A Unified Algorithm for Stochastic Path Problems

Christoph Dann   CDANN@CDANN.NET
Google Research

Chen-Yu Wei   CHENYU.WEI@USC.EDU
University of Southern California

Julian Zimmert   ZIMMERT@GOOGLE.COM
Google Research

Abstract

We study reinforcement learning in stochastic path (SP) problems. The goal in these problems is to maximize the expected sum of rewards until the agent reaches a terminal state. We provide the first regret guarantees in this general problem by analyzing a simple optimistic algorithm. Our regret bound matches the best known results for the well-studied special case of stochastic shortest path (SSP) with all non-positive rewards. For SSP, we present an adaptation procedure for the case when the scale of rewards $B_\star$ is unknown. We show that there is no price for adaptation, and our regret bound matches that with a known $B_\star$. We also provide a scale adaptation procedure for the special case of stochastic longest paths (SLP) where all rewards are non-negative. However, unlike in SSP, we show through a lower bound that there is an unavoidable price for adaptation.

1. Introduction

Imagine a web application with recurring user visits (epochs). During each visit, the app can choose from different contents to present to the user (actions), which might lead to a desired interaction such as a purchase or click on an ad (reward). The user's behaviour depends on their internal state, which is influenced by the content provided to them. Inevitably, at some point the user will abandon the session.

It is natural to model this as an episodic reinforcement-learning problem. However, the length of each episode is random and depends on the agent's actions. To deal with the random episode length, and hence potentially unbounded cumulative reward in a single episode, one could either consider a fixed horizon problem by clipping the length of each episode, or consider discounted rewards. Both approaches introduce biases to the actual objective of the agent and we consider a third option: the *stochastic longest path* (SLP) setting, which is analogous to the *stochastic shortest path* (SSP) problem (e.g., Rosenberg et al., 2020; Cohen et al., 2021; Tarbouriech et al., 2021b; Chen et al., 2021a) but with positive rewards instead of costs.

In more generality, we can assume a setting in which there are both negative rewards (cost), as well as positive rewards. For example, users without subscription using the free part of an application might induce an overall negative reward due to the cost of infrastructure. However, the hope is that the user is convinced by the free service to upgrade to a subscription, which provides revenue. We call this setting the *stochastic path* (SP) problem. Besides the example above, in RL applications such as video games, chess, robot navigation, etc., the termination of the task is usually not determined by the time or the umber of rounds, but by whether or not the agent meets a termination criteria. In these problems, SP can be a better formulation than the standard fixed-horizon or discounted ones.

We make the following contributions:

Table 1: Overview of regret bounds for stochastic path problems. See Section 2 for definitions.

Setting	Scale $B_\star$	Reg_K in $\tilde{O}(\cdot)$	
SP	known	$R\sqrt{SAK} + R_{\max}SA + B_\star S^2 A$	Theorem 2
		$R\sqrt{SAK}$ **(lower bound)**	Theorem 3, Theorem 4
SLP	known	$\sqrt{V_\star B_\star SAK} + B_\star S^2 A$	Theorem 6
	unknown	$B_\star S\sqrt{AK}$ or $\sqrt{V_\star B_\star SAK} + \frac{B_\star^2}{V_\star}S^3 A$	Theorem 8
		$B_\star\sqrt{SAK}$ or $\sqrt{V_\star B_\star SAK} + \frac{B_\star^2}{V_\star}SA$ **(lower bound)**	Corollary 10
SSP	known	$\sqrt{V_\star B_\star SAK} + B_\star S^2 A$	Tarbouriech et al. (2021b); Chen et al. (2021a)
	unknown	$\sqrt{V_\star B_\star SAK} + B_\star^3 S^3 A$	Tarbouriech et al. (2021b); Chen et al. (2021a)
		$\sqrt{V_\star B_\star SAK} + B_\star S^2 A$	Theorem 11

1. We formalize the general SP problem and provide a simple unified algorithm for it with SSP and SLP as special cases.

2. We present the first regret upper-bound for the SP and SLP problem and show through lower-bounds that they are minimax-optimal up to log-factors and lower-order terms. Technically, our analysis gives the first near-optimal near-horizon-free regret bound for episodic MDPs when the reward could be positive or negative. In comparison, previous analysis in Zhang et al. (2021); Tarbouriech et al. (2021b); Chen et al. (2021a) can only get near-horizon-free bounds for all-non-negative or all-non-positive reward (see Section 3 for more discussions).

3. For SSP, when the scale of the sum of rewards $B_\star$ is unknown, we derive an improved procedure to adapt to $B_\star$. Unlike prior work (Tarbouriech et al., 2021b; Chen et al., 2021a), our adaptation procedure allows us to recover the regret bound achieved with a known $B_\star$.

4. For SLP, we also derive an algorithm that adapts to unknown $B_\star$. This adaptation is qualitatively different than in the SSP case. In fact, we show through a lower bound that adaptivity to unkonwn $B_\star$ comes at an unavoidable price in SLP.

The contributions 3 and 4 above jointly formalize a distinction between SSP and SLP when the scale of cumulative rewards is unknown. An overview of the main regret bounds derived in this work and a comparison to existing results is available in Table 1.

Related work The SP problem and its special cases SSP and SLP are episodic reinforcement learning settings. When the horizon, the length of each episode is fixed and known, these problems have been extensively studied (Dann and Brunskill, 2015; Azar et al., 2017; Jin et al., 2018; Efroni et al., 2021; Zanette and Brunskill, 2019; Zhang et al., 2020). Among these work on finite-horizon tabular RL, the recent line of work on *horizon-free* algorithms (Wang et al., 2020; Zhang et al., 2020, 2022) is of particular interest. These works assume that rewards are non-negative and their cumulative sum are bounded by 1 and aim for regret that only incurs a logarithmic dependency on the horizon. Although many techniques developed there are useful for our setting as well, the SLP and general SP problem is more difficult since the reward sum is not bounded by 1, and there are potentially negative rewards.

The SSP problem has seen a number of publications recently (Rosenberg et al., 2020; Cohen et al., 2021; Tarbouriech et al., 2020, 2021a,b; Vial et al., 2022; Chen et al., 2021a,b, 2022a,b; Chen and Luo, 2021, 2022; Jafarnia-Jahromi et al., 2021; Min et al., 2022; Yin et al., 2022), for which we refer to Tarbouriech et al. (2021b); Chen et al. (2021a) for a detailed comparison.

2. Preliminaries

We consider a stochastic path (SP) problem with a finite state space $\mathcal{S}$, a finite action set $\mathcal{A}$, an initial state $s_{\text{init}} \in \mathcal{S}$, a terminal state g (for notational simplicity, we let $g \notin \mathcal{S}$), a transition kernel $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S} \cup \{g\}}$, and a reward function $r : \mathcal{S} \times \mathcal{A} \to [-1, 1]$. We define $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$. In an *episode*, the player starts from the initial state $s_1 = s_{\text{init}}$[1]. At the i-th step in an episode, the player sees the current state $s_i \in \mathcal{S}$, takes an action $a_i \in \mathcal{A}$, which leads to a reward value $r(s_i, a_i)$ and generates the next state s_{i+1} according to $s_{i+1} \sim P_{s_i, a_i}(\cdot)$. The episode terminates right after the player reaches state g (no action is taken on g). We assume that the reward r is known to the learner, while the transition P is not. We call the problem *stochastic longest path* (SLP) if $r(s, a) \geq 0$ for all s, a, and call it *stochastic shortest path* (SSP) if $r(s, a) \leq 0$ for all s, a.

A history-dependent deterministic policy $\pi = (\pi_1, \pi_2, \dots)$ is a mapping from state-action histories to actions, i.e., $\pi_i : (\mathcal{S} \times \mathcal{A})^{i-1} \times \mathcal{S} \to \mathcal{A}$; we use Π^{HD} to denote the set of all history-dependent deterministic policies. A stationary deterministic policy π is a mapping from states to actions, i.e., $\pi : \mathcal{S} \to \mathcal{A}$; we use Π^{SD} to denote the set of all stationary deterministic policies. The state value function of a policy $\pi \in \Pi^{\text{HD}}$ on state $s \in \mathcal{S}$ are defined as

$$V^{\pi}(s) \triangleq \mathbb{E}^{\pi} \left[\sum_{i=1}^{\tau} r(s_i, a_i) \,\middle|\, s_1 = s \right],$$

where $\tau = \min\{i : s_{i+1} = g\}$, i.e., the timestep right before reaching the terminal state g (or ∞ if g is never reached), and $\mathbb{E}^{\pi}$ denotes expectation under policy π. Naturally, $V^{\pi}(g) \triangleq 0$.

A policy π is called *proper* if g is reached with probability 1 under policy π starting from any state. In this paper, we make the following assumption:

Assumption 1 *All policies in Π^{HD} are proper.*

Assumption 1 is stronger than those in previous works on SSP (Rosenberg et al., 2020; Cohen et al., 2021; Tarbouriech et al., 2021b; Chen et al., 2021a), which only require the *existence* of a proper policy. We note that the algorithmic trick they developed (adding a small amount of cost to every step) can also help us weaken Assumption 1. More details on this are available in Appendix G.

By Theorem 7.1.9 of Puterman (2014), Assumption 1 implies that there is a stationary and deterministic optimal policy $\pi^{\star} \in \Pi^{\text{SD}}$ such that $V^{\pi^{\star}}(s) \geq V^{\pi}(s)$ for any $\pi \in \Pi^{\text{HD}}$ and any s. We let $V^{\star}(\cdot) \triangleq V^{\pi^{\star}}(\cdot)$ and define

$$V_{\star} \triangleq |V^{\star}(s_{\text{init}})|, \quad B_{\star} \triangleq \max_s |V^{\star}(s)|.$$

To establish our result, we also need the following definitions:

---

1. This assumption is for simplicity and is without loss of generality – if the initial state is drawn from a fixed distribution $\rho \in \Delta_{\mathcal{S}}$, we can create a virtual initial state s_{init} on which every action leads to zero reward and next state distribution ρ.

Definition 1 *Define*

$$R \triangleq \sup_{\pi \in \Pi^{\mathrm{HD}}} \sqrt{\mathbb{E}^{\pi}\left[\left(\sum_{i=1}^{\tau} r(s_i, a_i)\right)^2 \,\middle|\, s_1 = s_{init}\right]},$$

$$R_{\max} \triangleq \max_{s} \sup_{\pi \in \Pi^{\mathrm{HD}}} \sqrt{\mathbb{E}^{\pi}\left[\left(\sum_{i=1}^{\tau} r(s_i, a_i)\right)^2 \,\middle|\, s_1 = s\right]},$$

$$T_{\max} \triangleq \max_{s} \sup_{\pi \in \Pi^{\mathrm{HD}}} \mathbb{E}^{\pi}\left[\tau \mid s_1 = s\right],$$

where $\tau = \min\{i : s_{i+1} = g\}$, i.e., the timestep right before reaching the terminal state g.

In words, $T_{\max}$ is the maximum (over all policies and all states) expected time to reach the terminal state; R and $R_{\max}$ are two quantities that represent the *range* of the total reward in an episode. Notice that in the definition of R, the starting state is fixed as s_{init}, while in the definition of $R_{\max}$, a maximum is taken over all possible starting states.

Under Assumption 1, $V_\star$, $B_\star$, R, $R_{\max}$, and $T_{\max}$ are all bounded. For simplicity, we assume that they are all ≥ 1.

2.1. Learning Protocol

The learning procedure considered in this paper is the same as previous works on SSP. We let the learner interact with the SP for K episodes, each started from s_{init}. We define the regret as the difference between $KV^\star(s_{\mathrm{init}})$ (the expected total reward obtained by the optimal policy) and the total reward of the learner. We keep a time index t to track the number of steps executed by the learner, and let s_t denote the state the learner sees at time t. Episode k starts at time t_k, and thus $s_{t_k} = s_{\mathrm{init}}$. At time t, the learner takes an action a_t, and transitions to $s'_t \sim P_{s_t,a_t}(\cdot)$. If $s'_t \neq g$, we let $s_{t+1} = s'_t$; otherwise, we let $t_{k+1} = t + 1$ to be the first step of episode $k + 1$. The reader can refer to Algorithm 1 to see how the time indices are updated. The regret can be written as

$$\mathrm{Reg}_K = \sum_{k=1}^{K}\left(V^\star(s_{\mathrm{init}}) - \sum_{t=t_k}^{e_k} r(s_t, a_t)\right),$$

with $e_k = t_{k+1} - 1$. We let T to be the total number of steps during K episodes. That is, $T = e_K$.

2.2. Notation

For $x > 0$, define $\ln_+(x) \triangleq \ln(1 + x)$. We write $x = O(y)$ or $x \leq O(y)$ to mean that $x \leq cy$ for some universal constant c, and write $x = \tilde{O}(y)$ or $x \leq \tilde{O}(y)$ if $x \leq cy$ for some c that only contains logarithmic factors. $\mathbb{E}_t[\cdot]$ denotes expectation conditioned on history before time t. $[n]$ denotes the set $\{1, 2, , \ldots, n\}$. We define $\tilde{\iota}_{T,B,\delta} \triangleq (\ln(SA/\delta) + \ln\ln(BT)) \times \ln T$. $\mathbb{V}(P, V)$ where $P \in \Delta_S$ and $V \in \mathbb{R}^S$ denotes the variance of V under P, i.e., $\mathbb{V}(P, V) \triangleq \sum_{i=1}^{S} P(i)V(i)^2 - (\sum_{i=1}^{S} P(i)V(i))^2$.

Algorithm 1 VI-SP

---

1 **input**: $B \geq 1$, $0 < \delta < 1$, sufficiently large universal constants c_1, c_2 that satisfy $2c_1^2 \leq c_2$.

2 **Initialize**: $t \leftarrow 0$, $s_1 \leftarrow s_{\text{init}}$.

3 For all (s, a, s') where $s \neq g$, set

$$n(s, a, s') = n(s, a) \leftarrow 0, \quad Q(s, a) \leftarrow B, \quad V(s) \leftarrow B.$$

4 Set $V(g) \leftarrow 0$.

5 **for** $k = 1, \ldots, K$ **do**

6   **while** *true* **do**

7       $t \leftarrow t + 1$

8   `/* `$Q_t(s,a), V_t(s)$` are defined as the `$Q(s,a), V(s)$` at this point. */`

9   Take action $a_t = \operatorname{argmax}_a Q(s_t, a)$, receive reward $r(s_t, a_t)$, and transit to s'_t.

10   Update counters: $n_t \triangleq n(s_t, a_t) \leftarrow n(s_t, a_t) + 1$, $n(s_t, a_t, s'_t) \leftarrow n(s_t, a_t, s'_t) + 1$.

11   Define $\bar{P}_t(s') \triangleq \frac{n(s_t, a_t, s')}{n_t} \, \forall s'$.

12   Define $b_t \triangleq \max \left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_t, V) \iota_t}{n_t}}, \frac{c_2 B \iota_t}{n_t} \right\}$, where $\iota_t = \ln(SA/\delta) + \ln\ln(B n_t)$.

13       $Q(s_t, a_t) \leftarrow \min \left\{ r(s_t, a_t) + \bar{P}_t V + b_t, Q(s_t, a_t) \right\}$

14       $V(s_t) \leftarrow \max_a Q(s_t, a)$.

15   **if** $s'_t \neq g$ **then** then $s_{t+1} \leftarrow s'_t$;

16   **else** $s_{t+1} \leftarrow s_{\text{init}}$ and **break**;

---

3. An Algorithm for General Stochastic Path (SP)

Our algorithm for general SP is Algorithm 1, which is simplified from the SVI-SSP algorithm by Chen et al. (2021a). The inputs are a parameter B that is supposed to an upper bound of $B_\star$, and a confidence parameter δ. The algorithm maintains an optimistic estimator $Q(s, a)$ of $Q^\star(s, a) := r(s, a) + \mathbb{E}_{s' \sim P_{s,a}}[V^\star(s')]$ (i.e., with high probability, $Q(s, a) \geq Q^\star(s, a)$ always holds). In every step t, the learner chooses action $a_t = \operatorname{argmax}_a Q(s_t, a)$ based on the "optimism in the face uncertainty" principle (Line 9), and updates the entry $Q(s_t, a_t)$ after receiving the reward and the next state, with an additional exploration bonus b_t that keeps the optimism of $Q(s_t, a_t)$ (Line 11–Line 13). Although this algorithm is similar to the one in Chen et al. (2021a), the existing analysis only applies to SSP and SLP, and it is unclear how it handles general SP. Our main contribution in this section is to provide a regret guarantee for this algorithm in general SP.

The regret guarantee of Algorithm 1 is given by the following theorem.

Theorem 2 *If Assumption 1 holds, then Algorithm 1 with $B \geq B_\star$ ensures that with probability at least $1 - O(\delta)$, for all $K \geq 1$, with T being the total number of steps in K episodes,*

$$\text{Reg}_K = O\left(R\sqrt{SAK\tilde{\iota}_{T,B,\delta}} + R_{\max}SA \ln\left(\frac{R_{\max}K}{R\delta}\right) \tilde{\iota}_{T,B,\delta} + BS^2 A \tilde{\iota}_{T,B,\delta} \right),$$

where $\tilde{\iota}_{T,B,\delta} \triangleq (\ln(SA/\delta) + \ln\ln(BT)) \times \ln T.$[2]

---

2. Technically, the total number of steps T is a random quantity but can be replaced by $KT_{\max}$ with high probability if desired.

The proof of Theorem 2 can be found in Appendix A. Theorem 2 generalizes previous works on near-optimal near-horizon-free regret bounds for RL (Zhang et al., 2021; Tarbouriech et al., 2021b; Chen et al., 2021a). Specifically, with a closer look into their analysis, one can find that their analysis leads to a regret bound that depends on the magnitude of $\sum_{i \in \text{episode}} |r(s_i, a_i)|$, which can be much larger than $|\sum_{i \in \text{episode}} r(s_i, a_i)|$ if the rewards have mixed signs. To address this issue, we develop new analysis techniques to get a near-horizon-free regret bound, which only scales with $|\sum_{i \in \text{episode}} r(s_i, a_i)|$. Other than this, the rest of the proofs are similar to those in Chen et al. (2021a) with simplifications. Unlike prior work, our analysis does not involve the intricate "high-order expansion" as seen in Zhang et al. (2021); Tarbouriech et al. (2021b); Chen et al. (2021a), the possibility of which is hinted by Zhang et al. (2022).

Proof sketch for Theorem 2 We first connect the regret with the sum of *advantages*, $\sum_{t=1}^{T}(V^{\star}(s_t) - Q^{\star}(s_t, a_t))$. This is standard based on the performance difference lemma (Kakade and Langford, 2002).

We use the fact that $B \geq B_{\star}$ and the bonus construction to show that the value estimator $Q(s, a)$ always upper bounds $Q^{\star}(s, a)$ with high probability (Lemma 15). This relies on the monotonic value propagation idea developed by Zhang et al. (2021). Then following the analysis of Zhang et al. (2021) and Chen et al. (2021a), we can show the following high probability bound (Lemma 16):

$$\sum_{t=1}^{T}(V^{\star}(s_t) - Q^{\star}(s_t, a_t)) \leq \tilde{O}\left(\sqrt{SA \sum_{t=1}^{T} \mathbb{V}(P_{s_t, a_t}, V^{\star})} + BS^2 A\right). \tag{1}$$

The way we bound $\sum_{t=1}^{T} \mathbb{V}(P_{s_t, a_t}, V^{\star})$ is the key to handle the case where the rewards have mixed signs. Specifically, we show the following (Lemma 19):

$$\sum_{t=1}^{T} \mathbb{V}(P_{s_t, a_t}, V^{\star}) \leq \tilde{O}\left(R_{\max} \sum_{t=1}^{T}(V^{\star}(s_t) - Q^{\star}(s_t, a_t)) + R^2 K + R_{\max}^2\right). \tag{2}$$

Combining Eq. (1) and Eq. (2) and solving for $\sum_{t=1}^{T}(V^{\star}(s_t) - Q^{\star}(s_t, a_t))$, we get an upper bound for it, which in turn gives a high-probability regret bound. ■

3.1. Lower bound

In this subsection, we show that the upper bound established in Theorem 2 is nearly tight. The proofs for this subsection can be found in Appendix D.

Theorem 3 *For any $u \geq 2$, and $K \geq \Omega(SA)$, we can construct a set of SP instances such that $R \leq u$ for all instances, and there exists a distribution over these instances such that the expected regret of any algorithm is at least $\Omega(u\sqrt{SAK})$.*

Specially, in the lower bound construction of Theorem 3, $V_{\star}$ and $B_{\star}$ are of order $O(1)$ for all $u \leq \sqrt{K/(SA)}$, showing that $V_{\star}$ and $B_{\star}$ are insufficient to characterize the regret bound for general SP problems. As we will see in the following sections, this contrasts with the special cases SLP and SSP, where except for logarithmic terms, the coefficients in the regret bound can be completely characterized by $V_{\star}$ and $B_{\star}$.

