Learning Stochastic Shortest Path with Linear Function Approximation

Yifei Min [1] **Jiafan He** [2] **Tianhao Wang** [1] **Quanquan Gu** [2]

Abstract

We study the stochastic shortest path (SSP) problem in reinforcement learning with linear function approximation, where the transition kernel is represented as a linear mixture of unknown models. We call this class of SSP problems as linear mixture SSPs. We propose a novel algorithm with Hoeffding-type confidence sets for learning the linear mixture SSP, which can attain an $\widetilde{\mathcal{O}}(dB_\star^{1.5}\sqrt{K/c_{\min}})$ regret. Here K is the number of episodes, d is the dimension of the feature mapping in the mixture model, $B_\star$ bounds the expected cumulative cost of the optimal policy, and $c_{\min} > 0$ is the lower bound of the cost function. Our algorithm also applies to the case when $c_{\min} = 0$, and an $\widetilde{\mathcal{O}}(K^{2/3})$ regret is guaranteed. To the best of our knowledge, this is the first algorithm with a sublinear regret guarantee for learning linear mixture SSP. Moreover, we design a refined Bernstein-type confidence set and propose an improved algorithm, which provably achieves an $\widetilde{\mathcal{O}}(dB_\star\sqrt{K/c_{\min}})$ regret. In complement to the regret upper bounds, we also prove a lower bound of $\Omega(dB_\star\sqrt{K})$. Hence, our improved algorithm matches the lower bound up to a $1/\sqrt{c_{\min}}$ factor and poly-logarithmic factors, achieving a near-optimal regret guarantee.

1. Introduction

The Stochastic Shortest Path (SSP) model refers to a type of reinforcement learning (RL) problems where an agent repeatedly interacts with a stochastic environment and aims to reach some specific goal state while minimizing the cumulative cost. Compared with other popular RL settings such as episodic and infinite-horizon Markov Decision Processes (MDPs), the horizon length in SSP is random, varies

across different policies, and can potentially be infinite because the interaction only stops when arriving at the goal state. Therefore, the SSP model includes both episodic and infinite-horizon MDPs as special cases, and is comparably more general and of broader applicability. In particular, many goal-oriented real-world problems fit better into the SSP model, such as navigation and GO game (Andrychowicz et al., 2017; Nasiriany et al., 2019).

In recent years, there emerges a line of works on developing efficient algorithms and the corresponding analyses for learning SSP. Most of them consider the episodic setting, where the interaction between the agent and the environment proceeds in K episodes (Cohen et al., 2020; Tarbouriech et al., 2020a). For tabular SSP models where the sizes of the action and state space are finite, Cohen et al. (2021) developed a finite-horizon reduction algorithm that achieves the minimax regret $\widetilde{\mathcal{O}}(B_\star\sqrt{SAK})$, where $B_\star$ is the largest expected cost of the optimal policy starting from any state, S is the number of states and A is the number of actions. In a similar setting, Tarbouriech et al. (2021b) proposed the first algorithm that is minimax optimal, parameter-free and horizon-free at the same time. However, the algorithms mentioned above only apply to tabular SSP problems where the state and action space are small. In order to deal with SSP problems with large state and action spaces, function approximation techniques (Yang & Wang, 2019; Jin et al., 2020; Jia et al., 2020; Zhou et al., 2021b; Wang et al., 2020b;a) are needed.

Following the recent line of work on model-based reinforcement learning with linear function approximation (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b), we consider a linear mixture SSP model, which extends the tabular SSP. More specifically, we assume that the transition probability is parametrized by $\mathbb{P}(s'|s,a) = \langle\phi(s'|s,a), \theta^*\rangle$ for all triplet $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. Here we assume that $\phi \in \mathbb{R}^d$ is a known ternary feature mapping, and $\theta^* \in \mathbb{R}^d$ is an *unknown* model parameter vector that needs to be learned. Such a setting has been previously studied for episodic MDPs (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Cai et al., 2020) and infinite-horizon discounted MDPs (Zhou et al., 2021b). Nevertheless, algorithms developed in these works do not apply to SSP since the horizon length is random as mentioned above.

[1]Department of Statistics and Data Science, Yale University, CT 06520, USA [2]Department of Computer Science, University of California, Los Angeles, CA 90095, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

To tackle the challenge of varying horizon length, we propose a model-based optimistic algorithm with linear function approximation, dubbed LEVIS, for learning the linear mixture SSP. At the core of our algorithm are a confidence set of the model parameters and a Damped Extended Value Iteration (DEVI) subroutine for computing the optimistic estimate of the value function, which together guarantee that the algorithm will reach the goal state in every episode. Compared with the EVI subroutine developed for infinite-horizon discounted MDPs (Zhou et al., 2021b), we introduce a shrinking factor $q \approx 1/t$ in our DEVI with t being the cumulative number of time steps, which guarantees the convergence of DEVI. To compensate for the bias introduced by this shrinking factor, our algorithm performs lazy policy update, which is triggered by the doubling of the time interval between two policy updates or the doubling of the determinant of the covariance matrix. With all these algorithmic designs, our algorithm with Hoeffding-type bonus is guaranteed to achieve a $\widetilde{O}(dB_\star^{1.5}\sqrt{K/c_{\min}})$ regret when $c_{\min} > 0$. To the best of our knowledge, this is the first algorithm that enjoys a sublinear regret for linear mixture SSP. Moreover, we provide an improved algorithm LEVIS$^+$ based on Bernstein-type bonus that achieves an $\widetilde{O}(dB_\star\sqrt{K/c_{\min}})$ regret, which nearly matches our regret lower bound $\Omega(dB_\star\sqrt{K})$ in terms of $d, B_\star$ and K. This lower bound is proved via the construction of a hard-to-learn linear mixture SSP instance.

It is worth noting that a recent work by Vial et al. (2021) studied a different linear SSP model that is similar to the linear MDP (Yang & Wang, 2019; Jin et al., 2020), where both the underlying transition probability and the cost function are linear in a known d-dimensional feature mapping $\boldsymbol{\psi} \in \mathbb{R}^d$, i.e., $\mathbb{P}(s'|s,a) = \langle \boldsymbol{\psi}(s,a), \boldsymbol{\mu}(s)\rangle$ and $c(s,a) = \langle \boldsymbol{\psi}(s,a), \boldsymbol{\theta}\rangle$, and $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\theta}$ are unknown. For this model, their proposed algorithms can achieve an $\widetilde{O}(\sqrt{K})$ regret. Their algorithms are either computationally inefficient or under stronger assumptions such as orthonormal feature mappings. Their results are recently improved by Chen et al. (2021b) via a reduction to the finite-horizon MDP, yielding an efficient algorithm with an $\widetilde{O}(\sqrt{K})$ regret, and a computationally inefficient but "horizon free" algorithm. The linear SSP model is different from our linear mixture SSP model, and we refer the readers to Appendix A.1 for a detailed discussion. Chen et al. (2021b) also proposed algorithms for linear mixture SSPs via a reduction to learning finite-horizon linear mixture MDPs.

Our contributions are summarized as follows:

- We propose to study a linear mixture SSP model, and devise a novel and simple algorithm, dubbed **L**ower confidence **E**xtended **V**alue **I**teration for **S**SP (LEVIS), for learning SSP with linear function approximation.

- We prove that LEVIS achieves a regret of order $\widetilde{\mathcal{O}}(dB_\star^{1.5}\sqrt{K/c_{\min}})$ when $c_{\min} > 0$ and the agent has an order-accurate estimate $B \geq B_\star{}^1$. For the general case where $c_{\min} = 0$, our algorithm can achieve an $\widetilde{\mathcal{O}}(K^{2/3})$ regret guarantee by using a cost perturbation trick (Tarbouriech et al., 2021b).

- We further propose an improved version of LEVIS called LEVIS$^+$ using Bernstein-type bonus, and prove that LEVIS$^+$ achieves an $\widetilde{\mathcal{O}}(dB_\star\sqrt{K/c_{\min}})$ regret.

- We prove that for linear mixture SSP, the regret of any learning algorithms is at least $\Omega(dB_\star\sqrt{K})$. Hence, our LEVIS$^+$ algorithm nearly achieves the lower bound.

Notation We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. For any positive integer n, we denote by $[n]$ the set $\{1, \ldots, n\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote by $\|\mathbf{x}\|_1$ the Manhattan norm and denote by $\|\mathbf{x}\|_2$ the Euclidean norm. For a vector $\mathbf{x} \in \mathbb{R}^d$ and matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, we define $\|\mathbf{x}\|_{\boldsymbol{\Sigma}} = \sqrt{\mathbf{x}^\top \boldsymbol{\Sigma} \mathbf{x}}$. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an absolute constant C such that $a_n \leq Cb_n$. We use $\widetilde{O}(\cdot)$ to hide the logarithmic factors.

2. Related Work

Online learning in SSP SSP problems can be dated back to (Bertsekas & Tsitsiklis, 1991; Bertsekas & Yu, 2013; Bertsekas, 2012), but it is until recently that the regret minimization in online learning of SSP has been studied. In the tabular case, Tarbouriech et al. (2020a) proposed the first algorithm achieving an $\widetilde{\mathcal{O}}(D^{3/2}S\sqrt{AK/c_{\min}})$ regret where D is the diameter of SSP[2]. The regret was further improved to $\widetilde{O}(B_\star S\sqrt{AK})$ by Rosenberg et al. (2020); Cohen et al. (2020), with an extra $\sqrt{S}$ factor compared with the $\Omega(B_\star\sqrt{SAK})$ lower bound (Rosenberg et al., 2020). More recently, the $\widetilde{O}(B_\star\sqrt{SAK})$ minimax optimal regret were obtained by Cohen et al. (2021) and Tarbouriech et al. (2020b) independently using different approaches. Specifically, Cohen et al. (2021) reduced SSP to a finite-horizon MDP with a large terminal cost assuming $B_\star$ is known; while Tarbouriech et al. (2021b) avoid such requirement by adaptively estimating $B_\star$ with a doubling trick, together with a value iteration sub-routine ensuring the optimistic estimate of the value function. Our proposed method shares a similar spirit with the latter approach, but for learning SSP with linear function approximation.

The above algorithms are all model-based. Very recently,

---

[1]We say B is an order-accurate estimate of B^*, if there exists some unknown constant $\kappa \geq 1$ such that $B_\star \leq B \leq \kappa B_\star$.

[2]The diameter of an SSP is defined as the longest possible shortest path from any initial state to the goal state.

Chen et al. (2021a) developed the first model-free algorithm for SSP which achieves the minimax optimal regret when the minimum cost among all state-action pairs $c_{\min}$ is strictly positive. Their method is motivated by the `UCB-ADVANTAGE` algorithm (Zhang et al., 2020). Chen et al. (2022) proposed the first policy optimization algorithm for tabular SSP. For other settings of SSP, (Rosenberg & Mansour, 2020; Chen & Luo, 2021; Chen et al., 2021c) studied the case of adversarial costs. Also, the pioneering work by (Bertsekas & Tsitsiklis, 1991) studied the pure planning problem in SSP where the agent has full knowledge of all the model parameters, and is followed by a series of works (Bonet, 2007; Kolobov et al., 2011; Bertsekas & Yu, 2013; Guillot & Stauffer, 2020). On the other hand, Tarbouriech et al. (2021a) studied the sample complexity of SSP assuming the access to a generative model. Jafarnia-Jahromi et al. (2021) proposed the first posterior sampling algorithm for SSP. Multi-goal SSPs have also been studied by Lim & Auer (2012); Tarbouriech et al. (2020b).

Linear function approximation Linear MDP is one of the most widely studied models for RL with linear function approximation, which assumes both the transition probability and reward functions are linear functions of a known feature mapping (Yang & Wang, 2019; Jin et al., 2020). Representative work in this direction include Du et al. (2019); Zanette et al. (2020); Wang et al. (2020a); Fei et al. (2021b); He et al. (2021a), to mention a few.

Another popular model for RL with linear function approximation is the so-called linear mixture MDP/linear kernel MDP (Yang & Wang, 2020; Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Cai et al., 2020; Min et al., 2021; Zhou et al., 2021b;a). For the finite-horizon setting, Jia et al. (2020) proposed a `UCRL-VTR` algorithm that achieves a $\widetilde{\mathcal{O}}(d\sqrt{H^3 T})$ regret bound. Zhou et al. (2021a) further improve the result by proposing a `UCRL-VTR+` algorithm that attains the nearly minimax optimal regret $\widetilde{\mathcal{O}}(dH\sqrt{T})$ based on a novel Bernstein-type concentration inequality. For the discounted infinite horizon setting, Zhou et al. (2021b) proposed a `UCLK` algorithm with an $\mathcal{O}(d\sqrt{T}/(1-\gamma)^2)$ regret, and also give a $\widetilde{\mathcal{O}}(d\sqrt{T}/(1-\gamma)^{1.5})$ lower bound. The lower bound is later matched up to logarithmic factors by the `UCLK+` algorithm (Zhou et al., 2021a). The SSP model studied in this paper can be seen as an extension of linear mixture MDPs.

3. Preliminaries

Stochastic Shortest Path An SSP instance is an MDP $M := \{\mathcal{S}, \mathcal{A}, \mathbb{P}, c, s_{\text{init}}, g\}$, where $\mathcal{S}$ and $\mathcal{A}$ are the finite state space and action space respectively. Here s_{init} denotes the initial state and $g \in \mathcal{S}$ is the goal state. We denote the cost function by $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$, where $c(s, a)$ is the immediate cost of taking action a at state s. The goal

state g incurs zero cost, i.e., $c(g, a) = 0$ for all $a \in \mathcal{A}$. For any $(s', s, a) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $\mathbb{P}(s'|s, a)$ is the probability to transition to s' given the current state s and action a being taken. The goal state g is an absorbing state, i.e., $\mathbb{P}(g|g, a) = 1$ for all action $a \in \mathcal{A}$.

Linear mixture SSP In this work, we assume the transition probability function $\mathbb{P}$ to be a linear mixture of some basis kernels (Modi et al., 2020; Ayoub et al., 2020; Zhou et al., 2021a).

Assumption 3.1. Assume the feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is known and pregiven. There exists an *unknown* vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}^*\|_2 \leq \sqrt{d}$ such that $\mathbb{P}(s'|s, a) = \langle \phi(s'|s, a), \boldsymbol{\theta}^* \rangle$ for any state-action-state triplet $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Moreover, for any bounded function $V : \mathcal{S} \to [0, B]$, it holds that $\|\phi_V(s, a)\|_2 \leq B\sqrt{d}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $\phi_V(s, a) := \sum_{s' \in \mathcal{S}} \phi(s'|s, a)V(s')$.

For simplicity, for any function $V : \mathcal{S} \to \mathbb{R}$, we denote $\mathbb{P}V(s, a) = \sum_{s'} \mathbb{P}(s'|s, a)V(s')$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Therefore, under Assumption 3.1, we have

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

Proper policies A stationary and deterministic policy is a mapping $\pi : \mathcal{S} \to \mathcal{A}$ such that the action $\pi(s)$ is taken given the current state s. We denote by $T^\pi(s)$ the expected time that it takes by following π to reach the goal state g starting from s. We say a policy π is proper if $T^\pi(s) < \infty$ for any $s \in \mathcal{S}$ (otherwise it is improper). We denote by Π_{proper} the set of all stationary, deterministic and proper policies. We assume that Π_{proper} is non-empty, which is the common assumption in previous works on online learning of SSP (Rosenberg et al., 2020; Rosenberg & Mansour, 2020; Cohen et al., 2021; Tarbouriech et al., 2021b; Jafarnia-Jahromi et al., 2021; Chen et al., 2021a).

Assumption 3.2. The set of all stationary, deterministic and proper policies is non-empty, i.e., $\Pi_{\text{proper}} \neq \varnothing$.

Remark 3.3. The above assumption is weaker than Assumption 1 in Vial et al. (2021) which requires that all stationary policies are proper.

For any policy π, we define the cost-to-go function (a.k.a., value function) as

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

where $s_{t+1} \sim \mathbb{P}(\cdot|s_t, \pi(s_t))$. $V^\pi(s)$ can possibly be infinite if π is improper. The corresponding action-value

function of policy π is defined as

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

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

where $s_2 \sim \mathbb{P}(\cdot|s_1, a_1)$ and $s_{t+1} \sim \mathbb{P}(\cdot|s_t, \pi(s_t))$ for all $t \geq 2$. Since $c(\cdot, \cdot) \in [0, 1]$, for any proper policy $\pi \in \Pi_{\text{proper}}$, V^π and Q^π are both bounded functions.

Bellman optimality For any function $V : \mathcal{S} \to \mathbb{R}$, we define the optimal Bellman operator $\mathcal{L}$ as

$$\mathcal{L}V(s) := \min_{a \in \mathcal{A}}\{c(s, a) + \mathbb{P}V(s, a)\}. \tag{1}$$

Intuitively speaking, we want to learn the optimal policy $\pi^\star$ such that $V^\star(\cdot) := V^{\pi^\star}(\cdot)$ is the unique solution to the Bellman optimality equation $V = \mathcal{L}V$ and $\pi^\star$ minimizes the value function $V^\pi(s)$ component-wise over all policies. It is known that, in order for such $\pi^\star$ to exist, one sufficient condition is Assumption 3.2 together with an extra condition that any improper policy π has at least one infinite-value state, i.e., for any $\pi \notin \Pi_{\text{proper}}$, there exists some $s \in \mathcal{S}$ s.t. $V^\pi(s) = +\infty$ (Bertsekas & Tsitsiklis, 1991; Bertsekas & Yu, 2013; Tarbouriech et al., 2021b). Note that this additional condition is satisfied in the case of strictly positive cost , where for any state $s \neq g$ and $a \in \mathcal{A}$, it holds that $c(s, a) \geq c_{\min}$. To deal with the case of general cost function, one can adopt the cost perturbation trick (Tarbouriech et al., 2021b) and consider a modified problem with cost function $c_\rho(s, a) := \max\{c(s, a), \rho\}$ for some $\rho > 0$. This will introduce an additional cost of order $\mathcal{O}(\rho T)$ to the regret of the original problem, where T is the total number of steps. Therefore, the second condition can be avoided, and we can assume the existence of $\pi^\star$.

Throughout the paper, we denote by $B_\star$ the upper bound of the optimal value function $V^\star$, i.e., $B_\star := \max_{s \in \mathcal{S}} V^\star(s)$. Also, we define $T_\star := \max_{s \in \mathcal{S}} T^{\pi^\star}(s)$, which is finite under Assumption 3.2. Since the cost is bounded by 1, we have $B_\star \leq T_\star < +\infty$. Without loss of generality, we assume that $B_\star \geq 1$. Furthermore, we denote the corresponding optimal action-value function by $Q^\star := Q^{\pi^\star}$ which satisfies the following Bellman equation for all $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$Q^\star(s, a) = c(s, a) + \mathbb{P}V^\star(s, a),$$
$$V^\star(s) = \min_{a \in \mathcal{A}} Q^\star(s, a). \tag{2}$$

Learning objective Under Assumption 3.1, we assume c to be known for the ease of presentation. We study the episodic setting where each episode starts from a fixed initial state s_{init} and ends only if the agent reaches the goal state g. Given the total number of episodes, K, the objective of the

---

Algorithm 1 LEVIS

1: **Input:** regularization parameter λ, confidence radius $\{\beta_t\}$, cost perturbation $\rho \in [0, 1]$, an estimate $B \geq B_\star$
2: **Initialize:** set $t \leftarrow 1$, $j \leftarrow 0$, $t_0 = 0$, $\mathbf{\Sigma}_0 \leftarrow \lambda\mathbf{I}$, $\mathbf{b}_0 \leftarrow \mathbf{0}$, $Q_0(s, \cdot), V_0(s) \leftarrow 1 \,\forall s \neq g$ and 0 otherwise
3: **for** $k = 1, \ldots, K$ **do**
4:     Set $s_t = s_{\text{init}}$
5:     **while** $s_t \neq g$ **do**
6:         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 $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$
7:         Set $\mathbf{\Sigma}_t \leftarrow \mathbf{\Sigma}_{t-1} + \boldsymbol{\phi}_{V_j}(s_t, a_t)\boldsymbol{\phi}_{V_j}(s_t, a_t)^\top$
8:         Set $\mathbf{b}_t \leftarrow \mathbf{b}_{t-1} + \boldsymbol{\phi}_{V_j}(s_t, a_t)V_j(s_{t+1})$
9:         **if** $\det(\mathbf{\Sigma}_t) \geq 2\det(\mathbf{\Sigma}_{t_j})$ **or** $t \geq 2t_j$ **then**
10:             Set $j \leftarrow j + 1$
11:             Set $t_j \leftarrow t$, $\epsilon_j \leftarrow \frac{1}{t_j}$
12:             Set $\widehat{\boldsymbol{\theta}}_j \leftarrow \mathbf{\Sigma}_t^{-1}\mathbf{b}_t$
13:             Set $\mathcal{C}_j \leftarrow \left\{\boldsymbol{\theta} : \|\mathbf{\Sigma}_{t_j}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_j)\|_2 \leq \beta_{t_j}\right\}$
14:             Set $Q_j(\cdot, \cdot) \leftarrow \text{DEVI}(\mathcal{C}_j, \epsilon_j, \frac{1}{t_j}, \rho)$
15:             Set $V_j(\cdot) \leftarrow \min_{a \in \mathcal{A}} Q_j(\cdot, a)$
16:         **end if**
17:         Set $t \leftarrow t + 1$
18:     **end while**
19: **end for**

---

agent is to minimize the regret over K episodes defined as

$$R_K := \sum_{k=1}^{K} \sum_{i=1}^{I_k} c_{k,i} - K \cdot V^\star(s_{\text{init}}), \tag{3}$$

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. Note that R_K might be infinite if some episode never ends.

4. An Algorithm with Hoeffding-type Bonus

In this section, we propose a model-based algorithm for learning linear mixture SSPs, which is displayed in Algorithm 1. LEVIS is inspired by the UCLK-type of algorithms originally designed for discounted linear mixture MDPs (Zhou et al., 2021a;b). Our algorithm takes a multi-epoch form, where each episode is divided into epochs of different lengths (Jaksch et al., 2010; Lattimore & Hutter, 2012). Within each epoch, the agent executes the greedy policy induced by an optimistic estimator of the optimal Q-function. The switch between any two epochs is triggered by a doubling criterion, and then the estimated Q-function is updated through a Discounted Extend Value Iteration (DEVI) subroutine (Algorithm 2). We now give a detailed description of Algorithm 1.

In Algorithm 1, we maintain two global indices. Index t

Algorithm 2 `DEVI`

1: **Input:** confidence set $\mathcal{C}$, error parameter ϵ, transition bonus q, cost perturbation $\rho \in [0, 1]$
2: **Initialize:** $i \leftarrow 0$, and $Q^{(0)}(\cdot, \cdot)$, $V^{(0)}(\cdot) = 0$, and $V^{(-1)}(\cdot) = +\infty$
3: Set $Q(\cdot, \cdot) \leftarrow Q^{(0)}(\cdot, \cdot)$
4: **if** $\mathcal{C} \cap \mathcal{B} \neq \phi$ **then**
5:      **while** $\|V^{(i)} - V^{(i-1)}\|_\infty \geq \epsilon$ **do**
6:

$$Q^{(i+1)}(\cdot, \cdot) \leftarrow c_\rho(\cdot, \cdot) + (1 - q) \min_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(i)}}(\cdot, \cdot) \rangle \tag{4}$$

$$V^{(i+1)}(\cdot) \leftarrow \min_{a \in \mathcal{A}} Q^{(i+1)}(\cdot, a) \tag{5}$$

7:         Set $i \leftarrow i + 1$
8:      **end while**
9:      $Q(\cdot, \cdot) \leftarrow Q^{(i+1)}(\cdot, \cdot)$
10: **end if**
11: **Output:** $Q(\cdot, \cdot)$

---

represents the total number of steps, and index j tracks the number of calls to the `DEVI` sub-routine, where the output of `DEVI` is an optimistic estimator of the optimal action-value function. Each episode starts from a fixed initial state s_{init} (Line 4), ends when the goal state g is reached (Line 5) and is decomposed into epochs indexed by the global index j. Within epoch j, the agent repeatedly executes the policy induced by the current estimate Q_j (Line 6) and updates $\boldsymbol{\Sigma}_t$ and $\mathbf{b}_t$ (Lines 7 and 8). Each epoch ends when the either criterion in Line 9 is triggered, and the `DEVI` subroutine performs an optimistic planning to update the action-value function estimator (Lines 10 to 15).

4.1. Updating Criteria: Coupling Features with Time

As mentioned before, Algorithm 1 runs in epochs indexed by j, and one epoch ends when either of the two updating criteria is triggered (Line 9). The first updating criterion is satisfied once the determinant of $\boldsymbol{\Sigma}_t$ is at least doubled compared to its determinant at the end of the previous epoch. This is called lazy policy update that has been used in the linear bandits and RL literature (Abbasi-Yadkori et al., 2011; Zhou et al., 2021b; Wang et al., 2021), which reflects the diminishing return of learning the underlying transition. The intuition behind the determinant doubling criterion is that the determinant can be viewed as a surrogate measure of the exploration in the feature space. Thus, one only updates the policy when there is enough exploration being made since last update. Moreover, this update criterion reduces the computational cost as the total number of epochs would be bounded by $\mathcal{O}(\log T)$. Here T is the total number of steps through all K episodes. The doubling visitation criterion used in tabular SSP (Jafarnia-Jahromi et al., 2021;

Tarbouriech et al., 2021b) can be viewed as a special case of this doubling determinant-based criterion.

However, the above criterion alone cannot guarantee finite length for each epoch as we lack the boundedness of $\|\boldsymbol{\phi}_V(\cdot, \cdot)\|$, which holds for tabular SSP naturally since at most $|\mathcal{S}||\mathcal{A}| \max_{s \in \mathcal{S}, a \in \mathcal{A}} n(s, a)$ steps suffice to double $n(s, a)$ for at least one pair (s, a) by the pigeonhole principle. To address this issue, we introduce an extra triggering criterion: $t \geq 2t_j$. It turns out that despite of being extremely simple this criterion endows the algorithm with several nice properties. First, together with the `DEVI` error parameter $\epsilon_j = 1/t_j$, we can bound the cumulative error from value iterations in epoch j by a constant, i.e., $(2t_j - t_j) \cdot \epsilon_j = 1$. Second, it will not increase the total number of epochs since the time step doubling can happen at most $\mathcal{O}(\log T)$ times, which is consistent with the first criterion. These two properties together enable us to bound the total error from value iteration by $\mathcal{O}(\log T)$. Finally, this criterion is fairly easy to implement and has negligible time and space complexity.

In summary, our two updating criteria reflect a fine-grained characterization of the extent of exploration by coupling the feature space and time interval.

