Reaching Goals is Hard: Settling the Sample Complexity of the Stochastic Shortest Path

Liyu Chen[*] LIYUC@USC.EDU
University of Southern California

Andrea Tirinzoni TIRINZONI@META.COM
Meta

Matteo Pirotta PIROTTA@META.COM
Meta

Alessandro Lazaric LAZARIC@META.COM
Meta

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We study the sample complexity of learning an ϵ-optimal policy in the Stochastic Shortest Path (SSP) problem. We first derive sample complexity bounds when the learner has access to a generative model. We show that there exists a worst-case SSP instance with S states, A actions, minimum cost $c_{\min}$, and maximum expected cost of the optimal policy over all states $B_\star$, where any algorithm requires at least $\Omega(SAB_\star^3/(c_{\min}\epsilon^2))$ samples to return an ϵ-optimal policy with high probability. Surprisingly, this implies that whenever $c_{\min} = 0$ an SSP problem may not be learnable, thus revealing that learning in SSPs is strictly harder than in the finite-horizon and discounted settings. We complement this result with lower bounds when prior knowledge of the hitting time of the optimal policy is available and when we restrict optimality by competing against policies with bounded hitting time. Finally, we design an algorithm with matching upper bounds in these cases. This settles the sample complexity of learning ϵ-optimal polices in SSP with generative models.

We also initiate the study of learning ϵ-optimal policies without access to a generative model (i.e., the so-called best-policy identification problem), and show that sample-efficient learning is impossible in general. On the other hand, efficient learning can be made possible if we assume the agent can directly reach the goal state from any state by paying a fixed cost. We then establish the first upper and lower bounds under this assumption.

Finally, using similar analytic tools, we prove that horizon-free regret is impossible in SSPs under general costs, resolving an open problem in (Tarbouriech et al., 2021c).

Keywords: Stochastic Shortest Path, Markov Decision Process, PAC Learning

1. Introduction

The Stochastic Shortest Path (SSP) formalizes the problem of finding a policy that reaches a designated goal state while minimizing the cost accumulated over time. This setting subsumes many important application scenarios, such as indoor and car navigation, trade execution, and robotic manipulation. The SSP is strictly more general than the popular finite-horizon and discounted settings (see e.g., Bertsekas, 2012; Tarbouriech et al., 2020a) and it poses specific challenges due to the fact that no horizon is explicitly prescribed in the definition of the problem. In fact, different

[*] Research conducted when the author was an intern at Meta.

policies may have varying hitting times to the goal, e.g., the optimal policy may not be the policy with smallest hitting time whereas some policies may not even reach the goal.

While planning in SSPs is a widely studied and well-understood topic (Bertsekas and Tsitsiklis, 1991; Bertsekas and Yu, 2013), the problem of online learning in SSP, often referred to as *goal-oriented reinforcement learning (GRL)*, only recently became an active venue of research (Tarbouriech et al., 2020a, 2021b,c; Rosenberg and Mansour, 2020; Cohen et al., 2020, 2021; Chen et al., 2021a,b,c, 2022; Chen and Luo, 2021, 2022; Jafarnia-Jahromi et al., 2021; Vial et al., 2021; Min et al., 2021; Zhao et al., 2022). Most of the literature focuses on the regret minimization objective[1], for which learning algorithms with minimax-optimal performance are available even when no prior knowledge about the optimal policy is provided (e.g., its hitting time or the range of its value function). On the other hand, the *probably approximately correct* (PAC) objective, i.e., to learn an ϵ-optimal policy with high probability with as few samples as possible, has received little attention so far. One reason is that, as it is shown in (Tarbouriech et al., 2021b), in SSP it is not possible to convert regret into sample complexity bounds through an online-to-batch conversion (Jin et al., 2018) and PAC guarantees can only be derived by developing specific algorithmic and theoretical tools. Assuming access to a generative model, Tarbouriech et al. (2021b) derived the first PAC algorithm for SSP with sample complexity upper bounded as $\tilde{\mathcal{O}}(\frac{T_\ddagger B_\star^2 \Gamma S A}{\epsilon^2})$, where S is the number of states, A is the number of actions, Γ is the largest support of the transition distribution, $B_\star$ is the maximum expected cost of an optimal policy over all states, $T_\ddagger = B_\star / c_{\min}$, where $c_{\min}$ is the minimum cost over all state-action pairs, and ϵ is the desired accuracy. The most intriguing aspect of this bound is the dependency on $T_\ddagger$, which represents a worst-case bound on the hitting time $T_\star$ of the optimal policy (i.e., the horizon of the SSP) and it depends on the inverse of the minimum cost. While some dependency on the horizon may be unavoidable, as conjectured in (Tarbouriech et al., 2021b), we may expect the horizon to be independent of the cost function[2], as in finite-horizon and discounted problems. Moreover, in regret minimization, there are algorithms whose regret bound only scales with $T_\star$, with no dependency on $c_{\min}$, even when $c_{\min} = 0$ and no prior knowledge is available. It is thus reasonable to conjecture that the sample complexity should also scale with $T_\star$ instead of $T_\ddagger$. This leads us to the first question addressed in this paper:

Question 1: Is the dependency on $T_\ddagger = B_\star / c_{\min}$ in the sample complexity of learning with a generative model unavoidable?

Surprisingly, we derive a lower bound providing an affirmative answer to the question. In particular, we show that $\Omega(\frac{T_\ddagger B_\star^2 S A}{\epsilon^2})$ samples are needed to learn an ϵ-optimal policy, showing that a dependency on $T_\ddagger$ is indeed unavoidable and that it is not possible to adapt to the optimal policy hitting time $T_\star$[3]. This result also implies that there exist SSP instances with $c_{\min} = 0$ (i.e., $T_\ddagger = \infty$) that are *not learnable*. This shows for the first time that not only SSP is a strict generalization of the finite-horizon and discounted settings, but it is also strictly harder to learn. We then derive lower bounds when prior knowledge of the form $\overline{T} \geq T_\star$ is provided or when an optimality criterion restricted to policies with bounded hitting time is defined. Finally, we propose a simple algorithm based on a finite-horizon reduction argument and we prove upper bounds for its sample complexity matching the lower bound in each of the cases considered above; see Table 1.

1. We refer the reader to Appendix A for a detailed summary of prior works in related settings.
2. Notice that in general $T_\star \ll B_\star / c_{\min}$.
3. In our proof, we construct SSP instances where $T_\star < T_\ddagger$.

Performance (gen model)	Lower Bound	Upper Bound	Tarbouriech et al. (2021b)
(ϵ, δ) Definition 1	$\min\{T_\ddagger, \overline{T}\}\frac{B_\star^2 SA}{\epsilon^2}$	$\min\{T_\ddagger, \overline{T}\}\frac{B_\star^2 SA}{\epsilon^2}$	$\frac{T_\ddagger B_\star^2 \Gamma SA}{\epsilon^2}$
(ϵ, δ, T) Definition 5	$\frac{TB_{\star,T}^2 SA}{\epsilon^2}$ when $\min\{T_\ddagger, \overline{T}\} = \infty$	$\min\{T_\ddagger, T\}\frac{B_{\star,T}^2 SA}{\epsilon^2}$	$\frac{TB_{\star,T}^3 \Gamma SA}{\epsilon^3}$

Performance (BPI)	Assumption	Lower Bound	Upper Bound
(ϵ, δ) Definition 1	None	$\frac{A^{\Omega(\min\{B_\star, S\})}}{\epsilon}$	-
	Assumption 1	$\min\{T_\ddagger, \overline{T}\}\frac{B_\star^2 SA}{\epsilon^2} + \frac{J}{\epsilon}$	$\frac{T_\ddagger B_\star^2 SA}{\epsilon^2} + \frac{B_\star J^4 S^2 A^2}{c_{\min}^3 \epsilon}$

Table 1: Result summary *with (upper table) and without (lower table) a generative model*. Here, $\overline{T}$ is a known upper bound on the hitting time of the optimal policy ($\overline{T} = \infty$ when such a bound is unknown), $T_\ddagger = \frac{B_\star}{c_{\min}}$, $B_{\star,T}$ is the maximum expected cost over all starting states of the restricted optimal policy with hitting time bounded by T, and J is the cost to directly reach the goal from any state (Assumption 1). Operators $\tilde{\mathcal{O}}(\cdot)$ and $\Omega(\cdot)$ are hidden for simplicity.

When no access to a generative model is provided, the learner needs to directly execute a policy to collect samples to improve their estimate of the optimal policy. No result is currently available for this setting, often referred to as the *best-policy identification* (BPI) problem, and in this paper we address the following question:

Question 2: Is it possible to efficiently learn a near-optimal SSP policy when no access to a generative model is provided?

In this setting, we first derive a lower bound showing that in general sample efficient BPI is impossible. To resolve this negative result, we introduce an extra assumption requiring that the learner can reach the goal by paying a fixed cost J from any state. We then establish a $\Omega(\min\{T_\ddagger, \overline{T}\}\frac{B_\star^2 SA}{\epsilon^2} + \frac{J}{\epsilon})$ lower bound under this assumption. We also develop a finite-horizon reduction based algorithm with sample complexity $\tilde{\mathcal{O}}(\frac{T_\ddagger B_\star^2 SA}{\epsilon^2} + \frac{B_\star J^4 S^2 A^2}{c_{\min}^3 \epsilon})$, whose dominating term is minimax-optimal when $c_{\min} > 0$ and prior knowledge $\overline{T}$ is unavailable. This result is summarized in Table 1.

Finally, we show how similar technical tools derived for our lower bounds can be adapted to resolve an open question in Tarbouriech et al. (2021c) by showing that in regret minimization, a worst-case dependency on the hitting time of the optimal policy is indeed unavoidable without any prior knowledge (see Appendix G).

2. Preliminaries

An SSP instance is denoted by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, g, c, P)$, where $\mathcal{S}$ with $S = |\mathcal{S}|$ is the state space, $\mathcal{A}$ with $A = |\mathcal{A}|$ is the action space, $g \notin \mathcal{S}$ is the goal state, $c \in [c_{\min}, 1]^{\mathcal{S}_+ \times \mathcal{A}}$ with $c_{\min} \in [0, 1]$, $c(g, a) = 0$ for all a, and $\mathcal{S}_+ = \mathcal{S} \cup \{g\}$ is the cost function, and $P = \{P_{s,a}\}_{(s,a) \in \mathcal{S}_+ \times \mathcal{A}}$ with $P_{s,a} \in \Delta_{\mathcal{S}_+}$ and $P(g|g, a) = 1$ for all a is the transition function, where $\Delta_{\mathcal{S}_+}$ is the simplex over $\mathcal{S}_+$. All of these elements are known to the learner except the transition function.

A stationary policy π assigns an action distribution $\pi(\cdot|s) \in \Delta_{\mathcal{A}}$ to each state $s \in \mathcal{S}$. A policy is *deterministic* if $\pi(\cdot|s)$ concentrates on a single action (denoted by $\pi(s)$) for all s. Denote by $T^{\pi}(s)$ the expected number of steps it takes to reach g starting from state s and following π. A policy is *proper* if starting from any state it reaches the goal state with probability 1 (i.e., $T^{\pi}(s) < \infty$ for all $s \in \mathcal{S}$), and it is *improper* otherwise (i.e., there exists $s \in \mathcal{S}$ such that $T^{\pi}(s) = \infty$). We denote by Π the set of stationary policies, and Π_{∞} the set of stationary proper policies.

Given a cost function c and policy π, the value function of π, $V^{\pi} \in [0, \infty]^{\mathcal{S}_+}$ is defined as $V^{\pi}(s) = \mathbb{E}_{\pi}[\sum_{i=1}^{\infty} c(s_i, a_i)|s_1 = s]$, where the randomness is w.r.t. $a_i \sim \pi(\cdot|s_i)$ and $s_{i+1} \sim P_{s_i,a_i}$. We define the optimal proper policy $\pi^{\star} = \operatorname{argmin}_{\pi \in \Pi_{\infty}} V^{\pi}(s)$ for all $s \in \mathcal{S}$, and we write $V^{\pi^{\star}}$ as $V^{\star}$. It is known that $\pi^{\star}$ is stationary and deterministic.

We introduce a number of quantities that play a major role in characterizing the learning complexity in SSP: $B_{\star} = \max_s V^{\star}(s)$, the maximum expected cost of the optimal policy starting from any state, $T_{\star} = \max_s T^{\pi^{\star}}(s)$, and $D = \max_s \min_{\pi \in \Pi} T^{\pi}(s)$, the diameter of the SSP instance. Then, we have that

$$B_{\star} \le D \le T_{\star} \le \frac{B_{\star}}{c_{\min}} =: T_{\ddagger}.$$

Note that these inequalities may be strict and the gap arbitrarily large. Furthermore, this shows that the knowledge of $B_{\star}$ (or an upper bound) does not only provide an information about the range of the value function but also a worst-case bound $T_{\ddagger}$ on the horizon $T_{\star}$. We assume $B_{\star} \ge 1$, a commonly made assumption in previous work of SSP (Tarbouriech et al., 2021c; Chen et al., 2021a).[4]

Learning objective The goal of the learner is to identify a near-optimal policy of desired accuracy with high probability, with or without a generative model. We formalize each component below.

Sample Collection *With a generative model (PAC-SSP)*, the learner directly selects a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and collects a sample of the next state s' drawn from $P_{s,a}$. *Without a generative model (i.e., Best Policy Identification (BPI))*, the learner directly interacts with the environment through episodes starting from an initial state s_{init} and sequentially taking actions until g is reached.

ϵ-Optimality With a generative model, we say a policy π is ϵ-optimal if $V^{\pi}(s) - V^{\star}(s) \le \epsilon$ for all $s \in \mathcal{S}$. Without a generative model, a policy π is ϵ-optimal if $V^{\pi}(s_{\text{init}}) - V^{\star}(s_{\text{init}}) \le \epsilon$.

Definition 1 ((ϵ, δ)-Correctness) *Let $\mathcal{T}$ be the random stopping time by when an algorithm terminates its interaction with the environment and returns a policy $\widehat{\pi}$. We say that an algorithm is (ϵ, δ)-correct with sample complexity $n(\mathcal{M})$ if $\mathbb{P}_{\mathcal{M}}(\mathcal{T} \le n(\mathcal{M}), \widehat{\pi}$ is ϵ-optimal in $\mathcal{M}) \ge 1 - \delta$ for any SSP instance $\mathcal{M}$, where $n(\mathcal{M})$ is a deterministic function of the characteristic parameters of the problem (e.g., number of states and actions, inverse of the accuracy ϵ).*

Other notation We denote by $\overline{T}$ an upper bound of $T_{\star}$ known to the learner, and let $\overline{T} = \infty$ if such knowledge is unavailable. The $\tilde{\mathcal{O}}(\cdot)$ operator hides all logarithmic dependency including $\ln \frac{1}{\delta}$ for some confidence level $\delta \in (0, 1)$. For simplicity, we often write $a = \tilde{\mathcal{O}}(b)$ as $a \lesssim b$. Define $(x)_+ = \max\{0, x\}$. For $n \in \mathbb{N}_+$, define $[n] = \{1, \dots, n\}$.