The quantity R in the regret upper and lower bounds (Theorem 2 and Theorem 3) is undesirable because its definition involves a *supremum* over all policies, which might be very large. Is it possible to refine the upper bound so that it only depends on quantities that correspond to the *optimal policy*? Specifically, we define

$$R_\star \triangleq \max_s \sqrt{\mathbb{E}^{\pi^\star}\left[\left(\sum_{i=1}^\tau r(s_i, a_i)\right)^2 \,\middle|\, s_1 = s\right]},$$

and ask: can the regret bound only depends on $R_\star$? Notice that Theorem 3 is uninformative for this question because $R \approx R_\star$ in its construction. The next theorem gives a negative answer *when the learner is agnostic of the value of $R_\star$*.

Theorem 4 *Let $u \geq 2$ be arbitrarily chosen, and let $K \geq \Omega(SA)$. For any algorithm that obtains a expected regret bound of $\tilde{O}(u\sqrt{SAK})$ for all problem instances with $R_\star = R_{\max} \leq u$, there exists a problem instance with $R_\star = O(1)$ and $R_{\max} \leq u$ but the expected regret is at least $\tilde{\Omega}(u\sqrt{SAK})$.*

Given Theorem 4, a left open question is whether $\tilde{O}(R_\star\sqrt{SAK})$ is achievable when the learner has information about $R_\star$. Note that our algorithm Algorithm 1 only requires knowledge of $B_\star$, which, in general, does not provide information about $R_\star$ (e.g., in the construction of Theorem 4, $B_\star = O(1)$ for all $u \leq \sqrt{K/(SA)}$). Therefore, an algorithm with such a refined guarantee would be quite different from our algorithm.

4. Stochastic Longest Path (SLP)

For the special case SLP where $r(\cdot, \cdot) \geq 0$, we first demonstrate that our general result in Theorem 2 already gives a nearly tight bound. The following lemma connects the notion of $R, R_{\max}$ in general SP to $V_\star, B_\star$ in SLP.

Lemma 5 *If $r(s, a) \geq 0$ for all s, a, then $R = O(\sqrt{V_\star B_\star \ln_+(B_\star/V_\star)})$ and $R_{\max} = O(B_\star)$, where $\ln_+(x) \triangleq \ln(1 + x)$.*

Lemma 5 together with Theorem 2 immediately implies the regret guarantee for SLP. Specifically, assuming that $B \geq B_\star$, combining Theorem 2 and Lemma 5 yields

$$\text{Reg}_K = O\left(\sqrt{V_\star B_\star SAK \ln_+(B_\star/V_\star)\tilde{\iota}_{T,B,\delta}} + B_\star SA \ln\left(\frac{B_\star K}{V_\star \delta}\right)\tilde{\iota}_{T,B,\delta} + BS^2 A\tilde{\iota}_{T,B,\delta}\right). \quad (3)$$

The logarithmic terms in Eq. (3) can be slightly improved if we follow a different approach to bound the sum of variance $\sum_t \mathbb{V}(P_{s_t,a_t}, V^\star)$. That is, using Lemma 20 instead of using Lemma 19. Note that the proof of Lemma 20 is similar to those of previous works (Zhang et al., 2021; Tarbouriech et al., 2021b; Chen et al., 2021a), which leads to a regret bound that depends on the magnitude of $\sum_{i\in\text{episode}} |r(s_i, a_i)|$ instead of $|\sum_{i\in\text{episode}} r(s_i, a_i)|$. Therefore, while it does not work for general SP, we can use it for SLP. Comparing Eq. (3) and the bound in Theorem 6, we see that specializing our general result in Theorem 2 to SLP only leads to looseness in logarithmic factors.

Theorem 6 *If Assumption 1 holds and $r(\cdot, \cdot) \geq 0$, then Algorithm 1 with $B \geq B_\star$ ensures that with probability at least $1 - \delta$, for all $K \geq 1$, with T being the total number of steps in K episodes,*

$$\text{Reg}_K = O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,B,\delta}} + BS^2 A\tilde{\iota}_{T,B,\delta}\right). \quad (4)$$

The proof of Theorem 6 is in Appendix B.

Algorithm 2 Procedure to estimate $V_\star$ in SLP

---

input: $\zeta \geq 1, U > 1$.

for $i = 1, \ldots, \lceil \log_2 U \rceil$ **do**

> Initiate a Algorithm 1 with $B = 2^i \zeta$ and probability parameter as $\delta' = \delta/\lceil \log_2 U \rceil$ (call this instance ALG).
>
> Run ALG until $N \geq 16c^2 \zeta S^2 A \tilde{\iota}_{M,B,\delta'}$, where N is the number of episodes, M is the total number of steps, and c is the universal constant hidden in the $O(\cdot)$ notation in Eq. (4).
>
> Let $\hat{r}_i$ be average reward of ALG in these N episodes (i.e., the total reward divided by N).

return $\hat{V} \triangleq 2 \max_i \{\hat{r}_i\}$.

Algorithm 3 VI-SLP for unknown $B_\star$

---

input: $\zeta \geq 1, U > 1$.

Run Algorithm 2 with inputs ζ and U, and get output $\hat{V}$.

Run Algorithm 1 with input $B = \hat{V}\zeta$ in the rest of the episodes.

4.1. Algorithm without knowledge of $B_\star$

While Theorem 6 gives a near-optimal bound, it is unclear how to make the algorithm work if prior knowledge on $B_\star$, which we need in order to set the value of B, is unavailable. Here, we first present a passive way to set B that is simple but leads to a highly sub-optimal bound. Observe from Theorem 6 that B only appears in the "lower-order" term in the regret bound. Therefore, a simple idea is to set B to be something large (of order $\sqrt{K/S^3 A}$) with the hope that $B \geq B_\star$ will hold in a wide range of cases. With this choice, if $B_\star \leq B$ indeed holds, then we enjoy a regret bound of $\tilde{O}(\sqrt{V_\star B_\star SAK} + BS^2 A) = \tilde{O}(\sqrt{V_\star B_\star SAK} + \sqrt{SAK}) = \tilde{O}(\sqrt{V_\star B_\star SAK})$; if $B_\star > B$, then we simply bound the regret by $V_\star K \leq O(V_\star B_\star^2 S^3 A)$, where the last inequality is implied by $B_\star \geq B = \Theta(\sqrt{K/(S^3 A)})$. Overall, this simple approach gives a regret bound of

$$\tilde{O}\left(\sqrt{V_\star B_\star SAK} + V_\star B_\star^2 S^3 A\right). \tag{5}$$

While the dominant is optimal, the lower-order term has cubic dependency (i.e., $V_\star B_\star^2$) on the scale of the cumulative reward, which is unnatural, and can easily overwhelm the dominant term when $V_\star B_\star^2 S^3 A \gtrsim \sqrt{V_\star B_\star SAK}$, or $B_\star \geq (K/(V_\star S^5 A))^{1/3}$. Previous prior-knowledge-free algorithms for SSP (Cohen et al., 2021; Tarbouriech et al., 2021b; Chen et al., 2021a) also suffer from this issue and have at least cubic dependency on the scale of cumulative reward.

In this subsection, we introduce a way to obtain a regret guarantee that only (nearly) linearly depends on the scale. Observe that in SLP, $B_\star$ corresponds to the maximum total expected reward the learner can get starting from any state. Clearly, the learner needs to have some knowledge about the *optimal policy* in order to estimate this quantity. Fortunately, it needs not to be accurately estimated; an estimation up to a constant factor suffices. Therefore, a reasonable plan is to coarsely estimate $B_\star$ up to a constant factor, and then use the estimation to set B. Notice that estimating $B_\star$ requires estimating $V^\star(s)$ for all s since $B_\star = \max_s V^\star(s)$.

While we can indeed make this idea work (details omitted), we find that an even more economical solution is to just estimate $V_\star = V^\star(s_{\text{init}})$ and set B to be something large compared to this estimation. Below we explain this idea. Let's first assume that $\frac{B_\star}{V_\star} \leq \zeta$ for some fixed ζ (will be relaxed later). Now consider running Algorithm 1 for $N = \tilde{\Theta}(\zeta S^2 A)$ episodes with parameter

$B = V\zeta$ for some value V that we choose. If we happen to choose a $V \in [V_\star, 2V_\star]$, then we have $B = V\zeta \geq V_\star\zeta \geq B_\star$, and thus the regret bound in Theorem 6 holds. Let $\hat{r}$ be the average reward in these N episodes (i.e., total reward divided by N). Then Theorem 6 gives

$$V_\star - \hat{r} \leq \tilde{O}\left(\sqrt{\frac{V_\star B_\star SA}{N}} + \frac{BS^2A}{N}\right) = O\left(\sqrt{\frac{V_\star B_\star}{\zeta S}} + \frac{B}{\zeta}\right) = O\left(\frac{V_\star}{\sqrt{S}} + V\right) = O(V_\star), \quad (6)$$

where in the first equality we use $N = \tilde{\Theta}(\zeta S^2 A)$, in the second equality we use $\frac{B_\star}{V_\star} \leq \zeta$ and $B = V\zeta$, in the third equality we use $V \leq 2V_\star$. By setting N to be large enough, we can ensure that the $O(V_\star)$ on the right-hand side is no more than $\frac{1}{2}V_\star$, which then gives $\frac{1}{2}V_\star \leq \hat{r} \leq \frac{3}{2}V_\star$.

On the other hand, if we choose some V that is not in the range of $[V_\star, 2V_\star]$ and set $B = V\zeta$, we may not have a good guarantee like in Eq. (6). However, the following reversed inequality must hold no matter how large V is:

$$V_\star - \hat{r} \geq -\tilde{O}\left(\sqrt{\frac{V_\star B_\star}{N}} + \frac{B_\star}{N}\right) = -\tilde{O}\left(\sqrt{\frac{V_\star B_\star}{\zeta S^2 A}} + \frac{B_\star}{\zeta S^2 A}\right) = -\tilde{O}(V_\star), \quad (7)$$

where the first inequality is by the fact that $V_\star \geq \mathbb{E}[\hat{r}]$ (because $V_\star$ is the expected value of the *optimal* policy) and that we can use Freedman's inequality to lower bound $V_\star - \hat{r}$ (details given in the formal proof). This inequality gives $\hat{r} \leq O(V_\star)$ no matter what V we use.

With the observations from Eq. (6) and Eq. (7), we have the following strategy to estimate $V_\star$ *given that $\frac{B_\star}{V_\star} \leq \zeta$ holds*: we perform the procedure described above for every $V \in \{1, 2, 4, 8, \ldots\}$. Let $\hat{r}_i$ denote the average reward when we use $V = 2^i$. By the argument in Eq. (6), at least one of the $\hat{r}_i$'s is of order $\Theta(V_\star)$; by the argument in Eq. (7), all $\hat{r}_i$'s are of order $O(V_\star)$. Combining them, we have $V_\star = \Theta(\max_i\{\hat{r}_i\})$. This procedure to estimate $V_\star$ is formalized in Algorithm 2, with its guarantee given in the following lemma:

Lemma 7 *Suppose that $\frac{B_\star}{V_\star} \leq \zeta$, and $V_\star \leq U$. Then Algorithm 2 with inputs ζ and U ensures that its output $\hat{V}$ satisfies $V_\star \leq \hat{V} \leq 3V_\star$.*

With $\hat{V}$ from Algorithm 2 being a coarse estimation of $V_\star$, we can use it to set the parameter B in Algorithm 1 as $B = \hat{V}\zeta$. Our overall algorithm is presented in Algorithm 3. However, notice that Lemma 7 only gives a meaningful guarantee when the two conditions $\frac{B_\star}{V_\star} \leq \zeta$ and $V_\star \leq U$ hold. Apparently, they do not hold for all instances. In the theorem below, we show that with appropriate choices of ζ and U, the additional regret due to their violation is well-bounded.

Theorem 8 *In the case when the learner has access to an absolute upper bound for $V_\star$ (for example, $T_{\max}$ is an absolute upper bound for $V_\star$), by setting U to be that upper bound and setting $\zeta = \sqrt{K/(S^2 A \ln U)}$, Algorithm 3 ensures with probability at least $1 - O(\delta)$*

$$\text{Reg} = O\left(B_\star\sqrt{S^2 AK \ln U}\,\tilde{\iota}_{T,KU,\delta}\right).$$

Alternatively, with $\zeta = \sqrt{K/(S^3 A \ln U)}$, Algorithm 3 ensures with probability at least $1 - O(\delta)$

$$\text{Reg} = O\left(\sqrt{V_\star B_\star SAK \ln U}\,\tilde{\iota}_{T,KU,\delta} + \frac{B_\star^2 S^3 A \ln U}{V_\star}\,\tilde{\iota}_{T,KU,\delta}\right).$$

In the case when the learner has NO access to an absolute upper bound for $V_\star$, we set $U = K^{1/\epsilon}$ for some parameters $\epsilon \in (0, 1)$. With $\zeta = \sqrt{K/(S^2 A \ln U)}$ and $\zeta = \sqrt{K/(S^3 A \ln U)}$, Algorithm 3 ensures with probability at least $1 - O(\delta)$

$$\text{Reg} = O\left(B_\star \sqrt{\epsilon^{-1} S^2 AK \ln K} \tilde\iota_{T,K,\delta\epsilon} + V_\star^{1+\epsilon}\right)$$

and
$$\text{Reg} = O\left(\sqrt{\epsilon^{-1} V_\star B_\star SAK \ln K} \tilde\iota_{T,K,\delta\epsilon} + \frac{\epsilon^{-1} B_\star^2 S^3 A \ln K}{V_\star} \tilde\iota_{T,K,\delta\epsilon} + V_\star^{1+\epsilon}\right),$$

respectively.

The proof can be found in Appendix B. In Theorem 8, we obtain two regret bounds for SLP without knowledge of $B_\star$: $\tilde{O}(B_\star\sqrt{S^2 AK})$ and $\tilde{O}(\sqrt{V_\star B_\star SAK} + \frac{B_\star^2}{V_\star}S^3 A)$, both not matching the bound $O(\sqrt{V_\star B_\star SAK} + B_\star S^2 A)$ in Theorem 6 for the case with a known $B_\star$. Is it possible to close the gap between the "known $B_\star$" and the "unknown $B_\star$" cases? In Section 4.2, we will show that this is *impossible*, by giving a regret lower bound for algorithms agnostic of $B_\star$. The lower bound can be strictly larger than the upper bound with knowledge of $B_\star$, thus formally identifying the *price of information about $B_\star$* for SLP. Finally, we remark on the source of suboptimality in the bounds in Theorem 8. The bounds we get are $B_\star\sqrt{S^2 AK}$ and $\sqrt{V_\star B_\star SAK} + \frac{B_\star^2}{V_\star}S^3 A$. The additional S dependencies come from the S^2 in the lower-order term in Theorem 6. It is conjectured by previous work (Zhang et al., 2021) that this S^2 in the lower-order term can be improved to S. If this conjecture is true, then our bounds in Theorem 8 can be improved to $B_\star\sqrt{SAK}$ and $\sqrt{V_\star B_\star SAK} + \frac{B_\star^2}{V_\star}SA$. As we will see in Section 4.2, these bounds are unimprovable when $B_\star$ is unknown.

4.2. Lower bound for algorithms agnostic of $B_\star$

If the magnitude of $B_\star$ is known, Theorem 6 shows that a regret bound of $\tilde{O}(\sqrt{V_\star B_\star SAK} + B_\star S^2 A)$ is possible. In this section, we show that there is a price to pay for adaptivity.

Theorem 9 *In SLP, for any algorithm agnostic to $B_\star$ that obtains a regret bound of $\tilde{O}(\nu\sqrt{SAK})$ for any problem instance where $B_\star = V_\star = \nu$ and sufficiently large K, there exists a problem instance with $V_\star \leq 1 + 2\nu$, $B_\star = \tilde{O}(\nu\sqrt{K/(SA)})$ such that the regret is at least $\Omega(\nu K)$.*

See Appendix E for the proof. This theorem implies that being agnostic to $B_\star$ is fundamentally harder than knowing an order optimal bound on $B_\star$, since Theorem 6 obtains sub-linear regret of order $\tilde{O}(\nu(SA)^{\frac{1}{4}}K^{\frac{3}{4}})$ in the hard instance mentioned in Theorem 9. Considering two classes of upper bounds, one in which we always scale with $\sqrt{K}$ without dominating lower order terms, and one in which we allow a constant cost for adapting to an unknown $B_\star$, we directly derive the following results.

Corollary 10 *Any algorithm with an asymptotic upper bound of*

$$\tilde{O}\left(B_\star^\alpha V_\star^{1-\alpha}\sqrt{SAK}\right) + o\left(B_\star^2\right),$$

satisfies at least $\alpha \geq 1$ and any algorithm with an upper bound of

$$O\left(\sqrt{V_\star B_\star SAK} + \left(\frac{B_\star}{V_\star}\right)^2 poly(V_\star, S, A)\right)$$

requires the constant term to be at least $\tilde\Omega\left(\frac{B_\star^2 SA}{V_\star}\right)$.

Proof For the first part, note that for any $\alpha < 1$, the regret bound for the bad case in Theorem 9 with $\nu = O(1)$ reads $B_\star^\alpha V_\star^{1-\alpha} \sqrt{SAK} + o(B_\star^2) = O(K^{\frac{1}{2}(1+\alpha)}(SA)^{\frac{1}{2}(1-\alpha)}) + o(K)$, which is sublinear in K and hence constitutes a contradiction. Similarly, in the second case, we can make K large enough such that the constant term (i.e., $(\frac{B_\star}{V_\star})^2 \text{poly}(V_\star, S, A)$) is absorbed by the dominant term (i.e., $\sqrt{V_\star B_\star SAK}$) in the $V_\star = B_\star = \nu$ environment, which means we can apply Theorem 9 and the $\text{poly}(V_\star, S, A)$ term must be of order $\Omega(V_\star SA)$, to satisfy the $\Omega(\nu K)$ lower bound. ∎

5. Stochastic Shortest Path (SSP)

SSP has been studied extensively recently. The works by Tarbouriech et al. (2021b) and Chen et al. (2021a) have achieved a near-optimal regret bound $\tilde{O}(\sqrt{V_\star B_\star SAK} + B_\star S^2 A)$ when the knowledge on $B_\star$ is available to the learner.[3] When such knowledge is unavailable, they design a way to adjust B on the fly, and achieve a regret bound of $\tilde{O}(\sqrt{V_\star B_\star SAK} + B_\star^3 S^3 A)$.

In this section, we improve their results, showing that for SSP, a bound of $\tilde{O}(\sqrt{V_\star B_\star SAK} + B_\star S^2 A)$ is possible even without prior knowledge on $B_\star$. This is a contrast with Theorem 9 and Corollary 10, which show that for SLP, without prior knowledge on $B_\star$, this bound is unachievable.

Our algorithm is Algorithm 4. It is almost identical to Algorithm 1 with three main differences. First, B is no longer an input parameter in Algorithm 4, but is an internal variable updated on the fly. Second, the initial $Q(s, a)$, $V(s)$ values are initialized as 0 in Algorithm 4, instead of B, which is natural since $Q^\star(s, a), V^\star(s) \leq 0$ for SSP. Third, in Line 12–Line 16, the algorithm tries to find a large enough B to set the bonus term b_t, so that the resulted $|Q(s_t, a_t)|$ is upper bounded by B.

The operation in Line 12–Line 16 and the corresponding analysis are the keys to our improvement. Recall that we hope to always have $0 \geq Q_t(s, a) \geq Q^\star(s, a)$ to ensure optimism. Also, we want B to be not too much larger than $B_\star$ to avoid regret overhead. Let's assume $0 \geq Q_t(s, a) \geq Q^\star(s, a)$ for all s, a at time t. In Line 14, we attempt to calculate $Q_{t+1}(s_t, a_t)$ (denoted as $Q_{t+1}^{\text{tmp}}(s_t, a_t)$ below). If $B \geq B_\star$ holds, then since $r(s_t, a_t) + \bar{P}_t V_t + b_t \geq Q^\star(s_t, a_t)$ by the same argument as in the proof of Lemma 15 and $Q_t(s_t, a_t) \geq Q^\star(s_t, a_t)$ by assumption, we have $0 \geq Q_{t+1}^{\text{tmp}}(s_t, a_t) \geq Q^\star(s_t, a_t) \geq -B_\star \geq -B$ by the definition of $Q_{t+1}^{\text{tmp}}(s_t, a_t)$ in Line 14. Thus, B will not be increased in Line 16. In short, *if optimism always holds* (i.e., $Q_t(s, a) \geq Q^\star(s, a)$), we will only increase B in Line 16 when $B < B_\star$, and thus $B < 2B_\star$ all the time.