4.2. Optimistic Planning: Contraction via Perturbation

The optimism of Algorithm 1 is realized by the construction of the confidence set $\mathcal{C}_j$ (Line 11), which is fed into the `DEVI` subroutine. We now describe the estimation of the Q-function in the `DEVI` sub-routine (Algorithm 2). `DEVI` requires the access to a confidence region $\mathcal{C}_j$ that contains the true model parameter $\boldsymbol{\theta}^*$ with high probability (Line 13). We construct the confidence region $\mathcal{C}_j$ as an ellipsoid centered at $\widehat{\boldsymbol{\theta}}_j$ (Line 12) which can be viewed as an estimate of $\boldsymbol{\theta}^*$. The radius of the confidence region $\mathcal{C}_j$ is specified by a parameter β_t (Line 13). Since not every $\boldsymbol{\theta} \in \mathcal{C}_j$ defines a valid probability transition, we further take the intersection between $\mathcal{C}_j$ and a constraint set $\mathcal{B}$ defined as

$$\mathcal{B} := \Big\{ \boldsymbol{\theta} : \forall (s, a), \langle \boldsymbol{\phi}(\cdot | s, a), \boldsymbol{\theta} \rangle \text{ is a probability}$$

$$\text{distribution and } \langle \boldsymbol{\phi}(s' | g, a), \boldsymbol{\theta} \rangle = \mathbb{1}\{s' = g\} \Big\}.$$

Then $\mathcal{C}_j \cap \mathcal{B}$ is still a confidence region containing the true model parameter $\boldsymbol{\theta}^*$ with high probability since $\boldsymbol{\theta}^* \in \mathcal{B}$. Algorithm 2 requires two additional inputs: the optimality gap ϵ_j and the discount factor q. The use of ϵ_j is standard, while the use of the discount factor is new and the key to ensuring convergence of `DEVI`.

Specifically, (4) in Algorithm 2 repeatedly performs one-step value iteration by applying the Bellman operator to the set $\mathcal{C}_j \cap \mathcal{B}$. This is motivated by the Bellman optimality equation in (2), and uses $\min_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(i)}} \rangle$ as an optimistic estimate for $\mathbb{P}V^*$. However, using this estimate alone can-

not guarantee the convergence of `DEVI` because $\langle \cdot, \phi_{V^{(i)}} \rangle$ is not a contractive map, which holds for free in the discounted setting (Jaksch et al., 2010; Zhou et al., 2021b), but not in SSPs. More specifically, in the `EVI` algorithm for the discounted MDP (e.g., Algorithm 2 in (Zhou et al., 2021b)), there is an intrinsic discount factor $0 < \gamma < 1$, which ensures that the Bellman operator is a contraction. Consequently, the value iteration converges within a finite number of iterations. In contrast, the Bellman equation of SSP does not have such a discount factor. To address this issue, in (4), we introduce a $1 - q$ discount factor to ensure the contraction property. Although this causes an additional bias to the estimated transition probability function, we can alleviate it by choosing q properly. In particular, for each epoch j we set $q = 1/t_j$ (Line 14), and we can show that this bias will only cause an additive term of order $\mathcal{O}(\log T)$ in the final regret bound.

Besides the convergence guarantee, the $1 - q$ discount factor also brings an additional benefit that it biases the estimated transition kernel towards the goal state g, further encouraging optimism. Similar design can also be found in the `VISGO` algorithm proposed by Tarbouriech et al. (2021b). The intuition behind is to guarantee the existence of proper policies under the estimated transition probability function. As a result, the output of the value iteration, which solves $V = \widetilde{\mathcal{L}}V$ approximately for the Bellman operator $\widetilde{\mathcal{L}}$ induced by the estimated transition, can induce a greedy policy that is proper under the estimated transition.

The main computational overhead of `LEVIS` is from `DEVI`, which is quite efficient to implement. We discuss this in Appendix A.2.

5. Regret Bound of `LEVIS`

We present the main theoretical results for Algorithm 1 by giving regret upper bounds for both positive and general cost functions.

5.1. Upper Bound: Positive Cost Functions

We first consider a special case where the cost is strictly positive (except for the goal state g).

Assumption 5.1. We assume there exists an *unknown* constant $c_{\min} \in (0, 1)$ such that $c(s, a) \geq c_{\min}$ for all $s \in \mathcal{S} \setminus \{g\}$ and $a \in \mathcal{A}$.

Let T be the total number of steps in Algorithm 1, then the above assumption allows us to lower bound the total cumulative cost after the K episodes by $c_{\min} \cdot T$. Note that this provides a relation between the deterministic K and the random quantity T. For simplicity, we assume the agent has access to B, an order-accurate estimate of $B_\star$ satisfying $B_\star \leq B \leq \kappa B_\star$ for some unknown constant

$\kappa \geq 1$. Similar assumptions have also been imposed in previous works (Tarbouriech et al., 2021b; Vial et al., 2021).

Theorem 5.2. *Under Assumptions 3.1, 3.2 and 5.1, for any $\delta > 0$, let $\rho = 0$ and $\beta_t = B\sqrt{d \log(4(t^2 + t^3 B^2/\lambda)/\delta)} + \sqrt{\lambda d}$ for all $t \geq 1$, where $B \geq B_\star$ and $\lambda \geq 1$. Then with probability at least $1 - \delta$, the regret of Algorithm 1 satisfies*

$$R_K = \mathcal{O}\left(B^{1.5} d\sqrt{K/c_{\min}} \cdot \log^2\left(\frac{KBd}{c_{\min}\delta}\right) + \frac{B^2 d^2}{c_{\min}} \log^2\left(\frac{KBd}{c_{\min}\delta}\right)\right). \tag{6}$$

If $B = O(B_\star)$, Algorithm 1 attains an $\widetilde{O}(B_\star^{1.5} d\sqrt{K/c_{\min}})$ regret. The dominating term in (6) has a dependency on $1/c_{\min}$. For the tabular SSP, Cohen et al. (2021); Jafarnia-Jahromi et al. (2021); Tarbouriech et al. (2021b) avoid such a dependency by using a more delicate analysis. However, it remains an open question whether a similar result can be achieved for the linear mixture SSP.

Remark 5.3. Set the parameter δ in Theorem 5.2 as $\delta = 1/K$ and define the high probability event Ω as Theorem 5.2 holds. Then, we can obtain the expected regret bound:

$$\mathbb{E}[R_K] \leq \mathbb{E}[R_K|\Omega] \Pr[\Omega] + K \Pr[\bar{\Omega}]$$
$$= \mathcal{O}\left(B^{1.5} d\sqrt{K/c_{\min}} \cdot \log^2\left(\frac{KBd}{c_{\min}}\right) + \frac{B^2 d^2}{c_{\min}} \log^2\left(\frac{KBd}{c_{\min}}\right)\right),$$

which implies an $\widetilde{O}(B_\star^{1.5} d\sqrt{K/c_{\min}})$ expected regret.

The proof of Theorem 5.2 is in Appendix D.4.

5.2. Upper Bound: General Cost Functions

Without Assumption 5.1, an $\widetilde{O}(K^{2/3})$ regret can be achieved by running Algorithm 1 with $\rho = K^{-1/3}$.

Theorem 5.4. *Under Assumptions 3.1 and 3.2, for any $\delta > 0$, let $\rho = K^{-1/3}$ and $\beta_t = B\sqrt{d \log(4(t^2 + t^3 B^2/\lambda)/\delta)} + \sqrt{\lambda d}$ for all $t \geq 1$, where $B \geq B_\star$ and $\lambda \geq 1$. Then with probability at least $1 - \delta$, the regret of Algorithm 1 satisfies*

$$R_K = \mathcal{O}\left(\widetilde{B}^{1.5} dK^{2/3} \cdot \chi + T_\star K^{2/3} + \widetilde{B}^2 d^2 K^{1/3} \cdot \chi\right),$$

where $\widetilde{B} = B + T_\star/K^{1/3}$ and $\chi = \log^2\left((B + T_\star)Kd/\delta\right)$.

In Theorem 5.4, the regret depends on $\widetilde{B}$ instead of B. Note that $\widetilde{B}$ is approximately equal to $B_\star$ when $K = \Omega(T_\star^3)$ and $B = O(B_\star)$. Here $T_\star$ is defined in Section 3 as the maximum expected time it takes for the optimal policy to reach the goal state starting from any state.

The cost perturbation ρ is a common trick to deal with the case of general cost functions in the SSP literature (Tarbouriech et al., 2020a; Cohen et al., 2020; Tarbouriech et al.,

2021b). Similar to Tarbouriech et al. (2020a), the term $c_{\min}^{-1}$ is multiplicative with K in our regret bound given by Theorem 5.2, leading to an $\widetilde{\mathcal{O}}(K^{2/3})$ regret in the case of general cost functions. Similarly, the regret bound for linear SSP in Vial et al. (2021) also has a multiplicative $c_{\min}^{-1}$. Some later work on tabular SSP (Cohen et al., 2020; Tarbouriech et al., 2021b) has shown that it is possible to make the term $c_{\min}^{-1}$ additive and improve the regret to $\widetilde{\mathcal{O}}(K^{1/2})$ for general cost functions. How to get an additive $c_{\min}^{-1}$ term for the linear mixture SSP is an interesting future direction.

For the choice of the other parameters in Algorithm 1, by Theorems 5.2 and 5.4, we can set $\lambda = 1$ in both the positive and general cost cases. It is also not uncommon to assume a known upper bound $B \geq B_\star$ in existing SSP literature (Cohen et al., 2021; Vial et al., 2021). While is possible to deal with unknown B with a doubling trick for tabular SSP (Rosenberg et al., 2020; Tarbouriech et al., 2021b), it remains an open question for the linear SSP setting.

6. An Improved Algorithm with Bernstein-type Bonus

Despite its simple form, the major drawback of Algorithm 1 is that it only uses Hoeffding-type confidence sets, which possibly costs too much exploration as it ignores the variance information. Consequently, in the regret bound in Theorem 5.2, the dependence on $B_\star$ is not optimal. This is comparable to the situation for episodic linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020), where the size of Hoeffding-type confidence sets loosely scales with the horizon length H. It has been show that Bernstein-type bonus can sharpen the dependence on H for both tabular MDPs (Lattimore & Hutter, 2012; Azar et al., 2017; Jin et al., 2018; Zhang & Ji, 2019; Zhang et al., 2020; He et al., 2021b) and linear mixture MDPs (Zhou et al., 2021a; Wu et al., 2022; He et al., 2021b; Zhang et al., 2021). Therefore, to improve over the previous algorithm, we further incorporate the variance information and develop a improved algorithm called LEVIS⁺ (i.e., Algorithm 3).

6.1. The **LEVIS⁺** Algorithm

Our LEVIS⁺ algorithm is presented in Algorithm 3. Compared with the previous LEVIS algorithm, LEVIS⁺ shares the same structure but employs a more complicated estimation procedure which involves estimation of the variance of the value function (Lines 7 to 8). Based on the estimated variance, we then apply weighted ridge regression to obtain estimate $\widehat{\boldsymbol{\theta}}_j$ of the model parameter (Lines 9 to 11), where each data point (s_t, a_t, s_{t+1}) is weighted by $\widehat{\sigma}_t^2$ which upper bounds the conditional variance of $V_t(s_{t+1})$ given (s_t, a_t). In contrast to the (unweighted) ridge regression used in Algorithm 1, here we have a tighter concentration of $\widehat{\boldsymbol{\theta}}_j$

Algorithm 3 LEVIS⁺

1: **Input:** regularization parameter λ, confidence radius $\{\widehat{\beta}_t, \, \check{\beta}_t, \, \widetilde{\beta}_t\}$, cost perturbation $\rho \in [0, 1]$, an estimate $B \geq B_\star$

2: **Initialize:** set $t \leftarrow 1$, $j \leftarrow 0$, $t_0 = 0$, $\boldsymbol{\Sigma}_0 \leftarrow \lambda \mathbf{I}$, $\widetilde{\boldsymbol{\Sigma}}_0 \leftarrow \lambda \mathbf{I}$, $\mathbf{b}_0 \leftarrow \mathbf{0}$, $\widetilde{\mathbf{b}}_0 \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_0 \leftarrow 0$, $\widetilde{\boldsymbol{\theta}}_0 \leftarrow 0$, $Q_0(s, \cdot), V_0(s) \leftarrow 1$ for all $s \neq g$ and 0 otherwise

3: **for** $k = 1, \ldots, K$ **do**

4:     Set $s_t = s_{\text{init}}$

5:     **while** $s_t \neq g$ **do**

6:         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 $s_{t+1} \sim \mathbb{P}(\cdot | s_t, a_t)$

7:         Set $[\widehat{\mathbb{V}}_t V_j](s_t, a_t)$ and E_t by (29)

8:         Set $\widehat{\sigma}_t^2 \leftarrow \max\{B^2/d, [\widehat{\mathbb{V}}_t V_j](s_t, a_t) + E_t\}$

9:         Set $\boldsymbol{\Sigma}_t \leftarrow \boldsymbol{\Sigma}_{t-1} + \widehat{\sigma}_t^{-2} \boldsymbol{\phi}_{V_j}(s_t, a_t) \boldsymbol{\phi}_{V_j}(s_t, a_t)^\top$

10:         Set $\mathbf{b}_t \leftarrow \mathbf{b}_{t-1} + \widehat{\sigma}_t^{-2} \boldsymbol{\phi}_{V_j}(s_t, a_t) V_j(s_{t+1})$

11:         Set $\widehat{\boldsymbol{\theta}}_t \leftarrow \boldsymbol{\Sigma}_t^{-1} \mathbf{b}_t$

12:         Set $\widetilde{\boldsymbol{\Sigma}}_t \leftarrow \widetilde{\boldsymbol{\Sigma}}_{t-1} + \boldsymbol{\phi}_{V_j^2}(s_t, a_t) \boldsymbol{\phi}_{V_j^2}(s_t, a_t)^\top$

13:         Set $\widetilde{\mathbf{b}}_t \leftarrow \widetilde{\mathbf{b}}_{t-1} + \boldsymbol{\phi}_{V_j^2}(s_t, a_t) V_j^2(s_{t+1})$

14:         Set $\widetilde{\boldsymbol{\theta}}_t \leftarrow \widetilde{\boldsymbol{\Sigma}}_t^{-1} \widetilde{\mathbf{b}}_t$

15:         **if** $\det(\boldsymbol{\Sigma}_t) \geq 2 \det(\boldsymbol{\Sigma}_{t_j})$ **or** $t \geq 2t_j$ **then**

16:             Set $j \leftarrow j + 1$

17:             Set $t_j \leftarrow t, \epsilon_j \leftarrow \frac{1}{t_j}, q_j \leftarrow \frac{1}{t_j}$

18:             Set confidence ellipsoid $\widehat{\mathcal{C}}_j \leftarrow \left\{\boldsymbol{\theta} : \|\boldsymbol{\Sigma}_{t_j}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{t_j})\|_2 \leq \widehat{\beta}_{t_j}\right\}$

19:             Set $Q_j(\cdot, \cdot) \leftarrow \texttt{DEVI}(\widehat{\mathcal{C}}_{t_j}, \epsilon_j, q_j, \rho)$

20:             Set $V_j(\cdot) \leftarrow \min_{a \in \mathcal{A}} Q_j(\cdot, a)$

21:         **end if**

22:         Set $t \leftarrow t + 1$

23:     **end while**

24: **end for**

---

around the true parameter $\boldsymbol{\theta}^*$ as long as $\widehat{\sigma}_t^2$ improves upon the crude upper bound B^2, and hence obtain a tighter confidence set. Intuitively, the variance of the state-action pairs can be viewed as a surrogate measure of the data quality. Hence by weighting the data points using their (estimated) variance in the regression, the algorithm is able to learn the model more efficiently.

Due to the space limit, we give a detailed introduction of the algorithm design in Appendix G.1.

6.2. Regret Bound of **LEVIS⁺**

We now introduce the theoretical result for Algorithm 3. The following theorem establishes the regret upper bound of Algorithm 3. Compared to our lower bound in Theorem 7.1, it indicates that Algorithm 3 achieves a near-optimal regret with an appropriate choice of the parameters

$\{\check{\beta}_t, \widetilde{\beta}_t, \widehat{\beta}_t\}_{t \geq 1}$.

Specifically, one should choose $\check{\beta}_t = \widetilde{\mathcal{O}}(d)$, $\widetilde{\beta}_t = \widetilde{\mathcal{O}}(\sqrt{dB^4})$, and $\widehat{\beta}_t = \widetilde{\mathcal{O}}(\sqrt{d})$, where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic terms. The detailed choice of $\{\check{\beta}_t, \widetilde{\beta}_t, \widehat{\beta}_t\}_{t \geq 1}$ is given by (30) in Appendix G.1.

Theorem 6.1. *Under Assumptions 3.1, 3.2 and 5.1, for any $\delta > 0$, let $\rho = 0$, $\lambda = 1/B^2$ and $\{\check{\beta}_t, \widetilde{\beta}_t, \widehat{\beta}_t\}_{t \geq 1}$ be given by (30). Then with probability at least $1 - 7\delta$, for any sufficiently large K, the regret of Algorithm 3 satisfies*

$$R_K = \widetilde{\mathcal{O}}\left(d^2 B + dB\sqrt{K} + \sqrt{d}B^{1.5}\sqrt{\frac{K}{c_{\min}}}\right),$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides a factor polynomial in $\log(KB/(\lambda \delta c_{\min}))$.

Remark 6.2. Compared to the confidence sets $\{\widehat{\mathcal{C}}_j\}_{j \in [J]}$ used in Algorithm 3, the confidence sets $\{\mathcal{C}_j\}_{j \in [J]}$ in Algorithm 1 are too conservative, as they only use the crude upper bound B^2 on the variance of the estimated value functions that appear in the algorithm. In fact, the variance of these functions can be significantly smaller than the crude upper bound. As a result, by using the variance information of the data, $\{\widehat{\mathcal{C}}_j\}_{j \in [J]}$ are tighter confidence sets and still contain the true model parameter with high probability.

Remark 6.3. From Theorem 6.1, we can see that if $d \geq B$ and B is an order-accurate estimate of $B_\star$, i.e., $B = \mathcal{O}(B_\star)$, then the regret can be simplified to $\widetilde{\mathcal{O}}(dB_\star\sqrt{K/c_{\min}})$, matching the lower bound in Theorem 7.1 up to $1/\sqrt{c_{\min}}$ and poly-logarithmic factors. As a comparison, the concurrent result for linear mixture SSPs in Chen et al. (2021b) achieves an $\widetilde{\mathcal{O}}(B_\star\sqrt{dT_\star K} + B_\star d\sqrt{K})$ regret, where $T_\star$ is the expected time it takes the optimal policy to reach the goal state maximized over any initial state. Using the fact that $T_\star \leq B_\star/c_{\min}$ and Theorem 7.1, their bound is near-optimal when $d \geq T_\star$.

7. Lower Bound

We provide a hardness result for learning linear mixture SSPs by proving a lower bound for the expected regret suffered by any deterministic learning algorithm. For this purpose, we construct a class of hard-to-learn SSP instances with $|\mathcal{S}| = 2$ and a large action set, and show that any deterministic algorithm will at least suffer from one instance in the class. The detail is deferred to Appendix E.

Theorem 7.1. *Under Assumption 3.1, suppose $d \geq 2$, $B_\star \geq 2$ and $K > (d-1)^2/2^{12}$. Then for any possibly non-stationary history-dependent policy π, there exists a linear mixture SSP instance with parameter $\boldsymbol{\theta}^*$ such that $\mathbb{E}_{\pi, \boldsymbol{\theta}^*}[R_K] \geq dB_\star\sqrt{K}/1024$.*

Remark 7.2. The expectation in the lower bound is over the trajectory induced by policy π in the SSP instance parameterized by $\boldsymbol{\theta}^*$. Note that we allow the policy π to be

non-stationary and history-dependent. This is equivalent to assuming a deterministic learning algorithm, which is sufficient for establishing a lower bound (Cohen et al., 2020).

Remark 7.3. Our instance for the lower bound can be also adapted to a linear SSP instance (Vial et al., 2021; Chen et al., 2021b), which yields a same $\Omega(dB_\star\sqrt{K})$ lower bound. (See Remark E.1 for a detailed discussion.)

8. Overview of the Proof Technique

We explain the proof of the main results in Section 5. We first present a unifying master theorem that underlies Theorems 5.2 and 5.4, and then discuss the technical difficulty and corresponding proof techniques to deal with it.

8.1. A Master Theorem

The following master theorem serves as a backbone for proving the regret bound for specific classes of cost functions.

Theorem 8.1. *Under Assumptions 3.1 and 3.2, for any $\delta > 0$, let $\rho = 0$ and $\beta_t = B\sqrt{d\log\left(4(t^2 + t^3 B^2/\lambda)/\delta\right)} + \sqrt{\lambda}d$ for all $t \geq 1$, where $B \geq B_\star$ and $\lambda \geq 1$. Then with probability at least $1 - \delta$, the regret of Algorithm 1 satisfies*

$$R_K \leq 6\beta_T\sqrt{dT\log\left(1 + \frac{TB_\star^2}{\lambda}\right)} + 7dB_\star \cdot \iota,$$

where T is the total number of steps, $\iota = \log(T + \frac{T^2 B_\star^2 d}{\lambda})$.

Theorem 8.1 gives an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret upper bound with respect to the total number of steps T. However, for SSP problems, the horizon of each episode is unknown and T can be far greater than K, so we need to translate the dependence on T to K. For positive cost functions, this can be achieved by using Assumption 5.1, which imposes an implicit constraint that the total cost of an episode is bounded by $c_{\min}$ times the length of the episode. On the other hand, for general cost functions, it suffices to pick $\rho = K^{-1/3}$ and we can then apply the same reasoning as in the previous case. Please see Appendix D.4 for detailed calculations.

We discuss the proof of this master theorem below, which is one of our main technical contributions.

8.2. Proof Sketch of Theorem 8.1

Regret Decomposition In our analysis, instead of dealing with the regret in (3) directly, we first implicitly group the time steps into intervals and index them by $m = 1, \ldots, M$ as in Lemma 8.2. The basic idea here is to group all the time steps into disjoint intervals of which the end points are either the end of an episode or the time steps when the DEVI subroutine is called[3]. The purpose of such a

---

[3]The interval decomposition is implicit and only for the purpose of analysis. This is in contrast to the epoch decomposition, which

decomposition is to guarantee that within each interval the optimistic action-value function remains the same and so the induced policy. This has been used in several existing works on SSP (Rosenberg et al., 2020; Rosenberg & Mansour, 2020; Tarbouriech et al., 2021b) as well.

Lemma 8.2. *Assume the event in Lemma 8.3 holds, then the following holds for the regret defined in* (3)[4]*:*

$$R(M)$$
$$\leq \underbrace{\sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h}) \right]}_{E_1}$$
$$+ \underbrace{\sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[V_{j_m}(s_{m,h+1}) - \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) \right]}_{E_2}$$
$$+ \underbrace{\left[\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right) - \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}}) \right]}_{E_3} + 1, \quad (7)$$

where $j_m = j$ is the index of the value function estimate V_j used in the m-th interval[5] and H_m is the length of the m-th interval.

Bounding E_1. Controlling term E_1 is the essential and most difficult part. Roughly speaking, E_1 is the accumulated Bellman error of the DEVI outputs $Q_j(\cdot, \cdot)$ on the sample state-action trajectory. The ordinary method is to bound the sum of width of the confidence regions C_j. However, we face unique challenges. Recall from (4) that our value iteration sub-routine DEVI includes a transition bonus q for the sake of convergence. Such bonus leads to a biased Bellman operator, and the biases would accumulate over time and become an additive term in the regret bound of the order $\mathcal{O}(\sum_m q_{j_m} H_m)$. To bound the bias, we need to bound H_m and choose an appropriate q_{j_m}.

For H_m, we have $H_m = \mathcal{O}(t_{j_m+1} - t_{j_m})$, where $t_{j_m+1} - t_{j_m}$ is the length of epoch j, i.e., the number of time steps between two consecutive DEVI calls. As mentioned in Section 4.1, while the classical determinant-based criterion alone cannot guarantee finite epoch length, the additional time step doubling criterion fixes this issue by enforcing $t_{j_m+1} \leq 2t_{j_m}$. Furthermore, by picking each $q_j = \frac{1}{t_j}$, the total bias can be upper bounded as

$$\mathcal{O}\left(\sum_{j=1}^{J} t_j^{-1} \cdot (2t_j - t_j) \right) = \mathcal{O}(J).$$

is explicit and indexed by j in Algorithm 1. The difference is that an epoch ends when DEVI is called, while an interval ends when either DEVI is called or the goal state g is reached (i.e., an episode ends).

[4]$R(M)$ is the same as R_K. We use a different notation to emphasize the interval decomposition.

[5]This is well-defined since the same V_j is used at all time steps within one interval.

In this way, it suffices to bound the total number of calls to DEVI to bound the accumulative bias.

For the rest part of E_1, we can show that every time when DEVI is called, the output is an optimistic action-value function estimator with high probability (by Lemma 8.3). Finally, we need to bound the total difference between the estimated functions and the optimal action-value function. This follows from the elliptical potential lemma and the determinant-based doubling criterion. The details of bounding E_1 is deferred to Appendix D.1.

Bounding E_2 and E_3. Since E_2 is the sum of a martingale difference sequence, it can be bounded by $\mathcal{O}(\sqrt{T \log(T/\delta)})$ using standard martingale concentration inequality.

The term E_3 can be transformed into a telescoping sum under the interval decomposition. After the transformation, it can be shown that only those terms with $j_m \neq j_{m+1}$ would contribute to the final regret. Since the number of intervals with $j_m \neq j_{m+1}$ are at most $\mathcal{O}(J)$, the problem again reduces to bounding the number of DEVI calls.

Analysis of DEVI. By the algorithmic design we elaborated in Section 4, DEVI guarantees optimism and finite-time convergence, as summarized in Lemma 8.3 below.

Lemma 8.3. *For all $t \geq 1$, let $\rho = 0$ and $\beta_t = B\sqrt{d \log(4(t^2 + t^3 B^2/\lambda)/\delta)} + \sqrt{\lambda d}$, where $B \geq B_\star$. Then with probability at least $1 - \delta/2$, for all $j \geq 1$, DEVI converges in finite time, and it holds that $\theta^* \in \mathcal{C}_j \cap \mathcal{B}, 0 \leq Q_j(\cdot, \cdot) \leq Q^\star(\cdot, \cdot)$, and $0 \leq V_j(\cdot) \leq V^\star(\cdot)$.*

Note that in Lemma 8.3 the optimism only holds for the DEVI output, i.e., V_j for any $j \geq 1$. The initialization V_0 in Line 2 of the main Algorithm 1 is not necessarily satisfy optimistic since it is possible that $V^\star(s) < 1$ for some s. Still, such an initialization satisfies $\|V_0\|_\infty = 1 \leq B_\star$, which is sufficient to establish the optimism for $j \geq 1$.

