Implicit Finite-Horizon Approximation and Efficient Optimal Algorithms for Stochastic Shortest Path

Liyu Chen
University of Southern California
`liyuc@usc.edu`

Mehdi Jafarnia-Jahromi
University of Southern California
`mjafarni@usc.edu`

Rahul Jain
University of Southern California
`rahul.jain@usc.edu`

Haipeng Luo
University of Southern California
`haipengl@usc.edu`

Abstract

We introduce a generic template for developing regret minimization algorithms in the Stochastic Shortest Path (SSP) model, which achieves minimax optimal regret as long as certain properties are ensured. The key of our analysis is a new technique called implicit finite-horizon approximation, which approximates the SSP model by a finite-horizon counterpart *only in the analysis* without explicit implementation. Using this template, we develop two new algorithms: the first one is model-free (the first in the literature to our knowledge) and minimax optimal under strictly positive costs; the second one is model-based and minimax optimal even with zero-cost state-action pairs, matching the best existing result from [Tarbouriech et al., 2021b]. Importantly, both algorithms admit highly sparse updates, making them computationally more efficient than all existing algorithms. Moreover, both can be made completely parameter-free.

1 Introduction

We study the Stochastic Shortest Path (SSP) model, where an agent aims to reach a goal state with minimum cost in a stochastic environment. SSP is well-suited for modeling many real-world applications, such as robotic manipulation, car navigation, and others. Although it is widely studied empirically (e.g., [Andrychowicz et al., 2017, Nasiriany et al., 2019]) and in optimal control theory (e.g., [Bertsekas and Tsitsiklis, 1991, Bertsekas and Yu, 2013]), it has received less attention under the regret minimization setting where a learner needs to learn the environment and improve her policy on-the-fly through repeated interaction. Specifically, the problem proceeds in K episodes. In each episode, the learner starts at a fixed initial state, sequentially takes action, suffers some cost, and transits to the next state, until reaching a predefined goal state. The performance of the learner is measured by her regret, which is the difference between her total costs and that of the best policy.

Tarbouriech et al. [2020a] develop the first regret minimization algorithm for SSP with a regret bound of $\tilde{\mathcal{O}}(D^{3/2}S\sqrt{AK/c_{\min}})$, where D is the diameter, S is the number of states, A is the number of actions, and $c_{\min}$ is the minimum cost among all state-action pairs. Cohen et al. [2020] improve over their results and give a near optimal regret bound of $\tilde{\mathcal{O}}(B_\star S\sqrt{AK})$, where $B_\star \leq D$ is the largest expected cost of the optimal policy starting from any state. Even more recently, Cohen et al. [2021] achieve minimax regret of $\tilde{\mathcal{O}}(B_\star\sqrt{SAK})$ through a finite-horizon reduction technique, and concurrently Tarbouriech et al. [2021b] also propose minimax optimal and parameter-free algorithms. Notably, all existing algorithms are model-based with space complexity $\Omega(S^2A)$. Moreover, they all

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

update the learner's policy through full-planning (a term taken from [Efroni et al., 2019]), incurring a relatively high time complexity.

In this work, we further advance the state-of-the-art by proposing a generic template for regret minimization algorithms in SSP (Algorithm 1), which achieves minimax optimal regret as long as some properties are ensured. By instantiating our template differently, we make the following two key algorithmic contributions:

- In Section 4, we develop the *first model-free* SSP algorithm called LCB-ADVANTAGE-SSP (Algorithm 2). Similar to most model-free reinforcement learning algorithms, LCB-ADVANTAGE-SSP does not estimate the transition directly, enjoys a space complexity of $\tilde{\mathcal{O}}(SA)$, and also takes only $\mathcal{O}(1)$ time to update certain statistics in each step, making it a highly efficient algorithm. It achieves a regret bound of $\tilde{\mathcal{O}}(B_\star\sqrt{SAK} + B_\star^5 S^2 A / c_{\min}^4)$, which is minimax optimal when $c_{\min} > 0$. Moreover, it can be made parameter-free without worsening the regret bound.

- In Section 5, we develop another simple model-based algorithm called SVI-SSP (Algorithm 3), which achieves minimax regret $\tilde{\mathcal{O}}(B_\star\sqrt{SAK} + B_\star S^2 A)$ even when $c_{\min} = 0$, matching the best existing result by Tarbouriech et al. [2021b].[1] Notably, compared to their algorithm (as well as other model-based algorithms), SVI-SSP is computationally much more efficient since it updates each state-action pair only logarithmically many times, and each update only performs *one-step planning* (again, a term taken from [Efroni et al., 2019]) as opposed to full-planning (such as value iteration or extended value iteration); see more concrete time complexity comparisons in Section 5. SVI-SSP can also be made parameter-free following the idea of [Tarbouriech et al., 2021b].

We include a summary of regret bounds of all existing SSP algorithms as well as more complexity comparisons in Appendix A.

Techniques Our main technical contribution is a new analysis framework called *implicit finite-horizon approximation* (Section 3), which is the key to analyze algorithms developed from our template. The high level idea is to approximate an SSP instance by a finite-horizon counterpart. However, the approximation *only happens in the analysis*, a key difference compared to [Chen et al., 2021, Chen and Luo, 2021, Cohen et al., 2021] that explicitly implement such an approximation in their algorithms. As a result, our method not only avoids blowing up the space complexity by a factor of the horizon, but also allows one to derive a horizon-free regret bound (more explanation to follow).

In order to achieve the minimax optimal regret, our model-free algorithm LCB-ADVANTAGE-SSP uses a key variance reduction idea via a reference-advantage decomposition by [Zhang et al., 2020b]. However, crucial distinctions exist. For example, we update the reference value function more frequently instead of only one time, which helps reduce the sample complexity and improve the lower-order term in the regret bound. We also maintain an empirical upper bound on the value function in a doubling manner, which is the key to eventually make the algorithm parameter-free. On the other hand, for our model-based algorithm SVI-SSP, we adopt a special Bernstein-style bonus term and bound the learner's total variance via recursion, taking inspiration from [Tarbouriech et al., 2021b, Zhang et al., 2020a].

Empirical Evaluation We support our theoretical findings with experiments in Appendix H. Our model-free algorithm demonstrates a better convergence rate compared to vanilla Q learning with naive ϵ-greedy exploration. Our model-based algorithm has competitive performance compared to other model-based algorithms, while spending the least amount of time in updates.

Related Work For a detailed comparison of existing results for the same problem, we refer the readers to [Tarbouriech et al., 2021b, Table 1] as well as our Table 1. There are also several works [Rosenberg and Mansour, 2020, Chen et al., 2021, Chen and Luo, 2021] that consider the even more challenging SSP setting where the cost function is decided by an adversary and can change over time. Apart from regret minimization, Tarbouriech et al. [2021a] study the sample complexity of SSP with a generative model; Lim and Auer [2012] and Tarbouriech et al. [2020b] investigate exploration problems involving multiple goal states (multi-goal SSP).

---

[1] Depending on the available prior knowledge, the final bounds achieved by SVI-SSP are slightly different, but they all match that of EB-SSP. See [Tarbouriech et al., 2021b, Table 1] for more details.

The special case of SSP with a fixed horizon has been studied extensively, for both stochastic costs (e.g., [Azar et al., 2017, Jin et al., 2018, Efroni et al., 2019, Zanette and Brunskill, 2019, Zhang et al., 2020a]) and adversarial costs (e.g., [Neu et al., 2012, Zimin and Neu, 2013, Rosenberg and Mansour, 2019, Jin et al., 2020]). Importantly, recent works [Wang et al., 2020, Zhang et al., 2020a] find that when the cost for each episode is at most a constant, it is in fact possible to obtain a regret bound with only logarithmic dependency on the horizon. Tarbouriech et al. [2021b] generalize this concept to SSP and define horizon-free regret as a bound with only logarithmic dependence on the expected hitting time of the optimal policy starting from any state (which is bounded by $B_\star/c_{\min}$). They also propose the first algorithm with horizon-free regret for SSP, which is important for arguing minimax optimality even when $c_{\min}=0$. Notably, our model-based algorithm SVI-SSP also achieves horizon-free regret (but the model-free one does not).

2 Preliminaries

An SSP instance is defined by a Markov Decision Process (MDP) $M=(\mathcal{S},\mathcal{A},s_{\text{init}},g,c,P)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $s_{\text{init}}\in\mathcal{S}$ is the initial state, and $g\notin\mathcal{S}$ is the goal state. When taking action a in state s, the learner suffers a cost drawn in an i.i.d manner from an unknown distribution with mean $c(s,a)\in[0,1]$ and support $[c_{\min},1]$ ($c_{\min}\geq 0$), and then transits to the next state $s'\in\mathcal{S}^+=\mathcal{S}\cup\{g\}$ with probability $P_{s,a}(s')$. We assume that the transition P and the cost mean c are unknown to the learner, while all other parameters are known.

The learning process goes as follows: the learner interacts with the environment for K episodes. In the k-th episode, the learner starts in initial state s_{init}, sequentially takes an action, suffers a cost, and transits to the next state until reaching the goal state g. More formally, at the i-th step of the k-th episode, the learner observes the current state s_i^k (with $s_1^k=s_{\text{init}}$), takes action a_i^k, suffers a cost c_i^k, and transits to the next state $s_{i+1}^k\sim P_{s_i^k,a_i^k}$. An episode ends when the current state is g, and we define the length of episode k as I_k, such that $s_{I_k+1}^k=g$.

Learning Objective At a high level, the learner's goal is to reach the goal with a small total cost. To this end, we focus on *proper policies* — a (stationary and deterministic) policy $\pi:\mathcal{S}\rightarrow\mathcal{A}$ is a mapping that assigns an action $\pi(s)$ to each state $s\in\mathcal{S}$, and it is proper if the goal is reached with probability 1 when following π (that is, taking action $\pi(s)$ whenever in state s). Given a proper policy π, one can define the cost-to-go function $V^\pi:\mathcal{S}\rightarrow[0,\infty)$ as $V^\pi(s)=\mathbb{E}\left[\sum_{i=1}^{I}c_i\middle| P,\pi,s_1=s\right]$, where the expectation is with respect to the randomness of the cost c_i incurred at state-action pair $(s_i,\pi(s_i))$, next state $s_{i+1}\sim P_{s_i,\pi(s_i)}$, and the number of steps I before reaching g. The optimal proper policy $\pi^\star$ is then defined as a policy such that $V^{\pi^\star}(s)=\min_{\pi\in\Pi}V^\pi(s)$ for all $s\in\mathcal{S}$, where Π is the set of all proper policies assumed to be nonempty. The formal objective of the learner is then to minimize her regret against $\pi^\star$, the difference between her total cost and that of the optimal proper policy, defined as

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

where we use $V^\star$ as a shorthand for $V^{\pi^\star}$. The minimax optimal regret is known to be $\tilde{\mathcal{O}}(B_\star\sqrt{SAK})$, where $B_\star=\max_{s\in\mathcal{S}}V^\star(s)$, and $S=|\mathcal{S}^+|$ and $A=|\mathcal{A}|$ are the numbers of states (including the goal state) and actions respectively [Cohen et al., 2020].

Bellman Optimality Equation For a proper policy π, the corresponding action-value function $Q^\pi:\mathcal{S}\times\mathcal{A}\rightarrow[0,\infty)$ is defined as $Q^\pi(s,a)=c(s,a)+\mathbb{E}_{s'\sim P_{s,a}}[V^\pi(s')]$. Similarly, we use $Q^\star$ as a shorthand for $Q^{\pi^\star}$. it is known that $\pi^\star$ satisfies the Bellman optimality equation: $V^\star(s)=\min_{a\in\mathcal{A}}Q^\star(s,a)$ for all $s\in\mathcal{S}$ [Bertsekas and Tsitsiklis, 1991].

Assumption on $c_{\min}$ Similar to many previous works, our analysis requires $c_{\min}$ being known and strictly positive. When $c_{\min}$ is unknown or known to be 0, a simple workaround is to solve a modified SSP instance with all observed costs clipped to ϵ if they are below some $\epsilon>0$, so that $c_{\min}=\epsilon>0$. Then the regret in this modified SSP is similar to that in the original SSP up to an additive term of order $\mathcal{O}(\epsilon K)$ [Tarbouriech et al., 2020a]. Therefore, throughout the paper we assume that $c_{\min}$ is known and strictly positive unless explicitly stated otherwise.

Algorithm 1 A General Algorithmic Template for SSP

```
Initialize: $t \leftarrow 0$, $s_1 \leftarrow s_{\text{init}}$, $Q(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
for $k = 1, \ldots, K$ do
    repeat
1       Increment time step $t \overset{+}{\leftarrow} 1$.
2       Take action $a_t = \operatorname{argmin}_a Q(s_t, a)$, suffer cost $c_t$, transit to and observe $s'_t$.
3       Update $Q$ (so that it satisfies Property 1 and Property 2).
4       if $s'_t \neq g$ then $s_{t+1} \leftarrow s'_t$; else $s_{t+1} \leftarrow s_{\text{init}}$, break.
Record $T \leftarrow t$ (that is, the total number of steps).
```

Other Notations For simplicity, we use $C_K = \sum_{k=1}^{K} \sum_{i=1}^{I_k} c_i^k$ in the analysis to denote the total costs suffered by the learner over K episodes. For a function $X : \mathcal{S}^+ \rightarrow \mathbb{R}$ and a distribution P over $\mathcal{S}^+$, denote by $PX = \mathbb{E}_{S \sim P}[X(S)]$, $PX^2 = \mathbb{E}_{S \sim P}[X(S)^2]$, and $\mathbb{V}(P, X) = \textsc{Var}_{S \sim P}[X(S)]$ the expectation, second moment, and variance of $X(S)$ respectively where S is drawn from P. For a scalar x, define $(x)_+ = \max\{x, 0\}$, and denote by $\lceil x \rceil_2 = 2^{\lceil \log_2 x \rceil}$ and $\lfloor x \rfloor_2 = 2^{\lfloor \log_2 x \rfloor}$ the closest power of two upper and lower bounding x respectively. For an integer m, $[m]$ denotes the set $\{1, \ldots, m\}$. In pseudocode, $x \overset{+}{\leftarrow} y$ is a shorthand for the increment operation $x \leftarrow x + y$.

3 Implicit Finite-Horizon Approximation

In this section, we introduce our main analytical technique, that is, implicitly approximating the SSP problem with a finite-horizon counterpart. We start with a general template of our algorithms shown in Algorithm 1. For notational convenience, we concatenate state-action-cost trajectories of all episodes as one single sequence (s_t, a_t, c_t) for $t = 1, 2, \ldots, T$, where $s_t \in \mathcal{S}$ is one of the non-goal state, $a_t \in \mathcal{A}$ is the action taken at s_t, and c_t is the resulting cost incurred by the learner. Note that the goal state g is never included in this sequence (since no action is taken there), and we also use the notation $s'_t \in \mathcal{S}^+$ to denote the next-state following (s_t, a_t), so that s_{t+1} is simply s'_t unless $s'_t = g$ (in which case s_{t+1} is reset to the initial state s_{init}); see Line 4.

The template follows a rather standard idea for many reinforcement learning algorithms: maintain an (optimistic) estimate Q of the optimal action-value function $Q^\star$, and act greedily by taking the action with the smallest estimate: $a_t = \operatorname{argmin}_a Q(s_t, a)$; see Line 2. The key of the analysis is often to bound the estimation error $Q^\star(s_t, a_t) - Q(s_t, a_t)$, which is relatively straightforward in a discounted setting (where the discount factor controls the growth of the error) or a finite-horizon setting (where the error vanishes after a fixed number of steps), but becomes highly non-trivial for SSP due to the lack of similar structures.

A natural idea is to explicitly solve a discounted problem or a finite-horizon problem that approximates the original SSP well enough. Unfortunately, both approaches are problematic: approximating an undiscounted MDP by a discounted one often leads to suboptimal regret [Wei et al., 2020]; on the other hand, while explicitly approximating SSP with a finite-horizon problem can lead to optimal regret [Chen et al., 2021, Cohen et al., 2021], it greatly increases the space complexity of the algorithm, and also produces non-stationary policies, which is unnatural and introduces unnecessary complexity since the optimal policy in SSP is stationary.

Therefore, we propose to approximate the original SSP instance M with a finite-horizon counterpart $\widetilde{M}$ implicitly (that is, only in the analysis). We defer the formal definition of $\widetilde{M}$ to Appendix C, which is similar to those in [Chen et al., 2021, Cohen et al., 2021] and corresponds to interacting with the original SSP for H steps (for some integer H) and then teleporting to the goal. All we need in the analysis are the optimal value function $V_h^\star$ and optimal action-value function $Q_h^\star$ of $\widetilde{M}$ for each step $h \in [H]$, which can be defined recursively without resorting to the definition of $\widetilde{M}$:

$$Q_h^\star(s,a) = c(s,a) + P_{s,a} V_{h-1}^\star, \qquad V_h^\star(s) = \min_a Q_h^\star(s,a), \tag{1}$$

with $Q^\star_0(s,a)=0$ for all (s,a).[2] Intuitively, $Q^\star_H$ approximates $Q^\star$ well when H is large enough. This is formally summarized in the lemma below, whose proof is similar to prior works (see Appendix C).

Lemma 1. *For any value of H, $Q^\star_H(s,a)\le Q^\star(s,a)$ holds for all (s,a). For any $\beta\in(0,1)$, if $H\ge\frac{4B_\star}{c_{\min}}\ln(2/\beta)+1$, then $Q^\star(s,a)\le Q^\star_H(s,a)+B_\star\beta$ holds for all (s,a).*

In the remaining discussion, we fix a particular value of H. To carry out the regret analysis, we now specify two general requirements of the estimate Q. Let Q_t be the value of Q at the beginning of time step t (that is, the value used in finding a_t). Then Q_t needs to satisfy:

Property 1 (Optimism)**.** *With high probability, $Q_t(s,a)\le Q^\star(s,a)$ holds for all (s,a) and $t\ge1$.*

Property 2 (Recursion)**.** *There exists a "bonus overhead" $\xi_H>0$ and an absolute constant $d>0$ such that the following holds with high probability:*

$$\sum_{t=1}^{T}(\mathring{Q}(s_t,a_t)-Q_t(s_t,a_t))_+\le\xi_H+\left(1+\frac{d}{H}\right)\sum_{t=1}^{T}(\mathring{V}(s_t)-Q_t(s_t,a_t))_+,$$

for $\mathring{Q}=Q^\star_h$ and $\mathring{V}=V^\star_{h-1}$ $(h=1,\dots,H)$ as well as $\mathring{Q}=Q^\star$ and $\mathring{V}=V^\star$.[3]

Property 1 is standard and can usually be ensured by using a certain "bonus" term derived from concentration equalities in the update. These bonus terms on (s_t,a_t) accumulate into some bonus overhead in the final regret bound, which is exactly the role of ξ_H in Property 2. In both of our algorithms, ξ_H has a leading-order term $\tilde{\mathcal{O}}(\sqrt{B_\star SAC_K})$ and a lower-order term that increases in H.

Property 2 is a key property that provides a recursive form of the estimation error and allows us to connect it to the finite-horizon approximation. This is illustrated through the following two lemmas.

Lemma 2. *Property 2 implies $\sum_{t=1}^{T}(Q^\star_H(s_t,a_t)-Q_t(s_t,a_t))_+\le\mathcal{O}\left(H\xi_H\right)$.*

Proof. With $\mathring{Q}=Q^\star_H$ and $\mathring{V}=V^\star_{H-1}$, Property 2 implies

$$\begin{aligned}\sum_{t=1}^{T}(Q^\star_H(s_t,a_t)-Q_t(s_t,a_t))_+&\le\xi_H+\left(1+\frac{d}{H}\right)\sum_{t=1}^{T}(V^\star_{H-1}(s_t)-Q_t(s_t,a_t))_+\\&\le\xi_H+\left(1+\frac{d}{H}\right)\sum_{t=1}^{T}(Q^\star_{H-1}(s_t,a_t)-Q_t(s_t,a_t))_+,\end{aligned}$$

where in the last step we use the optimality of $V^\star_{H-1}$ from Eq. (1). Repeatedly applying this argument, we eventually arrive at $\sum_{t=1}^{T}(Q^\star_H(s_t,a_t)-Q_t(s_t,a_t))_+\le H\left(1+\frac{d}{H}\right)^H\xi_H+\left(1+\frac{d}{H}\right)^H\sum_{t=1}^{T}(Q^\star_0(s_t,a_t)-Q_t(s_t,a_t))_+=\mathcal{O}\left(H\xi_H\right)$, where the last step uses the facts $Q^\star_0(s_t,a_t)=0$ and $\left(1+\frac{d}{H}\right)^H\le e^d$ (an absolute constant). ☐

Lemma 3. *For any $\beta\in(0,1)$, if $H\ge\frac{4B_\star}{c_{\min}}\ln(2/\beta)+1$, then Property 1 and Property 2 together imply $\sum_{t=1}^{T}Q^\star(s_t,a_t)-V^\star(s_t)=\mathcal{O}\left(\beta C_K+\xi_H\right)$.*

Proof. Applying Property 2 with $\mathring{Q}=Q^\star$ and $\mathring{V}=V^\star$, we have $\sum_{t=1}^{T}(Q^\star(s_t,a_t)-Q_t(s_t,a_t))_+\le\xi_H+\left(1+\frac{d}{H}\right)\sum_{t=1}^{T}(V^\star(s_t)-Q_t(s_t,a_t))_+$. Now note that by Property 1, the Bellman optimality equation $V^\star(s_t)=\min_a Q^\star(s_t,a)$, and the fact $Q_t(s_t,a_t)=\min_a Q_t(s_t,a)$ (by the definition of a_t), the arguments within the clipping operation $(\cdot)_+$ are all non-negative and thus the clipping can be removed. Rearranging terms then gives

$$\begin{aligned}\sum_{t=1}^{T}Q^\star(s_t,a_t)-V^\star(s_t)&\le\xi_H+\frac{d}{H}\sum_{t=1}^{T}(V^\star(s_t)-Q_t(s_t,a_t))\\&\le\xi_H+\frac{d}{H}\sum_{t=1}^{T}(Q^\star(s_t,a_t)-Q_t(s_t,a_t)).\qquad\text{(optimality of }V^\star\text{)}\end{aligned}$$

---

[2] Note that our notation is perhaps unconventional compared to most works on finite-horizon MDPs, where $Q^\star_h$ usually refers to our $Q^\star_{H-h}$. We make this switch since we want to highlight the dependence on H for $Q^\star_H$.

[3] Note that ξ_H might be a random variable. In fact, it often depends on C_K.

It remains to bound the last term using the finite-horizon approximation $Q^\star_H$ as a proxy:

$$\begin{aligned}\sum_{t=1}^{T}(Q^\star(s_t,a_t)-Q_t(s_t,a_t)) &= \sum_{t=1}^{T}(Q^\star(s_t,a_t)-Q^\star_H(s_t,a_t)+Q^\star_H(s_t,a_t)-Q_t(s_t,a_t))\\ &= \mathcal{O}\left(TB_\star\beta + H\xi_H\right),\end{aligned}$$

where the last step uses Lemma 1 and Lemma 2. Importantly, this term is finally scaled by d/H, which, together with the fact $\frac{TB_\star}{H} \leq c_{\min}T \leq C_K$, proves the claimed bound. □

Readers familiar with the literature might already recognize the term $\sum_{t=1}^{T} Q^\star(s_t,a_t) - V^\star(s_t)$ considered in Lemma 3, which is closely related to the regret. Indeed, with this lemma, we can conclude a regret bound for our generic algorithm.

Theorem 1. *For any $\beta \in (0,1)$, if $H \geq \frac{4B_\star}{c_{\min}}\ln(2/\beta)+1$, then Algorithm 1 ensures (with high probability) $R_K = \tilde{\mathcal{O}}\left(\sqrt{B_\star C_K} + B_\star + \beta C_K + \xi_H\right)$.*

Proof. We first decompose the regret as follows, which holds generally for any algorithm:

$$\begin{aligned}R_K &= \sum_{k=1}^{K}\left(\sum_{i=1}^{I_k} c_i^k - V^\star(s_1^k)\right)\\ &\leq \sum_{k=1}^{K}\sum_{i=1}^{I_k}\left(c_i^k - V^\star(s_i^k) + V^\star(s_{i+1}^k)\right) = \sum_{t=1}^{T}(c_t - V^\star(s_t) + V^\star(s'_t))\\ &= \sum_{t=1}^{T}(c_t - c(s_t,a_t)) + \sum_{t=1}^{T}(V^\star(s'_t) - P_{s_t,a_t}V^\star) + \sum_{t=1}^{T}(Q^\star(s_t,a_t) - V^\star(s_t)).\end{aligned} \tag{2}$$

The first and the second term are the sum of a martingale difference sequence (since s'_t is drawn from P_{s_t,a_t}) and can be bounded by $\tilde{\mathcal{O}}\left(\sqrt{C_K}\right)$ and $\tilde{\mathcal{O}}\left(\sqrt{B_\star C_K} + B_\star\right)$ respectively using concentration inequalities; see Lemma 4, Lemma 35, and Lemma 5. The third term can be bounded using Lemma 3 directly, which finishes the proof. □

To get a sense of the regret bound in Theorem 1, first note that since $1/\beta$ only appears in a logarithmic term of the required lower bound of H, one can pick β to be small enough so that the term βC_K is dominated by others. Moreover, if ξ_H is $\tilde{\mathcal{O}}(\sqrt{B_\star SAC_K})$ plus some lower-order term ρ_H (which as mentioned is the case for our algorithms), then by solving a quadratic of $\sqrt{C_K}$, the regret bound of Theorem 1 implies $R_K = \tilde{\mathcal{O}}(B_\star\sqrt{SAK} + \rho_H)$, which is minimax optimal (ignoring ρ_H)!

Based on this analytical technique, it remains to design algorithms satisfying the two required properties. In the following sections, we provide two such examples, leading to the first model-free SSP algorithm and an improved model-based SSP algorithm.

4 The First Model-free Algorithm: LCB-ADVANTAGE-SSP

In this section, we present a model-free algorithm (the first in the literature) called LCB-ADVANTAGE-SSP that falls into our generic template and satisfies the required properties. It is largely inspired by the state-of-the-art model-free algorithm UCB-ADVANTAGE [Zhang et al., 2020b] for the finite-horizon problem. The pseudocode is shown in Algorithm 2, with only the lines instantiating the update rule of the Q estimates numbered. Importantly, the space complexity of this algorithm is only $\mathcal{O}(SA)$ since we do not estimate the transition directly or conduct explicit finite-horizon reduction, and the time complexity is only $\mathcal{O}(1)$ in each step.