The question then is how to show that optimism holds along the way. Because we start from $B = 1$ and only increases B when we are sure that $B < B_\star$, one might suspect that the bonus term b_t defined through B is insufficient at the beginning, and the optimism might fail. Because of this, Tarbouriech et al. (2021b) and Chen et al. (2021a) bound the regret in the $B < B_\star$ regime by a term linear in K. However, one key observation in the analysis is that the original purpose of the bonus term is to compensate the deviation of $(P_{s_t, a_t} - \bar{P}_t) V_t$, where $|V_t| \leq B$ by our algorithm. Since V_t is history-dependent, a common trick in the analysis (Azar et al., 2017; Zhang et al., 2021) is to replace V_t by $V^\star$ and bound the deviation of $(P_{s_t, a_t} - \bar{P}_t) V^\star$ using Freedman's inequality, for which a bonus term defined through $B_\star$ is required. To deal with our case, instead of replacing V_t

---

3. The upper bound reported in Tarbouriech et al. (2021b) and Chen et al. (2021a) is of order $B_\star \sqrt{SAK} + B_\star S^2 A$, which is larger than what we report here. This is simply because in their analysis, they upper bound $V_\star$ by $B_\star$. We redo their analysis and report their refined dependence on $V_\star$ here. Similarly, the lower bound obtained in Rosenberg et al. (2020) is $B_\star \sqrt{SAK}$ because in their lower bound construction, they only consider instances where $V_\star$ and $B_\star$ are of the same order. Hence, the upper bound we obtain here does not violate their lower bound.

Algorithm 4 VI-SSP for unknown $B_\star$

---

1 **input**: $0 < \delta < 1$, sufficiently large universal constants c_1, c_2 that satisfy $2c_1^2 \leq c_2$.

2 **Initialize**: $B \leftarrow 1, t \leftarrow 0, s_1 \leftarrow s_{\text{init}}$.

3 For all (s, a, s') where $s \neq g$, set

$$n(s, a, s') = n(s, a) \leftarrow 0, \quad Q(s, a) \leftarrow 0, \quad V(s) \leftarrow 0.$$

4 Set $V(g) \leftarrow 0$.

5 **for** $k = 1, \ldots, K$ **do**

6   **while** *true* **do**

7          $t \leftarrow t + 1$

8   /* $Q_t(s,a), V_t(s), B_t$ are defined as the $Q(s,a), V(s), B$ at this point. */

9   Take action $a_t = \text{argmax}_a Q(s_t, a)$, receive reward $r(s_t, a_t)$, and transit to s'_t.

10   Update counters: $n_t \triangleq n(s_t, a_t) \leftarrow n(s_t, a_t) + 1, n(s_t, a_t, s'_t) \leftarrow n(s_t, a_t, s'_t) + 1$.

11   Define $\bar{P}_t(s') \triangleq \frac{n(s_t, a_t, s')}{n_t} \forall s'$.

12   **while** *true* **do**

13   Define $b_t \triangleq \max\left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_t, V)\iota_t}{n_t}}, \frac{c_2 B\iota_t}{n_t} \right\}$, where $\iota_t = \ln(SA/\delta) + \ln\ln(Bn_t)$.

14              $Q^{\text{tmp}}(s_t, a_t) \leftarrow \min\left\{ r(s_t, a_t) + \bar{P}_t V + b_t, Q(s_t, a_t) \right\}$

15   **if** $|Q^{tmp}(s_t, a_t)| \leq B$ **then** $Q(s_t, a_t) \leftarrow Q^{\text{tmp}}(s_t, a_t)$ and **break** ;

16              $B \leftarrow 2B$.

17   /* b_t and ι_t are defined as the b_t and ι_t at this point. */

18          $V(s_t) \leftarrow \max_a Q(s_t, a)$.

19   **if** $s'_t \neq g$ **then** then $s_{t+1} \leftarrow s'_t$;

20   **else** $s_{t+1} \leftarrow s_{\text{init}}$ and **break**;

---

by $V^\star$, we replace it by $V^\star_{[B]} \triangleq \max\{-B, V^\star\}$ and use Freedman's inequality on $(P_{s_t, a_t} - \bar{P}_t)V^\star_{[B]}$, for which a bonus term defined through B suffices. We can further connect $V^\star_{[B]}$ back to $V^\star$ using the property $V^\star_{[B]} \geq V^\star$. The details are provided in Lemma 22, where we show that with high probability, $Q_t(s, a) \geq Q^\star(s, a)$ holds for all t, s, a, even if B is smaller than $B_\star$ along the learning process. The formal guarantee of our algorithm is given by the following theorem, with proof deferred to Appendix C.

Theorem 11 *Algorithm 4 guarantees for SSP problems that with probability at least $1 - O(\delta)$,* $\text{Reg}_K = \tilde{O}(\sqrt{V_\star B_\star SAK} + B_\star S^2 A)$.

6. Conclusions and Open Problems

In this work, we formulate the SP problem and give the first near-optimal regret bound for it. For special cases SLP and SSP, we further investigate the situation when the scale of the total reward $B_\star$ is unknown. By improving previous adaptation results for SSP, and giving new lower bounds for SLP, we formally show a distinction between these two cases when $B_\star$ is unknown.

    In the general case, although our algorithm achieves near-worst-case-optimal bounds in terms of R, there is still possibility of improving the bound using more refined quantities. We have ruled

out possibility of $V_\star$, $B_\star$, and the possibility of $R_\star$ when its value is unknown, but perhaps there are other candidate quantities. Further, there is a discrepancy in the analysis between the general SP / SLP setting and the SSP setting, i.e., while our result for the general case recovers that for the SLP case (up to logarithmic factors), it does not imply that for the SSP case. This also hints that R does not always capture the true difficulty of every instance.

In general SP when $B_\star$ is unknown, can we achieve the bound of order $\tilde{O}(R_{\max}\sqrt{\text{poly}(S, A)K})$, without any constant term that scales super-linearly with $R_{\max}$? This would be analogous to the $\tilde{O}(B_\star\sqrt{S^2AK})$ bound we get for SLP, but our technique there does not lead to this desired bound.

Finally, it remains an open question to prove or disprove Zhang et al. (2021)'s conjecture about the lower-order term, which has direct consequences for our adaptivity result in SLP.

References

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

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.

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*, pages 1651–1660. PMLR, 2021.

Liyu Chen and Haipeng Luo. Near-optimal goal-oriented reinforcement learning in non-stationary environments. *Advances in Neural Information Processing Systems*, 2022.

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

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

Liyu Chen, Rahul Jain, and Haipeng Luo. Improved no-regret algorithms for stochastic shortest path with linear mdp. In *International Conference on Machine Learning*, pages 3204–3245. PMLR, 2022a.

Liyu Chen, Haipeng Luo, and Aviv Rosenberg. Policy optimization for stochastic shortest path. *Conference on Learning Theory*, 2022b.

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

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *AAAI Conference on Artificial Intelligence*, 2021.

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Conference on Neural Information Processing Systems*, 2018.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. In *International Conference on Machine Learning*, pages 15584–15629. PMLR, 2022.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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

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

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. *Advances in Neural Information Processing Systems*, 34:6843–6855, 2021b.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Regret bounds for stochastic shortest path problems with linear function approximation. In *International Conference on Machine Learning*, pages 22203–22233. PMLR, 2022.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Conference on Neural Information Processing Systems*, 2020.

Ming Yin, Wenjing Chen, Mengdi Wang, and Yu-Xiang Wang. Offline stochastic shortest path: Learning, evaluation and towards optimality. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, 2019.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33: 15198–15207, 2020.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021.

Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pages 3858–3904. PMLR, 2022.

Appendix A. Upper bounds for General Stochastic Path

Definition 12 *Let Q_t, V_t be the Q, V at the beginning of round t (see the comments in Algorithm 1).*

Definition 13 *Define $\tilde{\iota}_{T,B,\delta} \triangleq (\ln(SA/\delta) + \ln\ln(BT)) \times \ln T$.*

A.1. Optimism and regret decomposition

Lemma 14 *Define*

$$f(P, V, n, \iota) = PV + \max\left\{ c_1 \sqrt{\frac{\mathbb{V}(P,V)\iota}{n}}, \frac{c_2 B \iota}{n} \right\}$$

If $2c_1^2 \leq c_2$ and $-B \leq V(\cdot) \leq B$, then $f(P, V, n, \iota)$ is increasing in V.

Proof We compute the derivative of $f(P, V, n, \iota)$ over $V(s^\star)$:

$$\frac{\partial f(P, V, n, \iota)}{\partial V(s^\star)} = P(s^\star) + \mathbf{1}\left\{ c_1 \sqrt{\frac{\mathbb{V}(P,V)\iota}{n}} > \frac{c_2 B \iota}{n} \right\} \times \frac{c_1 P(s^\star)(V(s^\star) - PV)\iota}{\sqrt{n\mathbb{V}(P,V)\iota}}$$

$$\geq P(s^\star) + \mathbf{1}\left\{ c_1 \sqrt{\frac{\mathbb{V}(P,V)\iota}{n}} > \frac{c_2 B \iota}{n} \right\} \times \frac{c_1^2 P(s^\star)(V(s^\star) - PV)}{c_2 B}$$

$$\geq P(s^\star) - \frac{2c_1^2}{c_2} P(s^\star)$$

$$\geq 0.$$

■

Lemma 15 *If $B \geq B_\star$, then with probability at least $1 - O(\delta)$, $Q_t(s, a) \geq Q^\star(s, a)$ for all (s, a) and t.*

Proof We use induction to prove this. When $t = 1$, $Q_1(s, a) = B \geq B_\star \geq Q^\star(s, a)$ for all s, a. Suppose that $Q_t(s, a) \geq Q^\star(s, a)$ for all for all s, a (which implies $V_t(s) \geq V^\star(s)$ for all s). Since Q_{t+1} and Q_t only differ in the entry (s_t, a_t), we only need to check $Q_{t+1}(s_t, a_t) \geq Q^\star(s_t, a_t)$. This can be seen from the calculation below:

$$r(s_t, a_t) + \bar{P}_t V_t + b_t = r(s_t, a_t) + \bar{P}_t V_t + \max \left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_t, V_t)\iota_t}{n_t}}, \frac{c_2 B \iota_t}{n_t} \right\}$$

(by the induction hypothesis $V_t(\cdot) \geq V^\star(\cdot)$ and the monotone property Lemma 14)

$$= r(s_t, a_t) + \bar{P}_t V^\star + \max \left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_t, V^\star)\iota_t}{n_t}}, \frac{c_2 B \iota_t}{n_t} \right\}$$

$$\geq r(s_t, a_t) + P_{s_t, a_t} V^\star \qquad \qquad \text{(By Lemma 28)}$$

$$= Q^\star(s_t, a_t).$$

Note that we can apply Lemma 28 only for $n_t \geq 4\iota_t$. If $n_t < 4\iota_t$, then the bias term itself is bounded by $\frac{c_2 B}{4} \geq B$ which ensures optimism. Besides, $Q_t(s_t, a_t) \geq Q^\star(s_t, a_t)$ by the induction hypothesis. Thus, $Q_{t+1}(s_t, a_t) \geq Q^\star(s_t, a_t)$, which finishes the induction. ∎

Lemma 16 *Suppose that $B \geq B_\star$. With probability at least $1 - O(\delta)$,*

$$\sum_{t=1}^{T} (V^\star(s_t) - Q^\star(s_t, a_t)) \leq O\left(\sqrt{SA \sum_{t=1}^{T} \mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B,\delta}} + BS^2 A \tilde{\iota}_{T,B,\delta} \right),$$

where $\tilde{\iota}_{T,B,\delta}$ is a logarithmic term defined in Definition 13.

Proof Below, we denote $P_t \triangleq P_{s_t, a_t}$.

$$\sum_{t=1}^{T} (Q_t(s_t, a_t) - Q^\star(s_t, a_t))$$

$$= \sum_{t=1}^{T} (Q_{t+1}(s_t, a_t) - Q^\star(s_t, a_t)) + \sum_{t=1}^{T} (Q_t(s_t, a_t) - Q_{t+1}(s_t, a_t))$$

$$\leq \sum_{t=1}^{T} \left(\bar{P}_t V_t - P_t V^\star \right) + \sum_{t=1}^{T} b_t + \sum_{t=1}^{T} \sum_{s,a} (Q_t(s, a) - Q_{t+1}(s, a))$$

$$\leq \sum_{t=1}^{T} P_t(V_t - V^\star) + \sum_{t=1}^{T} (\bar{P}_t - P_t)V^\star + \sum_{t=1}^{T} (\bar{P}_t - P_t)(V_t - V^\star) + \sum_{t=1}^{T} b_t + O(BSA)$$

$$\leq \underbrace{\sum_{t=1}^{T} \mathbf{1}_{s_t'}(V_t - V^\star)}_{\textbf{term}_1} + \underbrace{\sum_{t=1}^{T} (P_t - \mathbf{1}_{s_t'})(V_t - V^\star)}_{\textbf{term}_2} + \underbrace{\sum_{t=1}^{T} (\bar{P}_t - P_t)V^\star}_{\textbf{term}_3} + \underbrace{\sum_{t=1}^{T} (\bar{P}_t - P_t)(V_t - V^\star)}_{\textbf{term}_4}$$

$$+ \sum_{t=1}^{T} b_t + O(BSA)$$

$\underbrace{\phantom{+ \sum_{t=1}^{T} b_t}}_{\textbf{term}_5}$

where the first inequality is because $Q_{t+1}(s_t, a_t) \leq r(s_t, a_t) + \bar{P}_t V_t + b_t$ and $Q^\star(s_t, a_t) = r(s_t, a_t) + P_t V^\star$, and that $Q_t(s, a) \geq Q_{t+1}(s, a)$. Below, we bound the individual terms.

$$\begin{aligned}
\textbf{term}_1 &= \sum_{t=1}^{T} \mathbf{1}_{s'_t}(V_t - V^\star) \\
&\leq \sum_{t=1}^{T} \mathbf{1}_{s_{t+1}}(V_t - V^\star) \qquad (V_t(g) - V^\star(g) = 0 \text{ and } V_t(s) - V^\star(s) \geq 0 \text{ by Lemma 15}) \\
&\leq \sum_{t=1}^{T} \mathbf{1}_{s_t}(V_t - V^\star) + \sum_{t=1}^{T}(V_t(s_{t+1}) - V_t(s_t)) + \sum_{t=1}^{T}(V^\star(s_t) - V^\star(s_{t+1})) \\
&\leq \sum_{t=1}^{T} \mathbf{1}_{s_t}(V_t - V^\star) + \sum_{t=2}^{T}\sum_{s}(V_{t-1}(s) - V_t(s)) + O(B) \qquad (V_t \text{ decreases with } t) \\
&\leq \sum_{t=1}^{T} \mathbf{1}_{s_t}(V_t - V^\star) + O(BS).
\end{aligned} \tag{8}$$

We have $(P_t - \mathbf{1}_{s'_t})(V_t - V^\star)$ conditioned on step t is a zero mean random variable bounded in $[-B, B]$. By Lemma 26, with probability at least $1 - O(\delta)$, we have

$$\textbf{term}_2 = \sum_{t=1}^{T}(P_t - \mathbf{1}_{s'_t})(V_t - V^\star) = O\left(\sqrt{\sum_{t=1}^{T} \mathbb{V}(P_t, V_t - V^\star)\iota_t} + B\iota_t\right).$$

We have $(\bar{P}_t - P_t)V^\star = \frac{1}{n_t}\sum_{r:(s_r, a_r)=(s_t, a_t)}(\mathbf{1}_{s'_r} - P_r)V^\star$, which again by Lemma 26 is bounded for a fixed state, action pair simultaneously over all time-steps. Via union bound over all states and actions, we have with probability $1 - \mathcal{O}(\delta)$

$$\textbf{term}_3 = \sum_{t=1}^{T}(\bar{P}_t - P_t)V^\star = O\left(\sum_{t=1}^{T}\left(\sqrt{\frac{\mathbb{V}(P_t, V^\star)\iota_t}{n_t}} + \frac{B_\star \iota_t}{n_t}\right)\right). \tag{9}$$

For the next term, we use a union bound over all state action pairs and apply Lemma 27, to obtain with probability $1 - O(\delta)$

$$\textbf{term}_4 = \sum_{t=1}^{T}(\bar{P}_t - P_t)(V_t - V^\star) = O\left(\sum_{t=1}^{T}\left(\sqrt{\frac{\mathbb{V}(P_t, V_t - V^\star)\iota_t}{n_t}} + \frac{B\iota_t}{n_t}\right)\right). \tag{10}$$

Next, using Lemma 27 again, we have with probability $1 - O(\delta)$

$$\mathbb{V}(\bar{P}_t, V_t) = \bar{P}_t(V_t - \bar{P}_t V_t)^2$$

$$\leq \bar{P}_t(V_t - P_t V_t)^2$$

$$= \mathbb{V}(P_t, V_t) + (\bar{P}_t - P_t)(V_t - P_t V_t)^2$$

$$\leq \mathbb{V}(P_t, V_t) + 2B|(\bar{P}_t - P_t)(V_t - P_t V_t)|$$

$$\leq \mathbb{V}(P_t, V_t) + O\left(B\sqrt{S\frac{\mathbb{V}(P_t, V_t)\iota_t}{n_t}} + \frac{SB^2\iota_t}{n_t}\right)$$

$$\leq O\left(\mathbb{V}(P_t, V_t) + \frac{SB^2\iota_t}{n_t}\right). \qquad \text{(AM-GM inequality)}$$

By the definition of b_t, with probability at least $1 - O(\delta)$,

$$\mathbf{term}_5 = \sum_{t=1}^{T} b_t = O\left(\sum_{t=1}^{T}\left(\sqrt{\frac{\mathbb{V}(\bar{P}_t, V_t)\iota_t}{n_t}} + \frac{B\iota_t}{n_t}\right)\right)$$

$$= O\left(\sum_{t=1}^{T}\left(\sqrt{\frac{\mathbb{V}(P_t, V_t)\iota_t}{n_t}} + \frac{B\sqrt{S}\iota_t}{n_t}\right)\right)$$

$$= O\left(\sum_{t=1}^{T}\left(\sqrt{\frac{\mathbb{V}(P_t, V^\star)\iota_t}{n_t}} + \sqrt{\frac{\mathbb{V}(P_t, V_t - V^\star)\iota_t}{n_t}} + \frac{B\sqrt{S}\iota_t}{n_t}\right)\right)$$

Collecting terms and using Cauchy-Schwarz, we get

$$\sum_{t=1}^{T}(Q_t(s_t, a_t) - Q^\star(s_t, a_t)) \leq \sum_{t=1}^{T}(V_t(s_t) - V^\star(s_t))$$

$$+ O\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B,\delta}} + \sqrt{S^2 A\sum_{t=1}^{T}\mathbb{V}(P_t, V_t - V^\star)\tilde{\iota}_{T,B,\delta}} + BS^2 A\tilde{\iota}_{T,B,\delta}\right).$$

We further invoke Lemma 17 and bound the last expression by

$$\sum_{t=1}^{T}(V_t(s_t) - V^\star(s_t)) + O\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B,\delta}} + BS^2 A\tilde{\iota}_{T,B,\delta}\right).$$

Finally, noticing that $Q_t(s_t, a_t) = V_t(s_t)$ by the choice of a_t finishes the proof. ∎

Lemma 17 *With probability at least $1 - O(\delta)$,*

$$\sum_{t=1}^{T}\mathbb{V}(P_t, V_t - V^\star) = O\left(\frac{1}{S}\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star) + B^2 S^2 A\tilde{\iota}_{T,B,\delta}\right).$$

Proof Using Lemma 23 with $X_t = V_t - V^\star$, we get

$$\sum_{t=1}^{T} \mathbb{V}(P_t, V_t - V^\star)$$

$$= O\left(B\sum_{t=1}^{T}|V_t(s_t) - V^\star(s_t) - P_t(V_t - V^\star)| + B\sum_{t=1}^{T}\sum_s |V_t(s) - V_{t+1}(s)| + B^2\ln(1/\delta)\right)$$

$$= O\left(B\sum_{t=1}^{T}|(\bar{P}_t - P_t)V_t| + B\sum_{t=1}^{T}b_t + B^2S\ln(1/\delta)\right)$$

$$\leq O\left(B\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B,\delta}} + B\sqrt{S^2A\sum_{t=1}^{T}\mathbb{V}(P_t, V_t - V^\star)\tilde{\iota}_{T,B,\delta} + B^2S^2A\tilde{\iota}_{T,B,\delta}}\right).$$

(by the same argument as in Eq. (9) and Eq. (10))

Solving for $\sum_{t=1}^{T}\mathbb{V}(P_t, V_t - V^\star)$, we get

$$\sum_{t=1}^{T}\mathbb{V}(P_t, V_t - V^\star) = O\left(B\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B,\delta} + B^2S^2A\tilde{\iota}_T}\right)$$

$$= O\left(\frac{1}{S}\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star) + B^2S^2A\tilde{\iota}_{T,B,\delta}\right).$$

(by AM-GM)

∎

A.2. Bounding the sum of variance

In Appendix A.1, we have already shown that

$$\sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t)) = \tilde{O}\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)} + BS^2A\right)$$

in Lemma 16. In this subsection (Lemma 19), we close the loop and show

$$\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star) = \tilde{O}\left(R^2K + R_{\max}\sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t)) + R_{\max}^2\right). \tag{11}$$

We first establish some useful properties.