3. Lower Bounds with a Generative Model

In this section, we derive lower bounds on PAC-SSP in various cases.

4. Note that in regret minimization, the lower bounds for $B_{\star} \ge 1$ and $B_{\star} < 1$ are different (Cohen et al., 2021).

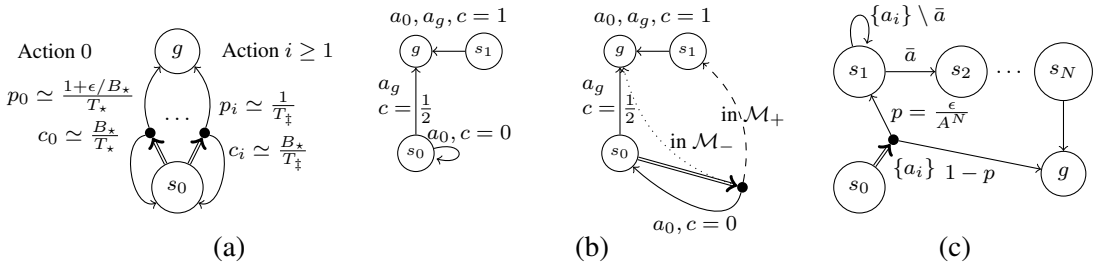

Figure 1: (a) hard instance (simplified for proof sketch) in Theorem 2 when $c_{\min} > 0$. (b) hard instance in Theorem 2 when $c_{\min} = 0$. (c) hard instance in Theorem 9. Here, c represents the cost of an action, while p represents the transition probability.

3.1. Lower Bound for ϵ-optimality

We first establish the sample complexity lower bound of any (ϵ, δ)-correct learning algorithm when no prior knowledge is available.

Theorem 2 *For any $S \geq 3$, $A \geq 3$, $c_{\min} > 0$, $B \geq 2$, $T_0 \geq \max\{B, \log_A S + 1\}$, $\epsilon \in (0, \frac{1}{32})$, and $\delta \in (0, \frac{1}{2e^4})$ such that $T_0 \leq B/c_{\min}$, there exists an MDP with S states, A actions, minimum cost $c_{\min}$, $B_\star = \Theta(B)$, and $T_\star = \Theta(T_0)$, such that any (ϵ, δ)-correct algorithm has sample complexity $\Omega\left(\frac{T_\ddagger B_\star^2 SA}{\epsilon^2} \ln \frac{1}{\delta}\right)$.[5] There also exists an MDP with $c_{\min} = 0$, $T_\star = 1$, $\overline{T} = \infty$, and $B_\star = 1$ in which every (ϵ, δ)-correct algorithm with $\epsilon \in (0, \frac{1}{2})$ and $\delta \in (0, \frac{1}{16})$ has infinite sample complexity.*

Details are deferred to Appendix C.1. We first remark that the lower bound qualitatively matches known PAC bounds for the discounted and finite-horizon settings in terms of its dependency on the size of the state-action space and on the inverse of the squared accuracy ϵ. As for the dependency on $B_\star$ and $c_{\min}$, it can be conveniently split in two terms: **1)** a term $B_\star^2$ and **2)** a factor $T_\ddagger = B_\star / c_{\min}$.

Dependency on $B_\star^2$ This term is connected to the range of the optimal value function $V^\star$. Interestingly, in finite-horizon and discounted settings H and $1/(1-\gamma)$ bound the range of the value function of *any* policy, whereas in SSP a more refined analysis is required to avoid dependencies on, e.g., $\max_\pi V^\pi$, which can be unbounded whenever an improper policy exists.

Dependency on $T_\ddagger$ While $T_\ddagger$ is an upper bound on the hitting time of the optimal policy, in the construction of the lower bound $T_\star$ is strictly smaller than $T_\ddagger$. For the case $c_{\min} > 0$, this shows that the algorithm proposed by Tarbouriech et al. (2021b) has an optimal dependency in $B_\star$ and $T_\ddagger$. On the other hand, this reveals that in certain SSP instances no algorithm can return an ϵ-optimal policy after collecting a finite number of samples. *This is the first evidence that learning in SSPs is strictly harder than the finite-horizon and discounted settings, where the sample complexity is always bounded.* This is also in striking contrast with results in regret minimization in SSP, where the regret is bounded even for $c_{\min} = 0$ and no prior knowledge about $B_\star$ or $T_\star$ is provided. This is due to the fact that the regret measures performance in the cost dimension and the algorithm is allowed to

5. Formally, for any $n \geq 0$, we say that an algorithm has sample complexity $\Omega(n)$ on an SSP instance $\mathcal{M}$ if $\mathbb{P}_{\mathcal{M}}(\mathcal{T} \leq n, \widehat{\pi}$ is ϵ-optimal in $\mathcal{M}) \geq 1 - \delta$.

change policies within and across episodes. On the other hand, in learning with a generative model the performance is evaluated in terms of the number of samples needed to confidently commit to a policy with performance ϵ-close to the optimal policy. This requires to distinguish between proper and improper policies, which can become arbitrarily hard in certain SSPs where $c_{\min} = 0$.

Proof Sketch In order to provide more insights about our result, here we present the main idea of our hard instances construction. We consider two cases separately: 1) $c_{\min} > 0$ and 2) $c_{\min} = 0$. When $c_{\min} > 0$, our construction is a variant of that in (Mannor and Tsitsiklis, 2004, Theorem 1); see an illustration in Figure 1 (a). Let's consider an MDP $\mathcal{M}$ with a multi-arm bandit structure: it has a single state s_0 and $N+1$ actions $\mathcal{A} = \{0, 1, \ldots, N\}$ (in the general case this corresponds to $N+1$ state-action pairs). Taking action 0 incurs a cost $\frac{B}{T_0}$ and transits to the goal state with probability $\frac{1+\bar{\epsilon}/2}{T_0}$ (stays in s_0 otherwise), where $\bar{\epsilon} = \frac{32\epsilon}{B}$. For each $i \in [N]$, taking action i incurs a cost $\frac{B}{T_1}$ and transits to the goal state with probability $\frac{1}{T_1}$. Note that in $\mathcal{M}$ the optimal action (deterministic policy) is 0, with $B_\star = \Theta(B)$, $T_\star = \Theta(T_0)$, $T_\ddagger = \Theta(T_1)$, whereas all other actions are more than ϵ suboptimal. Also note that it takes $\Omega(\frac{T_1 B_\star^2}{\epsilon^2})$ samples to estimate the expected cost of action $i \in [N]$ with accuracy ϵ. If an algorithm $\mathfrak{A}$ spends $o(\frac{T_1 B_\star^2}{\epsilon^2})$ samples on some action i', then we can consider an alternative MDP $\mathcal{M}'$, whose only difference compared to $\mathcal{M}$ is that taking action i' transits to the goal state with probability $\frac{1+\bar{\epsilon}}{T_1}$. Note that in $\mathcal{M}'$ the only ϵ-optimal action is i'. However, algorithm $\mathfrak{A}$ cannot distinguish between $\mathcal{M}$ and $\mathcal{M}'$ since it does not have enough samples on action i', and thus has a high probability on outputting the wrong action in either $\mathcal{M}$ or $\mathcal{M}'$. Applying this argument to each arm $i \in [N]$, we conclude that an (ϵ, δ)-correct algorithm needs at least $\Omega(\frac{N T_1 B_\star^2}{\epsilon^2}) = \Omega(\frac{T_\ddagger B_\star^2 SA}{\epsilon^2})$ samples.

We emphasize that in our construction, $T_\star$ (whose proxy is T_0) can be arbitrarily smaller than $T_\ddagger$ (whose proxy is T_1) in $\mathcal{M}$. However, the learner still needs $\Omega(\frac{T_\ddagger B_\star^2}{\epsilon^2})$ samples to exclude the alternative $\mathcal{M}'$ in which $T_\star = T_\ddagger$. A natural question to ask is what if we have prior knowledge on $T_\star$, which could potentially reduce the space of alternative MDPs. We answer this in Section 3.2.

When $c_{\min} = 0$, we consider a much simpler MDP $\mathcal{M}$ with two states $\{s_0, s_1\}$ and two actions $\{a_0, a_g\}$; see an illustration in Figure 1 (b). At s_0, taking a_0 transits to s_0 with cost 0 and taking a_g transits to g with cost $\frac{1}{2}$. At s_1, taking both actions transits to g with cost 1. Clearly $c_{\min} = 0$, $B_\star = T_\star = 1$, $V^\star(s_0) = \frac{1}{2}$, and both actions in s_0 are ϵ-optimal in $\mathcal{M}$. Now consider any algorithm $\mathfrak{A}$ with sample complexity $n < \infty$ on $\mathcal{M}$, and without loss of generality, assume that $\mathfrak{A}$ outputs a deterministic policy $\widehat{\pi}$. We consider two cases: 1) $\widehat{\pi}(s_0) = a_0$ and 2) $\widehat{\pi}(s_0) = a_g$. In the first case, consider an alternative MDP $\mathcal{M}_+$, whose only difference compared to $\mathcal{M}$ is that taking a_0 at s_0 transits to s_1 with probability $\frac{1}{n}$, and to s_0 otherwise. Note that the optimal action at s_0 is a_g in $\mathcal{M}_+$. Since $\mathfrak{A}$ uses at most n samples, with high probability it never observes transition (s_0, a_0, s_1) and is unable to distinguish between $\mathcal{M}$ and $\mathcal{M}_+$. Thus, it still outputs $\widehat{\pi}$ with $\widehat{\pi}(s_0) = a_0$ in $\mathcal{M}_+$. This gives $V^{\widehat{\pi}}(s_0) - V^\star(s_0) = 1 - \frac{1}{2}$ and $\widehat{\pi}$ is not ϵ-optimal for any $\epsilon \in (0, \frac{1}{2})$. In the second case, consider another alternative MDP $\mathcal{M}_-$, whose only difference compared to $\mathcal{M}$ is that taking a_0 at s_0 transits to g with probability $\frac{1}{n}$, and to s_0 otherwise. The optimal action at s_0 is a_0 in $\mathcal{M}_-$. Again, algorithm $\mathfrak{A}$ cannot distinguish between $\mathcal{M}$ and $\mathcal{M}_-$ and still output $\widehat{\pi}$ with $\widehat{\pi}(s_0) = a_g$ in $\mathcal{M}_-$. This gives $V^{\widehat{\pi}}(s_0) - V^\star(s_0) = \frac{1}{2}$ and $\widehat{\pi}$ is not ϵ-optimal for any $\epsilon \in (0, \frac{1}{2})$. Combining these two cases, we have that any (ϵ, δ)-correct algorithm with $\epsilon \in (0, \frac{1}{2})$ cannot have finite sample complexity.

Remark Our construction reveals that the potentially infinite horizon in SSP does bring hardness into learning when $c_{\min} = 0$. Indeed, we can treat $\mathcal{M}$ as an infinite-horizon MDP due to the presence

of the self-loop at s_0. Any algorithm that uses finite number of samples cannot identify all proper policies in $\mathcal{M}$, that is, it can never be sure whether (s_0, a_0) has non-zero probability of reaching states other than s_0.

3.2. Lower Bound for ϵ-optimality with Prior Knowledge on $T_\star$

Now we consider the case where the learning algorithm has some prior knowledge $\overline{T} \geq T_\star$ on the hitting time of the optimal proper policy. Intuitively, we expect the algorithm to exploit the knowledge of parameter $\overline{T}$ to focus on the set of policies $\{\pi : \|T^\pi\|_\infty \leq \overline{T}\}$ with bounded hitting time.[6]

Theorem 3 *For any $S \geq 3$, $A \geq 3$, $c_{\min} \geq 0$, $B \geq 2$, $T_0 \geq \max\{B, \log_A S + 1\}$, $\overline{T} \geq 0$, $\epsilon \in (0, \frac{1}{32})$, and $\delta \in (0, \frac{1}{2e^4})$ such that $T_0 \leq \min\{\overline{T}/2, B/c_{\min}\} < \infty$, there exist an MDP with S states, A actions, minimum cost $c_{\min}$, $B_\star = \Theta(B)$, and $T_\star = \Theta(T_0) \leq \overline{T}$, such that any (ϵ, δ)-correct algorithm has sample complexity $\Omega\left(\min\left\{T_\ddagger, \overline{T}\right\} \frac{B_\star^2 SA}{\epsilon^2} \ln \frac{1}{\delta}\right)$.*

Details are deferred to Appendix C.1. The main idea of proving Theorem 3 still follows from that of Theorem 2. Also note that the bound in Theorem 3 subsumes that of Theorem 2 since we let $\overline{T} = \infty$ when such knowledge is unavailable.

Dependency on $\min\{T_\ddagger, \overline{T}\}$ We distinguish two regimes: **1)** When $\overline{T} \leq T_\ddagger$ the bound reduces to $\Omega(\frac{\overline{T} B_\star^2 SA}{\epsilon^2})$ with no dependency on $c_{\min}$. In this case, an algorithm may benefit from its prior knowledge to effectively prune any policy with hitting time larger than T, thus reducing the sample complexity of the problem and avoiding infinite sample complexity when $c_{\min} = 0$. **2)** When $\overline{T} > T_\ddagger$, we recover the bound $\Omega(\frac{T_\ddagger B_\star^2 SA}{\epsilon^2})$. In this case, an algorithm does not pay the price of a loose upper bound on $T_\star$. Again, in our construction it is possible that $T_\star < \min\{T_\ddagger, \overline{T}\}$. This concludes that it is impossible to adapt to $T_\star$ for computing ϵ-optimal policies in SSPs.

3.3. Lower Bound for (ϵ, T)-optimality

Knowing that we cannot solve for an ϵ-optimal policy when $\min\{T_\ddagger, \overline{T}\} = \infty$, that is, $c_{\min} = 0$ and $\overline{T} = \infty$, we now consider a restricted optimality criterion where we only seek ϵ-optimality w.r.t. a set of proper policies.

Definition 4 (Restricted (ϵ, T)-Optimality) *For any $T \in [1, \overline{T}]$, we define the set $\Pi_T = \{\pi \in \Pi : \|T^\pi\|_\infty \leq T\}$. Also define $\pi^\star_{T,s} = \operatorname{argmin}_{\pi \in \Pi_T} V^\pi(s)$, $V^{\star,T}(s) = V^{\pi^\star_{T,s}}(s)$, and $B_{\star,T} = \max_s V^{\star,T}(s)$. We say that a policy π is (ϵ, T)-optimal if $V^\pi(s) - V^{\star,T}(s) \leq \epsilon$ for all $s \in \mathcal{S}$. We define $V^{\star,T}(s) = \infty$ for all s when $\Pi_T = \emptyset$.*[7]

When $T \geq T_\star$, we have $\pi^\star_{T,s} = \pi^\star$ for all s. When $D \leq T < T_\star$, the policy $\pi^\star_{T,s}$ exists and may vary for different starting state s due to the hitting time constraint. It can even be stochastic

6. Notice that $\{\pi : \|T^\pi\|_\infty \leq \overline{T}\}$ includes the optimal policy by definition since $\overline{T} \geq T_\star$.
7. Tarbouriech et al. (2021b) consider a slightly different notion of restricted optimality, where they let $T = \theta D$ with $\theta \in [1, \infty)$ as input to the algorithm, and D is unknown.

from the literature of constrained MDPs (Altman, 1999). [8] When $T < D$, we have $\Pi_T = \varnothing$, and $V^{\star,T}(s) = \infty$ for all s. Clearly, $V^{\star,T}(s) \geq V^\star(s)$ for any s and T.

