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 $O(d\sqrt{\sum_{k=1}^{K} \sigma_k^2} + d)$ regret, where $\sigma_k^2$ 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 $\tilde{O}(d)$ 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., $\{\sigma^2_i\}_{i=1}^K$ 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 $\tilde{O}(d \sqrt{\sum_{k=1}^K \sigma_k^2} + d)$, where $\sigma_k^2$ 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 $\tilde{O}(d)$, which corresponds to the benign regime. In the worst case when the noise is $R$-sub-Gaussian, the above regret reduces to $\tilde{O}(dR \sqrt{K} + d)$, 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Regret (General-Case)</th>
<th>Regret (Deterministic-Case)</th>
<th>Efficiency</th>
<th>Variances</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConfidenceBall(_2) (Dani et al., 2008)</td>
<td>(\tilde{O}(d\sqrt{K}))</td>
<td>(\tilde{O}(d\sqrt{K}))</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>OFUL (Abbasi-Yadkori et al., 2011)</td>
<td>(\tilde{O}(d\sqrt{K}))</td>
<td>(\tilde{O}(d\sqrt{K}))</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>Weighted OFUL (Zhou et al., 2021a)</td>
<td>(\tilde{O}\left(d\sqrt{\sum_{k=1}^{K} \sigma_k^2} + \sqrt{dK}\right))</td>
<td>(\tilde{O}(\sqrt{dK}))</td>
<td>Yes</td>
<td>Known</td>
</tr>
<tr>
<td>Weighted OFUL+ (Zhou and Gu, 2022)</td>
<td>(\tilde{O}\left(d\sqrt{\sum_{k=1}^{K} \sigma_k^2} + d\right))</td>
<td>(O(d \cdot \log(d, K)))</td>
<td>Yes</td>
<td>Known</td>
</tr>
<tr>
<td>VOFUL (Zhang et al., 2021b)</td>
<td>(O\left(d^{1.5}\sqrt{\sum_{k=1}^{K} \sigma_k^2} + d^5\right))</td>
<td>(O(d^5 \cdot \log(d, K)))</td>
<td>No</td>
<td>Unknown</td>
</tr>
<tr>
<td>VOFUL(_2) (Kim et al., 2021)</td>
<td>(O\left(d^{1.5}\sqrt{\sum_{k=1}^{K} \sigma_k^2} + d^3\right))</td>
<td>(O(d^2 \cdot \log(d, K)))</td>
<td>No</td>
<td>Unknown</td>
</tr>
<tr>
<td>SAVE (Theorem 2.3)</td>
<td>(\tilde{O}\left(d\sqrt{\sum_{k=1}^{K} \sigma_k^2} + d\right))</td>
<td>(O(d \cdot \log(d, K)))</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

- 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 planning 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. \( \text{Var}_{r_h} \) 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 \( \tilde{O}(d\sqrt{K} + d^2) \) in the worst case, which matches the nearly minimax optimal horizon-free regret in linear mixture MDPs.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Regret (General-Case)</th>
<th>Variance-Dependent</th>
<th>Assumption</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCRL-VTR</td>
<td>( \tilde{O}(dH^3K) )</td>
<td>No</td>
<td>Homogeneous( \sum_{h=1}^H r_h \leq H )</td>
<td>Yes</td>
</tr>
<tr>
<td>(Ayoub et al., 2020; Jia et al., 2020)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCRL-VTR+</td>
<td>( \tilde{O}(\sqrt{d^2H^3 + dH^4\sqrt{K}} + d^2H^3 + d^3H^2) )</td>
<td>No</td>
<td>Inhomogeneous( \sum_{h=1}^H r_h \leq H )</td>
<td>Yes</td>
</tr>
<tr>
<td>(Zhou et al., 2021a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarLin</td>
<td>( \tilde{O}\left(d^{4.5}\sqrt{K} + d^2\right) )</td>
<td>No</td>
<td>Homogeneous( \sum_{h=1}^H r_h \leq H )</td>
<td>No</td>
</tr>
<tr>
<td>(Zhang et al., 2021b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarLin2</td>
<td>( \tilde{O}(d\sqrt{K} + d^2) )</td>
<td>No</td>
<td>Homogeneous( \sum_{h=1}^H r_h \leq 1 )</td>
<td>Yes</td>
</tr>
<tr>
<td>(Kim et al., 2021)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HF-UCRL-VTR+</td>
<td>( \tilde{O}(d\sqrt{K} + d^2) )</td>
<td>No</td>
<td>Homogeneous( \sum_{h=1}^H r_h \leq 1 )</td>
<td>Yes</td>
</tr>
<tr>
<td>(Zhou and Gu, 2022)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCRL-AVE (Theorem 2.3)</td>
<td>( \tilde{O}\left(d\sqrt{\text{Var}K} + d^2\right) )</td>
<td>Yes</td>
<td>Homogeneous( \sum_{h=1}^H r_h \leq 1 )</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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 \( x \in \mathbb{R}^d \) and a positive semi-definite matrix \( \Sigma \in \mathbb{R}^{d \times d} \), we denote by \( \|x\|_\Sigma = \sqrt{x^\top \Sigma x} \) the vector’s Euclidean norm and define \( \|x\|_\Sigma = \sqrt{x^\top \Sigma x} \). For \( x, y \in \mathbb{R}^d \), let \( x \odot y \) be the Hadamard (componentwise) product of \( x \) and \( y \). 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 \( a_n \leq C b_n \) holds for all \( n \geq 1 \) and write \( a_n = \Omega(b_n) \) if there exists an absolute constant \( C > 0 \) such that \( a_n \geq C b_n \) holds for all \( n \geq 1 \). We use \( \tilde{O}() \) to further hide the polylogarithmic factors. We use \( \mathbb{I}\{\cdot\} \) to denote the indicator function. For \( a, b \in \mathbb{R} \) satisfying \( a \leq b \), we use \( [x]_{a,b} \) to denote the truncation function \( x \cdot \mathbb{I}\{a \leq x \leq b\} + a \cdot \mathbb{I}\{x < a\} + b \cdot \mathbb{I}\{x > b\} \).

2. Variance-Aware Learning for Heteroscedastic Linear Bandits

In this section, we propose a computationally efficient variance-aware algorithm, dubbed SAVE (Suplin + Adaptive Variance-aware Exploration), 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 $D_k \subseteq \mathbb{R}^d$ where each element represents a feasible action that can be selected by the agent; (2) the agent observes $D_k$ and selects $a_k \in D_k$; and (3) the environment generates the stochastic noise $\epsilon_k$ at round $k$ and reveal the stochastic reward $r_k = \langle \theta^*, a_k \rangle + \epsilon_k$ to the agent. We assume that there exists a uniform bound $A > 0$ for the $\ell_2$ norm of the feasible actions, i.e., for all $k \in [K], a \in D_k$, it holds that $\|a\|_2 \leq A$. Following Zhou et al. (2021a), we assume the following condition on the random noise $\epsilon_k$ at each round $k$:

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

Without loss of generality, we assume that the size of $D_k$ is finite and is bounded by $|D|$ for all $k \in [K]$. If the size of $D_k$ 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:

$$\text{Regret}(K) = \sum_{k \in [K]} (\langle a_k^*, \theta^* \rangle - \langle a_k, \theta^* \rangle), \quad \text{where } a_k^* = \underset{a \in D_k}{\text{argmax}} \langle a, \theta^* \rangle. \quad (2.2)$$

2.2. Technical Challenges

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

The first approach bounds $\langle a_k, \theta_k - \theta^* \rangle$ with $\|a_k\|_{\Sigma_k^{-1}} \|\theta_k - \theta^*\|_{\Sigma_k}$ by Cauchy-Schwarz inequality. Zhou et al. (2021a); Zhou and Gu (2022) constructed $C_k$ as an ellipsoid centering at $\hat{\theta}_k$, which is the solution to a weighted linear regression problem over the past contexts $a_i$, and their weight is based on the upper bound of the variance of heteroscedastic noise $\sigma_k^2$. Then they bound $\|\theta_k - \theta^*\|_{\Sigma_k}$ by $\|\theta_k - \hat{\theta}_k\|_{\Sigma_k}$ and $\|\theta^* - \hat{\theta}_k\|_{\Sigma_k}$ 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 $\sigma_k^2$ is known.

The second approach (Zhang et al., 2021b; Kim et al., 2021) follows the test-based framework. Instead of constructing $C_k$ as an ellipsoid centering at a least square estimator $\hat{\theta}_k$ for each round $k$, Zhang et al. (2021b); Kim et al. (2021) constructed $C_k$ 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 $a_k$. The limitation of their approach is that, in order to have a uniform upper bound on $\langle a_k, \theta_k - \theta^* \rangle$, they have to cover all possible directions of $a_k$, 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 \( \{G_k\}_{k=1}^{\infty} \) be a filtration, and \( \{x_k, \eta_k\}_{k \geq 1} \) be a stochastic process such that \( x_k \in \mathbb{R}^d \) is \( G_k \)-measurable and \( \eta_k \in \mathbb{R} \) is \( G_{k+1} \)-measurable. Let \( L, \sigma, \lambda, \epsilon > 0, \mu^* \in \mathbb{R}^d \). For \( k \geq 1 \), let \( y_k = \langle \mu^*, x_k \rangle + \eta_k \), where \( \eta_k, x_k \) satisfy

\[
\mathbb{E}[\eta_k | G_k] = 0, \quad |\eta_k| \leq R, \quad \sum_{i=1}^{k} \mathbb{E}[\eta_i^2 | G_i] \leq v_k, \quad \text{for } \forall k \geq 1.
\]

For \( k \geq 1 \), let \( Z_k = \lambda I + \sum_{i=1}^{k} x_i x_i^\top \), \( b_k = \sum_{i=1}^{k} y_i x_i \), \( \mu_k = Z_k^{-1} b_k \), and

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

where \( \rho \geq \sup_{k \geq 1} \|x_k\|_{Z_k^{-1}} \). Then, for any \( 0 < \delta < 1 \), we have with probability at least \( 1 - \delta \) that,

\[
\forall k \geq 1, \quad \|\sum_{i=1}^{k} x_i \eta_i\|_{Z_k^{-1}} \leq \beta_k, \quad \|\mu_k - \mu^*\|_{Z_k} \leq \beta_k + \sqrt{\lambda}\|\mu^*\|_2.
\]

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., \( \|\mu_k - \mu^*\|_{Z_k} \leq \tilde{O}(\sigma \sqrt{d} + R) \), where \( \sigma^2 \geq \sup_{k \geq 1} \mathbb{E}[\eta_i^2 | G_i] \). In contrast, the upper bound provided by Theorem 2.1 depends on the maximum of \( \|x_k\|_{Z_k^{-1}} \), which is of the order \( \tilde{O}(\sqrt{d}/k) \) under certain conditions (Carpentier et al., 2020). For these cases, our upper bound for \( \|\mu_k - \mu^*\|_{Z_k} \) scales as \( \tilde{O}(\sqrt{d} \cdot v_k/k + R \cdot \sqrt{d}/k) \), which is more fine-grained and strictly tighter than the previous upper bounds when \( k \geq d \).

