i \in [k]$, we have

$$\mathbb{E}\left[\eta_i \cdot \frac{\mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1 + \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}^2}\mathcal{I}_{i-1}\middle|\mathcal{G}_i\right] = 0,$$

and the summation of variance is upper bounded by

$$\sum_{i=1}^k \mathbb{E}\left[\left(\eta_i \cdot \frac{\mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1 + \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}^2}\mathcal{I}_{i-1}\right)^2\middle|\mathcal{G}_i\right] = \sum_{i=1}^k \left(\frac{\mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1 + \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}^2}\mathcal{I}_{i-1}\right)^2 \mathbb{E}[\eta_i^2|\mathcal{G}_i]$$

$$
\begin{aligned}
&\le \sum_{i=1}^k \bigg(\frac{\|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}} \|\mathbf{d}_{i-1}\|_{\mathbf{Z}_{i-1}^{-1}} \mathcal{I}_{i-1}}{1+\|\mathbf{x}_i\|^2_{\mathbf{Z}_{i-1}^{-1}}} \bigg)^2 \mathbb{E}[\eta_i^2|\mathcal{G}_i] \\
&\le \sum_{i=1}^k \bigg(\frac{\|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}} \beta_{i-1}}{1+\|\mathbf{x}_i\|^2_{\mathbf{Z}_{i-1}^{-1}}} \bigg)^2 \mathbb{E}[\eta_i^2|\mathcal{G}_i] \\
&\le \beta_k^2 \rho^2 \sum_{i=1}^k \mathbb{E}[\eta_i^2|\mathcal{G}_i] \\
&\le \beta_k^2 \rho^2 v_k.
\end{aligned}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to the definition of indicator function $\mathcal{I}_{i-1}$, the third inequality holds due to $\rho \ge \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}$ and the last inequality holds due to $\sum_{i=1}^k \mathbb{E}[\eta_i^2|\mathcal{G}_i] \le v_k$.

Therefore, using Freedman's inequality, for any $k \ge 1$, with probability $1-\delta/(4k^2)$, we have

$$
\begin{aligned}
\sum_{i=1}^k I_{1,i} \cdot \mathcal{I}_{i-1} &\le 2\sqrt{2\beta_k^2\rho^2 v_k \log(4k^2/\delta)} + 4/3 \cdot R\beta_k\rho \log(4k^2/\delta) \\
&\le \frac{1}{4}\beta_k^2 + 8\left(\rho^2 v_k \log(4k^2/\delta)\right) + \frac{1}{4}\beta_k^2 + (16/9)\cdot R^2\rho^2[\log(4k^2/\delta)]^2 \\
&\le \frac{3}{4}\beta_k^2,
\end{aligned}
$$

where the first inequality holds due to Lemma E.4 and the second inequality holds due to Young's inequality. After taking a union bound for all $k > 1$, it can then be further deduced that with probability $1-\delta/2$, for all $k \ge 1$, we have

$$
\sum_{i=1}^k I_{1,i} \cdot \mathcal{I}_{i-1} \le \frac{3}{4}\beta_k^2. \tag{B.3}
$$

For simplicity, let $\mathcal{E}_{I_1}$ be the events that (B.3) holds. Then we bound the summation of $I_{2,k}$ over k through the following calculation:

$$
\sum_{i=1}^k I_{2,i} \le \sum_{i=1}^k \eta_i^2\rho^2 = \rho^2 v_k + \rho^2 \sum_{i=1}^k \left[\eta_i^2 - \mathbb{E}[\eta_i^2|\mathcal{G}_i]\right], \tag{B.4}
$$

where the first inequality holds due to $\rho \ge \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}$. Still, we can bound the second term in (B.4) using Freedman's inequality in Lemma E.4. Notice that, for each round $i \in [k]$, we have

$$
\begin{aligned}
\big|\mathbb{E}[\eta_i^2|\mathcal{G}_i]\big] - \eta_i^2\big| \le R^2, \mathbb{E}\Big[\big(\mathbb{E}[\eta_i^2|\mathcal{G}_i] - \eta_i^2\big)|\mathcal{G}_i\Big] = 0, \\
\mathbb{E}\Big[\big(\mathbb{E}[\eta_i^2|\mathcal{G}_i] - \eta_i^2\big)^2|\mathcal{G}_i\Big] = \big(\mathbb{E}[\eta_i^2|\mathcal{G}_i]\big)^2 - 2\mathbb{E}[\eta_i^2|\mathcal{G}_i]\cdot\mathbb{E}[\eta_i^2|\mathcal{G}_i] + \mathbb{E}[\eta_i^4|\mathcal{G}_i] \le R^2\mathbb{E}[\eta_i^2|\mathcal{G}_i],
\end{aligned}
$$

According to Freedman's Inequality, for any k, with probability $1-\delta/(4k^2)$, we have

$$
\sum_{i=1}^k \left(\eta_i^2 - \mathbb{E}[\eta_i^2|\mathcal{G}_i]\right) \le \sqrt{2R^2\log(4k^2/\delta)v_k} + 2/3\cdot R^2\log(4k^2/\delta). \tag{B.5}
$$

Taking a union bound over all round $k \geq 1$, with probability at least $1-\delta/2$, for all $k \geq 1$, we have

$$\begin{aligned}\sum_{i=1}^{k} I_{2,k} &\leq \rho^2 v_k + \rho^2 \sum_{i=1}^{k}\big[\eta_i^2 - \mathbb{E}[\eta_i^2|\mathcal{G}_i]\big]\\ &\leq \rho^2 v_k + \rho^2 R\sqrt{2v_k\log(4k^2/\delta)} + \frac{2}{3}\cdot\rho^2\cdot R^2\log(4k^2/\delta)\\ &\leq 3\left(\rho^2 v_k\log(4k^2/\delta)\right) + 2R^2\rho^2[\log(4k^2/\delta)]^2\\ &\leq \frac{1}{4}\beta_k^2. \end{aligned} \tag{B.6}$$

where the first inequality holds due to (B.4), the second inequality holds due to (B.5) and the third inequality holds due to Young's inequality. For simplicity, let $\mathcal{E}_{I_2}$ be the events that (B.6) holds. In the remaining proof, we assume that events $\mathcal{E}_{I_1}$ and $\mathcal{E}_{I_2}$ holds, whose probability is no less than $1-\delta$ by the union bound. Under this situation, for any round $k \geq 0$, if $\mathcal{I}_{i-1} = 1$ holds for all $i \in [k]$, then according to (B.1), we have

$$\begin{aligned} q_{k+1}^2 &\leq \sum_{i=1}^{k+1} I_{1,i} + \sum_{i=1}^{k+1} I_{2,i}\\ &= \sum_{i=1}^{k+1} I_{1,i}\cdot\mathcal{I}_{i-1} + \sum_{i=1}^{k+1} I_{2,i}\\ &\leq \beta_{k+1}^2, \end{aligned}$$

where the last inequality holds due to the definition of events $\mathcal{E}_{I_1}$ and $\mathcal{E}_{I_2}$. This result indicates that $\mathcal{E}_{k+1} = 1$. Therefore, by induction, we can deduce that with probability at least $1-\delta$, for all $k \geq 1$, we have

$$\left\|\sum_{i=1}^{k}\mathbf{x}_i\eta_i\right\|_{\mathbf{Z}_k^{-1}} \leq \beta_k.$$

Furthermore, the estimation error between underlying vector $\boldsymbol{\mu}^*$ and estimator $\boldsymbol{\mu}_k$ can be upper bounded by:

$$\begin{aligned}\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_k} &= \|\mathbf{Z}_k^{-1}\mathbf{b}_k - \mathbf{Z}_k^{-1}\mathbf{Z}_k\boldsymbol{\mu}^*\|_{\mathbf{Z}_k}\\ &= \left\|\mathbf{Z}_k^{-1}\mathbf{b}_k - \mathbf{Z}_k^{-1}\sum_{i=1}^{k}\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\mu}^* - \lambda\mathbf{Z}_k^{-1}\boldsymbol{\mu}^*\right\|_{\mathbf{Z}_k}\\ &= \left\|\mathbf{Z}_k^{-1}\sum_{i=1}^{k}\mathbf{x}_i(y_i - \mathbf{x}_i^\top\boldsymbol{\mu}^*) - \lambda\mathbf{Z}_k^{-1}\boldsymbol{\mu}^*\right\|_{\mathbf{Z}_k}\\ &\leq \left\|\sum_{i=1}^{k}\mathbf{x}_i\eta_i\right\|_{\mathbf{Z}_k^{-1}} + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2\\ &\leq \beta_k + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \end{aligned}$$

where the first equality follows from the definition of $\boldsymbol{\mu}_k$, the second equality holds due to the definition of $\mathbf{Z}_k$ and the first inequality holds by triangle inequality with the fact that $\mathbf{Z}_k \succeq \lambda\mathbf{I}$. Thus, we complete the proof of Theorem 2.1. ■

Appendix C. Proofs from Section 2

C.1. Proof of Theorem 2.3

Lemma C.1 *Suppose that $\|\boldsymbol{\theta}^*\|_2 \leq 1$. In Algorithm 1, with probability at least $1-\delta$, the following statement holds for all round $k \geq 1$ and layer $\ell \in [L]$:*

$$\|\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}} \leq 16 \cdot 2^{-\ell}\sqrt{\sum_{i \in \Psi_{k,\ell}} w_i^2\sigma_i^2 \log(4k^2L/\delta)} + 6 \cdot 2^{-\ell}R\log(4k^2L/\delta) + 2^{-\ell+1}.$$

For simplicity, we denote $\mathcal{E}_{\text{conf}}$ as the event such that the result in Lemma C.1 holds in the remaining section.

Proof We first consider a fixed layer $\ell \in [L]$. Suppose that k is an arbitrary round satisfying $k \in \Psi_{k+1,\ell}$. Notice that in Line 14 (Algorithm 1), we introduce weight w_k to guarantee $\|w_k\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}} = 2^{-\ell}$.

Then we can apply Theorem 2.1 for the layer ℓ. In detail, for each $k \in \Psi_{K+1,\ell}$, we have

$$\|w_k\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}} = 2^{-\ell}, \quad \mathbb{E}[w_k^2\epsilon_k^2|\mathcal{F}_k] \leq w_k^2\mathbb{E}[\epsilon_k^2|\mathcal{F}_k] \leq w_k^2\sigma_k^2, \quad |w_k\epsilon_k| \leq |\epsilon_k| \leq R,$$

where the last inequality holds due to the fact that $w_k = 2^{-\ell}/\|\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}} \leq 1$. According to Theorem 2.1, we can deduce that with probability at least $1-\delta/L$, for all round $k \in \Psi_{K+1,\ell}$,

$$\|\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}} \leq 16 \cdot 2^{-\ell}\sqrt{\sum_{i \in \Psi_{k,\ell}} w_i^2\sigma_i^2 \log(4k^2L/\delta)} + 6 \cdot 2^{-\ell}R\log(4k^2L/\delta) + 2^{-\ell+1}.$$

Finally, after taking a union bound for all layer $\ell \in [L]$, we complete the proof of C.1. ■

Lemma C.2 *Suppose that the event $\mathcal{E}_{\text{conf}}$ defined in Lemma C.1 occurs. If $\{\widehat{\beta}_{k,\ell}\}_{k\geq 1, \ell \in [L]}$ satisfies*

$$\widehat{\beta}_{k,\ell} \geq 16 \cdot 2^{-\ell}\sqrt{\sum_{i \in \Psi_{k,\ell}} w_i^2\sigma_i^2 \log(4k^2L/\delta)} + 6 \cdot 2^{-\ell}R\log(4k^2L/\delta) + 2^{-\ell+1},$$

then for all $k \geq 1$ and $\ell \in [L]$ such that $\mathcal{A}_{k,\ell}$ exists, we have $\mathbf{a}_k^* \in \mathcal{A}_{k,\ell}$.

