LEARNING STOCHASTIC SHORTEST PATH WITH LINEAR FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

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 SSP. We propose a novel algorithm for learning the linear mixture SSP, which can attain a $\widetilde{\mathcal{O}}(dB_\star^{1.5}\sqrt{K/c_{\min}})$ 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 $c_{\min} = 0$, where a $\widetilde{\mathcal{O}}(K^{2/3})$ regret is guaranteed. To the best of our knowledge, this is the first algorithm with a sublinear regret guarantee for learning linear mixture SSP. In complement to the regret upper bounds, we also prove a lower bound of $\Omega(dB_\star\sqrt{K})$, which nearly matches our upper bound.

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 $\mathbb{P}(s'|s,a) = \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta}^* \rangle$ 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 $\boldsymbol{\phi} \in \mathbb{R}^d$ is a known ternary feature mapping, and $\boldsymbol{\theta}^* \in \mathbb{R}^d$ 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.

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 specially designed Extended Value Iteration (EVI) 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 EVI with t being the cumulative number of time steps, which guarantees the convergence of EVI. 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 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.

It is worth noting that a recent work by Vial et al. (2021) studied a different linear SSP model that is similar to linear MDPs (Yang & Wang, 2019; Jin et al., 2020), where both the underlying transition probability and 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. They proposed an algorithm with linear function approximation, which achieves $\widetilde{\mathcal{O}}(\sqrt{B_\star^3 d^3 K/c_{\min}})$ regret. The linear SSP model is different from our model, and we refer the interested readers to Ayoub et al. (2020); Zhou et al. (2021b) for a detailed comparison between these two assumptions. Besides the model difference, Vial et al. (2021) further assumed the feature mapping to be orthonormal in order to obtain the $\widetilde{O}(\sqrt{K})$ regret. We do not need such restrictive assumptions on the feature mapping, thus our algorithm provably works for more general cases.

Our contributions are summarized as follows:

- We propose to study a linear mixture SSP model, and devise a novel 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 $\widetilde{\mathcal{O}}(B_\star^{1.5} d\sqrt{K/c_{\min}})$ 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 $\widetilde{\mathcal{O}}(K^{2/3})$ regret guarantee by using a cost perturbation trick (Tarbouriech et al., 2021b).
- We prove that for linear mixture SSP, the regret of any learning algorithms is at least $\Omega(dB_\star\sqrt{K})$. This suggests that when $c_{\min} > 0$, our algorithm is optimal with regard to the dimension of the feature mapping d and number of episodes K.

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 $a_n \leq C b_n$. 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 a $\widetilde{\mathcal{O}}(D^{3/2}S\sqrt{AK/c_{\min}})$ 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

[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.

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, 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). For other settings of SSP, (Rosenberg & Mansour, 2020; Chen & Luo, 2021; Chen et al., 2021b) 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 SSP 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); He et al. (2021), 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; Zhou et al., 2021b;a). For the finite-horizon setting, Jia et al. (2020) proposed a UCRL-VTR algorithm that achieves a $\widetilde{\mathcal{O}}(d\sqrt{H^3T})$ 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 a $\widetilde{\mathcal{O}}(d\sqrt{T}/(1-\gamma)^2)$ 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 $M \coloneqq \{\mathcal{S}, \mathcal{A}, \mathbb{P}, c, s_{\text{init}}, g\}$, 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. Suppose the feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is known and pregiven. There exists an *unknown* vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}^*\|_2 \le \sqrt{d}$ such that $\mathbb{P}(s'|s, a) = \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}^* \rangle$ 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 $\|\boldsymbol{\phi}_V(s, a)\|_2 \le B\sqrt{d}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $\boldsymbol{\phi}_V(s, a) \coloneqq \sum_{s' \in \mathcal{S}} \boldsymbol{\phi}(s'|s, a)V(s')$.

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

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

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

$$V^{\pi}(s) \coloneqq \lim_{T \to +\infty} \mathbb{E}\left[\sum_{t=1}^{T} c(s_t, \pi(s_t)) \middle| s_1 = s\right], \quad \text{where } s_{t+1} \sim \mathbb{P}\big(\cdot \,|s_t, \pi(s_t)\big).$$

$V^{\pi}(s)$ can possibly be infinite if π is improper. The action-value function of policy π is defined as

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

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

$$\mathcal{L}V(s) \coloneqq \min_{a \in \mathcal{A}} \{c(s,a) + \mathbb{P}V(s,a)\}. \tag{3.1}$$

Intuitively speaking, we want to learn the optimal policy $\pi^{\star}$ such that $V^{\star}(\cdot) \coloneqq V^{\pi^{\star}}(\cdot)$ 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 $c_{\rho}(s,a) \coloneqq \max\{c(s,a), \rho\}$ 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., $B_{\star} \coloneqq \max_{s \in \mathcal{S}} V^{\star}(s)$. Also, we define $T_{\star} \coloneqq \max_{s \in \mathcal{S}} T^{\pi^{\star}}(s)$, 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 $Q^{\star} \coloneqq Q^{\pi^{\star}}$ which satisfies the following Bellman equation for all $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$Q^{\star}(s,a) = c(s,a) + \mathbb{P}V^{\star}(s,a), \quad V^{\star}(s) = \min_{a \in \mathcal{A}} Q^{\star}(s,a). \tag{3.2}$$

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 agent is to minimize the regret over K episodes defined as

$$R_K \coloneqq \sum_{k=1}^{K} \sum_{i=1}^{I_k} c_{k,i} - K \cdot V^{\star}(s_{\text{init}}), \tag{3.3}$$

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 during the k-th episode. Note that R_K might be infinite if some episode never ends.

4 Algorithms

In this section, we propose a model-based algorithm named `LEVIS`, as 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 some 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 an Extend Value Iteration (`EVI`) sub-routine (Algorithm 2). We now give a detailed description of Algorithm 1.