9. Conclusions

In this paper, we propose a novel algorithm for linear mixture SSP and prove its regret upper and lower bounds. For future work, there are several important directions. First, there is a $B_\star^{0.5}$ gap between the current upper and lower bounds. Second, it remains open to prove an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bound for linear mixture SSP for general cost functions.

Acknowledgements

References

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

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight experience replay. *arXiv preprint arXiv:1707.01495*, 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.

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

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.

Bonet, B. On the speed of convergence of value iteration on stochastic shortest-path problems. *Mathematics of Operations Research*, 32(2):365–373, 2007.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

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

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Chen, L. and Luo, H. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. *arXiv preprint arXiv:2102.05284*, 2021.

Chen, L., Jafarnia-Jahromi, M., Jain, R., and Luo, H. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *arXiv preprint arXiv:2106.08377*, 2021a.

Chen, L., Jain, R., and Luo, H. Improved no-regret algorithms for stochastic shortest path with linear mdp. *arXiv preprint arXiv:2112.09859*, 2021b.

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, 2021c.

Chen, L., Luo, H., and Rosenberg, A. Policy optimization for stochastic shortest path. *arXiv preprint arXiv:2202.03334*, 2022.

Cohen, A., Kaplan, H., Mansour, Y., and Rosenberg, A. Near-optimal regret bounds for stochastic shortest path. *arXiv preprint arXiv:2002.09869*, 2020.

Cohen, A., Efroni, Y., Mansour, Y., and Rosenberg, A. Minimax regret for stochastic shortest path. *arXiv preprint arXiv:2103.13056*, 2021.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2019.

Fei, Y. and Xu, R. Cascaded gaps: Towards gap-dependent regret for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2203.03110*, 2022.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Fei, Y., Yang, Z., Chen, Y., and Wang, Z. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

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, 2021b.

Greenberg, I., Chow, Y., Ghavamzadeh, M., and Mannor, S. Efficient risk-averse reinforcement learning. *arXiv preprint arXiv:2205.05138*, 2022.

Guillot, M. and Stauffer, G. The stochastic shortest path problem: a polyhedral combinatorics perspective. *European Journal of Operational Research*, 285(1):148–158, 2020.

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, 2021a.

He, J., Zhou, D., and Gu, Q. Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, 34, 2021b.

He, J., Zhou, D., Zhang, T., and Gu, Q. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2205.06811*, 2022.

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.

Jaimungal, S., Pesenti, S. M., Wang, Y. S., and Tatsat, H. Robust risk-aware reinforcement learning. *SIAM Journal on Financial Mathematics*, 13(1):213–226, 2022.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

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., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

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

Kolobov, A., Mausam, M., Weld, D. S., and Geffner, H. Heuristic search for generalized stochastic shortest path mdps. In *Twenty-First International Conference on Automated Planning and Scheduling*, 2011.

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

Lim, S. H. and Auer, P. Autonomous exploration for navigating in mdps. In *Conference on Learning Theory*, pp. 40–1. JMLR Workshop and Conference Proceedings, 2012.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pp. 3242–3245. PMLR, 2021.

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, 2021.

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

Nasiriany, S., Pong, V. H., Lin, S., and Levine, S. Planning with goal-conditioned policies. *arXiv preprint arXiv:1911.08453*, 2019.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

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, 2020a.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. Improved sample complexity for incremental autonomous exploration in mdps. In *NeurIPS*, 2020b.

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. *arXiv preprint arXiv:2104.11186*, 2021b.

Vial, D., Parulekar, A., Shakkottai, S., and Srikant, R. Regret bounds for stochastic shortest path problems with linear function approximation. *arXiv preprint arXiv:2105.01593*, 2021.

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

Wang, T., Zhou, D., and Gu, Q. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *arXiv preprint arXiv:2101.02195*, 2021.

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

Wu, Y., Zhou, D., and Gu, Q. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, 2022.

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

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

Yin, M., Chen, W., Wang, M., and Wang, Y.-X. Offline stochastic shortest path: Learning, evaluation and towards optimality. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

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, X., Chen, Y., Zhu, X., and Sun, W. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5757–5773. PMLR, 2022.

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

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 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, 2021.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. 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. Additional Discussions

A.1. Discussion on the Linear Mixture MDPS

The linear mixture MDP (Modi et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b) is a commonly considered model for linear function approximation, where one assumes the transition probability function $\mathbb{P}$ to be a linear mixture of some basis kernels. The linear mixture MDP covers several important MDP models studied in the literature. We briefly discuss them here.

Example A.1 (Tabular MDPs). For a tabular MDP $M(\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P})$ with $|\mathcal{S}|, |\mathcal{A}| \leq \infty$, the transition probability kernel can be represented by $|\mathcal{S}|^2 |\mathcal{A}|$ *unknown* parameters. The tabular MDP is a special case of linear mixture MDPs with the feature mapping $\phi(s'|s, a) = \mathbf{e}_{(s,a,s')} \in \mathbb{R}^d$ and parameter vector $\boldsymbol{\theta} = [\mathbb{P}(s'|s, a)] \in \mathbb{R}^d$, where $d = |\mathcal{S}|^2 |\mathcal{A}|$ and $\mathbf{e}_{(s,a,s')}$ denotes the corresponding natural basis in the d-dimensional Euclidean space.

Example A.2 (Linear combination of base models, Modi et al. 2020). For an MDP $M(\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P})$, suppose there exist m base transition probability kernels $\{p_i(s'|s, a)\}_{i=1}^m$, a feature mapping $\boldsymbol{\psi}(s, a) : \mathcal{S} \times \mathcal{A} \to \Delta^{d'}$ where $\Delta^{d'}$ is a $(d' - 1)$-dimensional simplex, and an *unknown* matrix $\mathbf{W} \in \mathbb{R}^{m \times d'} \in [0, 1]^{m \times d'}$ such that $\mathbb{P}(s'|s, a) = \sum_{k=1}^m [\mathbf{W}\boldsymbol{\psi}(s, a)]_k p_k(s'|s, a)$. Then it is a special case of linear mixture MDPs with feature mapping $\phi(s'|s, a) = \text{vec}(\mathbf{p}(s'|s, a)\boldsymbol{\psi}(s, a)^\top) \in \mathbb{R}^d$ and parameter vector $\boldsymbol{\theta} = \text{vec}(\mathbf{W}) \in \mathbb{R}^d$ where $d = md'$, $\text{vec}(\cdot)$ is the vectorization operator, and $\mathbf{p}(s'|s, a) = [p_k(s'|s, a)] \in \mathbb{R}^m$.

Example A.3 (linear-factored MDP, Yang & Wang 2019). For an MDP $M(\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P})$, suppose that there exist feature mappings $\boldsymbol{\psi}_1(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_1}$ satisfying $\|\boldsymbol{\psi}_1(s, a)\|_2 \leq \sqrt{d_1}$, $\boldsymbol{\psi}_2(s') : \mathcal{S} \to \mathbb{R}$ satisfying for any $V : \mathcal{S} \to [0, R]$, $\|\sum_s V(s)\boldsymbol{\psi}_2(s)\|_2 \leq R$ and an *unknown* matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ satisfying $\|\mathbf{M}\|_F \leq \sqrt{d_1}$ such that $\mathbb{P}(s'|s, a) = \boldsymbol{\psi}_1(s, a)^\top \mathbf{M} \boldsymbol{\psi}_2(s')$. Then it is a special case of linear mixture MDPs with feature mapping $\phi(s'|s, a) = \text{vec}(\boldsymbol{\psi}_2(s')\boldsymbol{\psi}_1(s, a)^\top) \in \mathbb{R}^d$ and parameter vector $\boldsymbol{\theta} = \text{vec}(\mathbf{M}) \in \mathbb{R}^d$, where $d = d_1 d_2$.

For more discussions, please refer to, for example, Section 2 in Ayoub et al. (2020), or Section 3 in Zhou et al. (2021b).

A.2. Computational Complexity of `DEVI`

In `DEVI`, we need to solve a sequence of optimization problems as given by (4) and (5). The computational complexity of solving (4) dominates that of (5). Fortunately, the objective function in (4) is convex and the constraint set $\mathcal{C} \cap \mathcal{B}$ is a convex set, so we can use projected gradient descent (Boyd et al., 2004) to solve it.

`DEVI` involves a number of iterations (i.e., the while-loop). The total number of iterations is $\mathcal{O}(t \log t)$, where t is the time step at which the `DEVI` is called. To see this, note that according to the stopping criterion $\|V^{(i)} - V^{(i-1)}\|_\infty \geq \epsilon$ and the update rule (4), we solve for n such that $(1 - q)^n \leq \epsilon$. Then since $\epsilon = 1/t$ and $q = 1/t$, we have $n = \mathcal{O}(t \log t)$.

The cost of each `DEVI` iteration is $\mathcal{O}(dRB|\mathcal{A}|)$, where d is the feature dimension, R is the computational cost for calculating ϕ_V, and B is the cost for solving an optimization problem of the form $\min_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \langle \phi_V, \mathbf{x} \rangle$. The optimization problem can be solved by using project gradient descent, as mentioned above. To calculate ϕ_V, Zhou et al. (2021b) proposed using Monte Carlo sampling to estimate ϕ_V with high accuracy, which turns out to be quite efficient. We refer interested readers to Appendix B of Zhou et al. (2021b) for more detail.

A.3. Disscussion on Future Directions

There are many promising future directions for the SSP problem. We discuss a few.

First, our current algorithm design is based on the principle of Optimism-in-Face-of-Uncertainty (OFU). It is possible to develop Thompson Sampling (TS) type algorithms by following the well-known `PSRL` algorithm (Osband et al., 2013). Jafarnia-Jahromi et al. (2021) studied this topic under the tabular case and gave an $\widetilde{\mathcal{O}}(B_\star S \sqrt{AK})$ regret bound. The empirical advantage of TS-type algorithms in this setting is still under-studied. In the bandit setting, Chapelle & Li (2011) showed that TS is more robust to delayed feedback and can outperform OFU-type algorithms. In the reinforcement learning setting, such a comprehensive empirical evaluation does not exist. We believe it is an important future research direction.

Furthermore, most existing works on SSP focus on the online episodic setting. The recent work by Yin et al. (2022) gives the first theoretical result for SSP in the offline setting. They studied both the policy evaluation problem and the policy learning problem. While they considered the tabular case, it is natural to extend offline SSP to the linear and general function approximation cases.

Also, while existing SSP works all focus on the regret minimization setting, it would be very meaningful if we can incorporate other useful factors into this framework. One example is the risk-sensitive RL, where the goal is to consider certain risk of the decisions made and minimize the regret simultaneously (Fei et al., 2020; 2021a; Jaimungal et al., 2022; Fei & Xu, 2022; Greenberg et al., 2022). This problem setting is very helpful for applications where the risk is a crucial factor, such as finance and AI medicine. Another example is the corruption-robust RL, where the goal is to find efficient algorithms in an environment with corrupted reward signals and transitions (e.g., rewards and transitions are picked by an adversary) (Lykouris et al., 2021; He et al., 2022; Zhang et al., 2022). This setting is very useful because data corruption is a big threat against the security of many ML systems.
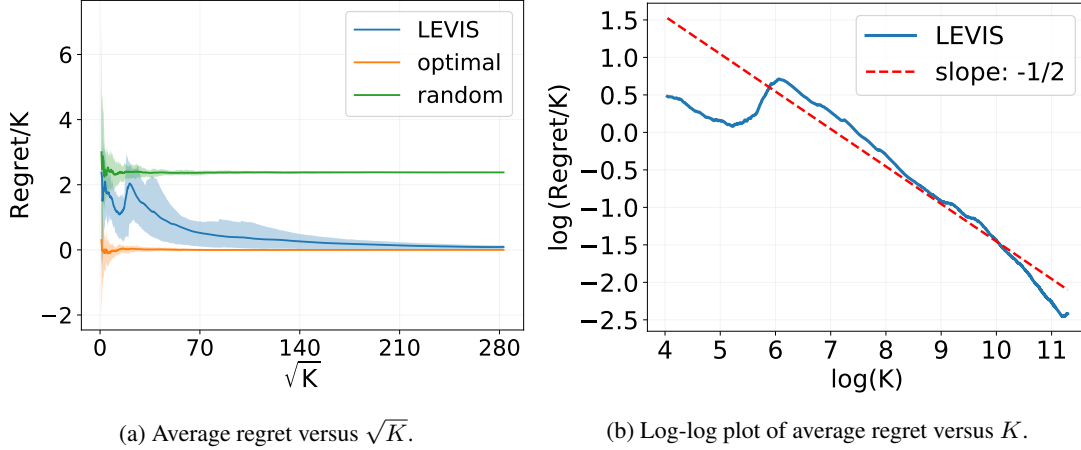
B. Numerical Simulations

(a) Average regret versus $\sqrt{K}$.

(b) Log-log plot of average regret versus K.

Figure 1. The left plot shows the average regret (i.e. R_K/K) of implementing Algorithm 1 on the SSP instance described in Appendix B with $\lambda = 1$, $\rho = 0$ and failing probability 0.01. The curve is the average of 40 trials. Colored areas indicate empirical [10%,90%] confidence intervals. The right plot is the log-log plot of R_K/K and K. The red dotted line has a slope equal to $-1/2$. It is clear that the curve has a slope very close to $-1/2$.

In this section, we present some results from numerical simulations, which corroborate our theory. We construct an SSP instance based on the example used in the proof of the lower bound. Specifically, we have the action space $\mathcal{A} = \{-1, 1\}^{d-1}$ with $|\mathcal{A}| = 2^{d-1}$. The state space is $\mathcal{S} = \{s_{\text{init}}, g\}$. We choose δ, Δ and $B_\star$ such that $\delta + \Delta = 1/B_\star$ and $\delta > \Delta$. The true model parameter $\boldsymbol{\theta}^*$ is given by

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

The feature mapping is defined as

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

Here we use $\mathbf{a}$ instead of a to emphasize that the action is vector-valued. One can verify that this is indeed a linear mixture SSP with the following transition function:

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

for all $\mathbf{a} \in \mathcal{A}$. For more details about this SSP instance, please refer to Appendix E. Note that this is a very hard SSP instance since it is difficult to distinguish between different actions, as we will later show in the proof of the lower bound.

The experimental results are shown in Fig. 1. We compare the performance of LEVIS with that of the optimal policy and the random policy. Here the optimal policy always chooses $\mathbf{a} = \mathbf{1}_{d-1}$ to maximize the probability of reaching g from s_{init} by the construction of the SSP, and the random policy picks $\mathbf{a} \in \mathcal{A}$ uniformly at random. We set $d = 5$ and $B_\star = 3$ in the simulation. In Fig. 1a, we plot the average regret R_K/K versus $\sqrt{K}$. It is evident that LEVIS has a sublinear regret, as opposed to the linear regret of the random policy. To further verify that the cumulative regret R_K indeed grows at an $\widetilde{\mathcal{O}}(\sqrt{K})$ rate, in Fig. 1b, we make the log-log plot of R_K/K and K. The red dotted line has a slope equal to $-1/2$. We see from Fig. 1b that the slope of the curve is very close to $-1/2$. This verifies the $\widetilde{\mathcal{O}}(\sqrt{K})$ regret of LEVIS. These results corroborate our theoretical findings.

C. Proof of Regret Decomposition

In this section, we prove the regret decomposition given by Lemma 8.2.

Proof of Lemma 8.2. We first explain the details of the interval decomposition. The first interval begin at $t = 1$, and an interval ends once either one of the two conditions is met: (1) the DEVI sub-routine is triggered (i.e., either the determinant of the covariance matrix or the time index is doubled); (2) the goal state g is reached, i.e., the current episode ends. We remark that this interval decomposition is only implicit since it is not implemented by the algorithm explicitly. Note that by the two conditions described above, each interval has bounded length almost surely. Indeed, even if the goal state is never reached or the determinant is never doubled due to ϕ_V having small norm, the time step only requires the number of iterations to be doubled.

We index the intervals by $m = 1, 2, \cdots$, and denote by M as the total number of intervals, which is possibly infinite. The length of the m-th interval is denoted by H_m. With a slight abuse of notation, we denote the trajectory for the m-th interval as $(s_{m,1}, a_{m,1}, \cdots, s_{m,H_m}, a_{m,H_m}, s_{m,H_m+1})$, where we have $s_{m,H_m+1} = g$ if interval m ends with condition (2) being met, and $s_{m,H_m+1} = s_{m+1,1}$ otherwise. We denote by $\mathcal{M}(M) \subseteq [M]$ the set of intervals which are the first interval of their corresponding episodes. We define the mapping j_m, such that for each $m \in [M]$, $j_m = j$ is the index of the value function estimate V_j used in the m-th interval.

Now let's see how the regret can be expressed under the interval decomposition introduced above. The regret can be written as

$$
\begin{aligned}
R(M) &\leq \sum_{m=1}^{M} \sum_{h=1}^{H_m} c_{m,h} - \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}}) + 1 \\
&= \sum_{m=1}^{M} \sum_{h=1}^{H_m} c_{m,h} + \sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h+1}) - V_{j_m}(s_{m,h}) \right) \\
&\quad + \sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right) - \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}}) + 1 \\
&= \underbrace{\sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h}) \right]}_{E_1} \\
&\quad + \underbrace{\sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[V_{j_m}(s_{m,h+1}) - \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) \right]}_{E_2} \\
&\quad + \underbrace{\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right) - \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}})}_{E_3} + 1.
\end{aligned}
\tag{8}
$$

The inequality in the above holds because of the optimism of V_j for $j \geq 1$. Here please note that, since V_0 is not the output of DEVI, optimism does not necessarily hold for V_0. Therefore, we simply add 1 at the RHS of the first inequality by the fact that $|V_0| \leq 1$ and the first interval has length equal to 1 according to the time step doubling updating criterion.

$\square$

D. Proof for Upper Bounds

In this section we finish the proof of the key result Theorem 8.1 by bounding the terms in the regret decomposition in Lemma 8.2.

D.1. Bounding E_1

Lemma D.1. *Assume the event of Lemma 8.3 holds. Then we have*

$$\sum_{m=1}^{M} \sum_{h=1}^{H_m} [c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h})]$$

$$\leq 4\beta_T \sqrt{2Td \cdot \log\left(1 + B_\star^2 T/\lambda\right)} + 5dB_\star \left[\log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + \log(T)\right] + 4.$$

Proof of Lemma D.1. By Line 6 and 15 in the algorithm, for any m and h, we have

$$V_{j_m}(s_{m,h}) = \min_{a \in \mathcal{A}} Q_{j_m}(s_{m,h}, a) = Q_{j_m}(s_{m,h}, a_{m,h}).$$

Therefore E_1 can be rewritten as

$$E_1 = \sum_{m=1}^{M} \sum_{h=1}^{H_m} [c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})]. \tag{9}$$

Denote by $\mathcal{M}_0(M)$ the set of m such that $j_m \geq 1$, i.e., $\mathcal{M}_0(M) = \{m \leq M : j_m \geq 1\}$. Then we see that $\mathcal{M}_0(M)$ is the collection of intervals such that Q_{j_m} is the output of DEVI instead of the initialization Q_0. Fix arbitrary $m \in \mathcal{M}_0(M)$ and h. Since Q_{j_m} is the output of DEVI, we have $Q_{j_m} = Q^{(l)}$ for some l, i.e., the l-th iteration in DEVI, and thus $V_{j_m}(\cdot) = \min_{a \in \mathcal{A}} Q^{(l)}(\cdot, a) = V^{(l)}(\cdot)$. By the design of DEVI, we have

$$Q^{(l)}(s_{m,h}, a_{m,h})$$
$$= c_{m,h} + (1 - q) \cdot \min_{\boldsymbol{\theta} \in \mathcal{C}_{j_m} \cap \mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(l-1)}}(s_{m,h}, a_{m,h}) \rangle$$
$$= c_{m,h} + (1 - q) \cdot \langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V^{(l-1)}}(s_{m,h}, a_{m,h}) \rangle$$
$$= c_{m,h} + (1 - q) \cdot \langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V^{(l)}}(s_{m,h}, a_{m,h}) \rangle + (1 - q) \cdot \langle \boldsymbol{\theta}_{m,h}, [\boldsymbol{\phi}_{V^{(l-1)}} - \boldsymbol{\phi}_{V^{(l)}}](s_{m,h}, a_{m,h}) \rangle,$$

where $\boldsymbol{\theta}_{m,h} = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}_j \cap \mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(l-1)}}(s_{m,h}, a_{m,h}) \rangle$ and its existence is guaranteed under the event of Lemma 8.3. Define $\mathbb{P}_{m,h}$ as the transition kernel parametrized by $\boldsymbol{\theta}_{m,h}$, i.e.,

$$\mathbb{P}_{m,h}(\cdot | \cdot, \cdot) = \langle \boldsymbol{\phi}(\cdot | \cdot, \cdot), \boldsymbol{\theta}_{m,h} \rangle.$$

Then from above we have

$$Q^{(l)}(s_{m,h}, a_{m,h})$$
$$= c_{m,h} + (1 - q) \cdot \langle \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V^{(l)}}(s_{m,h}, a_{m,h}) \rangle + (1 - q) \cdot \mathbb{P}_{m,h} \left[V^{(l-1)} - V^{(l)}\right](s_{m,h}, a_{m,h})$$
$$\geq c_{m,h} + (1 - q) \cdot \mathbb{P}_{m,h} V^{(l)}(s_{m,h}, a_{m,h}) - (1 - q) \cdot \frac{1}{t_{j_m}},$$

where the inequality is by the DEVI terminal condition that $\|V^{(l)} - V^{(l-1)}\|_\infty \leq \epsilon_j = 1/t_{j_m}$. Therefore we have

$$Q_{j_m}(s_{m,h}, a_{m,h}) \geq c_{m,h} + (1 - q) \cdot \mathbb{P}_{m,h} V_{j_m}(s_{m,h}, a_{m,h}) - (1 - q) \cdot \frac{1}{t_{j_m}},$$

and it follows that

$$c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})$$

$$\leq \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - (1-q) \cdot \mathbb{P}_{m,h}V_{j_m}(s_{m,h}, a_{m,h}) + (1-q) \cdot \frac{1}{t_{j_m}}$$

$$= [\mathbb{P} - \mathbb{P}_{m,h}]V_{j_m}(s_{m,h}, a_{m,h}) + q\mathbb{P}_{m,h}V_{j_m}(s_{m,h}, a_{m,h}) + (1-q) \cdot \frac{1}{t_{j_m}}$$

$$\leq [\mathbb{P} - \mathbb{P}_{m,h}]V_{j_m}(s_{m,h}, a_{m,h}) + \frac{B_\star}{t_{j_m}} + (1-q) \cdot \frac{1}{t_{j_m}}$$

$$= \langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle + \frac{B_\star + 1 - q}{t_{j_m}},$$

where the second inequality is by the optimism $V_{j_m} \leq V^\star \leq B_\star$ under the event of Lemma 8.3, and $q = 1/t_{j_m}$ according to Line 14 in Algorithm 1. We then conclude that

$$\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} [c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})]$$

$$\leq \underbrace{\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle}_{A_1} + \underbrace{(B_\star + 1) \cdot \sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \frac{1}{t_{j_m}}}_{A_2}. \quad (10)$$

To bound A_1: Recall that $\widehat{\boldsymbol{\theta}}_{j_m}$ given by Line 12 is the center of the confidence ellipsoid $\mathcal{C}_{j_m}$. First for each term $\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle$ in A_1, we write

$$\langle \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{j_m} + \widehat{\boldsymbol{\theta}}_{j_m} - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle$$

$$\leq \left(\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{j_m}\|_{\boldsymbol{\Sigma}_{t(m,h)}} + \|\widehat{\boldsymbol{\theta}}_{j_m} - \boldsymbol{\theta}_{m,h}\|_{\boldsymbol{\Sigma}_{t(m,h)}} \right) \cdot \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}}$$

$$\leq 2 \left(\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{j_m}\|_{\boldsymbol{\Sigma}_{t_{j_m}}} + \|\widehat{\boldsymbol{\theta}}_{j_m} - \boldsymbol{\theta}_{m,h}\|_{\boldsymbol{\Sigma}_{t_{j_m}}} \right) \cdot \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}}$$

$$\leq 4\beta_T \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}}. \quad (11)$$

Here the first inequality comes from the triangle inequality and Cauchy-Schwarz inequality. For the second inequality, recall that t_{j_m} given by Line 11 in Algorithm 1 is the time step when the j_m-th DEVI sub-routine is called, while $t(m,h)$ is the time step corresponds to the h-th step in the m-th interval and $t(m,h) \geq t_{j_m}$. Therefore, by the determinant-doubling triggering condition, we must have $\det(\boldsymbol{\Sigma}_{t(m,h)}) \leq 2\det(\boldsymbol{\Sigma}_{t_{j_m}})$, otherwise $t(m,h)$ and t_{j_m} would not belong to the same interval m. The second inequality then follows from $\lambda_i(\boldsymbol{\Sigma}_{t(m,h)}) \leq 2\lambda_i(\boldsymbol{\Sigma}_{t_{j_m}}) \, \forall i \in [d]$, where $\lambda_i(\cdot)$ is the i-th eigenvalue. The last inequality holds because under Lemma 8.3, $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_{m,h}$ belongs to the confidence ellipsoid $\mathcal{C}_{j_m}$ defined by Line 13.