Specifically, for each state-action pair (s,a), we divide the samples received when visiting (s,a) into consecutive stages of exponentially increasing length, and only update $Q(s,a)$ at the end of a stage. The number of samples e_j in stage j is defined through $e_1 = H$ and $e_{j+1} = \lfloor(1+1/H)e_j\rfloor$ for some parameter H. Further define $\mathcal{L}^\star = \{E_j\}_{j\in\mathbb{N}^+}$ with $E_j = \sum_{i=1}^{j} e_i$, which contains all the indices indicating the end of some stage. As mentioned, the algorithm only updates $Q(s,a)$ when the

Algorithm 2 LCB-ADVANTAGE-SSP

```
Parameters: horizon $H$, threshold $\theta^\star$, and failure probability $\delta \in (0,1)$.
Define: $\mathcal{L}^\star = \{E_j\}_{j \in \mathbb{N}^+}$ where $E_j = \sum_{i=1}^{j} e_i$, $e_1 = H$ and $e_{j+1} = \lfloor (1+1/H)e_j \rfloor$.
Initialize: $t \leftarrow 0$, $s_1 \leftarrow s_{\text{init}}$, $B \leftarrow 1$, for all $(s,a)$, $N(s,a) \leftarrow 0, M(s,a) \leftarrow 0$.
Initialize: for all $(s,a)$, $Q(s,a) \leftarrow 0, V(s) \leftarrow 0, V^{\text{ref}}(s) \leftarrow V(s), \widehat{C}(s,a) \leftarrow 0$.
Initialize: for all $(s,a)$, $\mu^{\text{ref}}(s,a) \leftarrow 0, \sigma^{\text{ref}}(s,a) \leftarrow 0, \mu(s,a) \leftarrow 0, \sigma(s,a) \leftarrow 0, v(s,a) \leftarrow 0$.
for $k = 1, \ldots, K$ do
    repeat
        Increment time step $t \overset{+}{\leftarrow} 1$.
        Take action $a_t = \operatorname{argmin}_a Q(s_t, a)$, suffer cost $c_t$, transit to and observe $s'_t$.
1       Increment visitation counters: $n = N(s_t,a_t) \overset{+}{\leftarrow} 1, m = M(s_t,a_t) \overset{+}{\leftarrow} 1$.
2       Update global accumulators: $\mu^{\text{ref}}(s_t,a_t) \overset{+}{\leftarrow} V^{\text{ref}}(s'_t)$, $\sigma^{\text{ref}}(s_t,a_t) \overset{+}{\leftarrow} V^{\text{ref}}(s'_t)^2$, $\widehat{C}(s_t,a_t) \overset{+}{\leftarrow} c_t$.
3       Update local accumulators: $v(s_t,a_t) \overset{+}{\leftarrow} V(s'_t)$, $\mu(s_t,a_t) \overset{+}{\leftarrow} V(s'_t) - V^{\text{ref}}(s'_t)$, $\sigma(s_t,a_t) \overset{+}{\leftarrow} (V(s'_t) - V^{\text{ref}}(s'_t))^2$.
4       if $n \in \mathcal{L}^\star$ then
5           Compute $\iota \leftarrow 256 \ln^6(4SAB_\star^8 n^5/\delta)$, cost estimator $\widehat{c} = \frac{\widehat{C}(s_t,a_t)}{n}$, bonuses $b' \leftarrow 2\sqrt{\frac{B^2\iota}{m}} + \sqrt{\frac{\widehat{c}\iota}{n}} + \frac{\iota}{n}$ and $b \leftarrow$
            $\sqrt{\frac{\sigma^{\text{ref}}(s_t,a_t)/n - (\mu^{\text{ref}}(s_t,a_t)/n)^2}{n}}\iota + \sqrt{\frac{\sigma(s_t,a_t)/m - (\mu(s_t,a_t)/m)^2}{m}}\iota + \left(\frac{4B}{n} + \frac{3B}{m}\right)\iota + \sqrt{\frac{\widehat{c}\iota}{n}}$.
6           $Q(s_t,a_t) \leftarrow \max\left\{\widehat{c} + \frac{v(s_t,a_t)}{m} - b', Q(s_t,a_t)\right\}$.
7           $Q(s_t,a_t) \leftarrow \max\left\{\widehat{c} + \frac{\mu^{\text{ref}}(s_t,a_t)}{n} + \frac{\mu(s_t,a_t)}{m} - b, Q(s_t,a_t)\right\}$.
8           $V(s_t) \leftarrow \min_a Q(s_t,a)$.
9           if $V(s_t) > B$ then $B \leftarrow 2V(s_t)$.
10          Reset local accumulators: $v(s_t,a_t) \leftarrow 0, \mu(s_t,a_t) \leftarrow 0, \sigma(s_t,a_t) \leftarrow 0, M(s_t,a_t) \leftarrow 0$.
11      if $\sum_a N(s_t,a)$ is a power of two not larger than $\theta^\star$ then $V^{\text{ref}}(s_t) \leftarrow V(s_t)$.
        if $s'_t \neq g$ then $s_{t+1} \leftarrow s'_t$; else $s_{t+1} \leftarrow s_{\text{init}}$, break.
```

total number of visits to (s,a) falls into the set $\mathcal{L}^\star$ (Line 4). The algorithm also maintains an estimate V for $V^\star$, which always satisfies $V(s) = \min_a Q(s,a)$ (Line 8), and importantly another reference value function V^{ref} whose role and update rule are to be discussed later.

In addition, some local and global accumulators are maintained in the algorithm. Local accumulators only store information related to the current stage. These include: $M(s,a)$, the number of visits to (s,a) within the current stage; $v(s,a)$, the cumulative value of $V(s')$ within the current stage, where s' represents the next state after each visit to (s,a); and finally $\mu(s,a)$ and $\sigma(s,a)$, the cumulative values of $V(s') - V^{\text{ref}}(s')$ and its square respectively within the current stage (Line 3). These local accumulators are reset to zero at the end of each stage (Line 10).

On the other hand, global accumulators store information related to all stages and are never reset. These include: $N(s,a)$, the number of visits to (s,a) from the beginning; $\widehat{C}(s,a)$, total cost incurs at (s,a) from the beginning; and $\mu^{\text{ref}}(s,a)$ and $\sigma^{\text{ref}}(s,a)$, the cumulative value of $V^{\text{ref}}(s')$ and its square respectively from the beginning, where again s' represents the next state after each visit to (s,a) (Line 2).

We are now ready to describe the update rule of Q. The first update, Line 6, is intuitively based on the equality $Q^\star(s,a) = c(s,a) + P_{s,a}V^\star$ and uses $v(s,a)/M(s,a)$ as an estimate for $P_{s,a}V^\star$ together with a (negative) bonus b' derived from Azuma's inequality (Line 5). As mentioned, the bonus is necessary to ensure Property 1 (optimism) so that Q is always a lower confidence bound of $Q^\star$ (hence the name "LCB"). Note that this update only uses data from the current stage (roughly $1/H$ fraction of the entire data collected so far), which leads to an extra $\sqrt{H}$ factor in the regret.

To address this issue, Zhang et al. [2020b] introduce a variance reduction technique via a reference-advantage decomposition, which we borrow here leading to the second update rule in Line 7. This is intuitively based on the decomposition $P_{s,a}V^\star = P_{s,a}V^{\text{ref}} + P_{s,a}(V^\star - V^{\text{ref}})$, where $P_{s,a}V^{\text{ref}}$ is approximated by $\mu^{\text{ref}}(s,a)/N(s,a)$ and $P_{s,a}(V^\star - V^{\text{ref}})$ is approximated by $\mu(s,a)/M(s,a)$. In addition, a "variance-aware" bonus term b is applied, which is derived from a tighter Freedman's inequality (Line 5). The reference function V^{ref} is some snapshot of the past value of V, and is guaranteed to be $\mathcal{O}(c_{\min})$ close to $V^\star$ on a particular state as long as the number of visits to this state exceeds some threshold $\theta^\star = \tilde{\mathcal{O}}\left(B_\star^2 H^3 SA/c_{\min}^2\right)$ (Line 11). Overall, this second update rule not only removes the extra $\sqrt{H}$ factor as in [Zhang et al., 2020b], but also turns some terms of order $\tilde{\mathcal{O}}(\sqrt{T})$ into $\tilde{\mathcal{O}}(\sqrt{C_K})$ in our context, which is important for obtaining the optimal regret.

Despite the similarity, we emphasize several key differences between our algorithm and that of [Zhang et al., 2020b]. First, [Zhang et al., 2020b] maintains a different Q estimate for each step of an episode (which is natural for a finite-horizon problem), while we only maintain one Q estimate (which is natural for SSP). Second, we update the reference function $V^{\text{ref}}(s)$ whenever the number of visits to s doubles (while still below the threshold $\theta^\star$; see Line 11), instead of only updating it once as in [Zhang et al., 2020b]. We show in Lemma 8 that this helps reduce the sample complexity and leads to a smaller lower-order term in the regret. Third, since there is no apriori known upper bound on V (unlike the finite-horizon setting), we maintain an empirical upper bound B (in a doubling manner) such that $V(s) \leq B \leq 2B_\star$ (Line 9), which is further used in computing the bonus terms b and b'. This is important for eventually developing a parameter-free algorithm.

In Appendix D, we show that Algorithm 2 indeed satisfies the two required properties.

Theorem 2. *Let* $H = \lceil \frac{4B_\star}{c_{\min}} \ln(\frac{2}{\beta}) + 1 \rceil_2$ *for* $\beta = \frac{c_{\min}}{2B_\star^2 SAK}$ *and* $\theta^\star = \tilde{\mathcal{O}}\left(\frac{B_\star^2 H^3 SA}{c_{\min}^2}\right)$ *be defined in Lemma 8, then Algorithm 2 satisfies Property 1 and Property 2 with* $d = 3$ *and* $\xi_H = \tilde{\mathcal{O}}\left(\sqrt{B_\star SAC_K} + \frac{B_\star^2 H^3 S^2 A}{c_{\min}}\right)$.

Proof Sketch. The proof of Property 1 largely follows the analysis of [Zhang et al., 2020b, Proposition 4] for the designed bonuses. To prove Property 2, similarly to [Zhang et al., 2020b] we can show:

$$\sum_{t=1}^{T} (\mathring{Q}(s_t,a_t) - Q_t(s_t,a_t))_+ \lesssim \xi_H + \sum_{t=1}^{T} \frac{1}{m_t} \sum_{i=1}^{m_t} P_{s_{\check{l}_{t,i}}, a_{\check{l}_{t,i}}} (\mathring{V} - V_{\check{l}_{t,i}})_+,$$

where m_t is the value of m used in computing $Q_t(s_t,a_t)$, and $\check{l}_{t,i}$ is the i-th time step the agent visits (s_t,a_t) among those m_t steps. Now it suffices to show that $\sum_{t=1}^{T} \frac{1}{m_t} \sum_{i=1}^{m_t} P_{s_{\check{l}_{t,i}}, a_{\check{l}_{t,i}}} (\mathring{V} - V_{\check{l}_{t,i}})_+ \lesssim (1 + \frac{3}{H}) \sum_{t=1}^{T} (\mathring{V}(s_t) - V_t(s_t))_+$, which is proven in Lemma 13. □

As a direct corollary of Theorem 1, we arrive at the following regret guarantee.

Theorem 3. *With the same parameters as in Theorem 2, with probability at least* $1 - 60\delta$, *Algorithm 2 ensures* $R_K = \tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + \frac{B_\star^5 S^2 A}{c_{\min}^4}\right)$.

We make several remarks on our results. First, while Algorithm 2 requires setting the two parameters H and $\theta^\star$ in terms of $B_\star$ to obtain the claimed regret bound, one can in fact achieve the exact same bound without knowing $B_\star$ by slightly changing the algorithm. The high level idea is to first apply the doubling trick from Tarbouriech et al. [2021b] to determine an upper bound on $B_\star$, then try logarithmically many different values of H and $\theta^\star$ simultaneously, each leading to a different update rule for Q and V^{ref}. This only increases the time and space complexity by a logarithmic factor, without hurting the regret (up to log factors). Details are deferred to Section D.5.

Second, as mentioned in Section 2, when $c_{\min}$ is unknown or $c_{\min} = 0$, one can clip all observed costs to ϵ if they are below $\epsilon > 0$, which introduces an additive regret term of order $\mathcal{O}(\epsilon K)$. By picking ϵ to be of order $K^{-1/5}$, our bound becomes $\tilde{\mathcal{O}}\left(K^{4/5}\right)$ ignoring other parameters. Although most existing works suffer the same issue, this is certainly undesirable, and our second algorithm to be introduced in the next section completely avoids this issue by having only logarithmic dependence on $1/c_{\min}$.

Algorithm 3 SVI-SSP

```
Parameters: horizon $H$, value function upper bound $B$, and failure probability $\delta \in (0,1)$.
Define: $\mathcal{L} = \{E_j\}_{j \in \mathbb{N}^+}$, where $E_j = \sum_{i=1}^{j} e_i$, $e_j = \lfloor \widetilde{e}_j \rfloor$, and $\widetilde{e}_1 = 1, \widetilde{e}_{j+1} = \widetilde{e}_j + \frac{1}{H} e_j$.
Initialize: $t \leftarrow 0, s_1 \leftarrow s_{\text{init}}$.
Initialize: for all $(s,a,s')$, $n(s,a,s') \leftarrow 0, n(s,a) \leftarrow 0, Q(s,a) \leftarrow 0, V(s) \leftarrow 0, \widehat{C}(s,a) \leftarrow 0$.
for $k = 1, \ldots, K$ do
    repeat
        Increment time step $t \overset{+}{\leftarrow} 1$.
        Take action $a_t = \mathrm{argmin}_a Q(s_t, a)$, suffer cost $c_t$, transit to and observe $s'_t$.
1       Update accumulators: $n = n(s_t,a_t) \overset{+}{\leftarrow} 1, n(s_t,a_t,s'_t) \overset{+}{\leftarrow} 1, \widehat{C}(s_t,a_t) \overset{+}{\leftarrow} c_t$.
2       if $n \in \mathcal{L}$ then
3           Update empirical transition: $\bar{P}_{s_t,a_t}(s') \leftarrow \frac{n(s_t,a_t,s')}{n}$ for all $s'$.
4           Compute $\iota \leftarrow 20 \ln \frac{2SAn}{\delta}$, cost estimator $\widehat{c} \leftarrow \frac{\widehat{C}(s,a)}{n}$, and bonus $b \leftarrow \max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{s_t,a_t}, V)\iota}{n}}, \frac{49B\iota}{n}\right\} + \sqrt{\frac{\widehat{c}\iota}{n}}$.
5           $Q(s_t,a_t) \leftarrow \max\{\widehat{c} + \bar{P}_{s_t,a_t} V - b, Q(s_t,a_t)\}$.
6           $V(s_t) \leftarrow \mathrm{argmin}_a Q(s_t,a)$.
        if $s'_t \neq g$ then $s_{t+1} \leftarrow s'_t$; else $s_{t+1} \leftarrow s_{\text{init}}$, break.
```

Finally, we point out that, just as in the finite-horizon case, the variance reduction technique is crucial for obtaining the minimax optimal regret. For example, if one instead uses an update rule similar to the (suboptimal) Q-learning algorithm of [Jin et al., 2018], then this is essentially equivalent to removing the second update (Line 7) of our algorithm. While this still satisfies Property 2, the bonus overhead ξ_H would be $\sqrt{H}$ times larger, resulting in a suboptimal leading term in the regret.

5 An Optimal and Efficient Model-based Algorithm: SVI-SSP

In this section, we propose a simple model-based algorithm called SVI-SSP (Sparse Value Iteration for SSP) following our template, which not only achieves the minimax optimal regret even when $c_{\min} = 0$, matching the state-of-the-art by a recent work [Tarbouriech et al., 2021b], but also admits highly sparse updates, making it more efficient than all existing model-based algorithms. The pseudocode is in Algorithm 3, again with only the lines instantiating the update rule for Q numbered.

Similar to Algorithm 2, SVI-SSP divides samples of each (s,a) into consecutive stages of (roughly) exponentially increasing length, and only update $Q(s,a)$ at the end of a stage (Line 2). However, the number of samples e_j in stage j is defined slightly differently through $e_j = \lfloor \widetilde{e}_j \rfloor, \widetilde{e}_1 = 1$, and $\widetilde{e}_{j+1} = \widetilde{e}_j + \frac{1}{H} e_j$ for some parameter H. In the long run, this is almost the same as the scheme used in Algorithm 2, but importantly, it forces more frequent updates at the beginning — for example, one can verify that $e_1 = \cdots = e_H = 1$, meaning that $Q(s,a)$ is updated every time (s,a) is visited for the first H visits. This slight difference turns out to be important to ensure that the lower-order term in the regret has no poly(H) dependence, as shown in Lemma 16 and further discussed in Remark 3. More intuition on the design of this update scheme is provided in Section E.1.

The update rule for Q is very simple (Line 5). It is again based on the equality $Q^\star(s,a) = c(s,a) + P_{s,a}V^\star$, but this time uses $\bar{P}_{s,a}V - b$ as an approximation for $P_{s,a}V^\star$, where $\bar{P}_{s,a}$ is the empirical transition directly calculated from two counters $n(s,a)$ and $n(s,a,s')$ (number of visits to (s,a) and (s,a,s') respectively), V is such that $V(s) = \min_a Q(s,a)$, and b is a special bonus term (Line 4) adopted from [Tarbouriech et al., 2021b, Zhang et al., 2020a] which ensures that Q is an optimistic estimate of $Q^\star$ and also helps remove poly(H) dependence in the regret.

SVI-SSP exhibits a unique structure compared to existing algorithms. In each update, it modifies only one entry of Q (similarly to model-free algorithms), while other model-based algorithms such as [Tarbouriech et al., 2021b] perform value iteration for every entry of Q repeatedly until convergence (concrete time complexity comparisons to follow). We emphasize that our implicit finite-horizon analysis is indeed the key to enable us to derive a regret guarantee for such a sparse value iteration algorithm. Specifically, in Appendix E, we show that SVI-SSP satisfies the two required properties.

Theorem 4. *If $B \geq B_\star$ and $H = \lceil \frac{4B}{c_{\min}} \ln(\frac{2}{\beta}) + 1 \rceil_2$ for $\beta = \frac{c_{\min}}{2B^2SAK}$, then Algorithm 3 satisfies Property 1 and Property 2 with $d = 1$ and $\xi_H = \tilde{\mathcal{O}}(\sqrt{B_\star SAC_K} + BS^2A + \beta C_K)$, where the dependence on H in ξ_H is hidden in logarithmic terms.*

Proof Sketch. The proof of Property 1 largely follows the analysis of [Tarbouriech et al., 2021b, Lemma 15]. To prove Property 2, we first show $\sum_{t=1}^T (\mathring{Q}(s_t, a_t) - Q_t(s_t, a_t))_+ \lesssim \xi_H + \sum_{t=1}^T P_t(\mathring{V} - V_{l_t})_+$, where l_t is the last time step $Q(s_t, a_t)$ is updated. Then, the remaining main steps are shown below with all details deferred to the corresponding key lemmas:

$$\begin{aligned}
\sum_{t=1}^T P_t(\mathring{V} - V_{l_t})_+ &\lesssim \left(1 + \frac{1}{H}\right) \sum_{t=1}^T P_t(\mathring{V} - V_t)_+ && \text{(Lemma 16)} \\
&\lesssim \left(1 + \frac{1}{H}\right) \sum_{t=1}^T (\mathring{V}(s_t) - V_t(s_t))_+ + \left(1 + \frac{1}{H}\right) \sum_{t=1}^T (P_t - \mathbb{I}_{s'_t})(\mathring{V} - V_t)_+ \\
&\lesssim \left(1 + \frac{1}{H}\right) \sum_{t=1}^T (\mathring{V}(s_t) - V_t(s_t))_+ + \xi_H, && \text{(Lemma 22 and Lemma 21)}
\end{aligned}$$

which completes the proof. ☐

Again, as a direct corollary of Theorem 1, we arrive at the following regret guarantee.

Theorem 5. *With the same parameters as in Theorem 4, with probability at least $1 - 12\delta$, Algorithm 3 ensures $R_K = \tilde{\mathcal{O}}(B_\star\sqrt{SAK} + BS^2A)$.*

Setting $B = B_\star$, our bound becomes $\tilde{\mathcal{O}}(B_\star\sqrt{SAK} + B_\star S^2A)$, which is minimax optimal even when $c_{\min}$ is unknown or $c_{\min} = 0$ (this is because the dependence on $1/c_{\min}$ is only logarithmic, and one can clip all observed costs to ϵ if they are below $\epsilon = 1/K$ in this case without introducing poly(K) overhead to the regret). When $B_\star$ is unknown, we can use the same doubling trick from Tarbouriech et al. [2021b] to obtain almost the same bound (with only the lower-order term increased to $\tilde{\mathcal{O}}\left(B_\star^3 S^3 A\right)$); see Section E.5 for details.[4]

Comparison with EB-SSP [Tarbouriech et al., 2021b] Our regret bounds match exactly the state-of-the-art by Tarbouriech et al. [2021b]. Thanks to the sparse update, however, SVI-SSP has a much better time complexity. Specifically, for SVI-SSP, each (s, a) is updated at most $\tilde{\mathcal{O}}(H) = \tilde{\mathcal{O}}({}^{B_\star}/_{c_{\min}})$ times (Lemma 16), and each update takes $\mathcal{O}(S)$ time, leading to total complexity $\tilde{\mathcal{O}}({}^{B_\star S^2 A}/_{c_{\min}})$. On the other hand, for EB-SSP, although each (s, a) only causes $\tilde{\mathcal{O}}(1)$ updates, each update runs value iteration on all entries of Q until convergence, which takes $\tilde{\mathcal{O}}({}^{B_\star^2 S^2}/_{c_{\min}^2})$ iterations (see their Appendix C) and leads to total complexity $\tilde{\mathcal{O}}({}^{B_\star^2 S^5 A}/_{c_{\min}^2})$, much larger than ours.

Comparison with ULCVI [Cohen et al., 2021] Another recent work by Cohen et al. [2021] using explicit finite-horizon approximation also achieves minimax regret but requires the knowledge of some hitting time of the optimal policy. Without this knowledge, their bound has a large $1/c_{\min}^4$ dependence in the lower-order term just as our model-free algorithm. Our results in this section show that implicit finite-horizon approximation has advantage over explicit approximation apart from reducing space complexity: the former does not necessarily introduce poly(H) dependence even for the lower-order term, while the latter does under the current analysis.

Acknowledgments and Disclosure of Funding

LC thanks Chen-Yu Wei for many helpful discussions. HL is supported by NSF Award IIS-1943607 and a Google Faculty Research Award. MJ and RJ's research is supported by NSF CCF-1817212, NSF ECCS-1810447 and ONR N00014-20-1-2258 awards.

---

[4] We note that this doubling trick is in fact also applicable to Algorithm 2. However, the specific approach we propose for this algorithm in Section D.5 is better in the sense that it does not worsen the regret at all.

References

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

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.

Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, 2021.

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*, 2021.

Alon Cohen, Haim Kaplan, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret bounds for stochastic shortest path. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8210–8219. PMLR, 2020.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. In *Neural Information Processing Systems*, 2021.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yonathan Efroni, Nadav Merlis, Aadirupa Saha, and Shie Mannor. Confidence-budget matching for sequential budgeted learning. In *International Conference on Machine Learning*, 2021.

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in neural information processing systems*, pages 4863–4873, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4860–4869, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 15522–15533. Curran Associates, Inc., 2020.

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

Soroush Nasiriany, Vitchyr Pong, Steven Lin, and Sergey Levine. Planning with goal-conditioned policies. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813, 2012.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5478–5486, 2019.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. *arXiv preprint arXiv:2006.11561*, 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*, pages 9428–9437. PMLR, 2020a.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Improved sample complexity for incremental autonomous exploration in MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 11273–11284. Curran Associates, Inc., 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*, pages 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. In *Neural Information Processing Systems*, 2021b.

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

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10170–10180. PMLR, 2020.

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

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7304–7312, 2019.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference On Learning Theory*, 2020a.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33, pages 15198–15207. Curran Associates, Inc., 2020b.

Alexander Zimin and Gergely Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2013.

Checklist

1. For all authors...
    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
    (b) Did you describe the limitations of your work? [Yes] see Related Work
    (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work is mainly theoretical, and we do not foresee any potential negative societal impacts.
    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
    (a) Did you state the full set of assumptions of all theoretical results? [Yes]
    (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] see Appendix H (same for questions (b)-(d))
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Table 1: Summary of existing regret minimization algorithms for SSP with their best achievable bounds (assuming necessary prior knowledge). Here, D, S, A are the diameter, number of states, and number of actions of the MDP, $T_\star$ is the maximum expected hitting time of the optimal policy over all states, $B_\star$ is the maximum expected costs of the optimal policy over all states, and K is the number of episodes.

Algorithm	Regret Bound
UC-SSP [Tarbouriech et al., 2020a]	$\tilde{\mathcal{O}}\left(DS\sqrt{DAK/c_{\min}} + S^2AD^2\right)$
Bernstein-SSP [Cohen et al., 2020]	$\tilde{\mathcal{O}}\left(B_\star S\sqrt{AK} + \sqrt{B_\star^3 S^2 A^2/c_{\min}}\right)$
ULCVI [Cohen et al., 2021]	$\tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + T_\star^4 S^2 A\right)$
EB-SSP [Tarbouriech et al., 2021b]	$\tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + B_\star S^2 A\right)$
LCB-ADVANTAGE-SSP **(Ours)**	$\tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + B_\star^5 S^2 A/c_{\min}^4\right)$
SVI-SSP **(Ours)**	$\tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + B_\star S^2 A\right)$

A A Summary of Existing Bounds

A summary of existing regret minimization algorithms for SSP and their regret bounds is shown in Table 1. Note that although LCB-ADVANTAGE-SSP has a larger lower order term depending on $\tilde{\mathcal{O}}(1/c_{\min}^4)$ among the minimax optimal algorithms, it actually nearly matches that of ULCVI when $T_\star$ is unknown, in which case their algorithm is run with $T_\star$ replaced by its upper bound $B_\star/c_{\min}$.

Time Complexity When $c_{\min} = 0$, the cost perturbation trick is applied (see paragraph "Assumption on $c_{\min}$" in Section 2 for more details) and $1/c_{\min}$ becomes a K-dependent quantity. This leads to a worse K-dependent time complexity for all algorithms in Table 1 except ULCVI. In fact, this seems to be a shared limitation of all algorithms that learns a stationary policy. On the other hand, when $T_\star$ is known, ULCVI (which learns a non-stationary policy) gives a better time complexity with no polynomial dependency on K. How to learn a stationary policy while avoiding K-dependent time complexity when $c_{\min} = 0$ is an interesting future direction.

B Preliminaries for the Appendix

Extra Notations in Appendix Denote by $\Delta_{\mathcal{X}}$ the simplex over set $\mathcal{X}$. For conciseness, throughout the appendix, we use the following notational shorthands:

- $\mathbb{I}_s(s') = \mathbb{I}\{s = s'\}$;
- $P_t = P_{s_t,a_t}$;
- for a function $f_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we often abuse the notation and use f_t to denote $f_t(s_t, a_t)$ when there is no confusion from the context; in fact, in Lemma 9 and Lemma 18, we also use f_t to denote $f_t(s, a)$ for a particular (s, a) pair;
- $\mathcal{V}_H = \{(Q^\star, V^\star)\} \cup \{(Q_h^\star, V_{h-1}^\star)\}_{h=1}^H$.

Note that for any $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$, we have $\mathring{Q}(s,a) = c(s,a) + P_{s,a}\mathring{V}$, $\mathring{V}(s) \in [0, B_\star]$, $\mathring{V}(g) = 0$ and $\mathring{V}(s) \leq \min_a \mathring{Q}(s,a)$. Throughout the paper, $\tilde{\mathcal{O}}\left(\cdot\right)$ also hides dependence on $\ln(1/\delta)$ and $\ln T$ where $\delta \in (0, e^{-1}]$ is some failure probability, and T is a random variable but can be bounded by $\frac{C_K}{c_{\min}}$ under strictly positive costs. We include a summary of most notations in Table 2.

Truncating the Interaction An important question in SSP is whether the algorithm halts in a finite number of steps. To implicitly show this, we do the following trick throughout the analysis. Fix any positive integer T' and explicitly stop the algorithm after T' steps. Our analysis will show that in

---

[5] In Table 2, "the current stage" means the current stage of (s, a) at time step t, and "the last stage" means the last stage of (s, a) before time step t.