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: $\widehat{\boldsymbol{\theta}}_j \leftarrow \mathbf{\Sigma}_t^{-1} \mathbf{b}_t$
13: Set confidence 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 $Q_j(\cdot,\cdot) \leftarrow \texttt{EVI}(\mathcal{C}_j, \epsilon_j, \frac{1}{t_j}, \rho)$
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**

Algorithm 2 EVI

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:

$$Q^{(i+1)}(\cdot,\cdot) \leftarrow c_\rho(\cdot,\cdot) + (1-q) \cdot \min_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(i)}}(\cdot,\cdot) \rangle \tag{4.1}$$

$$V^{(i+1)}(\cdot) \leftarrow \min_{a \in \mathcal{A}} Q^{(i+1)}(\cdot, a) \tag{4.2}$$

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

In Algorithm 1, we maintain two global indices. Index t represents the total number of steps, and index j tracks the number of calls to the EVI sub-routine, where the output of EVI is an updated 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 estimation Q_j of the action-value function (Line 6) and updates $\mathbf{\Sigma}_t$ and $\mathbf{b}_t$ (Lines 7 and 8). The current epoch ends when the either criterion in Line 9 is triggered, and the EVI subroutine performs an optimistic planning to update the action-value function estimator (Lines 10 to 15).

Update criteria As mentioned before, Algorithm 1 runs in epochs indexed by j, and one epoch ends when either of the two update criteria is triggered (Line 9). The first updating criterion is satisfied once the determinant of $\mathbf{\Sigma}_t$ is doubled compared to the 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), which reflects the diminishing return of learning the underlying transition. One 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 denotes 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 do not have that $\|\boldsymbol{\phi}_V(\cdot,\cdot)\|$ is bounded from below, 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 a pair of s,a by the pigeonhole principle. To address this problem, we show that we only need to add 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 `EVI` 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 allow 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.

Optimistic planning The optimism of Algorithm 1 is realized by the construction of the confidence set $\mathcal{C}_j$ (Line 11), which is fed into the `EVI` subroutine. We now describe the construction of the Q-function estimator in the `EVI` sub-routine (Algorithm 2). `EVI` requires the access to a confidence ellipsoid $\mathcal{C}_j$ that contains the true model parameter $\boldsymbol{\theta}^*$ with high probability (Line 13). Here we construct the confidence set $\mathcal{C}_j$ centered at the minimizer of the ridge regression problem with a confidence radius parameter β_t (Line 13). Since not every $\boldsymbol{\theta} \in \mathcal{C}_j$ defines a valid transition probability function, we further take the intersection between $\mathcal{C}_j$ and a constraint set $\mathcal{B}$ defined as follows

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

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

Specifically, (4.1) in Algorithm 2 repeatedly conducts one-step value iteration by applying the best possible Bellman operator to the set $\mathcal{C}_j \cap \mathcal{B}$. This is motivated by the Bellman optimality equation in (3.2), 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 cannot guarantee the convergence of `EVI` because $\langle \cdot, \boldsymbol{\phi}_{V^{(i)}}\rangle$ is not a contractive map, which holds for free in the discounted setting (Jaksch et al., 2010; Zhou et al., 2021b), but not in SSP. More specifically, in the `EVI` algorithm for the discounted setting (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. As a result, the value iteration converges in a finite number of iterations. However, the Bellman equation of SSP does not have a discount factor. To address this issue, in (4.1), we introduce an extra $1-q$ 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 as will be shown, this bias will only introduce an additive term of order $\mathcal{O}(\log T)$ in the final regret bound.

Besides the convergence guarantee, the $1-q$ 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` value iteration algorithm used by Tarbouriech et al. (2021b). The intuition behind such a design is to ensure 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.

Regarding the implementation of `LEVIS`, note that the main computational overhead is from `EVI`, where within each inner iteration we need to solve an optimization problem. Fortunately, the loss function is strongly convex, thus it can be efficiently solved by many convex optimization algorithms.

5 MAIN RESULTS

In this section, we present the main theoretical results for Algorithm 1. We provide regret upper bounds for both positive cost functions and general cost functions, followed by a lower bound.

5.1 UPPER BOUNDS: POSITIVE COST FUNCTIONS

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

Assumption 5.1. 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. To simplify the expression, 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 $\beta_t = B\sqrt{d \log\left(4(t^2 + t^3 B^2/\lambda)/\delta\right)} + \sqrt{\lambda d}$ 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

$$R_K = \mathcal{O}\left(B^{1.5} d \sqrt{K/c_{\min}} \cdot \log^2\left(\frac{KBd}{c_{\min}\delta}\right) + \frac{B^2 d^2}{c_{\min}} \log^2\left(\frac{KBd}{c_{\min}\delta}\right)\right). \tag{5.1}$$

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 (5.1) has an dependency on $1/c_{\min}$. For the tabular SSP, Cohen et al. (2021); Jafarnia-Jahromi et al. (2021); Tarbouriech et al. (2021b) avoid such dependency by using a Bernstein-type confidence set. However, it remains an open question whether a similar result can be achieved under the linear function approximation setting.

Remark 5.3. If we 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, for the expected regret, we have

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

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

5.2 UPPER BOUND: GENERAL COST FUNCTIONS

When Assumption 5.1 does not hold, 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 $\beta_t = B\sqrt{d \log\left(4(t^2 + t^3 B^2/\lambda)/\delta\right)} + \sqrt{\lambda d}$ 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

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

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

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. As a result, the perturbation can only give an $\widetilde{\mathcal{O}}(K^{2/3})$ regret in the case of general cost functions. Similarly, the regret bound of learning linear SSP (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 instead of multiplication, which improves the regret to $\widetilde{\mathcal{O}}(K^{1/2})$ for general cost functions. How to get an additive $c_{\min}^{-1}$ term in the linear function approximation setting 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. For the upper bound $B \geq B_\star$, note that assuming a known B is common in existing SSP literature (Cohen et al., 2021; Vial et al., 2021). Although it is possible to deal with an unknown B in the tabular SSP with a doubling trick (Rosenberg et al., 2020; Tarbouriech et al., 2021b), it remains an open question for SSP with linear function approximation.

5.3 Lower Bound

We also provide a hardness result for learning linear mixture SSP by proving the lower bound for the expected regret suffered by any deterministic learning algorithms.