Proof Fix an arbitrary round k. If layer $\ell = 1$, then $\mathbf{a}_k^* \in \mathcal{D}_k = \mathcal{A}_{k,\ell}$ trivially holds. Then for layer $\ell > 1$, we prove lemma C.2 by induction. Assume that $\mathbf{a}_k^* \in \mathcal{A}_{k,\ell_1}$ holds for some $\ell_1 \in \mathbb{Z}^+$ and $\mathcal{A}_{k,\ell_1+1}$ exists.

By Lemma C.1, for all $\mathbf{a} \in \mathcal{A}_{k,\ell_1}$, we have

$$\left|\langle \mathbf{a}, \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle - \langle \mathbf{a}, \boldsymbol{\theta}^*\rangle\right| \leq \|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell_1}^{-1}} \left\|\widehat{\boldsymbol{\theta}}_{k,\ell_1} - \boldsymbol{\theta}^*\right\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell_1}} \leq \widehat{\beta}_{k,\ell}\|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell_1}^{-1}}, \tag{C.1}$$

where the first inequality holds due to Cauchy-Schwarz inequality and the last inequality holds due to the definition of events $\mathcal{E}_{\text{conf}}$. According to Line 10 of Algorithm 1, $\mathcal{A}_{k,\ell_1+1}$ exists only if $\|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell_1}^{-1}} \leq 2^{-\ell_1}$ holds for all $\mathbf{a} \in \mathcal{A}_{k,\ell_1}$. Therefore, the sub-optimality gap in (C.1) can be further bounded as follows:

$$\left|\langle \mathbf{a}, \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle - \langle \mathbf{a}, \boldsymbol{\theta}^*\rangle\right| \leq \widehat{\beta}_{k,\ell}\|\mathbf{a}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell_1}^{-1}} \leq 2^{-\ell_1} \cdot \widehat{\beta}_{k,\ell_1}. \tag{C.2}$$

For short, let $\mathbf{a}_{\max} = \operatorname{argmax}_{\mathbf{a}' \in \mathcal{A}_{k,\ell_1}} \langle \mathbf{a}', \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle$. Then for the optimal action $\mathbf{a}_k^* \in \mathcal{A}_{k,l_1}$, we have

$$\begin{aligned}
&\langle \mathbf{a}_k^*, \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle - \max_{\mathbf{a}' \in \mathcal{A}_{k,\ell_1}} \langle \mathbf{a}', \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle \\
&= \langle \mathbf{a}_k^*, \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle - \langle \mathbf{a}_{\max}, \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle \\
&\geq \langle \mathbf{a}_k^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_{\max}, \boldsymbol{\theta}^*\rangle - \left|\langle \mathbf{a}_k^*, \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle - \langle \mathbf{a}_k^*, \boldsymbol{\theta}^*\rangle\right| - \left|\langle \mathbf{a}_{\max}, \widehat{\boldsymbol{\theta}}_{k,\ell_1}\rangle - \langle \mathbf{a}_{\max}, \boldsymbol{\theta}^*\rangle\right| \\
&\geq -2^{-\ell_1+1} \cdot \widehat{\beta}_{k,\ell_1},
\end{aligned}$$

where the last inequality holds due to (C.2) with the fact that $\langle \mathbf{a}_k^*, \boldsymbol{\theta}^*\rangle \geq \langle \mathbf{a}_{\max}, \boldsymbol{\theta}^*\rangle$. Therefore, according to the Line 11 (Algorithm 1), the optimal action $\mathbf{a}_k^* \in \mathcal{A}_{k,\ell_1+1}$. Therefore, by induction, we complete the proof of Lemma C.2 ■

Lemma C.3 *Suppose for all $k \geq 1$ and all $\mathbf{a} \in \mathcal{D}_k$, we have $\|\mathbf{a}\|_2 \leq A, \|\boldsymbol{\theta}^*\|_2 \leq 1$. If $\mathcal{E}_{\text{conf}}$ occurs and $\{\beta_{k,\ell}\}_{k\geq 1, \ell \in [L]}$ satisfies the requirement in Lemma C.2, then for all $\ell \in [L]\backslash\{1\}$, the regret incurred by the index set $\Psi_{K+1,\ell}$ is bounded as follows :*

$$\sum_{\tau \in \Psi_{K+1,\ell}} \left(\langle \mathbf{a}_\tau^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_\tau, \boldsymbol{\theta}^*\rangle\right) \leq \widetilde{O}\left(d \cdot 2^\ell \cdot \widehat{\beta}_{K,\ell-1}\right).$$

Proof For all round $\tau \in \Psi_{K+1,\ell}$, we can deduce that $\mathbf{a}_\tau, \mathbf{a}_\tau^* \in \mathcal{A}_{\tau,\ell}$ by Lemma C.2. Also, according to Line 11 of Algorithm 1, we have

$$\langle \mathbf{a}_\tau^*, \widehat{\boldsymbol{\theta}}_{\tau,\ell-1}\rangle - \langle \mathbf{a}_\tau, \widehat{\boldsymbol{\theta}}_{\tau,\ell-1}\rangle \leq 2^{-\ell+2}\widehat{\beta}_{\tau,\ell-1}. \tag{C.3}$$

Besides, from Line 10 and the round $\tau \in \Psi_{K+1,\ell}$, we have

$$\|\mathbf{a}_\tau\|_{\widehat{\boldsymbol{\Sigma}}_{\tau,\ell-1}^{-1}} \leq 2^{-\ell+1}, \quad \|\mathbf{a}_\tau^*\|_{\widehat{\boldsymbol{\Sigma}}_{\tau,\ell-1}^{-1}} \leq 2^{-\ell+1}. \tag{C.4}$$

We further compute

$$\begin{aligned}
\langle \mathbf{a}_\tau^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_\tau, \boldsymbol{\theta}^*\rangle &\leq \langle \mathbf{a}_\tau^*, \widehat{\boldsymbol{\theta}}_{\tau,\ell-1}\rangle + \left|\langle \mathbf{a}_\tau^*, \widehat{\boldsymbol{\theta}}_{\tau,\ell-1} - \boldsymbol{\theta}^*\rangle\right| - \langle \mathbf{a}_\tau, \widehat{\boldsymbol{\theta}}_{\tau,\ell-1}\rangle + \left|\langle \mathbf{a}_\tau, \widehat{\boldsymbol{\theta}}_{\tau,\ell-1} - \boldsymbol{\theta}^*\rangle\right| \\
&\leq \langle \mathbf{a}_\tau^*, \widehat{\boldsymbol{\theta}}_{\tau,\ell-1}\rangle - \langle \mathbf{a}_\tau, \widehat{\boldsymbol{\theta}}_{\tau,\ell-1}\rangle \\
&\quad + \|\mathbf{a}_\tau^*\|_{\widehat{\boldsymbol{\Sigma}}_{\tau,\ell-1}^{-1}} \left\|\widehat{\boldsymbol{\theta}}_{\tau,\ell-1} - \boldsymbol{\theta}^*\right\|_{\widehat{\boldsymbol{\Sigma}}_{\tau,\ell-1}} + \|\mathbf{a}_\tau\|_{\widehat{\boldsymbol{\Sigma}}_{\tau,\ell-1}^{-1}} \left\|\widehat{\boldsymbol{\theta}}_{\tau,\ell-1} - \boldsymbol{\theta}^*\right\|_{\widehat{\boldsymbol{\Sigma}}_{\tau,\ell-1}} \\
&\leq 2^{-\ell+2} \cdot \widehat{\beta}_{\tau,\ell-1} + 2^{-\ell+1} \cdot \widehat{\beta}_{\tau,\ell-1} + 2^{-\ell+1} \cdot \widehat{\beta}_{\tau,\ell-1} \\
&= 8 \cdot 2^{-\ell} \cdot \widehat{\beta}_{\tau,\ell-1},
\end{aligned} \tag{C.5}$$

where the second inequality holds due to Cauchy-Schwarz inequality and the last inequality holds due to Lemma C.1, (C.3) and (C.4). Taking the summation over $\tau \in \Psi_{K+1,\ell}$, we have

$$\begin{aligned}\sum_{\tau\in\Psi_{K+1,\ell}} \big(\langle \mathbf{a}_\tau^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_\tau, \boldsymbol{\theta}^*\rangle\big) &\le 8\cdot 2^{-\ell}\cdot \widehat{\beta}_{K,\ell-1}\,|\Psi_{K+1,\ell}| \\ &\le 8\cdot 2^{\ell}\cdot \widehat{\beta}_{K,\ell-1}\cdot \sum_{k\in\Psi_{K+1,\ell}} \|w_k\cdot \mathbf{a}_k\|^2_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}} \\ &\le 8\cdot 2^{\ell}\cdot \widehat{\beta}_{K,\ell-1}\cdot 2d\log\Big(1+2^{2\ell}K\cdot A^2/d\Big),\end{aligned}$$

where the first inequality holds due to (C.5), the second inequality holds since for all round $k \in \Psi_{k+1,\ell}$, the weight w_k satisfies $\|w_k\mathbf{a}_k\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}^{-1}} = 2^{-\ell}$, and the last inequality holds due to Lemma E.2. ■

Lemma C.4 *Let weight w_i be defined in Algorithm 1. With probability at least $1-2\delta$, for all $k \ge 1$, $\ell\in[L]$, the following two inequalities hold simultaneously:*

$$\begin{aligned}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 &\le 2\sum_{i\in\Psi_{k+1,\ell}} w_i^2\epsilon_i^2 + \frac{14}{3}R^2\log(4k^2L/\delta),\\ \sum_{i\in\Psi_{k+1,\ell}} w_i^2\epsilon_i^2 &\le \frac{3}{2}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 + \frac{7}{3}R^2\log(4k^2L/\delta).\end{aligned}$$

For simplicity, we denote $\mathcal{E}_{\text{var}}$ as the event such that the two inequalities in Lemma C.4 holds.

Proof We first consider a fixed layer $\ell\in[L]$. For the gap between $\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2$ and $\sum_{i\in\Psi_{k+1,\ell}} w_i^2\epsilon_i^2$, according to the definition, we have

$$\begin{aligned}&\text{for } \forall i\ge 1,\ \mathbb{E}\left[\epsilon_i^2-\sigma_i^2|\mathbf{a}_{1:i}, r_{1:i-1}\right]=0,\\ &\sum_{i\in\Psi_{k+1,\ell}} \mathbb{E}\left[w_i^2(\epsilon_i^2-\sigma_i^2)^2|\mathbf{a}_{1:i}, r_{1:i-1}\right] \le \sum_{i\in\Psi_{k+1,\ell}} \mathbb{E}\left[w_i^2\epsilon_i^4|\mathbf{a}_{1:i}, r_{1:i-1}\right] \le R^2\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2,\end{aligned}$$

where the first inequality holds due to $\text{Var}[x]\le\mathbb{E}[x^2]$ and the second inequality holds due to $|\epsilon_i|\le R$ and $\mathbb{E}\left[\epsilon_i^2|\mathbf{a}_{1:i}, r_{1:i-1}\right]=\sigma_i^2$. Applying Freedman's inequality (Lemma E.4) with $\{\epsilon_i^2\}_{i\in\Psi_{k+1,\ell}}$ and taking a union bound for all $k\ge 1$, with probability at least $1-2\delta/L$, for all $k\ge1$, the following inequality holds

$$\begin{aligned}\left|\sum_{i\in\Psi_{k+1,\ell}} w_i^2(\sigma_i^2-\epsilon_i^2)\right| &\le \sqrt{2R^2\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2\log(4k^2L/\delta)} + \frac{2}{3}\cdot 2R^2\log(4k^2L/\delta)\\ &\le \frac{1}{2}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 + \frac{7}{3}R^2\log(4k^2L/\delta),\end{aligned}$$

where the last inequality holds due to Young's inequality. Rearranging the above inequality, we conclude that $\mathbb{P}(\mathcal{E}_{\text{var}})\ge 1-2\delta$ by applying union bound over all $\ell\in[L]$. Thus, we complete the proof of Lemma C.4. ■

