Nearly Minimax Optimal Regret for Learning Linear Mixture Stochastic Shortest Path

Qiwei Di [1] **Jiafan He** [1] **Dongruo Zhou** [1] **Quanquan Gu** [1]

Abstract

We study the Stochastic Shortest Path (SSP) problem with a linear mixture transition kernel, where an agent repeatedly interacts with a stochastic environment and seeks to reach certain goal state while minimizing the cumulative cost. Existing works often assume a strictly positive lower bound of the cost function or an upper bound of the expected length for the optimal policy. In this paper, we propose a new algorithm to eliminate these restrictive assumptions. Our algorithm is based on extended value iteration with a fine-grained variance-aware confidence set, where the variance is estimated recursively from high-order moments. Our algorithm achieves an $\widetilde{\mathcal{O}}(dB_*\sqrt{K})$ regret bound, where d is the dimension of the feature mapping in the linear transition kernel, B_* is the upper bound of the total cumulative cost for the optimal policy, and K is the number of episodes. Our regret upper bound matches the $\Omega(dB_*\sqrt{K})$ lower bound of linear mixture SSPs in Min et al. (2022), which suggests that our algorithm is nearly minimax optimal.

1. Introduction

Stochastic Shortest Path (SSP) (Bertsekas, 2012) is a type of reinforcement learning problem where the agent aims to reach a predefined goal state while minimizing the total expected cost. In an SSP, for each episode, the agent starts at a specific initial state, chooses an action from the action set, receives some cost from the environment, and transits to the next state. The agent will stop at a fixed goal state (i.e., terminal state) and ends the current episode. Compared with episodic Markov Decision Processes (MDPs) and infinite-horizon MDPs, the SSP model is more general and thus more suitable to many modern applications such as Atari games, GO games, and navigation (Andrychowicz et al., 2017; Nasiriany et al., 2019).

For an SSP, since the agent only stops after reaching a goal state, the length of the episode usually depends on the current policy and can be different from episode to episode. Therefore, learning an SSP is usually more difficult than learning episodic MDPs and infinite-horizon MDPs. In recent years, there has been a sequence of works developing efficient algorithms for learning SSPs. We use regret to measure each algorithm, which is defined as the difference between the total cost and the lowest expected cost achieved by the optimal policy. For the tabular SSP setting where the state space and action space are finite, Tarbouriech et al. (2020) proposed an algorithm with a regret of $\widetilde{\mathcal{O}}(D^{3/2}S\sqrt{AK/c_{\min}})$, where D is the smallest expected hitting time from any starting state to the goal state and $c_{\min}$ is the assumed positive lower bound of the cost function. Rosenberg et al. (2020) proposed an algorithm with a regret of $\widetilde{\mathcal{O}}(B_*S\sqrt{AK})$ and showed that every algorithm should suffer from an $\Omega(B_*\sqrt{SAK})$ regret. Later, Cohen et al. (2021) developed an algorithm reduced from algorithms for episodic MDPs, which achieves the minimax lower bound. Tarbouriech et al. (2021b) made significant contributions to the study of SSP. One of their notable achievements is the development of an algorithm that does not rely on the assumption that $c_{\min} > 0$. This algorithm achieves a nearly optimal regret bound of $\widetilde{\mathcal{O}}(B_*\sqrt{SAK})$. Furthermore, they introduced the algorithms that do not require knowledge of T_* or B_* as well.

Many modern RL problems work with a large state and action spaces. In these cases, linear function approximation can be employed as a tool to make RL scalable to large state and action spaces (Bradtke & Barto, 1996). For the SSP setting, Vial et al. (2022) is the first one to consider linear function approximation on it. They proposed a computationally inefficient algorithm with an $\widetilde{\mathcal{O}}(\sqrt{d^3B_*^3K/c_{\min}})$ regret, where d is the dimension of the linear representation used in the algorithm. Later, Chen et al. (2022) proposed a computationally efficient algorithm with an improved regret of $\widetilde{\mathcal{O}}(\sqrt{d^3B_*^2T_*K})$ using the fact that $T_* \leq B_*/c_{\min}$. Min et al. (2022) considered the linear mixture SSP setting and proposed an algorithm LEVIS$^+$ with a regret of

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

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

$\widetilde{\mathcal{O}}(dB_* \sqrt{K/c_{\min}})$. Chen et al. (2022) also proposed an algorithm (UCRL-VTR-SSP) for linear mixture SSP with an $\widetilde{\mathcal{O}}(B_* \sqrt{dT_* K} + dB_* \sqrt{K})$ regret. On the other hand, Min et al. (2022) proved a lower bound of $\Omega(dB_* \sqrt{K})$.

However, the regret bounds of all the above works with linear function approximation depend on $c_{\min}$ or the expected length T_* polynomially, which prevents these algorithms from matching the lower bound, unlike their counterparts in the tabular setting. Therefore, a natural question arises:

Can we design an optimal algorithm for linear mixture SSPs, whose regret matches the lower bound?

Our work gives a positive answer to this question. We highlight our main contributions as follows,

- We propose a computationally-efficient algorithm for learning linear mixture SSPs. Our regret bound is $\widetilde{\mathcal{O}}(dB_* \sqrt{K})$, which matches the lower bound in Min et al. (2022) up to logarithmic factors. To the best of our knowledge, this is the first statistically near-optimal algorithm for learning SSPs with linear function approximation.
- Our algorithm has a component that estimates the optimal value function by solving a weighted regression problem, following (Min et al., 2022). The difference between our approach and the previous one is that the weights adapted in our weighted regression depend on both the variance of the estimated value function and the upper bound of the error between the optimal value function and the estimated value function, which has been studied in the design of horizon-free algorithms of linear mixture MDPs (Zhou & Gu, 2022). Our newly adapted weights enable us to obtain a more accurate estimate of the value function and improve the final regret.
- We also introduce a more delicate variance estimator. To do this, we introduce high-order moment estimates of the value function and build these estimates by solving multiple groups of weighted regressions. Compared with Min et al. (2022), our proposed variance estimates are more accurate, which help us eliminate the polynomial dependence of $c_{\min}$ in the regret.

Notation. For any positive number n, we denote by $[n] = \{1, 2, 3, \dots, n\}$. We use lowercase letters to denote scalars and use lower and uppercase bold face letters to denote vectors and matrices respectively. For a vector $\mathbf{x} \in \mathbb{R}^d$ and matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, we define $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma} \mathbf{x}}$ and define $\|\mathbf{x}\|_\infty = \max_i |x_i|$ to be the infinity norm of a vector. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an absolute constant C such that $a_n \le Cb_n$, and we write $a_n = \Omega(b_n)$ if there exists an absolute constant C such that $a_n \ge Cb_n$. We use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to further hide the logarithmic factors. We use $\mathbb{1}\{\}$ to denote the indicator function. For $a, b \in \mathbb{R}$ satisfying $a < b$, we use $[x]_{[a,b]}$ to denote the truncation function $x \cdot \mathbb{1}\{a \le x \le b\} + a \cdot \mathbb{1}\{x < a\} + b \cdot \mathbb{1}\{x > b\}$.

2. Related Work

Tabular SSP. Stochastic Shortest Path (SSP) is a popular variant of Markov Decision Process, which can be dated back to Bertsekas & Tsitsiklis (1991); Bertsekas & Yu (2013); Bertsekas (2012). The regret minimization problem of SSP was first studied by Tarbouriech et al. (2020), which proposed the first algorithm with a regret of $\widetilde{\mathcal{O}}(D^{3/2} S \sqrt{AK/c_{\min}})$, and a parameter-free algorithm with an $\mathcal{O}(K^{3/2})$ regret bound. Here D is the smallest expected hitting time from any starting state to the goal state and $c_{\min}$ is the assumed positive lower bound of the cost function. It was improved by Rosenberg et al. (2020) to $\mathcal{O}(B_* S \sqrt{AK})$ when B_* is known and $\mathcal{O}(B_*^{3/2} S \sqrt{AK})$ in the parameter-free case. There is still a $\sqrt{S}$ gap from the lower bound of $\Omega(B_* \sqrt{SAK})$ proved in the same paper. Later, Cohen et al. (2021) proposed an algorithm using the technique of reducing SSP to a finite-horizon MDP with a large terminal cost. This algorithm achieves the lower bound, but it requires some prior knowledge of T_*, which can be bypassed by using the trivial upper bound $T_* \le B_*/c_{\min}$, and B_* to properly tune the horizon and terminal cost in the reduction. As mentioned in Remark 2 of Cohen et al. (2021), this large dependence on $1/c_{\min}$ will not work well without the assumption $c_{\min} > 0$. Concurrently, Tarbouriech et al. (2021b) avoided this requirement. They first developed an algorithm that knows T_* without assuming $c_{\min} > 0$. This algorithm achieves an $\widetilde{\mathcal{O}}(B_* \sqrt{SAK})$ regret upper bound, matching the lower bound. They also introduced a parameter-free algorithm that does not require knowing T_* in advance. For the case where B_* is unknown, Tarbouriech et al. (2021b) proposed an algorithm with a 'doubling trick' to guess the unknown B_* from scratch. Using the analysis framework called implicit finite horizon approximation, Chen et al. (2021a) proposed the first model-free algorithm which is minimax optimal under strictly positive costs. They also introduced a model-based minimax optimal algorithm without this assumption that is computationally more efficient. In other aspects of the literature, Jafarnia-Jahromi et al. (2021) introduced the first posterior sampling algorithm for SSP. Tarbouriech et al. (2021a) studied the problem of SSP with access to a generative model. Moreover, Rosenberg & Mansour (2020); Chen & Luo (2021); Chen et al. (2021b) studied the problem with adversarial costs.

RL with Linear Function Approximation. There exists a large number of works studying RL with linear function approximation (Yang & Wang, 2019; Jin et al., 2020; Du et al., 2019; Zanette et al., 2020; Wang et al., 2020; Fei et al., 2021;

Table 1. Comparison of algorithms for learning SSP in terms of their regret guarantee.

Model	Algorithm	Regret
Tabular SSP	Bernstein-SSP (Rosenberg et al., 2020)	$\widetilde{O}\Big(B_* S\sqrt{AK}\Big)$
	ULCVI (Cohen et al., 2021)	$\widetilde{O}\Big(\sqrt{(B_*^2+B_*)SAK}\Big)$
	EB-SSP (Tarbouriech et al., 2021b)	$\widetilde{O}\Big(\sqrt{(B_*^2+B_*)SAK}+B_*S^2A\Big)$
	Lower Bound (Rosenberg et al., 2020)	$\Omega(B_*\sqrt{SAK})$
Linear Mixture SSP	LEVIS$^+$ (Min et al., 2022)	$\widetilde{O}\Big(dB_*\sqrt{K/c_{\text{min}}}\Big)$
	UCRL-VTR-SSP (Chen et al., 2022)	$\widetilde{O}\Big(B_*\sqrt{dT_*K}+dB_*\sqrt{K}\Big)$
	LEVIS^{++} (**Our work**)	$\widetilde{O}\Big(dB_*\sqrt{K}+d^2B_*\Big)$
	Lower Bound (Min et al., 2022)	$\Omega(dB_*\sqrt{K})$

Zhou et al., 2021b; He et al., 2021; Zhou & Gu, 2022). The counterpart of the SSP we study in episodic MDPs is called linear mixture MDPs, where the transition probability of the MDP is based on a linear mixture model (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Min et al., 2021; Zhou et al., 2021b). Zhou et al. (2021a) proposed an algorithm to achieve a nearly minimax optimal regret bound in episodic MDP. Recently, a new work can achieve horizon-free regret bound for linear mixture MDPs (Zhou & Gu, 2022). In the SSP setting, Vial et al. (2022) is the first to study a linear SSP model, which assumes there exist some feature maps and that both the cost function and the transition probability are linear in the feature maps. They proposed a computationally inefficient algorithm with a regret of $\widetilde{\mathcal{O}}(\sqrt{d^3B_*^3K/c_{\text{min}}})$. Chen et al. (2022) improved this result by a computationally efficient algorithm with an $\widetilde{\mathcal{O}}(\sqrt{d^3B_*^2T_*K})$ regret. To avoid the undesirable dependency on T_*, Chen et al. (2022) also proposed a computationally inefficient algorithm with a regret bound of $\widetilde{\mathcal{O}}(d^{3.5}B_*\sqrt{K})$ by constructing some confidence sets.

Linear Mixture SSP. Linear Mixture SSP is a different type of linear function approximation from linear SSP, which was first studied by Min et al. (2022). In their work, they proposed an algorithm (LEVIS) with a regret of $\widetilde{\mathcal{O}}(dB_*^{1.5}\sqrt{K/c_{\text{min}}})$ and an improved version (LEVIS$^+$) with a regret of $\widetilde{\mathcal{O}}(dB_*\sqrt{K/c_{\text{min}}})$. Chen et al. (2022) proposed another algorithm (UCRL-VTR-SSP) with an $\widetilde{\mathcal{O}}(B_*\sqrt{dT_*K}+dB_*\sqrt{K})$ regret. When $d \geq T_*$, the result is nearly optimal, but in other cases, the dependency on T_* is undesirable. Similar to the tabular setting, this dependency can be bypassed by replacing T_* with the upper bound $T_* \leq B_*/c_{\text{min}}$.

3. Preliminaries

Stochastic Shortest Path. An SSP is a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, c, s_{\text{init}}, g)$, where: $\mathcal{S}$ is the state space, $\mathcal{A}$ is a finite action space, s_{init} is the initial state and $g \in S$ is the goal state. $\mathbb{P}(s'|s,a)$ is the probability that action a in state s will lead to state s' at the next step, g is an absorbing state, $\mathbb{P}(g|g,a) = 1$ for all action $a \in \mathcal{A}$, c is a function from $\mathcal{S} \times \mathcal{A}$ to $[0,1]$, where $c(s,a)$ is the immediate cost function of taking action a in the state s. In addition, we assume that in the goal state g, the cost for any action $a \in \mathcal{A}$ satisfies $c(g,a) = 0$.

Linear Mixture SSP. We assume that the unknown transition probability $\mathbb{P}$ is a linear mixture function of feature mapping (Modi et al., 2020; Ayoub et al., 2020; Zhou et al., 2021a).

Assumption 3.1 (Linear Mixture SSP, Min et al. 2022)**.** We assume $\mathbb{P}(s'|s,a) = \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta}^* \rangle$, with $\|\boldsymbol{\theta}^*\|_2 \leq 1$, where the feature mapping $\boldsymbol{\phi}(s'|s,a) : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is known. For simplicity, for any bounded function $V : \mathcal{S} \to [0,1]$, we define the notation $\boldsymbol{\phi}_V$ as following: $\boldsymbol{\phi}_V(s,a) = \sum_{s' \in \mathcal{S}} \boldsymbol{\phi}(s'|s,a)V(s')$ and we also assume $\|\boldsymbol{\phi}_V(s,a)\|_2 \leq 1$.

Proper Policies. We consider stationary and deterministic policies in this work, where each of them is a mapping $\pi : \mathcal{S} \to \mathcal{A}$, such that in state s, the agent will take action $\pi(s) \in \mathcal{A}$. A policy π is proper if, with probability 1, it can get to the goal state in finite time. This definition of proper policies is the same as that in Tarbouriech et al. (2021b); Min et al. (2022). We define Π_{proper} to be set of all the proper policies. We make the assumption that Π_{proper} is non-empty.

Assumption 3.2. At least one proper stationary and deterministic policy exists. $\Pi_{\text{proper}} \neq \emptyset$

Value Function. We define the value function and the corresponding Q-function as below.

$$V^\pi(s) := \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^T c(s_t,\pi(s_t))\Big| s_1 = s\right],$$

$$Q^\pi(s,a) :=$$

$$\lim_{T\to\infty} \mathbb{E}\left[c(s_1,a_1) + \sum_{t=2}^T c(s_t,\pi(s_t))\Big| s_1 = s, a_1 = a\right].$$

For a proper policy π and all state-action pair (s,a), we have $V^\pi(s), Q^\pi(s,a) < \infty$. We define

$$\begin{aligned}\mathbb{P}V(s,a) &= \sum_{s'\in S} \mathbb{P}(s'|s,a)V(s') \\ &= \sum_{s'\in S} \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta}^*\rangle V(s') \\ &= \langle \boldsymbol{\phi}_V(s,a), \boldsymbol{\theta}^*\rangle,\end{aligned}$$

and the Bellman operator as

$$\mathcal{L}V(s) = \min_{a\in\mathcal{A}} \Big\{ c(s,a) + \mathbb{P}V(s,a)\Big\}.$$

By satisfying an additional assumption, the lemma presented here shows that we can derive an optimal policy denoted by π^*, which possesses numerous significant properties.