Theorem 5.5. 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

$$\mathbb{E}_{\pi,\boldsymbol{\theta}^*}[R_K] \geq \frac{dB_\star\sqrt{K}}{1024}. \tag{5.2}$$

Remark 5.6. The expectation in (5.2) is over the trajectories induced by executing the policy π in the SSP environment parameterized by $\boldsymbol{\theta}^*$. Note that here 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 5.7. Our instance for the lower bound can be also adapted to a linear SSP instance (Vial et al., 2021), which yields a $\Omega(dB_\star\sqrt{K})$ lower bound. (See Remark E.1 for a detailed discussion.)

6 Proof Sketch of the Main Results

In this section, we give a proof sketch of the main results in Section 5. Due to space limit, we defer the proof of the lemmas to the appendix.

6.1 Proof of Theorem 5.2

In this subsection, we prove Theorem 5.2, which gives the regret upper bound of Algorithm 1 for positive cost functions. The proof relies on the following intermediate result.

Theorem 6.1. Under Assumption 3.1 and 3.2, for any $\delta > 0$, let $\rho = 0$ and $\beta_t = B\sqrt{d\log\left(4(t^2+t^3B^2/\lambda)/\delta\right)} + \sqrt{\lambda d}$ for some $B \geq B_\star$ where $\lambda \geq 1$ and $\rho = 0$. Then with probability at least $1-\delta$, the regret of Algorithm 1 satisfies

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

where T is the total number of steps.

Remark 6.2. Theorem 6.1 gives an $\widetilde{O}(\sqrt{T})$ 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. Thus, Theorem 6.1 is not satisfactory due to its dependence on T. To deal with this problem, we further prove Theorem 5.2, which translates the dependence on T into the dependence on K but has a worse dependence on the dimension d and other logarithmic factors.

Theorem 6.1 applies to the general cost function with ρ set to 0. Note that the regret upper bound depends on the total number of time steps T, which is random. To replace the T-dependence by the K-dependence, it suffices to show that $T = \widetilde{\mathcal{O}}(K)$. As mentioned in Section 5.1, this can be easily derived under Assumption 5.1. We are now ready to prove Theorem 5.2.

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 6.1, with probability at least $1-\delta$, we have

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

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

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

Plugging this into Theorem 6.1 yields the desired result. □

Note that for the general cost functions, by simply picking $\rho = K^{-1/3}$ the result immediately follows from the case of positive costs, which is summarized in Theorem 5.4.

6.2 Proof Sketch of Theorem 6.1

The main steps in proving Theorem 6.1 include an analysis of EVI and a regret decomposition. The complete proof can be found in Appendix D.

Analysis of EVI. By the algorithmic design we elaborated in Section 4, EVI guarantees optimism and finite-time convergence, which is summarized in Lemma 6.3 below.

Lemma 6.3. Let $\rho = 0$ and $\beta_t = B\sqrt{d\log\left(4(t^2 + t^3B^2/\lambda)/\delta\right)} + \sqrt{\lambda d}$ for all $t \geq 1$, where $B \geq B_\star$. Then with probability at least $1 - \delta/2$, for all $j \geq 1$, EVI converges in finite time and the following holds

$$\boldsymbol{\theta}^* \in \mathcal{C}_j \cap \mathcal{B}\,, \quad 0 \leq Q_j(\cdot,\cdot) \leq Q^\star(\cdot,\cdot)\,, \quad \text{and} \quad 0 \leq V_j(\cdot) \leq V^\star(\cdot)\,.$$

Note that in Lemma 6.3 the optimism only holds for the EVI output, i.e., V_j for any $j \geq 1$. The initialization V_0 in Line 2 of the main Algorithm 1 does not necessarily satisfy the optimism since it is possible that $V^\star(s) < 1$ for some s. Still, such an initialization guarantees $\|V_0\|_\infty = 1 \leq B_\star$, which is crucial for establishing the optimism for $j \geq 1$. The proof of Lemma 6.3 is given in Appendix F.1.

Regret Decomposition. In our analysis, instead of dealing with (3.3) directly, we first implicitly decompose the times steps into intervals, which are indexed by $m = 1, \ldots, M$ in Lemma 6.4 below. The basic idea here is to decompose 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 EVI subroutine is triggered [3]. The purpose of such a regret decomposition is to guarantee that within each interval the optimistic action-value function remains the same, so the induced policy. This is a necessary and common requirement and can be found in the case of discounted infinite horizon MDPs (Zhou et al., 2021b). Similar decomposition trick has also been used in existing works on SSP (Rosenberg et al., 2020; Rosenberg & Mansour, 2020; Tarbouriech et al., 2021b).

Lemma 6.4. Assume the event in Lemma 6.3 holds, then we have the following upper bound for the regret defined in (3.3) [4]:

$$\begin{aligned} R(M) \leq & \underbrace{\textstyle\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[c_{m,h} + \mathbb{P}V_{j(m)}(s_{m,h},a_{m,h}) - V_{j(m)}(s_{m,h})\right]}_{E_1} \\ & + \underbrace{\textstyle\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j(m)}(s_{m,h+1}) - \mathbb{P}V_{j(m)}(s_{m,h},a_{m,h})\right]}_{E_2} \\ & + 2dB_\star\log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + 2B_\star\log(T) + 2. \end{aligned} \tag{6.1}$$

Bounding E_1 and E_2 We bound the terms E_1 and E_2 separately. Note that E_2 is the sum of a martingale difference sequence, and can be bounded by $\mathcal{O}(\sqrt{T\log(T/\delta)})$ using standard concentration. Bounding E_1 is more technical and it requires almost all the properties of our algorithmic design. In detail, we need to show that every time when EVI is triggered, it can output an optimistic action-value function estimator with high probability (by Lemma 6.3). Second, 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. Third, we need to bound the length of the epochs (i.e., the number of time steps between two EVIs), which is achieved by the time-step doubling criterion as explained in Section 4.