Lemma C.5 *Suppose that $\|\boldsymbol{\theta}^*\|_2 \le 1$. Let weight w_i be defined in Algorithm 1. On the event $\mathcal{E}_{\text{conf}}$ and $\mathcal{E}_{\text{var}}$ (defined in Lemma C.1, C.4), for all $k \ge 1$, $\ell \in [L]$ such that $2^\ell \ge 64\sqrt{\log\left(4(k+1)^2L/\delta\right)}$, we have the following inequalities:*

$$\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 \le 8\sum_{i\in\Psi_{k+1,\ell}} w_i^2\left(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell}, \mathbf{a}_i\rangle\right)^2 + 6R^2\log(4(k+1)^2L/\delta) + 2^{-2\ell+4},$$

$$\sum_{i\in\Psi_{k+1,\ell}} w_i^2\left(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell}, \mathbf{a}_i\rangle\right)^2 \le \frac{3}{2}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 + \frac{7}{3}R^2\log(4k^2L/\delta) + 2^{-2\ell}.$$

Proof Let ℓ be an arbitrary index in $[L]$. By the definition of events $\mathcal{E}_{\text{var}}$, we have

$$\begin{aligned}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 &\le 2\sum_{i\in\Psi_{k+1,\ell}} w_i^2\epsilon_i^2 + \frac{14}{3}R^2\log(4k^2L/\delta)\\
&\le 4\sum_{i\in\Psi_{k+1,\ell}} w_i^2\left(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell}, \mathbf{a}_i\rangle\right)^2 + 4\sum_{i\in\Psi_{k+1,\ell}} w_i^2\left[\epsilon_i - \left(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell}, \mathbf{a}_i\rangle\right)\right]^2\\
&\quad + \frac{14}{3}R^2\log(4k^2L/\delta), \end{aligned} \tag{C.6}$$

where the last inequality holds due to $(a+b)^2 \le 2a^2 + 2b^2$. In addition, the gap between ϵ_i and $r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell}, \mathbf{a}_i\rangle$ can be upper bounded by

$$\begin{aligned}&\sum_{i\in\Psi_{k+1,\ell}} w_i^2\left[\epsilon_i - \left(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell}, \mathbf{a}_i\rangle\right)\right]^2\\
&= \sum_{i\in\Psi_{k+1,\ell}} w_i^2\left(\langle\widehat{\boldsymbol{\theta}}_{k+1,\ell} - \boldsymbol{\theta}^*, \mathbf{a}_i\rangle\right)^2\\
&= \sum_{i\in\Psi_{k+1,\ell}} \left(\widehat{\boldsymbol{\theta}}_{k+1,\ell} - \boldsymbol{\theta}^*\right)^\top (w_i\mathbf{a}_i)\cdot(w_i\mathbf{a}_i)^\top\left(\widehat{\boldsymbol{\theta}}_{k+1,\ell} - \boldsymbol{\theta}^*\right)\\
&\le \left(\widehat{\boldsymbol{\theta}}_{k+1,\ell} - \boldsymbol{\theta}^*\right)^\top \widehat{\boldsymbol{\Sigma}}_{k+1,\ell}\left(\widehat{\boldsymbol{\theta}}_{k+1,\ell} - \boldsymbol{\theta}^*\right)\\
&\le \left(16\cdot 2^{-\ell}\sqrt{\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2\log(4(k+1)^2L/\delta)} + 6\cdot 2^{-\ell}R\log(4(k+1)^2L/\delta) + 2^{-\ell+1}\right)^2,\end{aligned} \tag{C.7}$$

where the first inequality holds due to $\widehat{\boldsymbol{\Sigma}}_{k+1,\ell} \succeq w_i^2\mathbf{a}_i\mathbf{a}_i^\top$ and the last inequality holds due to Lemma C.1. From (C.7), when $2^\ell \ge 64\sqrt{\log\left(4(k+1)^2L/\delta\right)}$, we have

$$\begin{aligned}&\sum_{i\in\Psi_{k+1,\ell}} w_i^2\left[\epsilon_i - \left(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell}, \mathbf{a}_i\rangle\right)\right]^2\\
&\qquad\le \frac{1}{8}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 + 2\left(6\cdot 2^{-\ell}R\log(4(k+1)^2L/\delta) + 2^{-\ell+1}\right)^2.\end{aligned} \tag{C.8}$$

where the inequality holds due to (C.7) with the fact that $(a+b)^2 \le 2a^2 + 2b^2$. Substituting (C.8) into (C.6), we have

$$\begin{aligned}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 &\le 4\sum_{i\in\Psi_{k+1,\ell}} w_i^2\Big(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell},\mathbf{a}_i\rangle\Big)^2 + 4\sum_{i\in\Psi_{k+1,\ell}} w_i^2\Big[\epsilon_i - \Big(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell},\mathbf{a}_i\rangle\Big)\Big]^2 \\ &\quad + \frac{14}{3}R^2\log(4k^2L/\delta) \\ &\le 4\sum_{i\in\Psi_{k+1,\ell}} w_i^2\Big(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell},\mathbf{a}_i\rangle\Big)^2 + \frac{1}{2}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 \\ &\quad + 2\Big(6\cdot 2^{-\ell}R\log(4(k+1)^2L/\delta) + 2^{-\ell+1}\Big)^2 + \frac{14}{3}R^2\log(4k^2L/\delta) \\ &\le 8\sum_{i\in\Psi_{k+1,\ell}} w_i^2\Big(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell},\mathbf{a}_i\rangle\Big)^2 + 6R^2\log(4(k+1)^2L/\delta) + 2^{-2\ell+4},\end{aligned}$$

where the last inequality holds due to the fact that $x \le x/2 + y$ implies $x \le 2y$. Thus, we complete the proof of the first part of Lemma C.5.

For the second part, note that $\boldsymbol{\theta}_{k+1,\ell}$ is the minimizer of the following weighted ridge regression

$$\boldsymbol{\theta}_{k+1,\ell} \leftarrow \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i\in\Psi_{k+1,\ell}} w_i^2\big(r_i - \langle\boldsymbol{\theta},\mathbf{a}_i\rangle\big)^2 + 2^{-2\ell}\|\boldsymbol{\theta}\|_2^2.$$

Thus, we have

$$\sum_{i\in\Psi_{k+1,\ell}} w_i^2\Big(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell},\mathbf{a}_i\rangle\Big)^2 \le \sum_{i\in\Psi_{k+1,\ell}} w_i^2\big(r_i - \langle\boldsymbol{\theta}^*,\mathbf{a}_i\rangle\big)^2 + 2^{-2\ell}\|\boldsymbol{\theta}^*\|_2^2 \le \sum_{i\in\Psi_{k+1,\ell}} w_i^2\epsilon_i^2 + 2^{-2\ell},$$

where the second inequality holds due to $\|\boldsymbol{\theta}^*\|_2 \le 1$. Combining the result in Lemma C.4, we can further conclude that

$$\begin{aligned}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\Big(r_i - \langle\widehat{\boldsymbol{\theta}}_{k+1,\ell},\mathbf{a}_i\rangle\Big)^2 &\le \sum_{i\in\Psi_{k+1,\ell}} w_i^2\epsilon_i^2 + 2^{-2\ell} \\ &\le \frac{3}{2}\sum_{i\in\Psi_{k+1,\ell}} w_i^2\sigma_i^2 + \frac{7}{3}R^2\log(4k^2L/\delta) + 2^{-2\ell}.\end{aligned}$$

Thus, we complete the proof of Lemma C.5. ■

Proof [Proof of Theorem 2.3] Applying a union bound on event $\mathcal{E}_{\text{conf}}$and $\mathcal{E}_{\text{var}}$ defined in Lemma C.1 and C.4, we have $P(\mathcal{E}_{\text{conf}} \cap \mathcal{E}_{\text{var}}) \ge 1 - 3\delta$. In the remaining proof, we suppose that $\mathcal{E}_{\text{conf}}, \mathcal{E}_{\text{var}}$ hold simultaneously. For simplicity, let $\ell^* = \lceil \frac{1}{2}\log_2\log\big(4(K+1)^2L/\delta\big)\rceil + 8$. By the definition of $\widehat{\beta}_{k,\ell}$ and Lemma C.5, we have for all $\ell^*+1 \le \ell \le L$,

$$\widehat{\beta}_{K,\ell-1} \ge 16\cdot 2^{-(\ell-1)}\sqrt{\sum_{i\in\Psi_{K,\ell-1}} w_i^2\sigma_i^2\log(4K^2L/\delta)} + 6\cdot 2^{-\ell}R\log(4K^2L/\delta) + 2^{-\ell},$$

which further implies

$$\begin{aligned}\sum_{\tau\in\Psi_{K+1,\ell}} \left(\langle \mathbf{a}_\tau^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_\tau, \boldsymbol{\theta}^*\rangle\right) &\le \widetilde{O}\left(d\cdot 2^\ell \cdot \widehat{\beta}_{K,\ell-1}\right)\\ &\le \widetilde{O}\left(d\sqrt{\sum_{k=1}^K w_k^2\left(r_k - \langle \widehat{\boldsymbol{\theta}}_{K+1,\ell}, \mathbf{a}_k\rangle\right)^2 + R^2 + 1} + R\right)\\ &\le \widetilde{O}\left(d\sqrt{\sum_{k=1}^K \sigma_k^2} + dR + d\right), \end{aligned} \tag{C.9}$$

where the first inequality holds due to Lemma C.3, the second inequality holds due to (2.3) and the last inequality follows from Lemma C.5.

For each round $k\in[K]\backslash\left(\bigcup_{\ell\in[L]}\Psi_{K+1,\ell}\right) := \Psi_{K+1,L+1}$, we set ℓ_k as the value of layer ℓ such that the while loop in Algorithm 1 stops. Therefore, we have

$$\begin{aligned}\sum_{k\in[K]\backslash\left(\bigcup_{\ell\in[L]}\Psi_{K+1,\ell}\right)} \left(\langle \mathbf{a}_k^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_k, \boldsymbol{\theta}^*\rangle\right) &\le \sum_{k\in\Psi_{K+1,L+1}} \left(\langle \mathbf{a}_k, \widehat{\boldsymbol{\theta}}_{k,\ell_k}\rangle + \widehat{\beta}_{k,\ell_k}\cdot\alpha - \langle \mathbf{a}_k, \boldsymbol{\theta}^*\rangle\right)\\ &\le \sum_{k\in\Psi_{K+1,L+1}} \left(\widehat{\beta}_{k,\ell_k}\cdot\alpha + \alpha\cdot\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{k,\ell_k}\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell_k}}\right)\\ &\le \sum_{k\in\Psi_{K+1,L+1}} 2\alpha\cdot\widehat{\beta}_{k,\ell_k}\\ &\le K\cdot\widetilde{O}\left(1/K\right) = \widetilde{O}(1), \end{aligned} \tag{C.10}$$

where the first inequality holds due to the selection rule of action $\mathbf{a}_k$ (Line 8 in Algorithm 1) with Lemma C.1, Lemma C.5 and the fact that $\mathbf{a}_k^*\in\mathcal{A}_{k,\ell_k}$ (Lemma C.2), the second inequality holds due to Cauchy-Schwarz inequality, the third inequality follows from Lemma C.1 and the last inequality follows from the definition of α and $\widehat{\beta}_{k,\ell}$.