Lemma 18 *Define the following notation:*

$$Y_k \triangleq \sum_{t=t_k}^{e_k}\mathbb{V}(P_t, V^\star), \quad Z_k \triangleq \sum_{t=t_k}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t)), \tag{12}$$

and recall that $\ln_+(x) \triangleq \ln(1 + x)$. *We have*

$$\mathbb{E}_{t_k}[Z_k^2] \leq O\left(R_{\max} \ln_+\left(\frac{R_{\max}}{R}\right) \mathbb{E}_{t_k}[Z_k] + 1\right), \tag{13}$$

$$\mathbb{E}_{t_k}[Y_k] \leq O\left(R_{\max} \ln_+\left(\frac{R_{\max}}{R}\right) \mathbb{E}_{t_k}[Z_k] + R^2\right), \tag{14}$$

$$\mathbb{E}_{t_k}[Y_k^2] \leq O\left(R_{\max}^3 \ln_+^2\left(\frac{R_{\max}}{R}\right) \mathbb{E}_{t_k}[Z_k] + R_{\max}^2 R^2 \ln_+\left(\frac{R_{\max}}{R}\right)\right), \tag{15}$$

$$Z_k \leq R_{\max} \ln\left(\frac{K}{\delta}\right) \text{ for all } k \in [K] \text{ w.p.} \geq 1 - \delta, \tag{16}$$

$$Y_k \leq R_{\max}^2 \ln\left(\frac{K}{\delta}\right) \text{ for all } k \in [K] \text{ w.p.} \geq 1 - \delta. \tag{17}$$

Before proving Lemma 18, we point out that the key is to show Eq. (14), which, after summing over k, will almost imply Eq. (11), but only *in expectation*. The other inequalities Eq. (13), Eq. (15), Eq. (16), Eq. (17) will be used in concentration inequalities that boost the expectation bound to a high-probability bound.

Proof [Lemma 18]

Proving Eq. (13) We use Lemma 24 with $X_t = V^\star(s_t) - Q^\star(s_t, a_t)$. First, note that $0 \leq V^\star(s_t) - Q^\star(s_t, a_t) \leq 2B_\star \leq 2R_{\max}$. Then note that for any t' in episode k,

$$\begin{aligned}
\mathbb{E}_{t'}\left[\sum_{t=t'}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t))\right] &= \mathbb{E}_{t'}\left[\sum_{t=t'}^{e_k}(V^\star(s_t) - r(s_t, a_t) - P_t V^\star)\right] \\
&= \mathbb{E}_{t'}\left[\sum_{t=t'}^{e_k}(V^\star(s_t) - r(s_t, a_t) - V^\star(s_t'))\right] \\
&= V^\star(s_{t'}) - \mathbb{E}_{t'}\left[\sum_{t=t'}^{e_k}r(s_t, a_t)\right] \leq 2R_{\max}. \tag{18}
\end{aligned}$$

Combining these two arguments and using Lemma 24 (b) (with c set to R), we get

$$\begin{aligned}
\mathbb{E}_{t_k}[Z_k^2] = \mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t))\right)^2\right] \\
\leq O\left(R_{\max} \ln_+\left(\frac{R_{\max}}{R}\right) \mathbb{E}_{t_k}\left[\sum_{t=t_k}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t))\right] + R^2\right) \\
= O\left(R_{\max} \ln_+\left(\frac{R_{\max}}{R}\right) \mathbb{E}_{t_k}[Z_k] + R^2\right).
\end{aligned}$$

Proving Eq. (14) Observe that

$$\mathbb{E}_{t_k}\left[\sum_{t=t_k}^{e_k}\mathbb{V}(P_t, V^\star)\right] = \mathbb{E}_{t_k}\left[\sum_{t=t_k}^{e_k}(V^\star(s_t') - P_t V^\star)^2\right] = \mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}(V^\star(s_t') - P_t V^\star)\right)^2\right], \tag{19}$$

where the last equality is because

$$\mathbb{E}_{t_k}\left[(V^\star(s_t') - P_t V^\star)(V^\star(s_u') - P_u V^\star)\right] = 0$$

for any $u > t \geq t_k$. We continue with the following:

$$\mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}(V^\star(s_t') - P_t V^\star)\right)^2\right]$$

$$= \mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}(V^\star(s_t) - P_t V^\star) - V^\star(s_{t_k})\right)^2\right]$$

$$= \mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t) + r(s_t, a_t)) - V^\star(s_{t_k})\right)^2\right]$$

$$\leq 3\mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t))\right)^2\right] + 3\mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}r(s_t, a_t)\right)^2\right] + 3V^\star(s_{t_k})^2$$

$$\leq 3\mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t))\right)^2\right] + 6R^2. \tag{20}$$

Thus, combining the arguments above, we have proven $\mathbb{E}_{t_k}[Y_k] \leq O\left(\mathbb{E}_{t_k}[Z_k^2] + R^2\right)$. Further combining this with Eq. (13), we get Eq. (14).

Proving Eq. (15) For any t' in episode k, by the same calculation as in Eq. (19), Eq. (20) (but instead of starting time from t_k, start from an arbitrary t' in episode k), we have

$$\mathbb{E}_{t'}\left[\sum_{t=t'}^{e_k}\mathbb{V}(P_t, V^\star)\right] \leq O\left(\mathbb{E}_{t'}\left[\left(\sum_{t=t'}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t))\right)^2\right] + R_{\max}^2\right) \leq O(R_{\max}^2), \tag{21}$$

where the last inequality is by Eq. (18) and Lemma 24 (c) with $X_t = V^\star(s_t) - Q^\star(s_t, a_t)$. Thus,

$$\mathbb{E}_{t_k}[Y_k^2] = \mathbb{E}_{t_k}\left[\left(\sum_{t=t_k}^{e_k}\mathbb{V}(P_t, V^\star)\right)^2\right]$$

$$\leq O\left(R_{\max}^2 \ln_+\left(\frac{R_{\max}^2}{R^2}\right)\mathbb{E}_{t_k}\left[\sum_{t=t_k}^{e_k}\mathbb{V}(P_t, V^\star)\right] + R^4\right)$$

$$\text{(by Eq. (21) and Lemma 24 (b) with } c = R^2)$$

$$= O\left(R_{\max}^3 \ln_+^2\left(\frac{R_{\max}}{R}\right)\mathbb{E}_{t_k}[Z_k] + R_{\max}^2 R^2 \ln_+\left(\frac{R_{\max}}{R}\right)\right). \quad \text{(by Eq. (14))}$$

Proving Eq. (16) This is directly by Eq. (18) and Lemma 24 (a).

Proving Eq. (17)   This is directly by Eq. (21) and Lemma 24 (a).   ∎

Lemma 19   *With probability at least $1 - O(\delta)$,*

$$\sum_{t=1}^{T} \mathbb{V}(P_t, V^\star) \leq$$

$$O\left(R_{\max} \ln_+ \left(\frac{R_{\max}}{R} \right) \sum_{t=1}^{T} (V^\star(s_t) - Q^\star(s_t, a_t)) + R^2 K + R_{\max}^2 \ln \left(\frac{R_{\max} K}{R \delta} \right) \ln \left(\frac{\ln(R_{\max} K)}{\delta} \right) \right).$$

Proof   Similar to Eq. (12), we define

$$Y_k \triangleq \sum_{t=t_k}^{e_k} \mathbb{V}(P_t, V^\star), \quad Z_k \triangleq \sum_{t=t_k}^{e_k} (V^\star(s_t) - Q^\star(s_t, a_t)).$$

By Lemma 26, with probability at least $1 - O(\delta)$,

$$\sum_{k=1}^{K} Y_k$$

$$\leq \sum_{k=1}^{K} \mathbb{E}_{t_k}[Y_k] + O\left(\sqrt{\sum_{k=1}^{K} \mathbb{E}_{t_k}[Y_k^2] \ln \frac{\ln\left(\sum_{k=1}^{K} \mathbb{E}_{t_k}[Y_k^2] \right)}{\delta}} + \left(\max_{k \in [K]} Y_k \right) \ln \frac{\ln\left(\max_{k \in [K]} Y_k \right)}{\delta} \right)$$

$$\leq O\left(R_{\max} \ln_+ \left(\frac{R_{\max}}{R} \right) \sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k] + R^2 K \right) \qquad \text{(by Eq. (14))}$$

$$+ O\left(\sqrt{\left(R_{\max}^3 \ln_+^2 \left(\frac{R_{\max}}{R} \right) \sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k] + R_{\max}^2 R^2 \ln_+ \left(\frac{R_{\max}}{R} \right) K \right) \ln \frac{\ln(R_{\max} K \sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k])}{\delta}} \right)$$

$$\text{(by Eq. (15))}$$

$$+ O\left(R_{\max}^2 \ln \frac{K}{\delta} \ln \frac{\ln\left(R_{\max}^2 \ln \frac{K}{\delta} \right)}{\delta} \right) \qquad \text{(by Eq. (17))}$$

$$\leq O\left(R_{\max} \ln_+ \left(\frac{R_{\max}}{R} \right) \sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k] + R^2 K + R_{\max}^2 \ln \frac{R_{\max} K}{R \delta} \ln \frac{\ln(R_{\max} K)}{\delta} \right).$$

$$\text{(AM-GM and that } \mathbb{E}_{t_k}[Z_k] \leq R_{\max})$$

$$\tag{22}$$

Next, we connect $\sum_k \mathbb{E}_{t_k}[Z_k]$ with $\sum_k Z_k$. By Lemma 26, with probability at least $1 - O(\delta)$,

$$\sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k]$$

$$\leq \sum_{k=1}^{K} Z_k + O\left(\sqrt{\sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k^2] \ln \frac{\ln\left(\sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k^2]\right)}{\delta}} + \left(\max_{k\in[K]} Z_k\right) \ln \frac{\ln\left(\max_{k\in[K]} Z_k\right)}{\delta}\right)$$

$$\leq \sum_{k=1}^{K} Z_k + O\left(\sqrt{R_{\max} \ln_+\left(\frac{R_{\max}}{R}\right) \sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k] \ln \frac{\ln(R_{\max} K)}{\delta}}\right)$$

$$\text{(by Eq. (13) and that } \mathbb{E}_{t_k}[Z_k^2] \leq R_{\max}^2)$$

$$+ O\left(R_{\max} \ln \frac{K}{\delta} \ln \frac{\ln(R_{\max} \ln \frac{K}{\delta})}{\delta}\right) \qquad \text{(by Eq. (16))}$$

$$\leq \sum_{k=1}^{K} Z_k + \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k] + O\left(R_{\max} \ln \frac{R_{\max} K}{R\delta} \ln \frac{\ln(R_{\max} K)}{\delta}\right). \qquad \text{(AM-GM)}$$

Solving for $\sum_{k=1}^{K} \mathbb{E}_{t_k}[Z_k]$ and plugging it to Eq. (22), we get that with probability at least $1 - O(\delta)$,

$$\sum_{k=1}^{K} Y_k \leq O\left(R_{\max} \ln_+\left(\frac{R_{\max}}{R}\right) \sum_{k=1}^{K} Z_k + R^2 K + R_{\max}^2 \ln \frac{R_{\max} K}{R\delta} \ln \frac{\ln(R_{\max} K)}{\delta}\right).$$

This finishes the proof.

$\blacksquare$

A.3. Bounding the regret

Proof [Theorem 2] We use the following notations:

$$Y_k \triangleq \sum_{t=t_k}^{e_k} \mathbb{V}(P_t, V^\star), \quad Z_k \triangleq \sum_{t=t_k}^{e_k} (V^\star(s_t) - Q^\star(s_t, a_t)).$$

$$\text{Reg}_K = \sum_{k=1}^{K} \left(V^\star(s_{t_k}) - \sum_{t=t_k}^{e_k} r(s_t, a_t)\right)$$

$$= \sum_{t=1}^{T} (V^\star(s_t) - V^\star(s_t') - r(s_t, a_t))$$

$$= \sum_{t=1}^{T} (V^\star(s_t) - Q^\star(s_t, a_t)) + \sum_{t=1}^{T} (P_t V^\star - V^\star(s_t'))$$

$$\leq \sum_{t=1}^{T} (V^\star(s_t) - Q^\star(s_t, a_t)) + O\left(\sqrt{\sum_{t=1}^{T} \mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B,\delta}} + B_\star \tilde{\iota}_{T,B,\delta}\right)$$

$$\leq O\left(\sum_{k=1}^{K} Z_k + \sqrt{\sum_{k=1}^{K} Y_k \tilde{\iota}_{T,B,\delta}} + B_\star \tilde{\iota}_{T,B,\delta}\right)$$

$$\leq O\left(\sum_{k=1}^{K} Z_k + \sqrt{\left(R_{\max}\ln_+\left(\frac{R_{\max}}{R}\right)\sum_{k=1}^{K} Z_k + R^2 K + R_{\max}^2 \ln\frac{R_{\max}K}{R\delta}\ln\frac{\ln(R_{\max}K)}{\delta}\right)\tilde{\iota}_{T,B,\delta}} + B_\star \tilde{\iota}_{T,B,\delta}\right)$$

<div align="right">(Lemma 19)</div>

$$\leq O\left(\sum_{k=1}^{K} Z_k + R\sqrt{K\tilde{\iota}_{T,B,\delta}} + R_{\max}\ln\left(\frac{R_{\max}K}{R\delta}\right)\tilde{\iota}_{T,B,\delta}\right).$$

<div align="right">(AM-GM)</div>

<div align="right">(23)</div>

It remains to bound $\sum_{k=1}^{K} Z_k$. By Lemma 16, we have with probability at least $1 - O(\delta)$,

$$\sum_{k=1}^{K} Z_k \leq O\left(\sqrt{SA\sum_{k=1}^{K} Y_k \tilde{\iota}_{T,B,\delta} + BS^2 A\tilde{\iota}_{T,B,\delta}}\right). \tag{24}$$

Further using Lemma 19 on the right-hand side,

$$\sum_{k=1}^{K} Z_k \leq$$

$$O\left(\sqrt{SA\left(R_{\max}\ln_+\left(\frac{R_{\max}}{R}\right)\sum_{k=1}^{K} Z_k + R^2 K + R_{\max}^2 \ln\frac{R_{\max}K}{R\delta}\ln\frac{\ln(R_{\max}K)}{\delta}\right)\tilde{\iota}_{T,B,\delta} + BS^2 A\tilde{\iota}_{T,B,\delta}}\right).$$

Solving for $\sum_{k=1}^{K} Z_k$, we get

$$\sum_{k=1}^{K} Z_k \leq O\left(R\sqrt{SAK\tilde{\iota}_{T,B,\delta}} + R_{\max}SA\ln\left(\frac{R_{\max}K}{R\delta}\right)\tilde{\iota}_{T,B,\delta} + BS^2 A\tilde{\iota}_{T,B,\delta}\right).$$

Plugging this to Eq. (23) finishes the proof.

<div align="right">■</div>

Appendix B. Upper Bound for Stochastic Longest Path

Proof [Lemma 5] Define

$$B_\pi = \max\left\{\sup_s \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} r(s_t, a_t) \,\Big|\, s_1 = s\right], 1\right\}$$

$$V_\pi = \max\left\{\mathbb{E}^\pi\left[\sum_{t=1}^{\infty} r(s_t, a_t) \,\Big|\, s_1 = s_{\text{init}}\right], 1\right\}$$

Since $r(\cdot, \cdot) \geq 0$, by Lemma 24 (b) (with $c \triangleq \min\{B_\pi, V_\star\}$), we have

$$R^2 \leq O\left(\sup_\pi V_\pi B_\pi \ln_+\left(\frac{B_\pi}{\min\{B_\pi, V_\star\}}\right) + V_\star^2\right) \leq O\left(V_\star B_\star \ln_+\frac{B_\star}{V_\star}\right).$$

By Lemma 24 (c), we have

$$R_{\max}^2 \leq O\left(\sup_\pi B_\pi^2\right) \leq O\left(B_\star^2\right).$$

∎

Lemma 20 *With probability at least $1 - \delta$, for all T,*

$$\sum_{t=1}^T \mathbb{V}(P_t, V^\star) \leq O\left(B_\star \sum_{t=1}^T (V^\star(s_t) - Q^\star(s_t, a_t)) + B_\star \sum_{t=1}^T |r(s_t, a_t)| + B_\star^2 \ln(T/\delta)\right).$$

Proof

$$\sum_{t=1}^T \mathbb{V}(P_t, V^\star)$$

$$= \sum_{t=1}^T \left(\mathbb{E}_{s' \sim P_t}[V^\star(s')^2] - (P_t V^\star)^2\right)$$

$$= \sum_{t=1}^T (V^\star(s_t')^2 - (P_t V^\star)^2) + \sum_{t=1}^T (\mathbb{E}_{s' \sim P_t}[V^\star(s')^2] - V^\star(s_t')^2)$$

$$\leq \sum_{t=1}^T \left(V^\star(s_t)^2 - (P_t V^\star)^2\right) + B_\star^2 + \sum_{t=1}^T (\mathbb{E}_{s' \sim P_t}[V^\star(s')^2] - V^\star(s_t')^2)$$

$$\text{(because } V^\star(s_t')^2 \leq V^\star(s_{t+1})^2)$$

$$= \sum_{t=1}^T (V^\star(s_t)^2 - Q^\star(s_t, a_t)^2) + \sum_{t=1}^T (Q^\star(s_t, a_t)^2 - (P_t V^\star)^2)$$

$$+ O\left(\sqrt{\sum_{t=1}^T \mathbb{V}(P_t, V^{\star 2}) \ln(T/\delta)} + B_\star^2 \ln(T/\delta)\right)$$

$$\leq O\left(B_\star \sum_{t=1}^T |V^\star(s_t) - Q^\star(s_t, a_t)| + B_\star \sum_{t=1}^T |Q^\star(s_t, a_t) - P_t V^\star|\right)$$

$$+ O\left(B_\star \sqrt{\sum_{t=1}^T \mathbb{V}(P_t, V^\star) \ln(T/\delta)} + B_\star^2 \ln(T/\delta)\right)$$

$$(a^2 - b^2 \leq |a + b||a - b| \text{ and Lemma 29})$$

$$\leq O\left(B_\star \sum_{t=1}^T (V^\star(s_t) - Q^\star(s_t, a_t)) + B_\star \sum_{t=1}^T |r(s_t, a_t)|\right) + \frac{1}{2} \sum_{t=1}^T \mathbb{V}(P_t, V^\star) + O\left(B_\star^2 \ln(T/\delta)\right)$$

$$(\text{AM-GM})$$

Solving for $\sum_{t=1}^T \mathbb{V}(P_t, V^\star)$ finishes the proof. ∎

Proof [Theorem 6] By the same calculation as in Eq. (23), we have

$$\text{Reg}_K \leq \sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t)) + O\left(\sqrt{\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B,\delta}} + B_\star \tilde{\iota}_{T,B,\delta}\right).$$

Using Lemma 20, we get

$$\text{Reg}_K \leq \sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t))$$
$$+ O\left(\sqrt{B_\star \sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t))\tilde{\iota}_{T,B,\delta} + B_\star \sum_{t=1}^{T}r(s_t, a_t)\tilde{\iota}_{T,B,\delta} + B_\star \tilde{\iota}_{T,B,\delta}}\right)$$
$$\leq O\left(\sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t)) + \sqrt{B_\star \sum_{t=1}^{T}r(s_t, a_t)\tilde{\iota}_{T,B,\delta} + B_\star \tilde{\iota}_{T,B,\delta}}\right). \quad (25)$$