7 Conclusions

In this paper, we proposed a novel algorithm for linear mixture SSP and proved 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. We believe this gap can be closed by using a Bernstein-type of confidence set (Zhou et al., 2021a). Second, it remains open to prove a $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bound for linear mixture SSPs for general cost functions when $c_{\min} = 0$.

[3] The interval decomposition is indexed by m in Lemma 6.4. It is implicit and only for the purpose of analysis. This is different from the epoch decomposition, which is explicit and indexed by j in Algorithm 1. The difference is that an epoch ends when EVI is triggered, while an interval ends when either EVI is triggered 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.

ETHICS STATEMENT

We don't see any potential ethical issues in our work.

REFERENCES

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

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *arXiv preprint arXiv:1707.01495*, 2017.

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

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

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

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

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

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

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

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *arXiv preprint arXiv:2106.08377*, 2021a.

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

Alon Cohen, Haim Kaplan, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret bounds for stochastic shortest path. *arXiv preprint arXiv:2002.09869*, 2020.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *arXiv preprint arXiv:2103.13056*, 2021.

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

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

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

Mehdi Jafarnia-Jahromi, Liyu Chen, Rahul Jain, and Haipeng Luo. Online learning for stochastic shortest path model via posterior sampling. *arXiv preprint arXiv:2106.05335*, 2021.

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

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

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

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

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

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

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

Soroush Nasiriany, Vitchyr H Pong, Steven Lin, and Sergey Levine. Planning with goal-conditioned policies. *arXiv preprint arXiv:1911.08453*, 2019.

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

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

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pp. 9428–9437. PMLR, 2020a.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Improved sample complexity for incremental autonomous exploration in mdps. In *NeurIPS*, 2020b.

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

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *arXiv preprint arXiv:2104.11186*, 2021b.

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

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

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

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

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

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

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, 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 $\boldsymbol{\phi}(s'|s,a) = \mathbf{e}_{(s,a,s')} \in \mathbb{R}^d$ 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 $\boldsymbol{\phi}(s'|s,a) = \mathrm{vec}(\mathbf{p}(s'|s,a)\boldsymbol{\psi}(s,a)^\top) \in \mathbb{R}^d$ and parameter vector $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{W}) \in \mathbb{R}^d$ where $d = md'$, $\mathrm{vec}(\cdot)$ 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 $\boldsymbol{\phi}(s'|s,a) = \mathrm{vec}\big(\boldsymbol{\psi}_2(s')\boldsymbol{\psi}_1(s,a)^\top\big) \in \mathbb{R}^d$ and parameter vector $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{M}) \in \mathbb{R}^d$, 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 EXTENSION TO BERNSTEIN-TYPE ALGORITHMS

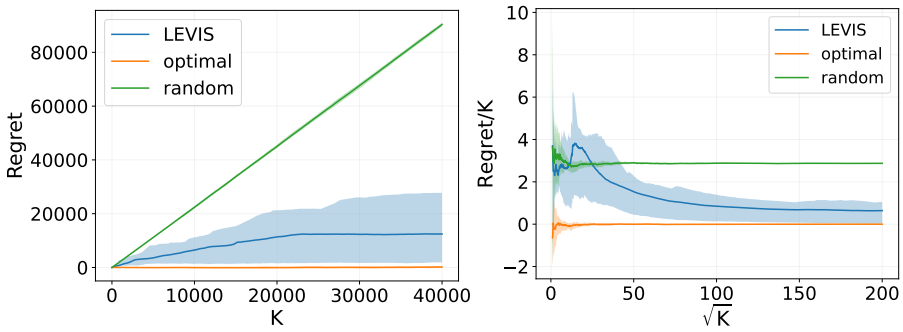
We believe it is possible to design a Bernstein-type algorithm to further improve the dependence on B from $\widetilde{\mathcal{O}}(B^{1.5})$ to $\widetilde{\mathcal{O}}(B)$, which is near-optimal according to the lower bound given by Theorem 5.5. Our belief is based on the following facts and analogy.

First, for the tabular SSP, previous work has shown that the near-optimal dependence on B is achievable by using Bernstein-type algorithms. For example, Algorithm 2 in Rosenberg et al. (2020) and Algorithm 1 in Tarbouriech et al. (2021b) are both Bernstein-type algorithms for tabular SSPs, which rely on the Bernstein-type bonus for exploration.

Second, for finite-horizon linear mixture MDPs, a Bernstein-type algorithm, `UCRL-VTR`$^+$, proposed in Zhou et al. (2021b) achieves $\mathcal{O}(H)$ dependence, where H is the horizon length. The key technique in their paper is to construct another linear estimator to estimate the variance of the value functions under the transition probability. Given this variance estimator, one can then use weighted ridge regression to estimate the transitional kernel parameter. Since B^* in SSPs can be viewed as a counterpart of H in finite-horizon MDPs, we think a similar result is achievable for the SSP problem by extending our algorithm in a way similar to that in Zhou et al. (2021b).

From another perspective, since at a high level our algorithmic design is more similar to the that of discounted MDPs than finite-horizon MDPs, one can also refer to the Bernstein-type algorithms for the discounted MDPs. For example, the `UCLK`$^+$ algorithm proposed in Zhou et al. (2021b) provably achieves a near-optimal regret for discounted MDPs by using the weighted linear regression technique. Notably, `UCLK`$^+$ uses a version of `EVI` algorithm along with a Bernstein-type confidence set.

Due to the above reason, we think an extension to the Bernstein-type algorithm for linear mixture SSPs is possible. We leave it as a future work.



(a) Cumulative regret R_K versus K.

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

Figure 1: Cumulative regret and average regret of implementing Algorithm 1 on the hard SSP instance described in Appendix B with $\lambda = 1$, $\rho = 0$ and failing probability 0.01. The curve is the average of 20 trials. Colored areas indicate empirical [10%,90%] confidence intervals.

B Numerical Simulations

In this section, we present some results from numerical simulations, which corroborate our theory. We construct an SSP instance based on the hard 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

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

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:

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

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. In Fig. 1a, we plot the cumulative regret R_K versus K. It is evident that `LEVIS` has a sublinear regret, as opposed to the linear regret of the random policy. In Fig. 1b, we plot the average reward versus $\sqrt{K}$, verifying the $\widetilde{\mathcal{O}}(\sqrt{K})$ regret of `LEVIS`. The results match our theoretical findings.