Lemma 3.3 (Bertsekas & Tsitsiklis 1991; Yu & Bertsekas 2013; Tarbouriech et al. 2021b)**.** *Suppose that Assumption 3.2 holds and for every improper policy π, there exists at least one state s, such that $V^\pi(s) = \infty$, then there exists an optimal policy π^*, which is a stationary, deterministic, and proper. What's more, $V^* = V^{\pi^*}$ is the unique solution of the equation $V = \mathcal{L}V$.*

Note that if the cost function is strictly positive, the second assumption inherently satisfied. In the absence of this assumption and with the knowledge of T_*, we employ the perturbation technique used in Tarbouriech et al. (2021b); Min et al. (2022) to bypass the second assumption. To clarify the discussion, we first propose an algorithm under the assumption that $c_{\text{min}} > 0$. The regret of this algorithm is proven in Theorem 5.1. In cases where this assumption does not hold, we introduce a positive parameter $\rho > 0$. We then apply our algorithm to a perturbed problem with a modified cost function defined as $c^\rho_{k,i} := \rho + c_{k,i}$, where $c_{k,i}$ represents the received cost. Simultaneously, we adjust the parameter $B_\rho := B + T_*\rho$. This adjustment ensures that our algorithm can handle an upper bound of the modified cost function. We prove the regret bound for this case in Theorem 5.3.

We denote by π^* the optimal policy. We define $V^*(s) = V^{\pi^*}(s) = \min_{\pi\in\Pi^*} V^\pi(s)$, $B_* = \max_{s\in\mathcal{S}} V^*(s)$ and $Q^*(s,a) = Q^{\pi^*}(s,a)$. We assume that we know an upper bound $B \geq B_*$. Denote by $T^\pi(s)$ the expected time that policy π takes to reach the goal state g starting from s. T_* is defined to be the expected time for the optimal policy to reach goal, i.e. $T_* = \max_{s\in\mathcal{S}} T^{\pi^*}(s)$.

Regret. The regret over the total K episodes is defined as

$$R_K = \sum_{k=1}^K \sum_{i=1}^{I_k} c_{k,i} - KV^*(s_{\text{init}}), \tag{3.1}$$

where I_k is the length of the k-th episode and $c_{k,i} = c(s_{k,i}, a_{k,i})$ is the cost triggered at the i-th step in the k-th episode. Our learning goal is to minimize this regret.

4. The Proposed Algorithm

In this section, we will propose our algorithm for linear mixture SSPs.

4.1. Algorithm Description

Our algorithm is displayed in Algorithm 1. Generally speaking, Algorithm 1 follows LEVIS$^+$ (Min et al., 2022) to construct $\widehat{\boldsymbol{\theta}}_{t,0}$ as the estimate of the model parameter $\boldsymbol{\theta}^*$ at the t-th step, using a weighted linear regression (Line 7 to 9) with weights $\bar{\sigma}_{t,0}$ (Line 6). In detail, $\widehat{\boldsymbol{\theta}}_{t,0}$ is the solution of the weighted regression problem

$$\widehat{\boldsymbol{\theta}}_{t,0} = \mathop{\text{argmin}}_{\boldsymbol{\theta}\in\mathbb{R}^d} \Big\{ \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^t \big[\langle \boldsymbol{\phi}_{V_j}(s_i,a_i), \boldsymbol{\theta}\rangle - V_j(s_{i+1})\big]^2 / \bar{\sigma}_{i,0}^{-2} \Big\},$$

where j is the index representing the update times of the confidence region. For simplicity of expression, the dependence of the index j on the specific time i within the summation is omitted in the equation. Given $\widehat{\boldsymbol{\theta}}_{t,0}$, Algorithm 1 occasionally updates the confidence region $\widehat{\mathcal{C}}_j$ (j is the index of the update times of the confidence region) in Line 16 and runs the subalgorithm LEVIS$^+$ (Algorithm 2) to obtain its value function estimates Q_j and V_j.

Our algorithm, similar to LEVIS$^+$ in Min et al. (2022), divides each episode into intervals of different length. The switch between two intervals is triggered by the updating criterion (Algorithm 1 Line 10). In Algorithm 1, we define two indices t and j, where the number of steps is indexed by t and the number of intervals is represented by j. The first updating criterion is based on the determinant of L groups of covariance matrices $\widetilde{\boldsymbol{\Sigma}}_{t,l}, l \in [L]$ of some given features. The definition of the L groups and the features will

Algorithm 1 LEVIS^{++}

```
Require: Regularization parameter $\lambda$, confidence radius $\{\widehat{\beta}_t\}_{t\geq 1}$, level $L$, variance parameters $\alpha_t$, $\gamma$, $[L] = \{0, 1, \cdots, L-1\}$.
 1: Initialize: set $t \leftarrow 1$, $j \leftarrow 0$, $t_0 = 0$,
    $\widetilde{\boldsymbol{\Sigma}}_{0,l} \leftarrow \lambda \boldsymbol{I}$, $\widehat{\boldsymbol{\theta}}_{0,l} \leftarrow 0$, $\widetilde{\mathbf{b}}_{0,l} \leftarrow 0$,
    $Q_0(s,\cdot), V_0(s) \leftarrow 1$ for all $s \neq g$ and 0 otherwise.
 2: for $k = 1, 2, \ldots, K$ do
 3:   Set $s_t = s_{\text{init}}$.
 4:   while $s_t \neq g$ do
 5:     Take action $a_t = \operatorname{argmin}_{a\in\mathcal{A}} Q_j(s_t, a)$,
        receive cost $c_t = c(s_t, a_t)$,
        and next state $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$.
 6:     Set $\bar{\sigma}_{t,l} \leftarrow$ Algorithm 3
        $\big(\{\boldsymbol{\phi}_{V_j^{2^l}}(s_t, a_t), \widehat{\boldsymbol{\theta}}_{t,l}, \widetilde{\boldsymbol{\Sigma}}_{t,l}, \widehat{\boldsymbol{\Sigma}}_{j,l}\}_{l\in[L]}, \widehat{\beta}_t, \alpha_t, \gamma\big)$.
 7:     For $l \in [L]$, set $\widetilde{\boldsymbol{\Sigma}}_{t,l} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{t-1,l}$
            $+\bar{\sigma}_{t,l}^{-2}\boldsymbol{\phi}_{V_j^{2^l}}(s_t, a_t)\boldsymbol{\phi}_{V_j^{2^l}}(s_t, a_t)^\top$.
 8:     For $l \in [L]$, set $\widetilde{\boldsymbol{b}}_{t,l} \leftarrow \widetilde{\boldsymbol{b}}_{t-1,l}$
            $+\bar{\sigma}_{t,l}^{-2}\boldsymbol{\phi}_{V_j^{2^l}}(s_t, a_t)V_j^{2^l}(s_{t+1})$.
 9:     Set $\widehat{\boldsymbol{\theta}}_{t,l} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{t,l}^{-1}\widetilde{\boldsymbol{b}}_{t,l}$.
10:     if $\exists l \in [L]$, $\det(\widetilde{\boldsymbol{\Sigma}}_{t,l}) \geq 2\det(\widetilde{\boldsymbol{\Sigma}}_{t_j,l})$ or $t \geq 2t_j$ then
11:       $j \leftarrow j + 1$.
12:       $t_j \leftarrow t, \epsilon_j \leftarrow \frac{1}{t_j}, q_j \leftarrow \frac{1}{t_j}$.
13:       for $l \in [L]$ do
14:         Set $\widehat{\boldsymbol{\Sigma}}_{j,l} = \widetilde{\boldsymbol{\Sigma}}_{t_j,l}$.
15:       end for
16:       Set confidence ellipsoid
              $\widehat{\mathcal{C}}_j \leftarrow \Big\{\boldsymbol{\theta} : \|\widehat{\boldsymbol{\Sigma}}_{j,0}^{\frac{1}{2}}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{t_j,0})\|_2 \leq \widehat{\beta}_{t_j}\Big\}$.
17:       Set $Q_j(\cdot,\cdot) \leftarrow \text{DEVI}(\widehat{\mathcal{C}}_j, \epsilon_j, q_j)$.
18:       Set $V_j(\cdot) \leftarrow \min_{a\in\mathcal{A}} \bar{Q}_j(\cdot, a)$.
19:     end if
20:     Set $t \leftarrow t + 1$.
21:   end while
22:   Set $j \leftarrow j + 1$.
23: end for
```

be revealed afterwards. If at least one of the determinant of covariance matrices is doubled compared with its determinant at the end of the previous step, we will trigger the DEVI process, update the value function, end the current interval and start a new one. (Line 10). The first updating criterion cannot guarantee the finite length for each interval, so we follow Min et al. (2022) and introduce the second updating criterion. If the number of steps t is doubled compared with the index of step t_j at the end of the previous interval, we will end the current interval and start a new one. This criterion can occur at most $\mathcal{O}(\log T)$ times and will not add too much complexity to the algorithm.

Algorithm 2 DEVI(Min et al., 2022)

```
Require: Confidence set $\mathcal{C}$, error parameter $\epsilon$, set $\mathcal{B}$ defined in (4.1), transition bonus $q$.
 1: Initialize: $i \leftarrow 0$, $Q^{(0)}(\cdot,\cdot) = 0$,
    $V^{(0)}(\cdot) = 0$ and $V^{(-1)}(\cdot) = \infty$.
 2: Set $Q(\cdot,\cdot) \leftarrow Q^{(0)}(\cdot,\cdot)$.
 3: if $\mathcal{C} \cap \mathcal{B} \neq \emptyset$ then
 4:   while $\|V^{(i)} - V^{(i+1)}\|_\infty \geq \epsilon$ do
 5:     $Q^{(i+1)}(\cdot,\cdot) \leftarrow c(\cdot,\cdot)$
            $+(1-q)\min_{\boldsymbol{\theta}\in\mathcal{C}\cap\mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(i)}}(\cdot,\cdot)\rangle$
 6:     $V^{(i+1)}(\cdot) \leftarrow \min_{a\in\mathcal{A}} Q^{(i+1)}(\cdot, a)$
 7:     Set $i \leftarrow i + 1$.
 8:   end while
 9:   $\bar{Q}(\cdot,\cdot) \leftarrow Q^{(i+1)}(\cdot,\cdot)$
10: end if
11: Output $\bar{Q}(\cdot,\cdot)$.
```

We construct the confidence ellipsoid $\widehat{C}_j$ in Line 16 and feed it to Algorithm 2 to get the estimation of the Q-function. We define a constraint set,

$$\mathcal{B} = \big\{\boldsymbol{\theta} : \forall (s,a), \langle \boldsymbol{\phi}(\cdot|s,a), \boldsymbol{\theta}\rangle \text{ is a probability distribution and } \langle \boldsymbol{\phi}(s'|g,a), \boldsymbol{\theta}\rangle = \mathbb{1}\{s' = g\}\big\}. \tag{4.1}$$

It can be shown that $\widehat{C}_j \cap \mathcal{B}$ contains the true parameter $\boldsymbol{\theta}^*$ with high probability. Then each one-step value iteration in DEVI (Algorithm 2, Line 5) applies the Bellman operator to the confidence set $\widehat{C}_j \cap \mathcal{B}$, which will find an optimistic estimate to the true optimal value function V^*. To prevent DEVI runs infinitely long (since the while loop condition in Line 4 may not be satisfied), we follow Min et al. (2022) to add the transition bonus $q_j = 1/t_j$ since the value iteration may not converge without such a bonus (Min et al., 2022). The additional bias caused by this bonus can be bounded by $\mathcal{O}(\log T)$ with our choice of q_j through our next following analysis.

4.2. Comparison with LEVIS$^+$ (Min et al., 2022)

In this subsection, we will compare our algorithm with LEVIS$^+$ (Min et al., 2022). Before telling the key difference between Algorithm 1 and LEVIS$^+$, we first recall the proof of the algorithm LEVIS$^+$ in Min et al. (2022) to see several key technical challenges we need to recover.

LEVIS$^+$ applies weighted ridge regression to obtain their estimate to the θ^* by the regression weights $\widehat{\sigma}_t$. First we rewrite the regret definition R_K by another formulation of double summation, which is

$$R_K + KV^*(s_{\text{init}}) = \sum_{k=1}^{K}\sum_{i=1}^{I_k} c_{k,i} = \sum_{m=1}^{M}\sum_{h=1}^{H_m} c_{m,h}, \tag{4.2}$$

Algorithm 3 High-order Moment Estimation (HOME)

```
Require: Features $\{\boldsymbol{\phi}_{t,l}\}_{l\in[L]}$, vector estimators $\{\widehat{\boldsymbol{\theta}}_{t,l}\}_{l\in[L]}$, covariance matrix $\{\widetilde{\boldsymbol{\Sigma}}_{t,l}, \widehat{\boldsymbol{\Sigma}}_{j,l}\}_{l\in[L]}$, confidence radius $\widehat{\beta}_t$, parameters $\alpha_t$, $\gamma$.
for $l = 0, 1, \dots, L-2$ do
    Set $[\bar{\mathbb{V}}_{t,l}V_{j+1}^{2^l}](s_t, a_t) \leftarrow$
        $\big[\langle \boldsymbol{\phi}_{t,l+1}, \widehat{\boldsymbol{\theta}}_{t,l+1}\rangle\big]_{[0,B^{2^{l+1}}]} - \big[\langle \boldsymbol{\phi}_{t,l}, \widehat{\boldsymbol{\theta}}_{t,l}\rangle\big]^2_{[0,B^{2^l}]}$.
    Set $E_{t,l} = \min\Big\{1, 2\widehat{\beta}_t \big\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{-\frac{1}{2}} \boldsymbol{\phi}_{t,l}/B^{2^l}\big\|_2\Big\}$
        $+ \min\Big\{1, \widehat{\beta}_t \big\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{-\frac{1}{2}} \boldsymbol{\phi}_{t,l+1}/B^{2^{l+1}}\big\|_2\Big\}$.
    Set $\sigma_{t,l}^2 \to [\bar{\mathbb{V}}_{t,l}V_{j+1}^{2^l}](s_t, a_t)/B^{2^{l+1}} + E_{t,l}$.
    Set $\bar{\sigma}_{t,l}^2 \leftarrow B^{2^{l+1}} \max\Big\{\sigma_{t,l}^2, \alpha_t^2,$
        $\gamma^2 \big\|\widetilde{\boldsymbol{\Sigma}}_{t,l}^{-\frac{1}{2}} \boldsymbol{\phi}_{t,l}/B^{2^l}\big\|_2\Big\}$.
end for
Set $\bar{\sigma}_{t,L-1}^2 \leftarrow B^{2^L} \max\Big\{1, \alpha_t^2,$
    $\gamma^2 \big\|\widetilde{\boldsymbol{\Sigma}}_{t,L-1}^{-\frac{1}{2}} \boldsymbol{\phi}_{t,L-1}/B^{2^{L-1}}\big\|_2\Big\}$
Ensure: $\{\bar{\sigma}_{t,l}^2\}_{l\in[L]}$.
```

where m is a regrouping index of k, where the m-th interval has the end points either from the end of an episode or the time steps when the confidence region is updated, H_m is the length of m-th interval. The following theorem plays a central role in Min et al. (2022)'s proof.

Theorem 4.1 (Theorem G.2 in Min et al. 2022)**.** *For any $\delta > 0$, let $\rho = 0, \lambda = 1/B^2$. Supposing that $c_{min} > 0$, $K \geq d^5 + B^2 d^4/c_{min}$, then with probability at least $1 - 7\delta$, The regret of algorithm LEVIS$^+$ satisfies*

$$R_K = \widetilde{\mathcal{O}}\left(\sqrt{B^2 dT + B^2 d^2 M}\right),$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides a term of $C \cdot \log^2\left(TB/\left(\lambda\delta c_{\min}\right)\right)$ for some problem-independent constant C, and $c_{\min}$ is a minimum of the cost function.

However, Theorem 4.1 cannot directly provide a $O(\sqrt{K})$ upper bound for the regret since in the SSP setting, the total number of steps T can be much greater than episode K. In order to control the number of steps T, Min et al. (2022) proved the following upper bound of the total length T,

$$T = \mathcal{O}\left(\frac{KB}{c_{\min}} \log^2\left(\frac{KB}{\lambda\delta c_{\min}}\right)\right),$$

which brings a $c_{\min}$ dependency to the regret. Therefore, if we want to achieve our goal to remove the polynomial dependency of $c_{\min}$ from our regret, one way is to remove the T dependency from the regret in Theorem 4.1.