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 \( \|a_k\|_{\Sigma_k^{-1}} \) to be small, where \( \Sigma_k \) is the covariance matrix of \( a_k \). 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 \( a_k \) which maximizes \( \langle a, \theta^* \rangle \). Since \( \theta^* \) is unknown, the selection process is based on \( L \) number of estimates of \( \theta^* \), which we denote them by \( \hat{\theta}_{k, \ell}, \ell \in [L] \). \( \hat{\theta}_{k, \ell} \) 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 $a$ 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: $\hat{\Sigma}_{1,\ell} \leftarrow 2^{-2\ell} \cdot I, \hat{b}_{1,\ell} \leftarrow 0, \hat{\theta}_{1,\ell} \leftarrow 0, \Psi_{k,\ell} \leftarrow \emptyset, \beta_{1,\ell} \leftarrow 2^{-\ell+1}$ for all $\ell \in [L]$.
3: for $k = 1, \ldots, K$ do
4:    Observe $D_k$.
5:    Let $A_{k,1} \leftarrow D_k, \ell \leftarrow 1$.
6:    while $a_k$ is not specified do
7:        if $\|a\|_{\hat{\Sigma}_{k,\ell}}^{-1} \leq \alpha$ for all $a \in A_{k,\ell}$ then
8:            Choose $a_k \leftarrow \arg\max_{a \in A_{k,\ell}} \langle a, \hat{\theta}_{k,\ell} \rangle + \hat{\beta}_{k,\ell} \|a\|_{\hat{\Sigma}_{k,\ell}}^{-1}$ and observe $r_k$.
9:        else if $\|a\|_{\hat{\Sigma}_{k,\ell}}^{-1} \leq 2^{-\ell}$ for all $a \in A_{k,\ell}$ then
10:           $A_{k,\ell+1} \leftarrow \{ a \in A_{k,\ell} | \langle a, \hat{\theta}_{k,\ell} \rangle \geq \max_{a' \in A_{k,\ell}} \langle a', \hat{\theta}_{k,\ell} \rangle - 2 \cdot 2^{-\ell} \hat{\beta}_{k,\ell} \}$.
11:        else
12:            Choose $a_k$ such that $\|a_k\|_{\hat{\Sigma}_{k,\ell}}^{-1} > 2^{-\ell}$ and observe $r_k$.
13:            Compute the weight: $w_k \leftarrow 2^{-\ell}/\|a_k\|_{\hat{\Sigma}_{k,\ell}}^{-1}$.
14:            Update the index sets: $\Psi_{k+1,\ell} \leftarrow \Psi_{k,\ell} \cup \{k\}$ and $\Psi_{k+1,\ell'} \leftarrow \Psi_{k,\ell'}$ for $\ell' \in [L] \setminus \{\ell\}$.
15:        end if
16:    $\ell \leftarrow \ell + 1$.
17: end while
18: For $\ell \in [L]$ such that $\Psi_{k+1,\ell} \neq \Psi_{k,\ell}$, update the estimators as follows:

$$\hat{\Sigma}_{k+1,\ell} + w_k^2 a_k a_k^\top, \hat{b}_{k+1,\ell} \leftarrow \hat{b}_{k,\ell} + w_k^2 r_k a_k, \hat{\theta}_{k+1,\ell} \leftarrow \hat{\Sigma}_{k+1,\ell}^{-1} \hat{b}_{k+1,\ell}.$$ 

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

their corresponding rewards. Starting from $\ell = 1$, the decision set $A_{k,\ell}$ will keep ‘shrinking’ by eliminating all $a \in A_{k,\ell}$ which are unlikely to be the maximizer of $\langle a, \theta^* \rangle$ (notably, since $\theta^*$ is unknown, here $\theta^*$ needs to be replaced by $\hat{\theta}_{k,\ell}$, as displayed in Line 11). The elimination process will not stop until some action $a$ with large uncertainty $\|a\|_{\hat{\Sigma}_{k,\ell}}^{-1}$ emerges. Then Algorithm 1 will either select the action $a$ 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 $a_k$ enjoys a large uncertainty.

Construction of the estimate $\hat{\theta}_{k,\ell}$ The first difference between our algorithm and SupLinUCB is the construction of the estimate $\hat{\theta}_{k,\ell}$. Unlike the unweighted ridge regression estimator applied in
SupLinUCB, we employ a weighted ridge-regression estimator as follows
\[
\forall k \in [K] \text{ and } \ell \in [L], \quad \widehat{\theta}_{k,\ell} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i \in \Psi_{k,\ell}} w_i^2 (r_i - (\theta, a_i))^2 + 2^{-2\ell} \lambda \|\theta\|_2^2,
\]
where the weight \(w_i\) is chosen such that for \(\forall \ell \in [L] \) and \(i \in \Psi_{k,\ell}\), \(\|w_i a_i\| \Sigma_{i,\ell}^{-1} = 2^{-\ell}\). We explain here why we want to adopt such a weighted regression scheme. In particular, the estimate \(\widehat{\theta}_{k,\ell}\) can be regarded as \(\mu_k\) in Theorem 2.1. By our construction of \(w_i\), we can ensure that the context \(x_i\) in Theorem 2.1, which is \(w_i a_i\) here, enjoys a uniform upper bound on the uncertainty, i.e., \(\|x_i\| \leq 2^{-\ell}\). Such a result can further imply that \(\|\widehat{\theta}_{k,\ell} - \theta^*\| \Sigma_{k,\ell}^{-1}\) 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 \(\theta_k = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{k-1} \frac{1}{\sigma_i^2} (r_i - (\theta, a_i))^2 + \lambda \|\theta\|_2^2\), 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 \(\|a_i\| \Sigma_{i,\ell}^{-1}\). 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 \(w_i = O(1/\|a_i\|^{1/2})\) 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 \(\widehat{\theta}_{k,\ell}\) following Theorem 2.1, which leads to a confidence bound of \(\theta^*\), i.e., \(\theta^* : \|\theta - \widehat{\theta}_{k,\ell}\| \Sigma_{k,\ell}^{-1} \leq O(2^{-\ell} \cdot \sqrt{\sum_{i \in \Psi_{k,\ell}} w_i^2 \sigma_i^2 + 2^{-\ell} R})\). This can be used in the arm selection step (Line 8 and Line 11). However, such a confidence set requires the knowledge of variances \(\sigma_i^2\) apriori. To address this issue, we need to replace \(\sigma_i^2\) with their empirical estimator. In detail, since \(\sigma_i^2 = \mathbb{E}[(r_i - \langle \theta^*, \vec{a}_i \rangle)^2| \vec{a}_{1:i-1}, \epsilon_{1:i-1}]\), we simply use an one-point plug-in estimator \((r_i - \langle \widehat{\theta}_{k,\ell}, \vec{a}_i \rangle)^2\). With such an estimator, we define the confidence radius \(\beta_{k+1,\ell}\) at round \(k + 1\) and layer \(\ell\) as
\[
\widehat{\beta}_{k+1,\ell} := 16 \cdot 2^{-\ell} \sqrt{8 \text{Var}_{k+1,\ell} + 6 R^2 \log(4(k + 1)^2 L/\delta) + 2^{-2\ell+4} \log(4k^2 L/\delta)} + 6 \cdot 2^{-\ell} R \log(4k^2 L/\delta) + 2^{-\ell+1},
\]
where \(\text{Var}_{k+1,\ell} := \sum_{i \in \Psi_{k+1,\ell}} w_i^2 (r_i - (\widehat{\theta}_{k+1,\ell}, a_i))^2, \quad 2^\ell \geq 64 \sqrt{\log(4(k + 1)^2 L/\delta)}\), otherwise.

We would like to emphasize that although our used one-point estimator \((r_i - \langle \widehat{\theta}_{k,\ell}, \vec{a}_i \rangle)^2\) might be an inaccurate estimator of the target \(\sigma_i^2\) for some round \(i\), the weighted summation of the one-point estimators \(\sum w_i^2 (r_i - \langle \theta_{k,\ell}, a_i \rangle)^2\) 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 \((r_i - \langle \theta_{k,\ell}, a_i \rangle)^2\) 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 \( \overline{\text{Var}}_{k+1, \ell} \) by \( \overline{\text{Var}}_{k+1, \ell} = \sum_{i \in \Psi_{k+1, \ell}} w_i^2 r_i^2 - 2\widehat{\theta}_{k+1, \ell}^\top \sum_{i \in \Psi_{k+1, \ell}} w_i r_i a_i + \widehat{\theta}_{k+1, \ell}^\top \left( \sum_{i \in \Psi_{k+1, \ell}} w_i^2 a_i a_i^\top \right) \widehat{\theta}_{k+1, \ell} \), 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 \( w_i r_i a_i \) 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 \( O(K|D|Ld^2) \), where \( L \) is actually a logarithmic term according to the choice of \( \alpha \) 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 \( a \in D_k, \|a\|_2 \leq A, \|\theta^*\|_2 \leq 1, \langle a, \theta^* \rangle \in [-1, 1] \). If \( \{\beta_{k, \ell}\}_{k, \ell \in [K]} \) 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 \( 1 - 3\delta \):

\[
\text{Regret}(K) = \tilde{O}\left(d \sqrt{\sum_{k=1}^{K} \sigma_k^2 + dR + d} \right).
\]

**Remark 2.4** If we treat \( R \) as a constant, the regret can be simplified as \( \tilde{O}\left(d \sqrt{\sum_{k=1}^{K} \sigma_k^2 + d} \right) \).

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 \( \sigma_k \). Compared with VOFUL2 (Kim et al., 2021), our SAVE algorithm improves the regret from \( \tilde{O}\left(d^{1.5} \sqrt{\sum_{k=1}^{K} \sigma_k^2 + d^2} \right) \) to \( \tilde{O}\left(d \sqrt{\sum_{k=1}^{K} \sigma_k^2 + d} \right) \). Furthermore, VOFUL2 needs to perform the arm elimination for each possible direction \( \mu \) 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 \( \tilde{O}(d) \) 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 \( M = (\mathcal{S}, \mathcal{A}, H, r, \mathbb{P}) \). 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 $\theta^*$.

1: Set $L = \lceil \log_2(1/\alpha) \rceil$.
2: Initialize: $\Sigma_{0,H+1,\ell} \leftarrow 2^{-2\ell} \lambda \cdot I$, $\hat{b}_{0,H+1,\ell} \leftarrow 0$, $\hat{\theta}_{1,\ell} \leftarrow 0$ for all $\ell \in [L]$.
3: for $k = 1, \ldots, K$ do
4:   $V_{k,H+1}(\cdot) \leftarrow 0$.
5:   Update the current estimators: $\hat{\Sigma}_{k,1,\ell} \leftarrow \hat{\Sigma}_{k-1,H+1,\ell}$, $\hat{b}_{k,1,\ell} \leftarrow \hat{b}_{k-1,H+1,\ell}$, $\hat{\theta}_{k,\ell} \leftarrow \frac{1}{\hat{\Sigma}_{k,1,\ell}} \hat{b}_{k,1,\ell}$ for all $\ell \in [L]$.
6:   Compute $\hat{\beta}_{k,\ell}$ according to (3.4).
7:   for $h = H, \ldots, 1$ do
8:     $Q_{k,h}(\cdot, \cdot) \leftarrow \min \left\{ 1, \min_{\ell \in [L]} \left[ r(\cdot, \cdot) + \langle \hat{\theta}_{k,\ell}, \phi_{V_{k,h+1}}(\cdot, \cdot) \rangle + \hat{\beta}_{k,\ell} \| \phi_{V_{k,h+1}}(\cdot, \cdot) \|_{\Sigma_{k,1,\ell}} \right] \right\}$.
9:     $\pi_k(\cdot, h) \leftarrow \arg\max_{a \in A} Q_{k,h}(\cdot, a)$, $V_{k,h}(\cdot) \leftarrow \max_{a \in A} Q_{k,h}(\cdot, a)$.
10: end for
11: Observe $s_k$.
12: for $h = 1, \ldots, H$ do
13:   Take action $a_k^h \leftarrow \pi_k(s_k^h, h)$ and observe $s_k^h$.
14:   $L_{k,h} \leftarrow \left\{ \ell \in [L] \left| \| \phi_{V_{k,h+1}}(s_k^h, a_k^h)\|_{\Sigma_{k,1,h,\ell}} \geq 2^{-\ell} \right. \right\}$.
15:   Set $\ell_{k,h} \leftarrow \left\{ \begin{array}{ll} L + 1, & \text{if } L_{k,h} = \emptyset, \\
16: \min (L_{k,h}), & \text{otherwise}. \end{array} \right.$
17:   if $\ell_{k,h} \neq L + 1$ then
18:     $w_{k,h} \leftarrow 2^{-\ell_{k,h}} / \| \phi_{V_{k,h+1}}(s_k^h, a_k^h)\|_{\Sigma_{k,1,h,\ell}}$.
19:     $\hat{\Sigma}_{k+1,h,\ell_{k,h}} \leftarrow \hat{\Sigma}_{k,h,\ell_k,h} + w_{k,h}^2 k_{h+1} \phi_{V_{k,h+1}}(s_k^h, a_k^h) \phi_{V_{k,h+1}}(s_k^h, a_k^h)^\top$.
20:     $\hat{b}_{k+1,h,\ell_{k,h}} \leftarrow \hat{b}_{k,h,\ell_k,h} + w_{k,h} k_{h+1} \phi_{V_{k,h+1}}(s_k^h, a_k^h) \phi_{V_{k,h+1}}(s_k^h, a_k^h)$.
21:   end if
22: end for
23: 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 $\pi$ 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