Table 2: Explanation of the notations

β	precision of the implicit finite horizon approximation;
Q_t, V_t	accumulators Q, V at the beginning of time step t;
$Q^\star, V^\star$	optimal value functions of the SSP instance;
$Q^\star_h, V^\star_h$	optimal value functions of taking h steps in the SSP instance and then teleporting to the goal state; see Eq. (1)
C_K	total costs the agent suffers in K episodes;
V_t^{ref}	reference value function at the beginning of time step t;
V^{REF}	reference value function at the end of learning; see Lemma 8
C_{REF}	costs in regret of using reference value function; see Lemma 8
$C_{\text{REF, 2}}$	another costs in the regret of using reference value function; see Lemma 8
B_t	an upper bound of estimated value function V_t;
$\widehat{c}_t(s,a)$	cost estimator used in the last update of $Q_t(s,a)$;
$n_t(s,a)$	the number of visits to (s,a) before the current stage;[5]
$m_t(s,a)$	the number of visits to (s,a) in the last stage;
$b_t(s,a), b'_t(s,a)$	bonus terms used in the last update of $Q_t(s,a)$;
$l_{t,i}(s,a)$	the i-th time step the agent visits (s,a) among those $n_t(s,a)$ steps before the current stage;
$\check{l}_{t,i}(s,a)$	the i-th time step the agent visits (s,a) among those $m_t(s,a)$ steps within the last stage;
$l_t(s,a)$	the last time step the agent visits (s,a) before the current stage;
$\iota_t(s,a)$	logarithmic terms used in the last update of $Q_t(s,a)$;
ε_t	indicator of whether time step t is in the first stage of (s_t, a_t);
ν_t	empirical variance of the advantage (i.e., the difference between the estimate value function and the reference value function) at time step t; see Eq. (4)
ν_t^{ref}	empirical variance of the reference value function at time step t; see Eq. (5)
$\bar{P}_{t,s,a}$	empirical transition at (s,a) at the beginning of the current stage of (s,a);
e_j, E_j	the length of the j-th stage and the total length of the first j stages;

this case the regret R_K is bounded by something independent of T', which then allows us to take T' to infinity and recover the original setting while maintaining the same bound. This also implicitly shows that the algorithm must halt in a finite number of steps.

C Omitted Details for Section 3

In this section, we provide omitted details and proofs for Section 3. We first introduce the class of finite horizon MDPs used in the approximation: given an SSP model $M = (\mathcal{S}, \mathcal{A}, s_{\text{init}}, g, c, P)$, we consider the costs of interacting with M for at most H steps and then directly teleporting to the goal state. Specifically, we define a finite-horizon SSP $\widetilde{M} = (\widetilde{\mathcal{S}}, \mathcal{A}, \widetilde{s}_{\text{init}}, g, \widetilde{c}, \widetilde{P})$ as follows:

- $\widetilde{\mathcal{S}} = \mathcal{S} \times [H]$, $\widetilde{s}_{\text{init}} = (s_{\text{init}}, 1)$ and the goal state g remains the same;
- transition from (s,h) to (s',h') is only possible when $h' = h+1$, and the transition follows the original MDP: $\widetilde{P}((s', h+1)|(s,h),a) = P(s'|s,a)$ for $h \in [H-1]$ and $\widetilde{P}(g|(s,H),a) = 1$;
- mean cost function also follows the original MDP: $\widetilde{c}_k((s,h),a) = c_k(s,a)$.

We also define $Q^\star_0(s,a)=V^\star_0(s)=0$, $Q^\star_h(s,a)=\widetilde{Q}^\star((s,H-h+1),a)$, $V^\star_h(s)=\widetilde{V}^\star(s,H-h+1)$ for $h\in[H]$, where $\widetilde{Q}^\star$ and $\widetilde{V}^\star$ are optimal state-action and state value functions in $\widetilde{M}$. Then, it is straightforward to verify that $Q^\star_h$ and $V^\star_h$ satisfy Eq. (1). Since M is equivalent to $\widetilde{M}$ with $H=\infty$, intuitively we should have $Q^\star(s,a)\approx Q^\star_H(s,a)$ for a sufficiently large H. The formal statement, shown in Lemma 1, is proven below:

Proof of Lemma 1. By definition $Q^\star_h(s,a)\leq Q^\star(s,a)$ holds for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $h\in[H]$, since $\widetilde{M}$ is a truncated version of M. Therefore, $V^\star_h(s)\leq B_\star$ holds, and the expected hitting time (the number of steps needed to reach the goal) of the optimal policy in $\widetilde{M}$ starting from any (s,h) is upper bounded by $\frac{B_\star}{c_{\min}}$. By [Rosenberg and Mansour, 2020, Lemma 6], when $h\geq\frac{4B_\star}{c_{\min}}\ln\frac{2}{\beta}$, the probability of not reaching g in h steps is at most β. Denote by $\widetilde{\pi}^\star_L$ the optimal policy of $\widetilde{M}$, and $\pi^\star_L$ a non-stationary policy in M which follows $\widetilde{\pi}^\star_L$ for the first H steps, and then follows $\pi^\star$ afterwards. We have for any $s\in\mathcal{S}$, $V^\star(s)-V^\star_{H-1}(s)\leq V^{\pi^\star_L}(s)-V^{\widetilde{\pi}^\star_L}_{H-1}(s)\leq B_\star\beta$, where we apply $H\geq\frac{4B_\star}{c_{\min}}\ln\frac{2}{\beta}+1$, $V^\star(s)\leq V^{\widetilde{\pi}^\star_L}(s)$ and $V^\star_{H-1}(s)=V^{\widetilde{\pi}^\star_L}_{H-1}(s)$. Finally, $Q^\star(s,a)-Q^\star_H(s,a)=P_{s,a}(V^\star-V^\star_{H-1})\leq B_\star\beta$. ☐

Lemma 4. *With probability at least* $1-2\delta$, $\sum_{t=1}^T c_t-c(s_t,a_t)=\tilde{\mathcal{O}}\left(\sqrt{C_K}\right)$.

Proof. By Eq. (24) of Lemma 35, $\|c\|_\infty\in[0,1]$, and Lemma 36 with $\alpha=1$, with probability at least $1-2\delta$:

$$\sum_{t=1}^T c_t-c(s_t,a_t)=\tilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^T\mathbb{E}[c_t^2]}\right)=\tilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^T c(s_t,a_t)}\right)=\tilde{\mathcal{O}}\left(\sqrt{C_K}\right).$$

☐

The next lemma is used in the proof of Theorem 1, which shows that the sum of the variances of the optimal value function is of order $\tilde{\mathcal{O}}(B_\star C_K)$. It is also useful in bounding the overhead of Bernstein-style confidence interval (see Lemma 11 and [Cohen et al., 2020, Lemma 4.7] for example).

Lemma 5. *With probability at least* $1-2\delta$, $\sum_{t=1}^T\mathbb{V}(P_{s_t,a_t},V^\star)=\tilde{\mathcal{O}}\left(B_\star^2+B_\star C_K\right)$.

Proof. Note that:

$$\begin{aligned}
\sum_{t=1}^T\mathbb{V}(P_{s_t,a_t},V^\star)&=\sum_{t=1}^T P_{s_t,a_t}(V^\star)^2-(P_{s_t,a_t}V^\star)^2\\
&=\sum_{k=1}^K\sum_{i=1}^{I_k}P_{s_i^k,a_i^k}(V^\star)^2-V^\star(s_i^k)^2+\sum_{k=1}^K\sum_{i=1}^{I_k}V^\star(s_i^k)^2-(P_{s_i^k,a_i^k}V^\star)^2\\
&\leq\sum_{k=1}^K\sum_{i=1}^{I_k}P_{s_i^k,a_i^k}(V^\star)^2-V^\star(s_{i+1}^k)^2+\sum_{k=1}^K\sum_{i=1}^{I_k}Q^\star(s_i^k,a_i^k)^2-(P_{s_i^k,a_i^k}V^\star)^2.\\
&\qquad\qquad(V^\star(s^k_{I_k+1})=0\text{ and }V^\star(s_i^k)\leq Q^\star(s_i^k,a_i^k))
\end{aligned}$$

For the first term, by Eq. (24) of Lemma 35 with $V^\star(s)\leq B_\star$ and Lemma 30 with $X=V^\star(S')$, $S'\sim P_{s_t,a_t}$, we have with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=1}^K\sum_{i=1}^{I_k}P_{s_i^k,a_i^k}(V^\star)^2-V^\star(s_{i+1}^k)^2&=\tilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^T\mathbb{V}(P_{s_t,a_t},(V^\star)^2)+B_\star^2}\right)\\
&=\tilde{\mathcal{O}}\left(B_\star\sqrt{\sum_{t=1}^T\mathbb{V}(P_{s_t,a_t},V^\star)+B_\star^2}\right).
\end{aligned}$$

For the second term, note that:

$$\sum_{k=1}^{K}\sum_{i=1}^{I_k} Q^\star(s_i^k,a_i^k)^2 - (P_{s_i^k,a_i^k}V^\star)^2 = \sum_{k=1}^{K}\sum_{i=1}^{I_k}\left(Q^\star(s_i^k,a_i^k) - P_{s_i^k,a_i^k}V^\star\right)\left(Q^\star(s_i^k,a_i^k) + P_{s_i^k,a_i^k}V^\star\right)$$

$$\leq \sum_{k=1}^{K}\sum_{i=1}^{I_k} 3B_\star c(s_i^k,a_i^k). \qquad (Q^\star(s,a)\leq 2B_\star \text{ and } V^\star(s)\leq B_\star \text{ for any } (s,a)\in\mathcal{S}\times\mathcal{A})$$

Therefore, $\sum_{t=1}^{T}\mathbb{V}(P_{s_t,a_t},V^\star) = \tilde{\mathcal{O}}\left(B_\star\sqrt{\sum_{t=1}^{T}\mathbb{V}(P_{s_t,a_t},V^\star)} + B_\star^2 + B_\star\sum_{k=1}^{K}\sum_{i=1}^{I_k}c(s_i^k,a_i^k)\right)$. By Lemma 25 with $x=\sum_{t=1}^{T}\mathbb{V}(P_{s_t,a_t},V^\star)$ and Lemma 36, we have with probability at least $1-\delta$,

$$\sum_{t=1}^{T}\mathbb{V}(P_{s_t,a_t},V^\star) = \tilde{\mathcal{O}}\left(B_\star^2 + B_\star\sum_{k=1}^{K}\sum_{i=1}^{I_k}c(s_i^k,a_i^k)\right) = \tilde{\mathcal{O}}\left(B_\star^2 + B_\star C_K\right).$$

□

D Omitted Details for Section 4

Before we present the proof of Theorem 3 (Section D.3), we first quantify the sample complexity of the reference value function (Section D.1) and prove the two required properties (Section D.2).

Extra Notations Denote by $Q_t(s,a)$, $V_t(s)$, $V_t^{\text{ref}}(s)$, B_t, $N_t(s,a)$ the value of $Q(s,a)$, $V(s)$, $V^{\text{ref}}(s)$, B, $N(s,a)$ at the beginning of time step t. Define $N_t(s)=\sum_a N_t(s,a)$. Denote by $n_t(s,a), m_t(s,a), b_t(s,a), b'_t(s,a), \iota_t(s,a), \widehat{c}_t(s,a)$ the value of $n, m, b, b', \iota, \widehat{c}$ used in computing $Q_t(s,a)$. Note that, these are *not* necessarily their values at time step t. For example, $n_t(s,a)$ is the number of visits to (s,a) before the current stage (not before time t); $m_t(s,a)$ the number of visits to (s,a) in the last stage; $b_t(s,a)$ and $b'_t(s,a)$ are the bonuses used in the last update of $Q_t(s,a)$; and $\widehat{c}_t(s,a)$ is the cost estimator used in the last update of $Q_t(s,a)$ ($b_t(s,a)$, $b'_t(s,a)$ and $\widehat{c}_t(s,a)$ are 0 when $n_t(s,a)=0$). Denote by $l_{t,i}(s,a)$ the i-th time step the agent visits (s,a) among those $n_t(s,a)$ steps before the current stage, and by $\check{l}_{t,i}(s,a)$ the i-th time step the agent visits (s,a) among those $m_t(s,a)$ steps within the last stage. With these notations, we have by the update rule of the algorithm:

$$\begin{aligned} Q_t(s,a) = \max\Bigg\{ & Q_{t-1}(s,a),\;\; \widehat{c}_t(s,a) + \frac{1}{m_t}\sum_{i=1}^{m_t} V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - b'_t, \\ & \widehat{c}_t(s,a) + \frac{1}{n_t}\sum_{i=1}^{n_t} V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}}) + \frac{1}{m_t}\sum_{i=1}^{m_t}(V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - V^{\text{ref}}_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}})) - b_t\Bigg\}, \end{aligned} \tag{3}$$

where m_t represents $m_t(s,a)$, $\check{l}_{t,i}$ represents $\check{l}_{t,i}(s,a)$, and similarly for n_t, $l_{t,i}$, b_t and b'_t.

We also define two empirical variances at time step t as:

$$\nu_t = \frac{1}{m_t}\sum_{i=1}^{m_t}(V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - V^{\text{ref}}_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}))^2 - \left(\frac{1}{m_t}\sum_{i=1}^{m_t}V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - V^{\text{ref}}_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}})\right)^2 \tag{4}$$

and

$$\nu_t^{\text{ref}} = \frac{1}{n_t}\sum_{i=1}^{n_t}V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}})^2 - \left(\frac{1}{n_t}\sum_{i=1}^{n_t}V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}})\right)^2. \tag{5}$$

Here, ν_t and ν_t^{ref} should be treated as a function of state-action pair (s,a), so that m_t, n_t, $\check{l}_{t,i}$, and $l_{t,i}$ in the formulas all represent $m_t(s,a)$, $n_t(s,a)$, $\check{l}_{t,i}(s,a)$, and $l_{t,i}(s,a)$. Except for Lemma 9, this input (s,a) is simply (s_t,a_t).

Further define $\varepsilon_t = \mathbb{I}\{n_t > 0\} = \mathbb{I}\{m_t > 0\}$, and $0/0$ to be 0 so that formula in the form $\frac{1}{n_t}\sum_{i=1}^{n_t}X_{l_{t,i}}$ is treated as 0 if $n_t=0$ (similarly for m_t).

D.1 Sample Complexity for Reference Value Function

In this section, we assume $H = \lceil \frac{4B_\star}{c_{\min}} \ln(\frac{2}{\beta}) + 1 \rceil_2$ for some $\beta > 0$ (the form used in Theorem 2). We show that to obtain a reference value with precision $\rho \geq 2B_\star\beta$ at state s (that is, $|V^{\text{ref}}(s) - V^\star(s)| \leq \rho$), $\tilde{\mathcal{O}}\left(\frac{B_\star^2 H^3 SA}{\rho^2}\right)$ number of visits to state s is sufficient (Corollary 6). Moreover, the total costs appeared in regret for a reference value function with maximum precision ρ is $\tilde{\mathcal{O}}\left(\frac{B_\star^2 H^3 S^2 A}{\rho}\right)$ (Lemma 8). Note that if we only update the reference value function once as in Zhang et al. [2020b], instead of applying our "smoother" update, the total costs become $\tilde{\mathcal{O}}\left(\frac{B_\star^2 H^3 S^2 A}{\rho^2}\right)$.

Lemma 6. *With probability at least $1 - 8\delta$, Algorithm 2 ensures for any non-negative weights $\{w_t\}_{t=1}^T$,*

$$\sum_{t=1}^T w_t(Q^\star(s_t, a_t) - Q_t(s_t, a_t)) \leq B_\star \left\|w\right\|_1 \beta + \tilde{\mathcal{O}}\left(H^2 SAB_\star \left\|w\right\|_\infty + B_\star\sqrt{H^3 SA \left\|w\right\|_\infty \left\|w\right\|_1}\right).$$

Proof. Define $w_t^{(0)} = w_t$ and $w_{t+1}^{(h+1)} = \sum_{t'=1}^T \sum_{i=1}^{m_{t'}} \frac{w_{t'}^{(h)}}{m_{t'}} \mathbb{I}\{t = \check{l}_{t',i}\}$. We first argue the following properties related to $w_t^{(h)}$ and vector $w^{(h)} = (w_1^{(h)}, \ldots, w_T^{(h)})$. Denote by j_t the stage to which time step t belongs. When $t = \check{l}_{t',i}$, we have $m_{t'} = e_{j_t}$. Therefore,

$$\sum_{t'=1}^T \sum_{i=1}^{m_{t'}} \frac{1}{m_{t'}} \mathbb{I}\{t = \check{l}_{t',i}\} \leq \frac{e_{j_t+1}}{e_{j_t}} \leq 1 + \frac{1}{H},$$

and thus, $\left\|w^{(h)}\right\|_\infty \leq (1 + \frac{1}{H}) \left\|w^{(h-1)}\right\|_\infty \leq \cdots \leq (1 + \frac{1}{H})^h \left\|w\right\|_\infty$. Moreover,

$$\left\|w^{(h+1)}\right\|_1 = \sum_{t=1}^T \sum_{t'=1}^T \sum_{i=1}^{m_{t'}} \frac{w_{t'}^{(h)}}{m_{t'}} \mathbb{I}\{t = \check{l}_{t',i}\} = \sum_{t'=1}^T w_{t'}^{(h)} \sum_{i=1}^{m_{t'}} \sum_{t=1}^T \frac{\mathbb{I}\{t = \check{l}_{t',i}\}}{m_{t'}} \leq \left\|w^{(h)}\right\|_1,$$

and thus $\left\|w^{(h)}\right\|_1 \leq \left\|w\right\|_1$ for any h. Also note that for any $\{X_t\}_t$ such that $X_t \geq 0$:

$$\sum_{t=1}^T \frac{w_t^{(h)}}{m_t} \sum_{i=1}^{m_t} X_{\check{l}_{t,i}} = \sum_{t'=1}^T \sum_{t=1}^T \frac{w_t^{(h)}}{m_t} \sum_{i=1}^{m_t} X_{t'} \mathbb{I}\{t' = \check{l}_{t,i}\} = \sum_{t'=1}^T w_{t'+1}^{(h+1)} X_{t'}. \tag{6}$$

Next, for a fixed (s, a), by Lemma 34, with probability at least $1 - \frac{\delta}{SA}$, when $n_t(s, a) > 0$:

$$|c(s,a) - \widehat{c}_t(s,a)| \leq 2\sqrt{\frac{2\widehat{c}_t(s,a)}{n_t(s,a)} \ln \frac{2SAn_t(s,a)}{\delta}} + \frac{19 \ln \frac{2SAn_t(s,a)}{\delta}}{n_t(s,a)} \leq \sqrt{\frac{\widehat{c}_t(s,a)\iota_t}{n_t(s,a)}} + \frac{\iota_t}{n_t(s,a)}. \tag{7}$$

Taking a union bound, we have Eq. (7) holds for all (s, a) when $n_t(s, a) > 0$ with probability at least $1 - \delta$. Then by definition of b'_t, we have

$$c(s_t, a_t) - \widehat{c}_t(s_t, a_t) \leq \mathbb{I}\{m_t = 0\} + b'_t. \tag{8}$$

Now we are ready to prove the lemma. First, we condition on Lemma 9, which happens with probability at least $1 - 7\delta$. Then for any $h \in \{0, \ldots, H-1\}$, $\mathring{Q} = Q_{H-h}$, $\mathring{V} = Q_{H-h-1}$ we have:

$$\begin{aligned}
&\sum_{t=1}^T w_t^{(h)} (\mathring{Q}(s_t, a_t) - Q_t(s_t, a_t))_+ \\
&\leq \sum_{t=1}^T w_t^{(h)} (c(s_t, a_t) - \widehat{c}_t(s_t, a_t))_+ + w_t^{(h)} \left(P_t \mathring{V} - \frac{1}{m_t} \sum_{i=1}^{m_t} V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}})\right)_+ + w_t^{(h)} b'_t \\
&\qquad\qquad \text{(by Eq. (3) and } \mathring{Q}(s,a) = c(s,a) + P_{s,a}\mathring{V}) \\
&\leq \sum_{t=1}^T 2B_\star w_t^{(h)} \mathbb{I}\{m_t = 0\} + \sum_{t=1}^T w_t^{(h)} \left(\frac{1}{m_t} \sum_{i=1}^{m_t} P_{\check{l}_{t,i}} \mathring{V} - \frac{1}{m_t} \sum_{i=1}^{m_t} V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}})\right)_+ + 2w_t^{(h)} b'_t. \\
&\qquad\qquad \text{(Eq. (8), } P_t = P_{\check{l}_{t,i}} \text{ and } P_t\mathring{V} \leq B_\star \mathbb{I}\{m_t = 0\} + \tfrac{1}{m_t} \textstyle\sum_{i=1}^{m_t} P_{\check{l}_{t,i}} \mathring{V})
\end{aligned}$$

Since $e_1 = H$, we have $\sum_{t=1}^{T} w_t^{(h)} \mathbb{I}\{m_t = 0\} \leq SAH \left\| w^{(h)} \right\|_\infty$. Moreover, by Eq. (24) of Lemma 35 with $X_t = \mathring{V}(s'_t)$, we have with probability at least $1 - \frac{\delta}{H}$: $\frac{1}{m_t} \sum_{i=1}^{m_t} P_{\check{l}_{t,i}} \mathring{V} \leq \frac{1}{m_t} \sum_{i=1}^{m_t} \mathring{V}(s'_{\check{l}_{t,i}}) + \tilde{\mathcal{O}}\left(\frac{B_\star \varepsilon_t}{\sqrt{m_t}} \right)$. Plugging these back to the previous inequality and using the definition of b'_t gives:

$$
\begin{aligned}
&\sum_{t=1}^{T} w_t^{(h)} (\mathring{Q}(s_t, a_t) - Q_t(s_t, a_t))_+ \\
&\leq 2HSAB_\star \left\| w^{(h)} \right\|_\infty + \sum_{t=1}^{T} \frac{w_t^{(h)}}{m_t} \sum_{i=1}^{m_t} \left(\mathring{V}(s'_{\check{l}_{t,i}}) - V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) \right)_+ + \tilde{\mathcal{O}}\left(\frac{B_\star w_t^{(h)} \varepsilon_t}{\sqrt{m_t}} + \frac{w_t^{(h)} \varepsilon_t}{n_t} \right) \\
&\leq 3HSAB_\star \left\| w^{(h)} \right\|_\infty + \tilde{\mathcal{O}}\left(B_\star \sqrt{HSA \left\| w^{(h)} \right\|_\infty \left\| w \right\|_1} \right) + \sum_{t=1}^{T} w_{t+1}^{(h+1)} \left(\mathring{V}(s'_t) - V_t(s'_t) \right)_+ \\
&\hfill \text{(Eq. (6) and Lemma 14)} \\
&\leq \tilde{\mathcal{O}}\left(HSAB_\star \left\| w^{(h)} \right\|_\infty + B_\star \sqrt{HSA \left\| w^{(h)} \right\|_\infty \left\| w \right\|_1} \right) + \sum_{t=1}^{T} w_t^{(h+1)} (\mathring{Q}(s_t, a_t) - Q_t(s_t, a_t))_+,
\end{aligned}
$$

where in the last inequality we apply:

$$
\begin{aligned}
\sum_{t=1}^{T} w_{t+1}^{(h+1)} \left(\mathring{V}(s'_t) - V_t(s'_t) \right)_+ &\leq \sum_{t=1}^{T} w_{t+1}^{(h+1)} (\mathring{V}(s'_t) - V_{t+1}(s'_t))_+ + \tilde{\mathcal{O}}\left(\left\| w^{(h)} \right\|_\infty SB_\star \right) \\
&\hfill \text{(apply Lemma 28 on } \textstyle\sum_{t=1}^{T} V_{t+1}(s'_t) - V_t(s'_t)\text{)} \\
&\leq \sum_{t=1}^{T} w_t^{(h+1)} (\mathring{V}(s_t) - V_t(s_t))_+ + \tilde{\mathcal{O}}\left(\left\| w^{(h)} \right\|_\infty SB_\star \right) \\
&\hfill \text{(}(\mathring{V}(s'_t) - V_{t+1}(s'_t))_+ \leq (\mathring{V}(s_{t+1}) - V_{t+1}(s_{t+1}))_+ \text{ and } w_{T+1}^{(h+1)} = 0\text{)} \\
&\leq \sum_{t=1}^{T} w_t^{(h+1)} (\mathring{Q}(s_t, a_t) - Q_t(s_t, a_t))_+ + \tilde{\mathcal{O}}\left(\left\| w^{(h)} \right\|_\infty SB_\star \right). \\
&\hfill \text{(}\mathring{V}(s_t) \leq \mathring{Q}(s_t, a_t) \text{ and } V_t(s_t) = Q_t(s_t, a_t)\text{)}
\end{aligned}
$$

By a union bound, the inequality above holds for $\mathring{Q} = Q_{H-h}, \mathring{V} = Q_{H-h-1}$ for all $h \in \{0, \ldots, H-1\}$ with probability at least $1 - \delta$. Applying the inequality above recursively starting from $h = 0$, and by $Q_0^\star(s, a) - Q_t(s, a) \leq 0$, $(1 + \frac{1}{H})^H \leq 3$:

$$
\sum_{t=1}^{T} w_t (Q_H^\star(s_t, a_t) - Q_t(s_t, a_t))_+ = \tilde{\mathcal{O}}\left(H^2 SAB_\star \left\| w \right\|_\infty + B_\star \sqrt{H^3 SA \left\| w \right\|_\infty \left\| w \right\|_1} \right).
$$

Therefore, by Lemma 1,

$$
\begin{aligned}
\sum_{t=1}^{T} w_t (Q^\star(s_t, a_t) - Q_t(s_t, a_t)) &= \sum_{t=1}^{T} w_t (Q^\star(s_t, a_t) - Q_H^\star(s_t, a_t) + Q_H^\star(s_t, a_t) - Q_t(s_t, a_t)) \\
&\leq B_\star \left\| w \right\|_1 \beta + \tilde{\mathcal{O}}\left(H^2 SAB_\star \left\| w \right\|_\infty + B_\star \sqrt{H^3 SA \left\| w \right\|_\infty \left\| w \right\|_1} \right).
\end{aligned}
$$

□

Now by Lemma 6 with $w_t = \mathbb{I}\{V^\star(s_t) - V_t(s_t) \geq \rho\}$ for some threshold ρ, we can bound the sample complexity of obtaining a value function with precision ρ (Corollary 6), which is used to determine the value of $\theta^\star$ (Lemma 8). However, one caveat here is that the bound in Lemma 6 has logarithmic dependency on T from ι_t, which should not appear in the definition of $\theta^\star$ since T is a random variable. To deal with this, we obtain a loose bound on T in the following lemma.

Lemma 7. *With probability at least* $1 - 13\delta$, $T = \tilde{\mathcal{O}}(B_\star K / c_{\min} + B_\star^2 H^3 SA / c_{\min}^2)$.

Proof. By Lemma 6 with $w_t = 1$, we have with probability at least $1 - 8\delta$:

$$\sum_{t=1}^{T} Q^\star(s_t, a_t) - Q_t(s_t, a_t) = B_\star T\beta + \tilde{\mathcal{O}}\left(H^2SAB_\star + B_\star\sqrt{H^3SAT}\right).$$

Now by Eq. (2), Lemma 4, Lemma 35, and Lemma 5, with probability at least $1 - 5\delta$,

$$\begin{aligned} R_K &\leq \sum_{t=1}^{T}(c_t - c(s_t, a_t)) + \sum_{t=1}^{T}(V^\star(s'_t) - P_{s_t,a_t}V^\star) + \sum_{t=1}^{T}(Q^\star(s_t, a_t) - V^\star(s_t)) \\ &\leq \tilde{\mathcal{O}}(\sqrt{B_\star C_K} + B_\star) + \sum_{t=1}^{T}(Q^\star(s_t, a_t) - Q_t(s_t, a_t)) \qquad (V_t = Q_t(s_t, a_t) \text{ and Lemma 9}) \\ &= B_\star T\beta + \tilde{\mathcal{O}}\left(H^2SAB_\star + B_\star\sqrt{H^3SAT}\right). \qquad (C_K \leq T) \end{aligned}$$

Further using $c_{\min}T - KB_\star \leq R_K$, $B_\star\beta \leq \frac{c_{\min}}{2}$, and Lemma 25 proves the statement. ☐

Corollary 6. *With probability at least $1 - 13\delta$, Algorithm 2 ensures for any $\rho \geq 2B_\star\beta$:*

$$\sum_{t=1}^{T} \mathbb{I}\left\{V^\star(s_t) - V_t(s_t) \geq \rho\right\} = \tilde{\mathcal{O}}\left(\frac{B_\star^2H^3SA}{\rho^2}\right) \triangleq U_\rho - 1,$$

and for any $s \in \mathcal{S}$, $N_t(s) \geq U_\rho$ implies $0 \leq V^\star(s) - V_t(s) \leq \rho$.