Definition 5 (**(ϵ, δ, T)-Correctness**) *Let $\mathcal{T}$ be the random stopping time by when an algorithm terminates its interaction with the environment and returns a policy $\widehat{\pi}$. We say that an algorithm is (ϵ, δ, T)-correct with sample complexity $n(\mathcal{M})$ if $\mathbb{P}_{\mathcal{M}}(\mathcal{T} \leq n(\mathcal{M}), \widehat{\pi}$ is (ϵ, T)-optimal in $\mathcal{M}) \geq 1 - \delta$ for any SSP instance $\mathcal{M}$, where $n(\mathcal{M})$ is a deterministic function of the characteristic parameters of the problem (e.g., number of states and actions, inverse of the accuracy ϵ).*

Note that π being (ϵ, T)-optimal does not require $\pi \in \Pi_T$. For example, $\pi^\star$ is (ϵ, T)-optimal for any $T \geq 1$. Similarly, policy output by an (ϵ, δ, T)-correct algorithm is not required to be in Π_T, and it is allowed return a better cost-oriented policy.

Now we establish a sample complexity lower bound of any (ϵ, δ, T)-correct algorithm when $\min\{T_\ddagger, \overline{T}\} = \infty$ (see Appendix C.2 for details).

Theorem 6 *For any $S \geq 6$, $A \geq 8$, $B_\star \geq 2$, $T \geq 6(\log_{A-1}(S/2) + 1)$, $B_T \geq 2$, $\epsilon \in (0, \frac{1}{32})$, and $\delta \in (0, \frac{1}{8e^4})$ such that $B_\star \leq B_T \leq B_\star(A-1)^{S/2-1}/4$ and $B_T \leq T/6$, and for any (ϵ, δ, T)-correct algorithm, there exist an MDP with $B_{\star,T} = \Theta(B_T)$, $c_{\min} = 0$, and parameters S, A, $B_\star$, such that with a generative model, the algorithm has sample complexity $\Omega\left(\frac{TB_{\star,T}^2 SA}{\epsilon^2} \ln \frac{1}{\delta}\right)$.*

Note that when $T \geq T_\star$, the lower bound reduces to $\frac{TB_\star^2 SA}{\epsilon^2}$, which coincides with that of Theorem 3. On the other hand, the sample complexity lower bound for computing (ϵ, T)-optimal policy when $\min\{T_\ddagger, \overline{T}\} < \infty$ is still unknown and it is an interesting open problem.

Proof Sketch We consider an MDP $\mathcal{M}$ with state space $\mathcal{S} = \mathcal{S}_T \cup \mathcal{S}_\star$. The learner can reach the goal state either through states in $\mathcal{S}_T$ or $\mathcal{S}_\star$, where in the first case the learner aims at learning an (ϵ, T)-optimal policy, and in the second case the learner aims at learning an ϵ-optimal policy. In $\mathcal{S}_T$, we follow the construction in Theorem 2 so that learning an ϵ-optimal policy on the sub-MDP restricted on $\mathcal{S}_T$ takes $\Omega(\frac{TB_{\star,T}^2 SA}{\epsilon^2})$ samples. In $\mathcal{S}_\star$, we consider a sub-MDP that forms a chain similar to (Strens, 2000, Figure 1), where the optimal policy suffer $B_\star$ cost but a bad policy could suffer $\Omega(B_\star A^S)$ cost. For each state s in $\mathcal{S}_\star$, we make the probability of transiting back to s by taking any action large enough, so that learner with sample complexity of order $\tilde{\mathcal{O}}(\frac{TB_{\star,T}^2 SA}{\epsilon^2})$ hardly receive any learning signals in $\mathcal{S}_\star$. Therefore, any algorithm with $\tilde{\mathcal{O}}(\frac{TB_{\star,T}^2 SA}{\epsilon^2})$ sample complexity should focus on learning the sub-MDP restricted on $\mathcal{S}_T$. This proves the statement.

8. Consider an MDP with one state and two actions. Taking action one suffers cost 1 and directly transits to the goal state. Taking action 2 suffers cost 0 and transits to the goal state with probability $1/3$. Now consider $T = 2$. Then the optimal constrained policy should take action 2 with probability $3/4$.

4. Algorithm with a Generative Model

Algorithm 1 Search Horizon

Input: hitting time bound T ($T = \overline{T}$ with prior knowledge), accuracy $\epsilon \in (0,1)$, and probability $\delta \in (0,1)$.
Initialize: $i \leftarrow 1$.
Let $B_i = 2^i$, $H_i = 4\min\{B_i/c_{\min}, T\}\ln(48B_i/\epsilon)$, $c_{f,i}(s) = 0.6B_i\mathbb{I}\{s \neq g\}$, $\delta_i = \delta/(40i^2)$, $N_i^\star = N^\star(B_i, H_i, \frac{\epsilon}{2}, \delta_i)$ and $N_i = \widehat{N}(B_i, H_i, 0.1B_i, \delta_i)$, where $N^\star$, $\widehat{N}$ are defined in Lemma 22 and Lemma 23 respectively.
// ESTIMATE $B_{\star,T}$
while *True* **do**
  Reset counter $\mathbf{N}$, and then draw N_i samples for each (s,a) to update $\mathbf{N}$.
  $\pi^i, V^i = \text{LCBVI}(H_i, \mathbf{N}, B_i, c_{f,i}, \delta_i)$ (refer to Algorithm 3).
  if $\left\|V_1^i\right\|_\infty \le 0.1B_i$ ***and*** $\max_{h\in[H+1]}\left\|V_h^i\right\|_\infty \le 0.7B_i$ **then break**.
  $i \leftarrow i+1$.
  if $B_i > 40T$ **then output** $\widehat{\pi} = \varnothing$. // I.E., $T < D$ (EVERY POLICY IS (ϵ, T)-OPTIMAL)
end
// COMPUTE ϵ-OPTIMAL POLICY
Reset counter $\mathbf{N}$, and then draw $N_i^\star$ samples for each (s,a) to update $\mathbf{N}$.
$\widehat{\pi}, \widehat{V} = \text{LCBVI}(H_i, \mathbf{N}, B_i, c_{f,i}, \delta_i)$ (refer to Algorithm 3).
Output: policy $\widehat{\pi}$ extended to infinite horizon.

In this section, we present an algorithm whose sample complexity matches all the lower bounds introduced in Section 3. We notice that the horizon (or hitting time) of the optimal policy plays an important role in the lower bounds. Thus, a natural algorithmic idea is to explicitly determine and control the horizon of the output policy. This leads us to the idea of finite-horizon reduction, which is frequently applied in the previous works on SSP (e.g., Chen et al., 2021b,c; Cohen et al., 2020).

Now we formally describe the finite-horizon reduction scheme. Given an SSP $\mathcal{M}$, let $\mathcal{M}_{H,c_f}$ be a time-homogeneous finite-horizon MDP with horizon H and terminal cost $c_f \in [0,\infty)^{\mathcal{S}_+}$, which has the same state space, action space, cost function, and transition function as $\mathcal{M}$. When interacting with $\mathcal{M}_{H,c_f}$, the learner starts in some initial state and stage h, it observes state s_h, takes action a_h, incurs cost $c(s_h,a_h)$, and transits to the next state $s_{h+1} \sim P_{s_h,a_h}$. It also suffers cost $c_f(s_{H+1})$ before ending the interaction. When the finite-horizon MDP $\mathcal{M}_{H,c_f}$ is clear from the context, we define $V_h^\pi(s)$ as the expected cost of following policy π starting from state s and stage h in $\mathcal{M}_{H,c_f}$.

Although the finite-horizon reduction has become a common technique in regret minimization for SSP, it is not straightforward to apply it in our setting. Indeed, even if we solve a near-optimal policy in the finite-horizon MDP, it is unclear how to apply the finite-horizon policy in SSP, where any trajectory may be much longer than H. Our key result is a lemma that resolves this. It turns out that when the terminal cost in the finite-horizon MDP is large enough, all we need for applying the finite-horizon policy to SSP is to repeat it periodically. Specifically, given a finite-horizon policy $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}\times[H]}$, we abuse the notation and define $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}\times\mathbb{N}_+}$ as an infinite-horizon non-stationary policy, such that $\pi(a|s,h+iH) = \pi(a|s,h), \forall i \in \mathbb{N}_+$. The following lemma relates the performance of π in $\mathcal{M}$ to its performance in $\mathcal{M}_{H,c_f}$ (see Appendix D.1 for details).

Lemma 7 *For any SSP $\mathcal{M}$, horizon H, and terminal cost function c_f, suppose π is a policy in $\mathcal{M}_{H,c_f}$ and $V_1^\pi(s) \le c_f(s)$ for all $s \in \mathcal{S}_+$. Then $V^\pi(s) \le V_1^\pi(s)$.*

Thanks to the lemma above, for a given horizon T, we can first learn an ϵ-optimal policy $\widehat{\pi}$ in $\mathcal{M}_{H,c_f}$ with $H = \tilde{\mathcal{O}}(T)$ and $c_f(s) = \mathcal{O}(B_{\star,T}\mathbb{I}\{s \neq g\})$, and then extend it to an SSP policy with performance $V^{\widehat{\pi}}(s) \leq V_1^{\widehat{\pi}}(s) \approx V_1^\star(s) \approx V^\star(s)$, where $V_1^\star$ is the optimal value function of stage 1 in $\mathcal{M}_{H,c_f}$, and the last step is by the fact that H is sufficiently large compared to T. As a result, $\widehat{\pi}$ would then be (ϵ, T)-optimal policy in the original SSP problem. Algorithm 1 builds on this idea. It takes a hitting time upper bound T as input, and aims at computing an (ϵ, T)-optimal policy. The main idea is to search the range of $B_{\star,T}$ and $\min\{T_\ddagger, T\}$ via a doubling trick on estimators B_i and H_i (Line 1). [9] For each possible value of B_i and H_i, we compute an optimal value function estimate with $0.1B_i$ accuracy using $SAN_i \lesssim S^2AH_i$ samples (Line 2), and stop if B_i becomes a proper upper bound on the estimated value function (Line 3). Here we need different conditions bounding V_1^i and V_h^i as the terminal cost c_f should be negligible starting from stage 1 but not for any stage. Once we determine their range, we compute an ϵ-optimal finite-horizon policy with final values of B_i and H_i using $SAN_i^\star \lesssim \frac{H_i B_i^2 SA}{\epsilon^2}$ samples (Line 5). On the other hand, if B_i becomes unreasonably large, then the algorithm claims that $T < D$ (Line 4), in which case $V^{\star,T}(s) = \infty$ for any s (see Definition 4), and any policy is (ϵ, T)-optimal by definition. In the procedure described above, we need to repeatedly compute a near-optimal policy with various accuracy and horizon. We use a simple variant of the UCBVI algorithm (Azar et al., 2017; Zhang et al., 2020b) to achieve this (see Algorithm 3 in Appendix D.2). The main idea is to compute an optimistic value function estimate by incorporating a Bernstein-style bonus (Line 1).

We state the guarantee of Algorithm 1 in the following theorem (see Appendix D.3 for details).

Theorem 8 *For any given $T \geq 1$, $\epsilon \in (0,1)$, and $\delta \in (0,1)$, with probability at least $1-\delta$, Algorithm 1 either uses $\tilde{\mathcal{O}}\left(S^2AT\right)$ samples to confirm that $T < D$, or uses $\tilde{\mathcal{O}}\left(\min\left\{T_\ddagger, T\right\}\frac{B_{\star,T}^2 SA}{\epsilon^2}\right)$ samples to output an (ϵ, T)-optimal policy (ignoring lower order terms).*

When prior knowledge $\infty > \overline{T} \geq T_\star$ is available, we simply set $T = \overline{T}$. In this case, we have that (ϵ, T)-optimality is equivalent to ϵ-optimality, $B_{\star,\overline{T}} = B_\star$ and Algorithm 1 matches the lower bound in Theorem 3. When $\min\{T_\ddagger, \overline{T}\} = \infty$, Algorithm 1 computes an (ϵ, T)-optimal policy with $\tilde{\mathcal{O}}(\frac{TB_{\star,T}^2 SA}{\epsilon^2})$ samples, which matches the lower bound in Theorem 6. Thus, our algorithm is minimax optimal in all cases considered in Section 3.

When comparing with the results in (Tarbouriech et al., 2021b), in terms of computing an ϵ-optimal policy, we remove a Γ factor and improve the dependency of $T_\ddagger$ to $\min\{T_\ddagger, \overline{T}\}$, that is, our algorithm is able to leverage a given bound on $T_\star$ to improve sample efficiency while theirs cannot. In terms of computing an (ϵ, T)-optimal policy, we greatly improve over their result by removing a $\frac{B_{\star,T}\Gamma}{\epsilon}$ factor and improving the dependency of T to $\min\{T_\ddagger, T\}$, that is, our algorithm automatically adapts to a smaller hitting time upper bound of the optimal policy.

Finally, it is interesting to notice that even though the (ϵ, T)-optimal policy is possibly stochastic, the policy output by Algorithm 1 is always deterministic, and it does not necessarily have hitting time bounded by T. In fact, Algorithm 1 puts no constraint on the hitting time of the output policy, except that the horizon for the reduction is $\tilde{\mathcal{O}}\left(T\right)$. Nevertheless, as shown in Theorem 8, we can still prove that the policy is (ϵ, T)-optimal since the requirement only evaluates the expected cost and not the constraint on the hitting time.

9. Note that $\left\|T^{\pi^\star_{T,s}}\right\|_\infty \leq T_\ddagger$ for any $T \geq D$ and state s since $\pi^\star_{T,s} = \pi^\star$ when $T \geq T_\ddagger \geq T_\star$.

5. Lower Bounds without a Generative Model

In this section, we consider the best policy identification problem in SSP, where the learner collects samples by interacting with the environment. This is a more challenging setting compared to having access to a generative model since the learner cannot "teleport" to any arbitrary state-action pair but it only observes trajectories obtained by playing policies from some initial state. Yet, this setting is more practical, and it naturally generalizes beyond tabular MDPs. Surprisingly, we find that BPI with polynomial number of samples is impossible in general.

Theorem 9 *For any $S \geq 4$, $A \geq 4$, $B_\star \geq 1$, $c_{\min} \geq 0$, $\epsilon \in (0, \frac{1}{4})$, and $\delta \in (0, \frac{1}{16})$, and any (ϵ, δ)-correct algorithm, there exists an MDP with parameters S, A, $B_\star$, and $c_{\min}$, such that without a generative model, the sample complexity of the algorithm is $\Omega\left(\frac{A^{\min\{\lfloor B_\star \rfloor, S-3\}}}{\epsilon}\right)$.*

Details are deferred to Appendix E.1. The exponential dependency $A^{\Omega(\min\{B_\star, S\})}$ implies that sample efficient BPI is impossible. The intuition of our construction is that if there are N unvisited states where the learner has no samples, then the learner may suffer $\Omega(A^N)$ cost on visiting these states. Therefore, the learner needs to spend $\Omega(A^N/\epsilon)$ samples on estimating the transition distribution to guarantee ϵ-optimality when there are N hardly reachable states; see an illustration in Figure 1 (c).