Also note that for each term $\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle$ in A_1, we have

$$\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle \leq \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle$$

$$= \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h})$$

$$\leq B_\star, \quad (12)$$

where both inequalities hold due to $0 \leq V_{j_m}(\cdot) \leq B_\star$. Combine (11) and (12) and we have

$$A_1 \leq 4\beta_T \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min \left\{ 1, \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}} \right\}$$

$$\leq 4\beta_T \sqrt{ \left(\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} 1 \right) \cdot \left(\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min \left\{ 1, \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}}^2 \right\} \right)}, \quad (13)$$

where the first inequality holds due to $B_\star < \beta_T$, and the second inequality is by Cauchy-Schwarz inequality. Note that

$$
\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min \left\{ 1, \, \|\phi_{V_{j_m}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}}^2 \right\}
$$

$$
\leq 2 \left[d \log \left(\frac{\operatorname{trace}(\lambda \mathbf{I}) + T \cdot \max_{m \in \mathcal{M}_0} \|\phi_{V_{j_m}}(\cdot, \cdot)\|_2^2}{d} \right) - \log \left(\det(\lambda \mathbf{I}) \right) \right]
$$

$$
\leq 2d \log \left(\frac{\lambda d + T B_\star^2 d}{\lambda d} \right)
$$

$$
= 2d \log \left(1 + T B_\star^2 / \lambda \right),
$$

where the first inequality holds by Lemma H.5, and the second inequality holds because $V_{j_m}(\cdot) \leq B_\star$ under Lemma 8.3 and thus $\max_{m \in \mathcal{M}_0} \|\phi_{V_{j_m}}(\cdot, \cdot)\|_2 \leq B_\star \sqrt{d}$ by Assumption 3.1. Combine the above inequality with (13) and we conclude that

$$
A_1 \leq 4\beta_T \sqrt{2Td \cdot \log \left(1 + B_\star^2 T / \lambda \right)}. \tag{14}
$$

To bound A_2: by the definition of $\mathcal{M}_0$ we can rewrite A_2 as

$$
A_2 = (B_\star + 1) \cdot \sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \frac{1}{t_{j_m}} = (B_\star + 1) \cdot \sum_{j=1}^{J} \sum_{t=t_j+1}^{t_{j+1}} \frac{1}{t_j}.
$$

Note that the time step doubling condition $t \geq 2t_j$ in Line 9 implies that $t_{j+1} \leq 2t_j$ for all j. Therefore we have

$$
A_2 \leq (B_\star + 1) \cdot \sum_{j=1}^{J} \frac{2t_j}{t_j} = 2(B_\star + 1)J \leq 4.5 d B_\star \left[\log \left(1 + \frac{T B_\star^2 d}{\lambda} \right) + \log(T) \right],
$$

where the last step is by Lemma D.3. Together with (10) and (14) we conclude that

$$
\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} [c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})]
$$

$$
\leq 4\beta_T \sqrt{2Td \cdot \log \left(1 + B_\star^2 T / \lambda \right)} + 5 d B_\star \left[\log \left(1 + \frac{T B_\star^2 d}{\lambda} \right) + \log(T) \right]. \tag{15}
$$

To bound E_1, it remains to bound the following

$$
\sum_{m \in \mathcal{M}_0^c} \sum_{h=1}^{H_m} [c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})].
$$

Note that by definition, $\mathcal{M}_0^c$ are all the intervals m such that $j_m = 0$, i.e., the intervals before the first call of the DEVI sub-routine. However, since $t_0 = 1$, by the triggering condition $t \geq 2t_0$, we know that the first DEVI is called at $t = 2$. Therefore we have

$$
\sum_{m \in \mathcal{M}_0^c} \sum_{h=1}^{H_m} [c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})] = \sum_{h=1}^{2} [c_{1,h} + \mathbb{P}V_0(s_{1,h}, a_{1,h}) - Q_0(s_{1,h}, a_{1,h})]
$$

$$
\leq 4,
$$

where the inequality holds because $c_{1,h}, V_0(\cdot) \leq 1$ and $0 \leq Q_0(\cdot, \cdot)$. Together with (15) we conclude that

$$
E_1 \leq 4\beta_T \sqrt{2Td \cdot \log \left(1 + B_\star^2 T / \lambda \right)} + 5 d B_\star \left[\log \left(1 + \frac{T B_\star^2 d}{\lambda} \right) + \log(T) \right] + 4. \tag{16}
$$

$\square$

D.2. Bounding E_2

The term E_2 is the sum of a martingale difference sequence. However, the function V_{j_m} is random and not necessarily bounded, which disqualifies us from applying tools like Azuma-Hoeffding inequality directly. To deal with this issue, we use an auxiliary sequence of functions. The result is summarized by the following lemma.

Lemma D.2. *With probability at least $1 - \delta$, both the event of Lemma 8.3 and the following hold*

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

Proof of Lemma D.2. We define the filtration $\{\mathcal{F}_{m,h}\}_{m,h}$ such that $\mathcal{F}_{m,h}$ is the σ-field of all the history up until $(s_{m,h}, a_{m,h})$ which contains $(s_{m,h}, a_{m,h})$ but does not contain $s_{m,h+1}$. Then $(s_{m,h}, a_{m,h})$ is $\mathcal{F}_{m,h}$-measurable. Also note that the time step t_{j_m} is no later than the time step $t(m, h)$, and thus the function V_{j_m} is also $\mathcal{F}_{m,h}$-measurable. By the definition of the operator $\mathbb{P}$, we have

$$\mathbb{E}\left[V_{j_m}(s_{m,h+1})|\mathcal{F}_{m,h}\right] = \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}),$$

which shows that the term E_2 is the sum of a martingale difference sequence. To deal with the problem that V_{j_m} might not be uniformly bounded, we define an auxiliary sequence of functions

$$\widetilde{V}_{j_m}(\cdot) := \min\{B_\star, V_{j_m}(\cdot)\},$$

and it immediately holds that $\widetilde{V}_{j_m}$ is $\mathcal{F}_{m,h}$-measurable. We now write E_2 as

$$
\begin{aligned}
E_2 = \sum_{m=1}^{M} \sum_{h=1}^{H_m} & \left[\widetilde{V}_{j_m}(s_{m,h+1}) - \mathbb{P}\widetilde{V}_{j_m}(s_{m,h}, a_{m,h})\right] \\
& + \sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[[V_{j_m} - \widetilde{V}_{j_m}](s_{m,h+1}) - \mathbb{P}[V_{j_m} - \widetilde{V}_{j_m}](s_{m,h}, a_{m,h})\right].
\end{aligned}
$$

Since $\widetilde{V}_{j_m}$ is bounded, we can apply Lemma H.2 and get that, with probability at least $1 - \delta/2$,

$$E_2 \leq 2B_\star \sqrt{2T \log\left(\frac{T}{\delta/2}\right)} + \sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[[V_{j_m} - \widetilde{V}_{j_m}](s_{m,h+1}) - \mathbb{P}[V_{j_m} - \widetilde{V}_{j_m}](s_{m,h}, a_{m,h})\right].$$

Now note that under the event of Lemma 8.3, we have $\widetilde{V}_{j_m} = V_{j_m}$ for all $j_m \geq 1$ by optimism and also $\widetilde{V}_0 = V_0$ by the initialization, which implies that the second term in the RHS is zero. Therefore, take the intersection of the two events and we conclude that, with probability at least $1 - \delta$, $E_2 \leq 2B_\star \sqrt{2T \log(2T/\delta)}$. $\qquad\square$

D.3. Bounding E_3

To bound E_3, we first need the following lemma which shows that the total calls to `DEVI` in Algorithm 1 can be bounded. The proof shows that our design of the update condition (i.e. Line 9 in Algorithm 1) is crucial to our regret analysis. Importantly, the determinant doubling criterion alone is not enough, and the novel time step doubling trick is necessary.

Lemma D.3. *Conditioned on the event in Lemma 8.3, the total number of calls to `DEVI` is bounded by $J \leq 2d \log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + 2\log(T)$.*

Proof of Lemma D.3. By Line 9 we have $J = J_1 + J_2$ 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. First we bound J_1. Note that V_0 is from the initialization

instead of the output of `DEVI` and it holds that $V_0 \leq B_\star$. By Line 7 of Algorithm 1 and the initialization $\Sigma_0 = \lambda \mathbf{I}$, we have

$$\|\Sigma_T\|_2 = \left\| \lambda \mathbf{I} + \sum_{j=0}^{J} \sum_{t=t_j+1}^{t_{j+1}} \phi_{V_j}(s_t, a_t) \phi_{V_j}(s_t, a_t)^\top \right\|_2$$

$$\leq \lambda + \sum_{j=0}^{J} \sum_{t=t_j+1}^{t_{j+1}} \|\phi_{V_j}(s_t, a_t)\|_2^2$$

$$\leq \lambda + TB_\star^2 d,$$

where the first inequality is by the triangle inequality and the second inequality holds by Assumption 3.1 and $V_j \leq B_\star$ for all $j \geq 0$ under the event of Lemma 8.3. We then have that $\det(\Sigma_T) \leq (\lambda + TB_\star^2 d)^d$. It follows that

$$\left(\lambda + TB_\star^2 d\right)^d \geq 2^{J_1} \cdot \det\left(\Sigma_0\right) = 2^{J_1} \cdot \lambda^d,$$

by the determinant-doubling trigger condition. From the above inequality we conclude that

$$J_1 \leq 2d \log \left(1 + \frac{TB_\star^2 d}{\lambda}\right).$$

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

$$J \leq 2d \log \left(1 + \frac{TB_\star^2 d}{\lambda}\right) + 2 \log(T).$$

$\square$

We are now ready to bound E_3 in (8).

Lemma D.4. *Assume the event in Lemma 8.3 holds. Then we have*

$$\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right) - \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}}) \leq 1 + 2dB_\star \log \left(1 + \frac{TB_\star^2 d}{\lambda}\right) + 2B_\star \log(T).$$

Proof of Lemma D.4. The proof resembles that of Lemma 31 in Tarbouriech et al. (2021b). We first consider the first term in the LHS. Rearrange the summation and we have

$$\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right)$$

$$= \sum_{m=1}^{M} V_{j_m}(s_{m,1}) - V_{j_m}(s_{m,H_m+1})$$

$$= \sum_{m=1}^{M-1} \left(V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1}) \right) + \sum_{m=1}^{M-1} \left(V_{j_m}(s_{m,1}) - V_{j_{m+1}}(s_{m+1,1}) \right)$$

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

Note that second sum in the above equation is a telescoping sum. Thus we have

$$
\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right)
$$
$$
= \sum_{m=1}^{M-1} \left(V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1}) \right) + V_{j_1}(s_{1,1}) - V_{j_M}(s_{M,1})
$$
$$
\quad + V_{j_M}(s_{M,1}) - V_{j_M}(s_{M,H_M+1})
$$
$$
= \sum_{m=1}^{M-1} \left(V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1}) \right) + V_{j_1}(s_{1,1}) - V_{j_M}(s_{M,H_M+1})
$$
$$
\leq \sum_{m=1}^{M-1} \left(V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1}) \right) + V_{j_1}(s_{1,1}), \tag{17}
$$

where the inequality holds because $V_j(\cdot)$ is non-negative for all j.

We now consider the term $V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1})$. Note that by the interval decomposition, interval m ends if and only if either of the two conditions are met. If interval m ends because goal is reached, then we have

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

If it ends because the `DEVI` sub-routine is triggered, then the value function estimator is updated by `DEVI` and $j_m \neq j_{m+1}$. In such case we simply apply the trivial upper bound $V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1}) \leq \max_j \|V_j\|_\infty$. By Lemma D.3, this happens at most $J \leq 2d \log \left(1 + \frac{TB_\star^2 d}{\lambda} \right) + 2 \log(T)$ times. Therefore, we can further bound the RHS of (17) as

$$
\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right)
$$
$$
\leq \sum_{m=1}^{M-1} V_{j_{m+1}}(s_{\text{init}}) \cdot \mathbb{1}\{m+1 \in \mathcal{M}(M)\} + V_{j_1}(s_{1,1}) + \left[2d \log \left(1 + \frac{TB_\star^2 d}{\lambda} \right) + 2 \log(T) \right] \cdot \max_j \|V_j\|_\infty
$$
$$
\leq \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}}) + V_0(s_{\text{init}}) + 2dB_\star \log \left(1 + \frac{TB_\star^2 d}{\lambda} \right) + 2B_\star \log(T)
$$
$$
\leq \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}}) + 1 + 2dB_\star \log \left(1 + \frac{TB_\star^2 d}{\lambda} \right) + 2B_\star \log(T),
$$

where the second inequality is by $\|V_j\|_\infty \leq B_\star$ and the last step is by the initialization $\|V_0\|_\infty \leq 1$. $\qquad\square$

D.4. Proof of Theorem 5.2 and Theorem 5.4

Proof of Theorem 5.2. The total cost in K episodes is upper bound by $R_K + KB_\star$ and is lower bounded by $T \cdot c_{\min}$. Together with Theorem 8.1, with probability at least $1 - \delta$, we have

$$
T \cdot c_{\min} \leq 6\beta_T \sqrt{dT \log \left(1 + \frac{TB_\star^2}{\lambda} \right)} + 7dB_\star \log \left(T + \frac{T^2 B_\star^2 d}{\lambda} \right) + KB_\star.
$$

Solving the above inequality for the total number of steps T, we obtain that

$$
T = \mathcal{O} \left(\log^2 \left(\frac{1}{\delta} \right) \cdot \left(\frac{KB_\star}{c_{\min}} + \frac{B^2 d^2}{c_{\min}^2} \right) \right).
$$

Plugging this into Theorem 8.1 yields the desired result. $\qquad\square$

Proof of Theorem 5.4. By picking $\rho = K^{-1/3}$, we have $c_{\min} \geq K^{-1/3}$. Replacing $c_{\min}$ with $K^{-1/3}$ in Theorem 5.2, the regret for the perturbed SSP is upper bounded by

$$
\mathcal{O}\left(\widetilde{B}^{1.5} d K^{2/3} \cdot \log^2\left(\frac{K\widetilde{B}d}{\delta} \right) + \widetilde{B}^2 d^2 K^{1/3} \log^2\left(\frac{K\widetilde{B}d}{\delta} \right) \right),
$$

where $\widetilde{B} = B + T_\star/\rho$. Since the difference between the optimal cost of the perturbed SSP and the original SSP is at most $T_\star \rho K$, the regret for the original SSP is upper bounded by

$$
\mathcal{O}\left(\widetilde{B}^{1.5} d K^{2/3} \cdot \log^2\left(\frac{K\widetilde{B}d}{\delta} \right) + \widetilde{B}^2 d^2 K^{1/3} \log^2\left(\frac{K\widetilde{B}d}{\delta} \right) \right) + T_\star K^{2/3},
$$

which completes the proof. $\qquad\square$

D.5. Proof of Theorem 8.1

Proof. Note that the regret decomposition (7) is proved under the condition that the event of Lemma 8.3 holds. Then together with Lemmas 8.3, D.1 and D.2, we conclude that with probability at least $1 - \delta$,

$$
R(M) \leq 4\beta_T \sqrt{2Td \cdot \log\left(1 + B_\star^2 T/\lambda\right)} + 5dB_\star \left[\log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + \log(T) \right]
$$

$$
+ 2B_\star \sqrt{2T \log\left(\frac{2T}{\delta} \right)}
$$

$$
+ 4 + 2dB_\star \log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + 2B_\star \log(T) + 2.
$$

Combining the lower order terms finishes the proof. $\qquad\square$

E. Lower Bound

E.1. Proof of the Lower Bound

Proof of Theorem 7.1. We now construct a class of challenging SSP instances. We denote these SSPs by $M = \{\mathcal{S}, \mathcal{A}, \mathbb{P}_{\boldsymbol{\theta}}, c, s_{\text{init}}, g\}$. The state space $\mathcal{S}$ contains two states, i.e., $\mathcal{S} = \{s_{\text{init}}, g\}$. The action space $\mathcal{A}$ contains 2^{d-1} actions where each action $\mathbf{a} \in \mathcal{A}$ is a $(d-1)$-dimensional vector $\mathbf{a} \in \{-1, 1\}^{d-1}$. Here we use the boldface notation $\mathbf{a}$ instead of a to emphasize the action is represented by a vector. The cost function is given as $c(s_{\text{init}}, \mathbf{a}) = 1$ and $c(g, \mathbf{a}) = 0$ for any $\mathbf{a} \in \mathcal{A}$. The transition kernel $\mathbb{P}_{\boldsymbol{\theta}}$ of this SSP class is parameterized by a $(d-1)$-dimensional vector $\boldsymbol{\theta} \in \{-\frac{\Delta}{d-1}, \frac{\Delta}{d-1}\}^{d-1} = \Theta$. Specifically, for any $\mathbf{a} \in \mathcal{A}$, we have

$$
\mathbb{P}_{\boldsymbol{\theta}}(s_{\text{init}}|s_{\text{init}}, \mathbf{a}) = 1 - \delta - \langle \mathbf{a}, \boldsymbol{\theta} \rangle, \quad \mathbb{P}_{\boldsymbol{\theta}}(g|s_{\text{init}}, \mathbf{a}) = \delta + \langle \mathbf{a}, \boldsymbol{\theta} \rangle, \quad \mathbb{P}_{\boldsymbol{\theta}}(g|g, \mathbf{a}) = 1,
$$

where δ and Δ are parameters to be determined later. It is easy to verify that this is indeed an instance of linear mixture SSP with the parameter $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^\top, 1)^\top \in \mathbb{R}^d$ and the feature mapping $\boldsymbol{\phi}(s_{\text{init}}|s_{\text{init}}, \mathbf{a}) = (-\mathbf{a}^\top, 1-\delta)^\top$, $\boldsymbol{\phi}(g|s_{\text{init}}, \mathbf{a}) = (\mathbf{a}^\top, \delta)^\top$, $\boldsymbol{\phi}(s_{\text{init}}|g, \mathbf{a}) = \mathbf{0_d}$, and $\boldsymbol{\phi}(g|g, \mathbf{a}) = (\mathbf{0}_{d-1}^\top, 1)^\top$.

Remark E.1. In addition, this hard-to-learn instance can be adapted into a linear SSP studied in Vial et al. (2021). More specifically, it suffices to set $\boldsymbol{\theta}^* = (1, \mathbf{0}_d^\top)^\top$, $\boldsymbol{\mu}(s_{\text{init}}) = (1 - \delta, -\sqrt{d}\boldsymbol{\theta}^\top, 0)$, $\boldsymbol{\phi}(s_{\text{init}}, \mathbf{a}) = (1, \mathbf{a}^\top/\sqrt{d}, 0)^\top$ and $\boldsymbol{\phi}(g, \mathbf{a}) = (0, \mathbf{0}_{d-1}^\top, 1)^\top$. Then the linear SSP defined by the cost function $c(s, \mathbf{a}) = \boldsymbol{\phi}(s, \mathbf{a})^\top \boldsymbol{\theta}^*$ and the transition probability function $\mathbb{P}_{\boldsymbol{\theta}}(s'|s, \mathbf{a}) = \boldsymbol{\phi}(s, \mathbf{a})^\top \boldsymbol{\mu}(s')$ indeed recovers our construction above. This suggests that our analysis also yields a $\Omega(dB_\star \sqrt{K})$ for linear SSP, further complementing the results in Vial et al. (2021).

Note that for this SSP instance, the optimal policy is to always choose $\mathbf{a}_{\boldsymbol{\theta}}$ in state s_{init}, where $\mathbf{a}_{\boldsymbol{\theta}}$ denote the vector whose entries has the same sign as the corresponding entries of $\boldsymbol{\theta}$, i.e., $\text{sgn}(\mathbf{a}_{\boldsymbol{\theta},j}) = \text{sgn}(\boldsymbol{\theta}_j)$ for $j = 1, \cdots, d-1$. Here $\mathbf{a}_{\boldsymbol{\theta},j}$ and $\boldsymbol{\theta}_j$ denote the j-th entry of the respective vectors. Then the expected cost under the optimal policy is

$$
V_1^{\pi_{\boldsymbol{\theta}}^\star}(s_{\text{init}}) = \sum_{t=1}^{\infty} (1 - \delta - \Delta)^{t-1} (\delta + \Delta) t = \frac{1}{\delta + \Delta}.
$$

Therefore we will choose δ and Δ such that

$$\delta + \Delta = \frac{1}{B_\star}. \tag{18}$$

It remains to show that for any history-dependent and possibly non-stationary policy $\pi = \{\pi_t\}_{t=1}^\infty$, there exists some valid choice of δ and Δ such that the corresponding SSP class is hard to learn.

Let's consider the regret in an arbitrary episode k. Let $s_1 = s_{\text{init}}$. The expected regret can be written as

$$
\begin{aligned}
R_{\boldsymbol{\theta},k} \\
&= V_1^\pi(s_1) - V_1^{\pi_{\boldsymbol{\theta}}^\star}(s_1) \\
&= V_1^\pi(s_1) - \mathbb{E}_{\mathbf{a}_1 \sim \pi}[Q_1^{\pi_{\boldsymbol{\theta}}^\star}(s_1, \mathbf{a}_1)] + \mathbb{E}_{a_1 \sim \pi}[Q_1^{\pi_{\boldsymbol{\theta}}^\star}(s_1, \mathbf{a}_1)] - V_1^{\pi^\star}(s_1) \\
&= \mathbb{E}_{\mathbf{a}_1}[c(s_1, \mathbf{a}_1)] + \mathbb{E}_{\mathbf{a}_1}\left\{\mathbb{E}_{s_2 \sim \mathbb{P}(\cdot|s_1, \mathbf{a}_1)}[V_2^\pi(s_2)]\right\} - \mathbb{E}_{a_1}[c(s_1, \mathbf{a}_1)] - \mathbb{E}_{a_1}\left\{\mathbb{E}_{s_2 \sim \mathbb{P}(\cdot|s_1, \mathbf{a}_1)}[V_2^{\pi_{\boldsymbol{\theta}}^\star}(s_2)]\right\} \\
&\quad + \mathbb{E}_{a_1}[Q_1^{\pi_{\boldsymbol{\theta}}^\star}(s_1, \mathbf{a}_1)] - V_1^{\pi_{\boldsymbol{\theta}}^\star}(s_1) \\
&= \mathbb{E}_{\mathbf{a}_1, s_2}[V_2^\pi(s_2) - V_2^{\pi_{\boldsymbol{\theta}}^\star}(s_2)] + \mathbb{E}_{\mathbf{a}_1}[Q_1^{\pi_{\boldsymbol{\theta}}^\star}(s_1, \mathbf{a}_1)] - V_1^{\pi_{\boldsymbol{\theta}}^\star}(s_1), \\
&= \mathbb{E}_{\mathbf{a}_1, s_2}[V_2^\pi(s_2) - V_2^{\pi_{\boldsymbol{\theta}}^\star}(s_2)] + \mathbb{E}_{\mathbf{a}_1}\left[\frac{2\Delta}{d-1}\mathbb{1}\{s_1 = s_{\text{init}}\}\sum_{j=1}^{d-1}\mathbb{1}\{\text{sgn}(a_{1,j}) \neq \text{sgn}(\theta_j)\}\right] \cdot B_\star, \tag{19}
\end{aligned}
$$

where the third equality is by the Bellman equation, and the last equality holds because choosing $\mathbf{a}_1$ at state $s_1 = s_{\text{init}}$ instead of $\mathbf{a}_{\boldsymbol{\theta}}$ results in an extra probability of $\frac{2\Delta}{d-1}\sum_{j=1}^d\mathbb{1}\{\text{sgn}(a_{1,j}) \neq \text{sgn}(\theta_j)\}$ to remain in s_{init} for step 2, which incurs an extra cost of 1 by our construction of the cost function. Now by recursion, we can write the regret in episode k as

$$R_{\boldsymbol{\theta},k} = \frac{2\Delta B_\star}{d-1} \cdot \sum_{i=1}^\infty \mathbb{E}_k\left[\mathbb{1}\{s_i = s_{\text{init}}\} \cdot \sum_{j=1}^{d-1}\mathbb{1}\{\text{sgn}(a_{i,j}) \neq \text{sgn}(\theta_j)\}\right],$$

where the expectation $\mathbb{E}_k$ is taken with respect to the trajectory induced by the transition kernel $\mathbb{P}_{\boldsymbol{\theta}}$ and history-dependent policy π given the history till the end of episode $k-1$.

We can now write the total expected regret of π in K episodes given $\boldsymbol{\theta}$ as

$$R_{\boldsymbol{\theta}}(K) = \frac{2\Delta B_\star}{d-1} \cdot \sum_{t=1}^\infty \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{1}\{s_t = s_{\text{init}}\} \cdot \sum_{j=1}^{d-1}\mathbb{1}\{\text{sgn}(a_{t,j}) \neq \text{sgn}(\theta_j)\}\right],$$

where the expectation is taken with respect to $\mathbb{P}_{\boldsymbol{\theta}}$ and π. Here we omit the subscript π since it is clear from the context.

We denote the total number of steps in s_{init} by $N := \sum_{t=1}^\infty \mathbb{1}\{s_t = s_{\text{init}}\}$, and for $j = 1, \cdots, d-1$,

$$N_j(\boldsymbol{\theta}) := \sum_{t=1}^\infty \mathbb{1}\{s_t = s_{\text{init}}\} \cdot \mathbb{1}\{\text{sgn}(a_{t,j}) \neq \text{sgn}(\theta_j)\}.$$

This allows us to write $R_{\boldsymbol{\theta}}(K) = \frac{2\Delta B_\star}{d-1}\mathbb{E}_{\boldsymbol{\theta}}[\sum_{j=1}^{d-1} N_j(\boldsymbol{\theta})]$. Now to bound the regret, we can rely on a standard technique using Pinsker's inequality (Jaksch et al., 2010). However, this would require each $N_j(\boldsymbol{\theta})$ to be almost surely bounded, which does not hold in the case of SSP. To circumvent this issue, we apply the "capping" trick from Cohen et al. (2020) that cap the learning process to contain only the first T steps for some pre-determined T. To be specific, if the K episodes are finished before the time T, then the agent remains in state g. In this case, the actual regret for this capped process is exactly equal to the uncapped process. On the other hand, if at time T the agent has not finished all the K episodes, it is stopped immediately. In this case the actual regret is smaller than that of the uncapped process. Therefore, we only need to lower bound the expected regret for this capped process.

Let $N^- := \sum_{t=1}^T \mathbb{1}\{s_t = s_{\text{init}}\}$, and

$$N_j^-(\boldsymbol{\theta}) := \sum_{t=1}^T \mathbb{1}\{s_t = s_{\text{init}}\} \cdot \mathbb{1}\{\text{sgn}(a_{t,j}) \neq \text{sgn}(\theta_j)\}.$$

Then we can lower bound the expected regret by $R_{\boldsymbol{\theta}}(K) \geq \frac{2\Delta B_\star}{d-1} \mathbb{E}_{\boldsymbol{\theta}}[\sum_{j=1}^{d-1} N_j^-(\boldsymbol{\theta})]$. For each $\boldsymbol{\theta} \in \{-\frac{\Delta}{d-1}, \frac{\Delta}{d-1}\}^{d-1}$, let $\boldsymbol{\theta}^j$ denote the vector which differs from $\boldsymbol{\theta}$ only at the j-th entry. Then we sum over $\boldsymbol{\theta}$ and get that

$$
\begin{aligned}
2 \sum_{\boldsymbol{\theta} \in \Theta} R_{\boldsymbol{\theta}}(K) &\geq \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\mathbb{E}_{\boldsymbol{\theta}}[N_j^-(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}^j}[N_j^-(\boldsymbol{\theta}^j)] \right) \\
&= \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\mathbb{E}_{\boldsymbol{\theta}^j}[N^-] + \mathbb{E}_{\boldsymbol{\theta}}[N_j^-(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}^j}[N_j^-(\boldsymbol{\theta})] \right) \\
&= \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\mathbb{E}_{\boldsymbol{\theta}}[N^-] + \mathbb{E}_{\boldsymbol{\theta}}[N_j^-(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}^j}[N_j^-(\boldsymbol{\theta})] \right).
\end{aligned}
\tag{20}
$$

The next shows that for large enough T, $\mathbb{E}_{\boldsymbol{\theta}}[N^-]$ is lower bounded for all $\boldsymbol{\theta}$.