\[
Q_{h}^{\pi}(s, a) = r(s, a) + \mathbb{E} \left[ \sum_{h' = h+1}^{H} r(s_{h'}, \pi(s_{h'}, h')) \left| s_h = s, a_h = a \right. \right], \quad V_{h}^{\pi}(s) = Q_{h}^{\pi}(s, \pi(s, h)),
\]

where $s_{h'+1} \sim \mathbb{P}(\cdot | s', a_h)$. We further define the optimal value function $V_{h}^{*}$ and the optimal action-value function $Q_{h}^{*}$ as $V_{h}^{*}(s) = \max_{a} V_{h}^{\pi}(s)$ 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 $[PV](s, a) = \mathbb{E}_{s' \sim \mathbb{P}(\cdot | s, a)} V(s')$. Therefore, for each stage $h \in [H]$ and policy $\pi$, we have the following Bellman equation, as well as the Bellman optimality equation:

\[
Q_{h}^{\pi}(s, a) = r(s, a) + [PV_{h}^{\pi}](s, a), \quad Q^{*}(s, a) = r(s, a) + [PV_{h+1}^{*}](s, a),
\]
where $V_{h+1}^π(\cdot) = V_{h+1}^*(\cdot) = 0$. At the beginning of episode $k$, the agent chooses a policy $\pi$ 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 $\pi$ 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 $\pi$, let $(s_h, a_h)_h^H_{h=1}$ be one trajectory following $\pi$, then $\sum_{h \in [H]} r(s_h, a_h) \leq 1$ almost surely.

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

$$\text{Var}_K = \sum_{k=1}^{K} \sum_{h=1}^{H} [\nabla V^*_k(s_h^k, a_h^k)].$$  \hspace{1cm} (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 $\text{Var}_K = 0$. 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 $\mathcal{M}(S, A, H, r, \mathbb{P})$ is a homogeneous, episodic $B$-bounded linear mixture MDP if there exists vectors $\theta^* \in \mathbb{R}^d$ with $\|\theta^*\|_2 \leq B$ and $\phi(\cdot, \cdot)$ satisfying (3.2), such that for each $(s, a) \in S \times A$, $s' \in S$ and stage $h \in [H]$, $\mathbb{P}(s'|s, a) = \langle \phi(s'|s, a), \theta^* \rangle$. Moreover, $\phi$ satisfies that for any bounded function $V : S \to [0, 1]$ and any tuple $(s, a) \in S \times A$,

$$\|\phi_V(s, a)\|_2 \leq 1,$$  \hspace{1cm} (3.2)

where $\phi_V(s, a) := \sum_{s' \in S} \phi(s'|s, a)V(s')$.

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

$$\text{Regret}(K) = \sum_{k \in [K]} [V^*_k(s^k_1) - V^*_k(s^k_1)].$$

### 3.2. The Proposed Algorithm

We present an adaptive variance-aware algorithm named UCRL with Adaptive Variance-Aware Exploration (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 $\mathbb{P}V_{k,h+1}(s, a) = \langle \phi_{V_{k,h+1}}(s, a), \theta^* \rangle$ is a linear function of the feature $\phi_{V_{k,h+1}}(s, a)$, and estimates $\mathbb{P}V_{k,h+1}(s, a)$ by a plug-in estimator $\langle \phi_{V_{k,h+1}}(s, a), \hat{\theta}_k \rangle$, where $\hat{\theta}_k$ is the estimate of $\theta^*$.

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 $\hat{\theta}_k$: instead of using a single estimate, Algorithm 2 maintains $L$ estimates $\hat{\theta}_{k,t}$, 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 \( \hat{\theta}_{k,\ell} \) to build the value function estimate \( Q_{k,h} \), then we show how Algorithm 2 updates \( \hat{\theta}_{k,\ell} \) accordingly. Algorithm 2 constructs \( Q_{k,h} \) as the minimum of \( L \) optimistic estimates computed by \( \langle \hat{\theta}_{k,\ell}, \phi_{V_{k,h}} \rangle \). The minimum step makes the estimate \( Q_{k,h} \) tighter than that in UCRL-VTR.

Similar to Algorithm 1, \( \hat{\theta}_{k,\ell} \) is the solution to some regression problem over the features \( \phi_{V_{k,h}} \) and their corresponding target values. For simplicity, we define the following subsets of \([K] \times [H]\):

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

which represents the indices of feature vectors in layer \( \ell \) at the beginning of round \( k \). Note that \( \hat{\theta}_{k,\ell} \) will be updated if the feature \( \phi_{V_{k,h+1}} \left( s_h^k, a_h^k \right) \) 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 \( \phi_{V_{k,h+1}} \left( s_h^k, a_h^k \right) \) 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, \( \hat{\theta}_{k,\ell} \) 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 \( \| w_{k,h} \phi_{V_{k,h+1}} \left( s_h^k, a_h^k \right) \| \Sigma_{k,h,\ell,k,h}^{-1} = 2^{-\ell_{k,h}}, \) 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 \( \theta^* \) without knowing the variance of value functions \( V_{k,h+1} \). Here we take the same approach: to replace the variance of \( V_{k,h+1} \), \( \mathbb{P}[V_{k,h+1} - \mathbb{E}V_{k,h+1}]^2 \) with its one-point empirical estimate \( (V_{i,h+1} - \langle \hat{\theta}_{k,\ell}, \phi_{V_{i,h+1}} \rangle)^2 \). In detail, the confidence radius is:

\[
\tilde{\beta}_{k,\ell} := 16 \cdot 2^{-\ell} \sqrt{8 \tilde{\text{var}}_{k,\ell} + 8 \log(4k^2H^2L/\delta) + 2^{-2\ell+5} \cdot \lambda B^2} \log(4k^2H^2L/\delta) + 6 \cdot 2^{-\ell} \log(4k^2H^2L/\delta) + 2^{-\ell} \sqrt{\lambda} \cdot B,
\]

where \( \tilde{\text{var}}_{k,\ell} = \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left( V_{i,h+1} - \langle \hat{\theta}_{k,\ell}, \phi_{V_{i,h+1}} \rangle \right)^2 \) \( |\Psi_{k,\ell}| \). We can adopt the method discussed in Subsection 2.4.2 to compute \( \tilde{\text{var}}_{k,\ell} \) in an efficient way. We call the construction of the confidence set along with its radius \( \tilde{\beta}_{k,\ell} \) 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 \( \tilde{\beta}_{k,\ell} \) as in (3.4), \( \alpha = 1/(KH)^{3/2} \) and \( \lambda = 1/B^2 \) in Algorithm 2. Then with probability at least \( 1 - (4|\log_2 2HK| + 9)\delta \), the regret of Algorithm 2 is bounded by:

\[
\text{Regret}(K) = \tilde{O}\left(d \sqrt{\text{Var}_{k}^* + d^2}\right).
\]
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: \(\text{Regret}(K) = \tilde{O}(d\sqrt{K} + d^2)\).

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 \(\tilde{O}(d^2)\), 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


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 $\epsilon$-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 $O(|S|^5|A|^4\epsilon^{-2}\text{polylog}(H))$ that only has a polylogarithmic dependence on $H$. Zhang et al. (2021a) then proposed a near-optimal algorithm with a regret of $O((|S||A|K + |S|^2|A|)\text{polylog}(H))$ 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 $\sigma_k$-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 $\sigma_k^2$ instead of $\sigma_k$-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 $\sigma_k$ 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:

$$d_0 = 0, d_k = \sum_{i=1}^{k} x_i \eta_i, q_0 = 0, q_k = \|d_k\|_{Z_{k-1}}, I_k = \mathbf{1}\{\forall 0 \leq s \leq k, q_s \leq \beta_s\}$$
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:

$$
q_k^2 = (d_{k-1} + x_k \eta_k)^\top Z_k^{-1}(d_{k-1} + x_k \eta_k)
= d_{k-1}^\top Z_k^{-1}d_{k-1} + 2\eta_k x_k^\top Z_k^{-1}d_{k-1} + \eta_k^2 x_k^\top Z_k^{-1}x_k
\leq q_{k-1}^2 + I_{1,k} + I_{2,k},
$$

where the inequality holds since $Z_k = Z_{k-1} + x_k x_k^\top \succeq Z_{k-1}$. For the term $I_{1,k}$, by the matrix inversion lemma, we have the following equation:

$$
I_{1,k} = 2\eta_k \left( x_k^\top Z_{k-1}^{-1}d_{k-1} - \frac{x_k Z_{k-1}^{-1} x_k^\top Z_{k-1}^{-1}d_{k-1}}{1 + \|x_k\|^2_{Z_{k-1}^{-1}}} \right)
= 2\eta_k \left( x_k^\top Z_{k-1}^{-1}d_{k-1} - \frac{\|x_k\|^2_{Z_{k-1}^{-1}} x_k^\top Z_{k-1}^{-1}d_{k-1}}{1 + \|x_k\|^2_{Z_{k-1}^{-1}}} \right)
= 2\eta_k \frac{x_k^\top Z_{k-1}^{-1}d_{k-1}}{1 + \|x_k\|^2_{Z_{k-1}^{-1}}}.
$$

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

$$
\sum_{i=1}^{k} I_{1,i} \cdot \mathcal{I}_{i-1} = 2 \sum_{i=1}^{k} \eta_i \frac{x_i^\top Z_{i-1}^{-1}d_{i-1}}{1 + \|x_i\|^2_{Z_{i-1}^{-1}}} \mathcal{I}_{i-1}.
$$

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

$$
\left| \eta_i \cdot \frac{x_i^\top Z_{i-1}^{-1}d_{i-1}}{1 + \|x_i\|^2_{Z_{i-1}^{-1}}} \mathcal{I}_{i-1} \right| \leq R \left| \frac{x_i Z_{i-1}^{-1} d_{i-1} \mathcal{I}_{i-1}}{1 + \|x_i\|^2_{Z_{i-1}^{-1}}} \right| \mathcal{I}_{i-1} \leq R \left| \frac{\|x_i\| Z_{i-1}^{-1} \mathcal{I}_{i-1}}{1 + \|x_i\|^2_{Z_{i-1}^{-1}}} \right| \leq R \beta_i \rho,
$$

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 $\rho \geq \|x_i\|_{Z_{i-1}^{-1}}$. In addition, for each round $i \in [k]$, we have

$$
\mathbb{E} \left[ \eta_i \cdot \frac{x_i^\top Z_{i-1}^{-1}d_{i-1}}{1 + \|x_i\|^2_{Z_{i-1}^{-1}}} \mathcal{I}_{i-1} \bigg| \mathcal{G}_i \right] = 0,
$$

and the summation of variance is upper bounded by

$$
\sum_{i=1}^{k} \mathbb{E} \left[ \left( \eta_i \cdot \frac{x_i^\top Z_{i-1}^{-1}d_{i-1}}{1 + \|x_i\|^2_{Z_{i-1}^{-1}}} \mathcal{I}_{i-1} \right)^2 \bigg| \mathcal{G}_i \right] = \sum_{i=1}^{k} \left( \frac{x_i^\top Z_{i-1}^{-1}d_{i-1}}{1 + \|x_i\|^2_{Z_{i-1}^{-1}}} \mathcal{I}_{i-1} \right)^2 \mathbb{E} [\eta_i^2 | \mathcal{G}_i].
$$
According to Freedman’s Inequality, for any $i$, where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to Freedman’s inequality in Lemma E.4. Notice that, for each round $i$, using Freedman’s inequality in Lemma E.4. For simplicity, let $E$ be the events that (B.3) holds. Then we bound the summation of $I_{2,k}$ over $k$ through the following calculation:

$$
\sum_{i=1}^{k} I_{2,i} \leq \sum_{i=1}^{k} \eta_i^2 \rho^2 = \rho^2 v_k + \rho^2 \sum_{i=1}^{k} [\eta_i^2 - \mathbb{E}[\eta_i^2 | G_i]],
$$

where the first inequality holds due to $\rho \geq \|x_i\|_{z_{i-1}}$ and the last inequality holds due to $\sum_{i=1}^{k} \mathbb{E}[\eta_i^2 | G_i] \leq v_k$.