Finally, for layer $\ell\in[\ell^*]$ and round $\tau\in\Psi_{K+1,\ell}$, we have

$$\sum_{\tau\in\Psi_{K+1,\ell}} \left(\langle \mathbf{a}_\tau^*, \boldsymbol{\theta}^*\rangle - \langle \mathbf{a}_\tau, \boldsymbol{\theta}^*\rangle\right) \le 2\left|\Psi_{K+1,\ell}\right| = 2^{2\ell+1}\sum_{\tau\in\Psi_{K+1,\ell}} \|w_\tau\mathbf{a}_\tau\|^2_{\widehat{\boldsymbol{\Sigma}}_{\tau,\ell}} \le \widetilde{O}(d), \tag{C.11}$$

where the first inequality holds since the reward is in the range $[-1,1]$, the equation follows from the fact that $\|w_\tau\mathbf{a}_\tau\|_{\widehat{\boldsymbol{\Sigma}}_{\tau,\ell}} = 2^{-\ell}$ holds for all $\tau\in\Psi_{K+1,\ell}$ and the last inequality follows from Lemma E.2 with the fact that $2^{\ell^*}\le 128\sqrt{\log(4(K+1)^2L/\delta)}$ is bounded by a logarithmic term. Putting (C.9), (C.10), (C.11) together, we have

$$\mathrm{Regret}(K) \le \widetilde{O}\left(d\sqrt{\sum_{k=1}^K \sigma_k^2} + dR + d\right).$$

Thus, we complete the proof of Theorem 2.3. ■

Appendix D. Proofs from Section 3

For $k \in [K]$, $h \in [H]$, let $\mathcal{F}_{k,h}$ be the σ-algebra generated by the random variables representing the state-action pairs up to and including those that appear stage h of episode k. More specifically, $\mathcal{F}_{k,h}$ is generated by

$$\begin{aligned} &s_1^1, a_1^1, \dots, s_h^1, a_h^1, \dots, s_H^1, a_H^1\,, \\ &s_1^2, a_1^2, \dots, s_h^2, a_h^2, \dots, s_H^2, a_H^2\,, \\ &\qquad\qquad\vdots \\ &s_1^k, a_1^k, \dots, s_h^k, a_h^k\,. \end{aligned}$$

For simplicity, we define the following indicator sequence I_h^k for all $(k,h) \in [K] \times [H]$:

$$I_h^k = \mathbb{1}\left\{\forall \ell \in [L], \det\left(\widehat{\boldsymbol{\Sigma}}_{k,h,\ell}\right) / \det\left(\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}\right) \le 4\right\}. \tag{D.1}$$

For each $1 \le h_1 \le h_2 \le H$, since $\widehat{\boldsymbol{\Sigma}}_{k,h_2,\ell} \succeq \widehat{\boldsymbol{\Sigma}}_{k,h_1,\ell}$, the indicator function is monotonic (e.g., $I_{h_1}^k \le I_{h_2}^k$). In addition, the following lemma provides an upper bound for the number of episodes when the determinant of covariance matrix grows sharply.

Lemma D.1 *If the indicator function I_h^k is defined as in (D.1), then for each $k \in [K]$, we have*

$$\sum_{i=1}^k (1 - I_H^i) \le \frac{dL}{2} \log \frac{\lambda + kH/d}{\lambda} + dL^2.$$

Proof For all layer $\ell \in [L]$, let $\mathcal{D}_\ell$ be the set of indices $i \in [k]$ such that

$$\det\left(\widehat{\boldsymbol{\Sigma}}_{i+1,1,\ell}\right) / \det\left(\widehat{\boldsymbol{\Sigma}}_{i,1,\ell}\right) > 4.$$

According to the update rule of $\boldsymbol{\Sigma}_{k,1,\ell}$, $\boldsymbol{\Sigma}_{k+1,1,\ell} \succeq \boldsymbol{\Sigma}_{k,1,\ell}$ holds for all episode $k \in [K]$. Therefore, we have

$$\det(\widehat{\boldsymbol{\Sigma}}_{k+1,1,\ell}) / \det(\widehat{\boldsymbol{\Sigma}}_{1,1,\ell}) = \prod_{i=1}^k \det\left(\widehat{\boldsymbol{\Sigma}}_{i+1,1,\ell}\right) / \det\left(\widehat{\boldsymbol{\Sigma}}_{i,1,\ell}\right) \ge 4^{|\mathcal{D}_\ell|}, \tag{D.2}$$

where the inequality holds due to the definition of set $\mathcal{D}_\ell$. In addition, the determinant of matrices $\widehat{\boldsymbol{\Sigma}}_{k+1,1,\ell}$ and $\widehat{\boldsymbol{\Sigma}}_{1,1,\ell}$ is bounded by:

$$\begin{aligned} \det(\widehat{\boldsymbol{\Sigma}}_{k+1,1,\ell}) &\le (\mathrm{tr}(\boldsymbol{\Sigma}_{k+1,1,\ell})/d)^d \le (2^{-2\ell}\lambda + kH/d)^d, \\ \det(\widehat{\boldsymbol{\Sigma}}_{1,1,\ell}) &= \left(2^{-2\ell} \cdot \lambda\right)^d, \end{aligned}$$

where the first inequality holds since $\widehat{\boldsymbol{\Sigma}}_{k+1,1,\ell} \succeq \mathbf{0}$, the last inequality holds due to $w_{k,i} \le 1$ and $\|\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\|_2 \le 1$. Combining these results, it holds that

$$|\mathcal{D}_\ell| \le \log_4\left(\frac{(\lambda + 2^{2\ell}kH/d)^d}{\lambda^d}\right) \le \frac{d}{2}\log_2 \frac{\lambda + 2^{2\ell}kH/d}{\lambda} \le \frac{d}{2}\log_2 \frac{\lambda + kH/d}{\lambda} + d \cdot \ell.$$

Finally, according to the definition of $\mathcal{D}_\ell$ and indicator function I_h^k, we have

$$\sum_{i=1}^{k}(1-I_H^i) \le \sum_{\ell\in[L]} |\mathcal{D}_\ell| \le \frac{dL}{2}\log_2\frac{\lambda + kH/d}{\lambda} + dL^2.$$

Thus, we complete the proof of Lemma D.1. ■

Lemma D.2 *Let $\Psi_{K+1,\ell}$ be defined in (3.3). Then for all layer $\ell \in [L]$, it holds that $|\Psi_{K+1,\ell}| \le 2d\log\big(1+KH/(2^{-2\ell}d\lambda)\big)$.*

Proof By the definition of $w_{k,h}$ in Algorithm 2,

$$\sum_{(k,h)\in\Psi_{K+1,\ell}} \|w_{k,h}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\|^2_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,h,\ell}} = |\Psi_{K+1,\ell}|\cdot 2^{-2\ell}.$$

On the other hand, by Lemma E.2, we have

$$\sum_{(k,h)\in\Psi_{K+1,\ell}} \|w_{k,h}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\|^2_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,h,\ell}} \le 2d\log\frac{2^{-2\ell}d\lambda + KH}{2^{-2\ell}d\lambda}.$$

Combining these results, we further conclude that $|\Psi_{K+1,\ell}| \le 2\cdot d\log\big(1+KH/(2^{-2\ell}d\lambda)\big)$. Thus, we complete the proof of Lemma D.1. ■

D.1. High-Probability Events

For simplicity, we define the stochastic transition noise $\epsilon_{k,h}$ and variance $\sigma_{k,h}$ as follows:

$$\begin{aligned}\epsilon_{k,h} &= V_{k,h+1}(s_{h+1}^k) - \Big\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\Big\rangle,\\ \sigma_{k,h} &= \sqrt{[\mathbb{V}V_{k,h+1}]\,(s_h^k,a_h^k)}.\end{aligned} \tag{D.3}$$

With these notations, we further define the following high-probability events:

$$\begin{aligned}\mathcal{E}_{\mathrm{c}} = \Bigg\{&\forall k\ge 1, \ell\in[L], \|\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}} \le 16\cdot 2^{-\ell}\sqrt{\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2\log(4k^2H^2L/\delta)}\\ &+ 6\cdot 2^{-\ell}\log(4k^2H^2L/\delta) + 2^{-\ell}\sqrt{\lambda}\cdot B\Bigg\},\end{aligned} \tag{D.4}$$

$$\mathcal{E}_{\mathrm{var}'} = \left\{\forall k\ge 1, \sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left|\epsilon_{i,h}^2 - \sigma_{i,h}^2\right| \le \frac{1}{2}\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2 + \frac{7}{3}\log\left(4k^2H^2/\delta\right)\right\}. \tag{D.5}$$

Lemma D.3 *Let $\mathcal{E}_{\mathrm{c}}$ be defined in (D.4). Then we have $\mathbb{P}(\mathcal{E}_{\mathrm{c}}) \ge 1-\delta$.*

Proof From the definition of $\ell_{k,h}$ and $w_{k,h}$ in Algorithm 2, we can deduce that for all $k \in [K], h \in [H]$, $\|w_{k,h}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,h,\ell_{k,h}}} \leq 2^{-\ell_{k,h}}$. According to Theorem 2.1, for layer $\ell \in [L]$, we have with probability at least $1 - \delta/L$, for all $k \in [K]$:

$$\|\widehat{\boldsymbol{\theta}}_{k,\ell}\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}} \leq 16 \cdot 2^{-\ell}\sqrt{\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2 \log(4k^2H^2L/\delta)} + 6 \cdot 2^{-\ell}\log(4k^2H^2L/\delta) + 2^{-\ell}\sqrt{\lambda}B.$$

After applying a union bound over $\ell \in [L]$, we complete the proof of Lemma D.3. ■

Lemma D.4 *Let $\mathcal{E}_{\text{var}'}$ be defined in* (D.5). *We have $\mathbb{P}(\mathcal{E}_{\text{var}'}) \geq 1 - 2\delta$.*

Proof By the definition of $\mathbb{V}$ and Definition 3.2, we have

$$\mathbb{E}[\epsilon_{k,h}^2|\mathcal{F}_{k,h}] = \sigma_{k,h}^2, \ \mathbb{P}(|\epsilon_{k,h}| \leq 1) = 1.$$

Equivalent as the proof of Lemma C.4, we can prove that with probability at least $1 - 2\delta$, for all episode $k \geq 1$ and layer $\ell \in [L]$,

$$\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left|\epsilon_{i,h}^2 - \sigma_{i,h}^2\right| \leq \frac{1}{2}\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2 + \frac{7}{3}\log\left(4k^2H^2L/\delta\right),$$

which completes the proof of Lemma D.4. ■

D.2. Proof of Optimism

Lemma D.5 *Let $w_{k,h}$ be defined in Algorithm 2. On the event $\mathcal{E}_{\text{c}}$ and $\mathcal{E}_{\text{var}'}$, for all $k \geq 1$, $\ell \in [L]$ such that $2^\ell \geq 64\sqrt{\log(4k^2H^2L/\delta)}$, the following inequalities hold:*

$$\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2 \leq 8\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left(V_{i,h+1}(s_{h+1}^i) - \langle\widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\rangle\right)^2 \\ + 8\log(4k^2H^2L/\delta) + 2^{-2\ell+5}\cdot\lambda B^2,$$

$$\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left(V_{i,h+1}(s_{h+1}^i) - \langle\widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\rangle\right)^2 \leq \frac{3}{2}\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2 + 2^{-2\ell}\lambda B^2 \\ + \frac{7}{3}\log\left(4k^2H^2L/\delta\right).$$