Proof. We can assume $\rho \leq B_\star$ since $\sum_{t=1}^{T}\mathbb{I}\{V^\star(s_t) - V_t(s_t) \geq \rho\} = 0$ when $\rho > B_\star$. By Lemma 6 with $w_t = \mathbb{I}\{V^\star(s_t) - V_t(s_t) \geq \rho\}$, $\rho w_t \leq w_t(V^\star(s_t) - V_t(s_t))$, $\rho \geq 2B_\star\beta$, and $V^\star(s_t) - V_t(s_t) \leq Q^\star(s_t, a_t) - Q_t(s_t, a_t)$, we have with probability at least $1 - 8\delta$:

$$\rho\left\|w\right\|_1 \leq \sum_{t=1}^{T} w_t(V^\star(s_t) - V_t(s_t)) \leq \frac{\rho}{2}\left\|w\right\|_1 + \tilde{\mathcal{O}}\left(H^2SAB_\star + B_\star\sqrt{H^3SA\left\|w\right\|_1}\right).$$

Therefore, by Lemma 25 and Lemma 7, $\left\|w\right\|_1 = \tilde{\mathcal{O}}\left(\frac{H^2SAB_\star}{\rho} + \frac{B_\star^2H^3SA}{\rho^2}\right)$, which has no logarithmic dependency on T. We prove the second statement by contradiction: suppose $N_t(s) \geq U_\rho$ and $V^\star(s) - V_t(s) > \rho$. Then since V_t is non-decreasing in t, $N_t(s) \leq \left\|w\right\|_1$. Thus, $U_\rho \leq N_t(s) \leq \left\|w\right\|_1 < U_\rho$, a contradiction. ☐

Lemma 8. *Define $\beta_i = \frac{B_\star}{2^i}$, $\widetilde{N}_0 = 0$, $\widetilde{N}_i = U_{\beta_i}$ (defined in Corollary 6) for $i \geq 1$ and $q^\star = \inf\{i : \beta_i \leq c_{\min}\}$. Define $V^{\text{REF}} = V^{\text{ref}}_{T+1}$, $\theta^\star = \lceil\widetilde{N}_{q^\star}\rceil_2$, and B^{ref}_t such that:*

$$B^{\text{ref}}_t(s) = \sum_{i=1}^{q^\star}\beta_{i-1}\mathbb{I}\{\lceil\widetilde{N}_{i-1}\rceil_2 \leq N_t(s) < \lceil\widetilde{N}_i\rceil_2\}.$$

Then with probability at least $1 - 13\delta$, $V^{\text{REF}}(s) - V^{\text{ref}}_t(s) \leq B^{\text{ref}}_t(s)$, and

$$\sum_{t=1}^{T} V^{\text{REF}}(s_t) - V^{\text{ref}}_t(s_t) \leq \sum_{t=1}^{T} B^{\text{ref}}_t(s_t) = \tilde{\mathcal{O}}\left(\frac{B_\star^2H^3S^2A}{c_{\min}}\right) \triangleq C_{\text{REF}},$$

$$\sum_{t=1}^{T}\left(V^{\text{REF}}(s_t) - V^{\text{ref}}_t(s_t)\right)^2 \leq \sum_{t=1}^{T} B^{\text{ref}}_t(s_t)^2 = \tilde{\mathcal{O}}\left(B_\star^2H^3S^2A\right) \triangleq C_{\text{REF, 2}}.$$

Proof. We condition on Corollary 6, which happens with probability at least $1 - 13\delta$. By Corollary 6 with $\rho = \beta_i$ for each $i \in [q^\star]$, we have $V^{\text{REF}}(s) - V^{\text{ref}}_t(s) \leq B^{\text{ref}}_t(s)$. Moreover, $B^{\text{ref}}_t(s)^2 = \sum_{i=1}^{q^\star}\beta_{i-1}^2\mathbb{I}\{\lceil\widetilde{N}_{i-1}\rceil_2 \leq N_t(s) < \lceil\widetilde{N}_i\rceil_2\}$. Thus,

$$\sum_{t=1}^{T} B^{\text{ref}}_t(s_t) \leq \sum_s\sum_{i=1}^{q^\star}\beta_{i-1}\lceil\widetilde{N}_i\rceil_2 = \tilde{\mathcal{O}}\left(\sum_s\sum_{i=1}^{q^\star}\frac{B_\star^2H^3SA}{\beta_i}\right) = \tilde{\mathcal{O}}\left(\frac{B_\star^2H^3S^2A}{\beta_{q^\star}}\right).$$

$$\sum_{t=1}^{T} B^{\text{ref}}_t(s_t)^2 \leq \sum_s\sum_{i=1}^{q^\star}\beta_{i-1}^2\lceil\widetilde{N}_i\rceil_2 = \tilde{\mathcal{O}}\left(\sum_s\sum_{i=1}^{q^\star}B_\star^2H^3SA\right) = \tilde{\mathcal{O}}\left(B_\star^2H^3S^2A\right).$$

□

D.2 Proofs of Required Properties

In this section, we prove Property 1 and Property 2 of Algorithm 2.

Lemma 9. *With probability at least* $1-7\delta$*, Algorithm 2 ensures* $Q_t(s,a) \le Q_{t+1}(s,a) \le Q^\star(s,a)$ *for any* $(s,a) \in \mathcal{S}\times\mathcal{A}, t \ge 1$.

Proof. We fix a pair (s,a), and denote $n_t, m_t, l_{t,i}, \check{l}_{t,i}, b_t, b'_t, \iota_t$ as shorthands of the corresponding functions evaluated at (s,a). The first inequality is by the update rule of Q_t. Next, we prove $Q_t(s,a) \le Q^\star(s,a)$ by induction on t. It is clearly true when $t=1$. For the induction step, the statement is clearly true when $n_t = m_t = 0$. When $n_t > 0$, it suffices to consider two update rules, that is, the last two terms in the max operator of Eq. (3). For the second update rule, note that,

$$\begin{aligned}
&\widehat{c}_t(s,a) + \frac{1}{n_t}\sum_{i=1}^{n_t} V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}}) + \frac{1}{m_t}\sum_{i=1}^{m_t}\left(V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - V^{\text{ref}}_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}})\right) - b_t\\
&= \widehat{c}_t(s,a) + \frac{1}{n_t}\sum_{i=1}^{n_t} P_{s,a}V^{\text{ref}}_{l_{t,i}} + \frac{1}{m_t}\sum_{i=1}^{m_t} P_{s,a}\left(V_{\check{l}_{t,i}} - V^{\text{ref}}_{\check{l}_{t,i}}\right)\\
&\qquad + \underbrace{\frac{1}{n_t}\sum_{i=1}^{n_t}\left(\mathbb{I}_{s'_{l_{t,i}}} - P_{s,a}\right)V^{\text{ref}}_{l_{t,i}}}_{\chi_1} + \underbrace{\frac{1}{m_t}\sum_{i=1}^{m_t}\left(\mathbb{I}_{s'_{\check{l}_{t,i}}} - P_{s,a}\right)\left(V_{\check{l}_{t,i}} - V^{\text{ref}}_{\check{l}_{t,i}}\right)}_{\chi_2} - b_t. \qquad (9)
\end{aligned}$$

Define $C'_t = \lceil \ln(B_\star^4 n_t)\rceil^2 \le \min\{4\ln^2(B_\star^4 n_t), B_\star^8 n_t^2\}$ (in general, we can set $C'_t = \lceil \ln(\widetilde{B}^4 n_t)\rceil^2$ for some $\widetilde{B} \ge B_\star$). For χ_1, by Eq. (24) of Lemma 35 with $b = B_\star^2$ and $C \le C'_t$, we have with probability at least $1 - \frac{\delta}{SA}$:

$$|\chi_1| = \left|\frac{1}{n_t}\sum_{i=1}^{n_t}\left(\mathbb{I}_{s'_{l_{t,i}}} - P_{s,a}\right)V^{\text{ref}}_{l_{t,i}}\right| \le 4\ln^3\left(\frac{4SAB_\star^8 n_t^5}{\delta}\right)\left(\sqrt{\frac{8\sum_{i=1}^{n_t}\mathbb{V}(P_{s,a}, V^{\text{ref}}_{l_{t,i}})}{n_t^2}} + \frac{5B_t}{n_t}\right),$$

Note that (recall that ν_t^{ref} represents $\nu_t^{\text{ref}}(s,a)$)

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\mathbb{V}(P_{s,a}, V^{\text{ref}}_{l_{t,i}}) - \nu_t^{\text{ref}} = \chi_3 + \chi_4 + \chi_5, \qquad (10)$$

where

$$\chi_3 = \frac{1}{n_t}\sum_{i=1}^{n_t}\left(P_{s,a}(V^{\text{ref}}_{l_{t,i}})^2 - V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}})^2\right), \quad \chi_4 = \left(\frac{1}{n_t}\sum_{i=1}^{n_t}V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}})\right)^2 - \left(\frac{1}{n_t}\sum_{i=1}^{n_t}P_{s,a}V^{\text{ref}}_{l_{t,i}}\right)^2,$$

$$\chi_5 = \left(\frac{1}{n_t}\sum_{i=1}^{n_t}P_{s,a}V^{\text{ref}}_{l_{t,i}}\right)^2 - \frac{1}{n_t}\sum_{i=1}^{n_t}(P_{s,a}V^{\text{ref}}_{l_{t,i}})^2.$$

By Eq. (24) of Lemma 35 with $b = B_\star^2$ and $C \le C'_t$, and Lemma 30 with $\left\|V^{\text{ref}}_{l_{t,i}}\right\|_\infty \le B_t$, with probability at least $1 - \frac{2\delta}{SA}$,

$$\begin{aligned}
|\chi_3| &\le \frac{4\ln^3(4SAB_\star^8 n_t^5/\delta)}{n_t}\left(\sqrt{8\sum_{i=1}^{n_t}\mathbb{V}(P_{s,a},(V^{\text{ref}}_{l_{t,i}})^2)} + 5B_t^2\right)\\
&\le \frac{4\ln^3(4SAB_\star^8 n_t^5/\delta)}{n_t}\left(2B_t\sqrt{8\sum_{i=1}^{n_t}\mathbb{V}(P_{s,a},V^{\text{ref}}_{l_{t,i}})} + 5B_t^2\right). \qquad (11)\\
|\chi_4| &\le \left|\frac{1}{n_t}\sum_{i=1}^{n_t}V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}}) + \frac{1}{n_t}\sum_{i=1}^{n_t}P_{s,a}V^{\text{ref}}_{l_{t,i}}\right|\left|\frac{1}{n_t}\sum_{i=1}^{n_t}V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}}) - \frac{1}{n_t}\sum_{i=1}^{n_t}P_{s,a}V^{\text{ref}}_{l_{t,i}}\right|\\
&\le 2B_t\cdot\frac{4\ln^3(4SAB_\star^8 n_t^5/\delta)}{n_t}\left(\sqrt{8\sum_{i=1}^{n_t}\mathbb{V}(P_{s,a},V^{\text{ref}}_{l_{t,i}})} + 5B_t\right). \qquad (12)
\end{aligned}$$

Moreover, $\chi_5 \le 0$ by Cauchy-Schwarz inequality. Therefore,

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\mathbb{V}(P_{s,a},V^{\text{ref}}_{l_{t,i}}) - \nu^{\text{ref}}_t \le \frac{4B_t\ln^3(4SAB_\star^8 n_t^5/\delta)}{n_t}\left(4\sqrt{8\sum_{i=1}^{n_t}\mathbb{V}(P_{s,a},V^{\text{ref}}_{l_{t,i}})} + 15B_t\right).$$

Applying Lemma 25 with $x = \sum_{i=1}^{n_t}\mathbb{V}(P_{s,a},V^{\text{ref}}_{l_{t,i}})$, we obtain:

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\mathbb{V}(P_{s,a},V^{\text{ref}}_{l_{t,i}}) \le 2\nu^{\text{ref}}_t + \frac{4216B_t^2\ln^6\frac{4SAB_\star^8 n_t^5}{\delta}}{n_t}.$$

Thus, $\left|\frac{1}{n_t}\sum_{i=1}^{n_t}\left(\mathbb{I}_{s'_{l_{t,i}}} - P_{s,a}\right)V^{\text{ref}}_{l_{t,i}}\right| \le \sqrt{\frac{\nu^{\text{ref}}_t}{n_t}\iota_t} + \frac{3B_t\iota_t}{n_t}$. By similar arguments, $|\chi_2| \le \sqrt{\frac{\nu_t}{m_t}\iota_t} + \frac{3B_t\iota_t}{m_t}$ with probability at least $1 - \frac{3\delta}{SA}$. Finally, by Eq. (7) and $B_t \ge 1$, we have $\widehat{c}_t(s,a) - c(s,a) \le \sqrt{\frac{\widehat{c}_t(s,a)\iota}{n_t}} + \frac{B_t\iota}{n_t}$. Therefore,

$$|\widehat{c}_t(s,a) - c(s,a)| + |\chi_1| + |\chi_2| \le b_t. \qquad (13)$$

Plugging Eq. (13) back to Eq. (9), and by the non-decreasing property of V^{ref}_t and $V_{\check{l}_{t,i}}(s) \le V^\star(s)$ for any $s \in \mathcal{S}^+$:

$$\begin{aligned}
&\widehat{c}_t(s,a) + \frac{1}{n_t}\sum_{i=1}^{n_t}V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}}) + \frac{1}{m_t}\sum_{i=1}^{m_t}\left(V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - V^{\text{ref}}_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}})\right) - b_t\\
&\le c(s,a) + \frac{1}{n_t}\sum_{i=1}^{n_t}P_{s,a}V^{\text{ref}}_{l_{t,i}} + \frac{1}{m_t}\sum_{i=1}^{m_t}P_{s,a}\left(V_{\check{l}_{t,i}} - V^{\text{ref}}_{\check{l}_{t,i}}\right) \le c(s,a) + P_{s,a}V^\star = Q^\star(s,a).
\end{aligned}$$

For the first update rule, by Eq. (24) of Lemma 35 with $b = K$ and $C \le C'_t$, with probability at least $1 - \frac{\delta}{SA}$, $\frac{1}{m_t}\sum_{i=1}^{m_t}V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - P_{\check{l}_{t,i}}V_{\check{l}_{t,i}} \le 2\sqrt{\frac{B_t^2\iota_t}{m_t}}$. Therefore, by Eq. (7):

$$\widehat{c}_t(s,a) + \frac{1}{m_t}\sum_{i=1}^{m_t}V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - b'_t \le c(s,a) + \frac{1}{m_t}\sum_{i=1}^{m_t}P_{\check{l}_{t,i}}V_{\check{l}_{t,i}} \le c(s,a) + P_{s,a}V^\star = Q^\star(s,a).$$

Combining two cases, we have $Q_t(s,a) \le Q^\star(s,a)$ for the fixed (s,a). By a union bound over $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have $Q_t(s,a) \le Q^\star(s,a)$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}, t \ge 1$. $\square$

Remark 1. *Note that the statement of Lemma 9 still holds if we use "compute* $\iota \leftarrow 256\ln^6(4SA\widetilde{B}^8 n^5/\delta)$*" in Line 5 of Algorithm 2 for some* $\widetilde{B} \ge B_\star$*. This is useful in deriving the parameter-free version of Algorithm 2 in Section D.5; see Line 1 of Algorithm 4.*

Proof of Theorem 2. Property 1 is satisfied by Lemma 9. For Property 2, we conditioned on Lemma 9, Lemma 8, Lemma 10, and Lemma 11, which holds with probability at least $1 - 50\delta$. Then, for any $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$:

$$
\begin{aligned}
&\sum_{t=1}^{T} (\mathring{Q}(s_t, a_t) - Q_t(s_t, a_t))_+ \\
&\leq \sum_{t=1}^{T} \left(c(s_t, a_t) - \widehat{c}_t(s_t, a_t) + P_t \mathring{V} - \frac{1}{n_t} \sum_{i=1}^{n_t} V^{\text{ref}}_{l_{t,i}}(s'_{l_{t,i}}) - \frac{1}{m_t} \sum_{i=1}^{m_t} \left(V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - V^{\text{ref}}_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) \right) + b_t \right)_+ \\
&\hfill \text{(by Eq. (3) and } \mathring{Q}(s, a) = c(s, a) + P_{s,a}\mathring{V}) \\
&\leq \sum_{t=1}^{T} 2B_\star \mathbb{I}\{m_t = 0\} + \sum_{t=1}^{T} \left(\frac{1}{m_t} \sum_{i=1}^{m_t} P_{\check{l}_{t,i}} \mathring{V} - \frac{1}{n_t} \sum_{i=1}^{n_t} P_{l_{t,i}} V^{\text{ref}}_{l_{t,i}} - \frac{1}{m_t} \sum_{i=1}^{m_t} P_{\check{l}_{t,i}} \left(V_{\check{l}_{t,i}} - V^{\text{ref}}_{\check{l}_{t,i}} \right) \right)_+ + 2b_t \\
&\hfill (P_t \mathring{V} \leq B_\star \mathbb{I}\{m_t = 0\} + \tfrac{1}{m_t} \textstyle\sum_{i=1}^{m_t} P_{\check{l}_{t,i}} \mathring{V} \text{ and Eq. (13))} \\
&\leq 2B_\star HSA + \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} P_{l_{t,i}} \left(V^{\text{REF}} - V^{\text{ref}}_{l_{t,i}} \right) + \frac{1}{m_t} \sum_{i=1}^{m_t} P_{\check{l}_{t,i}} (\mathring{V} - V_{\check{l}_{t,i}})_+ + 2b_t. \\
&(\textstyle\sum_{t=1}^{T} \mathbb{I}\{m_t = 0\} \leq SAH,\ P_t = P_{l_{t,i}} = P_{\check{l}_{t,i}}, \text{ and } V^{\text{ref}}_{\check{l}_{t,i}}(s) \leq V^{\text{REF}}(s) \text{ for any } s \in \mathcal{S} \text{ (Lemma 8))}
\end{aligned}
$$

By Lemma 12 and Lemma 10,

$$
\sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} P_{l_{t,i}} \left(V^{\text{REF}} - V^{\text{ref}}_{l_{t,i}} \right) = \tilde{\mathcal{O}} \left(\sum_{t=1}^{T} P_t (V^{\text{REF}} - V^{\text{ref}}_t) \right) = \tilde{\mathcal{O}} \left(C_{\text{REF}} \right).
$$

Moreover, by Lemma 13, with probability at least $1 - \frac{\delta}{H+1}$,

$$
\frac{1}{m_t} \sum_{i=1}^{m_t} P_{\check{l}_{t,i}} (\mathring{V} - V_{\check{l}_{t,i}})_+ \leq \left(1 + \frac{1}{H} \right)^2 \sum_{t=1}^{T} (\mathring{V}(s_t) - V_t(s_t))_+ + \tilde{\mathcal{O}} \left(B_\star (H + S) \right).
$$

Plugging these back, and by $(1 + \frac{1}{H})^2 \leq 1 + \frac{3}{H}$, Lemma 11 and Lemma 8, we get:

$$
\begin{aligned}
&\sum_{t=1}^{T} (\mathring{Q}(s_t, a_t) - Q_t(s_t, a_t))_+ \leq \tilde{\mathcal{O}} \left(B_\star HSA + C_{\text{REF}} \right) + \left(1 + \frac{1}{H} \right)^2 \sum_{t=1}^{T} (\mathring{V}(s_t) - V_t(s_t))_+ + 2 \sum_{t=1}^{T} b_t \\
&\leq \left(1 + \frac{3}{H} \right) \sum_{t=1}^{T} (\mathring{V}(s_t) - V_t(s_t))_+ + \tilde{\mathcal{O}} \left(\sqrt{B_\star SAC_K} + \sqrt{SAHc_{\min} C_K} + \frac{B_\star^2 H^3 S^2 A}{c_{\min}} \right).
\end{aligned}
$$

Taking a union bound over $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$ and using $H = \tilde{\mathcal{O}} \left(\frac{B_\star}{c_{\min}} \right)$ proves the claim. ☐

D.3 Proof of Theorem 3

Proof. By Theorem 1 and Theorem 2, with probability at least $1 - 60\delta$ and $\beta = \frac{c_{\min}}{2B_\star^2 SAK}$:

$$
C_K - KV^\star(s_{\text{init}}) = R_K \leq \tilde{\mathcal{O}} \left(\beta C_K + \sqrt{B_\star SAC_K} + \frac{B_\star^2 H^3 S^2 A}{c_{\min}} \right).
$$

Then by $V^\star(s_{\text{init}}) \leq B_\star, \beta \leq \frac{1}{2}$ and Lemma 25, we have $C_K = \tilde{\mathcal{O}} \left(B_\star K \right)$. Substituting this back and by $\beta \leq \frac{c_{\min}}{B_\star K}$, $H = \tilde{\mathcal{O}}(B_\star / c_{\min})$, we get $R_K = \tilde{\mathcal{O}} \left(B_\star \sqrt{SAK} + \frac{B_\star^5 S^2 A}{c_{\min}^4} \right)$. ☐

D.4 Extra Lemmas for Section 4

In this section, we gives proofs of auxiliary lemmas used in Section 4. Lemma 10 quantifies the cost of using reference value function. Lemma 11 quantifies the cost of using the variance-aware bonus terms b_t. Lemma 12, Lemma 13, and Lemma 14 deal with the bias induced by the sparse update scheme.

Lemma 10. *With probability at least* $1-9\delta$, $\sum_{t=1}^{T} P_t\left(V^{\text{REF}} - V_t^{\text{ref}}\right) \le \sum_{t=1}^{T} P_t B_t^{\text{ref}} = \tilde{\mathcal{O}}\left(C_{\text{REF}}\right)$, *where* C_{REF} *is defined in Lemma 8.*

Proof. By Lemma 8, Lemma 36, Lemma 28 and $B_{t+1}^{\text{ref}}(s_t') \le B_{t+1}^{\text{ref}}(s_{t+1})$ in each step:

$$\begin{aligned}\sum_{t=1}^{T} P_t\left(V^{\text{REF}} - V_t^{\text{ref}}\right) &\le \sum_{t=1}^{T} P_t B_t^{\text{ref}} \le 2\sum_{t=1}^{T} B_t^{\text{ref}}(s_t') + \tilde{\mathcal{O}}\left(B_\star\right)\\ &= \tilde{\mathcal{O}}\left(\sum_{t=1}^{T} B_t^{\text{ref}}(s_t) + SB_\star\right) = \tilde{\mathcal{O}}\left(C_{\text{REF}}\right).\end{aligned}$$

□

Lemma 11. *With probability at least* $1-21\delta$,

$$\sum_{t=1}^{T} b_t = \tilde{\mathcal{O}}\left(\sqrt{B_\star SAC_K} + B_\star H^2 S^{\frac{3}{2}} A + \sqrt{SAHc_{\min} C_K}\right).$$

Proof. We condition on Lemma 8, which holds with probability at least $1-8\delta$. By Eq. (14) and Eq. (15) of Lemma 14,

$$\begin{aligned}\sum_{t=1}^{T} b_t &\le \sum_{t=1}^{T} \sqrt{\frac{\nu_t^{\text{ref}}\varepsilon_t}{n_t}\iota_t} + \sqrt{\frac{\nu_t\varepsilon_t}{m_t}\iota_t} + B_\star \sum_t \left(\frac{4\varepsilon_t}{n_t} + \frac{3\varepsilon_t}{m_t}\right)\iota_t + \sqrt{\frac{\widehat{c}_t\varepsilon_t\iota_t}{n_t}}\\ &= \tilde{\mathcal{O}}\left(\sum_{t=1}^{T} \sqrt{\frac{\nu_t^{\text{ref}}\varepsilon_t}{n_t}} + \sqrt{\frac{\nu_t\varepsilon_t}{m_t}} + B_\star HSA + \sqrt{\frac{\widehat{c}_t\varepsilon_t}{n_t}}\right).\end{aligned}$$

Note that by Eq. (10), Eq. (11) and Eq. (12), when $n_t > 0$, with probability at least $1-2\delta$,

$$\begin{aligned}&\nu_t^{\text{ref}} - \frac{1}{n_t}\sum_{i=1}^{n_t} \mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}}) \le |\chi_3| + |\chi_4| - \chi_5\\ &\le \tilde{\mathcal{O}}\left(\frac{B_t}{n_t}\sqrt{\sum_{i=1}^{n_t} \mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}})} + \frac{B_t^2}{n_t}\right) + \frac{1}{n_t}\sum_{i=1}^{n_t}(P_{l_{t,i}}V_{l_{t,i}}^{\text{ref}})^2 - \left(\frac{1}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}}V_{l_{t,i}}^{\text{ref}}\right)^2\\ &\overset{\text{(i)}}{=} \tilde{\mathcal{O}}\left(\frac{B_t}{n_t}\sqrt{\sum_{i=1}^{n_t} \mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}})} + \frac{B_t^2}{n_t} + \frac{B_\star}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}}B_{l_{t,i}}^{\text{ref}}\right)\\ &\le \frac{1}{n_t}\sum_{i=1}^{n_t} \mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}}) + \tilde{\mathcal{O}}\left(\frac{B_t^2}{n_t} + \frac{B_\star}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}}B_{l_{t,i}}^{\text{ref}}\right), \qquad \text{(AM-GM Inequality)}\end{aligned}$$

where in (i) we apply:

$$\begin{aligned}&\frac{1}{n_t}\sum_{i=1}^{n_t}(P_{l_{t,i}}V_{l_{t,i}}^{\text{ref}})^2 - \left(\frac{1}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}}V_{l_{t,i}}^{\text{ref}}\right)^2 \le (P_t V^{\text{REF}})^2 - \left(\frac{1}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}}V_{l_{t,i}}^{\text{ref}}\right)^2\\ &\qquad\qquad (V_{l_{t,i}}^{\text{ref}}(s) \le V^{\text{REF}}(s) \text{ for any } s \in \mathcal{S})\\ &\le \frac{2B_\star}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}}\left(V^{\text{REF}} - V_{l_{t,i}}^{\text{ref}}\right) \le \frac{2B_\star}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}}B_{l_{t,i}}^{\text{ref}}. \qquad (\left\|V^{\text{REF}}\right\|_\infty \le B_\star \text{ and Lemma 8})\end{aligned}$$

Therefore, $\nu_t^{\text{ref}} - \frac{2}{n_t}\sum_{i=1}^{n_t} \mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}}) = \tilde{\mathcal{O}}\left(\frac{B_t^2}{n_t} + \frac{B_\star}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}} B_{l_{t,i}}^{\text{ref}}\right)$, and

$$\begin{aligned}
\nu_t^{\text{ref}} - 2\mathbb{V}(P_t, V^\star) &= \nu_t^{\text{ref}} - \frac{2}{n_t}\sum_{i=1}^{n_t} \mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}}) + \frac{2}{n_t}\sum_{i=1}^{n_t}(\mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}}) - \mathbb{V}(P_{l_{t,i}}, V^\star)) \\
&\qquad\qquad (P_t = P_{l_{t,i}}) \\
&\overset{\text{(i)}}{\leq} \tilde{\mathcal{O}}\left(\frac{B_\star^2}{n_t} + \frac{B_\star}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}} B_{l_{t,i}}^{\text{ref}}\right) + \frac{4B_\star}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}}\left(V^\star - V_{l_{t,i}}^{\text{ref}}\right) \\
&= \tilde{\mathcal{O}}\left(\frac{B_\star^2}{n_t} + \frac{B_\star}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}} B_{l_{t,i}}^{\text{ref}} + B_\star\beta_{q^\star}\right), \qquad (V^\star(s) - V_{l_{t,i}}^{\text{ref}}(s) \leq B_{l_{t,i}}^{\text{ref}}(s) + \beta_{q^\star}, \forall s)
\end{aligned}$$

where in (i) we apply the bound for $\nu_t^{\text{ref}} - \frac{2}{n_t}\sum_{i=1}^{n_t} \mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}})$, $B_t \leq B_\star$ and

$$\mathbb{V}(P_{l_{t,i}}, V_{l_{t,i}}^{\text{ref}}) - \mathbb{V}(P_{l_{t,i}}, V^\star) \leq (P_{l_{t,i}} V^\star)^2 - (P_{l_{t,i}} V_{l_{t,i}}^{\text{ref}})^2 \leq 2B_\star P_{l_{t,i}}(V^\star - V_{l_{t,i}}^{\text{ref}}).$$