Lemma E.2 (Lemma C.2 in Cohen et al. 2020). *If $T \geq 2KB_\star$, then it holds that $\mathbb{E}_{\boldsymbol{\theta}}[N^-] \geq KB_\star/4$ for all $\boldsymbol{\theta} \in \{-\frac{\Delta}{d-1}, \frac{\Delta}{d-1}\}^{d-1}$.*

We will also use the following lemma which is a version of Pinsker's inequality (Jaksch et al., 2010; Zhou et al., 2021b).

Lemma E.3 (Pinsker's inequality). *Fix T and denote the trajectory $\mathbf{s} = \{s_1, \cdots, s_T\} \in \mathcal{S}^T$. For any two probability distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ on $\mathcal{S}^T$ and any bounded function $f : \mathcal{S}^T \to [0, D]$, we have*

$$
\mathbb{E}_{\mathcal{P}_1} f(\mathbf{s}) - \mathbb{E}_{\mathcal{P}_2} f(\mathbf{s}) \leq D \cdot \sqrt{\frac{\log 2}{2}} \cdot \sqrt{\mathrm{KL}(\mathcal{P}_2 || \mathcal{P}_1)}.
$$

Then we pick $T = 2KB_\star$ and get

$$
\begin{aligned}
2 \sum_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(K) &\geq \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\frac{KB_\star}{4} + \mathbb{E}_{\boldsymbol{\theta}}[N_j^-(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}^j}[N_j^-(\boldsymbol{\theta})] \right) \\
&\geq \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\frac{KB_\star}{4} - T\sqrt{\frac{1}{2}} \sqrt{\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}} || \mathcal{P}_{\boldsymbol{\theta}^j})} \right),
\end{aligned}
$$

where the first inequality is by Lemma E.2, and the second inequality is by Lemma E.3. The next lemma shows that the KL-divergence can be related to the quantity N^-.

Lemma E.4. *Suppose $4\Delta < \delta \leq 1/3$. Then we have*

$$
\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}} || \mathcal{P}_{\boldsymbol{\theta}^j}) \leq \frac{16\Delta^2}{(d-1)^2 \delta} \mathbb{E}_{\boldsymbol{\theta}}[N^-].
$$

It follows from Lemma E.4 that

$$
\begin{aligned}
2 \sum_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(K) &\geq \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\frac{KB_\star}{4} - T\sqrt{\frac{1}{2}} \cdot \frac{4\Delta}{d-1} \cdot \frac{1}{\sqrt{\delta}} \sqrt{\mathbb{E}_{\boldsymbol{\theta}}[N^-]} \right) \\
&\geq \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\frac{KB_\star}{4} - T^{3/2}\sqrt{\frac{1}{2}} \cdot \frac{4\Delta}{d-1} \cdot \frac{1}{\sqrt{\delta}} \right) \\
&= \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\frac{KB_\star}{4} - (2KB_\star)^{3/2}\sqrt{\frac{1}{2}} \cdot \frac{4\Delta}{d-1} \cdot \frac{1}{\sqrt{\delta}} \right),
\end{aligned}
\tag{21}
$$

where the last inequality is by $N^- \leq T = 2KB_\star$. Simplify the expression and we get that

$$\frac{1}{|\Theta|} \sum_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(K) \geq B_\star \frac{1}{|\Theta|} \cdot \frac{1}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\frac{\Delta K B_\star}{4} - \frac{8\Delta^2}{(d-1)\sqrt{\delta}} (KB_\star)^{3/2} \right)$$

$$= B_\star \left[\frac{\Delta K B_\star}{4} - \frac{8\Delta^2}{(d-1)\sqrt{\delta}} (KB_\star)^{3/2} \right]. \tag{22}$$

We now pick

$$\Delta = \frac{(d-1)\sqrt{\delta}}{64\sqrt{KB_\star}}, \tag{23}$$

and δ such that $\delta + \Delta = 1/B_\star$, plug into (22) and get that

$$\frac{1}{|\Theta|} \sum_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(K) \geq \frac{dB_\star \sqrt{\delta}\sqrt{KB_\star}}{512} \geq \frac{dB_\star \sqrt{K}}{1024},$$

where the last step is by $\delta + \Delta = \frac{1}{B_\star}$ and $\Delta < \delta$. Therefore, there must exist some $\boldsymbol{\theta} \in \Theta$ such that the expected regret $R_{\boldsymbol{\theta}}(K)$ satisfies

$$R_{\boldsymbol{\theta}}(K) \geq \frac{dB_\star \sqrt{K}}{1024}.$$

Taking $\boldsymbol{\theta}^* = (\boldsymbol{\theta}, 1)^\top \in \mathbb{R}^d$ finishes the proof of the lower bound. It remains to check the conditions. Note that by (18) and (23), we have

$$\delta + \frac{(d-1)\sqrt{\delta}}{64\sqrt{KB_\star}} = \frac{1}{B_\star}.$$

Since we also have $\Delta < \delta$, we then require

$$\frac{d-1}{64\sqrt{KB_\star}} \leq \sqrt{\delta} < \frac{1}{\sqrt{B_\star}},$$

which implies that $K > (d-1)^2/2^{12}$. This finishes the proof of Theorem 7.1.

$\square$

E.2. Proof of Lemmas in Appendix E.1

Lemma E.2 is straightforward and we refer the reader to Lemma C.2 in Cohen et al. 2020. Lemma E.3 is a standard result. We thus omit their proof. Lemma E.4 can be easily adapted from Lemma 6.8 in Zhou et al. 2021b. However, since the MDP instance we construct under the SSP setting differs from theirs under the discounted setting, we present the proof here for completeness.

Proof of Lemma E.4. Denote the trajectory by $\mathbf{s}_t = \{s_1, s_2, \cdots, s_t\}$. The chain rule of the KL-divergence gives

$$\text{KL}(\mathcal{P}_{\boldsymbol{\theta}} || \mathcal{P}_{\boldsymbol{\theta}^j}) = \sum_{t=1}^{T-1} \text{KL}\left[\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t) \big|\big| \mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)\right], \tag{24}$$

where

$$\text{KL}\left[\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t) \big|\big| \mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)\right] := \sum_{\mathbf{s}_{t+1} \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)}.$$

Then we write

$$\sum_{s_{t+1}\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)}$$

$$= \sum_{\mathbf{s}_t\in\mathcal{S}^{\times t}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}_t) \sum_{s\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|\mathbf{s}_t) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}=s|\mathbf{s}_t)}$$

$$= \sum_{\mathbf{s}_{t-1}\in\mathcal{S}^{\times(t-1)}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}_{t-1}) \sum_{s'\in\mathcal{S},\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s',\mathbf{a}_t=\mathbf{a}|\mathbf{s}_{t-1})$$

$$\cdot \sum_{s\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s',\mathbf{a}_t=\mathbf{a},\mathbf{s}_{t-1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s',\mathbf{a}_t=\mathbf{a},\mathbf{s}_{t-1})}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}=s|s_t=s',\mathbf{a}_t=\mathbf{a},\mathbf{s}_{t-1})}.$$

Note that when $s' = g$, the transition is irrelevant of $\boldsymbol{\theta}$ and $\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s',\mathbf{a}_t=\mathbf{a},\mathbf{s}_{t-1}) = \mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}=s|s_t=s',\mathbf{a}_t=\mathbf{a},\mathbf{s}_{t-1})$ for all $\boldsymbol{\theta}$. Therefore the log-term in the above equation vanishes when $s' = g$. So we only need to consider the case where $s' = s_{\text{init}}$ in the summation, and it follows that

$$\sum_{s_{t+1}\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)}$$

$$= \sum_{\mathbf{s}_{t-1}\in\mathcal{S}^{\times(t-1)}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}_{t-1}) \sum_{\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a}|\mathbf{s}_{t-1})$$

$$\cdot \sum_{s\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a},\mathbf{s}_{t-1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a},\mathbf{s}_{t-1})}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}=s|s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a},\mathbf{s}_{t-1})}$$

$$= \sum_{\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a}) \cdot \sum_{s\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a})}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}=s|s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a})}. \tag{25}$$

Note that when $s_t = s_{\text{init}}$, s_{t+1} is either s_{init} or g with probability $1-\delta-\langle\mathbf{a},\boldsymbol{\theta}\rangle$ and $\delta+\langle\mathbf{a},\boldsymbol{\theta}\rangle$. Then we can further write (25) as

$$\sum_{s_{t+1}\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)}$$

$$= \sum_{\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a})$$

$$\cdot \left[(1-\delta-\langle\mathbf{a},\boldsymbol{\theta}\rangle) \cdot \log \frac{1-\delta-\langle\mathbf{a},\boldsymbol{\theta}\rangle}{1-\delta-\langle\mathbf{a},\boldsymbol{\theta}^j\rangle} + (\delta+\langle\mathbf{a},\boldsymbol{\theta}\rangle) \cdot \log \frac{\delta+\langle\mathbf{a},\boldsymbol{\theta}\rangle}{\delta+\langle\mathbf{a},\boldsymbol{\theta}^j\rangle} \right]$$

$$\le \sum_{\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s_{\text{init}},\mathbf{a}_t=\mathbf{a}) \cdot \frac{2\langle\mathbf{a},\boldsymbol{\theta}^j-\boldsymbol{\theta}\rangle^2}{\delta+\langle\mathbf{a},\boldsymbol{\theta}\rangle}, \tag{26}$$

where the last step holds due to the following inequality with $\delta' = \delta + \langle\mathbf{a},\boldsymbol{\theta}\rangle$, and $\epsilon' = \langle\mathbf{a},\boldsymbol{\theta}^j-\boldsymbol{\theta}\rangle$.

Lemma E.5 (Lemma 20, Jaksch et al. 2010). *For any real number δ' and ϵ' such that $0 \le \delta' \le 1/2$ and $\epsilon' \le 1 - 2\delta'$, we have*

$$\delta' \log \frac{\delta'}{\delta'+\Delta} + (1-\delta') \log \frac{1-\delta'}{1-\delta'-\epsilon'} \le \frac{2(\epsilon')^2}{\delta'}.$$

To verify the assumptions of Lemma E.5, note that $\delta' \le \delta+\Delta \le 1/12+1/3 < 1/2$ by $4\Delta \le \delta \le 1/3$ from the assumption of Lemma E.4. Also note that

$$\epsilon' = \langle a,\boldsymbol{\theta}^j-\boldsymbol{\theta}\rangle \le 2\Delta \le 1-2(\Delta+\delta) \le 1-2\delta',$$

where the first step is by the definition of $\boldsymbol{\theta}$, the second step is by $\delta \leq 1/12$ and $\delta + \Delta \leq 5/12$, and the last step is by $\delta' \leq \delta + \Delta$. Therefore, (26) holds and we have

$$
\begin{aligned}
\sum_{s_{t+1} \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)} &\leq \sum_{\mathbf{a} \in \mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t = s_{\text{init}}, \mathbf{a}_t = \mathbf{a}) \cdot \frac{2\langle \mathbf{a}, \boldsymbol{\theta}^j - \boldsymbol{\theta}\rangle^2}{\delta - \Delta} \\
&\leq \frac{2\langle \mathbf{a}, \boldsymbol{\theta}^j - \boldsymbol{\theta}\rangle^2}{\delta/2} \cdot \sum_{\mathbf{a} \in \mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t = s_{\text{init}}, \mathbf{a}_t = \mathbf{a}) \\
&= \frac{4(2\Delta)^2}{(d-1)^2 \delta} \cdot \sum_{\mathbf{a} \in \mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t = s_{\text{init}}, \mathbf{a}_t = \mathbf{a}) \\
&= \frac{16\Delta^2}{(d-1)^2 \delta} \cdot \mathcal{P}_{\boldsymbol{\theta}}(s_t = s_{\text{init}}).
\end{aligned}
$$

Together with (24) we have

$$
\begin{aligned}
\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}} \| \mathcal{P}_{\boldsymbol{\theta}^j}) = \sum_{t=1}^{T-1} \sum_{s_{t+1} \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)} \\
\leq \frac{16\Delta^2}{(d-1)^2 \delta} \sum_{t=1}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_t = s_{\text{init}}) \\
= \frac{16\Delta^2}{(d-1)^2 \delta} \mathbb{E}_{\boldsymbol{\theta}}[N^-],
\end{aligned}
$$

where the last step is by the definition of N^-.

$\square$

F. Lemmas for the Upper Bounds

F.1. Proof of Lemma 8.3

We first introduce the following classical result for self-normalized vector-valued martingales.

Lemma F.1 (Theorem 1, Abbasi-Yadkori et al. 2011). *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Suppose $\{\eta_t\}_{t=1}^{\infty}$ is a $\mathbb{R}$-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and $\eta_t | \mathcal{F}_{t-1}$ is B-sub-Gaussian. Let $\{\phi_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that ϕ_t is $\mathcal{F}_{t-1}$-measurable. Assume that $\boldsymbol{\Sigma}$ is an $d \times d$ positive definite matrix. For any $t \geq 1$, define*

$$
\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma} + \sum_{i=1}^{t} \phi_i \phi_i^\top, \quad \boldsymbol{a}_t = \sum_{i=1}^{t} \eta_i \phi_i.
$$

Then, for any $\delta > 0$, with probability at least δ, for all t, we have

$$
\|\boldsymbol{\Sigma}_t^{-1/2} \boldsymbol{a}_t\|_2 \leq B \sqrt{2 \log\left(\frac{\det(\boldsymbol{\Sigma}_t)^{1/2}}{\delta \cdot \det(\boldsymbol{\Sigma})^{1/2}}\right)}.
$$

In the following proof we will decompose t into different rounds. For all $j \geq 1$, round j corresponds to $t \in [t_j + 1, t_{j+1}]$, during which the action-value function estimator is the output Q_j of DEVI. We then apply an induction argument on the rounds to show that the optimism holds for all $j \geq 1$.

Proof of Lemma 8.3. From the initialization of Algorithm 1, we have $V_0 \leq B_\star$.
Let's consider **round 1**. We define $\eta_t = V_0(s_{t+1}) - \langle \phi_{V_0}(s_t, a_t), \boldsymbol{\theta}^* \rangle$ for $t \in [1, t_1]$. Then $\{\eta_t\}_{t=1}^{t_1}$ are $B_\star$-sub-Gaussian.

We then apply Lemma F.1 and conclude that the following holds with probability at least $1 - \frac{\delta}{t_1(t_1+1)}$, for all $t \in [1, t_1]$:

$$\left\| \mathbf{\Sigma}_t^{-1/2} \sum_{i=1}^{t} \boldsymbol{\phi}_{V_0}(s_i, a_i)\eta_i \right\|_2 \leq B_\star \sqrt{2\log\left(\frac{\det(\mathbf{\Sigma}_t)^{1/2}}{\delta \cdot \lambda^{d/2}/(t_1(t_1+1))}\right)}$$

$$\leq B_\star \sqrt{d\log\left(\frac{1 + td/(d\lambda)}{\delta/(t_1(t_1+1))}\right)}$$

$$\leq B_\star \sqrt{d\log\left(\frac{t_1(t_1+1) + t \cdot t_1(t_1+1)B_\star^2/\lambda}{\delta}\right)}, \tag{27}$$

where the second step is by Assumption 3.1, Lemma H.4 and the initialization $|V_0| \leq 1$. Consider the LHS of (27). We have

$$\left\| \mathbf{\Sigma}_t^{-1/2} \sum_{i=1}^{t} \boldsymbol{\phi}_{V_0}(s_i, a_i)\eta_i \right\|$$

$$= \left\| \mathbf{\Sigma}_t^{1/2}\mathbf{\Sigma}_t^{-1} \sum_{i=1}^{t} \boldsymbol{\phi}_{V_0}(s_i, a_i)V_0(s_{i+1}) - \mathbf{\Sigma}_t^{1/2}\mathbf{\Sigma}_t^{-1}\left(\mathbf{\Sigma}_t - \lambda\mathbf{I}\right)\boldsymbol{\theta}^* \right\|_2$$

$$= \left\| \mathbf{\Sigma}_t^{1/2}\widehat{\boldsymbol{\theta}}_t - \mathbf{\Sigma}_t^{1/2}\boldsymbol{\theta}^* + \lambda\mathbf{\Sigma}_t^{-1/2}\boldsymbol{\theta}^* \right\|_2$$

$$\geq \left\| \mathbf{\Sigma}_t^{1/2}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) \right\|_2 - \left\| \lambda\mathbf{\Sigma}_t^{-1/2}\boldsymbol{\theta}^* \right\|_2$$

$$\geq \left\| \mathbf{\Sigma}_t^{1/2}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) \right\|_2 - \lambda^{1/2} \cdot \sqrt{d},$$

where the first inequality holds by Cauchy-Schwarz inequality and the second inequality holds because $\|\boldsymbol{\theta}^*\|_2 \leq \sqrt{d}$. Together with (27) and the choice of β_t, we conclude that

$$\left\| \mathbf{\Sigma}_t^{1/2}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*) \right\|_2 \leq B_\star \sqrt{d\log\left(\frac{t_1(t_1+1) + t \cdot t_1(t_1+1)B_\star^2/\lambda}{\delta}\right)} + \sqrt{\lambda d} \leq \beta_{t_1}.$$

Since the above holds for all $t \in [1, t_1]$, it follows that with probability at least $1 - \frac{\delta}{t_1(t_1+1)}$, the true parameter $\boldsymbol{\theta}^*$ is in the set $\mathcal{C}_1 \cap \mathcal{B}$.

To show that the output Q_1 and V_1 of DEVI are optimistic, we apply a second induction argument on the loop of DEVI. For the base step, note that by non-negativity of $Q^\star$ and $V^\star$, we have $Q^{(0)} \leq Q^\star$ and $V^{(0)} \leq V^\star$. We now assume $Q^{(i)}$ and $V^{(i)}$ are optimistic. For the $i+1$-th iteration, we have

$$Q^{(i+1)}(\cdot, \cdot) = c(\cdot, \cdot) + (1-q) \cdot \min_{\boldsymbol{\theta} \in \mathcal{C}_1 \cap \mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(i)}}(\cdot, \cdot) \rangle$$

$$\leq c(\cdot, \cdot) + (1-q) \cdot \mathbb{P}V^{(i)}(\cdot, \cdot)$$

$$\leq c(\cdot, \cdot) + \mathbb{P}V^{(i)}(\cdot, \cdot)$$

$$\leq Q^\star(\cdot, \cdot),$$

where the first step is because we are considering the case where $\rho = 0$, the second step is because we are taking the minimum over a set that contains $\boldsymbol{\theta}^*$, the third step is by non-negativity of $\mathbb{P}V^{(i)}(\cdot, \cdot)$, and the last step is by the Bellman optimal condition (2) and the induction hypothesis that $V^{(i)}$ is optimistic. By induction, we conclude that $Q^{(i)}$ is optimistic for all i, and thus the final output $Q_1(\cdot, \cdot)$ and thus $V_1(\cdot)$ are both optimistic. We finish the proof for round 1.

Now for our outer induction, let's suppose that the event in Lemma 8.3 holds for round 1 to $j-1$ with high probability. That is, we define the event

$$\mathcal{E}_{j-1} := \{\boldsymbol{\theta}^* \in \mathcal{C}_i \cap \mathcal{B}, \quad V_i(\cdot) \leq V^\star(\cdot) \leq B_\star, \quad Q_i(\cdot, \cdot) \leq Q^\star(\cdot, \cdot) \quad \text{for all } i \in [1, j-1]\},$$

and assume that $\Pr(\mathcal{E}_{j-1}) \geq 1 - \delta'$ for some $\delta' > 0$. We now show that the event $\mathcal{E}_j$ also holds with high probability. Similar to the proof of Lemma D.2, we construct an auxiliary sequence of functions

$$\widetilde{V}_i(\cdot) := \min\{B_\star, V_i(\cdot)\}, \quad i \in [1, j-1].$$

We also denote, for any $i \in [1, j]$ and for any $t \in [t_{i-1} + 1, t_i]$,

$$\widetilde{\eta}_t = V_{i-1}(s_{t+1}) - \langle \phi_{\widetilde{V}_{i-1}}(s_t, a_t), \boldsymbol{\theta}^* \rangle,$$

$$\widetilde{\boldsymbol{\Sigma}}_t = \lambda \mathbf{I} + \sum_{l=1}^{t} \phi_{\widetilde{V}_{i(l)-1}}(s_l, a_l) \phi_{\widetilde{V}_{i(l)-1}}(s_l, a_l)^\top,$$

$$\widetilde{\boldsymbol{\theta}}_t = \widetilde{\boldsymbol{\Sigma}}_t^{-1} \sum_{l=1}^{t} \phi_{\widetilde{V}_{i(l)-1}}(s_l, a_l) \widetilde{V}_{i(l)-1}(s_{l+1}),$$

$$\widetilde{\mathcal{C}}_i = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \ \left\| \widetilde{\boldsymbol{\Sigma}}_{t_i}^{1/2}(\widetilde{\boldsymbol{\theta}}_{t_i} - \boldsymbol{\theta}^*) \right\|_2 \le \beta_{t_i} \right\},$$

where $i(l)$ is the round that contains the time step l, i.e., $l \in [t_{i-1} + 1, t_i]$. Observe that, by this construction $\{\widetilde{\eta}_t\}_{t=1}^{t_j}$ are almost surely $B_\star$-sub-Gaussian. This allows us to apply Lemma F.1 and do the similar computation as above, and get that, with probability at least $1 - \frac{\delta}{t_j(t_j+1)}$, we have the event $\widetilde{\mathcal{E}}_j$ holds where

$$\widetilde{\mathcal{E}}_j := \left\{ \boldsymbol{\theta}^* \in \widetilde{\mathcal{C}}_j \cap \mathcal{B}, \quad V_j(\cdot) \le V^\star(\cdot) \le B_\star, \quad Q_j(\cdot, \cdot) \le Q^\star(\cdot, \cdot) \right\},$$

and Q_j is the output of $\mathtt{DEVI}(\widetilde{\mathcal{C}}_j, \epsilon_j, \frac{1}{t_j}, \rho)$.

Now, observe that under the event $\mathcal{E}_{j-1}$, the optimism implies that $\widetilde{V}_i = V_i$ for all $i \in [1, j-1]$. It follows that under $\mathcal{E}_{j-1}$, we have $\widetilde{\eta}_t = \eta_t$, $\widetilde{\boldsymbol{\Sigma}}_t = \boldsymbol{\Sigma}_t$, $\widetilde{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_t$ for all $t \le t_j$, and thus $\widetilde{\mathcal{C}}_j = \mathcal{C}_j$. We then have

$$\mathcal{E}_j = \mathcal{E}_{j-1} \cap \widetilde{\mathcal{E}}_j,$$

and by the union bound we have that $\Pr(\mathcal{E}_j) \ge 1 - \delta' - \frac{\delta}{t_j(t_1+1)}$.

Now, by induction and taking the union bound

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

we conclude that with probability at least $1 - \delta$, the good event holds for all $j \le J$, where J is the total number of times $\mathtt{DEVI}$ being called. Note that compared with the analysis of $\mathtt{EVI}$ in the discounted MDPs setting (for example in Zhou et al. 2021b), our analysis of $\mathtt{DEVI}$ in SSP uses the induction argument and a union bound, which results in extra t factors in the logarithmic term in the confidence radius β_t. At last, replacing $t(t+1)$ with $2t^2$ and δ with $\delta/2$ gives the final expression for β_t.

It remains to argue that $\mathtt{DEVI}$ always converges in finite time. To begin with, note that it suffices to show that $\|V^{(i)} - V^{(i-1)}\|_\infty$ shrinks exponentially. We now claim that $\|Q^i - Q^{(i-1)}\|_\infty$ shrinks exponentially, which together with (5) gives the desired result since $\|V^{(i)} - V^{(i-1)}\|_\infty \le \|Q^{(i)} - Q^{(i-1)}\|_\infty$. To show this, first note that for any (s, a) pair,

$$
\begin{aligned}
|Q^{(i)}(s,a) - Q^{(i-1)}(s,a)| &= (1-q) \cdot \left| \min_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \langle \boldsymbol{\theta}, \phi_{V^{(i-1)}}(s,a) \rangle - \min_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \langle \boldsymbol{\theta}, \phi_{V^{(i-2)}}(s,a) \rangle \right| \\
&\le (1-q) \cdot \max_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} |\langle \boldsymbol{\theta}, \phi_{V^{(i-1)}}(s,a) - \phi_{V^{(i-2)}}(s,a) \rangle| \\
&= (1-q) \cdot \left| \langle \bar{\boldsymbol{\theta}}, \phi_{V^{(i-1)}}(s,a) - \phi_{V^{(i-2)}}(s,a) \rangle \right| \\
&= (1-q) \cdot \left| \bar{\mathbb{P}}(V^{(i-1)} - V^{(i-2)})(s,a) \right| \\
&\le (1-q) \cdot \max_{s' \in \mathcal{S}} \left| V^{(i-1)}(s') - V^{(i-2)}(s') \right| \\
&= (1-q) \cdot \max_{s' \in \mathcal{S}} \left| \min_{a'} Q^{(i-1)}(s',a') - \min_{a'} Q^{(i-2)}(s',a') \right| \\
&\le (1-q) \cdot \|Q^{(i-1)} - Q^{(i-2)}\|_\infty,
\end{aligned}
$$

where $\bar{\boldsymbol{\theta}}$ is the $\boldsymbol{\theta}$ in the non-empty set $\mathcal{C} \cap \mathcal{B}$ that achieves the maximum. Here the first inequality holds due to the maximum function, the second inequality holds because $\bar{\mathbb{P}}(\cdot|s,a)$ is a probability distribution, and the last inequality holds due to the same reason as the first one. Now, since s, a are arbitrary in the above, we conclude that $\|Q^{(i)} - Q^{(i-1)}\|_{\infty} \leq (1-q)\|Q^{(i-1)} - Q^{(i-2)}\|_{\infty}$. This finishes the proof.

$\square$

G. The Bernstein-type Algorithm

In this section, we give the full details of the Bernstein-type algorithm $\mathtt{LEVIS}^+$. We introduce the algorithm design in Appendix G.1. We then present the analysis of $\mathtt{LEVIS}^+$ and the corresponding proof of Theorem 6.1.