Proof Let ℓ be an arbitrary layer in $[L]$. According to the definition of event $\mathcal{E}_{\text{var}'}$, we have

$$\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2 \leq 2\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\epsilon_{i,h}^2 + \frac{14}{3}\log\left(4k^2H^2L/\delta\right) \\ \leq 4\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left(V_{i,h+1}(s_{h+1}^i) - \langle\widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\rangle\right)^2 \\ + 4\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left[\epsilon_{i,h} - \left(V_{i,h+1}(s_{h+1}^i) - \langle\widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\rangle\right)\right]^2$$

$$+\frac{14}{3}\log\left(4k^2H^2L/\delta\right), \tag{D.6}$$

where the last inequality holds due to the fact $(a+b)^2 \le 2a^2+2b^2$. Then we consider the second term and we have

$$\begin{aligned}
&\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left[\epsilon_{i,h}-\left(V_{i,h+1}(s_{h+1}^i)-\langle\widehat{\boldsymbol{\theta}}_{k,\ell},\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)\rangle\right)\right]^2\\
&\quad=\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left(\langle\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{k,\ell},\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)\rangle\right)^2\\
&\quad=\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left(\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{k,\ell}\right)^\top\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)^\top\left(\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{k,\ell}\right)\\
&\quad\le\left\|\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{k,\ell}\right\|_{\widehat{\boldsymbol{\Sigma}}_{k,\ell}}^2\\
&\quad\le\left(16\cdot 2^{-\ell}\sqrt{\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2\log(4k^2H^2L/\delta)}+6\cdot 2^{-\ell}\log(4k^2H^2L/\delta)+2^{-\ell}\sqrt{\lambda}\cdot B\right)^2,
\end{aligned} \tag{D.7}$$

where the inequality holds due to $\widehat{\boldsymbol{\Sigma}}_{k,\ell}\succeq\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)^\top$ and weight $w_{i,h}\le 1$, the last equality follows from the definition of $\mathcal{E}_c$. In addition, from (D.7), when $2^\ell\ge 64\sqrt{\log(4k^2H^2L/\delta)}$,

$$\begin{aligned}
&\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left[\epsilon_{i,h}-\left(V_{i,h+1}(s_{h+1}^i)-\langle\widehat{\boldsymbol{\theta}}_{k,\ell},\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)\rangle\right)\right]^2\\
&\quad\le\frac{1}{8}\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2+2\left(6\cdot 2^{-\ell}\log(4k^2H^2L/\delta)+2^{-\ell}\sqrt{\lambda}\cdot B\right)^2\\
&\quad\le\frac{1}{8}\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2+\log(4k^2H^2L/\delta)+2^{-2\ell+2}\cdot\lambda B^2,
\end{aligned} \tag{D.8}$$

where the first inequality and the second inequality hold due to the fact that $(a+b)^2\le 2a^2+2b^2$. Substituting (D.8) into (D.6), we have

$$\begin{aligned}
\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2 &\le 4\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left(V_{i,h+1}(s_{h+1}^i)-\langle\widehat{\boldsymbol{\theta}}_{k,\ell},\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)\rangle\right)^2\\
&\quad+\frac{1}{2}\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2+4\log(4k^2H^2L/\delta)+2^{-2\ell+4}\cdot\lambda B^2\\
&\le 8\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\left(V_{i,h+1}(s_{h+1}^i)-\langle\widehat{\boldsymbol{\theta}}_{k,\ell},\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i,a_h^i)\rangle\right)^2\\
&\quad+8\log(4k^2H^2L/\delta)+2^{-2\ell+5}\cdot\lambda B^2,
\end{aligned}$$

where the last inequality holds due to the fact that $x\le x/2+y$ implies $x\le 2y$. Thus, we complete the proof of the first inequality in this lemma.

Note that $\widehat{\boldsymbol{\theta}}_{k,\ell}$ is the minimizer of

$$\widehat{\boldsymbol{\theta}}_{k,\ell} \leftarrow \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2 \left(V_{i,h+1}(s_{h+1}^i) - \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\rangle\right)^2 + 2^{-2\ell}\lambda\|\boldsymbol{\theta}\|_2^2,$$

and we have

$$\begin{aligned}
&\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2 \left(V_{i,h+1}(s_{h+1}^i) - \left\langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\right\rangle\right)^2 \\
&\le \sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2 \left(V_{i,h+1}(s_{h+1}^i) - \left\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\right\rangle\right)^2 + 2^{-2\ell}\lambda\|\boldsymbol{\theta}^*\|_2^2 \\
&\le \sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\epsilon_{i,h}^2 + 2^{-2\ell}\lambda B^2 \\
&\le \frac{3}{2}\sum_{(i,h)\in\Psi_{k,\ell}} w_{i,h}^2\sigma_{i,h}^2 + 2^{-2\ell}\lambda B^2 + \frac{7}{3}\log\left(4k^2H^2L/\delta\right),
\end{aligned}$$

where the first inequality holds due to the definition of $\widehat{\boldsymbol{\theta}}_{k,\ell}$, the second inequality holds due to $\|\boldsymbol{\theta}^*\| \le B$ and the last inequality follows from the definition of $\mathcal{E}_{\mathrm{var}'}$. Therefore, we complete the proof of Lemma D.5. ■

Lemma D.6 *Let value function $Q_{k,h}, V_{k,h}$ and confidence radius $\widehat{\beta}_{k,\ell}$ be defined in Algorithm 2. Suppose that $\lambda = 1/B^2$ in Algorithm 2. Then, on the event $\mathcal{E}_{\mathrm{var}'} \cap \mathcal{E}_{\mathrm{c}}$, for any $(k,h) \in [K]\times[H]$, we have $[\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) \le V_{k,h}(s_h^k)$.*

Proof From the definition of event $\mathcal{E}_{\mathrm{var}'}$ and Lemma D.5, we can deduce that

$$\widehat{\beta}_{k,\ell} \ge 16\cdot 2^{-\ell}\sqrt{\sum_{i=1}^k \sigma_i^2\log(4k^2H^2L/\delta)} + 6\cdot 2^{-\ell}\log(4k^2H^2L/\delta) + 2^{-\ell}\sqrt{\lambda}\cdot B.$$

Therefore, by the definition of $\mathcal{E}_{\mathrm{c}}$, we have

$$\forall k \ge 1, \ell\in[L], \quad \|\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}} \le \widehat{\beta}_{k,\ell}. \tag{D.9}$$

According to Algorithm 2, we have

$$\begin{aligned}
V_{k,h}(s_h^k) &= \min\left\{1, \min_{\ell\in[L]}\left\{r(s_h^k, a_h^k) + \left\langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\right\rangle + \widehat{\beta}_{k,\ell}\left\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\right\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}}\right\}\right\} \\
&\ge \min\left\{1, \min_{\ell\in[L]}\left\{\left\langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\right\rangle + \widehat{\beta}_{k,\ell}\left\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\right\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}}\right\}\right\} \\
&\ge \min\left\{1, \min_{\ell\in[L]}\left\{\left\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\right\rangle\right\}\right\} \\
&= [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k),
\end{aligned}$$

where the first inequality holds due to $r(s_h^k, a_h^k) > 0$, the second one follows from (D.9), the last equality holds due to the definition of linear mixture MDPs and the fact that $V_{k,h+1}(s) \le 1$ for all $s \in \mathcal{S}$. Thus, we complete the proof of Lemma D.6. ■

Lemma D.7 *Let value function $Q_{k,h}, V_{k,h}$ and confidence radius $\widehat{\beta}_{k,\ell}$ be defined in Algorithm 2. Suppose that $\lambda = 1/B^2$ in Algorithm 2. Then, on the event $\mathcal{E}_{\mathrm{var}'} \cap \mathcal{E}_{\mathrm{c}}$, for any $(s, a, k, h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$, we have $Q_h^*(s, a) \le Q_{k,h}(s, a)$ and $V_h^*(s) \le V_{k,h}(s)$.*

Proof From the definition of event $\mathcal{E}_{\mathrm{var}'}$ and Lemma D.5, we can deduce that

$$\widehat{\beta}_{k,\ell} \ge 16 \cdot 2^{-\ell} \sqrt{\sum_{i=1}^{k} \sigma_i^2 \log(4k^2H^2L/\delta)} + 6 \cdot 2^{-\ell} \log(4k^2H^2L/\delta) + 2^{-\ell}\sqrt{\lambda} \cdot B.$$

Therefore, by the definition of $\mathcal{E}_{\mathrm{c}}$, we have

$$\forall k \ge 1, \ell \in [L], \quad \|\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}} \le \widehat{\beta}_{k,\ell}. \tag{D.10}$$

Consider an arbitrary episode $k \in [K]$ in the remaining proof. If for some stage $h > 1$, the following inequalities $Q_h^*(s, a) \le Q_{k,h}(s, a)$, $V_h^*(s) \le V_{k,h}(s)$ hold for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, then for any $(s, a) \in \mathcal{S} \times \mathcal{A}, \ell \in [L]$ and stage $h - 1$, we have

$$\begin{aligned}
Q_{h-1}^*(s, a) &= r(s, a) + \big\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_h^*}(s, a) \big\rangle \\
&\le r(s, a) + \big\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{k,h}}(s, a) \big\rangle \\
&\le r(s, a) + \big\langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{k,h}}(s, a) \big\rangle + \big\|\widehat{\boldsymbol{\theta}}_{k,\ell} - \boldsymbol{\theta}^*\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}} \big\|\boldsymbol{\phi}_{V_{k,h}}(s, a)\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}} \\
&\le r(s, a) + \big\langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{k,h}}(s, a) \big\rangle + \widehat{\beta}_{k,\ell} \big\|\boldsymbol{\phi}_{V_{k,h}}(s, a)\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}},
\end{aligned}$$

where the first inequality holds by our assumption that $V_h^*(s) \le V_{k,h}(s)$, the second inequality holds due to Cauchy-Schwarz inequality and the last inequality follows from (D.10). By the arbitrariness of layer ℓ, we have $Q_{h-1}^*(s, a) \le Q_{k,h-1}(s, a)$ holds for all state-action pair (s, a), which indicates that $V_{h-1}^*(s) \le V_{k,h-1}(s)$ holds for all $s \in \mathcal{S}$. Since $0 = V_{H+1}^*(\cdot) \le V_{k,H+1}(\cdot)$ holds trivially for stage $H + 1$, we complete the proof of Lemma D.7 by induction. ■

D.3. Sum of Bellman Errors

Lemma D.8 *Let $\widehat{\beta}_{k,\ell}$, $V_{k,h}$, $\boldsymbol{\phi}_{V_{k,h+1}}$ be defined in Algorithm 2 and set $\lambda = 1/B^2, \alpha = 1/(KH)^{3/2}$. Then on the event $\mathcal{E}_{\mathrm{var}'} \cap \mathcal{E}_{\mathrm{c}}$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \max\Big\{\big[V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)\big], 0\Big\} \le \widetilde{O}\left(d\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \sigma_{k,h}^2} + d\right).$$

Proof For simplicity, let ℓ^* be the smallest ℓ in L such that $2^\ell \ge 64\sqrt{\log\left(4K^2H^2L/\delta\right)}$. According to Algorithm 2, we have

$$V_{k,h}(s_h^k) = \min\left\{1, \min_{\ell \in [L]}\left\{r(s_h^k, a_h^k) + \Big\langle \widehat{\boldsymbol{\theta}}_{k,\ell}, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\rangle + \widehat{\beta}_{k,\ell}\left\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\right\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}}\right\}\right\}.$$