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

$$
\sum_{i=1}^{k} I_{1,i} \cdot \mathcal{I}_{i-1} \leq 2\sqrt{2\beta_k \rho^2 v_k \log(4k^2/\delta)} + 4/3 \cdot R \beta_k \rho \log(4k^2/\delta)
$$

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 $k \geq 1$, we have

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

For simplicity, let $E_{t_1}$ be the events that (B.3) holds. Then we bound the summation of $I_{2,k}$ over $k$ through the following calculation:

$$
\sum_{i=1}^{k} I_{2,i} \leq \sum_{i=1}^{k} \eta_i^2 \rho^2 = \rho^2 v_k + \rho^2 \sum_{i=1}^{k} [\eta_i^2 - \mathbb{E}[\eta_i^2 | G_i]],
$$

where the first inequality holds due to $\rho \geq \|x_i\|_{z_{i-1}}$. 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

$$
|\mathbb{E}[\eta_i^2 | G_i] - \eta_i^2| \leq R^2, \quad \mathbb{E}ig[(\mathbb{E}[\eta_i^2 | G_i] - \eta_i^2)^2 | G_i\big] = 0,
$$

$$
\mathbb{E}ig[(\mathbb{E}[\eta_i^2 | G_i] - \eta_i^2)^2 | G_i\big] = (\mathbb{E}[\eta_i^2 | G_i])^2 - 2\mathbb{E}[\eta_i^2 | G_i] \cdot \mathbb{E}[\eta_i^2 | G_i] + \mathbb{E}[\eta_i^2 | G_i] \leq R^2 \mathbb{E}[\eta_i^2 | G_i],
$$

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

$$
\sum_{i=1}^{k} (\eta_i^2 - \mathbb{E}[\eta_i^2 | G_i]) \leq \sqrt{2R^2 \log(4k^2/\delta)} v_k + 2/3 \cdot R^2 \log(4k^2/\delta).
$$

(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

$$
\sum_{i=1}^{k} I_{2,i} \leq \rho^2 v_k + \rho^2 \sum_{i=1}^{k} [\eta_i^2 - \mathbb{E}[\eta_i^2 | G_i]] \\
\leq \rho^2 v_k + \rho^2 R \sqrt{2v_k \log(4k^2/\delta)} + \frac{2}{3} \cdot \rho^2 \cdot R^2 \log(4k^2/\delta) \\
\leq 3 \left( \rho^2 v_k \log(4k^2/\delta) \right) + 2R^2 \rho^2 \log(4k^2/\delta)^2 \\
\leq \frac{1}{4} \beta_k^2.
$$

(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 $I_{i-1} = 1$ holds for all $i \in [k]$, then according to (B.1), we have

$$
q_{k+1}^2 \leq \sum_{i=1}^{k+1} I_{1,i} + \sum_{i=1}^{k+1} I_{2,i} \\
= \sum_{i=1}^{k+1} I_{1,i} \cdot I_{i-1} + \sum_{i=1}^{k+1} I_{2,i} \\
\leq \beta_{k+1}^2,
$$

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

$$
\left\| \sum_{i=1}^{k} x_i \eta_i \right\|_{Z_k^{-1}} \leq \beta_k.
$$

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

$$
\| \mu_k - \mu^* \|_{Z_k} = \| Z_k^{-1} b_k - Z_k^{-1} \sum_{i=1}^{k} x_i \eta_i \|_{Z_k} \\
= \left\| Z_k^{-1} b_k - Z_k^{-1} \sum_{i=1}^{k} x_i \mu^* - \lambda Z_k^{-1} \mu^* \right\|_{Z_k} \\
= \left\| Z_k^{-1} \sum_{i=1}^{k} (y_i - x_i^\top \mu^*) - \lambda Z_k^{-1} \mu^* \right\|_{Z_k} \\
\leq \left\| \sum_{i=1}^{k} x_i \eta_i \right\|_{Z_k^{-1}} + \sqrt{\lambda} \| \mu^* \|_2 \\
\leq \beta_k + \sqrt{\lambda} \| \mu^* \|_2,
$$

21
where the first equality follows from the definition of \( \mu_k \), the second equality holds due to the definition of \( Z_k \) and the first inequality holds by triangle inequality with the fact that \( Z_k \geq \lambda I \). 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 \( \| \theta^* \|_2 \leq 1 \). In Algorithm 1, with probability at least \( 1 - \delta \), the following statement holds for all round \( k \geq 1 \) and layer \( \ell \in [L] \):

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

For simplicity, we denote \( E_{\text{conf}} \) 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 \( \| w_k a_k \|_{\Sigma_{k,\ell}}^{-1} = 2^{-\ell} \).

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

\[
\| w_k a_k \|_{\Sigma_{k,\ell}}^{-1} = 2^{-\ell}, \quad \mathbb{E}[w_k^2 \sigma_k^2 | F_k] \leq w_k^2 \mathbb{E}[\sigma_k^2 | F_k] \leq w_k^2 \sigma_k^2, \quad |w_k \epsilon_k| \leq |\epsilon_k| \leq R,
\]

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

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

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 \( E_{\text{conf}} \) defined in Lemma C.1 occurs. If \( \{ \tilde{\beta}_{k,\ell} \}_{k \geq 1, \ell \in [L]} \) satisfies

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

then for all \( k \geq 1 \) and \( \ell \in [L] \) such that \( A_{k,\ell} \) exists, we have \( a_k^* \in A_{k,\ell} \).

Proof Fix an arbitrary round \( k \). If layer \( \ell = 1 \), then \( a_k^* \in D_k = A_{k,1} \) trivially holds. Then for layer \( \ell > 1 \), we prove lemma C.2 by induction. Assume that \( a_k^* \in A_{k,\ell_1} \) holds for some \( \ell_1 \in \mathbb{Z}^+ \) and \( A_{k,\ell_1+1} \) exists.

By Lemma C.1, for all \( a \in A_{k,\ell_1} \), we have

\[
\| a, \hat{\theta}_{k,\ell_1} - \langle a, \theta^* \rangle \|_{\Sigma_{k,\ell_1}} \| \hat{\theta}_{k,\ell_1} - \theta^* \|_{\Sigma_{k,\ell_1}} \leq \tilde{\beta}_{k,\ell} \| a \|_{\Sigma_{k,\ell_1}}^{-1}, \quad \text{(C.1)}
\]
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, $A_{k,\ell+1}$ exists only if $\|a\|_{\Sigma_{k,\ell}} \leq 2^{-\ell_1}$ holds for all $a \in A_{k,\ell_1}$. Therefore, the sub-optimality gap in (C.1) can be further bounded as follows:

$$\left| \langle a, \hat{\theta}_{k,\ell_1} \rangle - \langle a, \theta^* \rangle \right| \leq \hat{\beta}_{k,\ell} \|a\|_{\Sigma_{k,\ell_1}} \leq 2^{-\ell_1} \cdot \hat{\beta}_{k,\ell_1}. \quad (C.2)$$

For short, let $a_{\text{max}} = \arg\max_{a' \in A_{k,\ell_1}} \langle a', \hat{\theta}_{k,\ell_1} \rangle$. Then for the optimal action $a^*_k \in A_{k,\ell_1}$, we have

$$\langle a^*_k, \hat{\theta}_{k,\ell_1} \rangle - \max_{a' \in A_{k,\ell_1}} \langle a', \hat{\theta}_{k,\ell_1} \rangle = \langle a^*_k, \hat{\theta}_{k,\ell_1} \rangle - \langle a^*_k, \hat{\theta}_{k,\ell_1} \rangle$$

$$\geq \langle a^*_k, \theta^* \rangle - \langle a_{\text{max}}, \theta^* \rangle - \left| \langle a^*_k, \hat{\theta}_{k,\ell_1} \rangle - \langle a^*_k, \theta^* \rangle \right| - \left| \langle a_{\text{max}}, \hat{\theta}_{k,\ell_1} \rangle - \langle a_{\text{max}}, \theta^* \rangle \right|$$

$$\geq -2^{-\ell_1+1} \cdot \hat{\beta}_{k,\ell_1},$$

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

**Lemma C.3** Suppose for all $k \geq 1$ and all $a \in D_k$, we have $\|a\|_2 \leq A, \|\theta^*\|_2 \leq 1$. 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 $\ell \in [L] \setminus \{1\}$, the regret incurred by the index set $\Psi_{K+1,\ell}$ is bounded as follows:

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

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

$$\langle a^*_\tau, \hat{\theta}_{\tau,\ell-1} \rangle - \langle a_\tau, \hat{\theta}_{\tau,\ell-1} \rangle \leq 2^{-\ell+2} \hat{\beta}_{\tau,\ell-1}. \quad (C.3)$$

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

$$\|a_\tau\|_{\Sigma_{\tau,\ell-1}} \leq 2^{-\ell+1}, \quad \|a^*_\tau\|_{\Sigma_{\tau,\ell-1}} \leq 2^{-\ell+1}. \quad (C.4)$$

We further compute

$$\langle a^*_\tau, \theta^* \rangle - \langle a_\tau, \theta^* \rangle \leq \langle a^*_\tau, \hat{\theta}_{\tau,\ell-1} \rangle + \left| \langle a^*_\tau, \hat{\theta}_{\tau,\ell-1} - \theta^* \rangle - \langle a_\tau, \hat{\theta}_{\tau,\ell-1} \rangle + \langle a_\tau, \hat{\theta}_{\tau,\ell-1} - \theta^* \rangle \right|$$

$$\leq \langle a^*_\tau, \hat{\theta}_{\tau,\ell-1} \rangle - \langle a_\tau, \hat{\theta}_{\tau,\ell-1} \rangle$$

$$+ \|a^*_\tau\|_{\Sigma_{\tau,\ell-1}} \|\hat{\theta}_{\tau,\ell-1} - \theta^*\|_{\Sigma_{\tau,\ell-1}} + \|a_\tau\|_{\Sigma_{\tau,\ell-1}} \|\hat{\theta}_{\tau,\ell-1} - \theta^*\|_{\Sigma_{\tau,\ell-1}}$$

$$\leq 2^{-\ell+2} \cdot \hat{\beta}_{\tau,\ell-1} + 2^{-\ell+1} \cdot \hat{\beta}_{\tau,\ell-1} + 2^{-\ell+1} \cdot \hat{\beta}_{\tau,\ell-1}$$

$$= 8 \cdot 2^{-\ell} \cdot \hat{\beta}_{\tau,\ell-1}. \quad (C.5)$$
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

$$\sum_{\tau \in \Psi_{K+1,\ell}} \left( \langle a_\tau^*, \theta^* \rangle - \langle a_\tau, \theta^* \rangle \right) \leq 8 \cdot 2^{-\ell} \cdot \hat{\beta}_{K,\ell-1} |\Psi_{K+1,\ell}|$$

$$\leq 8 \cdot 2^\ell \cdot \hat{\beta}_{K,\ell-1} \cdot \sum_{k \in \Psi_{K+1,\ell}} \|w_k \cdot a_k\|^2_{S_k^2}$$

$$\leq 8 \cdot 2^\ell \cdot \hat{\beta}_{K,\ell-1} \cdot 2d \log \left( 1 + 2^{2\ell} K \cdot A^2/d \right),$$

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 $\|w_k \cdot a_k\|^2_{S_k^2} = 2^{-\ell}$, 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 $k \geq 1$, $\ell \in [L]$, the following two inequalities hold simultaneously:

$$\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 \leq 2 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \epsilon_i^2 + \frac{14}{3} R^2 \log(4k^2L/\delta),$$

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

For simplicity, we denote $E_{\text{var}}$ 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

$$\text{for } \forall i \geq 1, \mathbb{E} \left[ \epsilon_i^2 - \sigma_i^2 | a_{1:i}, r_{1:i-1} \right] = 0,$$

$$\sum_{i \in \Psi_{k+1,\ell}} \mathbb{E} \left[ w_i^2 (\epsilon_i^2 - \sigma_i^2) | a_{1:i}, r_{1:i-1} \right] \leq \sum_{i \in \Psi_{k+1,\ell}} \mathbb{E} \left[ w_i^2 \epsilon_i^4 | a_{1:i}, r_{1:i-1} \right] \leq R^2 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2,$$