To enable sample efficient BPI, we need to avoid the extreme event described above. One natural idea is to allow the learner to get out of "unfamiliar" states by paying a fixed cost. This assumption also appears in (Tarbouriech et al., 2020a, Section I.2) in the context of non-communicating SSPs.

Assumption 1 *There is an action $a_\dagger \in \mathcal{A}$ with $c(s, a_\dagger) = J$ and $P(g|s, a_\dagger) = 1$ for all $s \in \mathcal{S}$.*

This assumption guarantees that there is a proper policy π with $\|V^\pi\|_\infty = J$ and $\|T^\pi\|_\infty = 1$. We show in Section 6 that under Assumption 1, efficient BPI is indeed possible. To better understand the difficulty of BPI under this assumption, we also establish the following lower bound.

Theorem 10 *Under Assumption 1, for any $S \geq 8$, $A \geq 5$, $c_{\min} \geq 0$, $B \geq \max\{2, (\log_{A-1} S + 1)c_{\min}\}$, $\overline{T} \geq 2\max\{B, \log_{A-1} S + 1\}$, $J \geq 3B$, $\epsilon \in (0, \frac{1}{32})$, and $\delta \in (0, \frac{1}{2e^4})$ such that $\min\{\overline{T}/2, B/c_{\min}\} < \infty$ and $J \leq \frac{1}{2}(A-1)^N$ with $N = \min\{\lfloor B \rfloor, S-3\}$, and for any (ϵ, δ)-correct algorithm, there exist an MDP with parameters S, A, $c_{\min}$, $\overline{T}$, and $B_\star = \Theta(B)$, such that the algorithm has sample complexity $\Omega\left(\min\left\{T_\ddagger, \overline{T}\right\} \frac{B_\star^2 SA}{\epsilon^2} \ln\frac{1}{\delta} + \frac{J}{\epsilon}\right)$.*

Details are deferred to Appendix E.2. Compared to the lower bound in Theorem 3, it has an extra $\frac{J}{\epsilon}$ term, showing that BPI is harder even with the extra assumption. The first term in the lower bound implies that within easily reachable states, the sample complexity of BPI is similar to having access to a generative model. Compared to the lower bound in Theorem 9, we replace the exponential factor A^S by J, which reflects the worst case cost of encountering "unfamiliar" states. Whether the dependency on J is minimax optimal remains an important future direction.

6. Algorithm without a Generative Model

In this section, we develop an efficient algorithm for BPI under Assumption 1. It is actually unclear how to design such an algorithm at first glance. The main difficulty lies in deciding when to invoke the action $a_\dagger$. In fact, if we simply apply the commonly used optimism based algorithm, then $a_\dagger$ may

Algorithm 2 BPI-SSP

Define: $\mathcal{N} = \{2^j\}_{j\geq 0}$, $H = \frac{32J}{c_{\min}} \ln \frac{8J}{\epsilon}$, $c_f(s) = J\mathbb{I}\{s \neq g\}$.
Initialize: $B \leftarrow 1$, $m \leftarrow 1$, $\mathbf{N}(s,a) \leftarrow 0$ and $\mathbf{N}(s,a,s') \leftarrow 0$ for any $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}_+$.
1 **for** $r = 1, \ldots$ **do**
  while *True* **do**
2 $\pi^r, V^r \leftarrow \text{LCBVI}(H, \mathbf{N}, 2J, c_f, \frac{\delta}{2B})$.
  if $B \geq \max_{h \leq H/2+1, s} V_h^r(s)$ **then break**.
3 $B \leftarrow 2 \cdot \max_{h \leq H/2+1, s} V_h^r(s)$.
  end
4 $\widehat{\tau} \leftarrow 0$, $\lambda \leftarrow N_{\text{DEV}}(B, \frac{\epsilon}{4}, \frac{\delta}{2r^2})$ (defined in Lemma 27).
  for $m' = 1, \ldots, \lambda$ **do**
    for $h = 1, \ldots, H$ **do**
      Observe s_h^m, take action $a_h^m = \pi^r(s_h^m, h)$, and transit to s_{h+1}^m.
5 $\mathbf{N}(s_h^m, a_h^m) \overset{+}{\leftarrow} 1$, $\mathbf{N}(s_h^m, a_h^m, s_{h+1}^m) \overset{+}{\leftarrow} 1$, $\widehat{\tau} \overset{+}{\leftarrow} c(s_h^m, a_h^m)/\lambda$.
6 **if** $\mathbf{N}(s_h^m, a_h^m) \in \mathcal{N}$ **then**
        if $s_{h+1}^m \neq g$ **then** take action $a_\dagger$ to reach g, and suffer cost J.
        Return to Line 1 (skip round).
      end
      if $s_{h+1}^m = g$ **then break**.
      if $h = H$ **then** take action $a_\dagger$ to reach g, and suffer cost J.
    end
7 **if** $\widehat{\tau} > V^r(s_{init}) + \frac{\epsilon}{2}$ **then** return to Line 1 (failure round).
    $m \overset{+}{\leftarrow} 1$.
  end
8 **Return** policy $\widehat{\pi} = \pi^r$ (success round).
end

never be selected since $J \geq B_\star$. Intuitively, we want to involve $a_\dagger$ for states with large uncertainty, which conflicts with the principle of optimism in the face of uncertainty. Therefore, we need a more carefully designed scheme to balance exploration and exploitation.

It turns out that we can obtain a naive sample complexity bound by reducing BPI in SSP to BPI in a specific finite-horizon MDP. Consider a finite-horizon MDP $\mathcal{M}_{H,c_f}$ with $H = \tilde{\mathcal{O}}(\frac{J}{c_{\min}})$ and $c_f(s) = J\mathbb{I}\{s \neq g\}$. Executing policy π in $\mathcal{M}_{H,c_f}$ corresponds to following policy π in $\mathcal{M}$ for H steps and then taking action $a_\dagger$ if the goal state is not reached. Thus, any policy π in $\mathcal{M}_{H,c_f}$ can directly extend to $\mathcal{M}$ by defining $\pi(s, H+1) = a_\dagger$, and we have $V^\pi = V_1^\pi$. Moreover, by the choice of H and c_f, the optimal policy in $\mathcal{M}_{H,c_f}$ is also near-optimal in $\mathcal{M}$. Applying any minimax-optimal finite-horizon BPI algorithm (for example, a variant of (Tarbouriech et al., 2022, Algorithm 1)), we can solve an ϵ-optimal policy with $\tilde{\mathcal{O}}(\frac{H\|V_\cdot^\star\|_\infty^2 SA}{\epsilon^2}) = \tilde{\mathcal{O}}(\frac{J^3 SA}{c_{\min}\epsilon^2})$ samples, where $V_h^\star$ is the optimal value function in $\mathcal{M}_{H,c_f}$ and $\|V_\cdot^\star\|_\infty = \max_{h\in[H+1]} \|V_h^\star\|_\infty$.

The sample complexity bound above is undesirable as J appears in the dominating term, which is not the case in our established lower bound (Theorem 10). The main issue is that in $\mathcal{M}_{H,c_f}$, the range of optimal value function $\|V_\cdot^\star\|_\infty = \mathcal{O}(J)$. Thus, simply applying finite-horizon BPI algorithm with minimax rate is insufficient to adapt to $\|V^\star\|_\infty = B_\star$. Now we present a BPI algorithm that resolves this issue and achieves minimax optimal sample complexity in the dominating term when there is no prior knowledge. The pseudo-code is shown in Algorithm 2.

The main structure of Algorithm 2 follows that of (Lim and Auer, 2012; Cai et al., 2022) for autonomous exploration. The learning procedure is divided into rounds (Line 1) of three types: skip round, failure round, and success round. In each round, the learner follows a behavior policy for at most λ episodes to collect samples and estimate its empirical performance (Line 5). If in the current round the number of visits to some state-action pair is doubled, then the current round is classified as a skip round (Line 6). If the empirical performance of the current behavior policy is not close enough to its estimated performance, then the current round is classified as a failure round (Line 7). In both cases, we start a new round and the behavior policy is updated. Otherwise, the current round is a success round, and the algorithm returns an ϵ-optimal policy (Line 8).

To adapt to $B_\star$ instead of J, we maintain an estimator B that is an upper bound of the estimated value function in the first $H/2+1$ steps (Line 3). Intuitively, $\|V_h^\star\|_\infty \approx \|V^\star\|_\infty$ when $h \le \frac{H}{2}+1$. The estimator B is used to determine the number of episodes needed to obtain an accurate empirical performance estimate of current behavior policy (Line 4), where $\lambda \lesssim \frac{B^2}{\epsilon^2}$ (see Lemma 27). The sample complexity of Algorithm 2 is summarized in the following theorem.

Theorem 11 *Under Assumption 1 and assuming $c_{\min} > 0$, for any $\epsilon \in (0,1)$ and $\delta \in (0,1)$, Algorithm 2 is (ϵ,δ)-correct with sample complexity $\tilde{\mathcal{O}}(\frac{T_\ddagger B_\star^2 SA}{\epsilon^2} + \frac{B_\star J^4 S^2 A^2}{c_{\min}^3 \epsilon})$.*

Details are deferred to Appendix F. The achieved sample complexity has no J dependency in the dominating term, and the dominating term is minimax-optimal when there is no prior knowledge, that is, $\overline{T} = \infty$. On the other hand, the lower order term might be sub-optimal and the algorithm only works for strictly positive costs. Resolving these two issues is an important open question.

7. Conclusion

In this work, we study the sample complexity of the SSP problem. We provide an almost complete characterization of the minimax sample complexity with a generative model, and initiate the study of BPI in SSP. We derived two important negative results: 1) an ϵ-optimal policy may not be learnable in SSP even with a generative model; 2) best policy identification in SSP requires an exponential number of samples in general. We complemented the study of sample complexity with lower bounds for learnable settings with and without a generative model, and matching upper bounds. Many interesting problems remain open, such as the minimax optimal sample complexity of computing an (ϵ, T)-optimal policy when $\min\{T_\ddagger, \overline{T}\} < \infty$, and the minimax optimal sample complexity of BPI under Assumption 1. Furthermore, an important direction is to study BPI under weaker conditions than Assumption 1 (e.g., communicating SSP). We believe that similar results can be obtained, with a more complicated analysis, when a reset action to the initial state is available in every state, a common assumption in the literature.

References

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

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

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume II*, volume 2. Athena scientific, 2012.

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

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

Haoyuan Cai, Tengyu Ma, and Simon Du. Near-optimal algorithms for autonomous exploration and multi-goal stochastic shortest path. *arXiv preprint arXiv:2205.10729*, 2022.

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

Liyu Chen and Haipeng Luo. Near-optimal goal-oriented reinforcement learning in non-stationary environments. *arXiv preprint arXiv:2205.13044*, 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*, 2021a.

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

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, 2021c.

Liyu Chen, Haipeng Luo, and Aviv Rosenberg. Policy optimization for stochastic shortest path. *arXiv preprint arXiv:2202.03334*, 2022.

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

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

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

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 *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

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

Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. *arXiv preprint arXiv:2110.12727*, 2021.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34:15621–15634, 2021.

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

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018a.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM, 2018b.

Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pages 943–950, 2000.

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

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Improved sample complexity for incremental autonomous exploration in MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 11273–11284. Curran Associates, Inc., 2020b.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. A provably efficient sample collection strategy for reinforcement learning. *Advances in Neural Information Processing Systems*, 34:7611–7624, 2021a.

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, 2021b.

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*, 2021c.

Jean Tarbouriech, Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Adaptive multi-goal exploration. In *International Conference on Artificial Intelligence and Statistics*, pages 7349–7383. PMLR, 2022.

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

Mengdi Wang. Randomized linear programming solves the discounted markov decision problem in nearly-linear (sometimes sublinear) running time. *arXiv preprint arXiv:1704.01869*, 2017.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR, 2021.

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.

Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020a.

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

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.

Peng Zhao, Long-Fei Li, and Zhi-Hua Zhou. Dynamic regret of online markov decision processes. In *International Conference on Machine Learning*, pages 26865–26894. PMLR, 2022.

Contents of Appendix

Appendix A. Related Work

Sample complexity (with or without a generative model) is a well-studied topic in finite-horizon MDPs (Dann et al., 2017; Sidford et al., 2018a,b; Dann et al., 2019) and discounted MDPs (Kearns and Singh, 1998; Sidford et al., 2018a,b; Wang, 2017; Li et al., 2020). Apart from computing ϵ-optimal policy for a given cost function, researchers also study obtaining ϵ-optimal policies for an arbitrary sequence of cost functions after interacting with the environment, known as reward-free exploration (Jin et al., 2020; Ménard et al., 2021; Zhang et al., 2020a).

Instead of reaching a single goal state, another line of research considers exploration problems of discovering reachable states (Lim and Auer, 2012; Tarbouriech et al., 2020b, 2021a, 2022; Cai et al., 2022). Sample complexity of SSP is a building block for solving these problems, and existing results only consider strictly positive costs, that is, $c_{\min} > 0$.

Another active research area is learning ϵ-optimal policy purely from an offline dataset, known as offline reinforcement learning (Ren et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021; Yin and Wang, 2021; Yin et al., 2021, 2022). Offline SSP has been recently studied by Yin et al. (2022) where they provide a minimax optimal offline algorithm. While both offline RL and our setting aim to recover an ϵ-optimal policy, there are important differences. In offline SSP (Yin et al., 2022) , the samples are collected by a behavior policy with bounded coverage (i.e., maximum ratio between the state-action distribution of the optimal and behavior policy) while in sample complexity the algorithm is responsible of deciding the sample collection strategy. Furthermore, the analysis in (Yin et al., 2022) is limited to positive costs (i.e., $c_{\min} > 0$) and their sample complexity is measured in terms of number of trajectories and coverage bound. These terms hide both the dependence in terms of action space and, most importantly, the horizon. Our analysis provides a much more comprehensive understanding of sample complexity in SSP.

Appendix B. Preliminaries

Extra Notations For any distribution $P \in \Delta_{\mathcal{S}_+}$ and function $V \in \mathbb{R}^{\mathcal{S}_+}$, define $PV = \mathbb{E}_{S\sim P}[V(S)]$ and $\mathbb{V}(P, V) = \textsc{Var}_{S\sim P}[V(S)]$ as the expectation and the variance of $V(S)$ with S sampled from P respectively.

Table 2: The notation adopted in this paper.