It is worth mentioning that Bernstein-type confidence sets have been utilized in the existing literature on the tabular SSP (Rosenberg et al., 2020; Cohen et al., 2021; Tarbouriech et al., 2021b; Jafarnia-Jahromi et al., 2021; Chen et al., 2021a). The Bernstein technique can help get rid of the dependence on $c_{\min} > 0$ in the tabular setting (i.e., to achieve $\widetilde{\mathcal{O}}(\sqrt{K})$ for $c_{\min} = 0$), while it remains a challenge in the linear function approximation setting. Also, such variance-aware confidence sets allow 'horizon-free' algorithms for linear SSP (Chen et al., 2021b), though its regret suffers from worse dependence on the feature dimension d. Concurrent to our result, (Chen et al., 2021b) also proposed an algorithm for linear mixture SSP which utilizes Bernstein-type confidence sets. Similar to ours, they use the technique from the $\mathtt{UCRL-VTR}^+$ algorithm from Zhou et al. (2021a). The major difference is that their algorithm is based on the reduction to a finite-horizon linear mixture MDP while our algorithm is a direct algorithm without such a reduction.

G.1. A Detailed Introduction of $\mathtt{LEVIS}^+$

Here we go over the details of $\mathtt{LEVIS}^+$.

We define a variance operator $\mathbb{V}$ associated with the unknown underlying transition such that for any function $f : \mathcal{S} \to \mathbb{R}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$[\mathbb{V}f](s,a) := [\mathbb{P}f^2](s,a) - ([\mathbb{P}f](s,a))^2, \tag{28}$$

where $[\mathbb{P}f](s,a) = \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s,a)f(s')$. One can see from the definition that $[\mathbb{V}f](s,a)$ is the conditional variance of $f(s')$ where $s' \sim \mathbb{P}(\cdot|s,a)$. Further note that $[\mathbb{V}f](s,a)$ bears a nice form which allows us to estimate it by regression. To see this, we can calculate

$$[\mathbb{P}f^2](s,a) = \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s,a)f^2(s') = \sum_{s' \in \mathcal{S}} \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta}^* \rangle f^2(s') = \langle \boldsymbol{\phi}_{f^2}(s,a), \boldsymbol{\theta}^* \rangle,$$

$$[\mathbb{P}f](s,a) = \langle \boldsymbol{\phi}_f(s,a), \boldsymbol{\theta}^* \rangle,$$

which are linear functions of $\boldsymbol{\phi}_{f^2}$ and $\boldsymbol{\phi}_f$ respectively. Therefore, we can estimate $[\mathbb{V}f](s,a)$ by solving regression over $\mathbb{P}f$ and $\mathbb{P}f^2$.

Moreover, it turns out that, to apply the Bernstein-type concentration inequality, it suffices to use an upper bound of $\mathbb{V}V_j$ where V_j is the value function. Thus, the strategy here is to first construct an estimator $\widehat{\mathbb{V}}_t V_j$ (Line 7) and then add a deviation term to $\widehat{\mathbb{V}}_t V_j$ to get $\widehat{\sigma}_t^2$ (Line 8), which can be shown to be a tight upper bound of $\mathbb{V}V_j$ with high probability. Note that here we only need to apply the Hoeffding-type concentration inequality since this deviation term will not appear in the final regret bound.

Consequently, in Algorithm 3, we maintain two regularized Gram matrix matrices $\boldsymbol{\Sigma}_t$ (Line 9) and $\widetilde{\boldsymbol{\Sigma}}_t$ (Line 12). Here $\widetilde{\boldsymbol{\Sigma}}_t$ is the unweighted Gram matrix used to estimate the linear parameter of the $\mathbb{P}V_j^2$ term in the variance, and the estimator is given by $\widetilde{\boldsymbol{\theta}}_t$ (Line 14). On the other hand, $\boldsymbol{\Sigma}_t$ is a variance-weighted Gram matrix used to estimate the linear parameter of the $\mathbb{P}V_j$ term in the variance, and the estimator is denoted by $\widehat{\boldsymbol{\theta}}_t$ (Line 11). Furthermore, $\widehat{\boldsymbol{\theta}}_t$ is also the center of the confidence ellipsoid $\widehat{\mathcal{C}}_j$ (Line 18).

With $\widetilde{\boldsymbol{\theta}}_t$ and $\widehat{\boldsymbol{\theta}}_t$ in hand, for any $t \geq 1$ and $j = j_t$, we can compute the variance estimator by

$$[\widehat{\mathbb{V}}_t V_j](s_t, a_t) = [\langle \boldsymbol{\phi}_{V_j^2}(s_t, a_t), \widetilde{\boldsymbol{\theta}}_{t-1} \rangle]_{[0, B^2]} - (\langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \widehat{\boldsymbol{\theta}}_{t-1} \rangle_{[0, B]})^2,$$

$$E_t = \min\{B^2, 2B\check{\beta}_t \|\boldsymbol{\Sigma}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t)\|_2\} + \min\{B^2, \widetilde{\beta}_t \|\widetilde{\boldsymbol{\Sigma}}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j^2}(s_t, a_t)\|_2\}, \tag{29}$$

where $\widehat{\boldsymbol{\theta}}_0$ and $\widetilde{\boldsymbol{\theta}}_0$ are initialized to be $\mathbf{0}$. We then set $\widehat{\sigma}_t^2 = \max\{B^2/d, [\widehat{\mathbb{V}}_t V_j](s_t, a_t) + E_t\}$, and collecting all the components above yields the Bernstein-type confidence set. Finally, the rest of the procedures (Lines 15 to 20) are the same as those in Algorithm 1.

The Choice of Key Parameters. We set $\{\check{\beta}_t, \widetilde{\beta}_t, \widehat{\beta}_t\}_{t \geq 1}$ as follows:

$$\check{\beta}_t = 8d\sqrt{\log\left(1 + \frac{t}{\lambda}\right)\log\left(\frac{32t^4}{\delta}\right)} + 4\sqrt{d}\log\left(\frac{32t^4}{\delta}\right) + \sqrt{\lambda}d = \widetilde{\mathcal{O}}(d),$$

$$\widetilde{\beta}_t = 8\sqrt{dB^4\log\left(1 + \frac{tB^4}{d\lambda}\right)\log\left(\frac{32t^4}{\delta}\right)} + 4B^2\log\left(\frac{32t^4}{\delta}\right) + \sqrt{\lambda}d = \widetilde{\mathcal{O}}(\sqrt{dB^4}), \quad (30)$$

$$\widehat{\beta}_t = 8\sqrt{d\log\left(1 + \frac{t}{\lambda}\right)\log\left(\frac{32t^4}{\delta}\right)} + 4\sqrt{d}\log\left(\frac{32t^4}{\delta}\right) + \sqrt{\lambda}d = \widetilde{\mathcal{O}}(\sqrt{d}),$$

where we use $\widetilde{\mathcal{O}}(\cdot)$ to hide logarithmic terms.

G.2. Analysis of LEVIS$^+$

Here, we briefly introduce the main steps in establishing the regret upper bound for Algorithm 3. The detailed proof is in Appendix G.3.

First, we have the following result, which is the counterpart of Lemma 8.3. We define a map $j(t)$ such that for any $t \geq 1$, $j = j(t)$ is the index of the value function estimate V_j used for data collection (i.e. line 6 to 14) in step t of Algorithm 3.

Lemma G.1. *There exist choice of $\{\widehat{\beta}_t, \widetilde{\beta}_t, \check{\beta}_t\}_{t \geq 1}$, such that with probability at least $1 - 3\delta$, for all t and $j = j(t)$, the* DEVI *subroutine in Algorithm 3 converges in finite time and the following holds*

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t \cap \mathcal{B}, \quad 0 \leq Q_j(\cdot, \cdot) \leq Q^\star(\cdot, \cdot), \quad \text{and} \quad |\widehat{\mathbb{V}}_t V_j(s_t, a_t) - \mathbb{V}V_j(s_t, a_t)| \leq E_t. \quad (31)$$

Proof of Lemma G.1. See Section G.4. □

Using the Bernstein concentration inequality, we can choose $\{\check{\beta}_t, \widetilde{\beta}_t, \widehat{\beta}_t\}$ as given by (30).

Regret decomposition Compared to the interval decomposition for the Hoeffding case, we add an extra condition for triggering a new interval in the Bernstein case: a new interval start when either of the three conditions is met: (1) the cumulative cost in an interval exceeds $B_\star$; (2) DEVI is triggered; (3) the goal state g is reached. It is easy to check that, the regret decomposition has the same form as that of Lemma 8.2 with the extra condition. To see this, note that in the proof of Lemma 8.2 in Section C, we only require that the true parameter $\boldsymbol{\theta}^*$ is in the confidence sets and all V_j's are upper bounded by $V^\star$. These conditions hold under the event of Lemma G.1. Hence we have the following regret decomposition.

$$R(M) \leq \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - V_{j_m}(s_{m,h})]}_{E_1}$$

$$+ \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}[V_{j_m}(s_{m,h+1}) - \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h})]}_{E_2}$$

$$+ \underbrace{\sum_{m=1}^{M}\left(\sum_{h=1}^{H_m}V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1})\right) - \sum_{m \in \mathcal{M}(M)}V_{j_m}(s_{\text{init}})}_{E_3} + 1. \quad (32)$$

Given the above interval decomposition of the regret, we can get the following theorem which serves as a master theorem similar to Theorem 8.1.

Theorem G.2. *Under Assumption 3.1, 3.2 and 5.1, for any $\delta > 0$, let $\rho = 0$, $\lambda = 1/B^2$ and $\{\check{\beta}_t, \widetilde{\beta}_t, \widehat{\beta}_t\}_{t \geq 1}$ as given by (30). Then with probability at least $1 - 7\delta$,*

$$R(M) = \widetilde{\mathcal{O}} \left(\sqrt{B^2 dT + B^2 d^2 M + B^2 d^{3.5} T^{0.5} + \frac{B^3 d^4}{c_{\min}} M^{0.5}} \right),$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides a term of $C \cdot \log^2(TB/(\lambda \delta c_{\min}))$ for some problem-independent constant C.

Proof of Theorem G.2. Please see Appendix G.5. □

Theorem 6.1 can then be established using Theorem G.2.

G.3. Proof of Theorem 6.1

We first introduce two useful results. The following lemma bound the total number of `DEVI` calls in Algorithm 3.

Lemma G.3. *Let J denote the total number of `DEVI` calls by Algorithm 3. Then on the event of Lemma G.1,*

$$J \leq 2d \log \left(1 + \frac{T}{\lambda} \right) + 2 \log(T).$$

Proof. The proof follows from the same analysis as Lemma D.3, by noting that $\Sigma_T = \lambda \mathbf{I} + \sum_{t=1}^{T} \widehat{\sigma}_t^{-2} \phi_{V_j}(s_t, a_t) \phi_{V_j}(s_t, a_t)^{\top}$ with $\|\phi_{V_j}(s_t, a_t)/\widehat{\sigma}_t\|_2 \leq \sqrt{d}$ since $\widehat{\sigma}_t \geq B/\sqrt{d}$ and $|V_j| \leq B_\star \leq B$ on the event of Lemma G.1. □

The following lemma bounds the total number of intervals M, which is a direct result of the interval decomposition.

Lemma G.4. *Let C_M denote the total cost over M intervals. Then the total number of intervals M satisfies*

$$M \leq \frac{C_M}{B} + K + J.$$

Proof of Lemma G.4. This follows immediately from the three conditions in the interval decomposition (32). □

We are now ready to prove Theorem 6.1.

Proof of Theorem 6.1. Recall that $V^\star(s_{\text{init}})$ denotes the cost of the optimal policy. Then by (3) and Theorem G.2, we have

$$C_M = R(M) + K \cdot V^\star(s_{\text{init}})$$
$$= \widetilde{\mathcal{O}} \left(\sqrt{B^2 dT + B^2 d^2 M + B^2 d^{3.5} T^{0.5} + \frac{B^3 d^4}{c_{\min}} M^{0.5}} \right) + K \cdot V^\star(s_{\text{init}}),$$

where the first step holds since $R(M) = R_K$, and $\widetilde{\mathcal{O}}(\cdot)$ hides a term of $C \cdot \log^2(TB/(\lambda \delta c_{\min}))$. Suppose $K \geq d^5 + B^2 d^4/c_{\min}^2$. Then the above can be simplified to

$$C_M = \widetilde{\mathcal{O}} \left(\sqrt{B^2 dT + B^2 d^2 M} \right) + K \cdot V^\star(s_{\text{init}}).$$

We then have

$$C_M \leq C \log^2 \left(\frac{TB}{\lambda \delta c_{\min}} \right) \cdot \left(\sqrt{B^2 dT + \frac{B^2 d^2 C_M}{B} + B^2 d^2 K + B^2 d^3 \log(\frac{T}{\lambda})} \right) + K \cdot V^\star(s_{\text{init}})$$

$$\leq C \log^2 \left(\frac{TB}{\lambda \delta c_{\min}} \right) \left(d\sqrt{B} \sqrt{C_M} + \sqrt{B^2 dT + B^2 d^2 K + B^2 d^3 \log(\frac{T}{\lambda})} \right) + K \cdot V^\star(s_{\text{init}})$$

$$\leq C' \log^2 \left(\frac{TB}{\lambda \delta c_{\min}} \right) \left(d\sqrt{B} \sqrt{C_M} + \sqrt{B^2 dT + B^2 d^2 K} \right) + K \cdot V^\star(s_{\text{init}}), \tag{33}$$

where the first step is by Lemma G.3 and G.4, the second step is by $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$, and the last step is by $\lambda = 1/B^2$ and hence $B^2 d^3 \log(T/\lambda) = \mathcal{O}(B^2 dT)$ for all $T > K > B^2 d^4$.

We now apply the result that $c \le a\sqrt{c} + b \implies c \le (a + \sqrt{b})^2$ for $a, b \ge 0$, and get from (33) that

$$C_M = \mathcal{O}\left(d^2 B \log^4\left(\frac{TB}{\lambda \delta c_{\min}}\right) + \log^2\left(\frac{TB}{\lambda \delta c_{\min}}\right) \sqrt{B^2 dT + B^2 d^2 K}\right) + K \cdot V^\star(s_{\text{init}}). \tag{34}$$

Furthermore, since $C_M \ge c_{\min} T$, we have

$$c_{\min} \cdot T = \mathcal{O}\left(d^2 B \log^4\left(\frac{TB}{\lambda \delta c_{\min}}\right) + \log^2\left(\frac{TB}{\lambda \delta c_{\min}}\right) \sqrt{B^2 dT + B^2 d^2 K}\right) + K \cdot V^\star(s_{\text{init}}). \tag{35}$$

Applying $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $c \le a\sqrt{c} + b \implies c \le (a + \sqrt{b})^2$ to (35), we can further simplify and get that for all sufficiently large K and T,

$$\begin{aligned} T &= \mathcal{O}\left(\frac{B^2 d}{c_{\min}^2} \log^4\left(\frac{KB}{\lambda \delta c_{\min}}\right) + \frac{Bd\sqrt{K}}{c_{\min}} \log^2\left(\frac{KB}{\lambda \delta c_{\min}}\right)\right) + \frac{K \cdot V^\star(s_{\text{init}})}{c_{\min}} \\ &= \mathcal{O}\left(\frac{KB}{c_{\min}} \log^2\left(\frac{KB}{\lambda \delta c_{\min}}\right)\right), \end{aligned} \tag{36}$$

where the second step is by $V^\star(s_{\text{init}}) \le B_\star \le B$. Plug (36) into (34) and we get that

$$R_K = \mathcal{O}\left(d^2 B\zeta^2 + Bd\sqrt{K}\zeta + B^{1.5}\sqrt{d}\sqrt{\frac{K}{c_{\min}}}\zeta\right), \tag{37}$$

where $\zeta = \log^2(KB/(\lambda \delta c_{\min}))$. This finishes the proof.

$\square$

G.4. Proof of Lemma G.1

We need the following lemma.

Lemma G.5. *For any t and $j = j(t)$, let V_j, $\widehat{\boldsymbol{\theta}}_t$, $\boldsymbol{\sigma}_t$, $\widetilde{\boldsymbol{\theta}}_t$, $\widetilde{\boldsymbol{\Sigma}}_t$ be as given in Algorithm 3. If it further holds that $|V_j| \le B$, then we have*

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

Proof of Lemma G.5. By (29) we can write

$$\left|\widehat{\mathbb{V}}_t V_j(s_t, a_t) - \mathbb{V}V_j(s_t, a_t)\right|$$
$$= \left|[\langle \boldsymbol{\phi}_{V_j^2}(s_t, a_t), \widetilde{\boldsymbol{\theta}}_{t-1}\rangle]_{[0,B^2]} - \langle \boldsymbol{\phi}_{V_j^2}(s_t, a_t), \boldsymbol{\theta}^*\rangle + \langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \boldsymbol{\theta}^*\rangle^2 - (\langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \widehat{\boldsymbol{\theta}}_{t-1}\rangle_{[0,B]})^2\right|$$
$$\le \underbrace{\left|[\langle \boldsymbol{\phi}_{V_j^2}(s_t, a_t), \widetilde{\boldsymbol{\theta}}_{t-1}\rangle]_{[0,B^2]} - \langle \boldsymbol{\phi}_{V_j^2}(s_t, a_t), \boldsymbol{\theta}^*\rangle\right|}_{I_1} + \underbrace{\left|\langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \boldsymbol{\theta}^*\rangle^2 - (\langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \widehat{\boldsymbol{\theta}}_{t-1}\rangle_{[0,B]})^2\right|}_{I_2}.$$

To bound I_1, we have

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

The first step holds because both terms are in $[0, B^2]$: the first term is truncated to be in $[0, B^2]$, and the second term satisfies $|\langle \boldsymbol{\phi}_{V_j^2}(s_t, a_t), \boldsymbol{\theta}^*\rangle| = |\mathbb{P}V_j^2(s_t, a_t)| \le B^2$ since $|V_j| \le B$ by assumption. The second step is by Cauchy-Schwarz inequality. It follows that

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

To bound I_2, we have

$$I_2 = \left| \langle \phi_{V_j}(s_t, a_t), \boldsymbol{\theta}^* \rangle - [\langle \phi_{V_j}(s_t, a_t), \widehat{\boldsymbol{\theta}}_{t-1} \rangle]_{[0,B]} \right| \cdot \left| \langle \phi_{V_j}(s_t, a_t), \boldsymbol{\theta}^* \rangle + [\langle \phi_{V_j}(s_t, a_t), \widehat{\boldsymbol{\theta}}_{t-1} \rangle]_{[0,B]} \right|$$

$$\leq 2B \left| \langle \phi_{V_j}(s_t, a_t), \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t-1} \rangle \right|$$

$$\leq 2B \left\| \Sigma_{t-1}^{1/2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t-1}) \right\|_2 \left\| \Sigma_{t-1}^{-1/2} \phi_{V_j}(s_t, a_t) \right\|_2,$$

where the first inequality holds because both terms are bounded by B, and the second step is by Cauchy-Schwarz inequality. Thus we have

$$I_2 \leq \min \left\{ B^2, 2B \left\| \Sigma_{t-1}^{1/2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t-1}) \right\|_2 \left\| \Sigma_{t-1}^{-1/2} \phi_{V_j}(s_t, a_t) \right\|_2 \right\}.$$

The result follows from combining the bounds of I_1 and I_2. $\qquad\square$

Let's define to sequences of sets

$$\check{\mathcal{C}}_t := \left\{ \boldsymbol{\theta} : \left\| \Sigma_t^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t) \right\|_2 \leq \check{\beta}_t \right\},$$

$$\widetilde{\mathcal{C}}_t := \left\{ \boldsymbol{\theta} : \left\| \widetilde{\Sigma}_t^{1/2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_t) \right\|_2 \leq \widetilde{\beta}_t \right\}.$$

In the following proof, for each $t \geq 1$, we define the confidence ellipsoid $\widehat{\mathcal{C}}_{(t)}$ as

$$\widehat{\mathcal{C}}_{(t)} = \left\{ \boldsymbol{\theta} : \left\| \Sigma_t^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t) \right\|_2 \leq \widehat{\beta}_t \right\},$$

for all $t \geq 1$. Note that the confidence ellipsoid $\widehat{\mathcal{C}}_j$ in Algorithm 3 are indexed by j. By definition, we have $\widehat{\mathcal{C}}_j = \widehat{\mathcal{C}}_{(t_j)}$ for all j, where t_j defined by line 17 is the time step when the j-th DEVI is triggered.

We now prove Lemma G.1.

Proof of Lemma G.1. We prove by an induction argument.

Special case $t = 1$. This is a special case since the variance function estimate used for the step $t = 1$ is V_0, which is from the initialization of Algorithm 3 instead of the DEVI output. So we treat this case separately. From the initialization of Algorithm 3, we have $V_0 = 1 \leq B_\star$. Note that the initial stage $j = 0$ only contains one step $t = 1$. We first show that with probability at least $1 - 3\delta/2$, for $t = 1$ and $j = j(t)$, the event defined by (31) holds (except for the optimism, i.e. $0 \leq Q_j(\cdot, \cdot) \leq Q^\star(\cdot, \cdot)$ and hence $0 \leq V_j(\cdot) \leq V^\star(\cdot)$).

We show $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t$ for $t = 1$ by using Theorem H.3. Let $\mathbf{x}_t = \widehat{\sigma}_t^{-1} \phi_{V_j}(s_t, a_t)$ and $\eta_t = \widehat{\sigma}_t^{-1} V_j(s_{t+1}) - \widehat{\sigma}_t^{-1} \langle \phi_{V_j}(s_t, a_t), \boldsymbol{\theta}^* \rangle = 0$ since V_j is constant for $j = 0$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$, $\mathbf{Z}_t = \lambda \mathbf{I} + \sum_{t'=1}^t \mathbf{x}_{t'} \mathbf{x}_{t'}^\top$, $\mathbf{w}_t = \sum_{t'=1}^t \mathbf{x}_{t'} y_{t'}$ and $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1} \mathbf{w}_t$. Then we have $y_t = \widehat{\sigma}_t^{-1} V_j(s_{t+1})$ and $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_t$. Furthermore, the following holds almost surely:

$$\|\mathbf{x}_t\|_2 \leq \widehat{\sigma}_t^{-1} \cdot 1 \leq \sqrt{d}, \ |\eta_t| = 0, \ \mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0, \ \mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] = 0.$$

Then by Theorem H.3, with probability at least $1 - \delta/2$, for $t = 1$, it holds that

$$\left\| \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t \right\|_{\Sigma_t} \leq 8d\sqrt{\log(1 + t/\lambda)\log(8t^2/\delta)} + 4\sqrt{d}\log(8t^2/\delta) + \sqrt{\lambda d} \leq \check{\beta}_t,$$

which implies $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t$ for $t = 1$.

We then show $\boldsymbol{\theta}^* \in \widetilde{\mathcal{C}}_t$ for $t = 1$ with probability at least $1 - \delta/2$. We apply Theorem H.3 with $\mathbf{x}_t = \phi_{V_j^2}(s_t, a_t)$ and $\eta_t = V_j^2(s_{t+1}) - \langle \phi_{V_j^2}(s_t, a_t), \boldsymbol{\theta}^* \rangle = 0$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$, $\mathbf{Z}_t = \lambda \mathbf{I} + \sum_{t'=1}^t \mathbf{x}_{t'} \mathbf{x}_{t'}^\top$, $\mathbf{w}_t = \sum_{t'=1}^t \mathbf{x}_{t'} y_{t'}$ and $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1} \mathbf{w}_t$. Then we have $y_t = V_j^2(s_{t+1})$ and $\boldsymbol{\mu}_t = \widetilde{\boldsymbol{\theta}}_t$. Furthermore, it holds almost surely that:

$$\|\mathbf{x}_t\|_2 \leq B^2, \ |\eta_t| = 0, \ \mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0, \ \mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] = 0.$$

Then by Theorem H.3, with probability at least $1 - \delta/2$, for $t = 1$, it holds that

$$\left\| \boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}_t \right\|_{\widetilde{\mathbf{\Sigma}}_t} \leq 8\sqrt{dB^4 \log(1 + \frac{tB^4}{d\lambda}) \log(8t^2/\delta)} + 4B^2 \log(8t^2/\delta) + \sqrt{\lambda d} \leq \widetilde{\beta}_t,$$

which implies $\boldsymbol{\theta}^* \in \widetilde{\mathcal{C}}_t$.

We then show $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{(t)}$ for $t = 1$ with probability at least $1 - \delta/2$. Let $\mathbf{x}_t = \widehat{\sigma}_t^{-1} \boldsymbol{\phi}_{V_j}(s_t, a_t)$ and $\eta_t = \widehat{\sigma}_t^{-1} \left[V_j(s_{t+1}) - \langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \boldsymbol{\theta}^* \rangle \right]$. Let $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$, $\mathbf{Z}_t = \lambda \boldsymbol{I} + \sum_{t'=1}^t \mathbf{x}_{t'} \mathbf{x}_{t'}^\top$, $\mathbf{w}_t = \sum_{t'=1}^t \mathbf{x}_{t'} y_{t'}$ and $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1} \mathbf{w}_t$. Then we have $y_t = \widehat{\sigma}_t^{-1} V_j(s_{t+1})$ and $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_t$. Furthermore, the following holds almost surely:

$$\|\mathbf{x}_t\|_2 \leq \widehat{\sigma}_t^{-1} \cdot 1 \leq \sqrt{d}, \ |\eta_t| = 0, \ \mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0, \ \mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] = 0.$$

Then by Theorem H.3, with probability at least $1 - \delta/2$, for $t = 1$, it holds that

$$\left\| \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t \right\|_{\mathbf{\Sigma}_t} \leq 8\sqrt{d \log(1 + t/\lambda) \log(8t^2/\delta)} + 4\sqrt{d} \log(8t^2/\delta) + \sqrt{\lambda d} \leq \widehat{\beta}_t,$$

which implies $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{(t)}$ for $t = 1$. By a union bound, we get that, with probability at least $1 - 3\delta/2$, for $t = 1$, we have

$$\theta^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t \cap \widehat{\mathcal{C}}_{(t)} \cap \mathcal{B}.$$

It remains to show $\left| \widehat{\mathbb{V}}_t V_j(s_t, a_t) - \mathbb{V} V_j(s_t, a_t) \right| \leq E_t$. But this is trivial since $\widehat{\mathbb{V}} V_j = 0$ by the initialization of $\widetilde{\boldsymbol{\theta}}_t$ and $\widehat{\boldsymbol{\theta}}_t$, and $\mathbb{V} V_j = 0$ by $V_j(\cdot) = 0$, for $t = 1$ and $j = j(t) = 0$.