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

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

$$\leq \frac{1}{2} \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 + \frac{7}{3} R^2 \log(4k^2L/\delta),$$

where the last inequality holds due to Young’s inequality. Rearranging the above inequality, we conclude that $\mathbb{P}(E_{\text{var}}) \geq 1 - 2\delta$ by applying union bound over all $\ell \in [L]$. Thus, we complete the proof of Lemma C.4.
Lemma C.5  Suppose that $\|\theta^*\|_2 \leq 1$. 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 $k \geq 1$, $\ell \in [L]$ such that $2^\ell \geq 64 \sqrt{\log (4(k+1)^2L/\delta)}$, we have the following inequalities:

\[
\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 \leq 8 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right)^2 + 6R^2 \log(4(k+1)^2L/\delta) + 2^{-2\ell+4},
\]

\[
\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right)^2 \leq \frac{3}{2} \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 + \frac{7}{3} R^2 \log(4k^2L/\delta) + 2^{-2\ell}.
\]

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

\[
\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 \leq 2 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \epsilon_i^2 + \frac{14}{3} R^2 \log(4k^2L/\delta)
\]

\[
\leq 4 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right)^2 + 4 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left[ \epsilon_i - \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right) \right]^2
\]

\[
+ \frac{14}{3} R^2 \log(4k^2L/\delta),
\]

where the last inequality holds due to $(a + b)^2 \leq 2a^2 + 2b^2$. In addition, the gap between $\epsilon_i$ and $r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle$ can be upper bounded by

\[
\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left[ \epsilon_i - \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right) \right]^2
\]

\[
= \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( \langle \hat{\theta}_{k+1,\ell} - \theta^*, a_i \rangle \right)^2
\]

\[
= \sum_{i \in \Psi_{k+1,\ell}} \left( \hat{\theta}_{k+1,\ell} - \theta^* \right)^\top w_i a_i \left( \hat{\theta}_{k+1,\ell} - \theta^* \right)
\]

\[
\leq \left( \hat{\theta}_{k+1,\ell} - \theta^* \right)^\top \Sigma_{k+1,\ell} \left( \hat{\theta}_{k+1,\ell} - \theta^* \right)
\]

\[
\leq \left( 16 \cdot 2^{-\ell} \sqrt{\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 \log(4(k+1)^2L/\delta) + 6 \cdot 2^{-\ell} R \log(4(k+1)^2L/\delta) + 2^{-2\ell+1}} \right)^2,
\]

where the first inequality holds due to $\Sigma_{k+1,\ell} \succeq w_i^2 a_i a_i^\top$ and the last inequality holds due to Lemma C.1. From (C.7), when $2^\ell \geq 64 \sqrt{\log (4(k+1)^2L/\delta)}$, we have

\[
\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left[ \epsilon_i - \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right) \right]^2
\]

\[
\leq \frac{1}{8} \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 + 2 \left( 6 \cdot 2^{-\ell} R \log(4(k+1)^2L/\delta) + 2^{-2\ell+1} \right)^2.
\]
where the inequality holds due to (C.7) with the fact that \((a + b)^2 \leq 2a^2 + 2b^2\). Substituting (C.8) into (C.6), we have

\[
\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 \leq 4 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right)^2 + 4 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left[ \epsilon_i - \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right) \right]^2 + \frac{14}{3} R^2 \log(4k^2 L/\delta) \\
\leq 4 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right)^2 + \frac{1}{2} \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \sigma_i^2 \\
+ 2 \left( 6 \cdot 2^{-\ell} R \log(4(k + 1)^2 L/\delta) + 2^{-\ell + 1} \right)^2 + \frac{14}{3} R^2 \log(4k^2 L/\delta) \\
\leq 8 \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right)^2 + 26 R^2 \log(4(k + 1)^2 L/\delta) + 2^{-2\ell + 4},
\]

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

For the second part, note that \(\theta_{k+1,\ell}\) is the minimizer of the following weighted ridge regression

\[
\theta_{k+1,\ell} \leftarrow \arg \min_{\theta \in \mathbb{R}^{d}} \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \theta, a_i \rangle \right)^2 + 2^{-2\ell} \|\theta\|^2.
\]

Thus, we have

\[
\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right)^2 \leq \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \theta^*, a_i \rangle \right)^2 + 2^{-2\ell} \|\theta^*\|^2 \leq \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \epsilon_i^2 + 2^{-2\ell},
\]

where the second inequality holds due to \(\|\theta^*\|^2 \leq 1\). Combining the result in Lemma C.4, we can further conclude that

\[
\sum_{i \in \Psi_{k+1,\ell}} w_i^2 \left( r_i - \langle \hat{\theta}_{k+1,\ell}, a_i \rangle \right)^2 \leq \frac{3}{2} \sum_{i \in \Psi_{k+1,\ell}} w_i^2 \epsilon_i^2 + \frac{7}{3} R^2 \log(4k^2 L/\delta) + 2^{-2\ell}.
\]

Thus, we complete the proof of Lemma C.5. \(\blacksquare\)

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

\[
\hat{\beta}_{K,\ell-1} \geq 16 \cdot 2^{-(\ell-1)} \sqrt{\sum_{i \in \Psi_{K,\ell-1}} w_i^2 \sigma_i^2 \log(4K^2 L/\delta) + 6 \cdot 2^{-\ell} R \log(4K^2 L/\delta) + 2^{-\ell}},
\]
which further implies
\[
\sum_{\tau \in \Psi_{K+1,\ell}} (\langle a^*_\tau, \theta^* \rangle - \langle a_{\tau}, \theta^* \rangle) \leq \tilde{O} \left( d \cdot 2^\ell \cdot \beta_{K,\ell-1} \right)
\]
\[
\leq \tilde{O} \left( d \cdot \sqrt{\sum_{k=1}^{K} w_k^2 \left( r_k - \langle \hat{\theta}_{K+1,\ell}, a_k \rangle \right)^2 + R^2 + 1 + R} \right)
\]
\[
\leq \tilde{O} \left( d \cdot \sum_{k=1}^{K} \sigma_k^2 + dR + d \right),
\]  
(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 \( \ell_k \) as the value of layer \( \ell \) such that the while loop in Algorithm 1 stops. Therefore, we have
\[
\sum_{k \in [K] \backslash \left( \bigcup_{\ell \in [L]} \Psi_{K+1,\ell} \right)} (\langle a^*_k, \theta^* \rangle - \langle a_k, \theta^* \rangle) \leq \sum_{k \in \Psi_{K+1,L+1}} (\langle a_k, \hat{\theta}_{k,\ell_k} \rangle + \beta_{k,\ell_k} \cdot \alpha - \langle a_k, \theta^* \rangle)
\]
\[
\leq \sum_{k \in \Psi_{K+1,L+1}} \left( \beta_{k,\ell_k} \cdot \alpha + \alpha \cdot \| \theta^* - \hat{\theta}_{k,\ell_k} \| \Sigma_{k,\ell_k} \right)
\]
\[
\leq \sum_{k \in \Psi_{K+1,L+1}} 2\alpha \cdot \beta_{k,\ell_k}
\]
\[
\leq K \cdot \tilde{O} \left( 1/K \right) = \tilde{O}(1),
\]  
(C.10)

where the first inequality holds due to the selection rule of action \( a_k \) (Line 8 in Algorithm 1) with Lemma C.1, Lemma C.5 and the fact that \( a^*_k \in A_{k,\ell_k} \) (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 \( \alpha \) and \( \beta_{k,\ell_k} \).

Finally, for layer \( \ell \in [\ell^*] \) and round \( \tau \in \Psi_{K+1,\ell} \), we have
\[
\sum_{\tau \in \Psi_{K+1,\ell}} (\langle a^*_\tau, \theta^* \rangle - \langle a_{\tau}, \theta^* \rangle) \leq 2 \| w_{\tau} a_{\tau} \|_{\Sigma_{\tau,\ell}}^2 \leq \tilde{O}(d),
\]  
(C.11)

where the first inequality holds since the reward is in the range \([-1, 1]\), the equation follows from the fact that \( \| w_{\tau} a_{\tau} \|_{\Sigma_{\tau,\ell}} = 2^{-\ell} \) holds for all \( \tau \in \Psi_{K+1,\ell} \) and the last inequality follows from Lemma E.2 with the fact that \( 2^{\ell^*} \leq 128 \sqrt{\log(4(K+1)^2L/\delta)} \) is bounded by a logarithmic term. Putting (C.9), (C.10), (C.11) together, we have
\[
\text{Regret}(K) \leq \tilde{O} \left( d \cdot \sqrt{\sum_{k=1}^{K} \sigma_k^2 + dR + d} \right).
\]

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 \(\sigma\)-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

\[
\begin{align*}
&\{s_1, a_1, \ldots, s_h, a_h, \ldots, s_H, a_H, \\
&s_1^2, a_1^2, \ldots, s_h^2, a_h^2, \ldots, s_H^2, a_H^2, \\
&\cdots \\
&s_1^k, a_1^k, \ldots, s_h^k, a_h^k. 
\end{align*}
\]

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

\[
I^k_h = 1 \{ \forall \ell \in [L], \det \left( \hat{\Sigma}_{k,h,\ell} \right) / \det \left( \hat{\Sigma}_{k,1,\ell} \right) \leq 4 \}.
\]

(D.1)

For each \(1 \leq h_1 \leq h_2 \leq H\), since \(\hat{\Sigma}_{k,h_2,\ell} \succeq \hat{\Sigma}_{k,h_1,\ell}\), the indicator function is monotonic (e.g., \(I^k_{h_1} \leq I^k_{h_2}\)). 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 \(I^k_h\) is defined as in (D.1), then for each \(k \in [K]\), we have

\[
\sum_{i=1}^{k} (1 - I^k_h) \leq \frac{dL}{2} \log \frac{\lambda + kH/d}{\lambda} + dL^2.
\]

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

\[
\det \left( \hat{\Sigma}_{i+1,1,\ell} \right) / \det \left( \hat{\Sigma}_{i,1,\ell} \right) > 4.
\]

According to the update rule of \(\Sigma_{k,1,\ell}, \Sigma_{k+1,1,\ell} \succeq \Sigma_{k,1,\ell}\) holds for all episode \(k \in [K]\). Therefore, we have

\[
\det(\hat{\Sigma}_{k+1,1,\ell})/\det(\hat{\Sigma}_{1,1,\ell}) = \prod_{i=1}^{k} \det \left( \hat{\Sigma}_{i+1,1,\ell} \right) / \det \left( \hat{\Sigma}_{i,1,\ell} \right) \geq 4^{|\mathcal{D}_\ell|},
\]

(D.2)

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

\[
\det(\hat{\Sigma}_{k+1,1,\ell}) \leq \left( \text{tr}(\hat{\Sigma}_{k+1,1,\ell}) / d \right)^d \leq (2^{-2\ell} \lambda + kH/d)^d,
\]

\[
\det(\hat{\Sigma}_{1,1,\ell}) = (2^{-2\ell} \cdot \lambda)^d,
\]

where the first inequality holds since \(\hat{\Sigma}_{k+1,1,\ell} \succeq 0\), the last inequality holds due to \(w_{k,i} \leq 1\) and \(\|\phi_{V_{i+1}}(s_h^i, a_h^i)\|_2 \leq 1\). Combining these results, it holds that

\[
|\mathcal{D}_\ell| \leq \log_4 \left( \frac{(\lambda + 2^{2\ell} kH/d)^d}{\lambda^d} \right) \leq \frac{d}{2} \log_2 \frac{\lambda + 2^{2\ell} kH/d}{\lambda} \leq \frac{d}{2} \log_2 \frac{\lambda + kH/d}{\lambda} + d \cdot \ell.
\]
Finally, according to the definition of $D_\ell$ and indicator function $I_i$, we have

$$\sum_{i=1}^k (1 - I_i) \leq \sum_{\ell \in [L]} |D_\ell| \leq \frac{dL}{2} \log_2 \frac{\lambda + kH/d}{\lambda} + dL^2.$$ 

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 $|\Psi_{K+1,\ell}| \leq 2d \log (1 + KH/(2^{-2\ell}d\lambda))$.