Plugging the inequality above back, we have with probability at least $1 - 11\delta$,

$$\begin{aligned}
\sum_{t=1}^{T}\sqrt{\frac{\nu_t^{\text{ref}}}{n_t}} &= \tilde{\mathcal{O}}\left(\sum_{t=1}^{T}\sqrt{\frac{\mathbb{V}(P_t, V^\star)}{n_t}} + \frac{B_\star}{n_t} + \frac{1}{n_t}\sqrt{B_\star\sum_{i=1}^{n_t} P_{l_{t,i}} B_{l_{t,i}}^{\text{ref}}} + \sqrt{\frac{B_\star\beta_{q^\star}}{n_t}}\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + B_\star SA + \sqrt{\sum_{t=1}^{T}\frac{B_\star}{n_t}}\sqrt{\sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n_t} P_{l_{t,i}} B_{l_{t,i}}^{\text{ref}}} + \sqrt{B_\star\beta_{q^\star} SAT}\right) \\
&\qquad\qquad (\text{Lemma 14 and Cauchy-Schwarz inequality}) \\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star SAC_K} + B_\star SA + \sqrt{B_\star SAC_{\text{REF}}} + \sqrt{B_\star\beta_{q^\star} SAT}\right). \\
&\qquad\qquad (\text{Lemma 5, Lemma 14, Lemma 12 and Lemma 10})
\end{aligned}$$

Moreover,

$$\begin{aligned}
\sum_{t=1}^{T}\sqrt{\frac{\nu_t}{m_t}} &\leq \sum_{t=1}^{T}\frac{\sqrt{\sum_{i=1}^{m_t}(V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - V_{\check{l}_{t,i}}^{\text{ref}}(s'_{\check{l}_{t,i}}))^2}}{m_t} \leq \sum_{t=1}^{T}\frac{\sqrt{\sum_{i=1}^{m_t}(V^\star(s'_{\check{l}_{t,i}}) - V_{\check{l}_{t,i}}^{\text{ref}}(s'_{\check{l}_{t,i}}))^2}}{m_t} \\
&= \tilde{\mathcal{O}}\left(\sum_{t=1}^{T}\frac{\sqrt{\sum_{i=1}^{m_t} B_{\check{l}_{t,i}}^{\text{ref}}(s'_{\check{l}_{t,i}})^2}}{m_t} + \frac{\sqrt{\sum_{i=1}^{m_t}\beta_{q^\star}^2}}{m_t}\right) \\
&\qquad (V^\star(s'_{\check{l}_{t,i}}) - V_{\check{l}_{t,i}}^{\text{ref}}(s'_{\check{l}_{t,i}}) \leq B_{\check{l}_{t,i}}^{\text{ref}}(s'_{\check{l}_{t,i}}) + \beta_{q^\star}, (a+b)^2 \leq 2a^2 + 2b^2, \text{ and } \sqrt{x+y} \leq \sqrt{x} + \sqrt{y}) \\
&= \tilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^{T}\frac{1}{m_t}}\sqrt{\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t} B_{\check{l}_{t,i}}^{\text{ref}}(s'_{\check{l}_{t,i}})^2} + \sum_{t=1}^{T}\sqrt{\frac{\beta_{q^\star}^2}{m_t}}\right). \\
&\qquad\qquad (\text{Cauchy-Schwarz inequality})
\end{aligned}$$

Note that by Lemma 12, Lemma 28, $B_{t+1}^{\text{ref}}(s'_t) \leq B_{t+1}^{\text{ref}}(s_{t+1})$ and Lemma 8:

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t} B_{\check{l}_{t,i}}^{\text{ref}}(s'_{\check{l}_{t,i}})^2 &\leq \left(1 + \frac{1}{H}\right)\sum_{t=1}^{T} B_t^{\text{ref}}(s'_t)^2 \\
&= \tilde{\mathcal{O}}\left(\sum_{t=1}^{T} B_{t+1}^{\text{ref}}(s'_t) + SB_\star^2\right) = \tilde{\mathcal{O}}\left(\sum_{t=1}^{T} B_t^{\text{ref}}(s_t) + SB_\star^2\right) = \tilde{\mathcal{O}}\left(C_{\text{REF, 2}}\right).
\end{aligned}$$

Plugging this back to the last inequality, and by Lemma 14, we have:

$$\sum_{t=1}^{T}\sqrt{\frac{\nu_t}{m_t}} = \tilde{\mathcal{O}}\left(\sqrt{SAHC_{\text{REF, 2}}} + \sqrt{SAH\beta_{q^\star}^2 T}\right).$$

Finally, by Cauchy-Schwarz inequality, Eq. (15), Eq. (7) and Lemma 36:

$$\sum_{t=1}^{T}\sqrt{\frac{\widehat{c}_t\varepsilon_t}{n_t}} = \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\widehat{c}_t\varepsilon_t}\right) = \tilde{\mathcal{O}}\left(\sqrt{SA\left(\sum_{t=1}^{T}c(s_t,a_t)+\sum_{t=1}^{T}(\widehat{c}_t-c(s_t,a_t))\varepsilon_t\right)}\right)$$
$$= \tilde{\mathcal{O}}\left(\sqrt{SAC_K}+\sqrt{SA\sum_{t=1}^{T}\sqrt{\frac{\widehat{c}_t\varepsilon_t}{n_t}}}+SA\right).$$

Solving a quadratic equation gives $\sum_{t=1}^{T}\sqrt{\frac{\widehat{c}_t\varepsilon_t}{n_t}} = \tilde{\mathcal{O}}\left(\sqrt{SAC_K}+SA\right)$. Putting everything together, and by $\beta_{q^\star}=\mathcal{O}\left(c_{\min}\right), \beta_{q^\star}T=\mathcal{O}\left(c_{\min}T\right)=\mathcal{O}\left(C_K\right)$:

$$\sum_{t=1}^{T}b_t = \tilde{\mathcal{O}}\left(\sqrt{B_\star SAC_K}+\sqrt{B_\star SAC_{\text{REF}}}+\sqrt{SAHC_{\text{REF, 2}}}+\sqrt{SAHc_{\min}C_K}+B_\star HSA\right)$$
$$= \tilde{\mathcal{O}}\left(\sqrt{B_\star SAC_K}+B_\star H^2S^{\frac{3}{2}}A+\sqrt{SAHc_{\min}C_K}\right).$$

$\left(H=\Omega\left(\frac{B_\star}{c_{\min}}\right)\right.$ and definition of $C_{\text{REF}}, C_{\text{REF, 2}}$ (Lemma 8))

□

Lemma 12 (bias of the update scheme)**.** *Assuming* $X_t\geq 0$, *we have:*

$$\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}X_{\check{l}_{t,i}} \leq \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}X_t, \qquad \sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n_t}X_{l_{t,i}} = \mathcal{O}\left(\ln(T)\sum_{t=1}^{T}X_t\right).$$

Proof. For the first inequality, denote by j_t the stage to which time step t belongs. When $t'=\check{l}_{t,i}$, we have $m_t=e_{j_{t'}}$. Therefore, $\sum_{t=1}^{T}\sum_{i=1}^{m_t}\frac{1}{m_t}\mathbb{I}\{t'=\check{l}_{t,i}\}\leq\frac{e_{j_{t'}+1}}{e_{j_{t'}}}\leq 1+\frac{1}{H}$, and

$$\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}X_{\check{l}_{t,i}} = \sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}\sum_{t'=1}^{T}X_{t'}\mathbb{I}\{t'=\check{l}_{t,i}\} = \sum_{t'=1}^{T}X_{t'}\sum_{t=1}^{T}\sum_{i=1}^{m_t}\frac{\mathbb{I}\{t'=\check{l}_{t,i}\}}{m_t} \leq \left(1+\frac{1}{H}\right)\sum_{t'=1}^{T}X_{t'}.$$

For the second inequality:

$$\sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n_t}X_{l_{t,i}} = \sum_{t=1}^{T}\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{t'=1}^{T}X_{t'}\mathbb{I}\{t'=l_{t,i}\} = \sum_{t'=1}^{T}X_{t'}\sum_{t=1}^{T}\sum_{i=1}^{n_t}\frac{\mathbb{I}\{t'=l_{t,i}\}}{n_t}$$
$$\leq \sum_{t'=1}^{T}X_{t'}\sum_{z:t'\leq E_{z-1}\leq T}\frac{e_z}{E_{z-1}} = \mathcal{O}\left(\ln(T)\sum_{t'=1}^{T}X_{t'}\right).$$

□

Lemma 13. *Assuming* $X_t:\mathcal{S}^+\rightarrow[0,B]$ *is monotonic in* t *(i.e.,* $X_t(s)$ *is non-increasing or non-decreasing in* t *for any* $s\in\mathcal{S}^+$*) and* $X_t(g)=0$*, with probability at least* $1-\delta$*,*

$$\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}P_{\check{l}_{t,i}}X_{\check{l}_{t,i}} \leq \left(1+\frac{1}{H}\right)^2\sum_{t=1}^{T}X_t(s_t)+\tilde{\mathcal{O}}\left(B(H+S)\right).$$

Proof. By Lemma 12, Lemma 36 and Lemma 28, $X_{t+1}(s'_t)\leq X_{t+1}(s_{t+1})$ in each step,

$$\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}P_{\check{l}_{t,i}}X_{\check{l}_{t,i}} \leq \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}P_tX_t \leq \left(1+\frac{1}{H}\right)^2\sum_{t=1}^{T}X_t(s'_t)+\tilde{\mathcal{O}}\left(BH\right)$$
$$\leq \left(1+\frac{1}{H}\right)^2\sum_{t=1}^{T}X_t(s_t)+\tilde{\mathcal{O}}\left(B(H+S)\right).$$

□

Lemma 14. *For any non-negative weights* $\{w_t\}_t$*, and* $\alpha \in (0,1)$*, we have:*

$$\sum_{t=1}^{T} \frac{w_t \varepsilon_t}{n_t^{\alpha}} = \mathcal{O}\left(\left(\left\| w \right\|_{\infty} SA \right)^{\alpha} \left\| w \right\|_1^{1-\alpha} \right), \quad \sum_{t=1}^{T} \frac{w_t \varepsilon_t}{m_t^{\alpha}} = \mathcal{O}\left(\left(\left\| w \right\|_{\infty} HSA \right)^{\alpha} \left\| w \right\|_1^{1-\alpha} \ln \frac{\left\| w \right\|_{\infty}}{\left\| w \right\|_1} \right).$$

Moreover, when $w_t = v(s_t, a_t)$ *for some* v*,*

$$\sum_{t=1}^{T} \frac{w_t \varepsilon_t}{n_t^{\alpha}} = \tilde{\mathcal{O}}\left(\sum_{(s,a)} v(s,a) N_{T+1}(s,a)^{1-\alpha} \right), \quad \sum_{t=1}^{T} \frac{w_t \varepsilon_t}{m_t^{\alpha}} = \tilde{\mathcal{O}}\left(H^{\alpha} \sum_{(s,a)} v(s,a) N_{T+1}(s,a)^{1-\alpha} \right).$$

In case $w_t = 1$ *for all t, it holds that:*

$$\sum_{t=1}^{T} \frac{\varepsilon_t}{n_t^{\alpha}} = \tilde{\mathcal{O}}\left((SA)^{\alpha} T^{1-\alpha} \right), \quad \sum_{t=1}^{T} \frac{\varepsilon_t}{m_t^{\alpha}} = \tilde{\mathcal{O}}\left((SAH)^{\alpha} T^{1-\alpha} \right), \tag{14}$$

when $0 < \alpha < 1$*, and*

$$\sum_{t=1}^{T} \frac{\varepsilon_t}{n_t} = \mathcal{O}\left(SA \ln T \right), \quad \sum_{t=1}^{T} \frac{\varepsilon_t}{m_t} = \mathcal{O}\left(SAH \ln T \right), \tag{15}$$

when $\alpha = 1$*.*

Proof. Define $\mathfrak{n}(s,a,j) = \sum_{t:(s_t,a_t)=(s,a), n_t = E_j} w_t$, $\mathfrak{n}(s,a) = \sum_{j \geq 0} \mathfrak{n}(s,a,j)$. Then, $\sum_{(s,a)} \mathfrak{n}(s,a) = \left\| w \right\|_1$, $\mathfrak{n}(s,a,j) \leq \left\| w \right\|_{\infty} e_{j+1} \leq \left(1 + \frac{1}{H} \right) \left\| w \right\|_{\infty} e_j$. Moreover, by definitions of e_j and E_j,

$$\sum_{j \geq 1} \mathbb{I}\left\{ \left(1 + \frac{1}{H} \right) \left\| w \right\|_{\infty} E_{j-1} \leq \mathfrak{n}(s,a) \right\} = \mathcal{O}\left(H \ln \frac{\left\| w \right\|_1}{\left\| w \right\|_{\infty}} \right). \tag{16}$$

$$\sum_{j \geq 1} e_j \mathbb{I}\left\{ \left(1 + \frac{1}{H} \right) \left\| w \right\|_{\infty} E_{j-1} \leq \mathfrak{n}(s,a) \right\} = \mathcal{O}(\mathfrak{n}(s,a) / \left\| w \right\|_{\infty}). \tag{17}$$

Since $\frac{1}{E_j^{\alpha}}$ and $\frac{1}{e_j^{\alpha}}$ is decreasing, by "moving weights to earlier terms" (from $\mathfrak{n}(s,a,j)$ to $\mathfrak{n}(s,a,i)$ for $i < j$),

$$\begin{aligned}
\sum_{t=1}^{T} \frac{w_t \varepsilon_t}{n_t^{\alpha}} &= \sum_{(s,a)} \sum_{j \geq 1} \frac{\mathfrak{n}(s,a,j)}{E_j^{\alpha}} \leq \sum_{(s,a)} \sum_{j \geq 1} \left(1 + \frac{1}{H} \right) \left\| w \right\|_{\infty} \frac{e_j \mathbb{I}\left\{ \left(1 + \frac{1}{H} \right) \left\| w \right\|_{\infty} E_{j-1} \leq \mathfrak{n}(s,a) \right\}}{E_j^{\alpha}} \\
&= \mathcal{O}\left(\sum_{(s,a)} \left\| w \right\|_{\infty} \left(\frac{\mathfrak{n}(s,a)}{\left\| w \right\|_{\infty}} \right)^{1-\alpha} \right) && \left(\textstyle\sum_{j=1}^{J} \frac{e_j}{E_j^{\alpha}} = \mathcal{O}\left(E_J^{1-\alpha} \right) \text{ and Eq. (17)}\right) \\
&= \mathcal{O}\left(\left(\left\| w \right\|_{\infty} SA \right)^{\alpha} \left\| w \right\|_1^{1-\alpha} \right), && \text{(Hölder's inequality)}
\end{aligned}$$

$$\begin{aligned}
\sum_{t=1}^{T} \frac{w_t \varepsilon_t}{m_t^{\alpha}} &= \sum_{(s,a)} \sum_{j \geq 1} \frac{\mathfrak{n}(s,a,j)}{e_j^{\alpha}} \leq \sum_{(s,a)} \sum_{j \geq 1} \left(1 + \frac{1}{H} \right) \left\| w \right\|_{\infty} e_j^{1-\alpha} \mathbb{I}\left\{ \left(1 + \frac{1}{H} \right) \left\| w \right\|_{\infty} E_{j-1} \leq \mathfrak{n}(s,a) \right\} \\
&\leq \left(1 + \frac{1}{H} \right) \left\| w \right\|_{\infty} \left(\sum_{(s,a)} \sum_{j \geq 1} \mathbb{I}\left\{ \left\| w \right\|_{\infty} E_{j-1} \leq \mathfrak{n}(s,a) \right\} \right)^{\alpha} \left(\sum_{(s,a)} \frac{\mathfrak{n}(s,a)}{\left\| w \right\|_{\infty}} \right)^{1-\alpha} \\
& && \text{(Hölder's inequality and Eq. (17))} \\
&= \mathcal{O}\left(\left(\left\| w \right\|_{\infty} HSA \right)^{\alpha} \left\| w \right\|_1^{1-\alpha} \ln \frac{\left\| w \right\|_1}{\left\| w \right\|_{\infty}} \right). && \text{(Eq. (16))}
\end{aligned}$$

In case $w_t = 1$ and $\alpha \in (0,1)$, we have $\left\| w \right\|_{\infty} = 1$, $\left\| w \right\|_1 = T$, and Eq. (14) is proved. When $w_t = v(s_t, a_t)$ for some v, $\mathfrak{n}(s,a,j) \leq v(s,a) e_{j+1} \mathbb{I}\{j \leq J_{s,a}\}$, where $J_{s,a}$ is such that $E_{J_{s,a}} = n_T(s,a)$.

Thus,

$$\sum_{t=1}^{T}\frac{w_t\varepsilon_t}{n_t^{\alpha}} \leq \sum_{(s,a)} v(s,a)\sum_{j=1}^{J_{s,a}}\frac{e_{j+1}}{E_j^{\alpha}} = \mathcal{O}\left(\sum_{(s,a)} v(s,a)\sum_{j=1}^{J_{s,a}}\frac{e_j}{E_j^{\alpha}}\right) = \mathcal{O}\left(\sum_{(s,a)} v(s,a)N_{T+1}(s,a)^{1-\alpha}\right).$$

$$\begin{aligned}\sum_{t=1}^{T}\frac{w_t\varepsilon_t}{m_t^{\alpha}} &\leq \sum_{(s,a)} v(s,a)\sum_{j=1}^{J_{s,a}}\frac{e_{j+1}}{e_j^{\alpha}} = \mathcal{O}\left(\sum_{(s,a)} v(s,a)\sum_{j=1}^{J_{s,a}} e_j^{1-\alpha}\right)\\ &= \tilde{\mathcal{O}}\left(\sum_{(s,a)} v(s,a)J_{s,a}^{\alpha}\left(\sum_{j=1}^{J_{s,a}} e_j\right)^{1-\alpha}\right) = \tilde{\mathcal{O}}\left(H^{\alpha}\sum_{(s,a)} v(s,a)N_{T+1}(s,a)^{1-\alpha}\right).\end{aligned}$$

(Hölder's inequality and $J_{s,a} = \tilde{\mathcal{O}}\left(H\right)$ by how e_j grows)

In case $\alpha = 1$, we have:

$$\sum_{t=1}^{T}\frac{\varepsilon_t}{n_t} \leq \sum_{(s,a)}\sum_{j:0<E_{j-1}\leq T}\frac{e_j}{E_{j-1}} = \mathcal{O}\left(SA\ln T\right).$$

$$\sum_{t=1}^{T}\frac{\varepsilon_t}{m_t} \leq \sum_{(s,a)}\sum_{j:0<E_{j-1}\leq T}\left(1+\frac{1}{H}\right) = \mathcal{O}\left(SAH\ln T\right).$$

$\square$

D.5 Parameter free algorithm

In this section, we present a parameter-free model-free algorithm (Algorithm 5) that achieves the same regret guarantee as Algorithm 2 (up to log factors). The high level idea is to first apply the doubling trick from Tarbouriech et al. [2021b] to determine an upper bound on $B_\star$, then try logarithmically many different values of H and $\theta^\star$ simultaneously, each leading to a different update rule for Q and V^{ref}.

D.5.1 An upper bound on $B_\star$ is available

We first introduce Algorithm 4, which is a sub-algorithm that achieves the desired regret bound when we have an upper bound $\widetilde{B} \geq B_\star$. In this case, we only need to determine the appropriate value of H and $\theta^\star$. Define $N_\beta = \lceil \log_2(1/\beta) \rceil$ with $\beta = \frac{c_{\min}}{2\widetilde{B}^2SAK}$, $H_p = 2^p$ for $p \in \mathcal{P}$ with $\mathcal{P} = [N_\beta]$, and $\mathcal{H} = \{H_p\}_{p\in\mathcal{P}}$. Define $\mathcal{R} = [8N_\beta]$. Here, $\mathcal{H}$ and $\{2^r\}_{r\in\mathcal{R}}$ constitute the search range of H and $\theta^\star$.

For each p, r, we maintain accumulators $\mu^{\text{ref}}_{p,r}, \sigma^{\text{ref}}_{p,r}, \mu_{p,r}, \sigma_{p,r}, v_p, m_p$ similar to $\mu^{\text{ref}}, \sigma^{\text{ref}}, \mu, \sigma, v, m$ in Algorithm 2 (Line 2 and Line 3). For each $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $p \in \mathcal{P}$, we divide the samples received into consecutive stages, where the length of the j-th stage is $e_{p,j}$ with $e_{p,1} = H_p, e_{p,j+1} = \lfloor(1+\frac{1}{H_p})e_{p,j}\rfloor$. Also define the indices indicating the end of a stage for a given p as $\mathcal{L}_p = \{E_{p,j}\}_{j\in\mathbb{N}^+}$ with $E_{p,j} = \sum_{i=1}^{j} e_{p,i}$. We update $Q(s,a)$ only when the number of visits to (s,a) falls into $\mathcal{L}_p$ for some $p \in \mathcal{P}$ (Line 4), and there are two types of update rules similar to Algorithm 2 (Line 5 and Line 6). We also maintain $|\mathcal{R}|$ reference value functions, each with different final precision (Line 7). We show that the way we combine different update rules enable us to apply analysis of Algorithm 2 w.r.t any choice of $(p,r) \in \mathcal{P}\times\mathcal{R}$. Notably, we can proceed with $(p^\star, r^\star)$ with $H_{p^\star} = H, 2^{r^\star} = \theta^\star$, which gives us the same regret bound as Algorithm 2 without knowing $B_\star$.

Now we introduce some notations only used in this section. When it is clear from the context, we ignore dependency on p, and define $n_t(s,a), m_t(s,a), l_{t,i}(s,a), \check{l}_{t,i}(s,a), \widehat{c}_t(s,a)$ similarly as before for a given p. Denote by $V^{\text{ref}}_{r,t}(s)$ the value of $V^{\text{ref}}_r(s)$ at the beginning of time step t, and by

Algorithm 4 LCB-ADVANTAGE-SSP with an upper bound on $B_\star$

```
Parameter: initial value function upper bound $\widetilde{B} \geq B_\star$, failure probability $\delta \in (0,1)$.
Define: $\mathcal{L}_p = \{E_{p,j}\}_{j\in\mathbb{N}^+}$ where $E_{p,j} = \sum_{i=1}^{j} e_{p,i}$, $e_{p,1} = H_p$ and $e_{p,j+1} = \lfloor (1+1/H_p)e_{p,j} \rfloor$.
Initialize: $t \leftarrow 0$, $s_1 \leftarrow s_{\text{init}}$, $B \leftarrow 1$, for all $(s,a), p \in \mathcal{P}$, $N(s,a) \leftarrow 0$, $M_p(s,a) \leftarrow 0$.
Initialize: for all $(s,a), r \in \mathcal{R}$, $Q(s,a) \leftarrow 0$, $V(s) \leftarrow 0$, $V^{\text{ref}}_r(s) \leftarrow V(s)$, $\widehat{C}(s,a) \leftarrow 0$.
Initialize: for all $(s,a), p \in \mathcal{P}, r \in \mathcal{R}$, $\mu^{\text{ref}}_{p,r}(s,a) \leftarrow 0$, $\sigma^{\text{ref}}_{p,r}(s,a) \leftarrow 0$, $\mu_{p,r}(s,a) \leftarrow 0$, $\sigma_{p,r}(s,a) \leftarrow 0$, $v_p(s,a) \leftarrow 0$.
for $k = 1, \ldots, K$ do
    repeat
        Increment time step $t \overset{+}{\leftarrow} 1$.
        Take action $a_t = \operatorname{argmin}_a Q(s_t, a)$, suffer cost $c_t$, transit to and observe $s'_t$.
        Update global accumulators: $n = N(s_t,a_t) \overset{+}{\leftarrow} 1$, $\widehat{C}(s_t,a_t) \overset{+}{\leftarrow} c_t$.
        Compute $\iota \leftarrow 256\ln^6(4SA\widetilde{B}^8 n^5 \cdot 8N_\beta^2/\delta)$, $\widehat{c} \leftarrow \frac{\widehat{C}(s_t,a_t)}{n}$.
        for $p \in \mathcal{P}$ do
            for $r \in \mathcal{R}$ do
                Update reference value accumulators: $\mu^{\text{ref}}_{p,r}(s_t,a_t) \overset{+}{\leftarrow} V^{\text{ref}}_r(s'_t)$, $\sigma^{\text{ref}}_{p,r}(s_t,a_t) \overset{+}{\leftarrow} V^{\text{ref}}_r(s'_t)^2$, $\mu_{p,r}(s_t,a_t) \overset{+}{\leftarrow} V(s'_t) - V^{\text{ref}}_r(s'_t)$, $\sigma_{p,r}(s_t,a_t) \overset{+}{\leftarrow} (V(s'_t) - V^{\text{ref}}_r(s'_t))^2$.
            Update accumulators: $v_p(s_t,a_t) \overset{+}{\leftarrow} V(s'_t)$, $m_p = M_p(s_t,a_t) \overset{+}{\leftarrow} 1$.
            if $n \in \mathcal{L}_p$ then
                for $r \in \mathcal{R}$ do
                    $b_{p,r} \leftarrow \sqrt{\frac{\sigma^{\text{ref}}_{p,r}(s_t,a_t)/n - (\mu^{\text{ref}}_{p,r}(s_t,a_t)/n)^2}{n}\iota} + \sqrt{\frac{\sigma_{p,r}(s_t,a_t)/m_p - (\mu_{p,r}(s_t,a_t)/m_p)^2}{m_p}\iota} + \left(\frac{4B}{n} + \frac{3B}{m_p}\right)\iota + \sqrt{\frac{\widehat{c}\iota}{n}}$.
                    $Q(s_t,a_t) \leftarrow \max\left\{\widehat{c} + \frac{\mu^{\text{ref}}_{p,r}(s_t,a_t)}{n} + \frac{\mu_{p,r}(s_t,a_t)}{m_p} - b_{p,r}, Q(s_t,a_t)\right\}$.
                    Reset local accumulators: $\mu_{p,r}(s_t,a_t) \leftarrow 0$, $\sigma_{p,r}(s_t,a_t) \leftarrow 0$.
                Compute bonus $b'_p \leftarrow 2\sqrt{\frac{B^2\iota}{m_p}} + \sqrt{\frac{\widehat{c}\iota}{n}} + \frac{\iota}{n}$.
                $Q(s_t,a_t) \leftarrow \max\left\{\widehat{c} + \frac{v_p(s_t,a_t)}{m_p} - b'_p, Q(s_t,a_t)\right\}$.
                Reset local accumulators: $v_p(s_t,a_t) \leftarrow 0$, $M_p(s_t,a_t) \leftarrow 0$.
        $V(s_t) \leftarrow \min_a Q(s_t,a)$.
        if $V(s_t) > B$ then $B \leftarrow 2V(s_t)$.
        if $\sum_a N(s_t,a) = 2^r$ for some $r \in \mathcal{R}$ then $V^{\text{ref}}_{r'}(s_t) \leftarrow V(s_t)$, $\forall r' \geq r$.
        if $s'_t \neq g$ then $s_{t+1} \leftarrow s'_t$; else $s_{t+1} \leftarrow s_{\text{init}}$, break.
```

$b_{p,r,t}(s,a), b'_{p,t}(s,a)$ the value of $b_{p,r}(s,a), b'_p(s,a)$ in $Q_t(s,a)$. Also define:

$$\overline{Q}_{p,r,t}(s,a) = \widehat{c}_t(s,a) + \frac{1}{n_t}\sum_{i=1}^{n_t} V^{\text{ref}}_{r,l_{t,i}}(s'_{l_{t,i}}) + \frac{1}{m_t}\sum_{i=1}^{m_t}\left(V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - V^{\text{ref}}_{r,\check{l}_{t,i}}(s'_{\check{l}_{t,i}})\right) - b_{p,r,t}.$$

$$\overline{Q}'_{p,t}(s,a) = \widehat{c}_t(s,a) + \frac{1}{m_t}\sum_{i=1}^{m_t} V_{\check{l}_{t,i}}(s'_{\check{l}_{t,i}}) - b'_{p,t}.$$

Note that for any $(s,a) \in \mathcal{S} \times \mathcal{A}, t > 1$,

$$Q_t(s,a) = \max\left\{\max_{p,r} \overline{Q}_{p,r,t}(s,a), \max_p \overline{Q}'_{p,t}(s,a), Q_{t-1}(s,a)\right\}. \tag{18}$$

Next, we prove the key lemma of Algorithm 4, which shows that Q_t is an optimistic estimator of $Q^\star$.