C Proof of Regret Decomposition

In this section, we prove the regret decomposition given by Lemma 6.4.

Proof of Lemma 6.4. 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 EVI 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 $j(\cdot)$, such that for each $m \in [M]$, $j(m)$ the index of the value function estimate which is 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

$$
\begin{aligned}
R(M) &\le \sum_{m=1}^{M}\sum_{h=1}^{H_m} c_{m,h} - \sum_{m\in\mathcal{M}(M)} V_{j(m)}(s_{\text{init}}) + 1 \\
&\le \sum_{m=1}^{M}\sum_{h=1}^{H_m} c_{m,h} + \sum_{m=1}^{M}\left(\sum_{h=1}^{H_m} V_{j(m)}(s_{m,h+1}) - V_{j(m)}(s_{m,h})\right) \\
&\quad + 1 + 1 + 2dB_\star \log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + 2B_\star \log(T) \\
&= \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[c_{m,h} + \mathbb{P}V_{j(m)}(s_{m,h}, a_{m,h}) - V_{j(m)}(s_{m,h})\right]}_{E_1} \\
&\quad + \underbrace{\sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[V_{j(m)}(s_{m,h+1}) - \mathbb{P}V_{j(m)}(s_{m,h}, a_{m,h})\right]}_{E_2} \\
&\quad + 2dB_\star \log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + 2B_\star \log(T) + 2. \qquad \text{(C.1)}
\end{aligned}
$$

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

The second inequality in the above is given by Lemma C.2 below, which is proved by first bounding the total number of calls to EVI (see Lemma C.1). □

The following lemma shows that the total calls to EVI in the implementation of Algorithm 1 can be bounded. It turns out 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 C.1. Conditioned on the event in Lemma 6.3, the total number of calls to EVI is bounded by $J \le 2d \log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + 2\log(T)$.

Proof of Lemma C.1. 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 EVI and it holds that $V_0 \leq B_\star$. By Line 7 of Algorithm 1 and the initialization $\mathbf{\Sigma}_0 = \lambda \mathbf{I}$, we have

$$\begin{aligned}\|\mathbf{\Sigma}_T\|_2 &= \left\| \lambda \mathbf{I} + \sum_{j=0}^{J} \sum_{t=t_j+1}^{t_{j+1}} \boldsymbol{\phi}_{V_j}(s_t, a_t) \boldsymbol{\phi}_{V_j}(s_t, a_t)^\top \right\|_2 \\ &\leq \lambda + \sum_{j=0}^{J} \sum_{t=t_j+1}^{t_{j+1}} \|\boldsymbol{\phi}_{V_j}(s_t, a_t)\|_2^2 \\ &\leq \lambda + T B_\star^2 d,\end{aligned}$$

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 6.3. We then have that $\det(\mathbf{\Sigma}_T) \leq (\lambda + T B_\star^2 d)^d$. It follows that

$$\left(\lambda + T B_\star^2 d\right)^d \geq 2^{J_1} \cdot \det\left(\mathbf{\Sigma}_0\right) = 2^{J_1} \cdot \lambda^d,$$

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

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

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

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

□

The following Lemma C.2 is used to get the second inequality in (C.1).

Lemma C.2. Conditioned on the event in Lemma 6.3, for the interval decomposition, the following holds

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

Proof of Lemma C.2. 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

$$\begin{aligned}&\sum_{m=1}^{M} \left(\sum_{h=1}^{H_m} V_{j(m)}(s_{m,h}) - V_{j(m)}(s_{m,h+1}) \right) \\ &= \sum_{m=1}^{M} V_{j(m)}(s_{m,1}) - V_{j(m)}(s_{m,H_m+1}) \\ &= \sum_{m=1}^{M-1} \left(V_{j(m+1)}(s_{m+1,1}) - V_{j(m)}(s_{m,H_m+1}) \right) + \sum_{m=1}^{M-1} \left(V_{j(m)}(s_{m,1}) - V_{j(m+1)}(s_{m+1,1}) \right) \\ &\quad + V_{j(M)}(s_{M,1}) - V_{j(M)}(s_{M,H_M+1}).\end{aligned}$$

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

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

$$
\begin{aligned}
&= \sum_{m=1}^{M-1} \left(V_{j(m+1)}(s_{m+1,1}) - V_{j(m)}(s_{m,H_m+1})\right) + V_{j(1)}(s_{1,1}) - V_{j(M)}(s_{M,1}) \\
&\quad + V_{j(M)}(s_{M,1}) - V_{j(M)}(s_{M,H_M+1}) \\
&= \sum_{m=1}^{M-1} \left(V_{j(m+1)}(s_{m+1,1}) - V_{j(m)}(s_{m,H_m+1})\right) + V_{j(1)}(s_{1,1}) - V_{j(M)}(s_{M,H_M+1}) \\
&\le \sum_{m=1}^{M-1} \left(V_{j(m+1)}(s_{m+1,1}) - V_{j(m)}(s_{m,H_m+1})\right) + V_{j(1)}(s_{1,1}), \qquad \text{(C.2)}
\end{aligned}
$$

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

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

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

□

D Proof of Theorem 6.1

In this section we finish the proof of the key result Theorem 6.1 by bounding the terms in the regret decomposition in Lemma 6.4.