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

$$\sum_{(k,h) \in \Psi_{K+1,\ell}} \|w_{k,h} \phi V_{k,h+1}(s^{k}_{h},a^{k}_{h})\|_{\Sigma_{k,h,\ell}}^2 = |\Psi_{K+1,\ell}| \cdot 2^{-2\ell}.$$ 

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

$$\sum_{(k,h) \in \Psi_{K+1,\ell}} \|w_{k,h} \phi V_{k,h+1}(s^{k}_{h},a^{k}_{h})\|_{\Sigma_{k,h,\ell}}^2 \leq 2d \log \frac{2^{-2\ell}d\lambda + KH}{2^{-2\ell}d\lambda}.$$ 

Combining these results, we further conclude that $|\Psi_{K+1,\ell}| \leq 2 \cdot d \log (1 + KH/(2^{-2\ell}d\lambda))$. 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:

$$\epsilon_{k,h} = V_{k,h+1}(s^{k}_{h}) - \left< \theta^*, \phi V_{k,h+1}(s^{k}_{h},a^{k}_{h}) \right>,$$

$$\sigma_{k,h} = \sqrt{\nabla V_{k,h+1}(s^{k}_{h},a^{k}_{h})}.$$ 

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

$$\mathcal{E}_{c} = \left\{ \forall k \geq 1, \ell \in [L], \|\hat{\theta}_{k,\ell} - \theta^*\|_{\Sigma_{k,\ell}} \leq 16 \cdot 2^{-\ell} \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \sigma_{i,h}^2 \log(4k^2H^2L/\delta) + 6 \cdot 2^{-\ell} \log(4k^2H^2L/\delta) + 2^{-\ell} \sqrt{\lambda} \cdot B \right\},$$

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

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

29
Proof From the definition of $\ell_{k,h}$ and $w_{k,h}$ in Algorithm 2, we can deduce that for all $k \in [K]$, $h \in [H]$, $\|w_{k,h}\phi_{\Psi_{k,h}}(s_{k,h}^i, a_{k,h}^i)\|_{\Sigma^{-1}_{k,h}} \leq 2^{-\ell_{k,h}}$. According to Theorem 2.1, for layer $\ell \in [L]$, we have with probability at least $1 - \delta/L$, for all $k \in [K]$:

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

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 $\mathcal{V}$ and Definition 3.2, we have

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

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 $\ell \in [L]$,

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

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 $\mathcal{E}_c$ 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:

$$
\sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \sigma_{i,h}^2 \leq 8 \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left(V_{i,h+1}(s_{i,h+1}^i) - \langle \hat{\theta}_{k,\ell}, \phi_{\Psi_{i,h+1}}(s_{i,h+1}^i, a_{i,h+1}) \rangle \right)^2
$$

$$
+ 8 \log(4k^2H^2L/\delta) + 2^{-2\ell+5} \cdot \lambda B^2,
$$

$$
\sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left(V_{i,h+1}(s_{i,h+1}^i) - \langle \hat{\theta}_{k,\ell}, \phi_{\Psi_{i,h+1}}(s_{i,h+1}^i, a_{i,h+1}) \rangle \right)^2 \leq \frac{3}{2} \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \sigma_{i,h}^2 + 2^{-2\ell} \lambda B^2
$$

$$
+ \frac{7}{3} \log \left(4k^2H^2L/\delta\right).
$$

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

$$
\sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \sigma_{i,h}^2 \leq 2 \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \epsilon_{i,h}^2 + \frac{14}{3} \log \left(4k^2H^2L/\delta\right)
$$

$$
\leq 4 \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left(V_{i,h+1}(s_{i,h+1}^i) - \langle \hat{\theta}_{k,\ell}, \phi_{\Psi_{i,h+1}}(s_{i,h+1}^i, a_{i,h+1}) \rangle \right)^2
$$

$$
+ 4 \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left[\epsilon_{i,h} - \left(V_{i,h+1}(s_{i,h+1}^i) - \langle \hat{\theta}_{k,\ell}, \phi_{\Psi_{i,h+1}}(s_{i,h+1}^i, a_{i,h+1}) \rangle \right)\right]^2
$$
where the last inequality holds due to the fact \((a + b)^2 \leq 2a^2 + 2b^2\). Then we consider the second term and we have

\[
\sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \left[ \epsilon_{i,h} - \left(V_{i,h+1}(s_{h+1}^i) - \langle \tilde{\theta}_{k,l}, \phi_{V_{i,h+1}}(s_h^i, a_h^i) \rangle \right) \right]^2
\]

\[
= \sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \left( \theta^* - \tilde{\theta}_{k,l} \right) \phi_{V_{i,h+1}}(s_h^i, a_h^i) \phi_{V_{i,h+1}}(s_h^i, a_h^i) \top \left( \theta^* - \tilde{\theta}_{k,l} \right)
\]

\[
\leq \left\| \theta^* - \tilde{\theta}_{k,l} \right\|_2^2 \bar{\Sigma}_{k,l}
\]

\[
\leq \left( 16 \cdot 2^{-\ell} \sqrt{\sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \sigma_{i,h}^2 \log(4k^2H^2L/\delta)} + 6 \cdot 2^{-\ell} \log(4k^2H^2L/\delta) + 2^{-\ell} \sqrt{X \cdot B} \right)^2,
\]

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

\[
\sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \left[ \epsilon_{i,h} - \left(V_{i,h+1}(s_{h+1}^i) - \langle \tilde{\theta}_{k,l}, \phi_{V_{i,h+1}}(s_h^i, a_h^i) \rangle \right) \right]^2
\]

\[
\leq \frac{1}{8} \sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \sigma_{i,h}^2 + 2 \left( 6 \cdot 2^{-\ell} \log(4k^2H^2L/\delta) + 2^{-\ell} \sqrt{X \cdot B} \right)^2
\]

\[
\leq \frac{1}{8} \sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \sigma_{i,h}^2 + \log(4k^2H^2L/\delta) + 2^{-2\ell + 2} \cdot \lambda B^2,
\]

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

\[
\sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \sigma_{i,h}^2 \leq 4 \sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \left( V_{i,h+1}(s_{h+1}^i) - \langle \tilde{\theta}_{k,l}, \phi_{V_{i,h+1}}(s_h^i, a_h^i) \rangle \right)^2
\]

\[
+ \frac{1}{2} \sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \sigma_{i,h}^2 + 4 \log(4k^2H^2L/\delta) + 2^{-2\ell + 4} \cdot \lambda B^2
\]

\[
\leq 8 \sum_{(i,h)\in \Psi_{k,l}} w_{i,h}^2 \left( V_{i,h+1}(s_{h+1}^i) - \langle \tilde{\theta}_{k,l}, \phi_{V_{i,h+1}}(s_h^i, a_h^i) \rangle \right)^2
\]

\[
+ 8 \log(4k^2H^2L/\delta) + 2^{-2\ell + 5} \cdot \lambda B^2,
\]

where the last inequality holds due to the fact that \(x \leq x/2 + y\) implies \(x \leq 2y\). Thus, we complete the proof of the first inequality in this lemma.
Note that $\hat{\theta}_{k,\ell}$ is the minimizer of
\[
\hat{\theta}_{k,\ell} \leftarrow \arg \min_{\theta \in \mathbb{R}^d} \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left( V_{i,h+1}(s_{h+1}^i) - \langle \theta, \phi_{V_{i,h+1}}(s_{h+1}^i, a_{h+1}^i) \rangle \right)^2 + 2^{-2\ell} \lambda \|\theta\|^2,
\]
and we have
\[
\sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left( V_{i,h+1}(s_{h+1}^i) - \langle \hat{\theta}_{k,\ell}, \phi_{V_{i,h+1}}(s_{h+1}^i, a_{h+1}^i) \rangle \right)^2 \\
\leq \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left( V_{i,h+1}(s_{h+1}^i) - \langle \theta^*, \phi_{V_{i,h+1}}(s_{h+1}^i, a_{h+1}^i) \rangle \right)^2 + 2^{-2\ell} \lambda \|\theta^*\|^2 \\
\leq \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \left( V_{i,h+1}(s_{h+1}^i) - \langle \theta^*, \phi_{V_{i,h+1}}(s_{h+1}^i, a_{h+1}^i) \rangle \right)^2 + 2^{-2\ell} \lambda B^2 \\
\leq \frac{3}{2} \sum_{(i,h) \in \Psi_{k,\ell}} w_{i,h}^2 \sigma_{i,h}^2 + 2^{-2\ell} \lambda B^2 + \frac{7}{3} \log \left( 4k^2 H^2 L/\delta \right),
\]
where the first inequality holds due to the definition of $\hat{\theta}_{k,\ell}$, the second inequality holds due to $\|\theta^*\| \leq B$ and the last inequality follows from the definition of $\mathcal{E}_{var'}$. Therefore, we complete the proof of Lemma D.5.}

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

**Proof** From the definition of event $\mathcal{E}_{var'}$ and Lemma D.5, we can deduce that
\[
\hat{\beta}_{k,\ell} \geq 16 \cdot 2^{-\ell} \sqrt{\sum_{i=1}^k \sigma_i^2 \log(4k^2 H^2 L/\delta) + 6 \cdot 2^{-\ell} \log(4k^2 H^2 L/\delta) + 2^{-\ell} \sqrt{\lambda} \cdot B}.
\]
Therefore, by the definition of $\mathcal{E}_c$, we have
\[
\forall k \geq 1, \ell \in [L], \quad \|\hat{\theta}_{k,\ell} - \theta^*\|_{\Sigma_{k,1,\ell}} \leq \hat{\beta}_{k,\ell}. \tag{D.9}
\]

According to Algorithm 2, we have
\[
V_{k,h}(s_{h}^k) = \min \left\{ 1, \min_{\ell \in [L]} \left\{ r(s_{h}^k, a_{h}^k) + \langle \hat{\theta}_{k,\ell}, \phi_{V_{k,h+1}}(s_{h}^k, a_{h}^k) \rangle + \hat{\beta}_{k,\ell} \|\phi_{V_{k,h+1}}(s_{h}^k, a_{h}^k)\|_{\Sigma_{k,1,\ell}} \right\} \right\} \\
\geq \min \left\{ 1, \min_{\ell \in [L]} \left\{ \langle \hat{\theta}_{k,\ell}, \phi_{V_{k,h+1}}(s_{h}^k, a_{h}^k) \rangle + \hat{\beta}_{k,\ell} \|\phi_{V_{k,h+1}}(s_{h}^k, a_{h}^k)\|_{\Sigma_{k,1,\ell}} \right\} \right\} \\
\geq \min \left\{ 1, \min_{\ell \in [L]} \left\{ \langle \theta^*, \phi_{V_{k,h+1}}(s_{h}^k, a_{h}^k) \rangle \right\} \right\} \\
= [\mathbb{P}(V_{k,h+1})(s_{h}^k, a_{h}^k)],
\]
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 $V_{k,h+1}(s_{h+1}) \leq 1$ for all $s_{h+1} \in \mathcal{S}$. Thus, we complete the proof of Lemma D.6.
Lemma D.7 Let value function $Q_{k,h}, V_{k,h}$ and confidence radius $\hat{\beta}_{k,\ell}$ be defined in Algorithm 2. Suppose that $\lambda = 1/B^2$ in Algorithm 2. Then, on the event $\mathcal{E}_{\text{var}} \cap \mathcal{E}_c$, for any $(s, a, k, h) \in S \times A \times [K] \times [H]$, we have $Q_{k,h}^*(s, a) \leq Q_{k,h}(s, a)$ and $V_{k,h}^*(s) \leq V_{k,h}(s)$.