Symbol	Meaning
S	number of states
A	number of actions
$\mathcal{S}_+ = \mathcal{S} \cup \{g\}$	extended state space (g included)
$\pi \in \Delta_{\mathcal{A}}$	a (stationary) policy
$T^\pi(s)$	expected number of steps it takes to reach g (hitting time) starting from state s and following π
$V^\pi = \mathbb{E}_\pi[\sum_i^\infty c(s_i, a_i)\|s_1 = s]$	expected cost of following policy π starting from state s (value function of π)
Π	the set of stationary policies
Π_∞	the set of stationary proper policies
$\pi^\star = \operatorname{argmin}_{\pi\in\Pi_\infty} V^\pi(s)$ for all $s \in \mathcal{S}$	optimal proper policy
$V^\star = V^{\pi^\star}$	value function of optimal policy
$B_\star = \max_s V^\star(s)$	maximum expected cost of the optimal policy starting from any state
$T_\star = \max_s T^{\pi^\star}(s)$	maximum hitting time of the optimal policy starting from any state
$D = \max_s \min_{\pi\in\Pi} T^\pi(s)$	diameter of the SSP instance
$T_\ddagger = B_\star / c_{\min}$	a worst-case upper bound on the hitting time $T_\star$ of the optimal policy
$\overline{T}$	an upper bound of $T_\star$ known to the learner
$\Pi_T = \{\pi \in \Pi : \|T^\pi\|_\infty \le T\}$	set of policies with maximum hitting time upper bounded by T
$\pi^\star_{T,s} = \operatorname{argmin}_{\pi\in\Pi_T} V^\pi(s)$	optimal policy starting from state s restricted in Π_T
$V^{\star,T}(s) = V^{\pi^\star_{T,s}}(s)$	expected cost of $\pi^\star_{T,s}$ starting from state s
$B_{\star,T} = \max_s V^{\star,T}(s)$	maximum expected cost of optimal policies in Π_T starting from any state s
J	cost to directly reach the goal from any state under Assumption 1

Appendix C. Omitted Details in Section 3

In this section we provide omitted proofs and discussions in Section 3.

C.1. Proof of Theorem 2 and Theorem 3

It suffices to prove Theorem 3 and the second statement in Theorem 2, since Theorem 3 subsumes the first statement of Theorem 2. We decompose the proof into two cases: 1) $\min\{T_\ddagger, \overline{T}\} < \infty$, and 2) $\min\{T_\ddagger, \overline{T}\} = \infty$, and we prove each case in a separate theorem.

C.1.1. LOWER BOUND FOR $\min\{T_\ddagger, \overline{T}\} < \infty$

In case there is a finite upper bound on the hitting time of optimal policy, we construct a hard instance adapted from (Mannor and Tsitsiklis, 2004).

Proof [of Theorem 3] Without loss of generality, we assume $S = \frac{A^l - 1}{A-1}$ for some $l \geq 0$. It is clear that $l \leq \log_A S + 1$. We construct an MDP $\mathcal{M}_0$ of full A-ary tree structure: the root node is s_0, each action at a non-leaf node transits to one of its children with cost $c_{\min}$, and we denote the set of leaf nodes by $\mathcal{S}'$. Since $A \geq 3$, we have $|\mathcal{S}'| \geq \frac{S}{2}$. The action space is $\mathcal{A} = [A]$, and we partition the state-action pairs in $\mathcal{S}'$ into two parts: $\Lambda_0 = \mathcal{S}' \times [1]$ and $\Lambda = \mathcal{S}' \times \{2, \dots, A\} = \{(s_1, a_1), \dots, (s_N, a_N)\}$, where $N = |\mathcal{S}'|(A-1)$ (note that here we index state-action pair instead of state, so s_i, s_j with $i \neq j$ may refer to the same state in $\mathcal{S}'$). Now define $T_1 = \min\{\overline{T}/2, B/c_{\min}\}$. The cost function satisfies $c(s, 1) = \frac{B}{T_0}$ for $s \in \mathcal{S}'$, and $c(s_i, a_i) = \frac{B}{T_1}$ for $i \in [N]$. The transition function satisfies $P(g|s, 1) = \frac{1}{T_0} + \frac{T_1 \alpha}{2T_0}$, $P(s|s, 1) = 1 - P(g|s, 1)$ for $s \in \mathcal{S}'$, and $P(g|s_i, a_i) = \frac{1}{T_1}$, $P(s_i|s_i, a_i) = 1 - P(g|s_i, a_i)$ for $i \in [N]$, where $\alpha = \frac{32\epsilon}{T_1 B}$.

Now we consider a class of alternative MDPs $\{\mathcal{M}_i\}_{i=1}^N$. The only difference between $\mathcal{M}_0$ and $\mathcal{M}_i$ is that the transition of $\mathcal{M}_i$ at (s_i, a_i) satisfies $P(g|s_i, a_i) = \frac{1}{T_1} + \alpha$ and $P(s_i|s_i, a_i) = 1 - P(g|s_i, a_i)$, that is, (s_i, a_i) is a "good" state-action pair at $\mathcal{M}_i$. Denote by $B_\star^i$ and $T_\star^i$ the value of $B_\star$ and $T_\star$ in $\mathcal{M}_i$ respectively. For $i \in \{0, \dots, N\}$, we have $\frac{B}{2} \leq B_\star^i \leq c_{\min} \cdot l + B \leq 2B$ by $\alpha \leq \frac{1}{2T_1}$ and $B \geq 2$; $\frac{T_0}{2} \leq T_\star^i \leq l + T_1 \leq \overline{T}$; and $c_{\min}$ is indeed the minimum cost by $c_{\min} \leq \frac{B}{T_1}$. It is not hard to see that at s_0, the optimal behavior in $\mathcal{M}_0$ is to reach any leaf node and then take action 1 until g is reached; while in $\mathcal{M}_i$ for $i \in \{1, \dots, N\}$, the optimal behavior is to reach s_i and then take (s_i, a_i) until g is reached. Thus, $T_\star^0 = \Theta(T_0)$ and $T_\star^i = \Theta(T_1)$ for $i \in [N]$.

Without loss of generality, we consider learning algorithms that output a deterministic policy, which can be represented by $\widehat{v} \in \mathcal{S}' \times \mathcal{A}$ the unique state-action pair in $\mathcal{S}'$ reachable by following the output policy starting from s_0. Define event $\mathcal{E}_1 = \{\widehat{v} \in \Lambda_0\}$. Below we fix a $z \in [N]$. Let $\widehat{T}$ be the number of times the learner samples (s_z, a_z), and K_t be the number of times the agent observes (s_z, a_z, g) among the first t samples of (s_z, a_z). We introduce event

$$\mathcal{E}_2 = \left\{ \max_{1 \leq t \leq t^\star} |pt - K_t| \leq \varepsilon \right\},$$

where $\varepsilon = \sqrt{2p(1-p)t^\star \ln \frac{d}{\theta}}$, $p = \frac{1}{T_1}$, $d = e^4$, $\theta = \exp(-d'\alpha^2 t^\star / (p(1-p))) < 1$ for some $t^\star > 0$ to be specified later, and $d' = 128$. Also define events $\mathcal{E}_3 = \{\widehat{T} \leq t^\star\}$ and $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. For each $i \in \{0, \dots, N\}$, we denote by P_i and $\mathbb{E}_i$ the probability and expectation w.r.t $\mathcal{M}_i$ respectively.

Below we introduce two lemmas that characterize the behavior of the learner if it gathers insufficient samples on (s_z, a_z).

Lemma 12 *If $P_0(\mathcal{E}_3) \geq \frac{7}{8}$, then $P_0(\mathcal{E}_2 \cap \mathcal{E}_3) \geq \frac{3}{4}$.*

Proof Note that in $\mathcal{M}_0$, the probability of observing (s_z, a_z, g) is p for each sample of (s_z, a_z). Thus, $pt - K_t$ is a sum of i.i.d random variables, and the variance of $pt - K_t$ for $t = t^\star$ is $t^\star p(1-p)$. By Kolmogorov's inequality, we have

$$P_0(\mathcal{E}_2) = P_0\left(\max_{1 \leq t \leq t^\star} |pt - K_t| \leq \varepsilon\right) \geq 1 - \frac{t^\star p(1-p)}{2p(1-p)t^\star \ln \frac{d}{\theta}} = 1 - \frac{1}{2\ln \frac{d}{\theta}} \geq \frac{7}{8}.$$

Thus, $P_0(\mathcal{E}_2 \cap \mathcal{E}_3) = P_0(\mathcal{E}_2) + P_0(\mathcal{E}_3) - P_0(\mathcal{E}_2 \cup \mathcal{E}_3) \geq \frac{3}{4}$. ■

Lemma 13 *If $P_0(\mathcal{E}_3) \geq \frac{7}{8}$ and $P_0(\mathcal{E}_1) \geq 1 - \frac{\theta}{2d}$, then $P_z(\mathcal{E}_1) \geq \frac{\theta}{2d}$.*

Proof The range of ϵ ensures that $\alpha \leq \frac{p}{2} \leq \frac{1-p}{2}$. By the assumptions of this lemma and Lemma 12, we have $P_0(\mathcal{E}_1) \geq 1 - \frac{1}{2d} \geq \frac{7}{8}$ by $d \leq \frac{1}{16}$, and thus $P_0(\mathcal{E}) \geq \frac{1}{2}$. Let W be the interaction history of the learner and the generative model, and $L_j(w) = P_j(W = w)$ for $j \in \{0, \ldots, N\}$. Note that the next-state distribution is identical in $\mathcal{M}_0$ and $\mathcal{M}_z$ unless (s_z, a_z) is sampled. Define $K = K_{\widehat{T}}$. We have

$$\begin{aligned}\frac{L_z(W)}{L_0(W)} &= \frac{(p+\alpha)^K (1-p-\alpha)^{\widehat{T}-K}}{p^K (1-p)^{\widehat{T}-K}} = \left(1 + \frac{\alpha}{p}\right)^K \left(1 - \frac{\alpha}{1-p}\right)^{\widehat{T}-K} \\ &= \left(1 + \frac{\alpha}{p}\right)^K \left(1 - \frac{\alpha}{1-p}\right)^{K(\frac{1}{p}-1)} \left(1 - \frac{\alpha}{1-p}\right)^{\widehat{T}-\frac{K}{p}}.\end{aligned}$$

By $1 - u \geq e^{-u-u^2}$ for $u \in [0, \frac{1}{2}]$, $e^{-u} \geq 1 - u$, and $\alpha \leq \frac{1-p}{2}$, we have

$$\begin{aligned}\left(1 - \frac{\alpha}{1-p}\right)^{\frac{1}{p}-1} &\geq \exp\left(\frac{1-p}{p}\left(-\frac{\alpha}{1-p} - \left(\frac{\alpha}{1-p}\right)^2\right)\right) = \exp\left(-\frac{\alpha}{p}\right)\exp\left(-\frac{\alpha^2}{p(1-p)}\right) \\ &\geq \left(1 - \frac{\alpha}{p}\right)\left(1 - \frac{\alpha^2}{p(1-p)}\right).\end{aligned}$$

Therefore, conditioned on $\mathcal{E}$, we have

$$\begin{aligned}\frac{L_z(W)}{L_0(W)}\mathbb{I}_{\mathcal{E}} &\geq \left(1 - \frac{\alpha^2}{p^2}\right)^K \left(1 - \frac{\alpha^2}{p(1-p)}\right)^K \left(1 - \frac{\alpha}{1-p}\right)^{\widehat{T}-\frac{K}{p}} \mathbb{I}_{\mathcal{E}} \\ &\geq \left(1 - \frac{\alpha^2}{p^2}\right)^{p\widehat{T}+\varepsilon} \left(1 - \frac{\alpha^2}{p(1-p)}\right)^{p\widehat{T}+\varepsilon} \left(1 - \frac{\alpha}{1-p}\right)^{\frac{\varepsilon}{p}} \mathbb{I}_{\mathcal{E}},\end{aligned}$$

where in the last inequality we apply $|p\widehat{T} - K| \leq \varepsilon$ by $\mathcal{E}_2$ and $\mathcal{E}_3$. Then by $1 - u \geq e^{-2u}$ for $u \in [0, \frac{1}{2}]$ and $\alpha \leq \frac{p}{2} \leq \frac{1-p}{2}$, we have

$$\begin{aligned}
\frac{L_z(W)}{L_0(W)}\mathbb{I}_{\mathcal{E}} &\geq \exp\left(-2\left(\frac{\alpha^2}{p^2}(p\widehat{T}+\varepsilon) + \frac{\alpha^2}{p(1-p)}(p\widehat{T}+\varepsilon) + \frac{\alpha}{1-p}\frac{\varepsilon}{p}\right)\right) \\
&\geq \exp\left(-2\left(\frac{\widehat{T}\alpha^2}{p} + \frac{\widehat{T}\alpha^2}{1-p} + \frac{\alpha^2\varepsilon}{p^2} + \frac{\alpha^2\varepsilon}{p(1-p)} + \frac{\alpha\varepsilon}{p(1-p)}\right)\right)\mathbb{I}_{\mathcal{E}} \\
&\geq \exp\left(-2\left(\frac{1}{d'}\ln\frac{1}{\theta} + \frac{3\alpha\varepsilon}{p(1-p)}\right)\right)\mathbb{I}_{\mathcal{E}} \qquad (\widehat{T} \leq t^\star,\ \alpha^2 t^\star = \tfrac{p(1-p)}{d'}\ln\tfrac{1}{\theta}, \text{ and } \alpha < p) \\
&\overset{\text{(i)}}{\geq} \exp\left(-2\left(\frac{1}{d'}\ln\frac{1}{\theta} + 3\sqrt{\frac{2}{d'}}\ln\frac{d}{\theta}\right)\right)\mathbb{I}_{\mathcal{E}} \geq \left(\frac{\theta}{d}\right)^{2\left(1/d' + 3\sqrt{2/d'}\right)}\mathbb{I}_{\mathcal{E}} \geq \frac{\theta}{d}\mathbb{I}_{\mathcal{E}},
\end{aligned}$$

where in (i) we apply the definition of ε and $\alpha^2 t^\star = \frac{p(1-p)}{d'}\ln\frac{1}{\theta}$ to have

$$\frac{\alpha\varepsilon}{p(1-p)} = \sqrt{\frac{2t^\star\alpha^2}{p(1-p)}\ln\frac{d}{\theta}} = \sqrt{\frac{2}{d'}\ln\frac{1}{\theta}\ln\frac{d}{\theta}} \leq \sqrt{\frac{2}{d'}}\ln\frac{d}{\theta}.$$

Then we have

$$P_z(\mathcal{E}_1) \geq P_z(\mathcal{E}) = \mathbb{E}_z[\mathbb{I}_{\mathcal{E}}(W)] = \mathbb{E}_0\left[\frac{L_z(W)}{L_0(W)}\mathbb{I}_{\mathcal{E}}(W)\right] \geq \frac{\theta}{d}P_0(\mathcal{E}) \geq \frac{\theta}{2d}.$$

This completes the proof. ■

Now for any $\delta \in (0, \frac{1}{2d})$, let $t^\star = \frac{B^2(T_1-1)}{32^2 d'\epsilon^2}\ln\frac{1}{2d\delta}$. This gives $\frac{\theta}{2d} = \delta$.