In the proof of Theorem 4.1, Min et al. (2022) decomposed the regret in the following way: with high probability, we have the following decomposition of the regret,

$$\begin{aligned}
R_K \leq{} & \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j_m}(s_{m,h+1}) - \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h})\right]}_{I_1} \\
&+ \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h})\right]}_{I_2} \\
&+ \sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1})\right] \\
&- \sum_{m\in\mathcal{M}(m)} V_{j_m}(s_{\text{init}}) + 1.
\end{aligned}$$

Here the first and second terms are dominant, denoted by I_1 and I_2. The bounds of I_1 and I_2 in Min et al. (2022) bring a $\sqrt{T}$ dependency in the results. We will go through the process of their proof and see why the claims hold.

For term I_1, they derived the following result: with probability at least $1 - \delta$, the following inequality holds

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H_m}[V_{j_m}(s_{m,h+1}) - \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h})] \\
&\qquad \leq 2B_*\sqrt{2T\log\left(\frac{2T}{\delta}\right)}.
\end{aligned}$$

This inequality will result in the $\sqrt{T}$ dependency in Theorem 4.1. Thus, we need to make a more delicate estimation for term I_2.

For term I_2, Min et al. (2022) proved the following result. With high probability, the following inequality holds

$$\begin{aligned}
&\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\left[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h})\right] \\
&\leq \underbrace{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\left\{B_*, 4\widehat{\beta}_T\left\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\right\|_{\boldsymbol{\Sigma}_t^{-1}}\right\}}_{B_1} \\
&\quad \underbrace{+ (B_* + 1)\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\frac{1}{t_{j_m}}}_{B_2},
\end{aligned}$$

where term B_2 is non-dominant by the following inequality from Min et al. (2022)

$$B_2 \leq 4.5B_*\left[\log\left(1 + \frac{TB_*^2 d}{\lambda}\right) + \log T\right].$$

For term B_1, Min et al. (2022) proved that

$$B_1 \leq \sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\Big(B_* + 4\widehat{\beta}_T\widehat{\sigma}_t\Big)^2}$$
$$\cdot\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\Big\{1, \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}/\widehat{\sigma}_t\|^2_{\boldsymbol{\Sigma}_t^{-1}}\Big\}}$$
$$\leq \sqrt{2d\log(1+T/\lambda)\cdot\Big(2TB_*^2 + 32\widehat{\beta}_T^2\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\widehat{\sigma}_t^2\Big)}.$$

Firstly, this result has a $\sqrt{T}$ dependency. Furthermore, the term $\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\widehat{\sigma}_t^2 = \mathcal{O}(T)$ will also provide a $\sqrt{T}$ dependency. Therefore, if we want to improve the dependency of $c_{\min}$, we need to use more delicate bound for B_1. Furthermore, we will also need a new construction of weights $\widehat{\sigma}_t$ with a smaller upper bound.

4.3. Our Key Techniques

In this subsection, we will highlight the key techniques used in our algorithm, which tackle the technical challenges faced by LEVIS$^+$.

Design of Variance-aware and Uncertainty-aware Weights. From the above analysis, we can see that a 'good' selection of the weights $\bar{\sigma}_{t,0}$ can potentially help us to improve the dependency of T in the final regret. The first notable difference between our $\bar{\sigma}_{t,0}$ and $\widehat{\sigma}_t$ in Min et al. (2022) is that our weights are both *variance-aware* and *uncertainty-aware*. More specifically,

$$\bar{\sigma}_{t,0}^2 \leftarrow B^2\max\Big\{[\bar{\mathbb{V}}_{t,0}V_{j+1}](s_t,a_t)/B^2 + E_{t,0},$$
$$\alpha_t^2, \gamma^2\big\|\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-\frac{1}{2}}\boldsymbol{\phi}_{t,0}/B\big\|_2\Big\},$$

where $\bar{\mathbb{V}}_{t,0}$ is the estimated variance operator we will introduce in the next section, $E_{t,0}$ is low-order error correction term. To compare with, $\widehat{\sigma}_t$ only depends on the variance term. Zhou & Gu (2022) used the uncertainty-aware term to avoid the dependency of the worst-case range of the noise in a regression problem (Theorem 4.1 in Zhou & Gu (2022)). For our setting, the noise represents the uncertainty of the value function V_j, and its range has a worst-case upper bound which depends on $c_{\min}$. Thus, similar to Zhou & Gu (2022), our regret improves the $c_{\min}$ polynomially.

High-order Moment Variance Estimator. Besides the change of the definition we have mentioned , our adapted weights are more refined by using a recursive construction from the high-order moments estimates of the value function. A similar technique was used by Zhou & Gu (2022) to achieve a horizon-free result for linear mixture MDPs. Compared with LEVIS$^+$, from Algorithm 1 Line 6 to Algorithm 1 Line 9, we maintain L groups of weights $\bar{\sigma}_{t,l}$, $l \in [L]$ instead of 2. The l-th group includes the 2^l moment of the value function for each $l \in [L]$ and do the weighted regression according to each group to obtain different $\boldsymbol{\theta}_{t,l}$. A central part of our algorithm is the recursive design of the variance. Intuitively, the variance of a function V is defined as $\mathbb{V}V = \mathbb{E}[V^2] - \mathbb{E}[V]^2$. Since $\widehat{\boldsymbol{\theta}}_{t,1}$ is from the regression of V_j^2 and $\widehat{\boldsymbol{\theta}}_{t,0}$ is from the regression of V_j, it's natural to define the variance estimator

$$\bar{\mathbb{V}}_{t,0}V_j = [\langle\boldsymbol{\phi}_{V_j^2}, \widehat{\boldsymbol{\theta}}_{t,1}\rangle]_{[0,B^2]} - [\langle\boldsymbol{\phi}_{V_j}, \widehat{\boldsymbol{\theta}}_{t,0}\rangle]^2_{[0,B]}.$$

The truncation is used since the optimal value function should satisfy $V^* \leq B$ by our assumption. Similarly, when we try to estimate the variance of some high-order terms of V_j, we need to use the regression for some higher-order terms. We design the estimated variance of the high-order terms in Algorithm 3 Line 2, which is

$$[\bar{\mathbb{V}}_{t,l}V_{j+1}^{2^l}](s_t,a_t)$$
$$= \big[\big\langle\boldsymbol{\phi}_{t,l+1}, \widehat{\boldsymbol{\theta}}_{t,l+1}\big\rangle\big]_{[0,B^{2^{l+1}}]} - \big[\big\langle\boldsymbol{\phi}_{t,l}, \widehat{\boldsymbol{\theta}}_{t,l}\big\rangle\big]^2_{[0,B^{2^l}]}.$$

In this way, the information of high-order terms is transferred to the estimate of the lower-order terms, which makes our first-level estimate $\widehat{\boldsymbol{\theta}}_{t_j,0}$ affected by all the high-order information and different from its counterpart in Min et al. (2022). The details are shown in Algorithm 3.

To be consistent with our higher-order moment estimation technique, we modify the updating criterion, which will be triggered when any of the determinants of the covariance matrix is doubled or the time is doubled, shown in Algorithm 1 Line 10. In detail, the updating criterion will be triggered if there exists any $l \in [L]$, determinant $\widetilde{\boldsymbol{\Sigma}}_{t,l}$ is doubled. This is different from Min et al. (2022), which only considers $\widetilde{\boldsymbol{\Sigma}}_{t,0}$.

5. Main Results

We show the regret guarantee of Algorithm 1 as follows. Assume $c_{\min}$ is known in prior, then we have the following guarantees:

Theorem 5.1 (Known $c_{\min}$). *Set* $\alpha_t = 1/\sqrt{t}$, $\gamma = d^{-1/4}$, $\lambda = 1/B^2$, $L = \log(5B/c_{min})/\log 2$. *With the assumption of* $c_{min} > 0$, *for any* $\delta > 0$, *set* $\big\{\widehat{\beta}_t\big\}_{t\geq 1}$ *as*

$$\widehat{\beta}_t = 12\sqrt{d\log\left(1+t^2/(d\lambda)\right)\log\left(128(\log(t/d)+2)t^4/\delta\right)}$$
$$+ 30\sqrt{d}\log\big(128(\log(t/d)+2)t^4/\delta\big) + 1. \qquad (5.1)$$

then with probability at least $1-(2L+1)\delta$, *the regret of Algorithm 1 is bounded by*

$$R_K \leq \widetilde{\mathcal{O}}(d^2B + dB\sqrt{K}).$$

Here $\widetilde{\mathcal{O}}$ *hides some logarithmic factors of* K, B, $1/c_{min}$ *and* $1/\delta$.

Remark 5.2. If we know B_*, then we can set $B = B_*$ and get the regret bound of Regret$(K) \le \widetilde{\mathcal{O}}(d^2 B_* + dB_*\sqrt{K})$, which is nearly optimal when $K \ge d^2$. That nearly matches the $\mathcal{O}(dB_*\sqrt{K})$ lower bound proposed by Min et al. (2022).

With the knowledge of T^*, we can build a variant of Algorithm 1 which does not need know $c_{\min}$ in prior, by using the perturbing technique introduced in Tarbouriech et al. (2021b). Specifically, let $\rho > 0$ be some positive parameter. We run Algorithm 1 with the cost defined as $c^{\rho}_{k,i} := \rho + c_{k,i}$, where $c_{k,i}$ is the received cost. Meanwhile, we set $B^{\rho} := B + T_*\rho$. We call the algorithm as ρ-LEVIS^{++}. It is easy to see that $c^{\rho}_{k,i}$ enjoys a uniform lower bound ρ instead of $c_{\min}$. Thus, based on the regret over $c^{\rho}_{k,i}$ and B^{ρ} for ρ-LEVIS^{++} derived from Theorem 5.1, we have the following regret bound over $c_{k,i}$ and B:

Theorem 5.3 (Known T_*)**.** *Set $\rho = (T_*K)^{-1}$, $\alpha_t = 1/\sqrt{t}$, $\gamma = d^{-1/4}$, $\lambda = 1/B^2$, $L = \log(5B/\rho)/\log 2$. For any $\delta > 0$, set $\left\{\widehat{\beta}_t\right\}_{t\ge 1}$ the same as (5.1), then with probability at least $1-(2L+1)\delta$, the regret of ρ-LEVIS^{++} is bounded by*

$$R_K \le \widetilde{\mathcal{O}}(d^2 B + dB\sqrt{K}).$$

Here $\widetilde{\mathcal{O}}$ hides some logarithmic factors of K, B, and $1/\delta$.

6. Proof Sketch

In this section, we will provide the proof sketch of our main results.

6.1. Analysis of DEVI

To analyze DEVI, Min et al. (2022) proved an important fact that the DEVI process will converge and that the true parameter $\boldsymbol{\theta}^*$ will lie in the confidence ellipsoid we construct. Using this fact, we can get a bound of the error between the estimated variance and the true variance. Another important fact is that with high probability, the output of value function $V_j \le V^*$, thus $V_j \le B$ (optimism). But this result is not enough for our purpose because we add the variance of the high-order moment of V_j. So we need to bound the error between the estimation of the variance and the true variance of the higher moments. To deal with this problem, we first use the same argument in Min et al. (2022) to prove the optimism property. For the variance estimation part, we will do induction on l and j. It's similar to the technique used in Zhou & Gu (2022). Our result is summarized in the following lemma.

Lemma 6.1. *Set $\{\widehat{\beta}_t\}$ as (5.1), then with probability at least $1 - L\delta$, for all t and $j = j(t) \ge 1$ as the index of the value functions V at step t, and $l \in [L]$, DEVI converges in finite time and the following holds*

$$0 \le Q_j(\cdot,\cdot) \le Q^*(\cdot,\cdot), \qquad \boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,l},$$

$$\left|\left[\bar{\mathbb{V}}_{t,l} V_j^{2^l}\right](s_t,a_t) - \left[\mathbb{V} V_j^{2^l}\right](s_t,a_t)\right| \le B^{2^{l+1}} E_{t,l}.$$

6.2. Regret Decomposition

Following Min et al. (2022), we divide the time steps into intervals. We will divide a new interval every time the DEVI condition (Line 10) is triggered. Denoted by M, the total number of intervals, we decompose the regret based on intervals. This lemma follows Min et al. (2022).

Lemma 6.2. *Under the event of Lemma 6.1, the regret defined in (3.1) can be decomposed as*

$$\begin{aligned} R_K \le & \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h},a_{m,h}) - V_{j_m}(s_{m,h})\right]}_{I_1} \\ & + \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j_m}(s_{m,h+1}) - \mathbb{P}V_{j_m}(s_{m,h},a_{m,h})\right]}_{I_2} \\ & + \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1})\right] - \sum_{m\in\mathcal{M}(m)} V_{j_m}(s_{\mathit{init}}) + 1.}_{I_3} \end{aligned}$$

Bounding I_1 and I_2. Roughly speaking, I_1 is the accumulated Bellman error of the DEVI outputs. Recall from Line 5 in Algorithm 2 that we add a transition bonus q_j for the sake of convergence, which will cause some bias from the Bellman operator. We choose the same transition bonus $q_j = 1/t_j$ as Min et al. (2022). The result is shown in the following lemma,

Lemma 6.3. *Under the event of Lemma 6.1, we have the following bound for I_1*

$$\begin{aligned} & I_1 - 2 \\ & \le \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\left[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h},a_{m,h}) - V_{j_m}(s_{m,h})\right] \\ & \le \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\left\{B_*, 4\widehat{\beta}_T\left\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\right\|_{\widehat{\boldsymbol{\Sigma}}^{-1}_{t,0}}\right\} \\ & \quad + (B_*+1)\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\frac{1}{t_{j_m}}. \end{aligned}$$

Here the term B_2 is the bias brought by the transition bonus q_j. Min et al. (2022) proved that the bound of B_2 term can be reduced to bounding the total number of calls to DEVI.

For the rest of term I_1 (especially term B_1), and term I_2, we combine them together as the first term of a sequence so that we can use some recursive analysis techniques, which are similar to that in Lattimore & Hutter (2012) and Zhou & Gu (2022).

We define three sequences of quantities, which are

$$R_l = \sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \min\Big\{1, \widehat{\beta}_T \big\| \boldsymbol{\phi}_{V_{j_m}^{2^l}}(s_{m,h}, a_{m,h})/B^{2^l} \big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,l}^{-1}}\Big\},$$

$$S_l = \sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \mathbb{V}V_{j_m}^{2^l}(s_{m,h}, a_{m,h})/B^{2^{l+1}},$$

$$A_l = \sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \Big[\mathbb{P}V_{j_m}^{2^l}(s_{m,h}, a_{m,h}) - V_{j_m}^{2^l}(s_{m,h+1})\Big]/B^{2^l}.$$

Observe that $B_1 \le 4BR_0$ and $I_2 = BA_0$. Therefore, it suffices to bound $A_0 + 4R_0$. Note that if we have the optimism property ($V_j \le V^* \le B$), which is proved to occur with high probability in Section 6.1, these quantities will be less than T, $\forall l$. We can find an relationship of $A_l + 4R_l$ and $A_{l+1} + 4R_{l+1}$ by using the Bernstein concentration inequalities. Then, we can get the bound of $A_0 + 4R_0$; thus the bound of $B_1 + I_2$. The detailed proof is in the appendix.

Bounding I_3. Min et al. (2022) shows that the bound of E_3 is reduced to bounding the number of DEVI calls. Since in our algorithm, DEVI will also be triggered when the determinant of the covariance matrix for high-order terms is doubled, the number of DEVI calls is larger than that in Min et al. (2022). Fortunately, we prove that the error between the two numbers of DEVI calls only differs a logarithmic term, which does not hurt the final regret heavily. It is shown as the following lemma.

Lemma 6.4. *Conditioned on the event of Lemma 6.1, choose parameter $\alpha_t = 1/\sqrt{t}$, $L = \log(5B/c_{\text{min}})/\log 2$. The total number of calls to DEVI J is bounded by $J \le 4dL \log(1 + T/\lambda) + 2 \log T$.*

7. Conclusions

We study the problem of Shortest Stochastic Path, where the transition probability is approximated by a linear mixture model. We propose a novel algorithm and prove its regret upper bound. Our result nearly matches the lower bound. The hyperparameters of our algorithm still depends on $c_{\min}$ or T_*, and we leave the development of a parameter-free algorithm as that in Tarbouriech et al. (2021b) to future work.