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

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

Therefore, by the definition of $\mathcal{E}_c$, we have

$$\forall k \geq 1, \ell \in [L], \quad \|\hat{\theta}_{k,\ell} - \theta^*\|\Sigma_{k,1,\ell}^{-1} \leq \hat{\beta}_{k,\ell}. \quad (D.10)$$

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

$$Q_{h-1}^*(s, a) = r(s, a) + \langle \theta^*, \phi_{V_{k,h}}^*(s, a) \rangle \leq r(s, a) + \langle \theta^*, \phi_{V_{k,h}}(s, a) \rangle + \|\hat{\theta}_{k,\ell} - \theta^*\|\Sigma_{k,1,\ell}^{-1} \|\phi_{V_{k,h}}(s, a)\|\Sigma_{k,1,\ell}^{-1} \leq r(s, a) + \langle \hat{\theta}_{k,\ell}, \phi_{V_{k,h}}(s, a) \rangle + \hat{\beta}_{k,\ell} \|\phi_{V_{k,h}}(s, a)\|\Sigma_{k,1,\ell}^{-1},$$

where the first inequality holds by our assumption that $V_{k,h}^*(s) \leq V_{k,h}(s)$, the second inequality holds due to Cauchy-Schwarz inequality and the last inequality follows from (D.10). By the arbitrariness of layer $\ell$, we have $Q_{h-1}^*(s, a) \leq Q_{k,h-1}(s, a)$ holds for all state-action pair $(s, a)$, which indicates that $V_{h-1}^*(s) \leq V_{k,h-1}(s)$ holds for all $s \in S$. Since $0 = V_{H+1}^*(\cdot) \leq V_{k,H+1}(\cdot)$ 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 $\hat{\beta}_{k,\ell}, V_{k,h}, \phi_{V_{k,h+1}}$ be defined in Algorithm 2 and set $\lambda = 1/B^2, \alpha = 1/(KH)^{3/2}$. Then on the event $\mathcal{E}_{\text{var}} \cap \mathcal{E}_c$, we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} I^k_h \max \left\{ [V_{k,h}(s^k_h) - r(s^k_h, a^k_h) - [P_{V_{k,h+1}}](s^k_h, a^k_h)], 0 \right\} \leq \tilde{O} \left( d \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \sigma_{k,h}^2 + d} \right).$$

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

$$V_{k,h}(s^k_h) = \min \left\{ 1, \min_{\ell \leq \ell^*} \left\{ r(s^k_h, a^k_h) + \langle \hat{\theta}_{k,\ell}, \phi_{V_{k,h+1}}(s^k_h, a^k_h) \rangle + \hat{\beta}_{k,\ell} \|\phi_{V_{k,h+1}}(s^k_h, a^k_h)\|\Sigma_{k,1,\ell}^{-1} \right\} \right\}. \quad (D.11)$$

Therefore, we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} I^k_h \max \left\{ [V_{k,h}(s^k_h) - r(s^k_h, a^k_h) - [P_{V_{k,h+1}}](s^k_h, a^k_h)], 0 \right\} \quad (D.11)$$

33
where the first inequality holds due to the definition of value function $V_{k,h}(s^k_h)$, the second inequality holds due to Cauchy-Schwarz inequality with event $\mathcal{E}_c$ and the last inequality holds since indicator function $I_h \leq 1$. By the definition of indicator function $I_h$ and Lemma E.3, we further have

$$\|\Phi_{V_{k,h+1}}(s^k_h, a^k_h)\|_{\Sigma_{k,h-1}} \leq 2 \|\Phi_{V_{k,h+1}}(s^k_h, a^k_h)\|_{\Sigma_{k,h-1}} \leq 2 \cdot 2^{-\ell_{k,h}+1}, \tag{D.13}$$

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

$$\sum_{k=1}^{K} \sum_{h=1}^{H} I_h \max \left\{ [V_{k,h}(s^k_h) - r(s^k_h, a^k_h) - [\Pi V_{k,h+1}](s^k_h, a^k_h)], 0 \right\}$$

$$\leq \sum_{\ell = \ell^* + 1}^{L+1} |\Psi_{K+1,\ell}| \cdot \tilde{O} \left( 2^{-\ell} \cdot 2^{k \sum_{h=1}^{H} \sigma_{k,h}^2 + 2 \cdot 2^\ell} \right) + 2 \sum_{\ell = 1}^{\ell^*} |\Psi_{K+1,\ell}|$$

$$\leq \tilde{O} \left( d \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \sigma_{k,h}^2} \right),$$

where the first inequality follows from the definition of $\tilde{\beta}_{k,\ell}$ 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. \hfill \blacksquare

### D.4. Quantities in MDP

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

$$\tilde{V}_{k,h}(s) = V_{k,h}(s) - V^*_h(s), \quad \forall s \in \mathcal{S}, (k, h) \in [K] \times [H] \tag{D.14}$$

$$\tilde{V}_{k,h}(s) = V^*_h(s) - V^*_{\pi_k}(s), \quad \forall s \in \mathcal{S}, (k, h) \in [K] \times [H] \tag{D.15}$$

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

$$S_m = \sum_{k=1}^{K} \sum_{h=1}^{H} |\Psi V^*_{k,h+1}(s^k_h, a^k_h)|, \tag{D.16}$$
\[ \tilde{S}_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{V}^{2m}_{k,h+1}](s^k_h, a^k_h), \quad \text{(D.17)} \]

\[ \bar{S}_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{V}^{2m}_{k,h+1}](s^k_h, a^k_h), \quad \text{(D.18)} \]

\[ Q_0 = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{V}^*_h](s^k_h, a^k_h), \quad \text{(D.19)} \]

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

\[ A_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{P}V^{2m}_{k,h+1}](s^k_h, a^k_h) - V^{2m}_{k,h+1}(s^k_h), \quad \text{(D.20)} \]

\[ \bar{A}_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{P}\tilde{V}^{2m}_{k,h+1}](s^k_h, a^k_h) - \tilde{V}^{2m}_{k,h+1}(s^k_h), \quad \text{(D.21)} \]

\[ \bar{A}_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{P}\tilde{V}^{2m}_{k,h+1}](s^k_h, a^k_h) - \tilde{V}^{2m}_{k,h+1}(s^k_h), \quad \text{(D.22)} \]

\[ R_0 = \sum_{k=1}^{K} \sum_{h=1}^{H} I^k_h \max \left\{ \left[ V_{k,h}(s^k_h) - r(s^k_h, a^k_h) - [\mathbb{P}V_{k,h+1}](s^k_h, a^k_h) \right], 0 \right\}. \quad \text{(D.23)} \]

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

\[ G = \sum_{k=1}^{K} (1 - I^k_H), \quad \text{(D.24)} \]

where indicator function \( I^k_h \) 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 \( \tilde{S}_m, A_m, R_0, G \) be defined in (D.17), (D.20), (D.23), (D.24). On the event \( \mathcal{E}_{\text{var}} \cap \mathcal{E}_c \), we have the following inequalities for all \( m \in [M] \):

\[ \tilde{S}_m \leq \tilde{A}_{m+1} + G + 2^{m+1} \cdot (R_0 + G + A_0). \]

**Proof** Based on the definition of \( \tilde{S}_m \), we compute

\[ \tilde{S}_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [\mathbb{V}^{2m}_{k,h+1}](s^k_h, a^k_h) \]
the second inequality follows from the fact that \( a^x - b^x \leq x \max\{a - b, 0\} \) for \( a, b \in [0, 1] \) and \( x \geq 1 \), the third inequality follows from the monotonicity of indicator function \( I_h^k \) and the definition of function \( \tilde{V}_{k, h} \), the fourth holds since \( r(s^k_h, a^k_h) \geq 0 \) and the last inequality holds due to Lemma D.6.

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

\[
\tilde{S}_m \leq \tilde{A}_{m+1} + G + 2^{m+1} \cdot (R_0 + G + A_0).
\]

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

![ ]
Lemma D.10  Let $A_m, Q_0, \hat S_m$ be defined in (D.20), (D.19), (D.17). Then with probability at least $1 - 2\delta$, we have

$$A_0 \leq 2\sqrt{(Q_0 + \hat S_0) \log(1/\delta)} + (2/3) \cdot \log(1/\delta).$$

For simplicity, we denote the corresponding event by $\mathcal{E}_{t_1}$.

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

$$A_0 = \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \left[ \mathbb{E}[V_{k,h+1}(s^k_h, a^k_h)] \right] (s^k_h, a^k_h) - V_{k,h+1}(s^k_{h+1})} \leq \sqrt{2 \sum_{k=1}^{K} \sum_{h=1}^{H} \sigma_{k,h}^2 \log(1/\delta) + \frac{2}{3} \log(1/\delta)},$$

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

$$\sigma_{k,h}^2 = [\mathbb{E}[V_{k,h+1}]](s^k_h, a^k_h) \leq 2[\mathbb{E}[V^*_{k,h}]](s^k_h, a^k_h) + 2[\mathbb{E}[\hat V_{k,h+1}]](s^k_h, a^k_h),$$

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

**Lemma D.11** Let $\tilde A_m, \tilde S_m$ 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

$$\tilde A_m \leq \sqrt{2\tilde S_m \log(1/\delta)} + \frac{4}{3} \cdot \log(1/\delta).$$

We denote the corresponding event by $\mathcal{E}_{t_2}$.

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

$$\tilde A_m \leq \sqrt{2\tilde S_m \log(1/\delta)} + \frac{4}{3} \cdot \log(1/\delta),$$

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 $A_m, Q_0, \hat A_m, R_0, G$ be defined in (D.20), (D.19), (D.21), (D.23), (D.24). On the event $\mathcal{E}_{t_1} \cap \mathcal{E}_{\text{var}} \cap \mathcal{E}_c$, we have

$$A_0 \leq 4\sqrt{(Q_0 + \hat A_1 + G + 2(R_0 + G)) \log(1/\delta)} + 10 \cdot \log(1/\delta).$$

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

$$A_0 \leq 2\sqrt{(Q_0 + \hat S_0) \log(1/\delta)} + (2/3) \cdot \log(1/\delta) \leq 2\sqrt{(Q_0 + \hat A_1 + G + 2(R_0 + G + A_0)) \log(1/\delta)} + (2/3) \cdot \log(1/\delta)$$
\[ \leq 2 \sqrt{(Q_0 + \hat{A}_1 + G + 2(R_0 + G)) \log(1/\delta)} + (2/3) \cdot \log(1/\delta) + 2\sqrt{2A_0 \log(1/\delta)} \]
\[ \leq 4 \sqrt{(Q_0 + \hat{A}_1 + G + 2(R_0 + G)) \log(1/\delta)} + 10 \cdot \log(1/\delta), \]

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 \( x \leq a\sqrt{x} + b \Rightarrow x \leq a^2 + 2b \). 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 \( \mathcal{E}_{r_2} \cap \mathcal{E}_{\text{var}} \cap \mathcal{E}_c \), we have

\[ \hat{A}_1 \leq 4 \sqrt{(R_0 + 2G + A_0) \log(1/\delta)} + 11 \log(1/\delta). \]

**Proof**  By the definition of \( \mathcal{E}_{r_2} \) in Lemma D.11, we have

\[ \hat{A}_m \leq \sqrt{2\hat{S}_m \log(1/\delta)} + \frac{4}{3} \cdot \log(1/\delta). \]  \hspace{1cm} (D.29)

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

\[ \hat{A}_m \leq \sqrt{2(\hat{A}_{m+1} + 2^{m+1} \cdot (R_0 + 2G + A_0)) \log(1/\delta)} + \frac{4}{3} \cdot \log(1/\delta). \]

Applying Lemma E.5, we have

\[ \hat{A}_1 \leq \max\left\{11 \log(1/\delta), 4 \sqrt{(R_0 + 2G + A_0) \log(1/\delta)} + 2 \log(1/\delta)\right\} \]
\[ \leq 4 \sqrt{(R_0 + 2G + A_0) \log(1/\delta)} + 11 \log(1/\delta). \]

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 \( \mathcal{E}_{r_1} \cap \mathcal{E}_{r_2} \cap \mathcal{E}_{\text{var}} \cap \mathcal{E}_c \), we have