Lemma 14 *An (ϵ, δ)-correct algorithm with $\epsilon \in (0, \frac{1}{32})$ and $\delta \in (0, \frac{1}{2e^4})$ must have $P_0(\mathcal{E}_3) < \frac{7}{8}$.*

Proof Denote by π_0 a deterministic policy such that when following π_0 starting from s_0, it reaches some state in $\mathcal{S}'$ and then takes action 1 until reaching the goal state g. For any $j \in [N]$, denote by π_j a deterministic policy such that when following π_j starting from s_0, it reaches s_j and then takes action a_j until reaching the goal state g. It is not hard to see that π_j is an optimal policy in $\mathcal{M}_j$ for $j \in \{0, \ldots, N\}$ starting from s_0. Denote by V_i^j the value function of π_j in $\mathcal{M}_i$ and $V_i^\star$ the optimal value function of $\mathcal{M}_i$. By the choice of α, we have $V_0^j(s_0) - V_0^\star(s_0) = B - \frac{B}{1+T_1\alpha/2} > \epsilon$ for any $j \in [N]$. Thus, all ϵ-optimal deterministic policies in $\mathcal{M}_0$ have the same behavior as π_0 starting from s_0. On the other hand, $V_z^0(s_0) - V_z^\star(s_0) = B(\frac{1}{1+T_1\alpha/2} - \frac{1}{1+T_1\alpha}) > \epsilon$. Thus, all ϵ-optimal deterministic policies in $\mathcal{M}_z$ have the same behavior as π_z starting from s_0. Therefore, an (ϵ, δ)-correct algorithm should guarantee $P_0(\mathcal{E}_1) \geq 1 - \delta$ and $P_z(\mathcal{E}_1) < \delta$. When $P_0(\mathcal{E}_3) \geq \frac{7}{8}$, this leads to a contradiction by Lemma 13 and the choice of $t^\star$. Thus, $P_0(\mathcal{E}_3) < \frac{7}{8}$. ■

We are now ready to prove the main statement of Theorem 3. Note that in $\mathcal{M}_0$, an (ϵ, δ)-correct algorithm should guarantee $P_0(\widehat{T}_z \leq t^\star) < \frac{7}{8}$ for any $z \in [N]$ by Lemma 14, where $\widehat{T}_z$ is the number of times the learner samples (s_z, a_z). Define $\mathcal{N} = \sum_z \mathbb{I}\{\widehat{T}_z \leq t^\star\}$. Clearly, we have $\mathcal{N} \leq N$ and $\mathbb{E}_0[\mathcal{N}] < \frac{7N}{8}$. Moreover, $P_0(\mathcal{N} \geq \frac{8N}{9}) \leq \frac{63}{64}$ by Markov's inequality. This implies that with

probability at least $\frac{1}{64} > \frac{1}{2e^4}$, we have $|\{z \in [N] : \widehat{T}_z > t^\star\}| > \frac{N}{9}$ and thus the total number of samples used in $\mathcal{M}_0$ is at least $\frac{Nt^\star}{9}$. To conclude, there is no (ϵ, δ)-correct algorithm with $\epsilon \in (0, \frac{1}{32})$, $\delta \in (0, \frac{1}{2e^4})$, and sample complexity $\frac{Nt^\star}{9} = \Omega(\frac{NB^2T_1}{\epsilon^2} \ln \frac{1}{\delta}) = \Omega(\min\{\frac{B_\star}{c_{\min}}, \overline{T}\}\frac{B_\star^2 SA}{\epsilon^2} \ln \frac{1}{\delta})$ on $\mathcal{M}_0$ by $B_\star = \Theta(B)$ and the definition of T_1. This completes the proof. ■

Remark 15 *Note that $\overline{T}$ is both a parameter of the environment and the knowledge given to the learner. In fact, $\overline{T}$ constrains the hitting time of the optimal policy in all the possible alternative MDPs $\{\mathcal{M}_j\}_{j\in[N]}$, which affects the final lower bound. Also note that the lower bound holds even if the learner has access to an upper bound of $B_\star$ (which is $2B$ in the proof above).*

Why a faster rate is impossible with $\overline{T} \geq T_\star$? This result may seem unintuitive because when we have knowledge of $\overline{T} \geq T_\star$, a finite-horizon reduction with horizon $\tilde{\mathcal{O}}(\overline{T})$ ensures that the estimation error shrinks at rate $B_\star\sqrt{T_\star/n}$ (Yin and Wang, 2021, Figure 1), where n is the number of samples for each state-action pair. Then it seems that it might be possible to obtain a sample complexity of order $\frac{T_\star B_\star^2 SA}{\epsilon^2}$. However, our lower bound indicates that the sample complexity should scale with $\overline{T}$ instead of $T_\star$. An intuitive explanation is that even if the estimation error shrinks with rate $T_\star$ in hindsight, since the learner doesn't know the exact value of $T_\star$, it can only set n w.r.t $\overline{T}$ so that the output policy is ϵ-optimal even in the worst case of $\overline{T} = T_\star$.

C.1.2. LOWER BOUND FOR $\min\{T_\ddagger, \overline{T}\} = \infty$

Now we show that when there is no finite upper bound on $T_\star$, it really takes infinite number of samples to learn in the worst scenario.

Theorem 16 (Second statemnt of Theorem 2) *There exist an SSP instance with $c_{\min} = 0$, $T_\star = 1$, and $B_\star = 1$ in which every (ϵ, δ)-correct algorithm with $\epsilon \in (0, \frac{1}{2})$ and $\delta \in (0, \frac{1}{16})$ has infinite sample complexity.*

Proof Consider an SSP $\mathcal{M}_0$ with $\mathcal{S} = \{s_0, s_1\}$ and $\mathcal{A} = \{a_0, a_g\}$. The cost function satisfies $c(s_0, a_0) = 0$, $c(s_0, a_g) = \frac{1}{2}$, and $c(s_1, a) = 1$ for all a. The transition function satisfies $P(g|s_0, a_g) = 1$, $P(s_0|s_0, a_0) = 1$, and $P(g|s_1, a) = 1$ for all a; see Figure 1 (b). Clearly $c_{\min} = 0$, $B_\star = T_\star = 1$, and $V^\star(s_0) = \frac{1}{2}$ in $\mathcal{M}_0$. Without loss of generality, we consider learning algorithm that outputs deterministic policy $\widehat{\pi}$ and define events $\mathcal{E}_1 = \{\widehat{\pi}(s_0) = a_0\}$ and $\mathcal{E}_1' = \{\widehat{\pi}(s_0) = a_g\}$.

If a learning algorithm is (ϵ, δ)-correct with $\delta \in (0, \frac{1}{8})$ and has sample complexity $n \in [2, \infty)$ on $\mathcal{M}_0$, then consider two alternative MDPs $\mathcal{M}_+$ and $\mathcal{M}_-$. MDP $\mathcal{M}_+$ is the same as $\mathcal{M}_0$ except that $P(s_1|s_0, a_0) = \frac{1}{n}$ and $P(s_0|s_0, a_0) = 1 - \frac{1}{n}$. MDP $\mathcal{M}_-$ is the same as $\mathcal{M}_0$ except that $P(g|s_0, a_0) = \frac{1}{n}$ and $P(s_0|s_0, a_0) = 1 - \frac{1}{n}$. Note that in $\mathcal{M}_+$, the optimal proper policy takes a_g at s_0, and $V^\star(s_0) = \frac{1}{2}$; while in $\mathcal{M}_-$, the optimal proper policy takes a_0 at s_0, and $V^\star(s_0) = 0$. Let W be the interaction history between the learner and the generative model, and define $L_j(w) = P_j(W = w)$ for $j \in \{0, +, -\}$, where P_j is the probability w.r.t $\mathcal{M}_j$. Also let $\widehat{T}$ be the number of times the learner samples (s_0, a_0) before outputting $\widehat{\pi}$, and $\gamma(w) = \mathbb{I}\{L_0(w) > 0\}$. Define $\mathcal{E}_2 = \{\widehat{T} \leq n\}$, $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$ and $\mathcal{E}' = \mathcal{E}_1' \cap \mathcal{E}_2$. For any $j \in \{+, -\}$, we have $\frac{L_j(W)}{L_0(W)}\mathbb{I}_{\mathcal{E}}(W)\gamma(W) =$

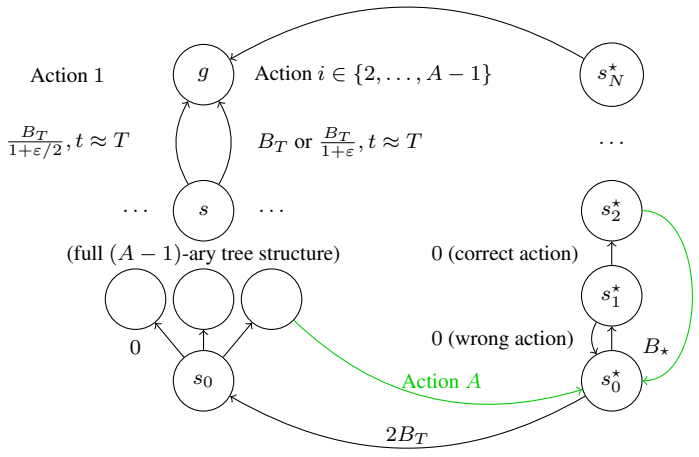

Figure 2: Hard instance in Theorem 6. Each arrow represents a possible transition of a state-action pair, and the value on the side is the expected cost of taking this state-action pair until the transition happens. Value t represents the expected number of steps needed for the transition to happen.

$(1-\frac{1}{n})^{\widehat{T}}\mathbb{I}_{\mathcal{E}}(W)\gamma(W) \geq (1-\frac{1}{n})^{n}\mathbb{I}_{\mathcal{E}}(W)\gamma(W) \geq \frac{\mathbb{I}_{\mathcal{E}}(W)\gamma(W)}{4}$. Thus,

$$P_j(\mathcal{E}) = \mathbb{E}_j[\mathbb{I}_{\mathcal{E}}(W)] \geq \mathbb{E}_j[\mathbb{I}_{\mathcal{E}}(W)\gamma(W)] = \mathbb{E}_0\left[\frac{L_j(W)}{L_0(W)}\mathbb{I}_{\mathcal{E}}(W)\gamma(W)\right] \geq \frac{P_0(\mathcal{E})}{4}.$$

By a similar arguments, we also have $P_j(\mathcal{E}') \geq P_0(\mathcal{E}')/4$ for $j \in \{+,-\}$. Now note that $P_0(\mathcal{E}_2) \geq \frac{7}{8}$ by the sample complexity of the learner. Since $\mathcal{E} \cup \mathcal{E}' = \mathcal{E}_2$ and $\mathcal{E} \cap \mathcal{E}' = \varnothing$, we have $P_0(\mathcal{E}) \geq \frac{7}{16}$ or $P_0(\mathcal{E}') \geq \frac{7}{16}$. Combining with $P_j(\mathcal{E}) \geq P_0(\mathcal{E})/4$ and $P_j(\mathcal{E}') \geq P_0(\mathcal{E}')/4$, we have either $P_j(\mathcal{E}) \geq \frac{7}{64}$ for $j \in \{+,-\}$, or $P_j(\mathcal{E}') \geq \frac{7}{64}$ for $j \in \{+,-\}$. In the first case, in $\mathcal{M}_+$, we have $V^{\widehat{\pi}}(s_0) - V^\star(s_0) = 1 - \frac{1}{2} = \frac{1}{2}$ with probability at least $\frac{7}{64}$. In the second case, in $\mathcal{M}_-$, we have $V^{\widehat{\pi}}(s_0) - V^\star(s_0) = \frac{1}{2} - 0 = \frac{1}{2}$ with probability at least $\frac{7}{64}$. Therefore, for any $\epsilon \in (0, \frac{1}{2})$ and $\delta \in (0, \frac{1}{16})$, there is a contradiction in both cases if the learner is (ϵ, δ)-correct and has finite sample complexity on $\mathcal{M}_0$. This completes the proof. ■

Remark 17 *Note that although $T_\star = 1$ in $\mathcal{M}_0$, the key of the analysis is that $T_\star$ can be arbitrarily large in the alternative MDPs. Indeed, if we have a finite upper bound $\overline{T}$ of $T_\star$, then the learning algorithm only requires finite number of samples as shown in Theorem 3.*

C.2. Proof of Theorem 6

Proof Without loss of generality, assume that $S = \frac{2((A-1)^l-1)}{A-2}$ for some $l \geq 0$. Consider a family of MDPs $\{\mathcal{M}_{i,j}\}_{i\in\{0,\ldots,N'\},j\in[A-1]^N}$ with state space $\mathcal{S} = \mathcal{S}_T \cup \mathcal{S}_\star$ where $|\mathcal{S}_T| = |\mathcal{S}_\star| = N+1$, $N' = (A-2)(A-1)^l$, and action space $\mathcal{A} = [A]$. States in $\mathcal{S}_T$ forms a full $(A-1)$-ary tree on action subset $[A-1]$ as in Theorem 3 with root s_0, $\overline{T} = T/3$, $T_0 = T/6$, $B = B_T$, and $c_{\min} = 0$. It is clear that $N' = |\Lambda|$ (defined in Theorem 3) in the tree formed by $\mathcal{S}_T$. The transition of $\mathcal{M}_{i,j}$ in

$\mathcal{S}_T$ corresponds to $\mathcal{M}_i$ in Theorem 3. We denote $\mathcal{S}_\star = \{s_0^\star, \dots, s_N^\star\}$, and for each state in $\mathcal{S}_T$, the remaining unspecified action transits to $s_0^\star$ with cost 0.