Acknowledgements

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

References

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

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

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

Bertsekas, D. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

Bertsekas, D. P. and Tsitsiklis, J. N. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

Bertsekas, D. P. and Yu, H. Stochastic shortest path problems under weak conditions. *Lab. for Information and Decision Systems Report LIDS-P-2909, MIT*, 2013.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Chen, L. and Luo, H. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, pp. 1651–1660. PMLR, 2021.

Chen, L., Jafarnia-Jahromi, M., Jain, R., and Luo, H. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34:10849–10861, 2021a.

Chen, L., Luo, H., and Wei, C.-Y. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference on Learning Theory*, pp. 1180–1215. PMLR, 2021b.

Chen, L., Jain, R., and Luo, H. Improved no-regret algorithms for stochastic shortest path with linear mdp. In *International Conference on Machine Learning*, pp. 3204–3245. PMLR, 2022.

Cohen, A., Efroni, Y., Mansour, Y., and Rosenberg, A. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34:28350–28361, 2021.

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

Fei, Y., Yang, Z., and Wang, Z. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pp. 3198–3207. PMLR, 2021.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

Jafarnia-Jahromi, M., Chen, L., Jain, R., and Luo, H. Online learning for stochastic shortest path model via posterior sampling. *arXiv preprint arXiv:2106.05335*, 2021.

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

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

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

Min, Y., Wang, T., Zhou, D., and Gu, Q. Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34: 7598–7610, 2021.

Min, Y., He, J., Wang, T., and Gu, Q. Learning stochastic shortest path with linear function approximation. In *International Conference on Machine Learning*, pp. 15584–15629. PMLR, 2022.

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

Nasiriany, S., Pong, V., Lin, S., and Levine, S. Planning with goal-conditioned policies. *Advances in Neural Information Processing Systems*, 32, 2019.

Rosenberg, A. and Mansour, Y. Stochastic shortest path with adversarially changing costs. *arXiv preprint arXiv:2006.11561*, 2020.

Rosenberg, A., Cohen, A., Mansour, Y., and Kaplan, H. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Tarbouriech, J., Garcelon, E., Valko, M., Pirotta, M., and Lazaric, A. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pp. 9428–9437. PMLR, 2020.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. Sample complexity bounds for stochastic shortest path with a generative model. In *Algorithmic Learning Theory*, pp. 1157–1178. PMLR, 2021a.

Tarbouriech, J., Zhou, R., Du, S. S., Pirotta, M., Valko, M., and Lazaric, A. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34: 6843–6855, 2021b.

Vial, D., Parulekar, A., Shakkottai, S., and Srikant, R. Regret bounds for stochastic shortest path problems with linear function approximation. In *International Conference on Machine Learning*, pp. 22203–22233. PMLR, 2022.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

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

Yu, H. and Bertsekas, D. P. On boundedness of q-learning iterates for stochastic shortest path problems. *Mathematics of Operations Research*, 38(2):209–227, 2013.

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

Zhang, Z., Yang, J., Ji, X., and Du, S. S. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021a.

Zhang, Z., Zhou, Y., and Ji, X. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pp. 12653–12662. PMLR, 2021b.

Zhou, D. and Gu, Q. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 2022.

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

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

A. Notations

Notation	Meaning
t,T	The number of steps./ The total number of steps.
j	The index of the update times of the confidence region.
s_t, a_t	States and actions our algorithm encounters at time t.
c_t	The cost obtained with state s_t and action a_t
l,L	The level of value function used in the regression./ The maximum level.
$Q_j(\cdot,\cdot), V_j(\cdot)$	The Q-function and value function obtained in the j-th update of the confidence region.
π^*	The optimal policy.
$Q^*(\cdot,\cdot), V^*(\cdot)$	The Q-function and value function obtained with the optimal policy.
$\widehat{\boldsymbol{\theta}}_{t,l}$	The estimated parameter of the linear mixture SSP model obtained by weighted regression at step t with level l.
$\boldsymbol{\theta}^*$	The ground-truth parameter of linear mixture SSP.
$\widehat{\mathcal{C}}_j$	The confidence set obtained in the j-th update of the confidence region, containing $\boldsymbol{\theta}^*$ with high probability.
$\widehat{\beta}_t$	Confidence radius at step t.
$\widetilde{\Sigma}_{t,l}$	The covariance matrix of step t and level l.
$\widehat{\Sigma}_{j,l}$	The covariance matrix used to construct $\widehat{\mathcal{C}}_j$, equal to $\widetilde{\Sigma}_{t_j,l}$.
$\bar{\sigma}_{t,l}$	The weights for regression problems of step t and level l , defined in Algorithm 3.
α_t, γ	Adjustable hyperparameters in the definition of $\bar{\sigma}_{t,l}$.

Table 2. Important Notations

B. Numerical Simulations

We follow the experiment setup in Min et al. (2022) and perform an experiment to compare our algorithm (LEVIS++) and the algorithm LEVIS in Min et al. (2022). The details are presented as follows.

The action space $\mathcal{A} = \{-1, 1\}^{d-1}$ with $|\mathcal{A}| = 2^{d-1}$. The state space is (s_{init}, g). We choose δ, Δ and B_* such that $\delta + \Delta = 1/B_*$ and $\delta > \Delta$. The true model parameter $\boldsymbol{\theta}^*$ is given by

$$\boldsymbol{\theta}^* = \left[\frac{\Delta}{d-1}, \dots, \frac{\Delta}{d-1}, 1\right]^\top.$$

The feature mapping is defined as

$$\begin{aligned}
\boldsymbol{\phi}(s_{\text{init}}, |s_{\text{init}}, \mathbf{a}) &= [-\mathbf{a}, 1-\delta]^\top, \\
\boldsymbol{\phi}(s_{\text{init}}|g, \mathbf{a}) &= \mathbf{0}, \\
\boldsymbol{\phi}(g|s_{\text{init}}, \mathbf{a}) &= [\mathbf{a}, \delta]^\top, \\
\boldsymbol{\phi}(g|g, a) &= [\mathbf{0}_{d-1}, 1]^\top.
\end{aligned}$$

This is a linear mixture SSP with transition function:

$$\begin{aligned}
\mathbb{P}(s_{\text{init}}|s_{\text{init}}, \mathbf{a}) &= 1 - \delta - \langle \mathbf{a}, \boldsymbol{\theta} \rangle, \\
\mathbb{P}(g|s_{\text{init}}, \mathbf{a}) &= \delta + \langle \mathbf{a}, \boldsymbol{\theta} \rangle, \\
\mathbb{P}(g|g, a) &= 1, \\
\mathbb{P}(s_{\text{init}}|g, a) &= 0.
\end{aligned}$$

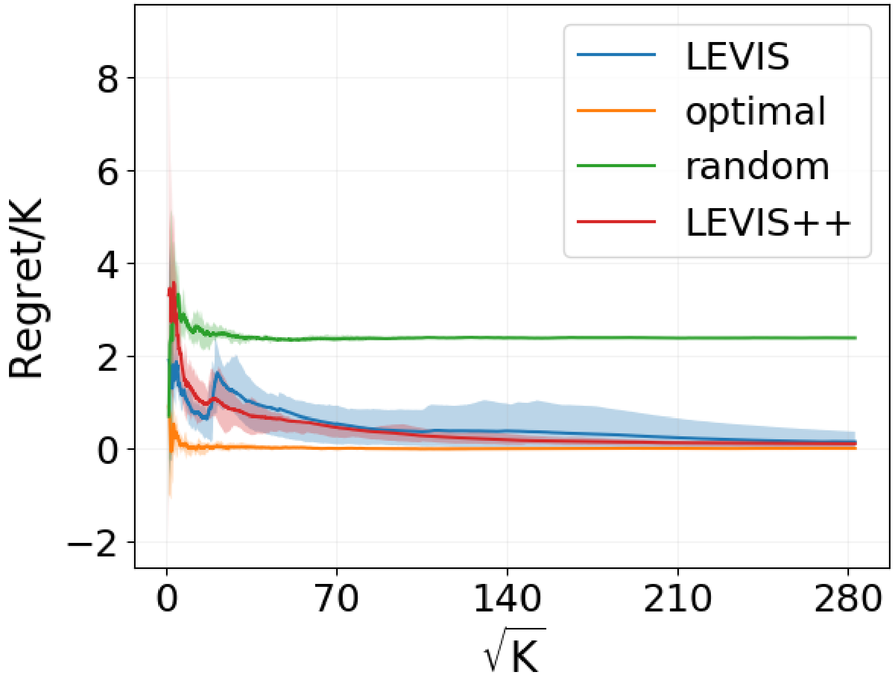

Figure 1. The plot shows the average regret (i.e. R_K/K) and compares the implementation results of Algorithm 1 and LEVIS in Min et al. (2022) on the SSP instance described in Appendix B with $\lambda = 1, \rho = 0$ and failing probability 0.01.

As demonstrated in the graph, our algorithm, LEVIS++, achieves consistently smaller regret than LEVIS in Min et al. (2022). It is worth noting that the expected length for the optimal policy on this synthetic data is relatively small. (There are only two states and the probability of the optimal policy reaching the goal state from the initial state is $1/B_*$, where B_* is set to be 3. Thus the expected total length for every episode is approximately 3.) One of our primary contributions is the elimination of the polynomial dependency on the expected length T_* for the optimal policy. As a result, our algorithm is likely to be more advantageous in numerous real-world applications where the expected length T_* for the optimal policy is even larger.

C. Analysis of Algorithm

C.1. Analysis of DEVI

We define some confidence ellipsoids of different levels. For each $j \in \mathbb{N}$, let $\widehat{\mathcal{C}}_{j,l} = \Big\{\boldsymbol{\theta} : \big\|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}_{t_j,l}(\widehat{\boldsymbol{\theta}}_{t_j,l} - \boldsymbol{\theta})\big\|_2 \le \widehat{\beta}_{t_j}\Big\}$, $l \in [L]$. $\widehat{C}_j = \cap_{l\in[L]}\widehat{C}_{j,l}$. The next lemma shows that with high probability, $\boldsymbol{\theta}^*$ lies in the confidence sets we construct.

Lemma C.1. *(Restatement of Lemma 6.1) Set $\{\widehat{\beta}_t\}$ as (5.1), then with probability at least $1 - L\delta$, for all t and $j = j(t) \ge 1$ as the index of the value functions V at step t, and $l \in [L]$, DEVI converges in finite time and the following holds*

$$0 \le Q_j(\cdot,\cdot) \le Q^*(\cdot,\cdot), \quad \boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,l}, \quad \textit{and} \quad \Big|\big[\bar{\mathbb{V}}_{t,l}V_j^{2^l}\big](s_t,a_t) - \big[\mathbb{V}V_j^{2^l}\big](s_t,a_t)\Big| \le B^{2^{l+1}}E_{t,l}.$$

C.2. Regret Decomposition

We prove the decomposition of regret following the structure in Min et al. (2022). First, we define some notations. The interval is indexed by $1, 2, 3, \ldots$ and the total number of intervals is denoted by M. For the m-th interval, the length is denoted by H_m. We denote by $\mathcal{M}(M)$ the set of intervals which are the first interval of their corresponding episodes. The regret is decomposed as the following lemma shows,

Lemma C.2. *(Restatement of Lemma 6.2) Under the event of Lemma 6.1, for the regret defined in (3.1), we have the* *following decomposition:*

$$R(M) \le \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\big[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h})\big]}_{I_1}$$
$$+ \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\big[V_{j_m}(s_{m,h+1}) - \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h})\big]}_{I_2}$$
$$+ \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\big[V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1})\big] - \sum_{m\in\mathcal{M}(M)} V_{j_m}(s_{\mathit{init}}) + 1}_{I_3}\,.$$

D. Proof of Theorem 5.1 and Theorem 5.3

Denote by $\mathcal{M}_0(M)$ the set of m such that $j_m \ge 2$.

We first deal with the term I_1. Without too much confusion, term I_1 can be seen as the accumulated Bellman error of the DEVI outputs. We first divide I_1 into two terms, the main term caused by the estimation error of vector $\boldsymbol{\theta}^*$ and the second term caused by the transition bonus $q = 1/t_j$, which is shown in the following lemma:

Lemma D.1. *Under the event of Lemma 6.1, we have the following inequality,*

$$\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\big|c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h})\big|$$
$$\le \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\Big\{B_*, 4\widehat{\beta}_T\big\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-1}}\Big\} + (B_*+1)\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\frac{1}{t_{j_m}}.$$

Furthermore, we have

$$|I_1| \le 2 + \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\Big\{B_*, 4\widehat{\beta}_T\big\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-1}}\Big\} + (B_*+1)\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\frac{1}{t_{j_m}}.$$

For simplicity, we define the two terms in the lemma D.1 as B_1 and B_2, where we have

$$\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\big[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h})\big]$$
$$\le \underbrace{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\Big\{B_*, 4\widehat{\beta}_T\big\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-1}}\Big\}}_{B_1} + \underbrace{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\frac{B_*+1}{t_{j_m}}}_{B_2}, \tag{D.1}$$

and we will bound them separately, To bound the B_2 term, we have the following lemmas.

Lemma D.2. *Conditioned on the event in Lemma 6.1, if we choose parameter α_t to be $\alpha_t = 1/\sqrt{t}$, then the total number of calls to DEVI algorithm is bounded by $J \le 4dL\log(1+T/\lambda) + 2\log T$. Furthermore, we have $|M_0| \le J \le 4dL\log(1+T/\lambda) + 2\log T$.*

Lemma D.3. *Using the same condition in Lemma D.2 and the definition of t_{j_m} in algorithm 1, we have an upper bound of B_2 term as follows,*

$$B_2 = \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\frac{B_*+1}{t_{j_m}} \le 5(B_*+1)\Big[2dL\log(1+T/\lambda) + \log T\Big].$$

Proof of Lemma D.3. By rewriting the summation using the index j, using Lemma D.2, we have

$$\begin{aligned} B_2 &\le (B_*+1)\sum_{j=0}^{J}\sum_{t=t_j+1}^{t_{j+1}}\frac{1}{t_j} \le (B_*+1)(J+1) \\ &\le 5(B_*+1)\bigg[2dL\log(1+T/\lambda)+\log T\bigg]. \end{aligned}$$

□

We then bound the term I_3, which follows the proof of Lemma D.4 in Min et al. (2022).

Lemma D.4. *Assuming the event in Lemma 6.1 holds, then we have the following inequality:*

$$\sum_{m=1}^{M}\sum_{h=1}^{H_m}\big[V_{j_m}(s_{m,h})-V_{j_m}(s_{m,h+1})\big]-\sum_{m\in\mathcal{M}(m)}V_{j_m}(s_{\textit{init}})+1 \le 2+4dB_*L\log\Big(1+\frac{T}{\lambda}\Big)+2B_*\log T.$$

For each level $l\in[L]$, we define the following sequences:

$$R_l = \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\Big\{1,\widehat{\beta}_T\big\|\boldsymbol{\phi}_{V_{j_m}^{2^l}}(s_{m,h},a_{m,h})/B^{2^l}\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,l}^{-1}}\Big\}. \tag{D.2}$$

$$S_l = \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\mathbb{V}V_{j_m}^{2^l}(s_{m,h},a_{m,h})/B^{2^{l+1}}. \tag{D.3}$$

$$A_l = \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\big(\mathbb{P}V_{j_m}^{2^l}(s_{m,h},a_{m,h})-V_{j_m}^{2^l}(s_{m,h+1})\big)/B^{2^l}. \tag{D.4}$$

Our goal is to construct the relationship between these sequences, and the following lemma deals with the sequence R_l.

Lemma D.5. *For the sequence R_l, the following inequality holds,*

$$R_l \le 2d\iota+2\widehat{\beta}_T\gamma^2 d\iota+2\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha_{t(m,h)}^2+\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\sigma_{t,l}^2},$$

where $\iota=\log(1+T/d\lambda\alpha_T^2)$. Next, we calculate the term of the sum of the variance in Lemma D.5.