Lemma 15. *With probability at least $1 - 7\delta$, Algorithm 4 with input $\widetilde{B} \geq B_\star$ ensures $Q_t(s,a) \leq Q_{t+1}(s,a) \leq Q^\star(s,a)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

Proof. The first inequality is by the update rule of Q_t. Next, we prove $Q_t(s,a) \le Q^\star(s,a)$ by induction on t. It is clearly true when when $t = 1$. For the induction step, note that for any p, r, the proof of Lemma 9 still proceeds to conclude that $\overline{Q}_{p,r,t}(s,a) \le Q^\star(s,a)$ and $\overline{Q}'_{p,t}(s,a) \le Q^\star(s,a)$, where we substitute b_t with $b_{p,r,t}$, b'_t with $b'_{p,t}$, and V^{ref}_t with $V^{\text{ref}}_{r,t}$ (also note Remark 1). Thus, by a union bound over $8N_\beta^2$ update rules, the computation of ι (Line 5), and Eq. (18), the claim is proved. □

Theorem 7. *With probability at least* $1 - 60\delta$, *Algorithm 4 with input* $\widetilde{B} \ge B_\star$ *ensures* $R_K = \tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + \frac{B_\star^5 S^2 A}{c_{\min}^4}\right)$.

Proof. Define $V^{\text{ref}} = V^{\text{ref}}_{r^\star}, V^{\text{REF}} = V^{\text{ref}}_{r^\star,T+1}, b_t = b_{p^\star,r^\star,t}, b'_t = b'_{p^\star,t}, H = H_{p^\star}$, and n_t, m_t, $l_{t,i}$, $\check{l}_{t,i}$ are defined for $p^\star$. We have Lemma 12, Lemma 6, Corollary 6, Lemma 8, Lemma 10, Lemma 11 and Theorem 2 holds for Algorithm 4. Following the steps in the proof of Theorem 3 gives the desired result. □

D.5.2 Without knowledge of $B_\star$

Now we introduce our parameter-free algorithm that achieves the desired regret bound without knowledge of $B_\star$. The main idea is to determine an upper bound on $B_\star$ using a doubling trick from [Tarbouriech et al., 2021b], and then run Algorithm 4 as a sub-algorithm. We divide the learning process into epochs indexed by ϕ. We maintain value function upper bound $\widetilde{B}$ and cost accumulator C recording the total costs suffered in current epoch. In epoch ϕ, we execute Algorithm 4 with value function upper bound $\widetilde{B}$. Moreover, we start a new epoch whenever:

1. $B > \widetilde{B}$,
2. or $C > \widetilde{B}K + x\left(\widetilde{B}\sqrt{SAK} + \frac{\widetilde{B}^5 S^2 A}{c_{\min}^4}\right)$.

Here, x is a large enough constant determined by Theorem 7, so that when $\widetilde{B} \ge B_\star$, we have with probability at least $1 - 60\delta$:

$$C - V^\star(s_{\text{init}}^\phi) - (K-1)V^\star(s_{\text{init}}) \le x\left(\widetilde{B}\sqrt{SAK} + \frac{\widetilde{B}^5 S^2 A}{c_{\min}^4}\right),$$

where s_{init}^ϕ is the initial state of epoch ϕ (note that Theorem 3 still holds when the initial state is changing over episodes). Moreover, we double the value of $\widetilde{B}$ whenever a new epoch starts. We summarize ideas above in Algorithm 5.

Theorem 8. *With probability at least* $1-60\delta$, *Algorithm 5 ensures* $R_K = \tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + \frac{B_\star^5 S^2 A}{c_{\min}^4}\right)$.

Proof. Denote by B_ϕ the value of B in epoch ϕ, and by C_ϕ the value of C at the end of epoch ϕ. Define $\phi^\star = \inf_\phi\{B_\phi \ge B_\star\}$. Clearly $B_\phi \le \max\{2B_\star, K\}$ for $\phi \le \phi^\star$. By Theorem 7, with probability at least $1 - 60\delta$, there is at most $\phi^\star$ epochs since the condition of starting a new epoch will never be triggered in epoch $\phi^\star$, and the regret in epoch $\phi^\star$ is properly bounded:

$$C_{\phi^\star} - V^\star(s_{\text{init}}^{\phi^\star}) - (K-1)V^\star(s_{\text{init}}) = \tilde{\mathcal{O}}\left(\widetilde{B}_{\phi^\star}\sqrt{SAK} + \frac{\widetilde{B}_{\phi^\star}^5 S^2 A}{c_{\min}^4}\right) = \tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + \frac{B_\star^5 S^2 A}{c_{\min}^4}\right).$$

Conditioned on the event that there are at most $\phi^\star$ epochs, we partition the regret into two parts: the total costs suffered before epoch $\phi^\star$, and the regret starting from epoch $\phi^\star$. It suffices to bound the total costs before epoch $\phi^\star$ assuming $K \le B_\star$ (otherwise $\phi^\star = 1$). By the update scheme of $\widetilde{B}$, we have at most $\lceil \log_2 B_\star \rceil + 1$ epochs before epoch $\phi^\star$. Moreover, by the second condition of starting a new epoch, the accumulated cost in epoch $\phi < \phi^\star$ is bounded by:

$$C_\phi \le K\widetilde{B}_\phi + \tilde{\mathcal{O}}\left(\widetilde{B}_\phi\sqrt{SAK} + \widetilde{B}_\phi S^2 A\right) = \tilde{\mathcal{O}}\left(\frac{B_\star^5 S^2 A}{c_{\min}^4}\right).$$

Algorithm 5 LCB-ADVANTAGE-SSP without knowledge of $B_\star$

```
Parameter: failure probability $\delta \in (0,1)$.
Define: $\mathcal{L}_p = \{E_{p,j}\}_{j\in\mathbb{N}^+}$ where $E_{p,j} = \sum_{i=1}^{j} e_{p,i}$, $e_{p,1} = H_p$ and $e_{p,j+1} = \lfloor (1+1/H_p)e_{p,j} \rfloor$.
Initialize: $\widetilde{B} \leftarrow K, C \leftarrow 0$.
Initialize: $t \leftarrow 0$, $s_1 \leftarrow s_{\text{init}}$, $B \leftarrow 1$, for all $(s,a), p \in \mathcal{P}, N(s,a) \leftarrow 0, M_p(s,a) \leftarrow 0$.
Initialize: for all $(s,a), r \in \mathcal{R}, Q(s,a) \leftarrow 0, V(s) \leftarrow 0, V_r^{\text{ref}}(s) \leftarrow V(s), \widehat{C}(s,a) \leftarrow 0$.
Initialize: for all $(s,a), p \in \mathcal{P}, r \in \mathcal{R}$, $\mu_{p,r}^{\text{ref}}(s,a) \leftarrow 0$, $\sigma_{p,r}^{\text{ref}}(s,a) \leftarrow 0$, $\mu_{p,r}(s,a) \leftarrow 0$, $\sigma_{p,r}(s,a) \leftarrow 0$, $v_p(s,a) \leftarrow 0$.
for $k = 1, \ldots, K$ do
    repeat
        Increment time step $t \overset{+}{\leftarrow} 1$.
        Take action $a_t = \operatorname{argmin}_a Q(s_t,a)$, suffer cost $c_t$, transit to and observe $s'_t$.
        Update global accumulators: $n = N(s_t,a_t) \overset{+}{\leftarrow} 1$, $\widehat{C}(s_t,a_t) \overset{+}{\leftarrow} c_t$, $C \overset{+}{\leftarrow} c_t$.
        Compute $\iota \leftarrow 256 \ln^6(4SA\widetilde{B}^8 n^5 \cdot 8N_\beta^2/\delta)$, $\widehat{c} \leftarrow \frac{\widehat{C}(s_t,a_t)}{n}$.
        for $p \in \mathcal{P}$ do
            for $r \in \mathcal{R}$ do
                Update reference value accumulators: $\mu_{p,r}^{\text{ref}}(s_t,a_t) \overset{+}{\leftarrow} V_r^{\text{ref}}(s'_t)$, $\sigma_{p,r}^{\text{ref}}(s_t,a_t) \overset{+}{\leftarrow} V_r^{\text{ref}}(s'_t)^2$, $\mu_{p,r}(s_t,a_t) \overset{+}{\leftarrow} V(s'_t) - V_r^{\text{ref}}(s'_t)$, $\sigma_{p,r}(s_t,a_t) \overset{+}{\leftarrow} (V(s'_t) - V_r^{\text{ref}}(s'_t))^2$.
            Update accumulators: $v_p(s_t,a_t) \overset{+}{\leftarrow} V(s'_t)$, $m_p = M_p(s_t,a_t) \overset{+}{\leftarrow} 1$.
            if $n \in \mathcal{L}_p$ then
                for $r \in \mathcal{R}$ do
                    $b_{p,r} \leftarrow \sqrt{\frac{\sigma_{p,r}^{\text{ref}}(s_t,a_t)/n - (\mu_{p,r}^{\text{ref}}(s_t,a_t)/n)^2}{n}\iota} + \sqrt{\frac{\sigma_{p,r}(s_t,a_t)/m_p - (\mu_{p,r}(s_t,a_t)/m_p)^2}{m_p}\iota} + \left(\frac{4B}{n} + \frac{3B}{m_p}\right)\iota + \sqrt{\frac{\widehat{c}\iota}{n}}$.
                    $Q(s_t,a_t) \leftarrow \max\left\{\widehat{c} + \frac{\mu_{p,r}^{\text{ref}}(s_t,a_t)}{n} + \frac{\mu_{p,r}(s_t,a_t)}{m_p} - b_{p,r}, Q(s_t,a_t)\right\}$.
                    Reset local accumulators: $\mu_{p,r}(s_t,a_t) \leftarrow 0$, $\sigma_{p,r}(s_t,a_t) \leftarrow 0$.
                Compute bonus $b'_p \leftarrow 2\sqrt{\frac{B^2\iota}{m_p}} + \sqrt{\frac{\widehat{c}\iota}{n}} + \frac{\iota}{n}$.
                $Q(s_t,a_t) \leftarrow \max\left\{\widehat{c} + \frac{v_p(s_t,a_t)}{m_p} - b'_p, Q(s_t,a_t)\right\}$.
                Reset local accumulators: $v_p(s_t,a_t) \leftarrow 0$, $M_p(s_t,a_t) \leftarrow 0$.
        $V(s_t) \leftarrow \min_a Q(s_t,a)$.
        if $V(s_t) > B$ then $B \leftarrow 2V(s_t)$.
        if $\sum_a N(s_t,a) = 2^r$ for some $r \in \mathcal{R}$ then $V_{r'}^{\text{ref}}(s_t) \leftarrow V(s_t)$, $\forall r' \geq r$.
        if $B > \widetilde{B}$ or $C > \widetilde{B}K + x\left(\widetilde{B}\sqrt{SAK} + \frac{\widetilde{B}^5 S^2 A}{c_{\min}^4}\right)$ then
            $\widetilde{B} \leftarrow 2\widetilde{B}$, $C \leftarrow 0$.
            $B \leftarrow 1$, for all $(s,a), p \in \mathcal{P}, N(s,a) \leftarrow 0, M_p(s,a) \leftarrow 0$.
            for all $(s,a), r \in \mathcal{R}, Q(s,a) \leftarrow 0, V(s) \leftarrow 0, V_r^{\text{ref}}(s) \leftarrow V(s), \widehat{C}(s,a) \leftarrow 0$.
            for all $(s,a), p \in \mathcal{P}, r \in \mathcal{R}$, $\mu_{p,r}^{\text{ref}}(s,a) \leftarrow 0$, $\sigma_{p,r}^{\text{ref}}(s,a) \leftarrow 0$, $\mu_{p,r}(s,a) \leftarrow 0$, $\sigma_{p,r}(s,a) \leftarrow 0$, $v_p(s,a) \leftarrow 0$.
        if $s'_t \neq g$ then $s_{t+1} \leftarrow s'_t$; else $s_{t+1} \leftarrow s_{\text{init}}$, break.
```

Combining these two parts, we get:

$$
\begin{aligned}
R_K &= \sum_{\phi=1}^{\phi^\star-1} C_\phi + \left(C_{\phi^\star} - V^\star(s_{\text{init}}^{\phi^\star}) - (K-1)V^\star(s_{\text{init}})\right) + \left(V^\star(s_{\text{init}}^{\phi^\star}) - V^\star(s_{\text{init}})\right) \\
&= \tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + \frac{B_\star^5 S^2 A}{c_{\min}^4}\right),
\end{aligned}
$$

where we assume $C_{\phi^\star} = 0$ and $s_{\text{init}}^{\phi^\star} = s_{\text{init}}$ if there are less than $\phi^\star$ epochs. $\square$

E Omitted Details for Section 5

Extra Notations Denote by $Q_t(s,a), V_t(s)$ the value of $Q(s,a), V(s)$ at the beginning of time step t, $V_0(s) = 0$, and $b_t(s,a), n_t(s,a), \bar{P}_{t,s,a}(s'), \iota_t(s,a), \widehat{c}_t(s,a)$ the value of $b, n, \bar{P}_{s,a}(s'), \iota, \widehat{c}$ used in computing $Q_t(s,a)$ (note that $b_t(s,a) = 0$ and $\widehat{c}_t(s,a) = 0$ if $n_t(s,a) = 0$). Denote by $l_t(s,a)$ the last time step the agent visits (s,a) among those $n_t(s,a)$ steps before the current stage, and $l_t(s,a) = t$ if the first visit to (s,a) is at time step t. Also define $\bar{P}_t = \bar{P}_{t,s_t,a_t}$ and $n_t^+(s,a) = \max\{1, n_t(s,a)\}$. With these notations, we have by the update rule of the algorithm:

$$Q_t(s,a) = \max\{Q_{t-1}(s,a), \widehat{c}_t(s,a) + \bar{P}_{t,s,a}V_{l_t} - b_t\}, \tag{19}$$

where b_t represents $b_t(s,a)$, and l_t represents $l_t(s,a)$ for notational convenience.

Before proving Theorem 5 (Section E.3), we first show some basic properties of our proposed update scheme (Section E.1), and proves the two required properties for Algorithm 3 (Section E.2).

E.1 Properties of Proposed Update Scheme

In this section, we prove that our proposed update scheme has the desired properties, that is, it suffers constant cost independent of H, while maintaining sparse update in the long run similar to the update scheme of Algorithm 2 (Lemma 16). We also quantify the bias induced by the sparse update compared to full-planning (that is, update every state-action pair at every time step) in Lemma 17.

Lemma 16. *The proposed update scheme satisfies the following:*

1. *For $\{X_t\}_{t\geq 0}$ such that $X_t \in [0, \mathring{B}]$ and $t < t', (s_t, a_t) = (s_{t'}, a_{t'})$ implies $X_t \geq X_{t'}$, we have: $\sum_{t=1}^T X_{l_t} \leq \mathring{B}SA + (1 + \frac{1}{H})\sum_{t=1}^T X_t$.*
2. *Denote $i_h^\star = \inf\{i \geq \mathbb{N}^+ : e_i \geq h\}$ for $h \in \mathbb{N}^+$. Then $i_h^\star = \mathcal{O}(H\ln(h))$.*

Proof. For any given $n \in \mathbb{N}^+$, define y_n as the index of the end of last stage, that is, the largest element in $\mathcal{L}$ that is smaller than n (also define $y_1 = 1$). For the first property, we first prove by induction that for any $j \in \mathbb{N}^+$, there exist non-negative weights $\{w_{n,i}\}_{n,i}$ such that:

1. For all $n \leq E_j$, $\sum_{i=1}^{y_n} w_{n,i} = \mathbb{I}\{n > 1\}$, and $w_{n,i} = 0$ for $i > y_n$.
2. $\sum_{n=1}^{E_j} w_{n,i} \leq 1 + \frac{1}{H}$ for any $i \leq E_j$.
3. $\widetilde{e}_{j+1} + \sum_{n=1}^{E_j}\sum_{n'=1}^{E_j} w_{n,n'} = (1 + 1/H)E_j$.

To give some intuition, we can imagine a continuous process where we process index n at time step n. Indices are divided into consecutive stages, and there are e_j indices in the j-th stage. At index n we need to consume 1 unit of energy accumulated up to the last stage (that is, up to index y_n) and then contributes $(1 + \frac{1}{H})$ energy to the future stages. We can think of $\widetilde{e}_j$ as the available amount of energy at the beginning of stage j (accumulated from indices up to E_{j-1}), and e_j as the amount of energy consumed in stage j (one unit by each index in stage j). The assignment of energy consumption is represented by $\{w_{n,i}\}$, where $w_{n,i}$ is the amount of energy consumed by index n which is contributed by index i. The result we are going to prove by induction states that the process described above can proceed indefinitely.

The base case of $j = 1$ is clearly true by $w_{1,i} = 0$ for any $i \in \mathbb{N}^+$ and $\widetilde{e}_2 = 1 + \frac{1}{H}$. For the induction step, by the third property, there are in total $(1 + \frac{1}{H})E_j$ energy contributed by indices up to E_j, where $\widetilde{e}_{j+1}$ is the amount of energy available to use for stages starting from $j+1$, and $\sum_{n=1}^{E_j}\sum_{n'=1}^{E_j} w_{n,n'}$ is the amount of energy consumed by indices up to E_j (we use one of the possible assignments of $\{w_{n,i}\}_{n,i}$ for $n \leq E_j$ from the previous induction step). We can easily distribute e_{j+1} weights (from $\widetilde{e}_{j+1}$) to indices in stage $j+1$ so that $\sum_{i=1}^{y_n} w_{n,i} = 1$ and $w_{n,i} = 0$ for $i > y_n$ for all $E_j < n \leq E_{j+1}$

(note that $y_n = E_j$ in this range), and $\sum_{n=1}^{E_{j+1}} w_{n,i} \leq 1 + \frac{1}{H}$ for any $i \leq E_{j+1}$. Moreover,

$$\begin{aligned}
\widetilde{e}_{j+2} + \sum_{n=1}^{E_{j+1}} \sum_{n'=1}^{E_{j+1}} w_{n,n'} &= \widetilde{e}_{j+1} + \frac{1}{H} e_{j+1} + \sum_{n=1}^{E_j} \sum_{n'=1}^{E_j} w_{n,n'} + e_{j+1} \\
&= \left(1 + \frac{1}{H}\right) E_j + \left(1 + \frac{1}{H}\right) e_{j+1} = \left(1 + \frac{1}{H}\right) E_{j+1}.
\end{aligned}$$

Thus, the induction step also holds. We are now ready to prove the first property. Denote by $t_i(s,a)$ the time step of the i-th visit to (s,a), and by $N(s,a)$ the total number of visits to (s,a) in K episodes. We have

$$\begin{aligned}
\sum_{t=1}^{T} X_{l_t} &= \sum_{(s,a)} \sum_{n=1}^{N(s,a)} X_{t_{y_n}(s,a)} \leq \sum_{(s,a)} X_{t_1(s,a)} + \sum_{(s,a)} \sum_{n=2}^{N(s,a)} \sum_{i=1}^{y_n} w_{n,i} X_{t_i(s,a)} \\
&\qquad (y_1 = 1,\ X_{t_i(s,a)} \text{ is non-increasing in } i, \text{ and } \{w_{n,i}\}_{n,i} \text{ is from the induction result}) \\
&\leq \mathring{B} SA + \sum_{(s,a)} \sum_{i=1}^{N(s,a)} X_{t_i(s,a)} \sum_{n=1}^{N(s,a)} w_{n,i} \leq \mathring{B} SA + \left(1 + \frac{1}{H}\right) \sum_{(s,a)} \sum_{i=1}^{N(s,a)} X_{t_i(s,a)} \\
&\qquad (X_{t_1(s,a)} \leq \mathring{B} \text{ and } \textstyle\sum_{n=1}^{N(s,a)} w_{n,i} \leq 1 + \frac{1}{H}) \\
&= \mathring{B} SA + \left(1 + \frac{1}{H}\right) \sum_{t=1}^{T} X_t.
\end{aligned}$$

For the second property, note that $i_h^\star = \inf\{i \in \mathbb{N}^+ : \widetilde{e}_i \geq h\}$ since h is an interger. Moreover,

$$\widetilde{e}_{i+1} = \left(1 + \frac{1}{H}\right) \widetilde{e}_i + \frac{1}{H}(e_i - \widetilde{e}_i) \geq \left(1 + \frac{1}{H}\right) \widetilde{e}_i - \frac{1}{H} \implies \widetilde{e}_{i+1} - 1 \geq \left(1 + \frac{1}{H}\right) (\widetilde{e}_i - 1)$$

$$\implies \widetilde{e}_i \geq (\widetilde{e}_{i_2^\star} - 1) \left(1 + \frac{1}{H}\right)^{i - i_2^\star} + 1 \geq \left(1 + \frac{1}{H}\right)^{i - i_2^\star} + 1, \quad \forall i \geq i_2^\star.$$

Therefore, $i_h^\star \leq \inf_i\{i \geq i_2^\star : (1 + 1/H)^{i - i_2^\star} + 1 \geq h\} = i_2^\star + \mathcal{O}(H \ln(h))$. Also, by inspecting e_i for small i we observe that $i_2^\star = \mathcal{O}(H)$, which implies that $i_h^\star = \mathcal{O}(H \ln(h))$. □

Remark 2. *Lemma 16 implies that there are at most $\mathcal{O}(\min\{SAH \ln T, ST\})$ updates in T steps.*

Remark 3. *Note that the update scheme in [Zhang et al., 2020b] (also used in Algorithm 2) induces a constant cost of order $\tilde{\mathcal{O}}(B_\star HSA)$, which ruins the horizon free regret. This is because their update scheme collects H samples before the first update. On the contrary, our update scheme updates frequently at the beginning, but has the same update frequency as that of [Zhang et al., 2020b] in the long run. This reduces the constant cost to $\tilde{\mathcal{O}}(B_\star SA)$ while maintaining the $\tilde{\mathcal{O}}(SAH)$ time complexity.*

The following lemma quantifies the dominating bias introduced by the sparse update.

Lemma 17 (bias of the update scheme)**.** $\sum_{t=1}^T P_t(V_t - V_{l_t}) \leq B_\star SA + \frac{1}{H} \sum_{t=1}^T P_t(V^\star - V_t)$ *and* $\sum_{t=1}^T \mathbb{V}(P_t, V_t - V_{l_t}) \leq \tilde{\mathcal{O}}\left(B_\star^2 SA\right) + \frac{B_\star}{H} \sum_{t=1}^T P_t(V^\star - V_t)$.

Proof. For the first statement, we apply Lemma 16 and $P_t = P_{l_t}$ to obtain

$$\sum_{t=1}^{T} P_t(V_t - V_{l_t}) = \sum_{t=1}^{T} P_{l_t}(V^\star - V_{l_t}) - \sum_{t=1}^{T} P_t(V^\star - V_t) \leq B_\star SA + \frac{1}{H} \sum_{t=1}^{T} P_t(V^\star - V_t).$$

Similarly, for the second statement

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{V}(P_t, V_t - V_{l_t}) &\leq \sum_{t=1}^{T} P_t(V_t - V_{l_t})^2 \leq B_\star \sum_{t=1}^{T} P_t(V_t - V_{l_t}) \\
&\leq B_\star^2 SA + \frac{B_\star}{H} \sum_{t=1}^{T} P_t(V^\star - V_t).
\end{aligned}$$

□

E.2 Proofs of Required Properties

In this section, we prove Property 1 (Lemma 18) and Property 2 of Algorithm 3, where Lemma 19 proves a preliminary form of Property 2.

Lemma 18. *With probability at least* $1-\delta$, $Q_t(s,a) \leq Q_{t+1}(s,a) \leq Q^\star(s,a)$, *for any* $(s,a) \in \mathcal{S}\times\mathcal{A}, t \geq 1$.