Therefore, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \max\Big\{\big[V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)\big], 0\Big\} \tag{D.11}$$

$$
\begin{aligned}
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \bigg[\min_{\ell\in[L]}\Big\{\Big\langle \widehat{\boldsymbol{\theta}}_{k,\ell}-\boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\Big\rangle + \widehat{\beta}_{k,\ell}\Big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\Big\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}}\Big\}\bigg]_{[0,2]} \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \min\Big\{2, \min_{\ell\in[L]}\Big\{2\widehat{\beta}_{k,\ell}\Big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\Big\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell}^{-1}}\Big\}\Big\} \\
&\le \sum_{\ell=\ell^*+1}^{L+1}\sum_{(k,h)\in\Psi_{K+1,\ell}} I_h^k \min\Big\{2, 2\widehat{\beta}_{k,\ell-1}\Big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\Big\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell-1}^{-1}}\Big\} + 2\sum_{\ell=1}^{\ell^*}|\Psi_{K+1,\ell}|, \qquad \text{(D.12)}
\end{aligned}
$$

where the first inequality holds due to the definition of value function $V_{k,h}(s_h^k)$, the second inequality holds due to Cauchy-Schwarz inequality with event $\mathcal{E}_c$ and the last inequality holds since indicator function $I_h^k \le 1$. By the definition of indicator function I_h^k and Lemma E.3, we further have

$$
\begin{aligned}
\Big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\Big\|_{\widehat{\boldsymbol{\Sigma}}_{k,1,\ell_{k,h}-1}^{-1}} &\le 2\Big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\Big\|_{\widehat{\boldsymbol{\Sigma}}_{k,h,\ell_{k,h}-1}^{-1}} \\
&\le 2\cdot 2^{-\ell_{k,h}+1}, \qquad \text{(D.13)}
\end{aligned}
$$

where the last inequality follows from the definition of $\ell_{k,h}$ in Algorithm 2. Substituting (D.13) into (D.12), we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \max\Big\{\big[V_{k,h}(s_h^k) - r(s_h^k,a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k,a_h^k)\big], 0\Big\} \\
&\le \sum_{\ell=\ell^*+1}^{L+1}|\Psi_{K+1,\ell}|\cdot\widetilde{O}\left(2^{-\ell}\cdot 2^{-\ell}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2 + 2^{-2\ell}}\right) + 2\sum_{\ell=1}^{\ell^*}|\Psi_{K+1,\ell}| \\
&\le \widetilde{O}\left(d\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2} + d\right),
\end{aligned}
$$

where the first inequality follows from the definition of $\widehat{\beta}_{k,\ell}$ in Algorithm 2 and Lemma D.5, the last inequality holds due to Lemma D.2 and the definition of L. Thus, we complete the proof of Lemma D.8. ■

D.4. Quantities in MDP

In this subsection, we define the following quantities: We use $\check{V}_{k,h}(s)$ to denote the estimation error between the optimistic value function and the actually optimal value function, and use $\widetilde{V}_{k,h}(s)$ to denote the sub-optimality gap of policy π_k at stage h:

$$
\begin{aligned}
\check{V}_{k,h}(s) &= V_{k,h}(s) - V_h^*(s), \quad \forall s\in\mathcal{S}, (k,h)\in[K]\times[H] \qquad \text{(D.14)} \\
\widetilde{V}_{k,h}(s) &= V_h^*(s) - V_h^{\pi_k}(s), \quad \forall s\in\mathcal{S}, (k,h)\in[K]\times[H] \qquad \text{(D.15)}
\end{aligned}
$$

We use $Q_0, S_m, \check{S}_m, \widetilde{S}_m$ to represent the total variances of optimal value function V_{h+1}^* and 2^m-th order value functions $(V_{k,h+1}^{2^m}, \check{V}_{k,h+1}^{2^m}, \widetilde{V}_{k,h+1}^{2^m})$:

$$
S_m = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}V_{k,h+1}^{2^m}](s_h^k,a_h^k), \qquad \text{(D.16)}
$$

$$\check{S}_m = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}\check{V}^{2^m}_{k,h+1}](s_h^k, a_h^k), \tag{D.17}$$

$$\widetilde{S}_m = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}\widetilde{V}^{2^m}_{k,h+1}](s_h^k, a_h^k), \tag{D.18}$$

$$Q_0 = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}V^*_{h+1}](s_h^k, a_h^k), \tag{D.19}$$

where Q_0 is introduced as a shorthand for Var_K^* for simplicity. In addition, for 2^m-th order value functions $(V^{2^m}_{k,h+1}, \check{V}^{2^m}_{k,h+1}, \widetilde{V}^{2^m}_{k,h+1})$ and optimistic value function $V_{k,h}$, we denote the summation of stochastic transition noise as follows:

$$A_m = \left|\sum_{k=1}^{K}\sum_{h=1}^{H}[[\mathbb{P}V^{2^m}_{k,h+1}](s_h^k, a_h^k) - V^{2^m}_{k,h+1}(s_{h+1}^k)]\right|, \tag{D.20}$$

$$\check{A}_m = \left|\sum_{k=1}^{K}\sum_{h=1}^{H}[[\mathbb{P}\check{V}^{2^m}_{k,h+1}](s_h^k, a_h^k) - \check{V}^{2^m}_{k,h+1}(s_{h+1}^k)]\right|, \tag{D.21}$$

$$\widetilde{A}_m = \left|\sum_{k=1}^{K}\sum_{h=1}^{H}[[\mathbb{P}\widetilde{V}^{2^m}_{k,h+1}](s_h^k, a_h^k) - \widetilde{V}^{2^m}_{k,h+1}(s_{h+1}^k)]\right|, \tag{D.22}$$

$$R_0 = \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \max\Big\{\big[V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)\big], 0\Big\}. \tag{D.23}$$

Finally, we use the quantity G to denote the number of episodes when the determinant of covariance matrix grows sharply:

$$G = \sum_{k=1}^{K}(1 - I_H^k), \tag{D.24}$$

where indicator function I_h^k is defined in (D.1). For the above quantities, we only consider $m \in [M]$ where $M := \lceil \log_2 2HK \rceil$. Now, we introduce the following lemmas to build the connection between these quantities.

To construct the connections and upper bounds of the quantities above, our proof in this subsection follows the previous approaches proposed by Zhang et al. (2021b) and Zhou and Gu (2022), but with a more fine-grained analysis to remove explicit K-dependence.

Lemma D.9 *Let $\check{S}_m$, A_m, R_0, G be defined in* (D.17), (D.20), (D.23), (D.24). *On the event $\mathcal{E}_{\mathrm{var}'} \cap \mathcal{E}_{\mathrm{c}}$, we have the following inequalities for all $m \in [M]$:*

$$\check{S}_m \le \check{A}_{m+1} + G + 2^{m+1} \cdot (R_0 + G + A_0).$$

Proof Based on the definition of $\check{S}_m$, we compute

$$\check{S}_m = \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}\check{V}^{2^m}_{k,h+1}](s_h^k, a_h^k)$$

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

$$= \sum_{k=1}^{K}\sum_{h=1}^{H}\left[[\mathbb{P}\check{V}_{k,h+1}^{2^{m+1}}](s_h^k,a_h^k) - \check{V}_{k,h+1}^{2^{m+1}}(s_{h+1}^k)\right] + \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - \Big([\mathbb{P}\check{V}_{k,h+1}^{2^m}](s_h^k,a_h^k)\Big)^2\right]. \tag{D.25}$$

For the second term, it can be further upper bounded by

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - \Big([\mathbb{P}\check{V}_{k,h+1}^{2^m}](s_h^k,a_h^k)\Big)^2\right]\\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\check{V}_{k,h}^{2^{m+1}}(s_h^k) - \Big([\mathbb{P}\check{V}_{k,h+1}](s_h^k,a_h^k)\Big)^{2^{m+1}}\right]\\
&\le 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H}\max\Big\{\check{V}_{k,h}(s_h^k) - [\mathbb{P}\check{V}_{k,h+1}](s_h^K,a_h^k),0\Big\}\\
&\le 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H}I_h^k\max\Big\{\big[V_{k,h}(s_h^k) - r(s_h^k,a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k,a_h^k)\big],0\Big\}\\
&\quad + 2^{m+1}\sum_{k=1}^{K}(1-I_H^k)\sum_{h=1}^{H}\max\Big\{\big[V_{k,h}(s_h^k) - r(s_h^k,a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k,a_h^k)\big],0\Big\}\\
&\le 2^{m+1}R_0 + 2^{m+1}\sum_{k=1}^{K}(1-I_H^k)\sum_{h=1}^{H}\big[V_{k,h+1}(s_{h+1}^k) - [\mathbb{P}V_{k,h+1}](s_h^k,a_h^k)\big]\\
&\quad + 2^{m+1}\sum_{k=1}^{K}(1-I_H^k)\sum_{h=1}^{H}\Big(V_{k,h}(s_h^k) - V_{k,h+1}(s_{h+1}^k)\Big)\\
&\le 2^{m+1}\cdot(R_0+G+A_0),
\end{aligned} \tag{D.26}$$

where the first inequality holds due to

$$\Big([\mathbb{P}\check{V}_{k,h+1}^{2^m}](s_h^k,a_h^k)\Big)^2 \ge \Big([\mathbb{P}\check{V}_{k,h+1}^{2^{m-1}}](s_h^k,a_h^k)\Big)^4 \ge \cdots \ge \Big([\mathbb{P}\check{V}_{k,h+1}](s_h^k,a_h^k)\Big)^{2^{m+1}},$$

the second inequality follows from the fact that $a^x - b^x \le x\max\{a-b,0\}$ for $a,b\in[0,1]$ and $x\ge 1$, the third inequality follows from the monotonicity of indicator function I_h^k and the definition of function $\check{V}_{k,h}$, the fourth holds since $r(s_h^k,a_h^k)\ge 0$ and the last inequality holds due to Lemma D.6.

Substituting (D.26) into (D.25), we have

$$\check{S}_m \le \check{A}_{m+1} + G + 2^{m+1}\cdot(R_0+G+A_0).$$

Thus, we complete the proof of Lemma D.9. ■

Lemma D.10 *Let A_m, Q_0, $\check{S}_m$ be defined in (D.20), (D.19), (D.17). Then with probability at least $1-2\delta$, we have*

$$A_0 \le 2\sqrt{(Q_0+\check{S}_0)\log(1/\delta)} + (2/3)\cdot\log(1/\delta).$$

For simplicity, we denote the corresponding event by $\mathcal{E}_{\mathrm{r}_1}$.

Proof Applying the Freedman's inequality in Lemma E.4, we have with probability at least $1-2\delta$,

$$A_0 = \left|\sum_{k=1}^{K}\sum_{h=1}^{H}\big[[\mathbb{P}V_{k,h+1}](s_h^k,a_h^k) - V_{k,h+1}(s_{h+1}^k)\big]\right| \le \sqrt{2\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{k,h}^2\log(1/\delta)} + \frac{2}{3}\log(1/\delta), \tag{D.27}$$

where the variance $\sigma_{k,h}$ is defined in (D.3). For variance $\sigma_{k,h}$, we further have

$$\sigma_{k,h}^2 = [\mathbb{V}V_{k,h+1}](s_h^k,a_h^k) \le 2[\mathbb{V}V_{h+1}^*](s_h^k,a_h^k) + 2[\mathbb{V}\check{V}_{k,h+1}](s_h^k,a_h^k), \tag{D.28}$$

where the inequality holds due to the fact that $\mathrm{Var}(x+y)\le 2\,\mathrm{Var}(x)+2\,\mathrm{Var}(y)$. Substituting (D.28) into (D.27), we complete the proof of Lemma D.10. ■

Lemma D.11 *Let $\check{A}_m$, $\check{S}_m$ be defined in (D.20), (D.17). With probability at least $1-2(M+1)\delta$, for all $m\in[M]\cup\{0\}$, we have*

$$\check{A}_m \le \sqrt{2\check{S}_m\log(1/\delta)} + \frac{4}{3}\cdot\log(1/\delta).$$

We denote the corresponding event by $\mathcal{E}_{\mathrm{r}_2}$.