Lemma D.6. *Using the variance $\sigma_{t,l}$ and confidence bonus $E_{t,l}$ defined in Algorithm 3, and under the event of Lemma 6.1, the following inequality holds,*

$$\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\sigma_{t,l}^2 \le 2\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}E_{t,l}+\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\mathbb{V}V_{j_m}^{2^l}(s_{m,h},a_{m,h})/B^{2^{l+1}}.$$

Note that

$$\begin{aligned} E_{t,l} &= \min\Big\{1,2\widehat{\beta}_t\big\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{-\frac{1}{2}}\boldsymbol{\phi}_{t,l}/B^{2^l}\big\|_2\Big\}+\min\Big\{1,\widehat{\beta}_t\big\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{-\frac{1}{2}}\boldsymbol{\phi}_{t,l+1}/B^{2^{l+1}}\big\|_2\Big\} \\ &\le \min\Big\{1,\sqrt{2}\widehat{\beta}_t\big\|\widetilde{\boldsymbol{\Sigma}}_{t,l}^{-\frac{1}{2}}\boldsymbol{\phi}_{t,l}/B^{2^l}\big\|_2\Big\}+\min\Big\{1,\frac{\sqrt{2}}{2}\widehat{\beta}_t\big\|\widetilde{\boldsymbol{\Sigma}}_{t,l+1}^{-\frac{1}{2}}\boldsymbol{\phi}_{t,l+1}/B^{2^{l+1}}\big\|_2\Big\}, \end{aligned}$$

where the last inequality holds due to Lemma H.6 and the fact $\det(\widetilde{\boldsymbol{\Sigma}}_{t,l})\le 2\det(\widehat{\boldsymbol{\Sigma}}_{j,l})$.

Thus, according to the definition of R_l, we get the following inequality,

$$\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}E_{t,l}\le 2R_l+R_{l+1}, \tag{D.5}$$

Combining the results in Lemma D.5, Lemma D.6, and (D.5) , we have

$$R_l \le 2d\iota + 2\widehat{\beta}_T\gamma^2 d\iota + 2\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha_{t(m,h)}^2 + S_l + 4R_l + 2R_{l+1}}. \tag{D.6}$$

In the argument above, we have connected the sequence $\{R_l\}$ with the sequence $\{S_l\}$, and the next two lemmas can connect the bound of S_l with the bound of A_l.

Lemma D.7. *If we define C_M to be the sum of the cost of all steps, i.e. $C_M = \sum_{m=1}^{M}\sum_{h=1}^{H_m} c(s_t, a_t)$. Then we have*

$$S_l \le A_{l+1} + |\mathcal{M}_0| + \frac{2^{l+1}}{B}C_M + \frac{2^{l+1}}{B}(4BR_0 + B_2),$$

where B_2 is the bias from the transition bonus and is defined in (D.1).

Lemma D.8. *Let $\{S_l, A_l\}_{l\in[L]}$ be defined in (D.3) and (D.4). Then we have $\mathbb{P}\left(\mathcal{E}_{D.8}\right) > 1 - L\delta$, where*

$$\mathcal{E}_{D.8} := \left\{\forall l \in [L], |A_l| \le \sqrt{2\zeta S_l} + \zeta\right\}.$$

and $\zeta = 4\log\left(2\log(T\log(1/\delta)) + 1)/\delta\right)$.

Proof of Theorem 5.1. Under the high-probability event $\mathcal{E}_{D.8}$, combining (D.6) with the result in Lemma D.7 , we have the following inequalities for the sequence A_l and R_l,

$$\begin{aligned}
|A_l| &\le 2\sqrt{2\Big[|A_{l+1}| + |\mathcal{M}_0| + \frac{2^{l+1}}{B}C_M + \frac{2^{l+1}}{B}(4BR_0 + B_2)\Big]\zeta} + \zeta,\\
R_l &\le 2d\iota + 2\widehat{\beta}_T\gamma^2 d\iota\\
&\quad + 2\sqrt{d\iota}\widehat{\beta}_T\sqrt{A_{l+1} + \frac{2^{l+1}}{B}(4BR_0 + B_2 + C_M) + 4R_l + 2R_{l+1}}\\
&\quad + 2\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha_{t(m,h)}^2 + |\mathcal{M}_0|},
\end{aligned}$$

where we use the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ when $a, b \ge 0$.

Our goal is to bound $|A_0| + 4R_0$. Using the fact that $\sqrt{a} + \sqrt{b} \le \sqrt{2(a+b)}$, we have

$$\begin{aligned}
|A_l| + 4R_l &\le 8d\iota + 8\widehat{\beta}_T\gamma^2 d\iota + \zeta\\
&\quad + 8\sqrt{d\iota}\widehat{\beta}_T\sqrt{A_{l+1} + \frac{2^{l+1}}{B}(4BR_0 + B_2 + C_M) + 4R_l + 2R_{l+1}}\\
&\quad + 8\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha_{t(m,h)}^2 + |\mathcal{M}_0|}\\
&\quad + 2\sqrt{2\Big[|A_{l+1}| + |\mathcal{M}_0| + \frac{2^{l+1}}{B}C_M + \frac{2^{l+1}}{B}(4BR_0 + B_2)\Big]\zeta}\\
&\le 8d\iota + 8\widehat{\beta}_T\gamma^2 d\iota + \zeta + 8\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha_{t(m,h)}^2 + |\mathcal{M}_0|}\\
&\quad + 2\max\{8\widehat{\beta}_T\sqrt{d\iota}, 2\sqrt{2\zeta}\}\sqrt{|A_l| + 4R_l + |A_{l+1}| + 4R_{l+1} + 2^{l+1}\Big(|A_0| + 4R_0 + \frac{B_2 + C_M}{B}\Big)}.
\end{aligned}$$

Set $a_l = A_l + 4R_l$ to Lemma H.4. Noting that $a_l = A_l + 4R_l \le 5T$ and $|A_0| + 4R_0 + \frac{B_2+C_M}{B} \ge C_M/B \ge c_{\text{min}}T/B$, our choice of $L = \log\left(5B/c_{\text{min}}\right)/\log 2$ can satisfy the condition of Lemma H.4. Thus, we can get an upper bound for

$A_0 + 4R_0$, which is

$$
\begin{aligned}
A_0 + 4R_0 \le 22\Big(2\max\Big\{8\widehat{\beta}_T\sqrt{d\iota}, 2\sqrt{2\zeta}\Big\}\Big)^2 \\
&+ 6\Big(8d\iota + 8\widehat{\beta}_T\gamma^2 d\iota + \zeta + 8\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha^2_{t(m,h)} + |\mathcal{M}_0|}\Big) \\
&+ 4\sqrt{2}\Big(2\max\{8\widehat{\beta}_T\sqrt{d\iota}, 2\sqrt{2\zeta}\}\Big)\sqrt{|A_0| + 4R_0 + \frac{B_2 + C_M}{B}}.
\end{aligned}
$$

Using the fact that $x \le a\sqrt{x} + b \Rightarrow x \le 2a^2 + 2b$, we can further bound $A_0 + 4R_0$ with

$$
\begin{aligned}
A_0 + 4R_0 \le 216\Big(\max\{8\widehat{\beta}_T\sqrt{d\iota}, 2\sqrt{2\zeta}\}\Big)^2 \\
&+ 12\Big(8d\iota + 8\widehat{\beta}_T\gamma^2 d\iota + \zeta + 8\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha^2_{t(m,h)} + |\mathcal{M}_0|}\Big) \\
&+ 8\sqrt{2}\Big(2\max\{8\widehat{\beta}_T\sqrt{d\iota}, 2\sqrt{2\zeta}\}\Big)\sqrt{\frac{B_2 + C_M}{B}}. \qquad \text{(D.7)}
\end{aligned}
$$

Finally, by the decomposition of regret, we the regret is upper bounded by $R(M) \le I_1 + I_2 + I_3$, where $I_1 \le 2 + 4BR_0 + B_2$ due to Lemma D.1 and the definition of R_0, $I_2 = BA_0$ by the definition of A_0, $I_3 \le 2 + 4dB_*L\log\Big(1 + T/\lambda\Big) + 2B_*\log T$ by Lemma D.4. Combining all of these results, we have

$$
R(M) \le B(A_0 + 4R_0) + B_2 + 4dB_*L\log\left(1 + \frac{T}{\lambda}\right) + 2B_*\log T + 5. \qquad \text{(D.8)}
$$

And by the initial definition of $R(M)$, we have

$$
\begin{aligned}
C_M &= R(M) + KV^*(s_{\text{init}}) \\
&\le B(A_0 + 4R_0) + B_2 + 4dB_*L\log\left(1 + \frac{T}{\lambda}\right) + 2B_*\log T + 5 + KV^*(s_{\text{init}}) \\
&\le B_2 + 4dB_*L\log\left(1 + \frac{T}{\lambda}\right) + 2B_*\log T + 5 + KV^*(s_{\text{init}}) \\
&\quad + 216B\big(\max\{8\widehat{\beta}_T\sqrt{d\iota}, 2\sqrt{2\zeta}\}\big)^2 \\
&\quad + 12B\big(8d\iota + 8\widehat{\beta}_T\gamma^2 d\iota + \zeta + 8\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha^2_{t(m,h)} + |\mathcal{M}_0|}\big) \\
&\quad + 8\sqrt{2}B\big(2\max\{8\widehat{\beta}_T\sqrt{d\iota}, 2\sqrt{2\zeta}\}\big)\sqrt{\frac{B_2}{B}} \\
&\quad + 8\sqrt{2}B\big(2\max\{8\widehat{\beta}_T\sqrt{d\iota}, 2\sqrt{2\zeta}\}\big)\sqrt{\frac{C_M}{B}} \\
&\le KV^*(s_{\text{init}}) + \widetilde{\mathcal{O}}(dB\widehat{\beta}_T^2\iota) + \widetilde{\mathcal{O}}(\sqrt{B}\widehat{\beta}_T\sqrt{d\iota}\sqrt{C_M}),
\end{aligned}
$$

where the first inequality holds due to (D.8), the second inequality holds due to (D.7), the last inequality holds due to Lemmas D.2 and D.3 with the choice of parameter, $\alpha_t = 1/t^2$ and $\gamma = d^{-1/4}$. Using the fact that $x \le a\sqrt{x} + b + z \Rightarrow x \le \frac{1}{2}a^2 + b + z + a\sqrt{b + a^2 + z}$, we have,

$$
\begin{aligned}
R(M) &= C_M - KV^*(s_{\text{init}}) \\
&\le \widetilde{\mathcal{O}}(d\iota B\widehat{\beta}_T^2) + \widetilde{\mathcal{O}}(\sqrt{B}\widehat{\beta}_T\sqrt{d\iota}) * \sqrt{KV^*(s_{\text{init}}) + \widetilde{\mathcal{O}}(d\iota B\widehat{\beta}_T^2)} \\
&\le \widetilde{\mathcal{O}}(d\iota B\widehat{\beta}_T^2) + \widetilde{\mathcal{O}}(\sqrt{d\iota}B\widehat{\beta}_T\sqrt{K}). \qquad \text{(D.9)}
\end{aligned}
$$

Here $\widetilde{\mathcal{O}}$ hides some logarithmic factors of T, K, B, $1/c_{\text{min}}$ and $1/\delta$. In addition, we have

$$\begin{aligned} c_{\text{min}}T \le C_M &= R(M) + KV^*(s_{\text{init}}) \\ &\le \widetilde{\mathcal{O}}(d\iota B\widehat{\beta}_T^2) + \widetilde{\mathcal{O}}(\sqrt{d\iota}B\widehat{\beta}_T\sqrt{K}) + KV^*(s_{\text{init}}). \end{aligned}$$

Therefore, we can get an upper bound of T, which is

$$T \le \widetilde{\mathcal{O}}\left(\frac{d^2B + dB\sqrt{K} + KB}{c_{\text{min}}}\right).$$

Here $\widetilde{\mathcal{O}}$ hides some logarithmic factors of K, B, $1/c_{\text{min}}$ and $1/\delta$. Putting the upper bound of T into (D.9), we finish the proof of Theorem 5.1. □

Proof of Theorem 5.3. The optimal policy will not change after we add the perturbation to the cost function. For every step, the perturbed cost function is ρ larger than the original one, thus for the optimal value function after the perturbation, we have $V_\rho^* = V^* + \rho T_*$. In addition, $B + \rho T_*$ can be an upper bound of the perturbed value function of optimal policy V^*. Furthermore, the perturbed cost function has a lower bound ρ. Therefore, we can use the algorithm and results for the perturbed cost function and the regret can be written in the following form,

$$\begin{aligned} R_K &= \sum_{k=1}^{K}\sum_{i=1}^{I_k} c_{k,i} - KV^*(s_{\text{init}}) \\ &\le \sum_{k=1}^{K}\sum_{i=1}^{I_k} c_{k,i}^{\rho} - KV_\rho^*(s_{\text{init}}) + \rho T_* K, \end{aligned}$$

where we use the inequality $c(\cdot,\cdot) \le c^\rho(\cdot,\cdot)$. Then using Theorem 5.1 with $c_{\text{min}} = \rho$, and the upper bound of the optimal value function $B_\rho = B + \rho T_*$, we have the inequality below,

$$\begin{aligned} R_K &\le \widetilde{\mathcal{O}}(d(B+\rho T_*)\widehat{\beta}_T^2 + \sqrt{d}(B+\rho T_*)\widehat{\beta}_T\sqrt{K}) + \rho T_* K \\ &= \widetilde{\mathcal{O}}(dB\widehat{\beta}_T^2 + \sqrt{d}B\widehat{\beta}_T\sqrt{K}). \end{aligned}$$

Here we use the choice of parameter $\rho = (T_*K)^{-1}$ and the $\widetilde{\mathcal{O}}$ hides some logarithmic term of T_*, K, B, $1/\delta$. □

E. Proof of Lemma 6.1

We first prove an upper bound of the error between the variance estimator and the true variance.

Lemma E.1. *For any t, $j = j(t)$ as the index of the value functions V at step t, and $l \in [L]$, let $V_j, \widehat{\boldsymbol{\theta}}_{t,l}, \sigma_t, \widehat{\boldsymbol{\Sigma}}_{t,l}$ be defined in Algorithms 1 and 3. We have the following inequality,*

$$\begin{aligned} \left|\bar{\mathbb{V}}_{t,l}V_j^{2^l}(s_t,a_t) - [\mathbb{V}V_j^{2^l}](s_t,a_t)\right| &\le \min\left\{B^{2^{l+1}}, \left\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{1/2}\left(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t,l+1}\right)\right\|_2 \left\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{-1/2}\boldsymbol{\phi}_{V_j^{2^{l+1}}}(s_t,a_t)\right\|_2\right\} \\ &\quad + \min\left\{B^{2^{l+1}}, 2B^{2^l}\left\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{1/2}\left(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t,l}\right)\right\|_2 \left\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{-1/2}\boldsymbol{\phi}_{V_j^{2^l}}(s_t,a_t)\right\|_2\right\}, \end{aligned}$$

if the inequality $|V_j(s)| \le B$ holds for all state $s \in \mathcal{S}$.