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

**Proof**  We compute

\[ A_0 \leq 4 \sqrt{(Q_0 + G + 2(R_0 + G)) \log(1/\delta)} + 10 \cdot \log(1/\delta) + 4 \sqrt{\hat{A}_1 \log(1/\delta)} \]
\[ \leq 4 \sqrt{(Q_0 + G + 2(R_0 + G)) \log(1/\delta)} + 2 \hat{A}_1 + 12 \log(1/\delta) \]
\[ \leq 8 \sqrt{(R_0 + 2G + A_0) \log(1/\delta)} + 34 \log(1/\delta) + 4 \sqrt{(Q_0 + G + 2(R_0 + G)) \log(1/\delta)} \]
\[ \leq 132 \log(1/\delta) + 28 \sqrt{R_0 \log(1/\delta)} + 40 \sqrt{G \log(1/\delta)} + 8 \sqrt{Q_0 \log(1/\delta)}, \]

where the first inequality follows from Lemma D.12, the second inequality holds due to the fact that \( 2ab \leq a^2 + b^2 \), the third inequality holds due to Lemma D.13 and the last inequality holds due to the fact that \( x \leq a\sqrt{x} + b \Rightarrow x \leq a^2 + 2b \). Thus, we complete the proof of Lemma D.14.
Lemma D.15  Let $\tilde{S}_m$, $A_m$, $R_0$, $G$ be defined in (D.18), (D.20), (D.23), (D.24). On the event $\mathcal{E}_{\text{var}} \cap \mathcal{E}_c$, we have the following inequalities for all $m \in [M]$:

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

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

$$\tilde{S}_m = \sum_{k=1}^{K} \sum_{h=1}^{H} [V_{k,h+1}^m(s_k^h, a_k^h)]$$

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

$$= \sum_{k=1}^{K} \sum_{h=1}^{H} \left[ \mathbb{P}[\mathbb{V}_{k,h+1}^m(s_k^h, a_k^h)] - \mathbb{E}_{k,h+1}(s_k^h) \right] + \sum_{k=1}^{K} \sum_{h=1}^{H} \left[ \mathbb{V}_{k,h+1}^m(s_k^h) - \left( \mathbb{P}[\mathbb{V}_{k,h+1}^m(s_k^h, a_k^h)] \right)^2 \right].$$

(D.30)

For the second term, we further have

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

$$\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \left[ \mathbb{V}_{k,h+1}^m(s_k^h) - \mathbb{E}_{k,h+1}(s_k^h) \right]$$

$$\leq 2^{m+1} \sum_{k=1}^{K} \sum_{h=1}^{H} \max \left\{ \mathbb{V}_{k,h+1}(s_k^h) - \mathbb{E}_{k,h+1}(s_k^h), 0 \right\}$$

$$\leq 2^{m+1} \sum_{k=1}^{K} \sum_{h=1}^{H} I_k^h \left[ \mathbb{V}_{k,h+1}(s_k^h) - \mathbb{E}_{k,h+1}(s_k^h) \right]$$

$$+ 2^{m+1} \sum_{k=1}^{K} (1 - I_k^h) \sum_{h=1}^{H} \left[ \mathbb{V}_{k,h+1}(s_k^h) - \mathbb{E}_{k,h+1}(s_k^h) \right]$$

$$\leq 2^{m+1} R_0 + 2^{m+1} \tilde{A}_0 + 2^{m+1} \sum_{k=1}^{K} (1 - I_k^h) \sum_{h=1}^{H} \left[ \mathbb{V}_{k,h+1}(s_k^h) - \mathbb{E}_{k,h+1}(s_k^h) \right]$$

(D.31)

where the first inequality holds since

$$\left( \mathbb{P}[\mathbb{V}_{k,h+1}^m(s_k^h, a_k^h)] \right)^2 \geq \left( \mathbb{P}[\mathbb{V}_{k,h+1}^{m-1}(s_k^h, a_k^h)] \right)^2 \geq \cdots \geq \left( \mathbb{P}[\mathbb{V}_{k,h+1}^m(s_k^h, a_k^h)] \right)^2.$$
the fourth inequality holds due to the definition of $R_0$ and $\tilde{A}_0$, the last inequality follows from the fact that $r(s_k^h, a_k^h) \geq 0$.

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

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

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

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

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

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

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

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

$$\tilde{A}_1 \leq 4\sqrt{(R_0 + 2G + A_0 + \tilde{A}_0) \log(1/\delta)} + 11 \cdot \log(1/\delta),$$

$$\tilde{A}_0 \leq 2\sqrt{(R_0 + 2G + A_0 + \tilde{A}_0) \log(1/\delta)} + 7 \cdot \log(1/\delta).$$

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

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

Applying Lemma E.5, we have

$$\tilde{A}_1 \leq \max \left\{ 11 \log(1/\delta), 4\sqrt{(R_0 + 2G + A_0 + \tilde{A}_0) \log(1/\delta)} + 2 \log(1/\delta) \right\}$$

$$\leq 4\sqrt{(R_0 + 2G + A_0 + \tilde{A}_0) \log(1/\delta)} + 11 \log(1/\delta).$$

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

$$\tilde{A}_0 \leq \sqrt{18 \log(1/\delta) + 4 \left( \sqrt{(R_0 + 2G + A_0 + \tilde{A}_0) \log(1/\delta)} \right)^2 \log(1/\delta) + \frac{4}{3} \cdot \log(1/\delta)}$$

$$\leq 2\sqrt{(R_0 + 2G + A_0 + \tilde{A}_0) \log(1/\delta)} + 7 \cdot \log(1/\delta).$$

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

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

We denote the corresponding event by $\mathcal{E}_{r4}$.

**Proof**  By the definition of $Q_0$, we have

$$Q_0 \leq 2\bar{S}_0 + 2 \sum_{k=1}^K \sum_{h=1}^H \mathbb{E} \left[ \mathbb{V} V_{h+1}^{\pi_k} (s_h^k, a_h^k) \right].$$

(D.33)

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

$$\mathbb{E} \left[ (s_h^k, a_h^k) \right] = \mathbb{V} V_{h+1}^{\pi_k} (s_h^k, a_h^k)$$

(D.34)

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

$$\sum_{h=1}^H \mathbb{V} V_{h+1}^{\pi_k} (s_h^k, a_h^k) \leq \sum_{h=1}^H \mathbb{V} V_{h+1}^{\pi_k} (s_h^k, a_h^k) \\ \leq \sum_{h=1}^H 1 \cdot \mathbb{E} \left[ \mathbb{V} V_{h+1}^{\pi_k} (s_h^k, a_h^k) \right] \\ \leq 1,$$

where the last inequality holds due to (D.34). By Freedman’s inequality (Lemma E.4), with probability at least $1 - \delta$, $K \times \sum_{h=1}^H \sum_{h=1}^H \mathbb{V} V_{h+1}^{\pi_k} (s_h^k, a_h^k) \leq \widetilde{O}(K)$.

(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 $\mathcal{E}_{r1} \cap \mathcal{E}_{r2} \cap \mathcal{E}_{r3} \cap \mathcal{E}_{c} \cap \mathcal{E}_{var}$, 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 $\text{Regret}(K)$,

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

41
\[
\leq \sum_{k=1}^{K} \left[ V_{k,1}(s_1^k) - V_1^{\pi_k}(s_1^k) \right]
\]
\[
\leq \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[ V_{k,h}(s_h^k) - V_{k,h+1}(s_{h+1}^k) \right] - \sum_{k=1}^{K} V_1^{\pi_k}(s_1^k) + G
\]
\[
= \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \cdot r(s_h^k, a_h^k) + \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[ V_{k,h}(s_h^k) - r(s_h^k, a_h^k) - [P V_{k,h+1}](s_h^k, a_h^k) \right]
\]
\[
+ \sum_{k=1}^{K} \sum_{h=1}^{H} I_h^k \left[ [P V_{k,h+1}](s_h^k, a_h^k) - V_{k,h+1}(s_{h+1}^k) \right] - \sum_{k=1}^{K} V_1^{\pi_k}(s_1^k) + G
\]
\[
\leq R_0 + A_0 + G + \sum_{k=1}^{K} \left( \sum_{h=1}^{H} r(s_h^k, a_h^k) - V_1^{\pi_k}(s_1^k) \right)
\]

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

For the term \( I_1 \), we have
\[
\sum_{k=1}^{K} \left( \sum_{h=1}^{H} r(s_h^k, a_h^k) - V_1^{\pi_k}(s_1^k) \right) = \sum_{k=1}^{K} \sum_{h=1}^{H} \left[ V_{h}^{\pi_k}(s_h^k) - [P V_{h+1}^{\pi_k}](s_h^k, a_h^k) \right] - \sum_{k=1}^{K} V_1^{\pi_k}(s_1^k)
\]
\[
= \sum_{k=1}^{K} \sum_{h=1}^{H} \left[ V_{h+1}^{\pi_k}(s_{h+1}^k) - [P V_{h+1}^{\pi_k}](s_h^k, a_h^k) \right]
\]
\[
\leq |A_0| + |\tilde{A}_0| + |\tilde{A}_0|,
\]

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
\[
R_0 \leq \tilde{O} \left( d \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \sigma^2_{h,k} + d} \right)
\]
\[
\leq \tilde{O} \left( d \sqrt{Q_0} + d \sqrt{S_0} + d \right)
\]
\[
\leq \tilde{O} \left( d \sqrt{Q_0} + d \sqrt{\tilde{A}_1 + G + R_0 + A_0} + d \right)
\]
\[
\leq \tilde{O} \left( d \sqrt{Q_0} + d \sqrt{G + R_0 + \sqrt{Q_0} + \sqrt{R_0} + d} \right)
\]
\[
\leq \tilde{O} \left( d \sqrt{Q_0} + d^2 \right),
\]

where the first inequality follows from Lemma D.8, the second inequality follows from the definition of \( S_0 \) 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

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

Putting everything together, we have

$$\text{Regret}(K) \leq \tilde{O}\left(d\sqrt{Q_0} + d^2\right).$$

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:

$$\text{Regret}(K) \leq \tilde{O}\left(d\sqrt{K} + d^2\right).$$

Proof We prove this corollary on the event $\mathcal{E}_{r_1} \cap \mathcal{E}_{r_2} \cap \mathcal{E}_{r_3} \cap \mathcal{E}_{r_4} \cap \mathcal{E}_c \cap \mathcal{E}_{\text{var}'}$, 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 $\mathcal{E}_{r_4}$ in Lemma D.18, we have

\[
Q_0 \leq 2S_0 + \tilde{O}(K)
\]

\[
\leq 2\tilde{A}_1 + G + 2(R_0 + G + A_0) + \tilde{O}(K)
\]

\[
\leq 8\sqrt{(R_0 + 2G + A_0 + \tilde{A}_0)\log(1/\delta)} + 22 \cdot \log(1/\delta) + G + 2(R_0 + G + A_0) + \tilde{O}(K)
\]

\[
\leq \tilde{O}\left(d\sqrt{Q_0} + d^2 + K\right)
\]

\[
\leq \tilde{O}(K + d^2),
\]

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 $x \leq a\sqrt{x} + b \Rightarrow x \leq a^2 + 2b$.

By Theorem 3.3, we can obtain

$$\text{Regret}(K) \leq \tilde{O}(d\sqrt{Q_0} + d^2) \leq \tilde{O}(d\sqrt{K} + d^2).$$

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 $\{G_i\}$ satisfying $|x_i| \leq M$ for some constant $M$, $x_i$ is $G_{i+1}$-measurable, $\mathbb{E}[x_i|G_i] = 0$. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\sum_{i=1}^n x_i \leq M\sqrt{2n \log(1/\delta)}.$$
Lemma E.2 (Lemma 11, Abbasi-Yadkori et al. 2011) For any $\lambda > 0$ and sequence $\{x_k\}_{k=1}^K \subset \mathbb{R}^d$ for $k \in [K]$, define $Z_k = \lambda I + \sum_{i=1}^{k-1} x_i x_i^T$. Then, provided that $\|x_k\|_2 \leq L$ holds for all $k \in [K]$, we have

$$
\sum_{k=1}^K \min \left\{1, \|x_k\|_2^2 / Z_k^{-1}\right\} \leq 2d \log \left(1 + KL^2 / (d\lambda)\right).
$$

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

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

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

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

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

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 $a_i \leq \lambda_2 \sqrt{a_{i+1}} + \lambda_4$ hold for any $1 \leq i \leq i'$. Then we have that

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