By Lemma 16,

$$\sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t))$$
$$\leq O\left(\sqrt{SA\sum_{t=1}^{T}\mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B,\delta} + BS^2 A\tilde{\iota}_{T,B,\delta}}\right)$$
$$\leq O\left(\sqrt{B_\star SA\sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t))\tilde{\iota}_{T,B,\delta} + B_\star SA\sum_{t=1}^{T}r(s_t, a_t)\tilde{\iota}_{T,B,\delta} + BS^2 A\tilde{\iota}_{T,B,\delta}}\right)$$

where in the last inequality we again use Lemma 20. Solving for $\sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t))$, we get

$$\sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t)) \leq O\left(\sqrt{B_\star SA\sum_{t=1}^{T}r(s_t, a_t)\tilde{\iota}_{T,B,\delta} + BS^2 A\tilde{\iota}_{T,B,\delta}}\right).$$

Using this in Eq. (25), we get

$$KV_\star - \sum_{t=1}^{T}r(s_t, a_t) \leq O\left(\sqrt{B_\star SA\sum_{t=1}^{T}r(s_t, a_t)\tilde{\iota}_{T,B,\delta} + BS^2 A\tilde{\iota}_{T,B,\delta}}\right).$$

If $\sum_{t=1}^{T}r(s_t, a_t) \geq KV_\star$, we have $KV_\star - \sum_{t=1}^{T}r(s_t, a_t) \leq 0$; if $\sum_{t=1}^{T}r(s_t, a_t) \leq KV_\star$, we can further bound the $\sum_{t=1}^{T}r(s_t, a_t)$ term on the right-hand side above by $KV_\star$. In both cases, we have

$$KV_\star - \sum_{t=1}^{T}r(s_t, a_t) \leq O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,B,\delta}} + BS^2 A\tilde{\iota}_{T,B,\delta}\right).$$

■

Lemma 21 *Let $r(\cdot, \cdot) \geq 0$. With probability at least $1 - \delta$, for all $K \geq 1$, with T being the total number of steps in K episodes,*

$$\text{Reg}_K \geq -O\left(\sqrt{V_\star B_\star K \ln(T/\delta)} + B_\star \ln(T/\delta)\right).$$

Proof

$$\text{Reg}_K = \sum_{k=1}^{K}\left(V^\star(s_{\text{init}}) - \sum_{t=t_k}^{e_k} r(s_t, a_t)\right)$$

$$= \sum_{t=1}^{T}\left(V^\star(s_t) - V^\star(s'_t) - r(s_t, a_t)\right)$$

$$= \sum_{t=1}^{T}\left(V^\star(s_t) - Q^\star(s_t, a_t)\right) + \sum_{t=1}^{T}(P_t V^\star - V^\star(s'_t))$$

$$\geq \sum_{t=1}^{T}\left(V^\star(s_t) - Q^\star(s_t, a_t)\right) - O\left(\sqrt{\sum_{t=1}^{T} \mathbb{V}(P_t, V^\star) \ln(T/\delta)} + B_\star \ln(T/\delta)\right).$$

By [Lemma 20](#), we can further lower bound the above expression by

$$\sum_{t=1}^{T}\left(V^\star(s_t) - Q^\star(s_t, a_t)\right)$$

$$- O\left(\sqrt{B_\star \sum_{t=1}^{T}(V^\star(s_t) - Q^\star(s_t, a_t)) \ln(T/\delta) + B_\star \sum_{t=1}^{T} r(s_t, a_t) \ln(T/\delta)} + B_\star \ln(T/\delta)\right)$$

$$\geq \sum_{t=1}^{T}\left(V^\star(s_t) - Q^\star(s_t, a_t)\right) - \frac{1}{2}\sum_{t=1}^{T}\left(V^\star(s_t) - Q^\star(s_t, a_t)\right)$$

$$- O\left(\sqrt{B_\star \sum_{t=1}^{T} r(s_t, a_t) \ln(T/\delta)} + B_\star \ln(T/\delta)\right) \qquad\qquad \text{(AM-GM)}$$

$$\geq -O\left(\sqrt{B_\star \sum_{t=1}^{T} r(s_t, a_t) \ln(T/\delta)} + B_\star \ln(T/\delta)\right). \qquad\qquad (V^\star(s_t) - Q^\star(s_t, a_t) \geq 0)$$

Hence we have

$$\text{Reg}_K = KV_\star - \sum_{t=1}^{T} r(s_t, a_t) \geq -O\left(\sqrt{B_\star \sum_{t=1}^{T} r(s_t, a_t) \ln(T/\delta)} + B_\star \ln(T/\delta)\right).$$

Solving for $\sum_{t=1}^{T} r(s_t, a_t)$, we get

$$KV_\star - \sum_{t=1}^{T} r(s_t, a_t) \geq -O\left(\sqrt{V_\star B_\star K \ln(T/\delta)} + B_\star \ln(T/\delta)\right).$$

27

■

Proof [Lemma 7] We consider a particular $i_\star \in \{1, \ldots, \lceil \log_2 U \rceil\}$ such that $2^{i-1} \leq V_\star < 2^i$. By the assumption that $1 \leq V_\star \leq U$, such $i_\star$ always exists. Notice that in the $i_\star$-th for-loop, the input is $B = 2^i \zeta \geq V_\star \zeta \geq B_\star$. Thus, according to Theorem 6, when the $i_\star$-th for-loop terminates, we have with probability at least $1 - \delta'$,

$$V_\star - \hat{r}_{i_\star} \leq \frac{1}{N} \times c \times \left(\sqrt{V_\star B_\star S A N \tilde{\iota}_{M,B,\delta'}} + BS^2 A \tilde{\iota}_{M,B,\delta'} \right)$$

$$= c \times \left(\sqrt{\frac{V_\star B_\star S A \tilde{\iota}_{M,B,\delta'}}{N}} + \frac{BS^2 A \tilde{\iota}_{M,B,\delta'}}{N} \right)$$

$$\leq c \times \left(\sqrt{\frac{V_\star B_\star}{16 c^2 \zeta S}} + \frac{2^i \zeta}{16 c^2 \zeta} \right)$$

(because the termination condition is $N \geq 16 c^2 \zeta S^2 A \tilde{\iota}_{M,B,\delta'}$)

$$\leq \frac{1}{4} V_\star + \frac{1}{8} V_\star, \qquad \text{(by the assumptions } B_\star \leq \zeta V_\star \text{ and } 2^{i-1} \leq V_\star)$$

which implies $\hat{r}_{i_\star} \geq \frac{1}{2} V_\star$.

Next, we consider an arbitrary $i \in \{1, \ldots, \lceil \log_2 U \rceil\}$. Because $V_\star$ is the expected reward of the optimal policy, in every episode, $V_\star$ is larger than the expected reward of the learner. By Lemma 21, we have with probability at least $1 - \delta'$, for all N,

$$V_\star - \hat{r}_i \geq -\frac{1}{N} \times c \times \left(\sqrt{V_\star B_\star N \ln(M/\delta')} + B_\star S^2 A \ln(M/\delta') \right)$$

$$\geq -\frac{1}{2} V_\star, \qquad \text{(by the same calculation as above and noticing that } \ln(M/\delta') \leq \tilde{\iota}_{M,B,\delta'})$$

which implies $\hat{r}_i \leq \frac{3}{2} V_\star$. With an union bound over i, the inequality holds for all i with probability at least $1 - \delta$. Combining the arguments, we conclude that with probability at least $1 - 2\delta$,

$$\frac{1}{2} V_\star \leq \max_i \{\hat{r}_i\} \leq \frac{3}{2} V_\star$$

The lemma is proven by noticing that $\hat{V} = 2 \max_i \{\hat{r}_i\}$. ■

Proof [Theorem 8] We first consider the case when $\frac{B_\star}{V_\star} \leq \zeta$ and $V_\star \leq U$.

Let $\{N_i\}_{i=1,2,\ldots,\lceil \log_2 U \rceil}$ be the number of episodes spent in the i-th for-loop in Algorithm 2. Thus, the total number of episodes the learner spends to estimate $\hat{V}$ is

$$\sum_{i=1}^{\lceil \log_2 U \rceil} N_i = O \left(\sum_{i=1}^{\lceil \log_2 U \rceil} \zeta S^2 A \tilde{\iota}_{M_i, 2^i \zeta, \delta'} \right)$$

where M_i is the number of steps spent in the i-th for-loop in Algorithm 2. By definition, $\tilde{\iota}_{M_i, 2^i \zeta, \delta'} = O((\ln(SA/\delta') + \ln\ln(2^i \zeta M_i)) \times \ln M_i) = O((\ln(SA/\delta) + \ln\ln T + \ln\ln(\zeta U)) \times \ln T) = O(\tilde{\iota}_{T, \zeta U, \delta})$, and thus

$$\sum_{i=1}^{\lceil \log_2 U \rceil} N_i = O \left(\zeta S^2 A \tilde{\iota}_{T, \zeta U, \delta} \times \ln U \right).$$

For these episodes, we simply bound the per-episode regret by $V_\star$. By Lemma 7, $\hat{V} \in [V_\star, 3V_\star]$, and so the main algorithm (Algorithm 1) is run with $B = \zeta\hat{V} \geq \zeta V_\star \geq B_\star$. The regret incurred when running Algorithm 1 with $B = \zeta\hat{V}$, according to Theorem 6, is thus upper bound by

$$O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + \zeta\hat{V}S^2A\tilde{\iota}_{T,\zeta U,\delta}\right) \leq O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + \zeta V_\star S^2A\tilde{\iota}_{T,\zeta U,\delta}\right).$$

Overall, the regret (including the estimation part and the main algorithm part), is upper bounded by

$$O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + \zeta V_\star S^2A\tilde{\iota}_{T,\zeta U,\delta}\ln U\right).$$

We remind that this is the regret bound we can achieve when $\frac{B_\star}{V_\star} \leq \zeta$ and $V_\star \leq U$. If either of them does not hold, we simply bound the total regret by $KV_\star$. Hence, the overall regret without these two assumptions is upper bounded by

$$O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + \zeta V_\star S^2A\tilde{\iota}_{T,\zeta U,\delta}\ln U\right) + KV_\star\mathbb{I}\left\{\frac{B_\star}{V_\star} > \zeta\right\} + KV_\star\mathbb{I}\{V_\star > U\}.$$

Case 1. U is a known absolute upper bound for $V_\star$ In this case, $\mathbb{I}\{V_\star > U\} = 0$, and we have

$$\text{Reg}_K = O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + \zeta V_\star S^2A\tilde{\iota}_{T,\zeta U,\delta}\ln U\right) + KV_\star\mathbb{I}\left\{\frac{B_\star}{V_\star} > \zeta\right\}$$

Let $\alpha > 0$ be a parameter and set $\zeta = \sqrt{K}/\alpha$. Then the last expression can be upper bounded by

$$O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + \frac{V_\star S^2A(\ln U)\sqrt{K}\tilde{\iota}_{T,\zeta U,\delta}}{\alpha} + KV_\star\mathbb{I}\left\{\frac{B_\star}{V_\star} > \frac{\sqrt{K}}{\alpha}\right\}\right)$$

$$\leq O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + \frac{V_\star S^2A(\ln U)\sqrt{K}\tilde{\iota}_{T,\zeta U,\delta}}{\alpha} + \min\left\{\alpha B_\star\sqrt{K}, \frac{\alpha^2 B_\star^2}{V_\star}\right\}\right)$$

where in the last inequality we use two different ways to bound $KV_\star$ under the inequality $\frac{B_\star}{V_\star} > \frac{\sqrt{K}}{\alpha}$: $KV_\star \leq K \times \frac{\alpha B_\star}{\sqrt{K}} = \alpha B_\star\sqrt{K}$, and $KV_\star \leq \left(\frac{\alpha B_\star}{V_\star}\right)^2 V_\star = \frac{\alpha^2 B_\star^2}{V_\star}$. If we set $\alpha = \sqrt{S^2A\ln U}$ and pick up the first term in $\min\{\cdot,\cdot\}$, then we get

$$\text{Reg}_K = O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + V_\star\sqrt{S^2AK\ln U}\tilde{\iota}_{T,\zeta U,\delta} + B_\star\sqrt{S^2AK\ln U}\right)$$

$$\leq O\left(B_\star\sqrt{S^2AK\ln U}\tilde{\iota}_{T,KU,\delta}\right).$$

If we set $\alpha = \sqrt{S^3A\ln U}$ and pick up the second term in $\min\{\cdot,\cdot\}$, then we get

$$\text{Reg}_K = O\left(\sqrt{V_\star B_\star SAK\tilde{\iota}_{T,\zeta U,\delta}} + V_\star\sqrt{SAK\ln U}\tilde{\iota}_{T,\zeta U,\delta} + \frac{B_\star^2 S^3A\ln U}{V_\star}\tilde{\iota}_{T,\zeta U,\delta}\right)$$

$$\leq O\left(\sqrt{V_\star B_\star SAK\ln U}\tilde{\iota}_{T,KU,\delta} + \frac{B_\star^2 S^3A\ln U}{V_\star}\tilde{\iota}_{T,KU,\delta}\right).$$

Case 2. Unknown range of $V_\star$ and set $U = K^{\frac{1}{\epsilon}}$ By the same argument above, if we set $\zeta = \sqrt{K/(S^2 A \ln U)}$, then we have

$$\text{Reg}_K = O\left(B_\star \sqrt{S^2 AK \ln U} \tilde{\iota}_{T,KU,\delta} + KV_\star \mathbb{I}\{V_\star > U\}\right)$$

$$\leq O\left(B_\star \sqrt{\epsilon^{-1} S^2 AK \ln K} \tilde{\iota}_{T,K,\delta\epsilon} + KV_\star \mathbb{I}\left\{V_\star > K^{\frac{1}{\epsilon}}\right\}\right)$$

$$\leq O\left(B_\star \sqrt{\epsilon^{-1} S^2 AK \ln K} \tilde{\iota}_{T,K,\delta\epsilon} + V_\star^{1+\epsilon}\right),$$

and if we set $\zeta = \sqrt{K/(S^3 A \ln U)}$, then

$$\text{Reg}_K = O\left(\sqrt{V_\star B_\star SAK \ln U} \tilde{\iota}_{T,KU,\delta} + \frac{B_\star^2 S^3 A \ln U}{V_\star} \tilde{\iota}_{T,KU,\delta} + KV_\star \mathbb{I}\{V_\star > U\}\right)$$

$$\leq O\left(\sqrt{\epsilon^{-1} V_\star B_\star SAK \ln K} \tilde{\iota}_{T,K,\delta\epsilon} + \frac{\epsilon^{-1} B_\star^2 S^3 A \ln K}{V_\star} \tilde{\iota}_{T,K,\delta\epsilon} + V_\star^{1+\epsilon}\right).$$

$\blacksquare$

Appendix C. Upper Bound for Stochastic Shortest Path

Lemma 22 *Algorithm 4 ensures $Q_t(s,a) \geq Q^\star(s,a)$ for all (s,a) and t with probability at least $1 - \delta$.*

Proof For any $B > 0$, we define $V_{[B]}^\star \in [-B, 0]^{\mathcal{S}}$ to be such that

$$V_{[B]}^\star(s) = \max\{-B, V^\star(s)\}.$$

We use induction to prove the lemma. When $t = 1$, $Q_1(s,a) = 0 \geq Q^\star(s,a)$ for all s, a since we are in the cost setting. Suppose that $0 \geq Q_t(s,a) \geq Q^\star(s,a)$ for all s, a (which implies $0 \geq V_t(s) \geq V^\star(s)$ for all s). Since Q_{t+1} and Q_t only differ in the entry (s_t, a_t), we only need to check $Q_{t+1}(s_t, a_t) \geq Q^\star(s_t, a_t)$. With the definition of b_t and ι_t specified in Line 17 of Algorithm 4, we have

$$r(s_t, a_t) + \bar{P}_t V_t + b_t = r(s_t, a_t) + \bar{P}_t V_t + \max\left\{c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_t, V_t)\iota_t}{n_t}}, \frac{c_2 B_{t+1}\iota_t}{n_t}\right\}$$

$$\geq r(s_t, a_t) + \bar{P}_t V_{[B_{t+1}]}^\star + \max\left\{c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_t, V_{[B_{t+1}]}^\star)\iota_t}{n_t}}, \frac{c_2 B_{t+1}\iota_t}{n_t}\right\}$$

$$\geq r(s_t, a_t) + P_{s_t, a_t} V_{[B_{t+1}]}^\star$$

$$\geq r(s_t, a_t) + P_{s_t, a_t} V^\star$$

$$= Q^\star(s_t, a_t),$$

where the first inequality is because $V_t(s) \geq V^\star(s)$ by the induction hypothesis and $V_t(s) \geq -B_t \geq -B_{t+1}$ by the algorithm, which jointly gives $V_t(s) \geq V_{[B_{t+1}]}^\star(s)$; then using the monotone property

in Lemma 14. Further notice that $Q_t(s_t, a_t) \geq Q^\star(s_t, a_t)$ by the induction hypothesis. Jointly, they imply $Q_{t+1}(s_t, a_t) \geq Q^\star(s_t, a_t)$, which finishes the induction.

We remark that to make the third inequality above hold for all possible $B = \{1, 2, 4, 8, \ldots\}$ with probability $1 - \delta$, we need a union bound over B's. Therefore, in the third inequality above, we actually apply Freedman's inequality for the $B = 2^i$ case with a probability parameter $\delta_i = \frac{\delta}{2i^2} = \frac{\delta}{2(\log_2 B)^2}$ so that $\sum_{i=1}^\infty \delta_i < \delta$. The additional $2(\log_2 B)^2$ factor in the log term is taken cared in the definition of ι_t.

■

Proof [Theorem 11] The proof is similar to that of Theorem 6. We require the combination of the bounds in Lemma 16 and Lemma 20. Notice that the proof of Lemma 16 requires the condition $B \geq B_\star$, the purpose of which is to ensure optimism $Q_t(\cdot, \cdot) \geq Q^\star(\cdot, \cdot)$. For the SSP case, since optimism is ensured by Lemma 22 without requiring $B \geq B_\star$, the conclusion of Lemma 16 still holds, with the B there replaced by B_{T+1} in Algorithm 4 (i.e., the maximum B used in Algorithm 4). Furthermore, with probability at least $1 - O(\delta)$, $B_{T+1} \leq 2B_\star$ according to the arguments in Section 5. Therefore, we have

$$\sum_{t=1}^T (V^\star(s_t) - Q^\star(s_t, a_t)) \leq O\left(\sqrt{SA\sum_{t=1}^T \mathbb{V}(P_t, V^\star)\tilde{\iota}_{T,B_\star,\delta}} + B_\star S^2 A \tilde{\iota}_{T,B_\star,\delta}\right).$$

The bound in Lemma 20 can be directly used. Combining them in a similar way as in the proof of Theorem 6, we get

$$\text{Reg}_K = KV^\star(s_{\text{init}}) - \sum_{t=1}^T r(s_t, a_t) \leq O\left(\sqrt{B_\star SA\sum_{t=1}^T |r(s_t, a_t)|\tilde{\iota}_{T,B_\star,\delta}} + B_\star S^2 A \tilde{\iota}_{T,B_\star,\delta}\right). \tag{26}$$

Recall that in SSP, $V^\star(\cdot) \leq 0$ and $r(\cdot, \cdot) \leq 0$, so Eq. (26) is equivalent to

$$\sum_{t=1}^T |r(s_t, a_t)| - KV_\star \leq O\left(\sqrt{B_\star SA\sum_{t=1}^T |r(s_t, a_t)|\tilde{\iota}_{T,B_\star,\delta}} + B_\star S^2 A \tilde{\iota}_{T,B_\star,\delta}\right).$$

Solving for $\sum_{t=1}^T |r(s_t, a_t)|$, we get

$$\sum_{t=1}^T |r(s_t, a_t)| \leq O\left(KV_\star + B_\star S^2 A \tilde{\iota}_{T,B_\star,\delta}\right).$$

Plugging this back to Eq. (26) finishes the proof. ■

Appendix D. Lower Bound for General Stochastic Path

Proof [Theorem 3] As mentioned in Footnote 1, we can convert between the cases of "deterministic initial state" and "random initial state" by just introducing one additional state. In our lower bound

proof, our construction is based on random initial states, but converting it to deterministic initial state is straightforward. In this proof, we use $P(\cdot|s,a) = P_{s,a}(\cdot)$ to denote the transition probability.