Proof Note that for all state s, $\check{V}_{k,h}(s) = V_{k,h}(s) - V_h^*(s) \in [-1,1]$. Applying Freedman's inequality in Lemma E.4, we have with probability at least $1-2\delta$:

$$\check{A}_m \le \sqrt{2\check{S}_m\log(1/\delta)} + \frac{4}{3}\cdot\log(1/\delta),$$

for each $m\in[M]$. Thus, we complete the proof of Lemma D.11 by using a union bound over $m\in[M]$. ■

Lemma D.12 *Let A_m, Q_0, $\check{A}_m$, R_0, G be defined in (D.20), (D.19), (D.21), (D.23), (D.24). On the event $\mathcal{E}_{\mathrm{r}_1}\cap\mathcal{E}_{\mathrm{var}'}\cap\mathcal{E}_{\mathrm{c}}$, we have*

$$A_0 \le 4\sqrt{\big(Q_0+\check{A}_1+G+2(R_0+G)\big)\log(1/\delta)} + 10\cdot\log(1/\delta).$$

Proof According to Lemma D.9 and Lemma D.10, we have

$$\begin{aligned}
A_0 &\le 2\sqrt{(Q_0+\check{S}_0)\log(1/\delta)} + (2/3)\cdot\log(1/\delta)\\
&\le 2\sqrt{\big(Q_0+\check{A}_1+G+2(R_0+G+A_0)\big)\log(1/\delta)} + (2/3)\cdot\log(1/\delta)
\end{aligned}$$

$$\begin{aligned}
&\le 2\sqrt{\big(Q_0 + \breve{A}_1 + G + 2(R_0 + G)\big)\log(1/\delta)} + (2/3)\cdot\log(1/\delta) + 2\sqrt{2A_0\log(1/\delta)}\\
&\le 4\sqrt{\big(Q_0 + \breve{A}_1 + G + 2(R_0 + G)\big)\log(1/\delta)} + 10\cdot\log(1/\delta),
\end{aligned}$$

where the first inequality holds due to Lemma D.9, the second inequality holds due to D.10 and the last inequality holds due to $x \le a\sqrt{x} + b \Rightarrow x \le a^2 + 2b$. Thus, we complete the proof of Lemma D.12. ■

Lemma D.13 *Let A_m, G, R_0 be defined in* (D.20), (D.24), (D.23). *On the event $\mathcal{E}_{\mathrm{r}_2} \cap \mathcal{E}_{\mathrm{var}'} \cap \mathcal{E}_{\mathrm{c}}$, we have*

$$\breve{A}_1 \le 4\sqrt{(R_0 + 2G + A_0)\log(1/\delta)} + 11\log(1/\delta).$$

Proof By the definition of $\mathcal{E}_{\mathrm{r}_2}$ in Lemma D.11, we have

$$\breve{A}_m \le \sqrt{2\breve{S}_m\log(1/\delta)} + \frac{4}{3}\cdot\log(1/\delta). \tag{D.29}$$

Substituting the bound of $\breve{S}_m$ in Lemma D.9 into (D.29),

$$\breve{A}_m \le \sqrt{2\left(\breve{A}_{m+1} + 2^{m+1}\cdot(R_0 + 2G + A_0)\right)\log(1/\delta)} + \frac{4}{3}\cdot\log(1/\delta).$$

Applying Lemma E.5, we have

$$\begin{aligned}
\breve{A}_1 &\le \max\Big\{11\log(1/\delta), 4\sqrt{(R_0 + 2G + A_0)\log(1/\delta)} + 2\log(1/\delta)\Big\}\\
&\le 4\sqrt{(R_0 + 2G + A_0)\log(1/\delta)} + 11\log(1/\delta).
\end{aligned}$$

Thus, we complete the proof of Lemma D.13. ■

Lemma D.14 *Let A_m, G, R_0, Q_0 be defined in* (D.20), (D.24), (D.23), (D.19). *On the event $\mathcal{E}_{\mathrm{r}_1} \cap \mathcal{E}_{\mathrm{r}_2} \cap \mathcal{E}_{\mathrm{var}'} \cap \mathcal{E}_{\mathrm{c}}$, we have*

$$A_0 \le 132\log(1/\delta) + 28\sqrt{R_0\log(1/\delta)} + 40\sqrt{G\log(1/\delta)} + 8\sqrt{Q_0\log(1/\delta)}.$$

Proof We compute

$$\begin{aligned}
A_0 &\le 4\sqrt{\big(Q_0 + G + 2(R_0 + G)\big)\log(1/\delta)} + 10\cdot\log(1/\delta) + 4\sqrt{\breve{A}_1\log(1/\delta)}\\
&\le 4\sqrt{\big(Q_0 + G + 2(R_0 + G)\big)\log(1/\delta)} + 2\breve{A}_1 + 12\log(1/\delta)\\
&\le 8\sqrt{(R_0 + 2G + A_0)\log(1/\delta)} + 34\log(1/\delta) + 4\sqrt{(Q_0 + G + 2(R_0 + G))\log(1/\delta)}\\
&\le 132\log(1/\delta) + 28\sqrt{R_0\log(1/\delta)} + 40\sqrt{G\log(1/\delta)} + 8\sqrt{Q_0\log(1/\delta)},
\end{aligned}$$

where the first inequality follows from Lemma D.12, the second inequality holds due to the fact that $2ab \le a^2 + b^2$, the third inequality holds due to Lemma D.13 and the last inequality holds due to the fact that $x \le a\sqrt{x} + b \Rightarrow x \le a^2 + 2b$. Thus, we complete the proof of Lemma D.14. ■

Lemma D.15 *Let $\widetilde{S}_m$, A_m, R_0, G be defined in (D.18), (D.20), (D.23), (D.24). On the event $\mathcal{E}_{\text{var}'} \cap \mathcal{E}_{\text{c}}$, we have the following inequalities for all $m \in [M]$:*

$$\widetilde{S}_m \le \widetilde{A}_{m+1} + G + 2^{m+1} \cdot (R_0 + G + A_0).$$

Proof Based on the definition of $\widetilde{S}_m$, we compute

$$\begin{aligned}
\widetilde{S}_m &= \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}\widetilde{V}^{2^m}_{k,h+1}](s_h^k, a_h^k) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\left[[\mathbb{P}\widetilde{V}^{2^{m+1}}_{k,h+1}](s_h^k, a_h^k) - \Big([\mathbb{P}\widetilde{V}^{2^m}_{k,h+1}](s_h^k, a_h^k)\Big)^2\right] \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\left[[\mathbb{P}\widetilde{V}^{2^{m+1}}_{k,h+1}](s_h^k, a_h^k) - \widetilde{V}^{2^{m+1}}_{k,h+1}(s_{h+1}^k)\right] + \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\widetilde{V}^{2^{m+1}}_{k,h}(s_h^k) - \Big([\mathbb{P}\widetilde{V}^{2^m}_{k,h+1}](s_h^k, a_h^k)\Big)^2\right],
\end{aligned} \tag{D.30}$$

For the second term, we further have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\widetilde{V}^{2^{m+1}}_{k,h}(s_h^k) - \Big([\mathbb{P}\widetilde{V}^{2^m}_{k,h+1}](s_h^k, a_h^k)\Big)^2\right] \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\widetilde{V}^{2^{m+1}}_{k,h}(s_h^k) - \Big([\mathbb{P}\widetilde{V}_{k,h+1}](s_h^k, a_h^k)\Big)^{2^{m+1}}\right] \\
&\le 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H}\max\Big\{\widetilde{V}_{k,h}(s_h^k) - [\mathbb{P}\widetilde{V}_{k,h+1}](s_h^k, a_h^k), 0\Big\} \\
&\le 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k\big[V_h^*(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k)\big] \\
&\quad + 2^{m+1}\sum_{k=1}^{K}(1 - I_H^k)\sum_{h=1}^{H}\big[V_h^*(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{h+1}^*](s_h^k, a_h^k)\big] \\
&\le 2^{m+1}R_0 + 2^{m+1}\check{A}_0 + 2^{m+1}\sum_{k=1}^{K}(1 - I_H^k)\sum_{h=1}^{H}\big[V_{k,h+1}(s_{h+1}^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)\big] \\
&\quad + 2^{m+1}\sum_{k=1}^{K}(1 - I_H^k)\sum_{h=1}^{H}\Big(V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - V_{k,h+1}(s_{h+1}^k)\Big) \\
&\le 2^{m+1} \cdot (R_0 + G + A_0 + \check{A}_0),
\end{aligned} \tag{D.31}$$

where the first inequality holds since

$$\Big([\mathbb{P}\widetilde{V}^{2^m}_{k,h+1}](s_h^k, a_h^k)\Big)^2 \ge \Big([\mathbb{P}\widetilde{V}^{2^{m-1}}_{k,h+1}](s_h^k, a_h^k)\Big)^4 \ge \cdots \ge \Big([\mathbb{P}\widetilde{V}_{k,h+1}](s_h^k, a_h^k)\Big)^{2^{m+1}},$$

the second inequality follows from the fact that $a^x - b^x \le x\max\{a - b, 0\}$ for $a, b \in [0, 1]$ and $x \ge 1$, the third inequality follows from the monotonicity of I_h^k and the definition of function $\widetilde{V}_{k,h}$, the fourth inequality holds due to the definition of R_0 and $\check{A}_0$, the last inequality follows from the fact that $r(s_h^k, a_h^k) \geq 0$.

Substituting (D.31) into (D.30), we have

$$\widetilde{S}_m \leq \widetilde{A}_{m+1} + G + 2^{m+1} \cdot (R_0 + G + A_0 + \check{A}_0).$$

Thus, we complete the proof of Lemma D.15. ■

Lemma D.16 *Let $\widetilde{A}_m$, $\widetilde{S}_m$ be defined in (D.22), (D.18). With probability at least $1 - 2(M+1)\delta$, for all $m \in [M] \cup \{0\}$,*

$$\widetilde{A}_m \leq \sqrt{2\widetilde{S}_m \log(1/\delta)} + \frac{4}{3} \cdot \log(1/\delta).$$

We denote the corresponding event by $\mathcal{E}_{\mathrm{r}_3}$.

Proof The proof is equivalent to the proof of Lemma D.11. ■

Lemma D.17 *Let A_m, $\check{A}_m$, R_0, G be defined in (D.20), (D.21), (D.23), (D.24). On the event $\mathcal{E}_{\mathrm{r}_1} \cap \mathcal{E}_{\mathrm{r}_2} \cap \mathcal{E}_{\mathrm{r}_3} \cap \mathcal{E}_{\mathrm{var}'} \cap \mathcal{E}_c$, we have*

$$\begin{aligned}\widetilde{A}_1 &\leq 4\sqrt{(R_0 + 2G + A_0 + \check{A}_0)\log(1/\delta)} + 11 \cdot \log(1/\delta),\\ \widetilde{A}_0 &\leq 2\sqrt{(R_0 + 2G + A_0 + \check{A}_0)\log(1/\delta)} + 7 \cdot \log(1/\delta).\end{aligned}$$

Proof By Lemma D.15 and Lemma D.16, we have for all $m \in [M] \cup \{0\}$,

$$\widetilde{A}_m \leq \sqrt{2\left(\widetilde{A}_{m+1} + 2^{m+1} \cdot (R_0 + 2G + A_0 + \check{A}_0)\right)\log(1/\delta)} + \frac{4}{3} \cdot \log(1/\delta). \tag{D.32}$$

Applying Lemma E.5, we have

$$\begin{aligned}\widetilde{A}_1 &\leq \max\left\{11\log(1/\delta), 4\sqrt{(R_0 + 2G + A_0 + \check{A}_0)\log(1/\delta)} + 2\log(1/\delta)\right\}\\ &\leq 4\sqrt{(R_0 + 2G + A_0 + \check{A}_0)\log(1/\delta)} + 11\log(1/\delta).\end{aligned}$$

By (D.32), it can be further deduced that

$$\begin{aligned}\widetilde{A}_0 &\leq \sqrt{18\log(1/\delta) + 4\left(\sqrt{R_0 + 2G + A_0 + \check{A}_0} + \sqrt{\log(1/\delta)}\right)^2}\sqrt{\log(1/\delta)} + \frac{4}{3} \cdot \log(1/\delta)\\ &\leq 2\sqrt{(R_0 + 2G + A_0 + \check{A}_0)\log(1/\delta)} + 7 \cdot \log(1/\delta).\end{aligned}$$

Thus, we complete the proof of Lemma D.17. ■

Lemma D.18 *Let Q_0, S_m be defined in (D.19), (D.16). With probability at least $1-\delta$, it holds that*

$$Q_0 \le 2\widetilde{S}_0 + \widetilde{O}(K).$$

We denote the corresponding event by $\mathcal{E}_{\mathrm{r}_4}$.