D.1 Bounding E_1

Lemma D.1. Assume the event of Lemma 6.3 holds. Then we have

$$
\begin{aligned}
&\sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[c_{m,h} + \mathbb{P}V_{j(m)}(s_{m,h}, a_{m,h}) - V_{j(m)}(s_{m,h})\right] \\
&\le 4\beta_T \sqrt{2Td \cdot \log\left(1 + B_\star^2 T/\lambda\right)} + 5dB_\star \left[\log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + \log(T)\right] + 4.
\end{aligned}
$$

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 rewrite as

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

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 `EVI` instead of the initialization Q_0. Fix arbitrary $m \in \mathcal{M}_0(M)$ and h. Since $Q_{j(m)}$ is the output of `EVI`, we have $Q_{j(m)} = Q^{(l)}$ for some l, i.e., the l-th iteration in `EVI`, and thus $V_{j(m)}(\cdot) = \min_{a \in \mathcal{A}} Q^{(l)}(\cdot, a) = V^{(l)}(\cdot)$. By the design of `EVI`, we have

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

where $\boldsymbol{\theta}_{m,h} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{C}_j \cap \mathcal{B}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(l-1)}}(s_{m,h}, a_{m,h}) \rangle$ and its existence is guaranteed under the event of Lemma 6.3. 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

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

where the inequality is by the `EVI` 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

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

where the second inequality is by the optimism $V_{j(m)} \leq V^\star \leq B_\star$ under the event of Lemma 6.3, and $q = 1/t_{j(m)}$ according to Line 14 in Algorithm 1. We then conclude that

$$\begin{aligned}
&\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \left[c_{m,h} + \mathbb{P} V_{j(m)}(s_{m,h}, a_{m,h}) - Q_{j(m)}(s_{m,h}, a_{m,h}) \right] \\
&\leq \underbrace{\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h}) \rangle}_{A_1} + \underbrace{(B_\star + 1) \cdot \sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \frac{1}{t_{j(m)}}}_{A_2}.
\end{aligned} \tag{D.2}$$

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

$$\begin{aligned}
&\langle \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{j(m)} + \widehat{\boldsymbol{\theta}}_{j(m)} - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h})\rangle \\
&\leq \left(\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{j(m)}\|_{\boldsymbol{\Sigma}_{t(m,h)}} + \|\widehat{\boldsymbol{\theta}}_{j(m)} - \boldsymbol{\theta}_{m,h}\|_{\boldsymbol{\Sigma}_{t(m,h)}} \right) \cdot \|\boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}^{-1}_{t(m,h)}} \\
&\leq 2 \left(\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{j(m)}\|_{\boldsymbol{\Sigma}_{t_{j(m)}}} + \|\widehat{\boldsymbol{\theta}}_{j(m)} - \boldsymbol{\theta}_{m,h}\|_{\boldsymbol{\Sigma}_{t_{j(m)}}} \right) \cdot \|\boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}^{-1}_{t(m,h)}} \\
&\leq 4\beta_T \|\boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}^{-1}_{t(m,h)}}. \qquad \text{(D.3)}
\end{aligned}$$

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 `EVI` 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 6.3, $\boldsymbol{\theta}^*$ 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

$$\begin{aligned}
\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}_{m,h}, \boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h})\rangle &\leq \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h})\rangle \\
&= \mathbb{P}V_{j(m)}(s_{m,h}, a_{m,h}) \\
&\leq B_\star, \qquad \text{(D.4)}
\end{aligned}$$

where both inequalities hold due to $0 \leq V_{j(m)}(\cdot) \leq B_\star$. Combine (D.3) and (D.4) and we have

$$\begin{aligned}
A_1 &\leq 4\beta_T \sum_{m\in\mathcal{M}_0} \sum_{h=1}^{H_m} \min\left\{1,\ \|\boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h})\|_{\boldsymbol{\Sigma}^{-1}_{t(m,h)}}\right\} \\
&\leq 4\beta_T \sqrt{\left(\sum_{m\in\mathcal{M}_0} \sum_{h=1}^{H_m} 1\right) \cdot \left(\sum_{m\in\mathcal{M}_0} \sum_{h=1}^{H_m} \min\left\{1,\ \|\boldsymbol{\phi}_{V_{j(m)}}(s_{m,h}, a_{m,h})\|^2_{\boldsymbol{\Sigma}^{-1}_{t(m,h)}}\right\}\right)}, \qquad \text{(D.5)}
\end{aligned}$$

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

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

where the first inequality holds by Lemma G.4, and the second inequality holds because $V_{j(m)}(\cdot) \leq B_\star$ under Lemma 6.3 and thus $\max_{m\in\mathcal{M}_0}\|\boldsymbol{\phi}_{V_{j(m)}}(\cdot,\cdot)\|_2 \leq B_\star\sqrt{d}$ by Assumption 3.1. Combine the above inequality with (D.5) and we conclude that

$$A_1 \leq 4\beta_T\sqrt{2Td\cdot\log\left(1 + B_\star^2 T/\lambda\right)}. \qquad \text{(D.6)}$$

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

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

where the last step is by Lemma C.1. Together with (D.2) and (D.6) we conclude that

$$\begin{aligned} &\sum_{m \in \mathcal{M}_0(M)} \sum_{h=1}^{H_m} \left[c_{m,h} + \mathbb{P}V_{j(m)}(s_{m,h}, a_{m,h}) - Q_{j(m)}(s_{m,h}, a_{m,h})\right] \\ &\leq 4\beta_T \sqrt{2Td \cdot \log\left(1 + B_\star^2 T/\lambda\right)} + 5dB_\star \left[\log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + \log(T)\right]. \end{aligned} \tag{D.7}$$

To bound E_1, it remains to bound the following

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

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 `EVI` sub-routine. However, since $t_0 = 1$, by the triggering condition $t \geq 2t_0$, we know that the first `EVI` is called at $t = 2$. Therefore we have

$$\begin{aligned} &\sum_{m \in \mathcal{M}_0^c} \sum_{h=1}^{H_m} \left[c_{m,h} + \mathbb{P}V_{j(m)}(s_{m,h}, a_{m,h}) - Q_{j(m)}(s_{m,h}, a_{m,h})\right] \\ &= \sum_{h=1}^{2} \left[c_{1,h} + \mathbb{P}V_0(s_{1,h}, a_{1,h}) - Q_0(s_{1,h}, a_{1,h})\right] \\ &\leq 4, \end{aligned}$$

where the inequality holds because $c_{1,h}, V_0(\cdot) \leq 1$ and $0 \leq Q_0(\cdot, \cdot)$. Together with (D.7) we conclude that

$$E_1 \leq 4\beta_T \sqrt{2Td \cdot \log\left(1 + B_\star^2 T/\lambda\right)} + 5dB_\star \left[\log\left(1 + \frac{TB_\star^2 d}{\lambda}\right) + \log(T)\right] + 4. \tag{D.8}$$