Proof of Lemma E.1. We first substitute the definition of variance estimator (Line 2 of Algorithm 3) into the left-hand side, and the error between the variance estimator and the true variance can be bounded by

$$\begin{aligned} &\left|[\bar{\mathbb{V}}_{t,l}V_j^{2^l}](s_t,a_t) - [\mathbb{V}V_j^{2^l}](s_t,a_t)\right| \\ &= \left|\left[\left\langle\boldsymbol{\phi}_{V_j^{2^{l+1}}}(s_t,a_t), \widehat{\boldsymbol{\theta}}_{t,l+1}\right\rangle\right]_{[0,B^{2^{l+1}}]} - \left\langle\boldsymbol{\phi}_{V_j^{2^{l+1}}}(s_t,a_t), \boldsymbol{\theta}^*\right\rangle\right. \end{aligned}$$

$$
\begin{aligned}
&+\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\boldsymbol{\theta}^*\right\rangle^2-\left[\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\widehat{\boldsymbol{\theta}}_{t,l}\right\rangle\right]^2_{[0,B^{2^l}]}\Bigg| \\
\le\ &\underbrace{\left|\left[\left\langle \boldsymbol{\phi}_{V_j^{2^{l+1}}}\left(s_t,a_t\right),\widehat{\boldsymbol{\theta}}_{t,l+1}\right\rangle\right]_{[0,B^{2^{l+1}}]}-\left\langle \boldsymbol{\phi}_{V_j^{2^{l+1}}}\left(s_t,a_t\right),\boldsymbol{\theta}^*\right\rangle\right|}_{I_1} \\
&+\underbrace{\left|\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\boldsymbol{\theta}^*\right\rangle^2-\left[\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\widehat{\boldsymbol{\theta}}_{t,l}\right\rangle\right]^2_{[0,B^{2^l}]}\right|}_{I_2},
\end{aligned}
$$

where the inequality holds due to the triangle inequality. To bound the term I_1, we have

$$
I_1\le\left|\left\langle \boldsymbol{\phi}_{V_j^{2^{l+1}}}\left(s_t,a_t\right),\widehat{\boldsymbol{\theta}}_{t,l+1}-\boldsymbol{\theta}^*\right\rangle\right|\le\left\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{1/2}\left(\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{t,l+1}\right)\right\|_2\left\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{-1/2}\boldsymbol{\phi}_{V_j^{2^{l+1}}}\left(s_t,a_t\right)\right\|_2,
$$

The first inequality holds because both terms are in $\left[0,B^{2^{l+1}}\right]$, which implied by the assumption of $|V_j(s)|\le B$. The second inequality holds due to Cauchy-Schwarz inequality. Furthermore, the facts that $|V_j(s)|\le B$ and both terms in I_1 lie in $\left[0,B^{2^{l+1}}\right]$ suggest that $I_1\le B^{2^{l+1}}$. Combining these two upper bounds, we have

$$
I_1\le\min\left\{B^{2^{l+1}},\left\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{1/2}\left(\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{t,l+1}\right)\right\|_2\left\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{-1/2}\boldsymbol{\phi}_{V_j^{2^{l+1}}}\left(s_t,a_t\right)\right\|_2\right\}.
$$

To bound the term I_2, we have

$$
\begin{aligned}
I_2=\ &\left|\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\boldsymbol{\theta}^*\right\rangle-\left[\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\widehat{\boldsymbol{\theta}}_{t,l}\right\rangle\right]_{[0,B^{2^l}]}\right| \\
&\cdot\left|\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\boldsymbol{\theta}^*\right\rangle+\left[\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\widehat{\boldsymbol{\theta}}_{t,l}\right\rangle\right]_{[0,B^{2^l}]}\right| \\
\le\ &2B^{2^l}\left|\left\langle \boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right),\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{t,l}\right\rangle\right| \\
\le\ &2B^{2^l}\left\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{1/2}\left(\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{t,l}\right)\right\|_2\left\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{-1/2}\boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right)\right\|_2,
\end{aligned}
$$

where the first inequality holds because $|V_j(s)|\le B$ and thus both terms in the second absolute value are bounded by B^{2^l}, and the second inequality holds by Cauchy-Schwarz inequality. Furthermore, both terms in I_2 lie in $[0,B^{2^{l+1}}]$. Combining these two upper bounds, we have

$$
I_2\le\min\left\{B^{2^{l+1}},2B^{2^l}\left\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{1/2}\left(\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{t,l}\right)\right\|_2\left\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{-1/2}\boldsymbol{\phi}_{V_j^{2^l}}\left(s_t,a_t\right)\right\|_2\right\}.
$$

Thus, we finish the proof of Lemma E.1 by combining the bounds of terms I_1 and I_2. □

Proof of Lemma 6.1. **Special case:** $t=1$

When $t=1$, we have the index of interval $j=1$. In this interval, V_0 is generated from our initialization instead of the DEVI process. For each $l\in[L-1]$, let $\mathbf{x}_{t,l}=\bar{\sigma}_{t,l}^{-1}\boldsymbol{\phi}_{V_j^{2^l}}$, $\eta_{t,l}=\bar{\sigma}_{t,l}^{-1}\left[V_j^{2^l}(s_{t+1})-\langle\boldsymbol{\phi}_{V_j^{2^l}},\boldsymbol{\theta}^*\rangle\right]=0$, $\forall l\in[L]$. $\boldsymbol{\mu}^*=\boldsymbol{\theta}^*$, $y_{t,l}=\langle\boldsymbol{\mu}^*,\mathbf{x}_{t,l}\rangle+\eta_{t,l}$. $\mathbf{Z}_{t,l}=\lambda\mathbf{I}+\sum_{t'=1}^{t}\mathbf{x}_{t',l}\mathbf{x}_{t',l}^\top$, $\mathbf{b}_{t,l}=\sum_{t'=1}^{t}\mathbf{x}_{t',l}y_{t',l}$ and $\boldsymbol{\mu}_{t,l}=\mathbf{Z}_{t,l}^{-1}\mathbf{b}_{t,l}$. Then the following inequalities hold,

$$
\mathbb{E}\left[\eta_{t,l}|\mathcal{G}_{t,l}\right]=0,|\eta_{t,l}|=0,\left|\eta_{t,l}\min\{1,\|\mathbf{x}_{t,l}\|_{\mathbf{Z}_{t-1}^{-1}}\}\right|=0,\|\mathbf{x}_{t,l}\|\le\bar{\sigma}_{t,l}^{-1}\cdot 1\le 1/(\alpha_t B^{2^l}),\mathbb{E}\left[\eta_{t,l}^2|\mathcal{G}_{t,l}\right]=0. \qquad \text{(E.1)}
$$

Using Theorem H.1 with (E.1), we can get $\|\boldsymbol{\mu}_{t,l}-\boldsymbol{\mu}^*\|_{\mathbf{Z}_{t,l}}\le\sqrt{\lambda}\,\|\boldsymbol{\mu}^*\|_2$ with probability at least $1-\delta/2$. After taking an union bound over all level $l\in[L]$, $\boldsymbol{\theta}^*\in\widehat{C}_1$ holds with probability at least $1-L\delta/2$.

At the end of step $t = 1$, the step number t will be doubled and DEVI will be triggered and output V_1. We next show that conditioned on the high-probability event $\{\boldsymbol{\theta}^* \in \widehat{C}_{1,0}\}$, the output of DEVI will satisfy $V_1 \le V^*$ (optimism).

We prove this by induction. First, the initialization $Q^{(0)}$ and $V^{(i)}$ are equal to 0, thus we have $Q^{(0)} \le Q^*$ and $V^{(0)} \le V^*$. For the induction hypothesis, suppose $Q^{(i)} \le Q^*$ and $V^{(i)} \le V^*$ for some i, we want to show $Q^{(i+1)} \le Q^*$ and $V^{(i+1)} \le V^*$. To prove this, by the iteration principle of $Q^{(i)}$, we have

$$\begin{aligned} Q^{(i+1)}(\cdot,\cdot) &= c(\cdot,\cdot) + (1-q) \cdot \min_{\boldsymbol{\theta} \in \widehat{\mathcal{C}}_{1,0} \cap \mathcal{B}} \big\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(i)}}(\cdot,\cdot) \big\rangle \\ &\le c(\cdot,\cdot) + (1-q) \cdot \mathbb{P}V^{(i)}(\cdot,\cdot) \\ &\le Q^{(i)} \\ &\le Q^*, \end{aligned} \tag{E.2}$$

where the first inequality is because we are taking a minimum on a set containing $\boldsymbol{\theta}^*$, the second inequality is by the Bellman equation and the last inequality is by our induction hypothesis.

Initial Step: We then go on to the next interval. From now on, the value function will be the output of DEVI. When the interval number $j = 1$ and $t \in [t_1 + 1, t_2]$ and we suppose the high-probability event $\{\boldsymbol{\theta}^* \in \widehat{C}_j\}$ occurs, we have $V_1 \le V^* \le B$.
For each $l \in [L-1]$, let $\mathbf{x}_{t,l} = \bar{\sigma}_{t,l}^{-1} \boldsymbol{\phi}_{V_j^{2^l}}$ $\eta_{t,l} = \bar{\sigma}_{t,l}^{-1} \mathbb{1}\left\{\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,l} \cap \widehat{\mathcal{C}}_{j,l+1}\right\} \big[V_j^{2^l}(s_{t+1}) - \langle \boldsymbol{\phi}_{V_j^{2^l}}, \boldsymbol{\theta}^* \rangle\big]$ for $l \in [L-1]$. $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_{t,l} = \big\langle \boldsymbol{\mu}^*, \mathbf{x}_{t,l} \big\rangle + \eta_{t,l}$. $\mathbf{Z}_{t,l} = \lambda \mathbf{I} + \sum_{t'=1}^t \mathbf{x}_{t',l}\mathbf{x}_{t',l}^\top$, $\mathbf{b}_{t,l} = \sum_{t'=1}^t \mathbf{x}_{t',l} y_{t',l}$ and $\boldsymbol{\mu}_{t,l} = \mathbf{Z}_{t,l}^{-1}\mathbf{b}_{t,l}$. Then the following inequalities hold,

$$\begin{aligned} &\mathbb{E}\big[\eta_{t,l} \big| \mathcal{G}_{t,l}\big] = 0, |\eta_{t,l}| \le \bar{\sigma}_{t,l}^{-1} B^{2^l} \le 1/\alpha_t, \|\mathbf{x}_{t,l}\|_2 \le \bar{\sigma}_{t,l}^{-1} \cdot B^{2^l} \le 1/\alpha_t, \\ &\big|\eta_{t,l} \min\{1, \|\mathbf{x}_{t,l}\|_{\mathbf{Z}_{t-1}^{-1}}\}\big| \le \bar{\sigma}_{t,l}^{-2} \|\boldsymbol{\phi}_{V_j^{2^l}}\|_{\mathbf{Z}_{t-1}^{-1}} B^{2^l} \le 1/\gamma^2. \end{aligned} \tag{E.3}$$

We also have

$$\begin{aligned} \mathbb{E}\big[\eta_{t,l}^2 \big| \mathcal{G}_{t,l}\big] &= \bar{\sigma}_{t,l}^{-2}\, \mathbb{1}\left\{\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,l} \cup \widehat{\mathcal{C}}_{j,l+1}\right\} \big[\mathbb{V}V_j^{2^l}\big](s_t, a_t) \\ &\le \bar{\sigma}_{t,l}^{-2}\, \mathbb{1}\left\{\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,l} \cup \widehat{\mathcal{C}}_{j,l+1}\right\} \bigg(\big[\mathbb{V}_{t,l}V_j^{2^l}\big](s_t, a_t) \\ &\qquad + \min\left\{B^{2^{l+1}}, 2B^{2^l} \left\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{-\frac{1}{2}} \boldsymbol{\phi}_{V_j^{2^l}}(s_t, a_t)\right\|_2 \cdot \left\|\widehat{\boldsymbol{\Sigma}}_{j,l}^{\frac{1}{2}}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t_j,l})\right\|_2\right\} \\ &\qquad + \min\left\{B^{2^{l+1}}, \left\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{-\frac{1}{2}} \boldsymbol{\phi}_{V_j^{2^{l+1}}}(s_t, a_t)\right\|_2 \cdot \left\|\widehat{\boldsymbol{\Sigma}}_{j,l+1}^{\frac{1}{2}}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t_j,l})\right\|_2\right\}\bigg) \\ &\le \bar{\sigma}_{t,l}^{-2}\Big(\big[\mathbb{V}_{t,l}V_j^{2^l}\big](s_t, a_t) + B^{2^{l+1}} E_{t,l}\Big) \\ &\le 1, \end{aligned} \tag{E.4}$$

where the first equation holds by the definition of $\eta_{t,l}$, the second inequality holds by Lemma E.1, and the third inequality holds by the definition of our confidence ellipsoid $\widehat{C}_{j,l}$, $\widehat{C}_{j,l+1}$ and $E_{t,l}$. The last inequality holds due to the definition of $\bar{\sigma}_{t,l}^{-2}$.

For $l = L-1$, let $\mathbf{x}_{t,L-1} = \bar{\sigma}_{t,L-1}^{-1} \boldsymbol{\phi}_{V_j^{2^{L-1}}}$, $\eta_{t,L-1} = \bar{\sigma}_{t,L-1}^{-1}\big[V_j^{2^{L-1}}(s_{t+1}) - \langle \boldsymbol{\phi}_{V_j^{2^{L-1}}}, \boldsymbol{\theta}^* \rangle\big]$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_{t,L-1} = \big\langle \boldsymbol{\mu}^*, \boldsymbol{x}_{t,L-1} \big\rangle + \eta_{t,L-1}$, $\mathbf{Z}_{t,L-1} = \lambda \mathbf{I} + \sum_{t'=1}^t \mathbf{x}_{t',L-1}\mathbf{x}_{t',L-1}^\top$, $\mathbf{b}_{t,L-1} = \sum_{t'=1}^t \mathbf{x}_{t',L-1} y_{t',L-1}$ and $\boldsymbol{\mu}_{t,L-1} = \mathbf{Z}_{t,L-1}^{-1}\mathbf{b}_{t,L-1}$. Then the following inequalities hold,

$$\begin{aligned} &\mathbb{E}\big[\eta_{t,L-1} \big| \mathcal{G}_{t,L-1}\big] = 0, |\eta_{t,L-1}| \le 1/\alpha_t, \|\boldsymbol{x}_{t,L-1}\| \le \bar{\sigma}_{t,L-1}^{-1} \cdot B^{2^{L-1}} \le 1/\alpha_t, \\ &\big|\eta_{t,L-1} \min\{1, \|\mathbf{x}_{t,L-1}\|_{\mathbf{Z}_{t,L-1}^{-1}}\}\big| \le 1/\gamma^2. \end{aligned} \tag{E.5}$$

And we can get directly by the definition of $\bar{\sigma}_{t,L-1}$ and the fact that $0 \le V_j(s) \le B$,

$$\mathbb{E}\big[\eta_{t,l}^2 \big| \mathcal{G}_{t,l}\big] \le 1. \tag{E.6}$$

Using Theorem H.1 with conditions (E.3) and (E.4) for $l \in [L-1]$ and (E.5) and (E.6) for $l = L$, choose $\epsilon = 1/\gamma^2$. With probability at least $1 - \delta/(t_2(t_2+1))$

$$\|\boldsymbol{\mu}_{t,l} - \boldsymbol{\mu}^*\|_{\mathbf{Z}_{t,l}} \leq \beta_t + \sqrt{\lambda}\,\|\boldsymbol{\mu}^*\|_2\,, \tag{E.7}$$

where

$$\begin{aligned}\beta_t =& 12\sqrt{d\log\left(1 + t/(d\lambda\alpha_t^2)\right)\log\left(256(\log(\gamma^2/\alpha_t)+1)t^4/\delta\right)}\\ &+ \frac{30}{\gamma^2}\log\left(256(\log(\gamma^2/\alpha_t)+1)t^4/\delta\right).\end{aligned}$$

Here we use the fact that $4t^2t_2(t_2+1) \leq 32t^4$ for $t \in [t_1+1, t_2]$, since $\forall t \in [t_1+1, t_2], t_2 \leq 2t, t_2 + 1 \leq 4t$ holds due to our criteria of doubling the number of steps.

We take a union bound over all level $l \in [L]$, and then we have the result that, with probability at least $1 - (L\delta)/(t_2(t_2+1))$, (E.7) holds for all level l simultaneously.

Induction step: Suppose that, with probability at least $1 - \delta'$, inequality (E.7) holds for all $t \in [1, t_{j-1}]$ and all l. For $t \in [t_{j-1}+1, t_j]$, we can define $\boldsymbol{\mu}_{t,l}$ and $\mathbf{Z}_{t,l}$ in the same way as the initial step. We claim that with probability of at least $1 - \delta' - (L\delta)/(t_j(t_j+1))$, inequality (E.7) holds for all $t \in [1, t_j]$ and all l simultaneously. Note that in the previous step, the only condition we use is the optimism $V_1 \leq V^* \leq B_*$. Using the same argument in (E.2), we can see that assuming the event of the true parameter $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,0}$ holds, the output V_j will satisfy $V_j \leq V^*$, thus $V_j \leq V^* \leq B$. Then we can follow the proof in the initial step and see (E.7) holds for $t \in [t_{j-1}+1, t_j]$ and $l \in [L]$ simultaneously with probability at least $1 - (L\delta)/(t_j(t_j+1))$. By taking a union bound, we make an induction from $t \in [1, t_{j-1}]$ to $t \in [1, t_j]$.

Finally, we use induction to see that with probability at least

$$1 - \sum_{j=1}^{J}\frac{L\delta}{t_j(t_j+1)} = 1 - \sum_{j=1}^{J}L\delta\left(\frac{1}{t_j} - \frac{1}{t_j+1}\right) \leq 1 - L\delta,$$

(E.7) holds for all t and $l \in [L]$ simultaneously.