We first consider a MDP with two non-terminal states x, y and a terminal state g. The number of actions is A. Let the initial state distribution be uniform$\{x, y\}$. For all action a, let $r(x, a) = 1$ and $r(y, a) = -1$. Let $\epsilon, \Delta \in (0, \frac{1}{4}]$ be quantities to be determined later. For all action a, let

$$P(\cdot \mid y, a) = (1 - \epsilon)\text{uniform}\{x, y\} + \epsilon \mathbf{1}_g.$$

For all but one single *good* action $a^\star$, let

$$P(\cdot \mid x, a) = (1 - \epsilon)\text{uniform}\{x, y\} + \epsilon \mathbf{1}_g.$$

For the good action $a^\star$, let

$$P(\cdot \mid x, a^\star) = \frac{1 - \epsilon + \Delta}{2}\mathbf{1}_x + \frac{1 - \epsilon - \Delta}{2}\mathbf{1}_y + \epsilon \mathbf{1}_g.$$

First, we calculate the optimal value function of this MDP.

Claim 1 $V^\star(x) = 1 + \frac{\Delta}{\epsilon}\frac{1+\epsilon}{2-\Delta}$, and $Q^\star(x, a) = 1 + \frac{\Delta}{\epsilon}\frac{1-\epsilon}{2-\Delta}$ for $a \neq a^\star$.

Proof [Claim 1] We first show that the optimal policy is to always choose $a^\star$ on state x. We only need to compare two deterministic policies: always choose $a^\star$, or always choose some other action $a \neq a^\star$. For the policy with $\pi(x) = a^\star$, we have

$$V^\pi(x) = r(x, a^\star) + P(x|x, a^\star)V^\pi(x) + P(y|x, a^\star)V^\pi(y) = 1 + \frac{1 - \epsilon + \Delta}{2}V^\pi(x) + \frac{1 - \epsilon - \Delta}{2}V^\pi(y),$$

$$V^\pi(y) = r(y, \cdot) + P(x|y, \cdot)V^\pi(x) + P(y|y, \cdot)V^\pi(y) = -1 + \frac{1 - \epsilon}{2}V^\pi(x) + \frac{1 - \epsilon}{2}V^\pi(y).$$

Solving the equations, we get $V^\pi(x) = 1 + \frac{\Delta}{\epsilon}\frac{1+\epsilon}{2-\Delta}$. On the other hand, if $\pi(x) \neq a$, then by similar calculation, we get $V^\pi(x) = 1$, which is smaller. This shows that the optimal policy is to always choose $a^\star$ on x. Thus, $V^\star(x) = Q^\star(x, a^\star) = 1 + \frac{\Delta}{\epsilon}\frac{1+\epsilon}{2-\Delta}$. Plugging this in another bellman equation

$$V^\star(y) = -1 + \frac{1 - \epsilon}{2}V^\star(x) + \frac{1 - \epsilon}{2}V^\star(y)$$

we get $V^\star(y) = -1 + \frac{\Delta}{\epsilon}\frac{1-\epsilon}{2-\Delta}$, and thus for $a \neq a^\star$,

$$Q^\star(x, a) = 1 + \frac{1 - \epsilon}{2}\left(V^\star(x) + V^\star(y)\right) = 1 + \frac{1 - \epsilon}{2}\left(\frac{\Delta}{\epsilon}\frac{2}{2 - \Delta}\right) = 1 + \frac{\Delta}{\epsilon}\frac{1 - \epsilon}{2 - \Delta}.$$

∎

Next, we follow the proof idea in Rosenberg et al. (2020) and consider a *truncated process*: First, we view the K episodes as a continuous process in which once the learner reaches g, a new state is drawn from the initial distribution and the learner restarts from there. Then we cap the process to make it contain at most $\frac{K}{\epsilon}$ step: if the learner has not finished all K episodes after $\frac{K}{\epsilon}$ steps, then we stop the learner before it finishes all K episodes.

In the original process, we let Reg to be the regret, T_x and T_y be the number of steps the learner visits x and y, respectively, $T_{x,a}$ be the number of steps the learner visits x and chooses action a, and let $T = T_x + T_y$. We define $T_x^-, T_y^-, T_{x,a}^-, T^-$ to be the corresponding quantities in the truncated process. We first show the following claim:

Claim 2 $\mathbb{E}[\text{Reg}] \geq \frac{2\Delta}{2-\Delta} \mathbb{E}[T_x^- - T_{x,a^\star}^-]$.

Proof [Claim 2] We first focus on episode k in the original process. Let the episode starts from $t = t_k$, and the last step in the episode before reaching g is $t = e_k$. Then the expected regret is given by

$$
\mathbb{E}\left[V^\star(s_{t_k}) - \sum_{t=t_k}^{e_k} r(s_t, a_t)\right] = \mathbb{E}\left[V^\star(s_{t_k}) - \sum_{t=t_k}^{e_k}(Q^\star(s_t, a_t) - P_{s_t,a_t} V^\star)\right]
$$

$$
= \mathbb{E}\left[V^\star(s_{t_k}) - \sum_{t=t_k}^{e_k}(Q^\star(s_t, a_t) - V^\star(s_{t+1}))\right]
$$

$$
= \mathbb{E}\left[\sum_{t=t_k}^{e_k}(V^\star(s_t) - Q^\star(s_t, a_t))\right].
$$

Notice that $V^\star(s_t) - Q^\star(s_t, a_t) = 0$ when $s_t = y$ or $(s_t, a_t) = (x, a^\star)$. When $s_t = x$ and $a_t \neq a^\star$, we have $V^\star(s_t) - Q^\star(s_t, a_t) = \frac{\Delta}{\epsilon} \frac{2\epsilon}{2-\Delta} = \frac{2\Delta}{2-\Delta}$ according to Claim 1. Thus, the expected regret in episode k is

$$
\mathbb{E}[\text{Reg}_k] = \mathbb{E}\left[\frac{2\Delta}{2-\Delta} \sum_{t=t_k}^{e_k} \mathbf{1}[s_t = x, a_t \neq a^\star]\right] = \mathbb{E}\left[\frac{2\Delta}{2-\Delta} \sum_{t=t_k}^{e_k}(\mathbf{1}[s_t = x] - \mathbf{1}[s_t = x, a_t = a^\star])\right].
$$

Summing this over episodes, we get

$$
\mathbb{E}[\text{Reg}] = \frac{2\Delta}{2-\Delta} \mathbb{E}[T_x - T_{x,a^\star}].
$$

Using the simple fact that $T_x - T_{x,a^\star} \geq T_x^- - T_{x,a^\star}^-$ finishes the proof. ∎

Claim 3 $\mathbb{E}[T_x^-] \geq \frac{K}{6\epsilon}$.

Proof [Claim 3] Since on every state-action pair (s, a), we have $P(x|s, a) \geq P(y|x, a)$ and the initial distribution is uniform between x and y, it holds that $\mathbb{E}[T_x^-] \geq \mathbb{E}[T_y^-]$. Thus it suffices to show that $\mathbb{E}[T^-] \geq \frac{K}{3\epsilon}$. Note that $\mathbb{E}[T^-] = \mathbb{E}[\min\{T, \frac{K}{\epsilon}\}] \geq \sum_{k=1}^{K} \mathbb{E}[\min\{L_k, \frac{1}{\epsilon}\}]$ where L_k is the length of episode k. If we can show that $L_k \geq \frac{1}{\epsilon}$ with probability at least $\frac{1}{3}$, then we have $\sum_{k=1}^{K} \mathbb{E}[\min\{L_k, \frac{1}{\epsilon}\}] \geq \frac{1}{3} \sum_{k=1}^{K} \min\{\frac{1}{\epsilon}, \frac{1}{\epsilon}\} = \frac{K}{3\epsilon}$. By the definition of the transition kernel, we indeed have $\Pr[L_k \geq \frac{1}{\epsilon}] \geq (1 - \epsilon)^{\frac{1}{\epsilon} - 1} \geq \frac{1}{e} \geq \frac{1}{3}$. This finishes the proof. ∎

The following proof follows the standard lower bound proof for multi-armed bandits (Auer et al., 2002). We create A different instances of MDPs, where in each of them, the choice of the good action is different. We use $\mathbb{P}_a$ and $\mathbb{E}_a$ to denote the probability measure and the expectation under the instance where a is chosen as the good action. We further introduce an instance where

every action behaves the same and there is no good action. We use $\mathbb{P}_{\text{unif}}$ and $\mathbb{E}_{\text{unif}}$ to denote the probability and expectation under this instance.

Claim 4 $\mathbb{E}_a[T_{x,a}^-] \leq \mathbb{E}_{\text{unif}}[T_{x,a}^-] + O\left(\frac{K\Delta}{\epsilon}\sqrt{\mathbb{E}_{\text{unif}}[T_{x,a}^-]}\right).$

Proof [Claim 4] With standard arguments from Auer et al. (2002), because $0 \leq T_{x,a}^- \leq \frac{K}{\epsilon}$, we have

$$\mathbb{E}_a[T_{x,a}^-] - \mathbb{E}_{\text{unif}}[T_{x,a}^-]$$

$$\leq \frac{K}{\epsilon}\|\mathbb{P}_a - \mathbb{P}_{\text{unif}}\|_1$$

$$\leq \frac{K}{\epsilon}\sqrt{\frac{1}{2}\text{KL}(\mathbb{P}_a, \mathbb{P}_{\text{unif}})}$$

$$= \frac{K}{\epsilon}\sqrt{\frac{1}{2}\sum_{t=1}^{K/\epsilon-1}\mathbb{P}_{\text{unif}}[(s_t, a_t) = (x,a)]\text{KL}\left(\text{Multi}\left(\frac{1-\epsilon}{2}, \frac{1-\epsilon}{2}, \epsilon\right), \text{Multi}\left(\frac{1-\epsilon+\Delta}{2}, \frac{1-\epsilon-\Delta}{2}, \epsilon\right)\right)}$$

(by the chain rule of KL; we use $\text{Multi}(a, b, c, \ldots)$ to denote a multinomial distribution)

$$\leq O\left(\frac{K}{\epsilon}\sqrt{\sum_{t=1}^{K/\epsilon}\mathbb{P}_{\text{unif}}[(s_t, a_t) = (x,a)] \times \frac{\Delta^2}{1-\epsilon}}\right)$$

$$= O\left(\frac{K\Delta}{\epsilon}\sqrt{\mathbb{E}_{\text{unif}}[T_{x,a}^-]}\right).$$

∎

Thus, the expected regret is

$$\frac{1}{A}\sum_a \mathbb{E}[\text{Reg}] \geq \frac{\Delta}{A}\sum_a \mathbb{E}_a[T_x^- - T_{x,a}^-]$$

$$\geq \frac{K\Delta}{6\epsilon} - \frac{\Delta}{A}\sum_a \mathbb{E}_a[T_{x,a}^-]$$

$$\geq \frac{K\Delta}{6\epsilon} - \frac{\Delta}{A}\sum_a\left(\mathbb{E}_{\text{unif}}[T_{x,a}^-] + O\left(\frac{K\Delta}{\epsilon}\sqrt{\mathbb{E}_{\text{unif}}[T_{x,a}^-]}\right)\right)$$

$$\geq \frac{K\Delta}{6\epsilon} - \frac{K\Delta}{A\epsilon} - O\left(\frac{\Delta}{A} \times \frac{K\Delta}{\epsilon} \times \sqrt{A\sum_a \mathbb{E}_{\text{unif}}[T_{x,a}^-]}\right)$$

$$\geq \frac{K\Delta}{6\epsilon} - \frac{K\Delta}{A\epsilon} - O\left(\frac{K\Delta^2}{\epsilon}\sqrt{\frac{K}{A\epsilon}}\right)$$

$$= \frac{K\Delta}{\epsilon}\left(\frac{1}{6} - \frac{1}{A} - O\left(\Delta\sqrt{\frac{K}{A\epsilon}}\right)\right).$$

Picking $\Delta = \Theta\left(\sqrt{\frac{A\epsilon}{K}}\right)$, we get that $\mathbb{E}[\text{Reg}] \geq \Omega\left(\sqrt{\frac{AK}{\epsilon}}\right)$.

Notice that for the MDP we constructed, we have

$$R^2 = \sup_{\pi} \mathbb{E}^{\pi} \left[\left(\sum_{t=1}^{\tau} r(s_t, a_t) \right)^2 \right]$$

$$= \sup_{\pi} \mathbb{E}^{\pi} \left[\left(\sum_{t=1}^{\tau} (r(s_t, a_t) - \mathbb{E}_t[r(s_t, a_t)]) \right)^2 + \left(\sum_{t=1}^{\tau} \mathbb{E}_t[r(s_t, a_t)] \right)^2 \right]$$

$$\leq \sup_{\pi} \mathbb{E}^{\pi} \left[\tau + (\Delta \tau)^2 \right]$$

$$\leq \frac{1}{\epsilon} + \frac{2\Delta^2}{\epsilon^2}.$$

We let $\epsilon = \frac{1}{u^2}$. With this choice, $R^2 \leq \frac{1}{\epsilon} + \frac{2\Delta^2}{\epsilon^2} \leq O\left(\frac{1}{\epsilon} + \frac{A}{\epsilon K}\right) \leq O\left(\frac{1}{\epsilon}\right) = O(u^2)$ if we assume $K \geq A$. Therefore, the condition in the lemma the satisfied, and the regret lower bound given by $\Omega\left(\sqrt{\frac{AK}{\epsilon}} \right) = \Omega(u\sqrt{AK})$.

To show the lower bound for general S, we make $\frac{S}{2}$ copies of the MDP we constructed, and let the initial state distribution be uniform over all states. When we run on this aggregated MDP for K episodes, a constant portion of the $\frac{S}{2}$ copies will be visited for $\Theta(\frac{K}{S})$ times. Using the lower bound we just established above (with K replaced by $\frac{K}{S}$), we have that the regret is lower bounded by

$$\Omega\left(S \times u \sqrt{\frac{AK}{S}} \right) = \Omega(u\sqrt{SAK}).$$

Notice that the assumption we need becomes $\frac{K}{S} \geq A$, or $K \geq SA$.

∎

Proof [Theorem 4] We first prove this theorem for an MDP with two non-terminal state x, y and a terminal state g. The number of actions is A. The construction is similar to that in Theorem 3. On state x, there is potentially a good action $a^\star \in [A-1]$. The reward function and transition kernel are chosen as below:

state	action	reward	$\to z$	$\to x$	$\to y$	$\to g$
x	$[A-1]\backslash\{a^\star\}$	1	0	$\frac{1-\epsilon-\Delta}{2}$	$\frac{1-\epsilon+\Delta}{2}$	ϵ
x	$a^\star$	1	0	$\frac{1-\epsilon+\Delta}{2}$	$\frac{1-\epsilon-\Delta}{2}$	ϵ
x	A	1	0	0	0	1
y	$[A]$	-1	0	$\frac{1-\epsilon}{2}$	$\frac{1-\epsilon}{2}$	ϵ

The choices of ϵ, Δ satisfy $\epsilon, \Delta \in (0, \frac{1}{8}]$ and $\Delta^2 \leq \epsilon$.

We consider two cases: one with the good action $a^\star$, and the other without the good action. We first find the optimal policy in each case. If there is a good action on x, then for the policy that always choose $a^\star$ on x, we have $V^{\pi}(x) = 1 + \frac{1-\epsilon+\Delta}{2} V^{\pi}(x) + \frac{1-\epsilon-\Delta}{2} V^{\pi}(y)$ and $V^{\pi}(y) = -1 + \frac{1-\epsilon}{2} V^{\pi}(x) + \frac{1-\epsilon}{2} V^{\pi}(y)$, which jointly give $V^{\pi}(x) = 1 + \frac{\Delta}{\epsilon} \frac{1+\epsilon}{2-\Delta}$; for the policy that always choose $[A-1]\backslash\{a^\star\}$ on x, we have (by similar calculation) $V^{\pi}(x) = 1 - \frac{\Delta}{\epsilon} \frac{1+\epsilon}{2+\Delta}$; for the policy

35

that always choose A on x, we have $V^\pi(x) = 1$. Clearly, the one that always choose $a^\star$ on x gives the highest expected total reward, so it is the optimal policy in this case. Therefore, when there is a good action,

$$V^\star(x) = Q^\star(x, a^\star) = 1 + \frac{\Delta}{\epsilon} \frac{1+\epsilon}{2-\Delta}, \tag{27}$$

$$V^\star(y) = -1 + \frac{1-\epsilon}{2} V^\star(x) + \frac{1-\epsilon}{2} V^\star(y) = -1 + \frac{\Delta}{\epsilon} \frac{1-\epsilon}{2-\Delta}, \tag{28}$$

$$V^\star(x) - Q^\star(x, a) = Q^\star(x, a^\star) - Q^\star(x, a),$$

$$= \Delta \left(V^\star(x) - V^\star(y) \right) = \Delta \left(2 + \frac{2\Delta}{2-\Delta} \right) \geq 2\Delta, \quad \text{for } a \neq a^\star, A \tag{29}$$

$$V^\star(x) - Q^\star(x, A) = \left(1 + \frac{\Delta}{\epsilon} \frac{1+\epsilon}{2-\Delta} \right) - 1 \geq \frac{\Delta}{2\epsilon}. \tag{30}$$

If there is no good action, then the optimal policy is to always choose action A on state x. In this case, we have

$$V^\star(x) = Q^\star(x, a^\star) = 1, \tag{31}$$

$$V^\star(y) = -1 + \frac{1-\epsilon}{2} V^\star(x) + \frac{1-\epsilon}{2} V^\star(y) = -1, \tag{32}$$

$$V^\star(x) - Q^\star(x, a) = 1 - \left[1 + \frac{1-\epsilon-\Delta}{2} V^\star(x) + \frac{1-\epsilon+\Delta}{2} V^\star(y) \right] = \Delta \quad \text{for } a \neq A \tag{33}$$

We use $\mathbb{P}_a$ and $\mathbb{E}_a$ to denote the probability measure and expectation under the environment where $a \in [A-1]$ is chosen as the good action on x; we use $\mathbb{P}_{\text{unif}}$ and $\mathbb{E}_{\text{unif}}$ to denote the probability and expectation under the environment where there is no good action.

For both kinds of MDPs, we have

$$R^2 = R_{\max}^2 = \sup_\pi \mathbb{E}^\pi \left[\left(\sum_{t=1}^\tau r(s_t, a_t) \right)^2 \right] \qquad (\tau \text{ is the last step before reaching } g)$$

$$= \Theta \left(\sup_\pi \mathbb{E}^\pi \left[\left(\sum_{t=1}^\tau (r(s_t, a_t) - \mathbb{E}_t[r(s_t, a_t)]) \right)^2 + \left(\sum_{t=1}^\tau \mathbb{E}_t[r(s_t, a_t)] \right)^2 \right] \right)$$

$$= \Theta \left(\sup_\pi \mathbb{E}^\pi \left[\tau + (\Delta\tau)^2 \right] \right)$$

$$= \Theta \left(\frac{1}{\epsilon} + \frac{\Delta^2}{\epsilon^2} \right) = \Theta \left(\frac{1}{\epsilon} \right), \tag{34}$$

where in the last equation we use the assumption $\Delta^2 \leq \epsilon$. For the MDP with good action, we have $R_\star = R$ since the optimal policy always choose $a^\star$ on $x^\star$; for the MDP without good action, $R_\star = \Theta(1)$ since the optimal policy will take action A on all state x, and directly go to g.

Like in the proof of Theorem 3, we consider the truncated process where the total number of steps is truncated to $\frac{K}{\epsilon}$ (and ignoring the regret incurred later). In the truncated process, we let T_x to denote the number of times the learner visits x, and $T_{x,a}$ denote the number of times the learner visits x and choose action a.

With all the calculations above, we have the following:

$$\mathbb{E}_{\text{unif}}[\text{Reg}_K] \geq \mathbb{E}\left[\sum_{t=1}^{\frac{K}{\epsilon}} (V^\star(s_t) - Q^\star(s_t, a_t))\right] \geq \mathbb{E}\left[(T_x - T_{x,A})\Delta\right], \tag{35}$$

where we use Eq. (33). On the other hand,

$$\mathbb{E}_a[\text{Reg}_K] \geq \mathbb{E}_a\left[T_{x,A} \times \frac{\Delta}{2\epsilon} + (T_x - T_{x,a} - T_{x,A}) \times 2\Delta\right] \geq \mathbb{E}_a\left[T_{x,A} \times \frac{\Delta}{4\epsilon} + (T_x - T_{x,a}) \times 2\Delta\right], \tag{36}$$

where we use Eq. (29) and Eq. (30) and that $\epsilon \leq \frac{1}{8}$.