Proof. The first inequality is clearly true by the update rule. Next, we prove $Q_t(s,a) \leq Q^\star(s,a)$. By Eq. (19), it is clearly true when $n_t(s,a)=0$. When $n_t(s,a)>0$, by Lemma 31: (here, l_t, ι_t is a shorthand of $l_t(s,a), \iota_t(s,a)$):

$$\widehat{c}_t(s,a) + \bar{P}_{t,s,a}V_{l_t} - b_t(s,a) = \widehat{c}_t(s,a) + f(\bar{P}_{t,s,a}, V_{l_t}, n_t(s,a), B, \iota_t) - \sqrt{\frac{\widehat{c}_t(s,a)\iota_t}{n_t(s,a)}}$$

$$\leq c(s,a) + f(\bar{P}_{t,s,a}, V^\star, n_t(s,a), B, \iota_t) + \frac{\iota_t}{n_t(s,a)} \qquad \text{(Eq. (20))}$$

$$= c(s,a) + \bar{P}_{t,s,a}V^\star - \max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{t,s,a},V^\star)\iota_t}{n_t(s,a)}}, \frac{49B\iota_t}{n_t(s,a)}\right\} + \frac{\iota_t}{n_t(s,a)}$$

$$\leq Q^\star(s,a) + (\bar{P}_{t,s,a} - P_{s,a})V^\star - 3\sqrt{\frac{\mathbb{V}(\bar{P}_{t,s,a},V^\star)\iota_t}{n_t(s,a)}} - \frac{24B\iota_t}{n_t(s,a)} + \frac{B\iota_t}{n_t(s,a)}$$

$$\left(B \geq B_\star \geq 1, Q^\star(s,a) = c(s,a) + P_{s,a}V^\star \text{ and } \max\{a,b\} \geq \tfrac{a+b}{2}\right)$$

$$\leq Q^\star(s,a) + (2\sqrt{2}-3)\sqrt{\frac{\mathbb{V}(\bar{P}_{t,s,a},V^\star)\iota_t}{n_t(s,a)}} + (20-24)\frac{B\iota_t}{n_t(s,a)} \leq Q^\star(s,a). \qquad \text{(Lemma 34)}$$

□

Lemma 19. *With probability at least* $1-9\delta$, *for all* $(\mathring{Q},\mathring{V}) \in \mathcal{V}_H$

$$\sum_{t=1}^T (\mathring{Q}(s_t,a_t) - Q_t(s_t,a_t))_+ \leq \left(1+\frac{1}{H}\right)\sum_{t=1}^T(\mathring{V}(s_t) - V_t(s_t))_+$$

$$+ \tilde{O}\left(\sqrt{B_\star SAC_K} + BS^2A + \sqrt{\frac{B_\star S^2A}{H}\sum_{t=1}^T V^\star(s_t) - V_t(s_t)}\right).$$

Proof. We first prove useful properties related to the cost estimator. For a fixed (s,a), by Lemma 34, with probability at least $1-\frac{\delta}{SA}$, when $n_t(s,a)>0$:

$$|c(s,a) - \widehat{c}_t(s,a)| \leq 2\sqrt{\frac{2\widehat{c}_t(s,a)}{n_t(s,a)}\ln\frac{2SA}{\delta}} + \frac{19\ln\frac{2SA}{\delta}}{n_t(s,a)} \leq \sqrt{\frac{\widehat{c}_t(s,a)\iota_t}{n_t(s,a)}} + \frac{\iota_t}{n_t(s,a)}. \tag{20}$$

Taking a union bound, we have Eq. (20) holds for all (s,a) when $n_t(s,a)>0$ with probability at least $1-\delta$. Then by definition of b_t, we have

$$c(s_t,a_t) - \widehat{c}_t(s_t,a_t) \leq \mathbb{I}\{n_t = 0\} + b_t. \tag{21}$$

Note that with probability at least $1-2\delta$, for all $(\mathring{Q},\mathring{V}) \in \mathcal{V}_H$,

$$\sum_{t=1}^T(\mathring{Q}(s_t,a_t) - Q_t(s_t,a_t))_+ \leq \sum_{t=1}^T (c(s_t,a_t) - \widehat{c}_t(s_t,a_t) + P_t\mathring{V} - \bar{P}_tV_{l_t})_+ + b_t$$

$$(\mathring{Q}(s_t,a_t) = c(s_t,a_t) + P_t\mathring{V} \text{ and Eq. (19)})$$

$$\leq \sum_{t=1}^T \mathbb{I}\{n_t=0\} + \sum_{t=1}^T\left[(P_t(\mathring{V} - V_{l_t}) + (P_t - \bar{P}_t)V^\star + (P_t - \bar{P}_t)(V_{l_t} - V^\star))_+ + 2b_t\right]$$

$$\leq SA + \sum_{t=1}^T\left[P_t(\mathring{V} - V_{l_t})_+ + \tilde{O}\left(\sqrt{\frac{\mathbb{V}(P_t,V^\star)}{n_t^+}} + \sqrt{\frac{S\mathbb{V}(P_t,V^\star - V_{l_t})}{n_t^+}} + \frac{SB_\star}{n_t^+}\right) + 2b_t\right].$$

$$((x+y)_+ \leq (x)_+ + (y)_+, \text{ Lemma 34, and Lemma 23})$$

Note that:

$$\begin{aligned}
\sum_{t=1}^{T} P_t(\mathring{V} - V_{l_t})_+ &\le \left(1 + \frac{1}{H}\right)\sum_{t=1}^{T} P_t(\mathring{V} - V_t)_+ + B_\star SA \qquad (P_{l_t} = P_t \text{ and Lemma 16})\\
&= B_\star SA + \left(1 + \frac{1}{H}\right)\sum_{t=1}^{T}\left((\mathring{V}(s'_t) - V_t(s'_t))_+ + (P_t - \mathbb{I}_{s'_t})(\mathring{V} - V_t)_+\right)\\
&\le \mathcal{O}\left(B_\star SA\right) + \left(1 + \frac{1}{H}\right)\sum_{t=1}^{T}\left((\mathring{V}(s_t) - V_t(s_t))_+ + (P_t - \mathbb{I}_{s'_t})(\mathring{V} - V_t)_+\right).\\
&\qquad (\text{Lemma 28 and } (\mathring{V}(s'_t) - V_{t+1}(s'_t))_+ \le (\mathring{V}(s_{t+1}) - V_{t+1}(s_{t+1}))_+)
\end{aligned}$$

Plugging this back to the previous inequality, and by Cauchy-Schwarz inequality and Lemma 24:

$$\begin{aligned}
\sum_{t=1}^{T}(\mathring{Q}(s_t, a_t) - Q_t(s_t, a_t))_+ &\le \left(1 + \frac{1}{H}\right)\sum_{t=1}^{T}\left((\mathring{V}(s_t) - V_t(s_t))_+ + (P_t - \mathbb{I}_{s'_t})(\mathring{V} - V_t)_+ + b_t\right)\\
&\quad + \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t}) + B_\star S^2 A}\right).
\end{aligned}$$

Next, we bound the term $\sum_{t=1}^{T}(P_t - \mathbb{I}_{s'_t})(\mathring{V} - V_t)_+$. We condition on Lemma 20, which holds with probability at least $1 - \delta$. Then, for a given $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$, by Lemma 22 with $X_t = (\mathring{V} - V_t)_+/B_\star$, we have with probability $1 - \frac{\delta}{H+1}$ (F_T, Y_T, and ζ_T are defined in Lemma 22):

$$\begin{aligned}
B_\star F_T(0) = \sum_{t=1}^{T}(P_t - \mathbb{I}_{s'_t})(\mathring{V} - V_t)_+ &\le B_\star(\sqrt{3Y_T\zeta_T} + 4\zeta_T) = \tilde{\mathcal{O}}\left(\sqrt{B_\star^2 Y_T} + B_\star\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star^2\left(S + 1 + \sum_{t=1}^{T}(X_t(s_t) - P_tX_t)_+\right)} + B_\star\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star^2 S + B_\star\sum_{t=1}^{T}(\mathring{V}(s_t) - V_t(s_t) - P_t(\mathring{V} - V_t))_+} + B_\star\right).\\
&\qquad ((x)_+ - (y)_+ \le (x - y)_+)\\
&\overset{\text{(i)}}{=} \tilde{\mathcal{O}}\left(\sum_{t=1}^{T} b_t + B_\star S\sqrt{A} + \sqrt{\frac{B_\star}{H}\sum_{t=1}^{T} P_t(V^\star - V_t)}\right)\\
&\quad + \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t})}\right),
\end{aligned}$$

where in (i) we apply:

$$\begin{aligned}
\sqrt{B_\star\sum_{t=1}^{T}(\mathring{V}(s_t) - V_t(s_t) - P_t(\mathring{V} - V_t))_+} &\le \sqrt{B_\star\left(\sum_{t=1}^{T} 2b_t + \frac{P_t(V^\star - V_t)}{H}\right)}\\
&\quad + \tilde{\mathcal{O}}\left(\sqrt{B_\star\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t})}\right) + B_\star S\sqrt{A}}\right)\\
&\qquad (\text{Lemma 20 and } \sqrt{x + y} \le \sqrt{x} + \sqrt{y})\\
&\le 2\sum_{t=1}^{T} b_t + \sqrt{\frac{B_\star}{H}\sum_{t=1}^{T} P_t(V^\star - V_t)} + \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t})} + B_\star S\sqrt{A}\right).\\
&\qquad (\text{AM-GM inequality and } \sqrt{x + y} \le \sqrt{x} + \sqrt{y})
\end{aligned}$$

Hence, by a union bound, the bound above for $\sum_{t=1}^{T}(P_t - \mathbb{I}_{s'_t})(\mathring{V} - V_t)_+$ holds for all $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$ with probability at least $1-\delta$, and with probability at least $1-4\delta$, for all $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$,

$$
\begin{aligned}
\sum_{t=1}^{T}(\mathring{Q}(s_t,a_t) - Q_t(s_t,a_t))_+ &\le \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}(\mathring{V}(s_t) - V_t(s_t))_+ + \tilde{\mathcal{O}}\left(B_\star S^2 A + \sum_{t=1}^{T} b_t\right)\\
&\quad + \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{S^2 A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t})} + \sqrt{\frac{B_\star}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)}\right)\\
&\le \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}(\mathring{V}(s_t) - V_t(s_t))_+ + \tilde{\mathcal{O}}\left(BS^2A + \sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)}\right)\\
&\quad + \tilde{\mathcal{O}}\left(\sqrt{S^2 A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t})} + \sqrt{\frac{B_\star SA}{H}\sum_{t=1}^{T}P_t(V^\star - V_t) + \sqrt{SAC_K}}\right). \\
&\hspace{20em}\text{(Lemma 21)}
\end{aligned}
$$

Note that:

$$
\begin{aligned}
&\sqrt{S^2 A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t})}\\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star S^2 A\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + B^2S^4A^2 + \frac{B_\star S^2 A}{H}\sum_{t=1}^{T}P_t(V^\star - V_t) + B_\star S^2 A\sqrt{SAC_K}}\right)\\
&\hspace{28em}\text{(Lemma 21)}\\
&= \tilde{\mathcal{O}}\left(\sqrt{B_\star S^2 A\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)}} + BS^2A + \sqrt{\frac{B_\star S^2 A}{H}\sum_{t=1}^{T}P_t(V^\star - V_t) + \sqrt{SAC_K}}\right)\\
&\hspace{18em}(\sqrt{x+y} \le \sqrt{x} + \sqrt{y} \text{ and AM-GM inequality})\\
&= \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + BS^2A + \sqrt{\frac{B_\star S^2 A}{H}\sum_{t=1}^{T}P_t(V^\star - V_t) + \sqrt{SAC_K}}\right).\\
&\hspace{26em}\text{(AM-GM inequality)}
\end{aligned}
$$

Plug this back to the previous inequality, and then by Lemma 5

$$
\begin{aligned}
\sum_{t=1}^{T}(\mathring{Q}(s_t,a_t) - Q_t(s_t,a_t))_+ &\le \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}(\mathring{V}(s_t) - V_t(s_t))_+\\
&\quad + \tilde{\mathcal{O}}\left(\sqrt{B_\star SAC_K} + BS^2A + \sqrt{\frac{B_\star S^2 A}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)}\right).
\end{aligned}
$$

Finally, applying Lemma 36, Lemma 28 and $(V^\star - V_{t+1})(s'_t) \le (V^\star - V_{t+1})(s_{t+1})$, the claim is proved by

$$
\sum_{t=1}^{T}P_t(V^\star - V_t) \le \tilde{\mathcal{O}}(B_\star) + 2\sum_{t=1}^{T}(V^\star(s'_t) - V_t(s'_t)) \le \tilde{\mathcal{O}}(SB_\star) + 2\sum_{t=1}^{T}(V^\star(s_t) - V_t(s_t)).
$$

□

Proof of Theorem 4. Property 1 is proved in Lemma 18. For Property 2, by Lemma 19, it suffices to bound $\sum_{t=1}^{T} V^\star(s_t) - V_t(s_t)$. By Lemma 19, $V^\star_{h-1}(s_t) \le Q^\star_h(s_t,a_t)$, and $V_t(s_t) = Q_t(s_t,a_t)$, we have with probability at least $1-9\delta$, for all $\mathring{Q} = Q^\star_h, \mathring{V} = V^\star_{h-1}, h \in [H]$:

$$\sum_{t=1}^{T}(Q^\star_h(s_t,a_t) - Q_t(s_t,a_t))_+ \leq \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}(Q^\star_{h-1}(s_t,a_t) - Q_t(s_t,a_t))_+ + \tilde{O}\left(\sqrt{B_\star SAC_K} + BS^2A + \sqrt{\frac{B_\star S^2 A}{H}\sum_{t=1}^{T}V^\star(s_t) - V_t(s_t)}\right), \quad \forall h \in [H].$$

Applying the inequality above recursively starting from $h = H$ and by $Q^\star_0(s,a) = 0$, $(1+\frac{1}{H})^H \leq 3$ we have:

$$\sum_{t=1}^{T}(Q^\star_H(s_t,a_t) - Q_t(s_t,a_t))_+ = \tilde{O}\left(H\left(\sqrt{B_\star SAC_K} + BS^2A\right) + \sqrt{B_\star HS^2A\sum_{t=1}^{T}V^\star(s_t) - V_t(s_t)}\right).$$

Then by Lemma 1 with $H = \lceil \frac{4B}{c_{\min}}\ln(\frac{2}{\beta}) + 1\rceil_2$:

$$\begin{aligned}\sum_{t=1}^{T}V^\star(s_t) - V_t(s_t) &\leq \sum_{t=1}^{T}(Q^\star(s_t,a_t) - Q^\star_H(s_t,a_t)) + \sum_{t=1}^{T}(Q^\star_H(s_t,a_t) - Q_t(s_t,a_t)) \\ &\leq B_\star\beta T + \tilde{O}\left(H\left(\sqrt{B_\star SAC_K} + BS^2A\right) + \sqrt{BHS^2A\sum_{t=1}^{T}V^\star(s_t) - V_t(s_t)}\right).\end{aligned}$$

Solving a quadratic equation w.r.t $\sum_{t=1}^{T}V^\star(s_t) - V_t(s_t)$ (Lemma 25), we have:

$$\sum_{t=1}^{T}V^\star(s_t) - V_t(s_t) \leq B_\star\beta T + \tilde{O}\left(H\left(\sqrt{B_\star SAC_K} + BS^2A\right)\right).$$

Plug this back to the bound of Lemma 19 and by AM-GM inequality, we have for all $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$:

$$\begin{aligned}&\sum_{t=1}^{T}(\mathring{Q}(s_t,a_t) - Q_t(s_t,a_t))_+ \\ &\leq \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}(\mathring{V}(s_t) - V_t(s_t))_+ + \frac{B_\star\beta T}{H} + \tilde{O}\left(\sqrt{B_\star SAC_K} + BS^2A\right).\end{aligned}$$

Moreover, by $H \geq \frac{B_\star}{c_{\min}}$, we have $\frac{B_\star\beta T}{H} \leq \beta c_{\min}T \leq \beta C_K$. Hence, Property 2 is satisfied with $d = 1, \xi_H = \beta C_K + \tilde{O}(\sqrt{B_\star SAC_K} + BS^2A)$ with probability at least $1-9\delta$. ☐

E.3 Proof of Theorem 5

Proof. By Theorem 1 and Theorem 4, with probability at least $1-12\delta$:

$$C_K - KV^\star(s_{\text{init}}) = R_K \leq \beta C_K + \tilde{O}\left(\sqrt{B_\star SAC_K} + BS^2A\right).$$

Then by $V^\star(s_{\text{init}}) \leq B_\star, \beta \leq \frac{1}{2}$ and Lemma 25, we have $C_K = \tilde{O}\left(B_\star K\right)$. Substituting this back and by $\beta \leq \frac{c_{\min}}{B_\star K}, H = \tilde{O}(B_\star/c_{\min})$, we get $R_K = \tilde{O}\left(B_\star\sqrt{SAK} + BS^2A\right)$. ☐

E.4 Extra Lemmas for Section 5

In this section, we give full proofs of auxiliary lemmas used in Section 5. Notably, Lemma 20 and Lemma 21 bound the additional terms appears in the recursion in Lemma 19. Lemma 22 gives recursion-based analysis on bounding the sum of martingale difference sequence, which is the key in obtaining horizon-free regret.

Lemma 20. *With probability at least* $1-\delta$*, we have for all* $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$*,*

$$\sum_{t=1}^{T}((\mathbb{I}_{s_t} - P_t)(\mathring{V} - V_t))_+ \leq \sum_{t=1}^{T} 2b_t + \frac{P_t(V^\star - V_t)}{H}$$
$$+ \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t}) + B_\star S^2 A}\right).$$

Proof. With probability at least $1-\delta$, for all $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$,

$$\sum_{t=1}^{T}(\mathring{V}(s_t) - V_t(s_t) - P_t(\mathring{V} - V_t))_+ \leq \sum_{t=1}^{T}(\mathring{Q}(s_t, a_t) - P_t\mathring{V} + P_tV_t - V_t(s_t))_+$$
$$\leq \sum_{t=1}^{T}(c(s_t, a_t) + P_tV_{l_t} - V_t(s_t))_+ + P_t(V_t - V_{l_t})$$

($\mathring{Q}(s_t, a_t) = c(s_t, a_t) + P_t\mathring{V}$, $(x+y)_+ \leq (x)_+ + (y)_+$, and V_t is increasing in t)

$$\leq B_\star SA + \sum_{t=1}^{T}(c(s_t, a_t) - \widehat{c}_t(s_t, a_t))_+ + ((P_t - \bar{P}_t)V_{l_t})_+ + b_t + \frac{1}{H}P_t(V^\star - V_t)$$

($V_t(s_t) = Q_t(s_t, a_t)$, Eq. (19), and Lemma 17)

$$\leq 2B_\star SA + \sum_{t=1}^{T}((P_t - \bar{P}_t)V^\star + (P_t - \bar{P}_t)(V_{l_t} - V^\star))_+ + 2b_t + \frac{1}{H}P_t(V^\star - V_t). \quad \text{(Eq. (21))}$$

Now by Lemma 34 and Lemma 23, we have with probability at least $1-\delta$: $(P_t - \bar{P}_t)V^\star = \mathcal{O}\left(\sqrt{\frac{\mathbb{V}(P_t, V^\star)}{n_t^+}} + \frac{B_\star}{n_t^+}\right)$ and $(P_t - \bar{P}_t)(V_{l_t} - V^\star) = \tilde{\mathcal{O}}\left(\sqrt{\frac{S\mathbb{V}(P_t, V^\star - V_{l_t})}{n_t^+}} + \frac{SB_\star}{n_t^+}\right)$. Plugging these back to the previous inequality, we have for all $(\mathring{Q}, \mathring{V}) \in \mathcal{V}_H$:

$$\sum_{t=1}^{T}(\mathring{V}(s_t) - V_t(s_t) - P_t(\mathring{V} - V_t))_+$$
$$\leq 2B_\star SA + \sum_{t=1}^{T}\tilde{\mathcal{O}}\left(\sqrt{\frac{\mathbb{V}(P_t, V^\star)}{n_t^+}} + \sqrt{\frac{S\mathbb{V}(P_t, V^\star - V_{l_t})}{n_t^+}} + \frac{SB_\star}{n_t^+}\right) + 2b_t + \frac{1}{H}P_t(V^\star - V_t)$$
$$\leq \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t}) + B_\star S^2A}\right) + \sum_{t=1}^{T}2b_t + \frac{P_t(V^\star - V_t)}{H}.$$

(Cauchy-Schwarz inequality and Lemma 24)

This completes the proof. ☐

Lemma 21. *With probability at least* $1 - 3\delta$*,*

$$\sum_{t=1}^{T} b_t = \tilde{\mathcal{O}}\left(BS^{3/2}A + \sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{\frac{B_\star SA}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)} + \sqrt{SAC_K}\right),$$
$$\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t}) = \tilde{\mathcal{O}}\left(B_\star\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star) + B^2S^2A + \frac{B_\star}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)} + B_\star\sqrt{SAC_K}\right).$$

Proof. First note that:

$$\sum_{t=1}^{T} b_t \overset{\text{(i)}}{=} \tilde{\mathcal{O}}\left(BSA + \sum_{t=1}^{T}\sqrt{\frac{\mathbb{V}(\bar{P}_t, V_{l_t})}{n_t^+}} + \sqrt{\frac{\widehat{c}_t}{n_t^+}}\right) \overset{\text{(ii)}}{=} \tilde{\mathcal{O}}\left(BSA + \sum_{t=1}^{T}\sqrt{\frac{\mathbb{V}(P_t, V_{l_t})}{n_t^+}} + \frac{B_\star\sqrt{S}}{n_t^+} + \sqrt{\frac{\widehat{c}_t}{n_t^+}}\right).$$

where in (i) we apply $\max\{a,b\} \le a+b$ and Lemma 24, and in (ii) we have with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{V}(\bar{P}_t, V_{l_t}) = \bar{P}_t(V_{l_t} - \bar{P}_t V_{l_t})^2 \le \bar{P}_t(V_{l_t} - P_t V_{l_t})^2 \qquad \left(\tfrac{\sum_i p_i x_i}{\sum_i p_i} = \operatorname{argmin}_z \textstyle\sum_i p_i (x_i - z)^2\right)\\
&= \mathbb{V}(P_t, V_{l_t}) + (P_t - \bar{P}_t)(V_{l_t} - P_t V_{l_t})^2\\
&\le \mathbb{V}(P_t, V_{l_t}) + \tilde{O}\left(\sum_{s'}\left(\sqrt{\frac{P_t(s')}{n_t^+}} + \frac{1}{n_t^+}\right)(V_{l_t}(s') - P_t V_{l_t})^2\right) \qquad \text{(Lemma 34)}\\
&\le \mathbb{V}(P_t, V_{l_t}) + \tilde{O}\left(B_\star\sqrt{\frac{S\mathbb{V}(P_t, V_{l_t})}{n_t^+}} + \frac{SB_\star^2}{n_t^+}\right) = \tilde{O}\left(\mathbb{V}(P_t, V_{l_t}) + \frac{SB_\star^2}{n_t^+}\right).\\
&\qquad\qquad \text{(Cauchy-Schwarz inequality and AM-GM inequality)}
\end{aligned}$$

Thus, by Lemma 29, Cauchy-Schwarz inequality, and Lemma 24, we have:

$$\begin{aligned}
\sum_{t=1}^T b_t &= \tilde{O}\left(BS^{3/2}A + \sum_{t=1}^T\sqrt{\frac{\mathbb{V}(P_t, V^\star)}{n_t^+}} + \sum_{t=1}^T\sqrt{\frac{\mathbb{V}(P_t, V^\star - V_{l_t})}{n_t^+}} + \sqrt{\frac{\widehat{c}_t}{n_t^+}}\right)\\
&= \tilde{O}\left(BS^{3/2}A + \sqrt{SA\sum_{t=1}^T\mathbb{V}(P_t, V^\star)} + \sqrt{SA\sum_{t=1}^T\mathbb{V}(P_t, V^\star - V_{l_t})} + \sqrt{SAC_K}\right),
\end{aligned} \tag{22}$$

where in the last inequality we apply:

$$\begin{aligned}
&\sum_{t=1}^T\sqrt{\frac{\widehat{c}_t}{n_t^+}} \le \sqrt{SA\left(\sum_{t=1}^T c(s_t,a_t) + \sum_{t=1}^T(c(s_t,a_t) - \widehat{c}_t)\right)}\\
&\qquad\qquad \text{(Cauchy-Schwarz inequality and Lemma 24)}\\
&\le \sqrt{SA\left(2C_K + \tilde{O}(1) + \sum_{t=1}^T\sqrt{\frac{\widehat{c}_t\iota_t}{n_t^+}} + \frac{\iota_t}{n_t^+}\right)} = \tilde{O}\left(\sqrt{SAC_K} + \sqrt{SA\sum_{t=1}^T\sqrt{\frac{\widehat{c}_t}{n_t^+}} + SA}\right),\\
&\qquad\qquad \text{(Lemma 36 and Eq. (20))}
\end{aligned}$$

and by Lemma 25 we obtain: $\sum_{t=1}^T\sqrt{\frac{\widehat{c}_t}{n_t^+}} = \tilde{O}(\sqrt{SAC_K} + SA)$. Applying Lemma 22 with $X_t(s) = (V^\star(s) - V_t(s))/B_\star$, we have with probability at least $1-\delta$ (G_T, Y_T, and ζ_T are defined in Lemma 22),

$$\sum_{t=1}^T\mathbb{V}(P_t, V^\star - V_t) = B_\star^2 G_T(0) \le 3B_\star^2 Y_T + 9B_\star^2\zeta_T \le 3B_\star\sum_{t=1}^T((\mathbb{I}_{s_t} - P_t)(V^\star - V_t))_+ + \tilde{O}\left(SB_\star^2\right).$$

By Lemma 20 and Eq. (22), with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{t=1}^T((\mathbb{I}_{s_t} - P_t)(V^\star - V_t))_+ \le \sum_{t=1}^T 2b_t + \frac{1}{H}P_t(V^\star - V_t)\\
&\quad + \tilde{O}\left(\sqrt{SA\sum_{t=1}^T\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^T\mathbb{V}(P_t, V^\star - V_{l_t}) + B_\star S^2 A}\right).\\
&= \tilde{O}\left(BS^2A + \sqrt{SA\sum_{t=1}^T\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^T\mathbb{V}(P_t, V^\star - V_{l_t})} + \frac{1}{H}\sum_{t=1}^T P_t(V^\star - V_t) + \sqrt{SAC_K}\right)\\
&\overset{\text{(i)}}{=} \tilde{O}\left(BS^2A + \sqrt{SA\sum_{t=1}^T\mathbb{V}(P_t, V^\star)} + \sqrt{S^2A\sum_{t=1}^T\mathbb{V}(P_t, V^\star - V_t)} + \frac{1}{H}\sum_{t=1}^T P_t(V^\star - V_t) + \sqrt{SAC_K}\right),
\end{aligned}$$

where in (i) we apply

$$\sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t})} = \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_t)} + \sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V_t - V_{l_t})}\right)$$

$$(\textsc{Var}[X+Y] \le 2\textsc{Var}[X] + 2\textsc{Var}[Y] \text{ and } \sqrt{x+y} \le \sqrt{x} + \sqrt{y})$$

$$= \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_t)} + \sqrt{S^2A\left(B_\star^2SA + \frac{B_\star}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)\right)}\right) \quad \text{(Lemma 17)}$$

$$= \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_t)} + B_\star S^2A + \frac{1}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)\right).$$

$$(\sqrt{x+y} \le \sqrt{x} + \sqrt{y} \text{ and AM-GM Inequality})$$

Plugging the bound on $\sum_{t=1}^{T}((\mathbb{I}_{s_t} - P_t)(V^\star - V_t))_+$ back, we have

$$\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_t) = \tilde{\mathcal{O}}\left(B^2S^2A + B_\star\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + B_\star\sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_t)}\right)$$
$$+ \tilde{\mathcal{O}}\left(\frac{B_\star}{H}\sum_{t=1}^{T}P_t(V^\star - V_t) + B_\star\sqrt{SAC_K}\right).$$

Solving a quadratic inequality w.r.t $\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_t)$ (Lemma 25), we obtain

$$\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_t) = \tilde{\mathcal{O}}\left(B^2S^2A + B_\star\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \frac{B_\star}{H}\sum_{t=1}^{T}P_t(V^\star - V_t) + B_\star\sqrt{SAC_K}\right),$$

and by $\textsc{Var}[X+Y] \le 2\textsc{Var}[X] + 2\textsc{Var}[Y]$ and Lemma 17,

$$\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t}) = \tilde{\mathcal{O}}\left(\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_t) + \mathbb{V}(P_t, V_t - V_{l_t})\right)$$
$$= \tilde{\mathcal{O}}\left(B_\star\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + B^2S^2A + \frac{B_\star}{H}\sum_{t=1}^{T}P_t(V^\star - V_t) + B_\star\sqrt{SAC_K}\right).$$

Moreover, by $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and AM-GM inequality:

$$\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star - V_{l_t})}$$
$$= \tilde{\mathcal{O}}\left(\sqrt{B_\star SA\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)}} + BS^{3/2}A + \sqrt{\frac{B_\star SA}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)} + \sqrt{B_\star SA\sqrt{SAC_K}}\right)$$
$$= \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + BS^{3/2}A + \sqrt{\frac{B_\star SA}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)} + \sqrt{SAC_K}\right).$$

Plug this back to Eq. (22):

$$\sum_{t=1}^{T}b_t = \tilde{\mathcal{O}}\left(BS^{3/2}A + \sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + \sqrt{\frac{B_\star SA}{H}\sum_{t=1}^{T}P_t(V^\star - V_t)} + \sqrt{SAC_K}\right).$$

□

Lemma 22. *Suppose $X_t : \mathcal{S}^+ \to [0,1]$ is monotonic in t (that is, $X_t(s)$ is non-decreasing or non-increasing in t for all $s \in \mathcal{S}^+$), and $X_t(g) = 0$. Define:*

$$F_n(d) = \sum_{t=1}^{n} P_t X_t^{2^d} - (X_t(s'_t))^{2^d}, \quad G_n(d) = \sum_{t=1}^{n} \mathbb{V}(P_t, X_t^{2^d}).$$

Then with probability at least $1-\delta$, for all $n \in \mathbb{N}^+$ simultaneously, $G_n(0) \le 3Y_n + 9\zeta_n$, $F_n(0) \le \sqrt{3Y_n\zeta_n} + 4\zeta_n$, where $Y_n = S + 1 + \sum_{t=1}^{n}(X_t(s_t) - P_tX_t)_+$, $\zeta_n = 32\ln^3\frac{4n^4}{\delta}$.