Note that by the time step doubling criterion, at the end of step $t = 1$, DEVI would be triggered and output the value function estimate V_j for $j = 1$. We now show that under the above event, the optimism holds for $j = 1$, i.e. $0 \leq Q_j \leq Q^\star$ and hence $0 \leq V_j \leq V^\star$. We prove by applying another induction argument on the loop of DEVI. For the base step, note that by non-negativity of $Q^\star$ and $V^\star$, we have $Q^{(0)} \leq Q^\star$ and $V^{(0)} \leq V^\star$. For the induction hypothesis, we assume $Q^{(i)}$ and $V^{(i)}$ are optimistic for some i. We want to show $Q^{(i+1)}$ and $V^{(i+1)}$ are optimistic. For the $(i+1)$-th iteration, we have

$$
\begin{aligned}
Q^{(i+1)}(\cdot, \cdot) &= c(\cdot, \cdot) + (1 - q) \cdot \min_{\boldsymbol{\theta} \in \widehat{\mathcal{C}}_1 \cap \mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(i)}}(\cdot, \cdot) \rangle \\
&\leq c(\cdot, \cdot) + (1 - q) \cdot \mathbb{P} V^{(i)}(\cdot, \cdot) \\
&\leq c(\cdot, \cdot) + \mathbb{P} V^{(i)}(\cdot, \cdot) \\
&\leq Q^\star(\cdot, \cdot),
\end{aligned} \tag{38}
$$

where the first step is because we are considering the case where $\rho = 0$, the second step is because we are taking the minimum over a set that contains $\boldsymbol{\theta}^*$, the third step is by non-negativity of $\mathbb{P} V^{(i)}(\cdot, \cdot)$, and the last step is by the Bellman optimal condition (2) and the induction hypothesis that $V^{(i)}$ is optimistic. By induction, we conclude that $Q^{(i)}$ is optimistic for all i, and thus the final output $Q_1(\cdot, \cdot)$ and thus $V_1(\cdot)$ are both optimistic.

Initial step $t \in [t_1 + 1, t_2]$. Recall that in this round, the value function estimate that is used to collect the data is V_1, which satisfies $V_1 \leq B_\star$ by the optimism proved above.

We first show that with probability at least $1 - \frac{3\delta}{t_2(t_2+1)}$, for all $t \in [t_1 + 1, t_2]$, the event in Lemma G.1 holds. The proof is actually similar to that of the $t = 1$ case. We show $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t$ by using Theorem H.3. Let $\mathbf{x}_t = \widehat{\sigma}_t^{-1} \boldsymbol{\phi}_{V_j}(s_t, a_t)$ and $\eta_t = \widehat{\sigma}_t^{-1} V_j(s_{t+1}) - \widehat{\sigma}_t^{-1} \langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \boldsymbol{\theta}^* \rangle$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$, $\mathbf{Z}_t = \lambda \boldsymbol{I} + \sum_{t'=1}^t \mathbf{x}_{t'} \mathbf{x}_{t'}^\top$, $\mathbf{w}_t = \sum_{t'=1}^t \mathbf{x}_{t'} y_{t'}$ and $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1} \mathbf{w}_t$. Then we have $y_t = \widehat{\sigma}_t^{-1} V_j(s_{t+1})$ and $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_t$, and the following holds almost surely:

$$\|\mathbf{x}_t\|_2 \leq \widehat{\sigma}_t^{-1} B_\star \leq \sqrt{d}, \ |\eta_t| \leq \sqrt{d}, \ \mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0, \ \mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] \leq d,$$

where the first and the second inequalities hold because $V_j \leq B_\star \leq B$ for $t \in [t_1 + 1, t_2]$ and thus $\|\boldsymbol{\phi}_{V_j}\| \leq B_\star$. By Theorem H.3, with probability at least $1 - \frac{\delta}{t_2(t_2+1)}$, for all $t \in [t_1 + 1, t_2]$, it holds that

$$\left\| \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t \right\|_{\mathbf{\Sigma}_t} \leq 8d\sqrt{\log(1 + t/\lambda) \log(32t^4/\delta)} + 4\sqrt{d} \log(32t^4/\delta) + \sqrt{\lambda d} = \check{\beta}_t,$$

where we use the fact that $4t^2 t_2(t_2 + 1) \leq 32t^4$ for $t \in [t_1 + 1, t_2]$, since $t_2 \leq 2t_1 \leq 2t$ by the doubling time step criterion. Thus we conclude that with the stated probability, $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t$.

To show $\boldsymbol{\theta}^* \in \widetilde{\mathcal{C}}_t$, we apply Theorem H.3 with $\mathbf{x}_t = \boldsymbol{\phi}_{V_j^2}(s_t, a_t)$ and $\eta_t = V_j^2(s_{t+1}) - \langle \boldsymbol{\phi}_{V_j^2}(s_t, a_t), \boldsymbol{\theta}^* \rangle$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$, $\mathbf{Z}_t = \lambda \boldsymbol{I} + \sum_{t'=1}^t \mathbf{x}_{t'} \mathbf{x}_{t'}^\top$, $\mathbf{w}_t = \sum_{t'=1}^t \mathbf{x}_{t'} y_{t'}$ and $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1} \mathbf{w}_t$. Then we have $y_t = V_j^2(s_{t+1})$ and $\boldsymbol{\mu}_t = \widetilde{\boldsymbol{\theta}}_t$, and the following holds almost surely:

$$\|\mathbf{x}_t\|_2 \leq B^2, \ |\eta_t| \leq B^2, \ \mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0, \ \mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] \leq B^4.$$

Then by Theorem H.3, with probability at least $1 - \frac{\delta}{t_2(t_2+1)}$, for all $t \in [t_1 + 1, t_2]$, it holds that

$$\left\| \boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}_t \right\|_{\widetilde{\boldsymbol{\Sigma}}_t} \leq 8\sqrt{dB^4 \log(1 + \frac{tB^4}{d\lambda}) \log(32t^4/\delta)} + 4B^2 \log(32t^4/\delta) + \sqrt{\lambda d} = \widetilde{\beta}_t.$$

To show $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{(t)}$, we apply Theorem H.3 with $\mathbf{x}_t = \widehat{\sigma}_t^{-1} \boldsymbol{\phi}_{V_j}(s_t, a_t)$ and

$$\eta_t = \widehat{\sigma}_t^{-1} \left[V_j(s_{t+1}) - \langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \boldsymbol{\theta}^* \rangle \right] \mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\},$$

$\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$, $\mathbf{Z}_t = \lambda \boldsymbol{I} + \sum_{t'=1}^t \mathbf{x}_{t'} \mathbf{x}_{t'}^\top$, $\mathbf{w}_t = \sum_{t'=1}^t \mathbf{x}_{t'} y_{t'}$ and $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1} \mathbf{w}_t$. Then we have $y_t = \widehat{\sigma}_t^{-1} V_j(s_{t+1})$ and $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_t$. Furthermore, it holds almost surely that

$$
\begin{aligned}
&\mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] \\
&= \widehat{\sigma}_t^{-2} \mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}[\mathbb{V}V_j](s_t, a_t) \\
&\leq \widehat{\sigma}_t^{-2} \mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\} \Big[\widehat{\mathbb{V}}_t V_j(s_t, a_t) + \min\Big\{B^2, \Big\| \widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}_t) \Big\|_2 \Big\| \widetilde{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{V_j^2}(s_t, a_t) \Big\|_2 \Big\} \\
&\qquad + \min\Big\{B^2, 2B \Big\| \boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t) \Big\|_2 \Big\| \boldsymbol{\Sigma}_t^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t) \Big\|_2 \Big\} \Big] \\
&\leq \widehat{\sigma}_t^{-2} \Big[\widehat{\mathbb{V}}_t V_j(s_t, a_t) + \min\Big\{B^2, \widetilde{\beta}_t \Big\| \widetilde{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{V_j^2}(s_t, a_t) \Big\|_2 \Big\} + \min\Big\{B^2, 2B\check{\beta}_t \Big\| \boldsymbol{\Sigma}_t^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t) \Big\|_2 \Big\} \Big] \\
&= 1,
\end{aligned}
$$

where the first inequality holds due to Lemma G.5 and $V_j \leq B$, the second inequality is due to the event in the indicator function, and the last step in by the definition of $\widehat{\sigma}_t$. We then use Theorem H.3 and get that with probability at least $1 - \frac{\delta}{t_2(t_2+1)}$, for all $t \in [t_1 + 1, t_2]$,

$$\left\| \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t \right\|_{\boldsymbol{\Sigma}_t} \leq 8\sqrt{d \log(1 + t/\lambda) \log(32t^4/\delta)} + 4\sqrt{d} \log(32t^4/\delta) + \sqrt{\lambda d} = \widehat{\beta}_t. \tag{39}$$

Denote by $\mathcal{E}_1'$ the event where $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t$ and (39) holds for all $t \in [t_1 + 1, t_2]$. Then on this event, we have

$$y_t = \langle \boldsymbol{\theta}^*, \widehat{\sigma}_t^{-1} \boldsymbol{\phi}_{V_j}(s_t, a_t) \rangle + \widehat{\sigma}_t^{-1} \left[V_j(s_{t+1}) - \langle \boldsymbol{\phi}_{V_j}(s_t, a_t), \boldsymbol{\theta}^* \rangle \right] = \widehat{\sigma}_t^{-1} V_j(s_{t+1}).$$

Therefore, the above shows that $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{(t)}$ for all $t \in [t_1 + 1, t_2]$ holds on the event $\mathcal{E}_1'$. Finally, by union bound, we get that with probability at least $1 - \frac{3\delta}{t_2(t_2+1)}$, for all $t \in [t_1 + 1, t_2]$, it holds that

$$\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t \cap \widehat{\mathcal{C}}_{(t)} \cap \mathcal{B}.$$

To show that the optimism holds under the above event, i.e. $0 \leq Q_j \leq Q^\star$ and $0 \leq V_j \leq V^\star$ for $j = 2$, note that this is the same as the proof for the $j = 1$ case. Indeed, the same induction trick on the loop of DEVI can be applied with the set $\widehat{\mathcal{C}}_{(t_1)}$ in (38) being replaced by $\widehat{\mathcal{C}}_{(t_2)}$.

To show $\left| \widehat{\mathbb{V}}_t V_j(s_t, a_t) - \mathbb{V}V_j(s_t, a_t) \right| \leq E_t$ under the above event, we apply the definition of E_t, Lemma G.5 and the fact that $V_1 \leq B_\star \leq B$.

From the case $t = 1$ and $t \in [t_1 + 1, t_2]$, we get that, with probability at least $1 - \frac{3\delta}{t_1(t_1+1)} - \frac{3\delta}{t_2(t_2+1)}$, for all $t \in [1, t_2]$ and $j = j(t)$, the following event holds

$$\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t \cap \widehat{\mathcal{C}}_{(t)} \cap \mathcal{B}, \quad 0 \le Q_j(\cdot, \cdot) \le Q^\star(\cdot, \cdot), \quad \left| \widehat{\mathbb{V}}_t V_j(s_t, a_t) - \mathbb{V} V_j(s_t, a_t) \right| \le E_t. \tag{40}$$

Induction step. Suppose that, with probability at least $1 - \delta'$, for some $j - 1 \ge 2$ and $t \in [1, t_{j-1}]$, the event in (40) holds. We want to show (40) also holds for j and $t \in [t_{j-1} + 1, t_j]$, with probability at least $1 - \delta' - \frac{3\delta}{t_j(t_j+1)}$. However, this immediately follows from the exact analysis of the initial step $t \in [t_1 + 1, t_2]$. Indeed, all we need is V_{j-1} is an optimistic estimate of $V_\star$, which holds by the induction hypothesis, and the probability comes from a union bound.

Finally, we conclude by mathematical induction that the event (40) holds with the probability at least the following:

$$1 - \sum_{j=1}^{J} \frac{3\delta}{t_j(t_j + 1)} = 1 - 3\delta \sum_{j=1}^{J} \left(\frac{1}{t_j} - \frac{1}{t_j + 1} \right),$$

which is lower bounded by $1 - 3\delta$. This implies the event (31) holds with probability at least $1 - 3\delta$, since the event (40) is a subset of the event (31). Furthermore, the finite time convergence of `DEVI` follows from the contraction property. This completes the proof. $\qquad \square$

G.5. Proof of Theorem G.2

We bound E_1, E_2 and E_3 separately.

G.5.1. BOUNDING E_1

Following the same reasoning as in Section D.1, we write E_1 as

$$E_1 = \sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h}) \right],$$

where

$$c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})$$
$$\le \langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle + \frac{B_\star + 1 - q}{t_{j_m}},$$

where we use the optimism $V_{j_m} \le V^\star \le B_\star$ under the event of Lemma G.1, and $q = 1/t_{j_m}$ according to Algorithm 3.

Recall the definition of $\mathcal{M}_0(M)$ being the set of m such that $j_m \ge 1$, i.e., $\mathcal{M}_0(M) = \{m \le M : j_m \ge 1\}$. Then we use the following

$$\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \left[c_{m,h} + \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h}) \right]$$
$$\le \underbrace{\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle}_{A_1} + (B_\star + 1) \cdot \underbrace{\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \frac{1}{t_{j_m}}}_{A_2}. \tag{41}$$

To bound A_1: Recall that $\widehat{\boldsymbol{\theta}}_{t_{j_m}}$ given by Line 12 is the center of the confidence ellipsoid $\mathcal{C}_{j_m}$. First for each term $\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle$ in A_1, we write

$$\langle \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{j_m} + \widehat{\boldsymbol{\theta}}_{j_m} - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle$$
$$\le 4\widehat{\beta}_T \| \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}}$$
$$= 4\widehat{\beta}_T \| \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) / \widehat{\sigma}_{t(m,h)} \|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}} \cdot \widehat{\sigma}_{t(m,h)}, \tag{42}$$

where the inequality follows from the same reasoning as in (11), with the confidence ellipsoids $\mathcal{C}_{j_m}$ being replaced by $\widehat{\mathcal{C}}_{j_m}$, and β_T replaced by $\widehat{\beta}_T$. Also note that for each term $\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle$ in A_1, we have

$$
\begin{aligned}
\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle &\leq \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h}) \rangle \\
&= \mathbb{P}V_{j_m}(s_{m,h}, a_{m,h}) \\
&\leq B_\star,
\end{aligned}
\tag{43}
$$

where both inequalities hold because $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_{m,h}$ are parameters of some transition kernels, and $0 \leq V_{j_m}(\cdot) \leq B_\star$ under the event of Lemma G.1.

Combining (42) and (43), we can bound A_1 as

$$
\begin{aligned}
A_1 &\leq \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min \left\{ B_\star, 4\widehat{\beta}_T \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})/\widehat{\sigma}_{t(m,h)}\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}} \widehat{\sigma}_{t(m,h)} \right\} \\
&\leq \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \left(B_\star + 4\widehat{\beta}_T \widehat{\sigma}_{t(m,h)} \right) \min \left\{ 1, \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})/\widehat{\sigma}_{t(m,h)}\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}} \right\} \\
&\leq \sqrt{\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \left(B_\star + 4\widehat{\beta}_T \widehat{\sigma}_{t(m,h)} \right)^2} \cdot \sqrt{\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min \left\{ 1, \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})/\widehat{\sigma}_{t(m,h)}\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}}^2 \right\}},
\end{aligned}
\tag{44}
$$

where the second inequality holds since $\min\{a_1 a_2, b_1 b_2\} \leq (a_1 + b_1) \min\{a_2, b_2\}$ for $a_1, a_2, b_1, b_2 > 0$, and the third inequality is by Cauchy-Schwarz inequality. To further bound the R.H.S. of (44), note that

$$
\begin{aligned}
&\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min \left\{ 1, \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})/\widehat{\sigma}_{t(m,h)}\|_{\boldsymbol{\Sigma}_{t(m,h)}^{-1}}^2 \right\} \\
&\leq 2 \left[d \log \left(\frac{\text{trace}(\lambda \mathbf{I}) + T \cdot \max_{m,h} \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})/\widehat{\sigma}_{t(m,h)}\|_2^2}{d} \right) - \log (\det(\lambda \mathbf{I})) \right] \\
&\leq 2d \log \left(\frac{\lambda d + Td}{\lambda d} \right) \\
&= 2d \log \left(1 + T/\lambda \right),
\end{aligned}
\tag{45}
$$

where the first inequality holds by Lemma H.5, and the second inequality holds because $V_{j_m}(\cdot) \leq B_\star$ under Lemma G.1, $\widehat{\sigma}_t \leq B/\sqrt{d}$ by Line 8, and thus $\max_{m,h} \|\boldsymbol{\phi}_{V_{j_m}}(s_{m,h}, a_{m,h})/\widehat{\sigma}_{t(m,h)}\|_2 \leq \sqrt{d}$ by Assumption 3.1. Furthermore, we have

$$
\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \left(B_\star + 4\widehat{\beta}_T \widehat{\sigma}_{t(m,h)} \right)^2 \leq 2TB_\star^2 + 32\widehat{\beta}_T^2 \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \widehat{\sigma}_{t(m,h)}^2,
\tag{46}
$$

and

$$
\begin{aligned}
&\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \widehat{\sigma}_{t(m,h)}^2 \\
&= \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \max \left\{ B^2/d, \ \widehat{\mathbb{V}}_t V_{j_m}(s_t, a_t) + E_t \right\} \\
&= \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \max \left\{ B^2/d, \ \mathbb{V}_t V_{j_m}(s_t, a_t) + 2E_t + \widehat{\mathbb{V}}_t V_{j_m}(s_t, a_t) - \mathbb{V}V_{j_m}(s_t, a_t) - E_t \right\} \\
&\leq \frac{B^2 T}{d} + 2 \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} E_t + \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \mathbb{V}V_{j_m}(s_t, a_t) + \underbrace{\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \widehat{\mathbb{V}}_t V_{j_m}(s_t, a_t) - E_t - \mathbb{V}V_{j_m}(s_t, a_t)}_{\leq 0},
\end{aligned}
\tag{47}
$$

where we write $t = t(m, h)$ for simplicity. Note that the last term is at most zero under the event of Lemma G.1.

For the second term $\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} E_t$, by (29), we have

$$
\begin{aligned}
&\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} E_t \\
&= \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min\{B^2, 2B\check{\beta}_t \widehat{\sigma}_t \|\boldsymbol{\Sigma}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t)/\widehat{\sigma}_t\|_2\} + \min\{B^2, \widetilde{\beta}_t \|\widetilde{\boldsymbol{\Sigma}}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j^2}(s_t, a_t)\|_2\} \\
&\leq 2B \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \check{\beta}_t \widehat{\sigma}_t \min\{1, \|\boldsymbol{\Sigma}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t)/\widehat{\sigma}_t\|_2\} \\
&\quad + \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \widetilde{\beta}_t \min\{1, \|\widetilde{\boldsymbol{\Sigma}}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j^2}(s_t, a_t)\|_2\} \\
&\leq 2\sqrt{3}B^2 \check{\beta}_T \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min\{1, \|\boldsymbol{\Sigma}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t)/\widehat{\sigma}_t\|_2\} \\
&\quad + \widetilde{\beta}_T \sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min\{1, \|\widetilde{\boldsymbol{\Sigma}}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j^2}(s_t, a_t)\|_2\},
\end{aligned}
\tag{48}
$$

where the first inequality holds since $\check{\beta}_t \widehat{\sigma}_t \geq B$ and $\widetilde{\beta}_t \geq B^2$, and the second inequality holds since $\check{\beta}_t \leq \check{\beta}_T$ and $\widehat{\sigma}_t \leq \sqrt{3}B$ by $\widehat{\sigma}_t^2 \leftarrow \max\{B^2/d, [\widehat{\mathbb{V}}_t V_j](s_t, a_t) + E_t\}$ and (29).

$$
\begin{aligned}
\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min\{1, \|\boldsymbol{\Sigma}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t)/\widehat{\sigma}_t\|_2\} &\leq \sqrt{T} \sqrt{\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min\{1, \|\boldsymbol{\Sigma}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t)/\widehat{\sigma}_t\|_2^2\}} \\
&\leq \sqrt{2dT \log\left(1 + T(B_\star \sqrt{d}/B)^2/(d\lambda)\right)} \\
&= \sqrt{2dT \log\left(1 + T/\lambda\right)},
\end{aligned}
\tag{49}
$$

where the first inequality is by Cauchy Schwarz inequality, and the second inequality is by Lemma H.5 and $\|\boldsymbol{\phi}_{V_j}(s_t, a_t)/\widehat{\sigma}_t\|_2 \leq B_\star/(B/\sqrt{d})$, since $\|\boldsymbol{\phi}_{V_j}\| \leq B_\star$ under the event of Lemma G.1 by Assumption 3.1 and $\widehat{\sigma}_t \geq B/\sqrt{d}$ by definition. Similarly, we have

$$
\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \min\{1, \|\widetilde{\boldsymbol{\Sigma}}_{t-1}^{-1/2} \boldsymbol{\phi}_{V_j^2}(s_t, a_t)\|_2\} \leq \sqrt{2dT \log\left(1 + TB_\star^4/(d\lambda)\right)}.
\tag{50}
$$

Combining (48), (49) and (50), we get

$$
\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} E_t \leq 5B^2 \check{\beta}_T \sqrt{dT \log(1 + T/\lambda)} + \sqrt{2} \widetilde{\beta}_T \sqrt{dT \log(1 + TB_\star^4/(d\lambda))}.
\tag{51}
$$

For the term $\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \mathbb{V} V_{j_m}(s_t, a_t)$, note that a trivial upper bound is $B_\star^2 T$ since $|V_{j_m}| \leq B_\star$ by the optimism. However, such bound is loose for the regret analysis. It turns out that we can use a total variance trick to bound the term by roughly $\mathcal{O}(B^2 M)$, where M is the number of intervals. This is summarized by Lemma G.6.

For any m, we define the event $\mathcal{E}_m$ as

$$
\mathcal{E}_m := \big\{ \text{For all } m' \leq m : \boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{j_{m'}} \cap \mathcal{B}, \quad 0 \leq Q_{j_{m'}}(\cdot, \cdot) \leq Q^\star(\cdot, \cdot), \quad \text{and} \quad 0 \leq V_{j_{m'}}(\cdot) \leq V^\star(\cdot) \big\}.
\tag{52}
$$

By the definition, it is clear that the event of Lemma G.1 is a subset of $\mathcal{E}_M$. Also we have $\mathcal{E}_m \subseteq \mathcal{E}_{m'}$ for any $m > m'$.

Lemma G.6 (Total variance bound). *With probability at least $1 - 3\delta$, it holds that*

$$\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \mathbb{V} V_{j_m}(s_t, a_t) \mathbb{1}\{\mathcal{E}_m\}$$

$$\leq 18B^2 M + \frac{338 B^3 d \widehat{\beta}_T^2}{c_{\min}} \left(\sqrt{M} \sqrt{\log(1/\delta)} \log \left(1 + \frac{2B}{c_{\min} \lambda} \right) + \log(1 + T/\lambda) \right).$$

The proof of Lemma G.6 is in Appendix G.6.

Combining (47), (51) and Lemma G.6, we can bound $\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \widehat{\sigma}_{t(m,h)}^2$ by

$$\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \widehat{\sigma}_{t(m,h)}^2$$

$$\leq \frac{B^2 T}{d} + 18B^2 M + 10B^2 \check{\beta}_T \sqrt{dT \log(1 + T/\lambda)} + 3\widetilde{\beta}_T \sqrt{dT \log(1 + TB_\star^4/(d\lambda))}$$

$$+ \frac{338 B^3 d \widehat{\beta}_T^2}{c_{\min}} \left(\sqrt{M} \sqrt{\log(1/\delta)} \log \left(1 + \frac{2B}{c_{\min} \lambda} \right) + \log(1 + T/\lambda) \right), \tag{53}$$

with probability at least $1 - 6\delta$ by a union bound over the event of Lemma G.1 and Lemma G.6. Plugging (53) into (46), and then using (44), (45) and (30), we conclude that A_1 can be bounded as

$$A_1 = \widetilde{\mathcal{O}} \left(\sqrt{B^2 dT + B^2 d^2 M + B^2 d^{3.5} T^{0.5} + \frac{B^3 d^4}{c_{\min}} M^{0.5}} \right), \tag{54}$$

where we use $\widetilde{\mathcal{O}}(\cdot)$ to hide a term polynomial in $\log^2(TB/(d\lambda\delta))$.

To bound A_2: By rewriting the summation using the index j, it immediately follows from Lemma G.3 that

$$A_2 \leq (B_\star + 1) \sum_{j=0}^{J} \sum_{t=t_j+1}^{t_{j+1}} \frac{1}{t_j} \leq 2(J+1) = 4d \log(1 + T/\lambda) + 2 \log(T) < A_1.$$

Furthermore, since DEVI is called at $t = 2t_0 = 2$ by the time step doubling condition, we have

$$\sum_{m \in \mathcal{M}_0^c} \sum_{h=1}^{H_m} [c_{m,h} + \mathbb{P} V_{j_m}(s_{m,h}, a_{m,h}) - Q_{j_m}(s_{m,h}, a_{m,h})] = \sum_{h=1}^{2} [c_{1,h} + \mathbb{P} V_0(s_{1,h}, a_{1,h}) - Q_0(s_{1,h}, a_{1,h})]$$

$$\leq 4,$$

where the inequality holds because $c_{1,h}, V_0(\cdot) \leq 1$ and $0 \leq Q_0(\cdot, \cdot)$. Together with (41) and (54), we conclude that

$$E_1 = \widetilde{\mathcal{O}} \left(\sqrt{B^2 dT + B^2 d^2 M + B^2 d^{3.5} T^{0.5} + \frac{B^3 d^4}{c_{\min}} M^{0.5}} \right). \tag{55}$$

G.5.2. BOUNDING E_2 AND E_3

To bound E_2, by the same reasoning as in Lemma D.2, we have that, with probability at least $1 - 4\delta$, the event of Lemma G.1 holds and

$$E_2 \leq 2B_\star \sqrt{2T \log \left(\frac{T}{\delta} \right)} = \widetilde{\mathcal{O}} \left(B_\star \sqrt{T} \right). \tag{56}$$

Here the $1 - 4\delta$ probability comes from a union bound of Lemma G.1 and Azuma-Hoeffding inequality.