Proof By the definition of Q_0, we have

$$Q_0 \le 2\widetilde{S}_0 + 2\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}V_{h+1}^{\pi_k}](s_h^k, a_h^k). \tag{D.33}$$

Note that for all $k \in [K]$,

$$\mathbb{E}_{\{(s_h,a_h)\}_{h\in[H]}\sim\pi_k}\left[\sum_{h=1}^{H}[\mathbb{V}V_{h+1}^{\pi_k}](s_h, a_h)\right] = \mathrm{Var}_{\{(s_h,a_h)\}_{h\in[H]}\sim\pi_k}\left[\sum_{h=1}^{H} r(s_h, a_h) - V_1^{\pi_k}(s_1^k)\right] \le 1, \tag{D.34}$$

where the last inequality holds due to the fact that $\sum_{h=1}^{H} r(s_h, a_h), V_1^{\pi_k}(s_1^k) \in [0, 1]$. In addition, the variance is upper bounded by:

$$\begin{aligned}\sum_{h=1}^{H}\mathrm{Var}\left[[\mathbb{V}V_{h+1}^{\pi_k}](s_h^k, a_h^k)\middle|\mathcal{F}_{k,1}\right] &\le \sum_{h=1}^{H}\mathbb{E}\left[\left([\mathbb{V}V_{h+1}^{\pi_k}](s_h^k, a_h^k)\right)^2\middle|\mathcal{F}_{k,1}\right]\\ &\le \sum_{h=1}^{H} 1\cdot\mathbb{E}\left[[\mathbb{V}V_{h+1}^{\pi_k}](s_h^k, a_h^k)\middle|\mathcal{F}_{k,1}\right]\\ &\le 1,\end{aligned}$$

where the last inequality holds due to (D.34). By Freedman's inequality (Lemma E.4), with probability at least $1-\delta/K$,

$$\sum_{h=1}^{H}[\mathbb{V}V_{h+1}^{\pi_k}](s_h^k, a_h^k) \le 1 + \sqrt{2\log(K/\delta)} + 2/3\cdot\log(K/\delta).$$

Using a union bound over $k \in [K]$, we can conclude that with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}V_{h+1}^{\pi_k}](s_h^k, a_h^k) \le \widetilde{O}(K). \tag{D.35}$$

Thus, we complete the proof of Lemma D.18 by substituting (D.35) into (D.33). ■

D.5. Regret Analysis

Proof [Proof of Theorem 3.3] We prove this theorem on the event $\mathcal{E}_{\mathrm{r}_1}\cap\mathcal{E}_{\mathrm{r}_2}\cap\mathcal{E}_{\mathrm{r}_3}\cap\mathcal{E}_{\mathrm{c}}\cap\mathcal{E}_{\mathrm{var}'}$, which occurs with probability at least $1-(4M+9)\delta$ by Lemmas D.10, D.11, D.16, D.3, D.4. On these events, we have the following decomposition of $\mathrm{Regret}(K)$,

$$\mathrm{Regret}(K) = \sum_{k=1}^{K}\left[V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\right]$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K} \Big[V_{k,1}(s_1^k) - V_1^{\pi_k}(s_1^k)\Big] \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \Big[V_{k,h}(s_h^k) - V_{k,h+1}(s_{h+1}^k)\Big] - \sum_{k=1}^{K} V_1^{\pi_k}(s_1^k) + G \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \cdot r(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \Big[V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [\mathbb{P}V_{k,h+1}](s_h^k, a_h^k)\Big] \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \Big[[\mathbb{P}V_{k,h+1}](s_h^k, a_h^k) - V_{k,h+1}(s_{h+1}^k)\Big] - \sum_{k=1}^{K} V_1^{\pi_k}(s_1^k) + G \\
&\leq R_0 + A_0 + G + \underbrace{\sum_{k=1}^{K}\left(\sum_{h=1}^{H} r(s_h^k, a_h^k) - V_1^{\pi_k}(s_1^k)\right)}_{I_1},
\end{aligned}
$$

where the first inequality holds due to Lemma D.7, the second inequality holds due to the monotonicity of indicator function I_h^k, the last inequality holds due to $I_h^k \leq 1$ and $r(s_h^k, a_h^k) \geq 0$.

For the term I_1, we have

$$
\begin{aligned}
\sum_{k=1}^{K}\left(\sum_{h=1}^{H} r(s_h^k, a_h^k) - V_1^{\pi_k}(s_1^k)\right) &= \sum_{k=1}^{K}\sum_{h=1}^{H}\Big[V_h^{\pi_k}(s_h^k) - [\mathbb{P}V_{h+1}^{\pi_k}](s_h^k, a_h^k)\Big] - \sum_{k=1}^{K} V_1^{\pi_k}(s_1^k) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\Big[V_{h+1}^{\pi_k}(s_{h+1}^k) - [\mathbb{P}V_{h+1}^{\pi_k}](s_h^k, a_h^k)\Big] \\
&\leq |A_0| + |\check{A}_0| + |\widetilde{A}_0|, \qquad \text{(D.36)}
\end{aligned}
$$

where the inequality holds due to $|x+y+z| \leq |x| + |y| + |z|$.

For the term R_0, according to Lemma D.8, we have

$$
\begin{aligned}
R_0 &\leq \widetilde{O}\left(d\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_{h,k}^2} + d\right) \\
&\leq \widetilde{O}\left(d\sqrt{Q_0} + d\sqrt{\check{S}_0} + d\right) \\
&\leq \widetilde{O}\left(d\sqrt{Q_0} + d\sqrt{\check{A}_1 + G + R_0 + A_0} + d\right) \\
&\leq \widetilde{O}\left(d\sqrt{Q_0} + d\sqrt{G + R_0 + \sqrt{Q_0} + \sqrt{R_0}} + d\right) \\
&\leq \widetilde{O}\left(d\sqrt{Q_0} + d^2\right), \qquad \text{(D.37)}
\end{aligned}
$$

where the first inequality follows from Lemma D.8, the second inequality follows from the definition of $\check{S}_0$ and Q_0, the third inequality holds due to Lemma D.9, the fourth inequality is obtained by applying Lemma D.13 and D.14, the last inequality follows from the fact that $x \leq a\sqrt{x} + b \Rightarrow x \leq a^2 + 2b$ and the upper bound of G in Lemma D.1.

For the term A_0, by Lemma D.14, we have

$$A_0 \le 132\log(1/\delta) + 28\sqrt{R_0\log(1/\delta)} + 40\sqrt{G\log(1/\delta)} + 8\sqrt{Q_0\log(1/\delta)}.$$

Putting everything together, we have

$$\mathrm{Regret}(K) \le \widetilde{O}\left(d\sqrt{Q_0} + d^2\right).$$

Thus, we complete the proof of Theorem 3.3. ■

Corollary D.19 *Under the same condition of Theorem 3.3, with probability at least $1-(4M+10)\delta$, the regret of Algorithm 2 is bounded by:*

$$\mathrm{Regret}(K) \le \widetilde{O}\left(d\sqrt{K} + d^2\right).$$

Proof We prove this corollary on the event $\mathcal{E}_{\mathrm{r}_1} \cap \mathcal{E}_{\mathrm{r}_2} \cap \mathcal{E}_{\mathrm{r}_3} \cap \mathcal{E}_{\mathrm{r}_4} \cap \mathcal{E}_{\mathrm{c}} \cap \mathcal{E}_{\mathrm{var}'}$, which occurs with probability at least $1-(4M+10)\delta$ by Lemmas D.10, D.11, D.16, D.18, D.3, D.4.

By the definition of $\mathcal{E}_{\mathrm{r}_4}$ in Lemma D.18, we have

$$\begin{aligned}
Q_0 &\le 2S_0 + \widetilde{O}(K)\\
&\le 2\widetilde{A}_1 + G + 2(R_0 + G + A_0) + \widetilde{O}(K)\\
&\le 8\sqrt{(R_0 + 2G + A_0 + \check{A}_0)\log(1/\delta)} + 22\cdot\log(1/\delta) + G + 2(R_0 + G + A_0) + \widetilde{O}(K)\\
&\le \widetilde{O}\left(d\sqrt{Q_0} + d^2 + K\right)\\
&\le \widetilde{O}(K + d^2),
\end{aligned}$$

where the second inequality follows from Lemma D.15, the third inequality holds due to D.17, the fourth inequality is derived by Lemma D.14, Lemma D.13, (D.37) and omitting the lower order terms, the last inequality holds due to the fact that $x \le a\sqrt{x} + b \Rightarrow x \le a^2 + 2b$.

By Theorem 3.3, we can obtain

$$\mathrm{Regret}(K) \le \widetilde{O}(d\sqrt{Q_0} + d^2) \le \widetilde{O}(d\sqrt{K} + d^2).$$

Thus, we complete the proof of Corollary 3.4. ■

Appendix E. Auxiliary Lemmas

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

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

Lemma E.2 (Lemma 11, Abbasi-Yadkori et al. 2011) *For any $\lambda > 0$ and sequence $\{\mathbf{x}_k\}_{k=1}^K \subset \mathbb{R}^d$ for $k \in [K]$, define $\mathbf{Z}_k = \lambda \mathbf{I} + \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{x}_i^\top$. Then, provided that $\|\mathbf{x}_k\|_2 \leq L$ holds for all $k \in [K]$, we have*

$$\sum_{k=1}^K \min\big\{1, \|\mathbf{x}_k\|_{\mathbf{Z}_k^{-1}}^2\big\} \leq 2d \log\big(1 + KL^2/(d\lambda)\big).$$

Lemma E.3 (Lemma 12, Abbasi-Yadkori et al. 2011) *Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$ are two positive definite matrices satisfying that $\mathbf{A} \succeq \mathbf{B}$, then for any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_{\mathbf{A}} \leq \|\mathbf{x}\|_{\mathbf{B}} \cdot \sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.*

Lemma E.4 (Freedman 1975) *Let $M, v > 0$ be fixed constants. Let $\{x_i\}_{i=1}^n$ be a stochastic process, $\{\mathcal{G}_i\}_i$ be a filtration so that for all $i \in [n]$, x_i is $\mathcal{G}_i$-measurable, while almost surely*

$$\mathbb{E}\left[x_i|\mathcal{G}_{i-1}\right] = 0, \quad |x_i| \leq M, \quad \sum_{i=1}^n \mathbb{E}[x_i^2|\mathcal{G}_{i-1}] \leq v.$$

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

$$\sum_{i=1}^n x_i \leq \sqrt{2v \log(1/\delta)} + 2/3 \cdot M \log(1/\delta).$$

Lemma E.5 (Lemma 2, Zhang et al. 2021a) *Let $\lambda_1, \lambda_2, \lambda_4 > 0$, $\lambda_3 \geq 1$, and $i' = \log_2 \lambda_1$. Let $a_1, a_2, \cdots, a_{i'}$ be non-negative reals such that $a_i \leq \lambda_1$ and $a_i \leq \lambda_2 \sqrt{a_{i+1} + 2^{i+1}\lambda_3} + \lambda_4$ hold for any $1 \leq i \leq i'$. Then we have that*

$$a_1 \leq \max\left\{\left(\lambda_2 + \sqrt{\lambda_2^2 + \lambda_4}\right)^2, \lambda_2\sqrt{8\lambda_3} + \lambda_4\right\}.$$