We define $\mathcal{E} = \cup_{t,l}\left\{\, \|\boldsymbol{\mu}_{t,l} - \boldsymbol{\mu}^*\|_{\mathbf{Z}_{t,l}} \leq \beta_t + \sqrt{\lambda}\,\|\boldsymbol{\mu}^*\|_2 \,\right\}$, and we have shown that $\mathcal{E}$ holds with probability of at least $1 - L\delta$. Conditioned on the event $\mathcal{E}$, recalling our definition of $\widehat{\mathcal{C}}_{j,l} = \left\{\boldsymbol{\theta} : \left\|\widehat{\boldsymbol{\Sigma}}_{t_j,l}^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{t_j,l} - \boldsymbol{\theta})\right\|_2 \leq \widehat{\beta}_{t_j}\right\}$, we have the following results.

For $t = 1$, $l \in [L]$, the definition of $\boldsymbol{\mu}_{1,l}$ is just the same as $\widehat{\boldsymbol{\theta}}_{1,l}$ and $\widehat{\boldsymbol{\Sigma}}_{t_j,l} = \mathbf{Z}_{t,l}$ when $j = 1$ and $t = 1$. Thus, we have $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{1,l}, \forall l \in [L]$. Then we consider the term with the highest order. For all j and $l = L-1$, we directly have $\boldsymbol{\mu}_{t_j,L-1} = \widehat{\boldsymbol{\theta}}_{j,L-1}$. Thus, we have $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,L-1}$. For all j and $l \in [L-1]$, we have the following induction argument,

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,l} \cap \widehat{\mathcal{C}}_{j,l+1} \Rightarrow \mathbb{1}\{\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,l} \cap \widehat{\mathcal{C}}_{j,l+1}\} = 1 \Rightarrow \boldsymbol{\mu}_{t,l} = \widehat{\boldsymbol{\theta}}_{t,l}, \mathbf{Z}_{t,l} = \widetilde{\boldsymbol{\Sigma}}_{t,l} \Rightarrow \boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j+1,l}.$$

By induction on l and j, conditioned on event $\mathcal{E}$, we have

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j,l}, \quad 0 \leq Q_j(\cdot,\cdot) \leq Q^*(\cdot,\cdot), \tag{E.8}$$

and according to Lemma E.1 and the definition of the confidence ellipsoid of $\widehat{C}_{j,l}$, we have

$$\left|[\bar{\mathbb{V}}_{t,l}V_j^{2^l}](s_t,a_t) - [\mathbb{V}V_j^{2^l}](s_t,a_t)\right| \leq B^{2^{l+1}}E_{t,l}.$$

□

F. Proof of Lemma D.1

Proof of Lemma D.1. According to the DEVI algorithm, the output Q can be denoted by some iteration of $Q^{(n)}$, i.e.

$$Q_{j_m}(\cdot,\cdot) = Q^{(n)}(\cdot,\cdot) \quad \text{for some iteration } n \in \mathbb{N}$$

$$V_{j_m}(\cdot) = \min_{a \in \mathcal{A}} Q^{(n)}(\cdot, a) = V^{(n)}(\cdot).$$

Through the design of the DEVI algorithm,

$$\begin{aligned}
Q^{(n)}(s_{m,h}, a_{m,h}) &= c_{m,h} + (1-q) \cdot \min_{\boldsymbol{\theta} \in \mathcal{C}_{j_m,0} \cap \mathcal{B}} \big\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(n-1)}}(s_{m,h}, a_{m,h}) \big\rangle \\
&= c_{m,h} + (1-q) \cdot \big\langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V^{(n-1)}}(s_{m,h}, a_{m,h}) \big\rangle \\
&= c_{m,h} + (1-q) \cdot \big\langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V^{(n)}}(s_{m,h}, a_{m,h}) \big\rangle \\
&\quad + (1-q) \cdot \big\langle \boldsymbol{\theta}_{m,h}, \big[\boldsymbol{\phi}_{V^{(n-1)}} - \boldsymbol{\phi}_{V^{(n)}}\big](s_{m,h}, a_{m,h}) \big\rangle \\
&\geq c_{m,h} + (1-q) \cdot \big\langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V^{(n)}}(s_{m,h}, a_{m,h}) \big\rangle - (1-q)\frac{1}{t_{j_m}}, \qquad \text{(F.1)}
\end{aligned}$$

where $\boldsymbol{\theta}_{m,h} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{C}_{j_m,0} \cap \mathcal{B}} \big\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(n-1)}}(s_{m,h}, a_{m,h}) \big\rangle$. The last inequality is because of the terminal condition of our DEVI algorithm.

With a similar argument,

$$Q^{(n)}(s_{m,h}, a_{m,h}) \leq c_{m,h} + (1-q)\big\langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V^{(l)}}(s_{m,h}, a_{m,h}) \big\rangle + (1-q)\frac{1}{t_{j_m}}. \qquad \text{(F.2)}$$

According to inequality (F.1) , we have

$$\begin{aligned}
&c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h}) \\
&\leq c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - \big[c_{m,h} + (1-q) \cdot \big\langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle - (1-q)\frac{1}{t_{j_m}}\big] \\
&\leq \big\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle + q \cdot \big\langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle + \frac{1-q}{t_{j_m}} \\
&\leq \big\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle + \frac{B_* + 1 - q}{t_{j_m}} \\
&\leq \big\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle + \frac{B_* + 1}{t_{j_m}},
\end{aligned}$$

where the second inequality holds due to the definition $\mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) = \big\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle$ and the result $V_{j_m} \leq V^* \leq B_*$ proved in Lemma 6.1 and the third inequality holds because of our choice of parameter q in Algorithm 1.

We also bound the negative part, so we can get an upper bound of its absolute value, which is

$$\begin{aligned}
&- \big[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})\big] \\
&\leq -c_{m,h} - \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) + \big[c_{m,h} + (1-q) \cdot \big\langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle + (1-q)\frac{1}{t_{j_m}}\big] \\
&\leq \big\langle \boldsymbol{\theta}_{m,h} - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle - q \cdot \big\langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle + \frac{1-q}{t_{j_m}} \\
&\leq \big\langle \boldsymbol{\theta}_{m,h} - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle + \frac{B_* + 1 - q}{t_{j_m}} \\
&\leq \big\langle \boldsymbol{\theta}_{m,h} - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle + \frac{B_* + 1}{t_{j_m}},
\end{aligned}$$

where the first inequality holds because of (F.2), the second inequality is due to the definition $\mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) = \big\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle$ and $V_{j_m} \leq V^* \leq B_*$ by Lemma 6.1 and the third inequality holds because of our choice of parameter q in algorithm 1. Thus we can the upper bound of the absolute value:

$$\big|c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})\big| \leq \big|\big\langle \boldsymbol{\theta}_{m,h} - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \big\rangle\big| + \frac{B_* + 1}{t_{j_m}}. \qquad \text{(F.3)}$$

To bound the first item of the right-hand side of (F.3), we have

$$\begin{aligned}\Big|\big\langle \boldsymbol{\theta}_{m,h}-\boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\rangle\Big| &\le \Big(\big\|\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{j_m}\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}}+\big\|\boldsymbol{\theta}_{m,h}-\widehat{\boldsymbol{\theta}}_{j_m}\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}}\Big)\big\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-1}}\\ &\le 2\Big(\big\|\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{j_m}\big\|_{\widehat{\boldsymbol{\Sigma}}_{j_m,0}}+\big\|\boldsymbol{\theta}_{m,h}-\widehat{\boldsymbol{\theta}}_{j_m}\big\|_{\widehat{\boldsymbol{\Sigma}}_{j_m,0}}\Big)\big\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-1}}\\ &\le 4\widehat{\beta}_T\big\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-1}}.\end{aligned}$$

The first inequality uses the triangle inequality. The second inequality uses our partition of intervals. Since the condition is not triggered, we have $\det(\widetilde{\boldsymbol{\Sigma}}_{t,0}) \le 2\det(\widehat{\boldsymbol{\Sigma}}_{j_m,0})$. The third inequality holds because the high-probability event in 6.1 shows $\boldsymbol{\theta}_{m,h}$ lies in $\widehat{\mathcal{C}}_{j_m,0}$ and Lemma 6.1.

Meanwhile, due to the fact that $0 \le V_{j_m} \le B_*$,

$$\Big|\big\langle \boldsymbol{\theta}_{m,h}-\boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\rangle\Big| \le B_*.$$

Combining these two upper bounds, we have

$$\Big|\big\langle \boldsymbol{\theta}_{m,h}-\boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\rangle\Big| \le \min\Big\{B_*, 4\widehat{\beta}_T\big\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-1}}\Big\}. \tag{F.4}$$

Combining all of these, we can finish the proof of lemma D.1, which is

$$\begin{aligned}&\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\big[c_{m,h}+\mathbb{P}V_{j_m}(s_{m,h},a_{m,h})-V_{j_m}(s_{m,h})\big]\\ &\le \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\Big|\big\langle \boldsymbol{\theta}_{m,h}-\boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\rangle\Big| + \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\frac{B_*+1}{t_{j_m}}\\ &\le \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\Big\{B_*, 4\widehat{\beta}_T\big\|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h},a_{m,h})\big\|_{\widetilde{\boldsymbol{\Sigma}}_{t,0}^{-1}}\Big\} + \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\frac{B_*+1}{t_{j_m}},\end{aligned}$$

where the first inequality is from (F.3) and the second is from (F.4). □

G. Proof of Other Technical Lemmas

Proof of Lemma 6.2. The regret can be written as

$$\begin{aligned}R(M) &\le \sum_{m=1}^{M}\sum_{h=1}^{H_m}c_{m,h} - \sum_{m\in\mathcal{M}(M)}V_{j_m}(s_{\text{init}})+1\\ &= \sum_{m=1}^{M}\sum_{h=1}^{H_m}c_{m,h} + \sum_{m=1}^{M}\sum_{h=1}^{H_m}\big[V_{j_m}(s_{m,h+1})-V_{j_m}(s_{m,h})\big]\\ &\quad + \sum_{m=1}^{M}\sum_{h=1}^{H_m}\big[V_{j_m}(s_{m,h})-V_{j_m}(s_{m,h+1})\big] - \sum_{m\in\mathcal{M}(m)}V_{j_m}(s_{\text{init}})+1\\ &= \sum_{m=1}^{M}\sum_{h=1}^{H_m}\big[c_{m,h}+\mathbb{P}V_{j_m}(s_{m,h},a_{m,h})-V_{j_m}(s_{m,h})\big]\\ &\quad + \sum_{m=1}^{M}\sum_{h=1}^{H_m}\big[V_{j_m}(s_{m,h+1})-\mathbb{P}V_{j_m}(s_{m,h},a_{m,h})\big]\end{aligned}$$

$$+\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j_m}(s_{m,h})-V_{j_m}(s_{m,h+1})\right]-\sum_{m\in\mathcal{M}(M)}V_{j_m}(s_{\text{init}})+1,$$

where the inequality holds due to the optimism of V_{j_m},i.e $V_{j_m}(s)\leq V^*(s),\forall s\in\mathcal{S}$ under the event of Lemma 6.1. □

Proof of Lemma D.2. We divide the calls to DEVI into two parts J_1 and J_2, where J_1 is the total number of times that the determinant is doubled and J_2 is the total number of times that the time step is doubled. We divide J_1 into $J_1=\sum_{l=0}^{L-1}J_{1,l}$, where $J_{1,l}$ is the total number of times that the determinant of moment order l is doubled. For $\forall l\in[L]$, we then give the bound of $J_{1,l}$.

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\Sigma}}_{T,l}\right\|_2&=\left\|\lambda\mathbf{I}+\sum_{t=1}^{T}\bar{\sigma}_{t,l}^{-2}\boldsymbol{\phi}_{V_j^{2^l}}(s_t,a_t)\,\boldsymbol{\phi}_{V_j^{2^l}}(s_t,a_t)^\top\right\|_2\\&\leq\lambda+\sum_{t=1}^{T}\left\|\bar{\sigma}_{t,l}^{-1}\boldsymbol{\phi}_{V_j^{2^l}}(s_t,a_t)\right\|_2^2\\&\leq\lambda+\frac{T}{\alpha_t^2}\\&=\lambda+T^2,\end{aligned}$$

where the first inequality is by the triangle inequality and the second inequality holds by $0\leq V_j(s)\leq B,\forall j$, under the event of Lemma 6.1 and the definition of $\bar{\sigma}_{t,l}^{-2}$. Following the inequality that $\det\mathbf{A}\leq\|\mathbf{A}\|_2^n$, where $\mathbf{A}$ is any $n\times n$ matrix, we have $\det\left(\widetilde{\boldsymbol{\Sigma}}_{T,l}\right)\leq\left(\lambda+T^2\right)^d$. Furthermore, we have

$$\left(\lambda+T^2\right)^d\geq 2^{J_{1,l}}\cdot\det\left(\boldsymbol{\Sigma}_0\right)=2^{J_{1,l}}\cdot\lambda^d,$$

From the above inequality, we conclude that

$$J_{1,l}\leq 2d\log\left(1+\frac{T^2}{\lambda}\right)\leq 4d\log\left(1+\frac{T}{\lambda}\right).$$

Note that this bound does not depend on l, so we take a summation over all $l\in[L]$ and get the bound of J_1 as

$$J_1\leq 4dL\log\left(1+\frac{T}{\lambda}\right).$$

To bound J_2, note that $t_0=1$ and thus $2^{J_2}\leq T$, which immediately gives $J_2\leq\log_2 T\leq 2\log T$. Altogether we conclude that

$$J=J_1+J_2\leq 4dL\log\left(1+\frac{T}{\lambda}\right)+2\log T.$$

Thus, we complete the proof of Lemma D.2. □

Proof of Lemma D.4. We first consider the first term on the left-hand side. After rearranging the summation, we can find that the following equation holds,

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j_m}(s_{m,h})-V_{j_m}(s_{m,h+1})\right]\\&=\sum_{m=1}^{M}V_{j_m}(s_{m,1})-V_{j_m}(s_{m,H_m+1})\\&=\sum_{m=1}^{M-1}\left(V_{j_{m+1}}(s_{m+1,1})-V_{j_m}(s_{m,H_m+1})\right)+\sum_{m=1}^{M-1}\left(V_{j_m}(s_{m,1})-V_{j_{m+1}}(s_{m+1,1})\right)\end{aligned}$$

$$+ V_{j_M}\left(s_{M,1}\right) - V_{j_M}\left(s_{M,H_M+1}\right).$$

Using the telescope argument to the second term in the above equation, we have the following inequality,

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j_m}\left(s_{m,h}\right) - V_{j_m}\left(s_{m,h+1}\right)\right] \\
&= \sum_{m=1}^{M-1}\left(V_{j_{m+1}}\left(s_{m+1,1}\right) - V_{j_m}\left(s_{m,H_m+1}\right)\right) + V_{j_1}\left(s_{1,1}\right) - V_{j_M}\left(s_{M,1}\right) \\
&\qquad + V_{j_M}\left(s_{M,1}\right) - V_{j_M}\left(s_{M,H_M+1}\right) \\
&= \sum_{m=1}^{M-1}\left(V_{j_{m+1}}\left(s_{m+1,1}\right) - V_{j_m}\left(s_{m,H_m+1}\right)\right) + V_{j_1}\left(s_{1,1}\right) - V_{j_M}\left(s_{M,H_M+1}\right) \\
&\le \sum_{m=1}^{M-1}\left(V_{j_{m+1}}\left(s_{m+1,1}\right) - V_{j_m}\left(s_{m,H_m+1}\right)\right) + V_{j_1}\left(s_{1,1}\right),
\end{aligned}$$

where the last inequality holds because $V_j(\cdot)$ is non-negative for all $j \in \mathbb{N}$ from the design of DEVI algorithm.