□

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 6.3 and the following hold

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

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

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

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

$$\widetilde{V}_{j(m)}(\cdot) \coloneqq \min\{B_\star, V_{j(m)}(\cdot)\},$$

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 G.2 and get that, with probability at least $1-\delta/2$,

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

Now note that under the event of Lemma 6.3, we have $\widetilde{V}_{j(m)} = V_{j(m)}$ for all $j(m) \ge 1$ 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 $1-\delta$, $E_2 \le 2B_\star\sqrt{2T\log(2T/\delta)}$. □

D.3 Proof of Theorem 6.1

Proof. Note that the regret decomposition (6.1) is proved under the condition that the event of Lemma 6.3 holds. Then together with Lemmas 6.3, D.1 and D.2, we conclude that with probability at least $1-\delta$,

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

Combining the lower order terms finishes the proof. □

E Lower Bound

E.1 Proof of the Lower Bound

Proof of Theorem 5.5. 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 $\boldsymbol{\theta} \in \{-\frac{\Delta}{d-1}, \frac{\Delta}{d-1}\}^{d-1}$. 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

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

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

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

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 $N \coloneqq \sum_{t=1}^{\infty} \mathbb{1}\{s_t = s_{\text{init}}\}$, and for $j = 1, \cdots, d-1$,

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

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

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

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

$$\begin{aligned}
2\sum_{\boldsymbol{\theta} \in \boldsymbol{\theta}} R_{\boldsymbol{\theta}}(K) &\geq \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\mathbb{E}_{\boldsymbol{\theta}}[N_j^-(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}^j}[N_j^-(\boldsymbol{\theta}^j)] \right) \\
&= \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\mathbb{E}_{\boldsymbol{\theta}^j}[N^-] + \mathbb{E}_{\boldsymbol{\theta}}[N_j^-(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}^j}[N_j^-(\boldsymbol{\theta})] \right) \\
&= \frac{2\Delta B_\star}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\mathbb{E}_{\boldsymbol{\theta}}[N^-] + \mathbb{E}_{\boldsymbol{\theta}}[N_j^-(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}^j}[N_j^-(\boldsymbol{\theta})] \right). \qquad \text{(E.3)}
\end{aligned}$$

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

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

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

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

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

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

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

$$\begin{aligned}
\frac{1}{|\boldsymbol{\theta}|} \sum_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(K) &\geq B_\star \frac{1}{|\boldsymbol{\theta}|} \cdot \frac{1}{d-1} \sum_{\boldsymbol{\theta}} \sum_{j=1}^{d-1} \left(\frac{\Delta K B_\star}{4} - \frac{8\Delta^2}{(d-1)\sqrt{\delta}} (KB_\star)^{3/2} \right) \\
&= B_\star \left[\frac{\Delta K B_\star}{4} - \frac{8\Delta^2}{(d-1)\sqrt{\delta}} (KB_\star)^{3/2} \right]. \qquad \text{(E.5)}
\end{aligned}$$

We now pick

$$\Delta = \frac{(d-1)\sqrt{\delta}}{64\sqrt{KB_\star}}, \qquad \text{(E.6)}$$

and δ such that $\delta + \Delta = 1/B_\star$, plug into (E.5) and get that

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

where the last step is by $\delta + \Delta = \frac{1}{B_\star}$ and $\Delta < \delta$. Therefore, there must exist some $\boldsymbol{\theta} \in \boldsymbol{\theta}$ 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 (E.1) and (E.6), 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 5.5.

□

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

$$\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}} || \mathcal{P}_{\boldsymbol{\theta}^j}) = \sum_{t=1}^{T-1} \mathrm{KL}\left[\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t) \big|\big| \mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)\right], \tag{E.7}$$

where

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

Then we write

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

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

$$\begin{aligned}
&\sum_{s_{t+1}\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)} \\
&= \sum_{\mathbf{s}_{t-1}\in\mathcal{S}^{\times(t-1)}} \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{s}_{t-1}) \sum_{\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a}|\mathbf{s}_{t-1}) \\
&\quad \cdot \sum_{s\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a}, \mathbf{s}_{t-1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a}, \mathbf{s}_{t-1})}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}=s|s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a}, \mathbf{s}_{t-1})} \\
&= \sum_{\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a}) \cdot \sum_{s\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}=s|s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a})}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}=s|s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a})}.
\end{aligned} \tag{E.8}$$

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 (E.8) as

$$\begin{aligned}
&\sum_{s_{t+1}\in\mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}^j}(s_{t+1}|\mathbf{s}_t)} \\
&= \sum_{\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a}) \\
&\quad \cdot \left[(1-\delta-\langle \mathbf{a}, \boldsymbol{\theta}\rangle) \cdot \log \frac{1-\delta-\langle \mathbf{a}, \boldsymbol{\theta}\rangle}{1-\delta-\langle \mathbf{a}, \boldsymbol{\theta}^j\rangle} + (\delta + \langle \mathbf{a}, \boldsymbol{\theta}\rangle) \cdot \log \frac{\delta + \langle \mathbf{a}, \boldsymbol{\theta}\rangle}{\delta + \langle \mathbf{a}, \boldsymbol{\theta}^j\rangle}\right] \\
&\le \sum_{\mathbf{a}\in\mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}}(s_t=s_{\text{init}}, \mathbf{a}_t=\mathbf{a}) \cdot \frac{2\langle \mathbf{a}, \boldsymbol{\theta}^j - \boldsymbol{\theta}\rangle^2}{\delta + \langle \mathbf{a}, \boldsymbol{\theta}\rangle},
\end{aligned} \tag{E.9}$$

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 $\delta' \leq \delta+\Delta \leq 1/12+1/3 < 1/2$ by $4\Delta \leq \delta \leq 1/3$ from the assumption of Lemma E.4. Also note that

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

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, (E.9) 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 (E.7) we have

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

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

□

F LEMMAS FOR THE UPPER BOUNDS

F.1 PROOF OF LEMMA 6.3

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 $\{\boldsymbol{\phi}_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $\boldsymbol{\phi}_t$ is $\mathcal{F}_{t-1}$-measurable. Assume that $\boldsymbol{\Sigma}$ is an $d \times d$ positive definite matrix. For any $t \geq 1$, define

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

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 $t \in [t_{j-1}+1, t_j]$, during which the action-value function estimator is the output Q_j of `EVI`. We then apply an induction argument on the rounds to show that the optimism holds for all $j \geq 1$.