Consider another set of MDPs $\{\mathcal{M}_i'\}_{i\in\{0,\dots,N'\}}$ with state space $\mathcal{S}_T$. The transition and cost functions of $\mathcal{M}_i'$ is the same as $\mathcal{M}_i$ in $\mathcal{S}_T$ except that its action space is restricted to $[A-1]$. Theorem 3 implies that there exists constants α_1, α_2, such that any (ϵ', δ')-correct algorithm with $\epsilon' \in (0, \frac{1}{32})$, $\delta' \in (0, \frac{1}{2e^4})$ has sample complexity at least $C(\epsilon', \delta') = \frac{\alpha_1 B_T^2 TSA}{\epsilon'^2} \ln \frac{\alpha_2}{\delta'}$ on $\{\mathcal{M}_i'\}_i$ (note that in Theorem 3 we only show the sample complexity lower bound in $\mathcal{M}_0'$, but it not hard to show a similar bound for other $\mathcal{M}_i'$ following similar arguments). Now we specify the transition and cost functions in $\mathcal{S}_\star$ for each $\mathcal{M}_{i,j}$ such that learning $\{\mathcal{M}_{i,j}\}_{i,j}$ is as hard as learning $\{\mathcal{M}_i'\}_i$. At $s_0^\star$, taking any action suffers cost 1; taking any action in $[A-1]$ transits to $s_1^\star$ with probability $\frac{1}{B_\star}$ and stays at $s_0^\star$ otherwise; taking action A transits to s_0 with probability $\frac{1}{2B_T}$ and stays at $s_0^\star$ otherwise. At $s_n^\star$ for $n \in [N]$, taking any action suffers cost 0; taking action j_n (recall that $j \in [A-1]^N$) transits to $s_{n+1}^\star$ (define $s_{N+1}^\star = g$) with probability $p = \min\{\frac{1}{2T}, \frac{\delta}{4C(\epsilon,4\delta)}\}$ and stays at $s_n^\star$ otherwise; taking any other action in $[A-1]$ transits to $s_0^\star$ with probability p and stays at $s_n^\star$ otherwise; taking action A directly transits to $s_0^\star$; see illustration in Figure 2. Note that any $\mathcal{M}_{i,j}$ has parameters $B_\star$ (transiting to $s_0^\star$ from any state and then reaching g through $\mathcal{S}_\star$) and satisfies $B_{\star,T} \in [\frac{B_T}{2}, 3B_T]$ (transiting from $s_0^\star$ to s_0 and then reaching g through $\mathcal{S}_T$). From now on we fix the learner as an (ϵ, δ, T)-correct algorithm with sample complexity $C(\epsilon, 4\delta) - 1$ on $\{\mathcal{M}_{i,j}\}_{i,j}$. Define $\mathcal{E}_1$ as the event that the first $C(\epsilon, 4\delta)$ samples drawn by the learner from any $(s_n^\star, a)$ with $n \in [N]$ and $a \in [A-1]$ transits to $s_n^\star$, and denote by $P_{i,j}$ the probability distribution w.r.t $\mathcal{M}_{i,j}$. By $1 + x \geq e^{\frac{x}{1+x}}$ for $x \geq -1$ and $e^x \geq 1 + x$, we have for any i, j,

$$P_{i,j}(\mathcal{E}_1) = (1-p)^{C(\epsilon,4\delta)} \geq e^{\frac{-pC(\epsilon,4\delta)}{1-p}} \geq e^{-2pC(\epsilon,4\delta)} \geq e^{-\frac{\delta}{2}} \geq 1 - \frac{\delta}{2}.$$

Also define $\mathcal{E}_2$ as the event that the learner uses at most $C(\epsilon, 4\delta) - 1$ samples, and $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. We have $P_{i,j}(\mathcal{E}_2) \geq 1 - \delta$ by the sample complexity of the learner, and thus $P_{i,j}(\mathcal{E}) \geq 1 - \frac{3}{2}\delta$ for any i, j. We first bound the expected cost of the learner in $\mathcal{S}_\star$ conditioned on $\mathcal{E}$. Denote by $V_{\mathcal{M}}^\pi$ the value function of policy π in $\mathcal{M}$.

Lemma 18 *Given any policy distribution ρ, there exists $j^\star$ such that $\mathbb{E}_{\pi\sim\rho}[\mathbb{I}\{V_{\mathcal{M}_{i,j^\star}}^\pi(s_0^\star) \geq 2B_T\}] \geq \frac{1}{2}$ for any i.*

Proof Below we fix an $i \in [N']$. For any policy π and $j \in [A-1]^N$, define $x_j^\pi = \prod_{n=1}^N p^\pi(j_n|s_n^\star)$ and $y^\pi = \pi(A|s_0^\star)$, where $p^\pi(a|s_n^\star)$ is the probability that when following policy π starting from $s_n^\star$, the last action taken before leaving $s_n^\star$ is a. It is not hard to see that in our construction, p^π is independent of j. Also denote by V_j^π the value function of policy π in $\mathcal{M}_{i,j}$. Note that

$$\begin{aligned}
V_j^\pi(s_0^\star) &\geq 1 + \frac{1-y^\pi}{B_\star} V_j^\pi(s_1^\star) + \left(1 - \frac{y^\pi}{2B_T} - \frac{1-y^\pi}{B_\star}\right) V_j^\pi(s_0^\star) \\
&= 1 + \frac{(1-y^\pi)(1-x_j^\pi)}{B_\star} V_j^\pi(s_0^\star) + \left(1 - \frac{y^\pi}{2B_T} - \frac{1-y^\pi}{B_\star}\right) V_j^\pi(s_0^\star) \\
&= 1 + \left(1 - \frac{y^\pi}{2B_T} - \frac{(1-y^\pi)x_j^\pi}{B_\star}\right) V_j^\pi(s_0^\star).
\end{aligned}$$

Reorganizing terms gives $V_j^\pi(s_0^\star) \geq \frac{1}{y^\pi/(2B_T)+(1-y^\pi)x_j^\pi/B_\star}$. Now if $V_j^\pi(s_0^\star) < 2B_T$, then we have $y^\pi + (1-y^\pi)\frac{2B_T x_j^\pi}{B_\star} > 1$, which gives $x_j^\pi > \frac{B_\star}{2B_T}$. Let $\mathcal{X}^\pi$ be the set of $j \in [A-1]^N$ such that $V_j^\pi(s_0^\star) < 2B_T$. By $\frac{B_\star}{2B_T}|\mathcal{X}^\pi| \leq \sum_j x_j^\pi \leq 1$, we have $|\mathcal{X}^\pi| \leq \frac{2B_T}{B_\star}$. Define $z^\pi(j) = \mathbb{I}\{j \in \mathcal{X}^\pi\}$. We have $\sum_j z^\pi(j) = |\mathcal{X}^\pi| \leq \frac{2B_T}{B_\star}$ for any π, and thus $\sum_j \int_\pi z^\pi(j)\rho(\pi)d\pi \leq \frac{2B_T}{B_\star}$. Therefore, there exists $j^\star$ such that $\int_\pi z^\pi(j^\star)\rho(\pi)d\pi \leq \frac{2B_T}{B_\star(A-1)^N}$, which implies that

$$\mathbb{E}_{\pi\sim\rho}[\mathbb{I}\{V^\pi_{\mathcal{M}_{i,j^\star}}(s_0^\star) \geq 2B_T\}] = 1 - \int_\pi z^\pi(j^\star)\rho(\pi)d\pi \geq 1 - \frac{2B_T}{B_\star(A-1)^N} \geq \frac{1}{2}.$$

The proof is completed by noting that for the picked $j^\star$, the bound above holds for any i, since the lower bound on $V_j^\pi(s_0^\star)$ we applied above is independent of i. ■

Now consider another set of MDPs $\{\mathcal{M}_i''\}_{i\in\{0,\dots,N'\}}$ with state space $\mathcal{S}_T$. The transition and cost functions of $\mathcal{M}_i''$ is the same as $\mathcal{M}_{i,j}$ restricted on $\mathcal{S}_T$ for any j, except that taking action A at any state directly transits to g with cost $2B_T$. We show that any (ϵ',δ')-correct algorithm with $\epsilon' \in (0, \frac{1}{32})$, $\delta' \in (0, \frac{1}{2e^4})$ has sample complexity at least $C(\epsilon',\delta')$ on $\{\mathcal{M}_i''\}_i$ Given any policy π on $\mathcal{M}_i''$, define g_π as a policy on $\mathcal{M}_i'$ and $\mathcal{M}_i''$ such that $g_\pi(a|s) \propto \pi(a|s)$ and $\sum_{a=1}^{A-1} g_\pi(a|s) = 1$. It is straightforward to see that $V^{g_\pi}_{\mathcal{M}_i'}(s) = V^{g_\pi}_{\mathcal{M}_i''}(s) \leq V^\pi_{\mathcal{M}_i''}(s)$ and $V^\star_{\mathcal{M}_i'}(s) = V^\star_{\mathcal{M}_i''}(s)$, where $V^\star_{\mathcal{M}}$ is the optimal value function in $\mathcal{M}$. Thus, if there exists an algorithm $\mathfrak{A}$ that is (ϵ',δ')-correct with sample complexity less than $C(\epsilon',\delta')$ on $\{\mathcal{M}_i''\}_i$, then we can obtain an (ϵ',δ')-correct algorithm on $\{\mathcal{M}_i'\}_i$ with sample complexity less than $C(\epsilon',\delta')$ as follows: executing $\mathfrak{A}$ on $\{\mathcal{M}_i''\}_i$ to obtain policy $\widehat{\pi}$, and then output $g_{\widehat{\pi}}$. This leads to a contradiction to the definition of $C(\cdot,\cdot)$, and thus any (ϵ',δ')-correct algorithm with $\epsilon' \in (0, \frac{1}{32})$, $\delta' \in (0, \frac{1}{2e^4})$ has sample complexity at least $C(\epsilon',\delta')$ on $\{\mathcal{M}_i''\}_i$.

Since we assume that the learner has sample complexity less than $C(\epsilon, 4\delta)$ on $\mathcal{M}_{i,j}$, for a fixed j_0, there exists $i^\star$ such that $P_{i^\star,j_0}(\mathcal{E}_3) > 4\delta$, where $\mathcal{E}_3 = \{\exists s : V^{\widehat{\pi}}_{\mathcal{M}_{i^\star}''}(s) - V^\star_{\mathcal{M}_{i^\star}''}(s) > \epsilon\}$ (note that $\widehat{\pi}$ is computed on $\mathcal{M}_{i^\star,j_0}$, but we can apply $\widehat{\pi}$ restricted on $\mathcal{S}_T$ to $\mathcal{M}_{i^\star}''$). This also implies that $P_{i^\star,j}(\mathcal{E}\cap\mathcal{E}_3) = P_{i^\star,j_0}(\mathcal{E}\cap\mathcal{E}_3) \geq \frac{5\delta}{2}$ for any j, since the value of $P_{i,j}(\omega)$ is independent of j when $\omega \in \mathcal{E}$. Define $\mathcal{E}_4 = \{\exists s : V^{\widehat{\pi}}_{\mathcal{M}}(s) - V^{\star,T}_{\mathcal{M}}(s) > \epsilon\}$. By Lemma 18, there exists $j^\star$ such that

$$P_{i^\star,j^\star}(\mathcal{E}_4|\mathcal{E}\cap\mathcal{E}_3) \geq \mathbb{E}_{\widehat{\pi}\sim P_{i^\star}(\cdot|\mathcal{E}\cap\mathcal{E}_3)}[\mathbb{I}\{V^{\widehat{\pi}}_{\mathcal{M}_{i^\star,j^\star}}(s_0^\star) \geq 2B_T\}] \geq \frac{1}{2},$$

since the distribution of $\widehat{\pi}$ is independent of j under $\mathcal{E}\cap\mathcal{E}_3$, $V^{\star,T}_{\mathcal{M}_{i^\star,j}}(s) = V^\star_{\mathcal{M}_{i^\star}''}(s)$ for any j and $s \in \mathcal{S}_T$, and $V^{\widehat{\pi}}_{\mathcal{M}_{i^\star,j}}(s) \geq V^{\widehat{\pi}}_{\mathcal{M}_{i^\star}''}(s)$ for $s \in \mathcal{S}_T$ when $V^{\widehat{\pi}}_{\mathcal{M}_{i^\star,j}}(s_0^\star) \geq 2B_T$. Putting everything together, we have $P_{i^\star,j^\star}(\mathcal{E}_4) \geq P_{i^\star,j^\star}(\mathcal{E}_4\cap\mathcal{E}\cap\mathcal{E}_3) > \delta$, a contradiction. Therefore, there is no (ϵ,δ,T)-correct algorithm with sample complexity less than $C(\epsilon,4\delta)$ on $\{\mathcal{M}_{i,j}\}_{i,j}$. In other words, for any (ϵ,δ,T)-correct algorithm, there exists $\mathcal{M} \in \{\mathcal{M}_{i,j}\}_{i,j}$ such that this algorithm has sample complexity at least $C(\epsilon,4\delta)$ on $\mathcal{M}$. This completes the proof. ■

Appendix D. Omitted Details in Section 4

In this section, we present the omitted proofs of Lemma 7 and Theorem 8. To prove Theorem 8, we first discuss the guarantee of the finite-horizon algorithm in Appendix D.2. Then, we bound the sample complexity of Algorithm 1 in Appendix D.3.

Algorithm 3 LCBVI $(H, \mathbf{N}, B, c_f, \delta)$

Input: horizon H, counter $\mathbf{N}$, optimal value function upper bound B, terminal cost c_f, failure probability δ, and cost function c,

Define: $\bar{P}_{s,a}(s') = \frac{\mathbf{N}(s,a,s')}{\mathbf{N}^+(s,a)}$ and $b(s,a,V) = \max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a},V)\iota}{\mathbf{N}^+(s,a)}}, \frac{49B\iota}{\mathbf{N}^+(s,a)}\right\}$, where $\iota = \ln\frac{2SAHn}{\delta}$, $n = \sum_{s,a}\mathbf{N}(s,a)$, and $\mathbf{N}^+(s,a) = \max\{1, \mathbf{N}(s,a)\}$.

Initialize: $\widehat{V}_{H+1} = c_f$.

for $h = H, \dots, 1$ **do**

1 $\quad \widehat{Q}_h(s,a) = \left(c(s,a) + \bar{P}_{s,a}\widehat{V}_{h+1} - b(s,a,\widehat{V}_{h+1})\right)_+$.

$\quad \widehat{V}_h(s) = \min_a \widehat{Q}_h(s,a)$.

end

Output: $(\widehat{\pi}, \widehat{V})$ with $\widehat{\pi}(s,h) = \operatorname{argmin}_a \widehat{Q}_h(s,a)$.

D.1. Proof of Lemma 7

Proof Let $V^\pi_{1,h}$ be the value function V^π_1 in $\mathcal{M}_{h,c_f}$. For any $n \geq 0$, we have

$$\begin{aligned} V^\pi_{1,(n+1)H}(s) &= \mathbb{E}_\pi\left[\sum_{i=1}^{nH} c(s_i,a_i) + V^\pi_{1,H}(s_{nH+1}) \,\middle|\, s_1 = s\right] \\ &\leq \mathbb{E}_\pi\left[\sum_{i=1}^{nH} c(s_i,a_i) + c_f(s_{nH+1}) \,\middle|\, s_1 = s\right] = V^\pi_{1,nH}(s). \end{aligned}$$

Therefore, $V^\pi(s) \leq \lim_{n\to\infty} V^\pi_{1,nH}(s) \leq V^\pi_{1,H}(s)$ and this completes the proof. Note that the first inequality may be strict. Indeed, $V^\pi = \lim_{H\to\infty} V^\pi_{1,H}$ in $\mathcal{M}_{H,0}$. Consider an improper policy π behaving in a loop with zero cost. Then, $V^\pi = 0$ but $\lim_{H\to\infty} V^\pi_{1,H} = c_f$ in $\mathcal{M}_{H,c_f}$. ■

D.2. Guarantee of the Finite-Horizon Algorithm LCBVI

In this section, we discuss and prove the guarantee of Algorithm 3.