Proof. Note that:

$$G_n(d) = \sum_{t=1}^{n} P_t X_t^{2^{d+1}} - (P_t X_t^{2^d})^2 \le \sum_{t=1}^{n} P_t X_t^{2^{d+1}} - (P_tX_t)^{2^{d+1}} \qquad (x^p \text{ is convex for } p > 1)$$

$$= \sum_{t=1}^{n} P_t X_t^{2^{d+1}} - X_t(s'_t)^{2^{d+1}} + \sum_{t=1}^{n} X_t(s'_t)^{2^{d+1}} - X_t(s_t)^{2^{d+1}} + \sum_{t=1}^{n} X_t(s_t)^{2^{d+1}} - (P_tX_t)^{2^{d+1}}$$

$$\overset{\text{(i)}}{\le} F_n(d+1) + S + 1 + 2^{d+1}(X_t(s_t) - P_tX_t)_+ \le F(d+1) + 2^{d+1}Y_n,$$

where in (i) we apply Lemma 26 and,

$$\sum_{t=1}^{n} X_t(s'_t)^{2^{d+1}} - X_t(s_t)^{2^{d+1}} = \sum_{t=1}^{n} X_t(s'_t)^{2^{d+1}} - X_{t+1}(s'_t)^{2^{d+1}} + \sum_{t=1}^{n} X_{t+1}(s'_t)^{2^{d+1}} - X_t(s_t)^{2^{d+1}}$$

$$\le S + \sum_{t=1}^{n} X_{t+1}(s_{t+1})^{2^{d+1}} - X_t(s_t)^{2^{d+1}} = S + X_{n+1}(s_{n+1})^{2^{d+1}} - X_1(s_1)^{2^{d+1}} \le S + 1.$$

(Lemma 28 and $X_{t+1}(s'_t) \le X_{t+1}(s_{t+1})$)

For a fixed d, n, by Eq. (23) of Lemma 35, with probability $1 - \frac{\delta}{2n^2\lceil\log_2 n + 1\rceil}$,

$$F_n(d) \le \sqrt{G_n(d)\zeta_n} + \zeta_n \le \sqrt{(F_n(d+1) + 2^{d+1}Y_n)\zeta_n} + \zeta_n.$$

Taking a union bound on $d = 0, \ldots, \lceil\log_2 n\rceil$, and by Lemma 27 with $\lambda_1 = n, \lambda_2 = \sqrt{\zeta_n}, \lambda_3 = Y_n, \lambda_4 = \zeta_n$, we have:

$$F_n(1) \le \max\{(\sqrt{\zeta_n} + \sqrt{2\zeta_n})^2, \sqrt{8Y_n\zeta_n} + \zeta_n\} \le \max\{6\zeta_n, \sqrt{8Y_n\zeta_n} + \zeta_n\}.$$

Therefore, $G_n(0) \le F_n(1) + 2Y_n \le \max\{6\zeta_n, Y_n + 9\zeta_n\} + 2Y_n \le 3Y_n + 9\zeta_n$, and $F_n(0) \le \sqrt{G_n(0)\zeta_n} + \zeta_n \le \sqrt{3Y_n\zeta_n} + 4\zeta_n$. Taking a union bound over $n \in \mathbb{N}^+$ proves the claim. ☐

Lemma 23. *Given $X_t : \mathcal{S}^+ \to \mathbb{R}$ with $\|X_t\|_\infty \le B$, with probability at least $1-\delta$, it holds that for all $t \ge 1$ simultaneously: $(P_t - \bar{P}_t)X_t = \tilde{\mathcal{O}}\left(\sqrt{\frac{S\mathbb{V}(P_t, X_t)}{n_t^+}} + \frac{SB}{n_t^+}\right)$.*

Proof. For a fixed $(s,a) \in \mathcal{S} \times \mathcal{A}$, by Lemma 34, with probability $1 - \frac{\delta}{SA}$, for any $t \ge 1$ such that $(s_t, a_t) = (s, a)$:

$$(P_t - \bar{P}_t)X_t = \sum_{s'}(P_t(s') - \bar{P}_t(s'))(X_t(s') - P_tX_t) \qquad \left(\textstyle\sum_{s'} P_t(s') - \bar{P}_t(s') = 0\right)$$

$$= \tilde{\mathcal{O}}\left(\sum_{s'}\left(\sqrt{\frac{P_t(s')}{n_t^+}} + \frac{1}{n_t^+}\right)|X_t(s') - P_tX_t|\right) = \tilde{\mathcal{O}}\left(\sqrt{\frac{S\mathbb{V}(P_t, X_t)}{n_t^+}} + \frac{SB}{n_t^+}\right).$$

Taking a union bound over $(s,a) \in \mathcal{S} \times \mathcal{A}$, the statement is proved. ☐

Lemma 24. $\sum_{t=1}^{T} \frac{1}{n_t^+} = \mathcal{O}(SA\ln T)$.

Proof. Define $J_{s,a}$ such that $E_{J_{s,a}} = n_T(s,a)$. It is easy to see that $e_{j+1}/e_j \le 2$. Then,

$$\sum_{t=1}^{T} \frac{1}{n_t^+} \le SA + \sum_{(s,a)}\sum_{j=1}^{J_{s,a}} \frac{e_{j+1}}{E_j} \le SA + 2\sum_{(s,a)}\sum_{j=1}^{J_{s,a}} \frac{e_j}{E_j} = \mathcal{O}(SA\ln T).$$

☐

Algorithm 6 SVI-SSP without knowledge of $B_\star$

```
Parameters: failure probability $\delta \in (0,1)$.
Define: $\mathcal{L} = \{E_j\}_{j\in\mathbb{N}^+}$, where $E_j = \sum_{i=1}^{j} e_i, e_j = \lfloor \widetilde{e}_j \rfloor$, and $\widetilde{e}_1 = 1, \widetilde{e}_{j+1} = \widetilde{e}_j + \frac{1}{H} e_j$.
Initialize: $B \leftarrow \frac{\sqrt{K}}{S^{3/2}A^{1/2}}, H \leftarrow \lceil \frac{4B}{c_{\min}} \ln \frac{4B^2SAK}{c_{\min}} \rceil_2, C \leftarrow 0, t \leftarrow 0, s_1 \leftarrow s_{\text{init}}$.
Initialize: for all $(s,a,s')$, $n(s,a,s') \leftarrow 0, n(s,a) \leftarrow 0$, $Q(s,a) \leftarrow 0$, $V(s) \leftarrow 0$, $\widehat{C}(s,a) \leftarrow 0$.
for $k = 1, \ldots, K$ do
    repeat
        Increment time step $t \overset{+}{\leftarrow} 1$.
        Take action $a_t = \operatorname{argmin}_a Q(s_t, a)$, suffer cost $c_t$, transit to and observe $s'_t$.
        Update visitation counters: $n = n(s_t, a_t) \overset{+}{\leftarrow} 1, n(s_t, a_t, s'_t) \overset{+}{\leftarrow} 1$.
        Update cost accumulator $C \overset{+}{\leftarrow} c_t$, $\widehat{C}(s,a) \leftarrow c_t$.
        if $n \in \mathcal{L}$ then
            Update empirical transition: $\bar{P}_{s_t,a_t}(s') \leftarrow \frac{n(s_t,a_t,s')}{n}$ for all $s'$.
            Compute $\iota \leftarrow \ln \frac{2SAn}{\delta}$, $\widehat{c} \leftarrow \frac{\widehat{C}(s_t,a_t)}{n}$, and bonus $b \leftarrow \max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{s_t,a_t},V)\iota}{n}}, \frac{49B\iota}{n}\right\}$.
            $Q(s_t, a_t) \leftarrow \max\{\widehat{c} + \bar{P}_{s_t,a_t}V - b, Q(s_t, a_t)\}$.
            $V(s_t) \leftarrow \operatorname{argmin}_a Q(s_t, a)$.
        if $\|V\|_\infty > B$ or $C > KB + x(B\sqrt{SAK} + BS^2A)$ then
            $B \leftarrow 2B, H \leftarrow \lceil \frac{4B}{c_{\min}} \ln \frac{4B^2SAK}{c_{\min}} \rceil_2, C \leftarrow 0$, and update $x$.
            $n(s,a,s') \leftarrow 0, n(s,a) \leftarrow 0$, $Q(s,a) \leftarrow 0$, $V(s) \leftarrow 0$, $\widehat{C}(s,a) \leftarrow 0$ for all $(s,a,s')$.
        if $s'_t \neq g$ then $s_{t+1} \leftarrow s'_t$; else $s_{t+1} \leftarrow s_{\text{init}}$, break.
```

E.5 Parameter-free Algorithm

Following [Tarbouriech et al., 2021b], we divide the learning process into epochs indexed by ϕ. We maintain value function upper bound B initialized with $\frac{\sqrt{K}}{S^{3/2}A^{1/2}}$ and cost accumulator C recording the total costs suffered in the current epoch. In epoch ϕ, we execute an instance of Algorithm 3 with value function upper bound B. Moreover, we start a new epoch whenever:

1. $\|V\|_\infty > B$,
2. or $C > KB + x(B\sqrt{SAK} + BS^2A)$.

Here, x is a large enough constant determined by Theorem 5, so that when $B \geq B_\star$, we have with probability at least $1 - 12\delta$:

$$C - V^\star(s^\phi_{\text{init}}) - (K-1)V^\star(s_{\text{init}}) \leq x(B_\star\sqrt{SAK} + BS^2A),$$

where s^ϕ_{init} is the initial state of epoch ϕ (note that Theorem 5 still holds when the initial state is changing over episodes). Moreover, we double the value of B whenever a new epoch starts. We summarize ideas above in Algorithm 6.

Theorem 9. *With probability at least* $1 - 12\delta$*, Algorithm 6 ensures* $R_K = \tilde{O}(B_\star\sqrt{SAK} + B_\star^3 S^3 A)$.

Proof. Denote by B_ϕ the value of B in epoch ϕ, and by C_ϕ the value of C at the end of epoch ϕ. Define $\phi^\star = \inf_\phi\{B_\phi \geq B_\star\}$. Clearly $B_\phi \leq \max\{2B_\star, \sqrt{K}/S^{3/2}A^{1/2}\}$ for $\phi \leq \phi^\star$. By Theorem 5, with probability at least $1 - 12\delta$, there is at most $\phi^\star$ epochs since the condition of starting a new epoch will never be triggered in epoch $\phi^\star$, and the regret in epoch $\phi^\star$ is properly bounded:

$$C_{\phi^\star} - V^\star(s^{\phi^\star}_{\text{init}}) - (K-1)V^\star(s_{\text{init}}) = \tilde{O}\left(B_\star\sqrt{SAK} + B_{\phi^\star}S^2A\right) = \tilde{O}\left(B_\star\sqrt{SAK} + B_\star S^2 A\right).$$

Conditioned on the event that there are at most $\phi^\star$ epochs, we partition the regret into two parts: the total costs suffered before epoch $\phi^\star$, and the regret starting from epoch $\phi^\star$. It suffices to bound the total costs before epoch $\phi^\star$ assuming $K \leq B_\star^2 S^3 A$ (otherwise $\phi^\star = 1$). By the update scheme of

B, we have at most $\lceil \log_2 B_\star \rceil + 1$ epochs before epoch $\phi^\star$. Moreover, by the second condition of starting a new epoch, the accumulated cost in epoch $\phi < \phi^\star$ is bounded by:

$$C_\phi \leq KB_\phi + \tilde{\mathcal{O}}\left(B_\phi\sqrt{SAK} + B_\phi S^2 A\right) = \tilde{\mathcal{O}}\left(B_\star^3 S^3 A\right).$$

Combining these two parts, we get:

$$\begin{aligned} R_K &= \sum_{\phi=1}^{\phi^\star - 1} C_\phi + (C_{\phi^\star} - V^\star(s_{\text{init}}^{\phi^\star}) - (K-1)V^\star(s_{\text{init}})) + (V^\star(s_{\text{init}}^{\phi^\star}) - V^\star(s_{\text{init}})) \\ &= \tilde{\mathcal{O}}\left(B_\star\sqrt{SAK} + B_\star^3 S^3 A\right), \end{aligned}$$

where we assume $C_{\phi^\star} = 0$ and $s_{\text{init}}^{\phi^\star} = s_{\text{init}}$ if there are less than $\phi^\star$ epochs. □

F Auxiliary Lemmas

Lemma 25. *If $x \leq (a\sqrt{x} + b)\ln^p(cx)$ for some $a, b, c > 0$ and absolute constant $p \geq 0$, then $x = \tilde{\mathcal{O}}(a^2 + b)$. Specifically, $x \leq a\sqrt{x} + b$ implies $x \leq (a + \sqrt{b})^2 \leq 2a^2 + 2b$.*

Lemma 26. *For any $a, b \in [0, 1]$ and $k \in \mathbb{N}^+$, we have: $a^k - b^k \leq k(a-b)_+$.*

Proof. $a^k - b^k = (a-b)(\sum_{i=1}^k a^{i-1}b^{k-i}) \leq (a-b)_+ \cdot \sum_{i=1}^k 1 = k(a-b)_+$. □

Lemma 27. *([Zhang et al., 2020a, Lemma 11]) Let $\lambda_1, \lambda_2, \lambda_4 \geq 0, \lambda_3 \geq 1$ and $i' = \log_2(\lambda_1)$. Let $a_1, a_2, \ldots, a_{i'}$ be non-negative reals such that $a_i \leq \lambda_1$ and $a_i \leq \lambda_2\sqrt{a_{i+1} + 2^{i+1}\lambda_3} + \lambda_4$ for any $1 \leq i \leq i'$. Then, $a_1 \leq \max\{(\lambda_2 + \sqrt{\lambda_2^2 + \lambda_4})^2, \lambda_2\sqrt{8\lambda_3} + \lambda_4\}$.*

Lemma 28. *Assume $v_t : \mathcal{S}^+ \to [0, B]$ is monotonic in t (i.e., $v_t(s)$ is non-increasing or non-decreasing in t for any $s \in \mathcal{S}^+$). Then, for any state sequence $\{s_t\}_{t=1}^n, n \in \mathbb{N}^+$, we have: $|\sum_{t=1}^n v_{t+1}(s_t) - v_t(s_t)| \leq SB$.*

Proof.

$$\begin{aligned} \left|\sum_{t=1}^n v_{t+1}(s_t) - v_t(s_t)\right| &\leq \sum_{s\in\mathcal{S}^+}\left|\sum_{t=1}^n (v_{t+1}(s) - v_t(s))\mathbb{I}\{s_t = s\}\right| \\ &\leq \sum_{s\in\mathcal{S}^+}\left|\sum_{t=1}^n v_{t+1}(s) - v_t(s)\right| \leq \sum_{s\in\mathcal{S}^+}|v_{n+1}(s) - v_1(s)| \leq SB. \qquad (v_t(s) \text{ is monotonic in } t) \end{aligned}$$

□

Lemma 29. *([Cohen et al., 2021, Lemma C.3]) For any two random variables X, Y with $\textsc{Var}[X] < \infty, \textsc{Var}[Y] < \infty$. We have: $\sqrt{\textsc{Var}[X]} - \sqrt{\textsc{Var}[Y]} \leq \sqrt{\textsc{Var}[X - Y]}$.*

Lemma 30. *For any two random variables X, Y, we have:*

$$\textsc{Var}[XY] \leq 2\textsc{Var}[X]\left\|Y\right\|_\infty^2 + 2(\mathbb{E}[X])^2\textsc{Var}[Y].$$

Consequently, $\|X\|_\infty \leq C$ implies $\textsc{Var}[X^2] \leq 4C^2\textsc{Var}[X]$.

Proof. First note that for any two random variables U, V, we have $\textsc{Var}[U + V] \leq 2\textsc{Var}[U] + 2\textsc{Var}[V]$. Now let $U = (X - \mathbb{E}[X])Y$ and $V = \mathbb{E}[X]Y$, we have:

$$\begin{aligned} \textsc{Var}[XY] &\leq 2\textsc{Var}[(X - \mathbb{E}[X])Y] + 2\textsc{Var}[\mathbb{E}[X]Y] \leq 2\mathbb{E}[(X - \mathbb{E}[X])^2Y^2] + 2(\mathbb{E}[X])^2\textsc{Var}[Y] \\ &\leq 2\textsc{Var}[X]\left\|Y\right\|_\infty^2 + 2(\mathbb{E}[X])^2\textsc{Var}[Y]. \end{aligned}$$

□

Lemma 31. *([Tarbouriech et al., 2021b, Lemma 14]) Define $\Upsilon = \{v \in [0, B]^{\mathcal{S}^+} : v(g) = 0\}$. Let $f : \Delta_{\mathcal{S}^+} \times \Upsilon \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ with $f(p, v, n, B, \iota) = pv - \max\left\{c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}}, c_2\frac{B\iota}{n}\right\}$, with $c_1 = 7$ and $c_2 = 49$. Then f satisfies for all $p \in \Delta_{\mathcal{S}^+}, v \in \Upsilon$ and $n, \iota > 0$,*

1. $f(p, v, n, B, \iota)$ *is non-decreasing in* $v(s)$*, that is,*
$$\forall v, v' \in \Upsilon, v(s) \leq v'(s), \forall s \in \mathcal{S}^+ \implies f(p, v, n, B, \iota) \leq f(p, v', n, B, \iota);$$
2. $f(p, v, n, B, \iota) \leq pv - \frac{c_1}{2}\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} - \frac{c_2}{2}\frac{B\iota}{n} \leq pv - 3\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} - 24\frac{B\iota}{n}$.

Lemma 32. *([Jaksch et al., 2010, Lemma 19], [Cohen et al., 2020, Lemma B.18]) For any sequence of numbers* $z_1, \ldots, z_n$ *with* $0 \leq z_t \leq Z_{t-1} = \max\{1, \sum_{i=1}^{t-1} z_i\}$*:*
$$\sum_{t=1}^{n} \frac{z_t}{Z_{t-1}} \leq 2\ln Z_n, \quad \sum_{t=1}^{n} \frac{z_t}{\sqrt{Z_{t-1}}} \leq 3\sqrt{Z_n}.$$

G Concentration Inequalities

Lemma 33. *([Cohen et al., 2020, Theorem D.1]) Let* $\{X_t\}_t$ *be a martingale difference sequence such that* $|X_t| \leq B$*. Then with probability at least* $1-\delta$*,*
$$\left|\sum_{t=1}^{n} X_t\right| \leq B\sqrt{n\ln\frac{2n}{\delta}}, \quad \forall n \geq 1.$$

Lemma 34. *Let* $\{X_t\}_t$ *be a sequence of i.i.d random variables with mean* μ*, variance* σ^2*, and* $0 \leq X_t \leq B$*. Then with probability at least* $1-\delta$*, the following holds for all* $n \geq 1$ *simultaneously:*
$$\left|\sum_{t=1}^{n}(X_t - \mu)\right| \leq 2\sqrt{2\sigma^2 n\ln\frac{2n}{\delta}} + 2B\ln\frac{2n}{\delta}.$$
$$\left|\sum_{t=1}^{n}(X_t - \mu)\right| \leq 2\sqrt{2\hat{\sigma}_n^2 n\ln\frac{2n}{\delta}} + 19B\ln\frac{2n}{\delta}.$$
where $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{t=1}^{n} X_t^2 - (\frac{1}{n}\sum_{t=1}^{n} X_t)^2$*.*

Proof. For a fixed n, the first inequality holds with probability at least $1 - \frac{\delta}{4n^2}$ by Freedman's inequality. Then by [Efroni et al., 2021, Lemma 19], with probability at least $1 - \frac{\delta}{4n^2}$, $|\sigma - \hat{\sigma}_n| \leq \sqrt{\frac{36B^2\ln(2n/\delta)}{n^+}}$. Therefore, $\sqrt{n}\sigma = \sqrt{n}\hat{\sigma}_n + \sqrt{n}(\sigma - \hat{\sigma}_n) \leq \sqrt{n}\hat{\sigma}_n + 6B\sqrt{\ln(2n/\delta)}$. Plugging this back to the first inequality gives the second inequality. ☐

Lemma 35. *(Strengthened Freedman's inequality) Let* $X_{1:\infty}$ *be a martingale difference sequence with respect to a filtration* $\{\mathcal{F}_t\}_t$ *such that* $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$*. Suppose* $B_t \in [1, b]$ *for a fixed constant* b*,* $B_t \in \mathcal{F}_{t-1}$ *and* $X_t \leq B_t$ *almost surely. Then for a given* n*, with probability at least* $1-\delta$*:*
$$\left|\sum_{t=1}^{n} X_t\right| \leq C\Big(\sqrt{8V_{1,n}\ln(2C/\delta)} + 5B_{1,n}\ln(2C/\delta)\Big), \tag{23}$$
and with probability at least $1-\delta$ *we have for all* $1 \leq l \leq n$ *simultaneously*
$$\left|\sum_{t=l}^{l+n-1} X_t\right| \leq C\Big(\sqrt{8V_{l,n}\ln(4Cn^3/\delta)} + 5B_{l,n}\ln\left(4Cn^3/\delta\right)\Big) \leq 8CB_{l,n}\sqrt{n}\ln(4Cn^3/\delta), \tag{24}$$
where $V_{l,n} = \sum_{t=l}^{l+n-1}\mathbb{E}[X_t^2|\mathcal{F}_{t-1}]$*,* $B_{l,n} = \max_{l \leq t < l+n} B_t$*, and* $C = \lceil\ln(b)\rceil\lceil\ln(nb^2)\rceil$*.*

Proof. Eq. (23) is simply from applying [Lee et al., 2020, Theorem 2.2] to $\{X_t\}_t$ and $\{-X_t\}_t$. Fix some $l, n \geq 1$. Eq. (24) holds with probability at least $1 - \frac{\delta}{2n^3}$ by Eq. (23). By a union bound (first sum over l, then sum over n), the statement is proved. ☐

Lemma 36. *Given* $\alpha \geq 1$ *and a martingale sequence* $\{X_t\}_t$ *such that* $X_t \in \mathcal{F}_t, 0 \leq X_t \leq B$*, with probability at least* $1-\delta$*:*
$$\sum_{t=1}^{n}\mathbb{E}[X_t|\mathcal{F}_{t-1}] \leq \left(1 + \frac{1}{\alpha}\right)\sum_{t=1}^{n} X_t + 8B\alpha\ln\frac{2n}{\delta}, \quad \forall n \geq 1.$$

Proof. Define $Y_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] - X_t$. For a given n, by Freedman's inequality, with probability at least $1 - \frac{\delta}{2n^2}$:

$$\sum_{t=1}^{n} Y_t \leq \eta \sum_{t=1}^{n} \mathbb{E}[(X_t - \mathbb{E}[X_t|\mathcal{F}_{t-1}])^2|\mathcal{F}_{t-1}] + \frac{2\ln(2n/\delta)}{\eta} \leq B\eta\mathbb{E}[X_t|\mathcal{F}_{t-1}] + \frac{2\ln(2n/\delta)}{\eta},$$

for some $\eta < \frac{1}{B}$. Reorganizng terms, we get when $\eta = \frac{1}{2B\alpha} < \frac{1}{B}$ (note that $B\eta \leq \frac{1}{2}$):

$$\begin{aligned}\sum_{t=1}^{n} \mathbb{E}[X_t|\mathcal{F}_{t-1}] &\leq \frac{1}{1 - B\eta}\left(\sum_{t=1}^{n} X_t + \frac{2\ln(2n/\delta)}{\eta}\right) \leq (1 + 2B\eta)\sum_{t=1}^{n} X_t + \frac{4\ln(2n/\delta)}{\eta} \\ &\leq \left(1 + \frac{1}{\alpha}\right)\sum_{t=1}^{n} X_t + 8B\alpha\ln\frac{2n}{\delta}. \qquad \left(\tfrac{1}{1-x} \leq 1 + 2x \text{ when } x \in [0, \tfrac{1}{2}]\right)\end{aligned}$$

By a union bound over n, we obtain the desired bound. ☐

H Experiments

In this section, we benchmark known SSP algorithms empirically. We consider two environments, RandomMDP and GridWorld. In RandomMDP, there are 5 states and 2 actions, and both transition and cost function are chosen uniformly at random. In GridWorld, there are 12 states (including the goal state) and 4 actions (LEFT, RIGHT, UP, DOWN) forming a 3×4 grid. The agent starts at the upper left corner of the grid, and the goal state is at the lower right corner of the grid. Taking each action initiates an attempt to moves one step towards the indicated direction with probability 0.85, and moves randomly towards the other three directions with probability 0.15. The movement attempt fails if the agent tries to move out of the grid, and in this case the agent stays at the same position. The cost is 1 for each state-action pair. In our experiments, $B_\star \approx 1.5$ and $c_{\min} \approx 0.04$ in RandomMDP, and $B_\star \approx 6$ and $c_{\min} = 1$ in GridWorld.

We implement two model-free algorithms: Q-learning with ϵ-greedy exploration [Yu and Bertsekas, 2013] and LCB-ADVANTAGE-SSP, and five model-based algorithms: UC-SSP [Tarbouriech et al., 2020a][6], Bernstein-SSP [Cohen et al., 2020], ULCVI [Cohen et al., 2021], EB-SSP [Tarbouriech et al., 2021b], and SVI-SSP. For each algorithm, we optimize hyper-parameters for the best possible results. Moreover, instead of incorporating the logarithmic terms from confidence intervals suggested by the theory, we treat it as a hyper-parameter ι and search its best value. The hyper-parameters used in the experiments are shown in Table 4. All experiments are performed in Google Cloud Platform on a compute engine with machine type "e2-medium".

The plot of accumulated regret is shown in Figure 1. Q-learning with ϵ-greedy exploration suffers linear regret, indicating that naive ϵ-greedy exploration is inefficient. UC-SSP and SVI-SSP show competitive results in both environments. SVI-SSP also consistently outperforms EB-SSP, both of which are minimax-optimal and horizon-free.

In Table 3, we also show the time spent in updates (policy, accumulators, etc) in the whole learning process for each algorithm. Our model-based algorithm SVI-SSP spends least time in updates among all algorithms, confirming our theoretical arguments. ULCVI and UC-SSP spend most time in updates, which is reasonable since these two algorithms computes a new policy in each episode, instead of exponentially sparse updates.

---

[6] we implement a variant of UC-SSP with a fixed pivot horizon for a much better empirical performance, where $\gamma_{k,j} = 10^{-6}$ always (see their Algorithm 2 for the definition of $\gamma_{k,j}$)

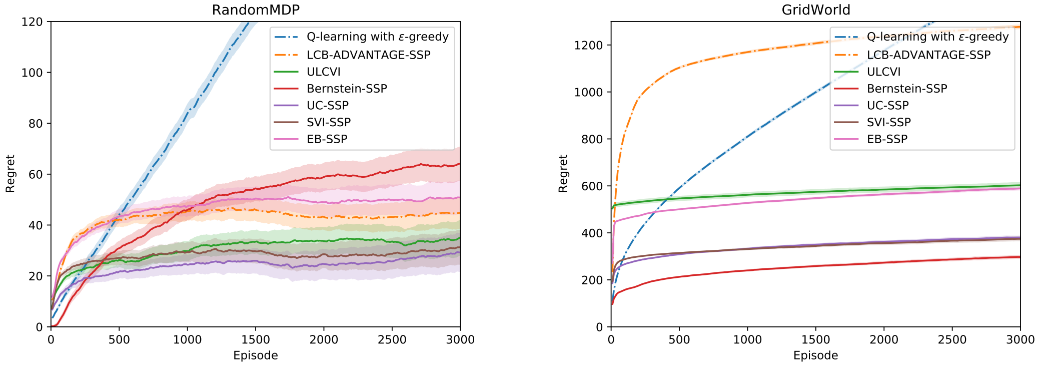

Figure 1: Accumulated regret of each algorithm on RandomMDP (left) and GridWorld (right) in 3000 episodes. Each plot is an average of 500 repeated runs, and the shaded area is 95% confidence interval. Dotted lines represent model-free algorithms and solid lines represent model-based algorithms.

Table 3: Average time (in seconds) spent in updates in 3000 episodes for each algorithm. Our model-based algorithm SVI-SSP is the most efficient algorithm.

	RandomMDP	GridWorld
Q-learning with ϵ-greedy	0.3385	0.3773
LCB-ADVANTAGE-SSP	0.3517	0.3982
UC-SSP	14.4472	8.6886
Bernstein-SSP	0.2918	0.4656
ULCVI	15.7128	22.8062
EB-SSP	0.2319	0.4619
SVI-SSP	**0.1207**	**0.1419**

Table 4: Hyper-parameters used in the experiments. We search the best parameters for each algorithm.

	Algorithm	Parameters
RandomMDP	Q-learning with ϵ-greedy	$\epsilon = 0.05$
	LCB-ADVANTAGE-SSP	$H = 5, \iota = 0.05, \theta^\star = 4096$
	UC-SSP	$\iota = 1.0$
	Bernstein-SSP	$\iota = 2.0$
	ULCVI	$H = 80, \iota = 2.0$
	EB-SSP	$\iota = 0.05$
	SVI-SSP	$H = 15, \iota = 0.05$
GridWorld	Q-learning with ϵ-greedy	$\epsilon = 0.05$
	LCB-ADVANTAGE-SSP	$H = 5, \iota = 0.1, \theta^\star = 4096$
	UC-SSP	$\iota = 0.5$
	Bernstein-SSP	$\iota = 0.5$
	ULCVI	$H = 100, \iota = 1.0$
	EB-SSP	$\iota = 0.01$
	SVI-SSP	$H = 10, \iota = 0.01$