By the same arguments as in Claim 4 in the proof of Theorem 3, we have that for $a \in [A-1]$,

$$\left|\mathbb{E}_a[T_x - T_{x,a}] - \mathbb{E}_{\text{unif}}[T_x - T_{x,a}]\right| \leq O\left(\frac{K\Delta}{\epsilon}\sqrt{\mathbb{E}_{\text{unif}}[T_{x,a}]}\right) \tag{37}$$

and

$$\left|\mathbb{E}_{\text{unif}}[T_{x,A}] - \mathbb{E}_a[T_{x,A}]\right| \leq O\left(K\Delta\sqrt{\mathbb{E}_{\text{unif}}[T_{x,a}]}\right) \tag{38}$$

where we use that $T_x - T_{x,a} \in [0, \frac{K}{\epsilon}]$ and $T_{x,A} \in [0, K]$. In an environment where $a^\star$ is chosen randomly from $[A-1]$, the expected regret is

$$\frac{1}{A-1}\sum_{a=1}^{A-1} \mathbb{E}_a[\text{Reg}_K]$$

$$\geq \frac{1}{A-1}\sum_{a=1}^{A-1} \mathbb{E}_a\left[\frac{T_{x,A}}{4\epsilon}\Delta + 2(T_x - T_{x,a})\Delta\right] \qquad \text{(by Eq. (36))}$$

$$= \frac{\Delta}{4\epsilon}\frac{1}{A-1}\sum_{a=1}^{A-1}\mathbb{E}_a[T_{x,A}] + 2\Delta \times \frac{1}{A-1}\sum_{a=1}^{A-1}\mathbb{E}_a[T_x - T_{x,a}]$$

$$\geq \frac{\Delta}{4\epsilon}\frac{1}{A-1}\sum_{a=1}^{A-1}\left(\mathbb{E}_{\text{unif}}[T_{x,A}] - O\left(K\Delta\sqrt{\mathbb{E}_{\text{unif}}[T_{x,a}]}\right)\right)$$

$$\qquad + 2\Delta \times \frac{1}{A-1}\sum_{a=1}^{A-1}\left(\mathbb{E}_{\text{unif}}[T_x - T_{x,a}] - O\left(\frac{K\Delta}{\epsilon}\sqrt{\mathbb{E}_{\text{unif}}[T_{x,a}]}\right)\right)$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Eq. (37) and Eq. (38))}$$

$$\geq \frac{\Delta}{4\epsilon}\mathbb{E}_{\text{unif}}[T_{x,A}] + 2\Delta\mathbb{E}_{\text{unif}}[T_x] - \frac{2\Delta}{A-1}\mathbb{E}_{\text{unif}}[T_x - T_{x,A}] - O\left(\frac{K\Delta^2}{\epsilon}\sqrt{\frac{1}{A-1}\sum_{a=1}^{A-1}\mathbb{E}_{\text{unif}}[T_{x,a}]}\right)$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(AM-GM)}$$

$$\geq \frac{\Delta}{4\epsilon}\mathbb{E}_{\text{unif}}[T_{x,A}] + \Delta\mathbb{E}_{\text{unif}}[T_x] - O\left(\frac{K\Delta^2}{\epsilon}\sqrt{\frac{1}{A-1}\sum_{a=1}^{A-1}\mathbb{E}_{\text{unif}}[T_{x,a}]}\right).$$

Before continuing, we prove a property:

Claim 1 $\mathbb{E}[T_x] \geq \frac{K-2\mathbb{E}[T_{x,A}]}{24\epsilon}$.

Proof [Claim 1] We focus on a single episode. Let N_x be the total number of steps the learner visits x in an episode, $N_{x,A}$ be the number of steps the learner visits x and chooses A. Clearly, $N_{x,A} \leq 1$ since once the learner chooses A, the episodes ends.

We prove the following statement: if $\mathbb{E}[N_{x,A}] \leq \frac{1}{2}$, then $\mathbb{E}[N_x] \geq \frac{1}{24\epsilon}$. This can be seen by the following: let $\mathcal{E}_A$ be the event that in the episode, the learner *ever* chooses action A when visiting x, and let $\mathcal{E}'_A$ be its complement event (i.e., replacing *ever* by *never*). Then

$$\Pr\left[N_x \geq \frac{1}{3\epsilon}\right] \geq \Pr\left[\mathcal{E}'_A\right] \times \Pr\left[N_x \geq \frac{1}{3\epsilon} \,\Big|\, \mathcal{E}'_A\right]$$

Notice that $\Pr[\mathcal{E}_A] = \mathbb{E}[N_{x,A}] \leq \frac{1}{2}$, so $\Pr[\mathcal{E}'_A] \geq \frac{1}{2}$. Then notice that

$$\Pr\left[N_x \geq \frac{1}{3\epsilon} \,\Big|\, \mathcal{E}'_A\right] \geq (1 - 3\epsilon)^{\frac{1}{3\epsilon}} \geq \left(1 - \frac{1}{2}\right)^2 = \frac{1}{4}$$

because every time the learner select an action in $[A - 1]$, with probability at least

$$\frac{1 - \epsilon - \Delta}{2} + \frac{1 - \epsilon + \Delta}{2}\left(\frac{1 - \epsilon}{2} + \left(\frac{1 - \epsilon}{2}\right)^2 + \cdots\right) = \frac{1 - \epsilon - \epsilon\Delta}{1 + \epsilon} \geq 1 - 3\epsilon$$

he will visit state x again. Hence, $\mathbb{E}[N_x] \geq \frac{1}{3\epsilon}\Pr\left[N_x \geq \frac{1}{3\epsilon}\right] \geq \frac{1}{3\epsilon} \times \frac{1}{2} \times \frac{1}{4} = \frac{1}{24\epsilon}$.

Now we consider all K episodes, and denote $N_x^{(k)}$, $N_{x,a}^{(k)}$ denotes the visitation counts that correspond to episode k. By the discussion above, we have

$$\mathbb{E}\left[T_x\right] = \sum_{k=1}^{K}\mathbb{E}\left[N_x^{(k)}\right] \geq \sum_{k=1}^{K}\frac{1}{24\epsilon}\mathbb{I}\left\{\mathbb{E}\left[N_{x,A}^{(k)}\right] \leq \frac{1}{2}\right\}$$

$$= \frac{K}{24\epsilon} - \sum_{k=1}^{K}\frac{1}{24\epsilon}\mathbb{I}\left\{\mathbb{E}\left[N_{x,A}^{(k)}\right] > \frac{1}{2}\right\}$$

$$\geq \frac{K}{24\epsilon} - \sum_{k=1}^{K}\frac{2}{24\epsilon}\mathbb{E}\left[N_{x,A}^{(k)}\right]$$

$$= \frac{K}{24\epsilon} - \frac{1}{12\epsilon}\mathbb{E}[T_{x,A}].$$

∎

With Claim 1, we continue with the previous calculation:

$$\frac{1}{A-1}\sum_{a=1}^{A-1}\mathbb{E}_a[\text{Reg}_K] \geq \frac{\Delta}{4\epsilon}\mathbb{E}_{\text{unif}}[T_{x,A}] + \Delta\mathbb{E}_{\text{unif}}[T_x] - O\left(\frac{K\Delta^2}{\epsilon}\sqrt{\frac{1}{A-1}\sum_{a=1}^{A-1}\mathbb{E}_{\text{unif}}[T_{x,a}]}\right)$$

$$\geq \frac{\Delta}{12\epsilon}\mathbb{E}_{\text{unif}}[T_{x,A}] + \Delta\mathbb{E}_{\text{unif}}[T_x] - O\left(\frac{K\Delta^2}{\epsilon}\sqrt{\frac{1}{A}\mathbb{E}_{\text{unif}}[T_x - T_{x,A}]}\right)$$

$$\geq \frac{K\Delta}{24\epsilon} - O\left(\frac{K\Delta^2}{\epsilon}\sqrt{\frac{1}{A}\mathbb{E}_{\text{unif}}[T_x - T_{x,A}]}\right)$$

Suppose that the algorithm can ensure $\mathbb{E}[\text{Reg}_K] \leq O(u\sqrt{AK})$ for all instances with $R_{\max} \leq u$. By the bound in Eq. (34), there exists universal constants $c_2 > 0$ and a term c_1 that only involves logarithmic factors such that

$$\frac{K\Delta}{24\epsilon} - \frac{c_2 K\Delta^2}{\epsilon}\sqrt{\frac{1}{A}\mathbb{E}_{\text{unif}}[T_x - T_{x,A}]} \leq c_1\sqrt{\frac{1}{\epsilon}AK}$$

We pick $\Delta = 48c_1\sqrt{\frac{\epsilon A}{K}}$ (one can verify that this satisfies $\Delta^2 \leq \epsilon$ as long as $K \geq \Omega(c_1^2 A)$). Then the inequality above reads

$$2c_1\sqrt{\frac{1}{\epsilon}AK} - 48^2 c_1^2 c_2\sqrt{A\mathbb{E}_{\text{unif}}[T_x - T_{x,A}]} \leq c_1\sqrt{\frac{1}{\epsilon}AK},$$

which is equivalent to

$$\mathbb{E}_{\text{unif}}[T_x - T_{x,A}] \geq \frac{K}{48^4 c_1^2 c_2^2 \epsilon}.$$

This, together with Eq. (35), implies

$$\mathbb{E}_{\text{unif}}[\text{Reg}_K] \geq \frac{K\Delta}{48^4 c_1^2 c_2^2 \epsilon} = \frac{1}{48^3 c_1 c_2^2}\sqrt{\frac{AK}{\epsilon}}. \tag{39}$$

Recall that unif specifies the environment where there is no good action, and in this case $R_\star = \Theta(1)$. However, the bound Eq. (39) scales with $R = R_{\max} = \Theta(\sqrt{1/\epsilon}) \gg R_\star$.

Now we generalize our construction to general number of states S. We construct an MDP with S non-terminating states that consists of $\frac{S}{2}$ copies of the two-state MDP we just constructed, and equip every state x with two additional actions. To connect these $\frac{S}{2}$ copies, we create a balanced binary tree with $\frac{S}{2}$ nodes; each node of the tree is the x in the two-state MDP. The two additional actions on every state x will lead to a reward of zero and deterministic transitions to the left and the right child of the node, respectively. Furthermore, we only let at most one of the copies to have the optimal action $a^\star$. The initial state for this tree-structured MDP is its root.

Below, we argue that this tree-structured MDP is at least as hard as the original two-state MDP with $\Theta(SA)$ actions (to create this two-state MDP, we let its actions on state x be the union over the actions on all states x in the tree-structured MDPs; similar for y). We can see that for any algorithm in the tree-structured MDP, there is a corresponding algorithm for the two-state MDP that achieves the same expected reward. Besides, the expected reward for the optimal policy on the tree-structured MDP and the two-state MDP are the same. Therefore, our lower for general S can be simply obtained through the two-state construction with SA actions. This finishes the proof. ■

state	action	$\to x$	$\to y$	$\to g$
x	b	0	$\frac{\sqrt{A}}{c\sqrt{K}\ln^2(vK)}$	$1 - \frac{\sqrt{A}}{c\sqrt{K}\ln^2(vK)}$
x	$[A]\backslash\{b\}$	$1 - \frac{1}{v}$	0	$\frac{1}{v}$
y	$a^\star$	0	$1 - \frac{\sqrt{A}}{2cv\sqrt{K}\ln^2(vK)}$	$\frac{\sqrt{A}}{2cv\sqrt{K}\ln^2(vK)}$
y	$[A]\backslash\{a^\star\}$	0	$\frac{1}{2}$	$\frac{1}{2}$

Table 2: Transition kernel for Theorem 9

Appendix E. Lower Bound for Stochastic Longest Path

Proof [Theorem 9] In this proof, we use $P(\cdot|s,a) = P_{s,a}(\cdot)$ to denote transition probability. We first prove this theorem for $S = 2$ (excluding goal state). We assume that the regret bound claimed by the algorithm for the $V_\star = B_\star \leq v$ case is

$$\mathbb{E}[\text{Reg}_K] \leq cv\sqrt{AK} \tag{40}$$

for some c that only involves logarithmic terms.

We create an SLP with an two non-terminal states x, y, a terminal state g, and A actions $\{1, 2, \ldots, A\}$. The initial state is x. The reward function is a constant 1 for all actions on all non-terminating state (i.e., the total reward is the total number of steps before reaching g). On state x, there is a special action b such that $P(g|x,b) = 1 - \frac{\sqrt{A}}{c\sqrt{K}\ln^2(vK)}$ and $P(y|x,b) = \frac{\sqrt{A}}{c\sqrt{K}\ln^2(vK)}$; for all other actions $a \in [A]\backslash\{b\}$, $P(g|x,a) = P(x|x,a) = \frac{1}{2}$. On state y, there is potentially a good action $a^\star$ such that $P(g|y,a^\star) = \frac{\sqrt{A}}{2cv\sqrt{K}\log^2 K}$ and $P(y|y,a^\star) = 1 - \frac{\sqrt{A}}{2cv\sqrt{K}\ln^2(vK)}$; for other actions $a \in [A]\backslash\{a^\star\}$, $P(g|y,a) = P(y|y,a) = \frac{1}{2}$. The transition kernel is summarized in Table 2.

Let $\mathbb{P}_a$ and $\mathbb{E}_a$ denote the probability measure and expectation in the instance where $a^\star = a$, and let $\mathbb{P}_{\text{unif}}$ and $\mathbb{E}_{\text{unif}}$ denote those in the instance where $a^\star$ does not exist.

Let N_a be the total number of times (in K episodes) the learner chooses action a on state y, N_b be the total number of times the learner chooses action b on state x, and N_y be the total number of times the learner visits state y.

In the environment where $a^\star$ does not exist, the optimal policy on state x is to choose any action in $[A]\backslash\{b\}$. This is because for any policy such that $\pi(x) = b$, we have $V^\pi(x) = 1 + P(y|x,b)V^\pi(y) \leq 3$, and for any policy such that $\pi(x) \in [A]\backslash\{b\}$, we have $V^\pi(x) = v$. Therefore,

$$\mathbb{E}_{\text{unif}}[\text{Reg}_K] = (v-3)\mathbb{E}_{\text{unif}}[N_b] \geq \frac{v}{2}\mathbb{E}_{\text{unif}}[N_b].$$

In this environment, $V^\star(x) = v$ and $V^\star(y) = 2$, and thus $V_\star = B_\star = v$. By the assumption for the algorithm, we have

$$\frac{v}{2}\mathbb{E}_{\text{unif}}[N_b] \leq \mathbb{E}_{\text{unif}}[\text{Reg}_K] \leq cv\sqrt{AK},$$

or $\mathbb{E}_{\text{unif}}[N_b] \leq 2c\sqrt{AK}$.

On the other hand, in the environment where $a^\star = a$, the optimal policy is to choose action b on state x until transitioning to state y and then choose $a^\star$ on y. This is because for this policy, $V^\pi(x) \geq$

$$P(y|x,b)V^\pi(y) = \frac{\sqrt{A}}{c\sqrt{K}\ln^2(vK)} \times \frac{2cv\sqrt{K}\ln^2(vK)}{\sqrt{A}} = 2v, \text{ while for all other policies, } V^\pi(x) \leq v. \text{ For}$$

this environment, we have $V_\star = V^\star(x) = Q^\star(x,b) = 2v$, and $B_\star = 2cv\ln^2(vK)\sqrt{\frac{K}{A}}$.

Next, we bound the KL divergence between the two environments. Note that we have

$$\frac{1}{A}\sum_{a=1}^{A}\mathbb{E}_{\text{unif}}[N_a] = \frac{1}{A}\mathbb{E}_{\text{unif}}[N_y] = \frac{1}{A}P(y|x,b)\mathbb{E}_{\text{unif}}[N_b] = \frac{1}{c\sqrt{AK}\ln^2(vK)}\mathbb{E}_{\text{unif}}[N_b] \leq \frac{2}{\log^2(vK)}$$

where in the last inequality we use $\mathbb{E}_{\text{unif}}[N_b] \leq 2c\sqrt{AK}$ which we just derived above. Hence,

$$\frac{1}{A}\sum_{a=1}^{A}\text{KL}(\mathbb{P}_a, \mathbb{P}_{\text{unif}}) = \frac{1}{A}\sum_{a=1}^{A}\mathbb{E}_{\text{unif}}[N_a]\text{KL}\left(\text{Bernoulli}\left(\frac{\sqrt{A}}{2cv\sqrt{K}\ln^2(vK)}\right), \text{Bernoulli}\left(\frac{1}{2}\right)\right)$$

$$\leq O\left(\frac{1}{A}\sum_{a=1}^{A}\mathbb{E}_{\text{unif}}[N_a]\ln(vK)\right)$$

$$= O\left(\frac{1}{\ln(vK)}\right).$$

We consider the environment where $a^\star$ is chosen uniformly randomly from $[A]$, and denote its probability measure and expectation as $\overline{\mathbb{P}}$ and $\overline{\mathbb{E}}$. Clearly, $\overline{\mathbb{P}} = \frac{1}{A}\sum_{a=1}^{A}\mathbb{P}_a$ and $\overline{\mathbb{E}} = \frac{1}{A}\sum_{a=1}^{A}\mathbb{E}_a$. By the convexity of KL divergence, we have

$$\text{KL}(\overline{\mathbb{P}}, \mathbb{P}_{\text{unif}}) \leq \frac{1}{A}\sum_{a=1}^{A}\text{KL}(\mathbb{P}_a, \mathbb{P}_{\text{unif}}) = O\left(\frac{1}{\ln(vK)}\right).$$

By the same argument as in the proof of Claim 3, Theorem 3, we have

$$\overline{\mathbb{P}}(N_b > K/2) \leq \mathbb{P}_{\text{unif}}(N_b > K/2) + O\left(\sqrt{\text{KL}(\overline{\mathbb{P}}, \mathbb{P}_{\text{unif}})}\right)$$

$$\leq \frac{2}{K} \times \mathbb{E}_{\text{unif}}[N_b] + O\left(\frac{1}{\sqrt{\ln(vK)}}\right)$$

$$\leq O\left(\frac{c\sqrt{A}}{\sqrt{K}} + \frac{1}{\sqrt{\ln(vK)}}\right)$$

Thus, for large enough K, we have $\overline{\mathbb{P}}(N_b > K/2) \leq \frac{1}{2}$ and hence the regret in this environment is bounded by $\Omega(vK)$.

To generalize the result to general number of states, we leave the initial state x unchanged, but replace the state y by a binary tree with $S-1$ nodes with root y_1 and leaves $y_{S/2}, \ldots, y_{S-1}$. The transition probability $P(y_1|x,b)$ takes the value of $P(y|x,b)$ specified in Table 2. For all non-leaf nodes, there are only two actions that can deterministically transition to its two children with zero reward. For all leaf nodes, the transition and reward are same as the y node in Table 2. Among all leaves, there is at most one action on one node being the good action $a^\star$.

Similar to the proof of Theorem 4, it is not hard to see that this MDP is at least as hard as the original two-state MDP with $\Omega(SA)$ actions on y, by letting the action set on y in the two-state

MDP to be the union of actions on the leaves in the tree-structed MDP. This is because the optimal value on these two cases are the same, and every policy in the tree-structured MDP can find a corresponding policy in the two-state MDP with the same expected reward. Hence our lower bound for general S can be obtained by our original constant S construction with $\Theta(SA)$ actions.