Proof of Lemma 6.3. From the initialization of Algorithm 1, we have $V_0 \leq B_\star$.
Let's consider **round 1**. We define $\eta_t = V_0(s_{t+1}) - \langle \boldsymbol{\phi}_{V_0}(s_t, a_t), \boldsymbol{\theta}^* \rangle$ 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 $t \in [1, t_1]$:

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_t^{-1/2}\sum_{i=1}^{t}\boldsymbol{\phi}_{V_0}(s_i,a_i)\eta_i\right\|_2 &\le B_\star\sqrt{2\log\left(\frac{\det(\boldsymbol{\Sigma}_t)^{1/2}}{\delta\cdot\lambda^{d/2}/(t_1(t_1+1))}\right)}\\ &\le B_\star\sqrt{d\log\left(\frac{1+td/(d\lambda)}{\delta/(t_1(t_1+1))}\right)}\\ &\le B_\star\sqrt{d\log\left(\frac{t_1(t_1+1)+t\cdot t_1(t_1+1)B_\star^2/\lambda}{\delta}\right)}, \qquad \text{(F.1)}\end{aligned}$$

where the second step is by Assumption 3.1, Lemma G.3 and the initialization $|V_0| \le 1$. Consider the LHS of (F.1). We have

$$\begin{aligned}&\left\|\boldsymbol{\Sigma}_t^{-1/2}\sum_{i=1}^{t}\boldsymbol{\phi}_{V_0}(s_i,a_i)\eta_i\right\|\\ &= \left\|\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\Sigma}_t^{-1}\sum_{i=1}^{t}\boldsymbol{\phi}_{V_0}(s_i,a_i)V_0(s_{i+1}) - \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\Sigma}_t^{-1}\left(\boldsymbol{\Sigma}_t-\lambda\mathbf{I}\right)\boldsymbol{\theta}^*\right\|_2\\ &= \left\|\boldsymbol{\Sigma}_t^{1/2}\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\theta}^* + \lambda\boldsymbol{\Sigma}_t^{-1/2}\boldsymbol{\theta}^*\right\|_2\\ &\ge \left\|\boldsymbol{\Sigma}_t^{1/2}(\widehat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}^*)\right\|_2 - \left\|\lambda\boldsymbol{\Sigma}_t^{-1/2}\boldsymbol{\theta}^*\right\|_2\\ &\ge \left\|\boldsymbol{\Sigma}_t^{1/2}(\widehat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}^*)\right\|_2 - \lambda^{1/2}\cdot\sqrt{d},\end{aligned}$$

where the first inequality holds by Cauchy-Schwarz inequality and the second inequality holds because $\|\boldsymbol{\theta}^*\|_2 \le \sqrt{d}$. Together with (F.1) and the choice of β_t, we conclude that

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

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

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

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 (3.2) 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 6.3 holds for round 1 to $j-1$ with high probability. That is, we define the event

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

and assume that $\Pr(\mathcal{E}_{j-1}) \ge 1-\delta'$ 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

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

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

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

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

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

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

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 $\texttt{EVI}(\widetilde{\mathcal{C}}_j, \epsilon_j, \frac{1}{t_j}, \rho)$.

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 $\texttt{EVI}$ being called. Note that compared with the analysis of $\texttt{EVI}$ in the discounted MDPs setting (for example in Zhou et al. 2021b), our analysis of $\texttt{EVI}$ 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 $\texttt{EVI}$ always converges in finite time. This actually follows directly from the results established above by using an argument similar to the analysis of $\texttt{EVI}$ for MDPs with constant discount factor. 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 (4.2) 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,

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

$$\leq (1-q)\cdot\|Q^{(i-1)}-Q^{(i-2)}\|_\infty,$$

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.

□

G AUXILIARY LEMMAS

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

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

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

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

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

Proof of Lemma G.2. By Lemma G.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 $\delta\leq 1/2$.

□

Lemma G.3 (Determinant-Trace inequality, Lemma 10 in Abbasi-Yadkori et al. 2011)**.** Assume $\boldsymbol{\phi}_1,\cdots,\boldsymbol{\phi}_t\in\mathbb{R}^d$ and for any $s\leq t$, $\|\boldsymbol{\phi}_s\|_2\leq L$. Let $\lambda>0$ and $\boldsymbol{\Sigma}_t = \lambda\mathbf{I}+\sum_{s=1}^t \boldsymbol{\phi}_s\boldsymbol{\phi}_s^\top$. Then

$$\det(\boldsymbol{\Sigma}_t) \leq \left(\lambda + tL^2/d\right)^d.$$

Lemma G.4 (Lemma 11 in Abbasi-Yadkori et al. 2011)**.** Let $\{\boldsymbol{\phi}_t\}_{t=1}^\infty$ be in $\mathbb{R}^d$ such that $\|\boldsymbol{\phi}_t\|\leq L$ for all t. Assume $\boldsymbol{\Sigma}_0$ is a PSD matrix in $\mathbb{R}^{d\times d}$, and let $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_0 + \sum_{s=1}^t \boldsymbol{\phi}_s\boldsymbol{\phi}_s^\top$. Then we have

$$\sum_{s=1}^t \min\left\{1, \|\boldsymbol{\phi}_s\|_{\boldsymbol{\Sigma}_{s-1}^{-1}}\right\} \leq 2\left[d\log\left(\frac{\mathrm{trace}(\boldsymbol{\Sigma}_0)+tL^2}{d}\right) - \log\det(\boldsymbol{\Sigma}_0)\right].$$