We now consider the term $V_{j_{m+1}}\left(s_{m+1,1}\right) - V_{j_m}\left(s_{m,H_m+1}\right)$. Note that by the interval decomposition, interval m ends if and only if the updating criterion of the DEVI algorithm is met or the goal state is reached. In addition, if interval m ends because the goal state is reached, then we have

$$V_{j_{m+1}}\left(s_{m+1,1}\right) - V_{j_m}\left(s_{m,H_m+1}\right) = V_{j_{m+1}}\left(s_{\text{init}}\right) - V_{j_m}(g) = V_{j_{m+1}}\left(s_{\text{init}}\right).$$

If it ends because the updating criterion of the DEVI algorithm is triggered, then the value function is updated by DEVI and $j_m \neq j_{m+1}$. In such case, we simply apply the trivial upper bound $V_{j_{m+1}}\left(s_{m+1,1}\right) - V_{j_m}\left(s_{m,H_m+1}\right) \le \max_j \left\|V_j\right\|_\infty$. According to Lemma D.2, this happens at most $J \le 4dL\log\left(1 + T/\lambda\right) + 2\log T$ times. Therefore, we can further bound the term $\sum_{m=1}^{M-1}\left(V_{j_{m+1}}\left(s_{m+1,1}\right) - V_{j_m}\left(s_{m,H_m+1}\right)\right)$ as

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left(V_{j_m}\left(s_{m,h}\right) - V_{j_m}\left(s_{m,h+1}\right)\right) \\
&\le \sum_{m=1}^{M-1} V_{j_{m+1}}\left(s_{\text{init}}\right)\cdot \mathbb{1}\{m+1 \in \mathcal{M}(M)\} + V_{j_1}\left(s_{1,1}\right) + \left[4dL\log\left(1 + \frac{T}{\lambda}\right) + 2\log T\right]\cdot \max_j \left\|V_j\right\|_\infty \\
&\le \sum_{m\in\mathcal{M}(M)} V_{j_m}\left(s_{\text{init}}\right) + V_0\left(s_{\text{init}}\right) + 4dB_*L\log\left(1 + \frac{T}{\lambda}\right) + 2B_*\log T \\
&\le \sum_{m\in\mathcal{M}(M)} V_{j_m}\left(s_{\text{init}}\right) + 1 + 4dB_*\log\left(1 + \frac{T}{\lambda}\right) + 2B_*\log T,
\end{aligned}$$

where the second inequality holds due to $\left\|V_j\right\|_\infty \le B_*$, under the event of Lemma 6.1, and the last inequality holds because of our initialization $\left\|V_0\right\|_\infty \le 1$. Thus, we complete the proof of Lemma D.4. □

Proof of Lemma D.5. For any level l, applying Lemma H.2 with $\mathbf{x}_t = \boldsymbol{\phi}_{V_j^{2^l}}(s_t, a_t)/B^{2^l}$, $\sigma_t^2 = \Big([\bar{\mathbb{V}}_{t,l}V_{j+1}^{2^l}](s_t, a_t)/B^{2^{l+1}} + E_{t,l}\Big)$, $\alpha_t'^2 = \alpha_t^2$ and $\gamma'^2 = \gamma^2$, where α_t,γ are parameters used to construct the weights $\bar{\sigma}_{t,l}$ in Algorithm 3, $\beta_t = \widehat{\beta}_T$ is defined in (5.1), we have the following results,

$$\begin{aligned}
R_l &= \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\min\left\{1, \widehat{\beta}_T\left\|\boldsymbol{\phi}_{V_{j_m}^{2^l}}\left(s_{m,h}, a_{m,h}\right)/B^{2^l}\right\|_{\widetilde{\boldsymbol{\Sigma}}_{t,l}^{-1}}\right\} \\
&\le 2d\iota + 2\widehat{\beta}_T\gamma^2 d\iota + 2\sqrt{d\iota}\widehat{\beta}_T\sqrt{\sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\alpha_{t(m,h)}^2 + \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m}\sigma_{t,l}^2}.
\end{aligned}$$

Thus, we complete the proof of Lemma D.5. □

Proof of Lemma D.6. Use the definition of variance $\sigma_{t,l}$, we have the following inequality,

$$
\begin{aligned}
\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\sigma_{t,l}^2 &= \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m} E_{t,l} + \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}[\bar{\mathbb{V}}_t V_{j_m}^{2^l}](s_{m,h}, a_{m,h})/B^{2^{l+1}} \\
&= 2\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m} E_{t,l} + \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\mathbb{V}V_{j_m}^{2^l}(s_{m,h}, a_{m,h})/B^{2^{l+1}} \\
&\quad + B^{-2^{l+1}}\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\big([\bar{\mathbb{V}}_t V_{j_m}^{2^l}](s_{m,h}, a_{m,h}) - B^{2^{l+1}}E_{t,l} - \mathbb{V}V_{j_m}^{2^l}(s_{m,h}, a_{m,h})\big) \\
&\le 2\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m} E_{t,l} + \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\mathbb{V}V_{j_m}^{2^l}(s_{m,h}, a_{m,h})/B^{2^{l+1}},
\end{aligned}
$$

where the last inequality holds since $\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\big([\bar{\mathbb{V}}_t V_{j_m}^{2^l}](s_{m,h}, a_{m,h}) - B^{2^{l+1}}E_{t,l} - \mathbb{V}V_{j_m}^{2^l}(s_{m,h}, a_{m,h})\big) \le 0$ under the event of Lemma 6.1. □

Proof of Lemma D.7. According the definition of S_l and the convexity of square function x^2, we can rearrange the summation in E_l into three terms and obtain the following inequality,

$$
\begin{aligned}
S_l &= \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\Big[\mathbb{P}V_{j_m}^{2^{l+1}}(s_{m,h}, a_{m,h})/B^{2^{l+1}} - \big(\mathbb{P}V_{j_m}^{2^l}(s_{m,h}, a_{m,h})/B^{2^l}\big)^2\Big] \\
&\le \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\Big[\mathbb{P}V_{j_m}^{2^{l+1}}(s_{m,h}, a_{m,h})/B^{2^{l+1}} - \big(\mathbb{P}V_{j_m}(s_{m,h}, a_{m,h})/B\big)^{2^{l+1}}\Big] \\
&= \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\Big[\mathbb{P}V_{j_m}^{2^{l+1}}(s_{m,h}, a_{m,h})/B^{2^{l+1}} - V_{j_m}^{2^{l+1}}(s_{m,h+1})/B^{2^{l+1}}\Big] \\
&\quad + \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\Big[V_{j_m}^{2^{l+1}}(s_{m,h})/B^{2^{l+1}} - \big(\mathbb{P}V_{j_m}(s_{m,h}, a_{m,h})/B\big)^{2^{l+1}}\Big] \\
&\quad + \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\Big[V_{j_m}^{2^{l+1}}(s_{m,h+1})/B^{2^{l+1}} - V_{j_m}^{2^{l+1}}(s_{m,h})/B^{2^{l+1}}\Big].
\end{aligned}
$$

Recalling the definition of A_l, the first term is simply equal to A_{l+1}. For the third term, we can use the telescope argument and get

$$
\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\Big[V_{j_m}^{2^{l+1}}(s_{m,h+1})/B^{2^{l+1}} - V_{j_m}^{2^{l+1}}(s_{m,h})/B^{2^{l+1}}\Big] \le |\mathcal{M}_0|.
$$

For the second term, we can inductively degrade the exponents of the value function. Combining the bound of these three terms, we have the following inequality

$$
\begin{aligned}
S_l &= A_{l+1} + \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\Big[V_{j_m}^{2^{l+1}}(s_{m,h})/B^{2^{l+1}} - \big(\mathbb{P}V_{j_m}(s_{m,h+1})/B\big)^{2^{l+1}}\Big] \\
&\quad + \sum_{m\in\mathcal{M}_0}\Big[V_{j_m}^{2^{l+1}}(s_{m,H_m+1})/B^{2^{l+1}} - V_{j_m}^{2^{l+1}}(s_{m,1})/B^{2^{l+1}}\Big] \\
&\le A_{l+1} + |\mathcal{M}_0| + \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\Big[V_{j_m}^{2^l}(s_{m,h})/B^{2^l} + \big(\mathbb{P}V_{j_m}(s_{m,h+1})/B\big)^{2^l}\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \left[V_{j_m}^{2^l}(s_{m,h})/B^{2^l} - \left(\mathbb{P}V_{j_m}(s_{m,h+1})/B\right)^{2^l}\right] \\
&\le A_{l+1} + |\mathcal{M}_0| + 2 \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\left[V_{j_m}^{2^l}(s_{m,h})/B^{2^l} - \left(\mathbb{P}V_{j_m}(s_{m,h},a_{m,h})/B\right)^{2^l}\right] \\
&\le \dots \\
&\le A_{l+1} + |\mathcal{M}_0| + 2^{l+1} \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\left[V_{j_m}(s_{m,h})/B - \mathbb{P}V_{j_m}(s_{m,h},a_{m,h})/B\right],
\end{aligned}
$$

where we use the fact $V_j(s) \le B, \forall j \ge 1$ under the event of Lemma 6.1. Note that the sum is just similar to the term I_1 in Lemma 6.2 when we first decompose the regret. So we have the following inequality,

$$
\begin{aligned}
S_l &\le A_{l+1} + |\mathcal{M}_0| + 2^{l+1}/B \underbrace{\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m} c(s_{m,h},a_{m,h})}_{\le C_M} \\
&\quad + 2^{l+1} \underbrace{\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\left[V_{j_m}(s_{m,h})/B - \mathbb{P}V_{j_m}(s_{m,h},a_{m,h})/B - c(s_{m,h},a_{m,h})/B\right]}_{-I_1/B} \\
&\le A_{l+1} + |\mathcal{M}_0| + \frac{2^{l+1}}{B}C_M + \frac{2^{l+1}}{B}(B_1 + B_2) \\
&\le A_{l+1} + |\mathcal{M}_0| + \frac{2^{l+1}}{B}C_M + \frac{2^{l+1}}{B}(4BR_0 + B_2),
\end{aligned}
$$

where we use Lemma D.1 and the observation $B_1 \le 4BR_0$. □

Proof of Lemma D.8. We follow the proof of Lemma 25 in Zhang et al. (2021a). We use Lemma H.3 for each fixed level l. Let $x_{m,h} = \left[\left[\mathbb{P}V_{j_m}^{2^l}\right](s_{m,h},a_{m,h}) - V_{j_m}^{2^l}(s_{m,h+1})\right]/B^{2^l}$, then we have $\mathbb{E}\left[x_{m,h} \mid \mathcal{G}_{m,h}\right] = 0$ and $\mathbb{E}\left[x_{m,h}^2 \middle| \mathcal{G}_{m,h}\right] = \left[\mathbb{V}V_{j_m}^{2^l}\right](s_{m,h},a_{m,h})/B^{2^{l+1}}$. Therefore, for each level $l \in [L]$, with probability at least $1-\delta$, we have

$$
\begin{aligned}
A_l = \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m} x_{m,h} &\le \sqrt{2\zeta \sum_{m\in\mathcal{M}_0(M)}\sum_{h=1}^{H_m} \mathbb{V}V_{j_m}^{2^l}(s_{m,h},a_{m,h})/B^{2^{l+1}}} + \zeta \\
&= \sqrt{2\zeta S_l} + \zeta,
\end{aligned}
$$

where $\zeta = 4\log\left(2\log(T\log(1/\delta)) + 1)/\delta\right)$. Taking a union bound over $l \in [L]$, we complete the proof of Lemma D.8. □

H. Auxiliary Lemmas

Theorem H.1 (Theorem 4.3 in Zhou & Gu 2022)**.** *Let $\left\{\mathcal{G}_k\right\}_{k=1}^{\infty}$ be a filtration, and $\{\mathbf{x}_k, \eta_k\}_{k\ge1}$ be a stochastic process such that $\mathbf{x}_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma, \lambda, \epsilon > 0, \boldsymbol{\mu}^* \in \mathbb{R}^d$. For $k \ge 1$, let $y_k = \langle\boldsymbol{\mu}^*, \mathbf{x}_k\rangle + \eta_k$ and suppose that $\eta_k, \mathbf{x}_k$ also satisfy*

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

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

$$
\begin{aligned}
\beta_k =& 12\sqrt{\sigma^2 d\log\left(1 + kL^2/(d\lambda)\right)\log\left(32(\log(R/\epsilon)+1)k^2/\delta\right)} \\
&+ 24\log\left(32(\log(R/\epsilon)+1)k^2/\delta\right)\max_{1\le i\le k}\left\{|\eta_i|\min\left\{1, \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\right\}\right\} + 6\log\left(32(\log(R/\epsilon)+1)k^2/\delta\right)\epsilon.
\end{aligned}
$$

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

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

Lemma H.2 (Lemma B.1 in Zhou & Gu 2022)**.** *Let $\{\sigma_t, \beta_t\}_{t\geq 1}$ be a sequence of non-negative numbers, $\alpha'_t > 0$ decreasing, $\gamma' > 0, \{\mathbf{x}_t\}_{t\geq 1} \subset \mathbb{R}^d$ and $\|\mathbf{x}_t\|_2 \leq L$. Let $\{\mathbf{Z}_t\}_{t\geq 1}$ and $\{\bar{\sigma}_t\}_{t\geq 1}$ be recursively defined as follows: $\mathbf{Z}_1 = \lambda \mathbf{I}$*

$$\forall t \geq 1, \bar{\sigma}_t = \max\left\{\sigma_t, \alpha'_t, \gamma' \|\mathbf{x}_t\|_{\mathbf{Z}_t^{-1}}^{1/2}\right\}, \mathbf{Z}_{t+1} = \mathbf{Z}_t + \mathbf{x}_t \mathbf{x}_t^\top / \bar{\sigma}_t^2.$$

Let $\iota = \log\left(1 + TL^2 / \left(d\lambda\alpha_T'^2\right)\right)$. Then we have

$$\sum_{t=1}^{T} \min\left\{1, \beta_t \|\mathbf{x}_t\|_{\mathbf{Z}_t^{-1}}\right\} \leq 2d\iota + 2 \max_t \beta_t \gamma'^2 d\iota + 2\sqrt{d\iota} \sqrt{\sum_t \beta_t^2 \left(\sigma_t^2 + \alpha_t'^2\right)}.$$

Lemma H.3 (Lemma 11 in Zhang et al. 2021b)**.** *Let $M > 0$ be a constant. Let $\{x_i\}_{i=1}^n$ be a stochastic process, $\mathcal{G}_i = \sigma(x_1, \ldots, x_i)$ be the σ-algebra of $x_1, \ldots, x_i$. Suppose $\mathbb{E}[x_i \mid \mathcal{G}_{i-1}] = 0, |x_i| \leq M$ and $\mathbb{E}\left[x_i^2 \middle| \mathcal{G}_{i-1}\right] < \infty$ almost surely. Then, for any $\delta, \epsilon > 0$, we have*

$$\begin{aligned} &\mathbb{P}\left(\left|\sum_{i=1}^{n} x_i\right| \leq 2\sqrt{2\log(1/\delta)\sum_{i=1}^{n}\mathbb{E}\left[x_i^2 \mid \mathcal{G}_{i-1}\right] + 2\sqrt{\log(1/\delta)}\epsilon + 2M\log(1/\delta)}\right) \\ &\quad > 1 - 2\left(\log\left(M^2 n/\epsilon^2\right) + 1\right)\delta. \end{aligned}$$

Lemma H.4. *Let $\lambda_1, \lambda_2, \lambda_3, \lambda_4 > 0$ and $\kappa \geq \max\{\log_2(\lambda_1/\lambda_3), 1\}$. Let $a_1, \ldots, a_n$ be non-negative real numbers such that $a_i \leq \min\left\{\lambda_1, \lambda_2\sqrt{a_i + a_{i+1} + 2^{i+1}\lambda_3} + \lambda_4\right\}$ for any $1 \leq i \leq \kappa$. Let $a_{\kappa+1} = \lambda_1$. Then we have $a_1 \leq 22\lambda_2^2 + 6\lambda_4 + 4\lambda_2\sqrt{2\lambda_3}$.*

Proof of Lemma H.4. Define a new sequence $b_l = a_l/\lambda_3$. Using Lemma H.5 with sequence $\{b_l\}$ and parameters $\lambda_1/\lambda_3, \lambda_2/\lambda_3, 1, \lambda_4/\lambda_3$, we finish the proof of Lemma H.4. □

Lemma H.5 (Lemma 12 in Zhang et al. 2021a)**.** *Let $\lambda_1, \lambda_2, \lambda_4 > 0, \lambda_3 \geq 1$ and $\kappa = \max\{\log_2 \lambda_1, 1\}$. Let $a_1, \ldots, a_n$ be non-negative real numbers such that $a_i \leq \min\left\{\lambda_1, \lambda_2\sqrt{a_i + a_{i+1} + 2^{i+1}\lambda_3} + \lambda_4\right\}$ for any $1 \leq i \leq \kappa$. Let $a_{\kappa+1} = \lambda_1$. Then we have $a_1 \leq 22\lambda_2^2 + 6\lambda_4 + 4\lambda_2\sqrt{2\lambda_3}$.*

Lemma H.6 (Lemma 12 in Abbasi-Yadkori et al. 2011)**.** *Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are two positive definite matrices satisfying $\mathbf{A} \succeq \mathbf{B}$, then for any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_{\mathbf{A}} \leq \|\mathbf{x}\|_{\mathbf{B}} \cdot \sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.*