∎

Appendix F. Auxiliary Lemmas

Lemma 23 *Let $X_t \in [-c, c]^S$ be in the filtration of $(s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ for some $c > 0$ with $X_t(g) \triangleq 0$. Then with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \mathbb{V}(P_t, X_t) \leq O\left(c \sum_{t=1}^{T} |X_t(s_t) - P_t X_t| + c \sum_{t=1}^{T} \|X_t - X_{t+1}\|_\infty + c^2 \ln(1/\delta)\right).$$

Proof Denote $P_t := P_{s_t, a_t}$.

$$\sum_{t=1}^{T} \mathbb{V}(P_t, X_t)$$

$$= \sum_{t=1}^{T} \left(P_t X_t^2 - (P_t X_t)^2\right)$$

$$= \sum_{t=1}^{T} \left(P_t X_t^2 - X_t(s_t')^2\right) + \sum_{t=1}^{T} \left(X_t(s_t')^2 - X_t(s_t)^2\right) + \sum_{t=1}^{T} \left(X_t(s_t)^2 - (P_t X_t)^2\right)$$

$$\leq O\left(\sqrt{\sum_{t=1}^{T} \mathbb{V}(P_t, X_t^2) \ln(1/\delta)} + c^2 \ln(1/\delta)\right) + \sum_{t=1}^{T}(X_t(s_{t+1})^2 - X_t(s_t)^2) + \sum_{t=1}^{T} \left(X_t(s_t)^2 - (P_t X_t)^2\right)$$

$$\text{(because } X_t(g) = 0)$$

$$\leq O\left(c\sqrt{\sum_{t=1}^{T} \mathbb{V}(P_t, X_t) \ln(1/\delta)} + c^2 \ln(1/\delta)\right) + 2c \sum_{t=1}^{T} \|X_t - X_{t+1}\|_\infty + 2c \sum_{t=1}^{T} |X_t(s_t) - P_t X_t|$$

$$\leq \frac{1}{2} \sum_{t=1}^{T} \mathbb{V}(P_t, X_t) + 2c \sum_{t=1}^{T} \|X_t - X_{t+1}\|_\infty + 2c \sum_{t=1}^{T} |X_t(s_t) - P_t X_t| + O\left(c^2 \ln(1/\delta)\right).$$

$$\text{(AM-GM inequality)}$$

Solving the inequality we get the desired inequality. ∎

Lemma 24 *Let $X_1, X_2, \ldots, X_\tau \subset [0, b]$ be a sequence with a random stopping time τ, where X_i is in the filtration of $\mathcal{F}_i = (X_1, \ldots, X_{i-1})$, for some $b \geq 1$. Suppose that for any i, $\mathbb{E}\left[\sum_{t=i}^{\tau} X_t \mid \mathcal{F}_i\right] \leq B$. Then*

   (a) with probability at least $1 - \delta$, $\sum_{t=1}^{\tau} X_t \leq O((B + b) \ln(1/\delta))$,

   (b) $\mathbb{E}\left[\left(\sum_{t=1}^{\tau} X_t\right)^2\right] \leq O\left((B + b) \ln_+(\frac{B+b}{c}) \mathbb{E}\left[\sum_{t=1}^{\tau} X_t\right] + c^2\right)$ for any $1 \leq c \leq B + b$.

(c) $\mathbb{E}\big[\left(\sum_{t=1}^{\tau} X_t\right)^2 \big] \leq O\left((B+b)^2\right).$

Proof For a sequence $X_1, X_2, \ldots, X_\tau$, define

$$\tau_1 = \min \left\{ n \leq \tau : \ \sum_{t=1}^{n} X_t \geq 2B \right\},$$

and for $m \geq 2$, define

$$\tau_m = \min \left\{ n \leq \tau : \ \sum_{t=\tau_{m-1}+1}^{n} X_t \geq 2B \right\}$$

If such τ_m does not exist, let $\tau_m = \infty$. Naturally, we define $\tau_0 = 0$.

By the condition stated in the lemma and Markov's inequality, we have

$$\Pr\left[\tau_{m+1} < \infty \mid \tau_m < \infty\right] \leq \Pr\left[\sum_{t=\tau_m+1}^{\tau} X_t \geq 2B \right] \leq \frac{1}{2}.$$

Therefore, $\Pr[\tau_m < \infty] \leq 2^{-m}$ and $\Pr[\tau_m = \infty] \geq 1 - 2^{-m}$. Also, notice that by the definition of τ_i, we have $\sum_{t=\tau_{i-1}+1}^{\tau_i} X_t \leq 2B+b$ for all i (otherwise, $\sum_{t=\tau_{i-1}+1}^{\tau_i-1} X_t > 2B$, contradicting the definition of τ_i). Thus, if $\tau_m = \infty$, then

$$\sum_{t=1}^{\tau} X_t \leq \sum_{i=1}^{m-1} \sum_{t=\tau_{i-1}+1}^{\tau_i} X_t + \sum_{t=\tau_{m-1}+1}^{\tau} X_t \leq (2B+b)(m-1) + 2B \leq (2B+b)m.$$

Combining the arguments above, we have the following: for any $\delta < 0.5$ (letting $m = \lceil \log_2(1/\delta) \rceil$), with probability at least $1 - 2^{-m} \geq 1 - \delta$,

$$\sum_{t=1}^{\tau} X_t \leq (2B+b)m = (2B+b)\lceil \log_2(1/\delta) \rceil \leq 8(B+b)\ln(1/\delta).$$

This proves (a). Below we bound the second moment:

$$\mathbb{E}\left[\left(\sum_{t=1}^{\tau} X_t\right)^2 \right]$$

$$\leq \mathbb{E}\left[\left(\sum_{t=1}^{\tau} X_t\right)^2 \ \middle| \ \tau_M = \infty \right] \Pr\left[\tau_M = \infty\right]$$

$$+ \sum_{m=M+1}^{\infty} \mathbb{E}\left[\left(\sum_{t=1}^{\tau} X_t\right)^2 \ \middle| \ \tau_m = \infty \right] \Pr[\tau_{m-1} < \infty, \tau_m = \infty]$$

$$\leq (2B+b)M\mathbb{E}\left[\sum_{t=1}^{\tau} X_t \ \middle| \ \tau_M = \infty \right] + \sum_{m=M+1}^{\infty} \left((2B+b)m\right)^2 \Pr[\tau_{m-1} < \infty]$$

$$\leq 2M(B+b)\mathbb{E}\left[\sum_{t=1}^{\tau} X_t\right] + (2B+b)^2 \sum_{m=M+1}^{\infty} 2^{-m+1}m^2.$$

If we pick $M = \lceil 4\log_2(\frac{2B+b}{c})\rceil$, then $(2B+b)^2 2^{-\frac{M}{2}} \leq c^2$, and the last expression can be further upper bounded by

$$O\left((B+b)\ln_+\left(\frac{B+b}{c}\right)\mathbb{E}\left[\sum_{t=1}^{\tau} X_t\right] + c^2 \sum_{m=M+1}^{\infty} 2^{-\frac{m}{2}+1}m^2\right)$$

$$= O\left((B+b)\ln_+\left(\frac{B+b}{c}\right)\mathbb{E}\left[\sum_{t=1}^{\tau} X_t\right] + c^2\right),$$

which proves (b). (c) is an immediate result of (b) by picking $c = B + b$ and noticing that $\mathbb{E}[\sum_{t=1}^{\tau} X_t] \leq B$. ∎

Lemma 25 (Exercise 5.15 in Lattimore and Szepesvári (2020)) *Let X_t be a real valued random variable in the filtration of $\mathcal{F}_t = (X_1, \ldots X_{t-1})$ such that $\mathbb{E}[X_t \,|\, \mathcal{F}_t] = 0$ and assume $\eta > 0$, $\mathbb{E}[X_t \,|\, \mathcal{F}_t] < \eta^{-1}$ a.s. Then with probability at least $1 - \delta$ for all $0 < t \leq T$,*

$$\sum_{s=1}^{t} X_s \leq \eta \sum_{s=1}^{t} \mathbb{E}[X_s^2 \,|\, \mathcal{F}_s] + \eta^{-1}\ln(\delta^{-1}).$$

Proof The claim is stated for a fixed horizon T in Lattimore and Szepesvári (2020). However, we can define the surrogate random variable

$$X_i' \triangleq X_i \cdot \left(1 - \mathbb{I}\left\{\exists 0 \leq j < i : \sum_{s=1}^{j} X_s > \eta \sum_{s=1}^{j} \mathbb{E}[X_s^2 \,|\, \mathcal{F}_s] + \eta^{-1}\ln(\delta^{-1})\right\}\right).$$

X_i' is adapted to the filtration and it holds with probability $1 - \delta$,

$$\sum_{s=1}^{T} X_s' \leq \eta \sum_{s=1}^{T} \mathbb{E}[X_s'^2 \,|\, \mathcal{F}_s] + \eta^{-1}\ln(\delta^{-1}),$$

which is equivalent to the anytime result for X_i. ∎

Lemma 26 *Let X_t be a real valued random variable in the filtration of $\mathcal{F}_t = (X_1, \ldots X_{t-1})$ such that $\mathbb{E}[X_t \,|\, \mathcal{F}_t] = 0$ and assume $\mathbb{E}[|X_t| \,|\, \mathcal{F}_t] < \infty$ a.s. Then with probability at least $1 - \delta$ uniformly over all $T > 0$,*

$$\sum_{t=1}^{T} X_t \leq 4\sqrt{V_T\left(\ln(\delta^{-1}) + 2\ln\ln(V_T)\right)} + eU_T \ln\left(\frac{2\ln^2(eU_T)}{\delta}\right),$$

where $V_T = \sum_{t=1}^{T} \mathbb{E}[X_t^2 | \mathcal{F}_t]$, $U_T = \max\{1, \max_{t \in [T]} X_t\}$.

Proof Define $Z_t^{(i)} = X_t \cdot \mathbb{I}\{U_t \le \exp(i)\}$, then $Z_t^{(i)}\exp(-i) \le 1$ almost surely. By Exercise 5.15 of Lattimore and Szepesvári (2020), with probability at least $1 - \delta/(2i^2)$, we have

$$\sum_{t=1}^{T} Z_t^{(i)} \le \sum_{t=1}^{T}(Z_t^{(i)} - \mathbb{E}[Z_t^{(i)} \mid \mathcal{F}_t]) \le \exp(-i)\sum_{t=1}^{T}\mathbb{E}\left[\left(Z_t^{(i)}\right)^2 \;\middle|\; \mathcal{F}_t\right] + \exp(i)\ln\left(\frac{2i^2}{\delta}\right).$$

By a union bound, this holds with probability $1-\delta$ uniformly over all $i \ge 1$. Note that $\sum_{t=1}^{T}\mathbb{E}\left[\left(Z_t^{(i)}\right)^2 \mid \mathcal{F}_t\right] \le \sum_{t=1}^{T}\mathbb{E}[X_t^2 \mid \mathcal{F}_t] = V_T$ and for any i such that $\exp(i) \ge U_T$, we have $\sum_{t=1}^{T} Z_t^{(i)} = \sum_{t=1}^{T} X_t$. Hence with probability $1 - \delta$,

$$\begin{aligned}
\sum_{t=1}^{T} X_t &\le \min_{i:\exp(i)\ge U_T} \exp(-i)V_T + \exp(i)\ln\left(\frac{2i^2}{\delta}\right) \\
&\le eU_T\ln\left(\frac{2\ln(eU_T)^2}{\delta}\right) + \min_{i\in\mathbb{Z}}\exp(-i)V_T + \exp(i)\ln\left(\frac{2i^2}{\delta}\right) \\
&\le eU_T\ln\left(\frac{2\ln(eU_T)^2}{\delta}\right) + 2\min_{\gamma>0}\gamma^{-1}V_T + \gamma\ln\left(\frac{2\ln(\gamma)^2}{\delta}\right) \\
&\le eU_T\ln\left(\frac{2\ln(eU_T)^2}{\delta}\right) + 4\sqrt{V_T\ln\left(\frac{\ln^2(V_T)}{\delta}\right)}. \quad \text{(choosing } \gamma = \sqrt{\frac{V_T}{\ln(2\delta^{-1}\ln^2(V_T))}})
\end{aligned}$$

The reasoning for the second inequality is the following. We are computing $\min_{i\ge i_0} f(i) + g(i)$, where $f(i)$ is monotonically decreasing in i and $g(i)$ is monotonically increasing. Let $i^* = \operatorname{argmin}_{i\in\mathbb{Z}} f(i) + g(i)$. If $i^* \ge i_0$, the equation is obviously true. Otherwise the optimal solution is $\min_{i\ge i_0} f(i) + g(i) = f(i_0) + g(i_0) \le f(i^*) + g(i_0)$ by the monotonicity. ∎

Lemma 27 *Let $X_1, \cdots \in \mathbb{R}^S$ be a sequence of i.i.d. random vectors with mean μ, variance Σ such that $\|X_i\|_1 < c$ almost surely. Then with probability at least $1 - 2S\delta$, it holds for all $w \in \mathbb{R}^S$ such that $\|w\|_\infty < C$ and $T > 0$ simultaneously:*

$$\sum_{t=1}^{T}\langle X_t, w\rangle \le 4\sqrt{STw^\top\Sigma w\left(\ln(\delta^{-1}) + 2\ln\ln(Tc^2)\right)} + eScC\ln\left(\frac{2\ln^2(ecC)}{\delta}\right).$$

Proof The proof is extends Lemma 26. Let $v_1, \ldots, v_S$ be the eigenvectors of Σ, then we have with probability $1 - 2S\delta$ for all v_i simultaneously

$$\left|\sum_{t=1}^{T}\langle X_t, v_i\rangle\right| \le 4\sqrt{Tv_i^\top\Sigma v_i\left(\ln(\delta^{-1}) + 2\ln\ln(Tc^2)\right)} + ec\ln\left(\frac{2\ln^2(ec)}{\delta}\right).$$

Let $w = \sum_{i=1}^{S} a_i v_i$, where we know that $\|w\|_2 \le \sqrt{S}C$ and $\sum_{i=1}^{S}|a_i| \le SC$. This implies

$$\sum_{t=1}^{T}\langle X_t, w\rangle \le \sum_{i=1}^{S} 4\sqrt{Ta_i^2 v_i^\top\Sigma v_i\left(\ln(\delta^{-1}) + 2\ln\ln(Tc^2)\right)} + |a_i|ec\ln\left(\frac{2\ln^2(ec)}{\delta}\right)$$

$$\leq 4\sqrt{ST w^\top \Sigma w \left(\ln(\delta^{-1}) + 2\ln\ln(Tc^2)\right)} + ecCS\ln\left(\frac{2\ln^2(ec)}{\delta}\right).$$

∎

Lemma 28 *Let $X_1, \cdots \in [0, B]$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Then with probability at least $1 - \delta$, it holds for all $T \geq 8\ln(2\delta^{-1})$ simultaneously:*

$$\sum_{t=1}^{T} X_t - T\mu \leq \sqrt{8\sum_{i=1}^{T}\left(X_i - \frac{1}{T}\sum_{j=1}^{T}X_j\right)^2 \left(\ln(2\delta^{-1}) + 4\ln\ln(T)\right)} + 12B(\ln(2\delta^{-1}) + 4\ln\ln(T)).$$

Proof By Lemma 26, we have that with probability $1 - \delta/2$, for all $T > 0$ simultaneously,

$$\sum_{t=1}^{T} X_t - T\mu \leq \sqrt{T\sigma^2(\ln(2\delta^{-1}) + 4\ln\ln(T))} + B(\ln(2\delta^{-1}) + 4\ln\ln(T)).$$

Applying the same Lemma to the sequence $Z_i = -(X_i - \mu)^2 + \sigma^2$, we have with probability $1 - \delta/2$ for all $T > 0$ simultaneously

$$T\sigma^2 - \sum_{t=1}^{T}(X_t - \mu)^2$$

$$\leq \sqrt{\sum_{t=1}^{T}\mathbb{E}_{t-1}[((X_t - \mu)^2 - \sigma^2)^2](\ln(2\delta^{-1}) + 4\ln\ln(T)) + \sigma^2(\ln(2\delta^{-1}) + 4\ln\ln(T))}$$

$$\leq B\sqrt{T\sigma^2(\ln(2\delta^{-1}) + 4\ln\ln(T))} + \sigma^2(\ln(2\delta^{-1}) + 4\ln\ln(T))$$

$$\leq \frac{T}{2}\sigma^2 + 5B^2(\ln(2\delta^{-1}) + 4\ln\ln(T)).$$

Finally we have

$$\sum_{t=1}^{T}(X_t - \mu)^2 = \sum_{t=1}^{T}\left(X_t - \frac{1}{T}\sum_{j=1}^{T}X_j\right)^2 + T\left(\mu - \frac{1}{T}\sum_{t=1}^{T}X_t\right)^2.$$

Combining everything and taking a union bounds, leads to with probability $1 - \delta$

$$\sum_{t=1}^{T} X_t - T\mu$$

$$\leq \sqrt{2\left(\sum_{t=1}^{T}\left(X_t - \frac{1}{T}\sum_{j=1}^{T}X_j\right)^2 + \frac{1}{T}\left(\sum_{t=1}^{T}X_t - T\mu\right)^2 + 10B^2\right)(\ln(2\delta^{-1}) + 4\ln\ln(T))}$$

$$\qquad + B(\ln(2\delta^{-1}) + 4\ln\ln(T))$$

$$\leq \left| \sum_{t=1}^{T} X_t - T\mu \right| \sqrt{\frac{2\ln(2\delta^{-1}) + 4\ln\ln(T)}{T}} + \sqrt{2\sum_{t=1}^{T}\left(X_t - \frac{1}{T}\sum_{j=1}^{T}X_j\right)^2} + 6B(\ln(2\delta^{-1}) + 4\ln\ln(T)).$$

For $T \geq 8\ln(2\delta^{-1})$, we have $\frac{2\ln(2\delta^{-1})+4\ln\ln(T)}{T} \leq \frac{1}{4}$ and rearranging finishes the proof. ∎

Lemma 29 (Lemma 30, Chen et al. (2021a)) *Let $\|X\|_\infty \leq C$, then $\mathbb{V}(P, X^2) \leq 4C\mathbb{V}(P, X)$ for any P.*

Appendix G. Weakening the assumption on proper policies

Assumption 1 can be weakened to the following:

Assumption 2 *There exists a policy $\pi^\star \in \Pi^{SD}$ such that*

- $V^{\pi^\star}(s) \geq V^\pi(s)$ *for all $s \in \mathcal{S}$ and $\pi \in \Pi^{HD}$.*

- $T_\star \triangleq \max_s \mathbb{E}^{\pi^\star}[\tau \mid s_1 = s] < \infty$, *where τ is the time index right before reaching g.*

In words, Assumption 2 assumes that there exists an optimal policy that is stationary and proper. If such an optimal policy is not unique, we can take $T_\star$ to be the minimum over all such policies. Notice that the first part of Assumption 2 is sufficient for all our algorithms to work, though the regret bound has a $\ln T$ factor, which could be unbounded. That is why in the main text we introduced the stronger Assumption 1 and upper bound T by the order of $KT_{\max}$. Below we show that with the additional second part of Assumption 2 and the algorithmic trick introduced in Tarbouriech et al. (2021b), the $T_{\max}$ dependency can be replaced by $T_\star$.

Assuming that an order optimal bound of $T_\star$ is known, the agent modifies the MDP by modifying all rewards $\tilde{r}(s, a) = r(s, a) - \frac{1}{KT_\star}$. The value of the optimal policy in the modified MDP is smaller than that in the original MDP by at most $\frac{1}{K}$. It is then sufficient to bound the regret for the modified MDP, since

$$\sum_{k=1}^{K}\left(V^\star(s_{\text{init}}) - \sum_{t=t_k}^{e_k} r(s_t, a_t)\right) \leq 1 + \sum_{k=1}^{K}\left(\tilde{V}^\star(s_{\text{init}}) - \sum_{t=t_k}^{e_k}\tilde{r}(s_t, a_t)\right).$$

For the modified MDP, we can show that the total time horizon T is bounded. By a trivial bound of $\mathbb{V}(P_t, V^\star) \leq B_\star^2$, combined with Lemma 16, we get that with probability at least $1 - \delta$,

$$\sum_{k=1}^{K}\left(\tilde{V}^\star(s_{\text{init}}) - \sum_{t=t_k}^{e_k}\tilde{r}(s_t, a_t)\right) \leq O\left(\sqrt{SAB_\star^2 T \tilde{\iota}_{T,B,\delta}} + BS^2 A\tilde{\iota}_{T,B,\delta}\right).$$

Notice that $K\tilde{V}^\star(s_{\text{init}}) \geq KV^\star(s_{\text{init}}) - 1$ and $\sum_{t=1}^{T}\tilde{r}(s_t, a_t) = -\frac{T}{KT_\star} + \sum_{t=1}^{T} r(s_t, a_t) \leq -\frac{T}{KT_\star} + KV^\star(s_{\text{init}}) + R\sqrt{K\ln(KR/\delta)} + R_{\max}\ln(KR_{\max}/\delta)$ with probability $\geq 1 - O(\delta)$ by Lemma 26 (notice that $\mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t)\right] \leq KV^\star(s_{\text{init}})$). Rearranging by T leads to with probability at least $1 - O(\delta)$,

$$T \leq O\left(\text{poly}(S, A, B, B_\star, R, R_{\max}, K, T_\star, \delta^{-1})\right).$$

This upper bound on T helps us to replace the $\ln T$ dependency in the regret by a log term that only involves algorithm-independent quantities.

If an order optimal bound of $T_\star$ is not known, we can follow the arguments in Tarbouriech et al. (2021b) that sets $\tilde{r}(s, a) = r(s, a) - \frac{1}{K^n}$ for some $n \gg 1$. By this, we can also remove the dependency on $T_{\max}$, with the price of an additional $\frac{T_\star}{K^{n-1}}$ regret.