To bound E_3, following the proof of Lemma D.4 and (17), we can write

$$\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right) \leq \sum_{m=1}^{M-1} \left(V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1}) \right) + V_{j_1}(s_{1,1}). \tag{57}$$

Now consider the term $V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1})$. Note that by the interval decomposition, interval m ends if and only if either of the three conditions are met. If interval m ends because goal is reached, then we have

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

If it ends because the `DEVI` sub-routine is triggered, then the value function estimator is updated by `DEVI` and $j_m \neq j_{m+1}$. In such case we simply apply the trivial upper bound $V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1}) \leq \max_j \|V_j\|_\infty$. By Lemma G.3, this happens at most $J \leq 2d \log\left(1 + \frac{T}{\lambda}\right) + 2\log(T)$ times. If the interval ends because the cumulative cost reaches B, then $s_{m+1,1} = s_{m,H_m+1}$ and $V_{j_m} = V_{j_{m+1}}$ and hence $V_{j_{m+1}}(s_{m+1,1}) - V_{j_m}(s_{m,H_m+1}) = 0$. Therefore, we can bound the RHS of (57) as

$$\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j_m}(s_{m,h}) - V_{j_m}(s_{m,h+1}) \right)$$
$$\leq \sum_{m=1}^{M-1} V_{j_{m+1}}(s_{\text{init}}) \cdot \mathbb{1}\{m+1 \in \mathcal{M}(M)\} + V_{j_1}(s_{1,1}) + \left[2d \log\left(1 + \frac{T}{\lambda}\right) + 2\log(T) \right] \cdot \max_j \|V_j\|_\infty$$
$$\leq \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}}) + V_0(s_{\text{init}}) + 2dB_\star \log\left(1 + \frac{T}{\lambda}\right) + 2B_\star \log(T)$$
$$\leq \sum_{m \in \mathcal{M}(M)} V_{j_m}(s_{\text{init}}) + 1 + 2dB_\star \log\left(1 + \frac{T}{\lambda}\right) + 2B_\star \log(T),$$

where the second inequality is by $\|V_j\|_\infty \leq B_\star$ and the last step is by the initialization $\|V_0\|_\infty \leq 1$. Rearranging the terms, we conclude that under the event of Lemma G.1,

$$E_3 \leq 1 + 2dB_\star \log\left(1 + \frac{T}{\lambda}\right) + 2B_\star \log(T) = \widetilde{\mathcal{O}}(Bd). \tag{58}$$

G.5.3. BOUNDING $R(M)$

Combining (32), (55), (56) and (58), we get the final bound for $R(M)$:

$$R(M) = \widetilde{\mathcal{O}}\left(\sqrt{B^2 dT + B^2 d^2 M + B^2 d^{3.5} T^{0.5} + \frac{B^3 d^4}{c_{\min}} M^{0.5}} \right),$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides a term of $C \cdot \log^2(TB/(\lambda \delta c_{\min}))$ for some problem-independent constant C. This holds with probability at least $1 - 7\delta$ by a union bound over the event of Lemma G.1, Lemma G.6 and (56).

G.6. Proof of Lemma G.6

In this section, for any interval m, we define F_m as the trajectory until the end of the interval m, i.e.,

$$F_m = \cup_{m'=1}^{m} \{ s_{t(m',1)}, a_{t(m',1)}, \cdots, s_{t(m',H_{m'})}, a_{t(m',H_{m'})}, s_{t(m',H_{m'}+1)} \}.$$

We first prove the following result which bounds the expected sum of variance for an arbitrary interval by using the technique from Azar et al. (2017).

Lemma G.7. *For any interval m,*

$$\mathbb{E}\left[\sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t)\,\mathbb{1}\{\mathcal{E}_m\}\,\middle|\, F_{m-1}\right] \leq 18B^2 + 24\xi,$$

$$\xi = 24\frac{B}{c_{\min}}\mathbb{E}\left[\sum_{h=1}^{H_m}\left(\widehat{\beta}_t^2\|\phi_{V_j}(s_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}}^2 + \frac{2B_\star^2}{t_j^2}\right)\mathbb{1}\{\mathcal{E}_m\}\,\middle|\, F_{m-1}\right].$$

To prove Lemma G.7, we need the following lemma.

Lemma G.8 (Lemma B.15 in Cohen et al. 2020). *Let $\{X_t\}_{t=1}^\infty$ be a martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^\infty$, such that X_t is $\mathcal{F}_t$-measurable. Let $Y_n = (\sum_{t=1}^n X_t)^2 - \sum_{t=1}^n \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}]$. Then $\{Y_n\}_{n=0}^\infty$ is a martingale. If we further assume H is a stopping time such that $H < C$ for some fixed C almost surely, then $\mathbb{E}[Y_H] = 0$.*

Proof of Lemma G.7. We define $X_{m,h} = [\mathbb{P}V_j(s_t, a_t) - V_j(s_{t+1})]\,\mathbb{1}\{\mathcal{E}_m\}$, where $t = t(m, h)$ and $j = j_m$ as a simplified notation. Then conditioned on F_{m-1}, $\mathbb{1}\{\mathcal{E}_m\}$ is determined, and thus $\{X_{m,h}\}_{h=1}^\infty$ is a martingale difference sequence with respect to $\{F_{m,h}\}_{h=1}^\infty$, where $F_{m,h}$ is the trajectory from the from the beginning of interval m to time h. Furthermore, H_m is a stopping time w.r.t. the trajectory which is upper bounded by $2B_\star/c_{\min}$ since by the interval decomposition a new interval would start if the cumulative cost exceeds $B_\star$. Therefore, we have

$$\mathbb{E}\left[\sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t)\,\mathbb{1}\{\mathcal{E}_m\}\,\middle|\, F_{m-1}\right] := \mathbb{E}\left[\sum_{h=1}^{H_m} (\mathbb{P}V_j(s_t, a_t) - V_j(s_{t+1}))^2\,\mathbb{1}\{\mathcal{E}_m\}\,\middle|\, F_{m-1}\right]$$

$$= \mathbb{E}\left[\sum_{h=1}^{H_m} (X_{m,h})^2\,\middle|\, F_{m-1}\right]$$

$$= \mathbb{E}\left[\left(\sum_{h=1}^{H_m} X_{m,h}\right)^2\,\middle|\, F_{m-1}\right]$$

$$= \mathbb{E}\left[\left(\sum_{h=1}^{H_m} [\mathbb{P}V_j(s_t, a_t) - V_j(s_{t+1})]\right)^2\,\mathbb{1}\{\mathcal{E}_m\}\,\middle|\, F_{m-1}\right], \tag{59}$$

where the third step is by Lemma G.8.

By Assumption 3.1, we can write $\mathbb{P}V_j(s_t, a_t) = \langle \boldsymbol{\theta}^*, \phi_{V_j}(s_t, a_t)\rangle$. Furthermore, since V_j and Q_j are the output of DEVI, we can write $Q_j = Q^{(l)}$ and $V_j(\cdot) = \min_a Q^{(l)}(\cdot, a)$ for some l, which denote the l-th iteration in the implementation of DEVI. It follows that

$$Q^{(l)}(s_t, a_t) = c(s_t, a_t) + (1 - q_j)\min_{\boldsymbol{\theta} \in \widehat{\mathcal{C}}_{j_m} \cap \mathcal{B}} \langle \boldsymbol{\theta}, \phi_{V^{(l-1)}}(s_t, a_t)\rangle$$

$$= c_t + (1 - q_j)\langle \boldsymbol{\theta}_{m,h}, \phi_{V^{(l)}}(s_t, a_t)\rangle + (1 - q_j)\langle \boldsymbol{\theta}_{m,h}, \phi_{V^{(l-1)}}(s_t, a_t) - \phi_{V^{(l)}}(s_t, a_t)\rangle$$

$$= c_t + (1 - q_j)\langle \boldsymbol{\theta}_{m,h}, \phi_{V^{(l)}}(s_t, a_t)\rangle + (1 - q_j)\mathbb{P}_{m,h}[V^{(l-1)} - V^{(l)}](s_t, a_t),$$

where $\boldsymbol{\theta}_{m,h}$ in the second step denotes the minimizer of the first step, $\mathbb{P}_{m,h}$ is the transition kernel defined by $\boldsymbol{\theta}_{m,h}$, and $c_t = c(s_t, a_t)$. Since $Q_j(s_t, a_t) = \min_a Q_j(s_t, a) = V_j(s_t)$, we get

$$V_j(s_t) - c_t = Q^{(l)}(s_t, a_t) - c_t$$

$$= (1 - q_j)\langle \boldsymbol{\theta}_{m,h}, \phi_{V^{(l)}}(s_t, a_t)\rangle + (1 - q_j)\mathbb{P}_{m,h}[V^{(l-1)} - V^{(l)}](s_t, a_t)$$

$$= (1 - q_j)\langle \boldsymbol{\theta}_{m,h} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^*, \phi_{V^{(l)}}(s_t, a_t)\rangle + (1 - q_j)\mathbb{P}_{m,h}[V^{(l-1)} - V^{(l)}](s_t, a_t)$$

$$= (1 - q_j)\langle \boldsymbol{\theta}_{m,h} - \boldsymbol{\theta}^*, \phi_{V_j}(s_t, a_t)\rangle + (1 - q_j)\mathbb{P}V_j(s_t, a_t)$$

$$\quad + (1 - q_j)\mathbb{P}_{m,h}[V^{(l-1)} - V^{(l)}](s_t, a_t)$$

$$= \mathbb{P}V_j(s_t, a_t) + (1 - q_j)\langle \boldsymbol{\theta}_{m,h} - \boldsymbol{\theta}^*, \phi_{V_j}(s_t, a_t)\rangle$$

$$\quad + (1 - q_j)\mathbb{P}_{m,h}[V^{(l-1)} - V^{(l)}](s_t, a_t) - q_j\mathbb{P}V_j(s_t, a_t) \tag{60}$$

Note that when event $\mathcal{E}_m$ holds, we can write

$$
\begin{aligned}
&\left|\langle \boldsymbol{\theta}_{m,h} - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_j}(s_t, a_t)\rangle\right| \\
&= \left|\langle \boldsymbol{\theta}_{m,h} - \widehat{\boldsymbol{\theta}}_j + \widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_j}(s_t, a_t)\rangle\right| \\
&\leq |(\boldsymbol{\theta}_{m,h} - \widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{\Sigma}_t^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t)| + |(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_j)^\top \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{\Sigma}_t^{-1/2} \boldsymbol{\phi}_{V_j}(s_t, a_t)| \\
&\leq \|\boldsymbol{\theta}_{m,h} - \widehat{\boldsymbol{\theta}}_j\|_{\boldsymbol{\Sigma}_t} \cdot \|\boldsymbol{\phi}_{V_j}(s_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\quad + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_j\|_{\boldsymbol{\Sigma}_t} \cdot \|\boldsymbol{\phi}_{V_j}(s_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\leq 2\widehat{\beta}_t \cdot \|\boldsymbol{\phi}_{V_j}(s_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}},
\end{aligned}
\tag{61}
$$

where the first inequality is by the triangular inequality, the second inequality is by the Cauchy Schwarz inequality, and the third step holds because $\boldsymbol{\theta}^* \in \mathcal{C}_{j_m}$ under the event $\mathcal{E}_m$.

Combining (60) and (61), under the event $\mathcal{E}_m$, we have

$$
V_j(s_t) - c_t = \mathbb{P}V_j(s_t, a_t) + e_t,
$$

where

$$
|e_t| \leq 2(1 - q_j)\widehat{\beta}_t \cdot \|\boldsymbol{\phi}_{V_j}(s_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}} + q_j B_\star + (1 - q_j)\epsilon_j,
\tag{62}
$$

where we use $|V_j| \leq B_\star$ under $\mathcal{E}_m$ for $j = j_m$. Together with (59), we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t)\, \mathbb{1}\{\mathcal{E}_m\}\,\Bigg|\, F_{m-1}\right] \\
&= \mathbb{E}\left[\left(\sum_{h=1}^{H_m} [V_j(s_t) - c_t - e_t - V_j(s_{t+1})]\right)^2 \mathbb{1}\{\mathcal{E}_m\}\,\Bigg|\, F_{m-1}\right] \\
&\leq \mathbb{E}\left[2\left(\sum_{h=1}^{H_m} [V_j(s_t) - c_t - V_j(s_{t+1})]\right)^2 \mathbb{1}\{\mathcal{E}_m\} + 2\left(\sum_{h=1}^{H_m} e_t\right)^2 \mathbb{1}\{\mathcal{E}_m\}\,\Bigg|\, F_{m-1}\right] \\
&= \mathbb{E}\left[2\left(V_j(s_{t(m,1)}) - V_j(s_{t(m,H_m+1)}) - \sum_{h=1}^{H_m} c_t\right)^2 \mathbb{1}\{\mathcal{E}_m\} + 2\left(\sum_{h=1}^{H_m} e_t\right)^2 \mathbb{1}\{\mathcal{E}_m\}\,\Bigg|\, F_{m-1}\right] \\
&\leq \mathbb{E}\left[2\,(3B)^2\, \mathbb{1}\{\mathcal{E}_m\} + 2\left(\sum_{h=1}^{H_m} e_t\right)^2 \mathbb{1}\{\mathcal{E}_m\}\,\Bigg|\, F_{m-1}\right],
\end{aligned}
\tag{63}
$$

where the second step is by $(a+b)^2 \leq 2a^2 + 2b^2$, the third step is by canceling the terms in the telescoping sum, and the last step holds since $|V_j(s_{t(m,1)}) - V_j(s_{t(m,H_m+1)})| \leq B_\star \leq B$ by optimism and $\sum_{h=1}^{H_m} c_t \leq 2B$ by the interval decomposition. Furthermore, we can write

$$
\begin{aligned}
&\mathbb{E}\left[\left(\sum_{h=1}^{H_m} e_t\right)^2 \Bigg|\, F_{m-1}\right] \\
&\leq \mathbb{E}\left[H_m \sum_{h=1}^{H_m} e_t^2\, \mathbb{1}\{\mathcal{E}_m\}\,\Bigg|\, F_{m-1}\right] \\
&\leq \frac{2B}{c_{\min}} \cdot \mathbb{E}\left[\sum_{h=1}^{H_m} e_t^2\, \mathbb{1}\{\mathcal{E}_m\}\,\Bigg|\, F_{m-1}\right] \\
&\leq \frac{2B}{c_{\min}} \cdot \mathbb{E}\left[6 \sum_{h=1}^{H_m} \left(\widehat{\beta}_t^2 \|\boldsymbol{\phi}_{V_j}(s_t, a_t)\|_{\boldsymbol{\Sigma}_t^{-1}}^2 + q_j^2 B_\star^2 + (1 - q_j)^2 \epsilon_j^2\right) \mathbb{1}\{\mathcal{E}_m\}\,\Bigg|\, F_{m-1}\right],
\end{aligned}
\tag{64}
$$

where the first and third steps are by $(\sum_{i=1}^{k} a_i)^2 \le k \sum_{i=1}^{k} a_i^2$ for $k \ge 1$, and the second step is by $H_m \le 2B/c_{\min}$.

Combining (63) and (64), we have

$$\mathbb{E}\left[\sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t)\, \mathbb{1}\{\mathcal{E}_m\}\middle| F_{m-1}\right]$$

$$\le 18B^2 + 24\frac{B}{c_{\min}}\mathbb{E}\left[\sum_{h=1}^{H_m}\left(\widehat{\beta}_t^2\|\phi_{V_j}(s_t, a_t)\|_{\Sigma_t^{-1}}^2 + B_\star^2(q_j^2 + \epsilon_j^2)\right)\mathbb{1}\{\mathcal{E}_m\}\middle| F_{m-1}\right].$$

Plugging in $q_j = \epsilon_j = 1/t_j$ finishes the proof.

$\square$

We are now ready to prove Lemma G.6 by using Lemma G.7 and applying Azuma-Hoeffding inequality multiple times.

Proof of Lemma G.6. We first bound $\sum_{m\in\mathcal{M}_0} \mathbb{E}\left[\sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t)\, \mathbb{1}\{\mathcal{E}_m\}\middle| F_{m-1}\right]$. By Lemma G.7, we have

$$\sum_{m\in\mathcal{M}_0} \mathbb{E}\left[\sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t)\, \mathbb{1}\{\mathcal{E}_m\}\middle| F_{m-1}\right]$$

$$\le 18B^2 M + \underbrace{\frac{24B}{c_{\min}}\sum_{m\in\mathcal{M}_0}\mathbb{E}\left[\sum_{h=1}^{H_m}\left(\widehat{\beta}_t^2\|\phi_{V_j}(s_t, a_t)\|_{\Sigma_t^{-1}}^2\right)\mathbb{1}\{\mathcal{E}_m\}\middle| F_{m-1}\right]}_{\text{I}}$$

$$+ \underbrace{\frac{48BB_\star^2}{c_{\min}}\sum_{m\in\mathcal{M}_0}\mathbb{E}\left[\sum_{h=1}^{H_m}\frac{1}{t_j^2}\mathbb{1}\{\mathcal{E}_m\}\middle| F_{m-1}\right]}_{\text{II}}. \tag{65}$$

To bound I, first note that, by picking $\lambda = 1/B^2$,

$$\sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\left(\widehat{\beta}_t^2\|\phi_{V_j}(s_t, a_t)\|_{\Sigma_t^{-1}}^2\right)\mathbb{1}\{\mathcal{E}_m\}$$

$$= \sum_{m\in\mathcal{M}_0}\sum_{h=1}^{H_m}\left(\widehat{\beta}_t^2\widehat{\sigma}_t^2\|\phi_{V_j}(s_t, a_t)/\widehat{\sigma}_t\|_{\Sigma_t^{-1}}^2\right)\mathbb{1}\{\mathcal{E}_m\}$$

$$\le 3\widehat{\beta}_T^2 \cdot B^2 d \cdot \log\left(1 + T/\lambda\right), \tag{66}$$

where we use Lemma H.4, Lemma H.5 and $B/\sqrt{d} \le \widehat{\sigma}_t \le \sqrt{3}B$. Define X_m as

$$X_m = \sum_{h=1}^{H_m}\left(\widehat{\beta}_t^2\|\phi_{V_j}(s_t, a_t)\|_{\Sigma_t^{-1}}^2\right)\mathbb{1}\{\mathcal{E}_m\} - \mathbb{E}[\sum_{h=1}^{H_m}(\widehat{\beta}_t^2\|\phi_{V_j}(s_t, a_t)\|_{\Sigma_t^{-1}}^2)\mathbb{1}\{\mathcal{E}_m\} \mid F_{m-1}].$$

Then $\{X_m\}_{m=1}^{\infty}$ is a martingale difference sequence since $\mathbb{E}[X_m|F_{m-1}] = 0$. Furthermore, by Lemma H.5 again we have $|X_m| \le 3\widehat{\beta}_t^2 B^2 d \log(1 + 2B/(c_{\min}\lambda))$, since $H_m \le 2B/c_{\min}$. By Azuma-Hoeffding inequality, we get that, with probability at least $1 - \delta$,

$$\left|\sum_{m\in\mathcal{M}_0} X_m\right| \le 6\widehat{\beta}_T^2 \cdot B^2 d \cdot \log\left(1 + \frac{2B}{c_{\min}\lambda}\right)\sqrt{M\log(1/\delta)}.$$

Combining with (66), we have that, with probability at least $1 - \delta$,

$$\text{I} \le \frac{144B^3 d\widehat{\beta}_T^2}{c_{\min}}\left(\sqrt{M}\cdot\log(1 + 2B/(c_{\min}\lambda))\cdot\sqrt{\log(1/\delta)} + \log(1 + T/\lambda)\right). \tag{67}$$

Similarly, to bound II, we have

$$\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \frac{1}{t_j^2} \mathbb{1}\{\mathcal{E}_m\} \le \sum_{j=1}^{J} \sum_{t=t_j+1}^{t_{j+1}} \frac{1}{t_j^2} \le \sum_{j=1}^{J} \frac{2t_j}{t_j^2} \le J + 1,$$

since $2t_j \le t_j^2$ for all $t_j > 1$. We then apply Azuma-Hoeffding inequality to the martingale difference sequence $Y_m := \sum_{h=1}^{H_m} \frac{1}{t_j^2} \mathbb{1}\{\mathcal{E}_m\} - \mathbb{E}[\sum_{h=1}^{H_m} \frac{1}{t_j^2} \mathbb{1}\{\mathcal{E}_m\} \mid F_{m-1}]$ with $|Y_m| \le 2$ since $H_m \le t_{j+1} - t_j \le 2t_j$, and get that, with probability at least $1 - \delta$, $|\sum_{m \in \mathcal{M}_0} Y_m| \le 4\sqrt{M \log(1/\delta)}$. It follows that, with probability at least $1 - \delta$,

$$\text{II} \le \frac{48BB_\star^2}{c_{\min}} \left(J + 4\sqrt{M \log(1/\delta)} + 1 \right). \tag{68}$$

Plugging (67) and (68) into (65) and by a union bound, we conclude that, with probability at least $1 - 2\delta$,

$$\sum_{m \in \mathcal{M}_0} \mathbb{E}\left[\sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t) \mathbb{1}\{\mathcal{E}_m\} \middle| F_{m-1} \right]$$
$$\le 18B^2 M + \frac{336B^3 d\widehat{\beta}_T^2}{c_{\min}} \left(\sqrt{M}\sqrt{\log(1/\delta)} \log\left(1 + \frac{2B}{c_{\min}\lambda}\right) + \log(1 + T/\lambda) \right), \tag{69}$$

where we use the bound for J from Lemma G.3.

Now, to bound $\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t) \mathbb{1}\{\mathcal{E}_m\}$, we apply Azuma-Hoeffding inequality once again. Specifically, we define Z_m as

$$Z_m := \sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t) \mathbb{1}\{\mathcal{E}_m\} - \mathbb{E}\left[\sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t) \mathbb{1}\{\mathcal{E}_m\} \middle| F_{m-1} \right],$$

and we have $|Z_m| \le 2BB_\star^2/c_{\min}$ since $H_m \le 2B/c_{\min}$ and $|V_j| \le B_\star$ on the event $\mathcal{E}_m$. Then with probability at least $1 - \delta$, $|\sum_{m \in \mathcal{M}_0} Z_m| \le (2BB_\star^2/c_{\min})\sqrt{M \log(1/\delta)}$. Together with (69) and a union bound, we conclude that, with probability at least $1 - 3\delta$,

$$\sum_{m \in \mathcal{M}_0} \sum_{h=1}^{H_m} \mathbb{V}V_j(s_t, a_t) \mathbb{1}\{\mathcal{E}_m\}$$
$$\le 18B^2 M + \frac{338B^3 d\widehat{\beta}_T^2}{c_{\min}} \left(\sqrt{M}\sqrt{\log(1/\delta)} \log\left(1 + \frac{2B}{c_{\min}\lambda}\right) + \log(1 + T/\lambda) \right).$$

$\square$

H. Auxiliary Lemmas

In this subsection we introduce the auxiliary lemmas used in the analysis.

Lemma H.1 (Azuma-Hoeffding inequality). *Let $\{X_t\}_{t=0}^\infty$ be a real-valued martingale such that for every $t \ge 1$, it holds that $|X_t - X_{t-1}| \le B$ for some $B \ge 0$. Then with probability at least $1 - \delta$, the following holds*

$$|X_t - X_0| \le 2B\sqrt{t \log\left(\frac{1}{\delta}\right)}.$$

Lemma H.2 (Azuma-Hoeffding inequality, anytime version). *Let $\{X_t\}_{t=0}^\infty$ be a real-valued martingale such that for every $t \ge 1$, it holds that $|X_t - X_{t-1}| \le B$ for some $B \ge 0$. Then for any $0 < \delta \le 1/2$, with probability at least $1 - \delta$, the following holds for all $t \ge 0$*

$$|X_t - X_0| \le 2B\sqrt{2t \log\left(\frac{t}{\delta}\right)}.$$

Proof of Lemma H.2. By Lemma H.1, for any t, with probability at least $1 - \frac{\delta}{t(t+1)}$, we have

$$|X_t - X_0| \leq 2B\sqrt{t\log\left(\frac{t(t+1)}{\delta}\right)}.$$

Note that since

$$\sum_{t=1}^{\infty} \frac{\delta}{t(t+1)} = \sum_{t=1}^{\infty}\left(\frac{1}{t} - \frac{1}{t+1}\right)\delta = \delta,$$

we take an union bound and get that, with probability at least $1 - \delta$, for all t, the following holds

$$|X_t - X_0| \leq 2B\sqrt{t\log\left(\frac{t(t+1)}{\delta}\right)} \leq 2B\sqrt{t\log\left(\frac{t^2}{\delta^2}\right)},$$

where the second step is by $\delta \leq 1/2$.

$\square$

Theorem H.3 (Bernstein inequality for vector-valued martingales, Theorem 4.1 in Zhou et al. 2021a)**.** *Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration, and $\{\mathbf{x}_t, \eta_t\}_{t\geq 1}$ be a stochastic process such that $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{F}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{F}_{t+1}$-measurable. Define $\mathbf{y}_t = \langle\mu^*, \mathbf{x}_t\rangle + \eta_t$. Assume the following holds:*

$$|\eta_t| \leq R, \ \mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0, \ \mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] \leq \sigma^2, \ \|\mathbf{x}_t\|_2 \leq L.$$

And for all $t \geq 1$, let

$$\beta_t = 8\sigma\sqrt{d\log(1 + tL^2/(d\lambda)) \cdot \log(4t^2/\delta)} + 4R\log(4t^2/\delta).$$

Then for any $0 < \delta < 1$, with probability at least $1 - \delta$, for all $t \geq 1$, it holds that

$$\left\|\sum_{i=1}^{t}\mathbf{x}_i\eta_i\right\|_{\mathbf{Z}_t^{-1}} \leq \beta_t, \ \|\mu_t - \mu^*\|_{\mathbf{Z}_t} \leq \beta_t + \sqrt{\lambda}\|\mu^*\|_2,$$

where $\mu_t = \mathbf{Z}_t^{-1}\mathbf{w}_t$, $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^{\top}$, $\mathbf{w}_t = \sum_{i=1}^{t}y_i\mathbf{x}_i$.

Lemma H.4 (Determinant-trace inequality, Lemma 10 in Abbasi-Yadkori et al. 2011)**.** *Assume $\phi_1, \cdots, \phi_t \in \mathbb{R}^d$ and for any $s \leq t$, $\|\phi_s\|_2 \leq L$. Let $\lambda > 0$ and $\Sigma_t = \lambda\mathbf{I} + \sum_{s=1}^{t}\phi_s\phi_s^{\top}$. Then*

$$\det(\Sigma_t) \leq \left(\lambda + tL^2/d\right)^d.$$

Lemma H.5 (Lemma 11 in Abbasi-Yadkori et al. 2011)**.** *Let $\{\phi_t\}_{t=1}^{\infty}$ be in $\mathbb{R}^d$ such that $\|\phi_t\| \leq L$ for all t. Assume Σ_0 is a PSD matrix in $\mathbb{R}^{d\times d}$, and let $\Sigma_t = \Sigma_0 + \sum_{s=1}^{t}\phi_s\phi_s^{\top}$. Then we have*

$$\sum_{s=1}^{t}\min\left\{1, \|\phi_s\|_{\Sigma_{s-1}^{-1}}^2\right\} \leq 2\left[d\log\left(\frac{\text{trace}(\Sigma_0) + tL^2}{d}\right) - \log\det(\Sigma_0)\right].$$

Furthermore, if $\lambda_{\min}(\Sigma_0) \geq \max\{1, L^2\}$, then

$$\sum_{s=1}^{t}\|\phi_s\|_{\Sigma_{s-1}^{-1}}^2 \leq 2\left[d\log\left(\frac{\text{trace}(\Sigma_0) + tL^2}{d}\right) - \log\det(\Sigma_0)\right].$$