Notations Within this section, H, $\mathbf{N}$, B, c_f, δ are inputs of Algorithm 3, and $\widehat{\pi}$, $\widehat{Q}$, $\widehat{V}$, $\bar{P}_{s,a}$, $\mathbf{N}$, $\mathbf{N}^+$, ι, and b are defined in Algorithm 3. Value function V^π_h is w.r.t MDP $\mathcal{M}_{H,c_f}$, and we denote by $V^\star_h, Q^\star_h$ the optimal value function and action-value function, such that $V^\star_h(s) = \operatorname{argmin}_\pi V^\pi_h(s)$ and $Q^\star_h(s,a) = c(s,a) + P_{s,a}V^\star_{h+1}$ for $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. We also define $V^\pi_{H+1} = V^\star_{H+1} = c_f$ for any policy π, and $B^\star_H = \max_{h\in[H+1]}\|V^\star_h\|_\infty$. For any $(\bar{s},\bar{h}) \in \mathcal{S}\times[H]$ and $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, denote by $q_{\pi,(\bar{s},\bar{h})}(s,a,h)$ the probability of visiting (s,a) in stage h if the learner starts in state $\bar{s}$ in stage $\bar{h}$ and follows policy π afterwards. For any value function $V \in \mathbb{R}^{\mathcal{S}\times[H+1]}$, define $\|V_\cdot\|_\infty = \max_{h\in[H+1]}\|V_h\|_\infty$.

We first prove optimism of the estimated value functions.

Lemma 19 *When $B \geq B^\star_H$, we have $\widehat{Q}_h(s,a) \leq Q^\star_h(s,a)$ for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ with probability at least $1-\delta$.*

Proof We prove this by induction. The case of $h = H + 1$ is clearly true. For $h \le H$, note that

$$\begin{aligned}
&c(s,a) + \bar{P}_{s,a}\widehat{V}_{h+1} - b(s,a,\widehat{V}_{h+1}) \le c(s,a) + \bar{P}_{s,a}V^\star_{h+1} - b(s,a,V^\star_{h+1}) && \text{(Lemma 39)}\\
&= c(s,a) + P_{s,a}V^\star_{h+1} + (\bar{P}_{s,a} - P_{s,a})V^\star_{h+1} - \max\left\{7\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^\star_{h+1})\iota}{\mathbf{N}^+(s,a)}}, \frac{49B\iota}{\mathbf{N}^+(s,a)}\right\}\\
&\le c(s,a) + P_{s,a}V^\star_{h+1} + (2\sqrt{2}-3)\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^\star_{h+1})\iota}{\mathbf{N}^+(s,a)}} + (19-24)\frac{B\iota}{\mathbf{N}^+(s,a)}\\
&\qquad\qquad \text{(Lemma 44 and } \max\{a,b\} \ge \tfrac{a+b}{2}\text{)}\\
&\le c(s,a) + P_{s,a}V^\star_{h+1} = Q^\star_h(s,a).
\end{aligned}$$

The proof is then completed by the definition of $\widehat{Q}$. ■

Lemma 20 *For any state $\bar{s} \in \mathcal{S}$ and $\bar{h} \in [H]$, we have*

$$\left|V^{\widehat{\pi}}_{\bar{h}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s})\right| \le \sum_{s,a,h} q_{\widehat{\pi},(\bar{s},\bar{h})}(s,a,h)\left(\left|(P_{s,a} - \bar{P}_{s,a})\widehat{V}_{h+1}\right| + b(s,a,\widehat{V}_{h+1})\right).$$

Proof First note that

$$\begin{aligned}
V^{\widehat{\pi}}_{\bar{h}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s}) &\le P_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})}V^{\widehat{\pi}}_{\bar{h}+1} - \bar{P}_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})}\widehat{V}_{\bar{h}+1} + b(\bar{s},\widehat{\pi}(\bar{s},\bar{h}),\widehat{V}_{\bar{h}+1}) && \text{(definition of } \widehat{\pi} \text{ and } \widehat{V}\text{)}\\
&= P_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})}(V^{\widehat{\pi}}_{\bar{h}+1} - \widehat{V}_{\bar{h}+1}) + (P_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})} - \bar{P}_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})})\widehat{V}_{\bar{h}+1} + b(\bar{s},\widehat{\pi}(\bar{s},\bar{h}),\widehat{V}_{\bar{h}+1})\\
&= \sum_{s,a,h} q_{\widehat{\pi},(\bar{s},\bar{h})}(s,a,h)\left((P_{s,a} - \bar{P}_{s,a})\widehat{V}_{h+1} + b(s,a,\widehat{V}_{h+1})\right).\\
&\qquad\qquad \text{(expand } P_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})}(V^{\widehat{\pi}}_{\bar{h}+1} - \widehat{V}_{\bar{h}+1}) \text{ recursively and } V^{\widehat{\pi}}_{H+1} = \widehat{V}_{H+1}\text{)}
\end{aligned}$$

For the other direction,

$$\begin{aligned}
(\widehat{V}_{\bar{h}}(\bar{s}) - V^{\widehat{\pi}}_{\bar{h}}(\bar{s}))_+ &\le \left(\bar{P}_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})}\widehat{V}_{\bar{h}+1} - P_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})}V^{\widehat{\pi}}_{\bar{h}+1}\right)_+ && ((a)_+ - (b)_+ \le (a-b)_+)\\
&\le P_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})}(\widehat{V}_{\bar{h}+1} - V^{\widehat{\pi}}_{\bar{h}+1})_+ + \left|(P_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})} - \bar{P}_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})})\widehat{V}_{\bar{h}+1}\right| && ((a+b)_+ \le (a)_+ + (b)_+)\\
&\le \sum_{s,a,h} q_{\widehat{\pi},(\bar{s},\bar{h})}(s,a,h)\left|(P_{s,a} - \bar{P}_{s,a})\widehat{V}_{h+1}\right|. && \text{(expand } P_{\bar{s},\widehat{\pi}(\bar{s},\bar{h})}(\widehat{V}_{\bar{h}+1} - V^{\widehat{\pi}}_{\bar{h}+1})_+ \text{ recursively)}
\end{aligned}$$

Combining both directions completes the proof. ■

Remark 21 *Note that the inequality in Lemma 20 holds even if optimism (Lemma 19) does not hold, which is very important for estimating $B_\star$.*

Lemma 22 *There exists a function $N^\star(B', H', \epsilon', \delta') \lesssim \frac{B'^2H'}{\epsilon'^2} + \frac{SB'H'}{\epsilon'} + SH'^2$ such that when $B \ge B^\star_H$ and $\mathbf{N}(s,a) = N \ge N^\star(B,H,\epsilon,\delta)$ for all s,a for some integer N, we have $\left\|V^{\widehat{\pi}}_{\cdot} - V^\star_{\cdot}\right\|_\infty \le \epsilon$ with probability at least $1-\delta$.*

Proof Below we assume that $B \geq B_H^\star$. Fix any state $\bar{s} \in \mathcal{S}$ and $\bar{h} \in [H]$, we write $q_{\widehat{\pi},(\bar{s},\bar{h})}$ as $q_{\widehat{\pi}}$ for simplicity. We have with probability at least $1-4\delta$,

$$\begin{aligned}
&V_{\bar{h}}^{\widehat{\pi}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s}) \\
&\leq \sum_{s,a,h} q_{\widehat{\pi}}(s,a,h) \left(\left|(P_{s,a} - \bar{P}_{s,a}) V_{h+1}^\star\right| + \left|(P_{s,a} - \bar{P}_{s,a})(\widehat{V}_{h+1} - V_{h+1}^\star)\right| + b(s,a,\widehat{V}_{h+1}) \right) \\
&\qquad\qquad (|a+b| \leq |a| + |b| \text{ and Lemma 20}) \\
&\lesssim \sum_{s,a,h} q_{\widehat{\pi}}(s,a,h) \left(\sqrt{\frac{\mathbb{V}(P_{s,a}, V_{h+1}^\star)}{N}} + \frac{SB}{N} + \sqrt{\frac{S\mathbb{V}(P_{s,a}, \widehat{V}_{h+1} - V_{h+1}^\star)}{N}} + \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, \widehat{V}_{h+1})}{N}} \right) \\
&\qquad\qquad (\text{Lemma 25 and } \max\{a,b\} \leq (a)_+ + (b)_+) \\
&\lesssim \sqrt{\frac{H}{N} \sum_{s,a,h} q_{\widehat{\pi}}(s,a,h) \mathbb{V}(P_{s,a}, V_{h+1}^\star)} + \sqrt{\frac{SH}{N} \sum_{s,a,h} q_{\widehat{\pi}}(s,a,h) \mathbb{V}(P_{s,a}, \widehat{V}_{h+1} - V_{h+1}^\star)} + \frac{SBH}{N},
\end{aligned}$$

where in the last inequality we apply $\textsc{Var}[X+Y] \leq 2(\textsc{Var}[X] + \textsc{Var}[Y])$, Cauchy-Schwarz inequality, Lemma 24, and $\sum_{s,a,h} q_{\widehat{\pi}}(s,a,h) \leq H$. Now note that:

$$\begin{aligned}
&\sum_{s,a,h} q_{\widehat{\pi}}(s,a,h) \mathbb{V}(P_{s,a}, V_{h+1}^\star) = \mathbb{E}_{\widehat{\pi}} \left[\sum_{h=\bar{h}}^{H} \mathbb{V}(P_{s_h,a_h}, V_{h+1}^\star) \,\middle|\, s_{\bar{h}} = \bar{s} \right] \\
&= \mathbb{E}_{\widehat{\pi}} \left[\sum_{h=\bar{h}}^{H} \left(P_{s_h,a_h}(V_{h+1}^\star)^2 - (P_{s_h,a_h} V_{h+1}^\star)^2 \right) \,\middle|\, s_{\bar{h}} = \bar{s} \right] \\
&= \mathbb{E}_{\widehat{\pi}} \left[\sum_{h=\bar{h}}^{H} \left(V_{h+1}^\star(s_{h+1})^2 - V_h^\star(s_h)^2 \right) + \sum_{h=\bar{h}}^{H} \left(V_h^\star(s_h)^2 - (P_{s_h,a_h} V_{h+1}^\star)^2 \right) \,\middle|\, s_{\bar{h}} = \bar{s} \right] \\
&\leq B^2 + 3B\mathbb{E}_{\widehat{\pi}} \left[\sum_{h=\bar{h}}^{H} \left(Q_h^\star(s_h,a_h) - P_{s_h,a_h} V_{h+1}^\star \right)_+ \right] \\
&\qquad\qquad (a^2 - b^2 \leq (a+b)(a-b)_+ \text{ for } a,b>0 \text{ and } V^\star(s_h) \leq Q_h^\star(s_h,a_h)) \\
&= B^2 + 3B\mathbb{E}_{\widehat{\pi}} \left[\sum_{h=\bar{h}}^{H} c(s_h,a_h) \,\middle|\, s_{\bar{h}} = \bar{s} \right] = B^2 + 3BV_{\bar{h}}^{\widehat{\pi}}(\bar{s}).
\end{aligned}$$

Plugging this back and by $\mathbb{V}(P_{s,a}, \widehat{V}_{h+1} - V_{h+1}^\star) \leq \left\|\widehat{V}_{h+1} - V_{h+1}^\star\right\|_\infty^2$, we have with probability at least $1-\delta$,

$$\begin{aligned}
0 \leq V_{\bar{h}}^{\widehat{\pi}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s}) &\lesssim B\sqrt{\frac{H}{N}} + \sqrt{\frac{BHV_{\bar{h}}^{\widehat{\pi}}(\bar{s})}{N}} + \sqrt{\frac{SH^2}{N}} \left\|\widehat{V}_\cdot - V_\cdot^\star\right\|_\infty + \frac{SBH}{N} \qquad \text{(Lemma 19)} \\
&\lesssim B\sqrt{\frac{H}{N}} + \sqrt{\frac{BH(V_{\bar{h}}^{\widehat{\pi}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s}))}{N}} + \sqrt{\frac{SH^2}{N}} \left\|\widehat{V}_\cdot - V_\cdot^{\widehat{\pi}}\right\|_\infty + \frac{SBH}{N},
\end{aligned}$$

where in the last step we apply $\widehat{V}_{\bar{h}}(\bar{s}) \le B$ and $\left\|\widehat{V}_\cdot - V_\cdot^\star\right\|_\infty \le \left\|\widehat{V}_\cdot - V_\cdot^{\widehat{\pi}}\right\|_\infty$ since $\widehat{V}(s) \le V^\star(s) \le V^{\widehat{\pi}}(s)$ for all $s \in \mathcal{S}$ by Lemma 19. Solving a quadratic inequality w.r.t $V_{\bar{h}}^{\widehat{\pi}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s})$, we have

$$V_{\bar{h}}^{\widehat{\pi}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s}) \lesssim B\sqrt{\frac{H}{N}} + \sqrt{\frac{SH^2}{N}}\left\|\widehat{V}_\cdot - V_\cdot^{\widehat{\pi}}\right\|_\infty + \frac{SBH}{N}.$$

The inequality above implies that there exist quantity $\overline{N}^\star \lesssim SH^2$, such that when $N \ge \overline{N}^\star$, we have

$$V_{\bar{h}}^{\widehat{\pi}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s}) \lesssim B\sqrt{\frac{H}{N}} + \frac{1}{2}\left\|\widehat{V}_\cdot - V_\cdot^{\widehat{\pi}}\right\|_\infty + \frac{SBH}{N},$$

for any $(\bar{s}, \bar{h})$. Taking maximum of the left-hand-side over $(\bar{s}, \bar{h})$, reorganizing terms and by Lemma 19, we obtain

$$\left\|V_\cdot^{\widehat{\pi}} - V_\cdot^\star\right\|_\infty \le \left\|V_\cdot^{\widehat{\pi}} - \widehat{V}_\cdot\right\|_\infty \lesssim B\sqrt{\frac{H}{N}} + \frac{SBH}{N}. \tag{1}$$

Now define $n^\star = \overline{N}^\star + \inf_n\{\text{right-hand-side of Eq. (1)} \le \epsilon \text{ when } N = n\}$. We have $n^\star \lesssim \frac{B^2H}{\epsilon^2} + \frac{SBH}{\epsilon} + SH^2$. This implies that when $B \ge B_H^\star$ and $\mathbf{N}(s,a) = N \ge n^\star$ for all s, a, we have $\left\|V_\cdot^{\widehat{\pi}} - V_\cdot^\star\right\|_\infty \le \epsilon$ with probability at least $1 - 5\delta$. The proof is then completed by treating $n^\star$ as a function with input B, H, ϵ, δ and replace δ by $\delta/5$ in the arguments above. ■

Lemma 23 *There exists functions $\widehat{N}(B', H', \epsilon', \delta') \lesssim \frac{B'^2SH'}{\epsilon'^2} + \frac{SB'H'}{\epsilon'}$ such that when $\mathbf{N}(s,a) = N \ge \widehat{N}(B, H, \epsilon, \delta)$ for all s, a for some N and $\left\|\widehat{V}_\cdot\right\|_\infty \le B$, we have $\left\|V_\cdot^{\widehat{\pi}} - \widehat{V}_\cdot\right\|_\infty \le \epsilon$ with probability at least $1 - \delta$.*

Proof Below we assume that $\left\|\widehat{V}_\cdot\right\|_\infty \le B$. For any state fixed $\bar{s} \in \mathcal{S}$ and $\bar{h} \in [H]$, we write $q_{\widehat{\pi},(\bar{s},\bar{h})}$ as $q_{\widehat{\pi}}$ for simplicity. Note that with probability at least $1 - 2\delta$,

$$\begin{aligned}
\left|V_{\bar{h}}^{\widehat{\pi}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s})\right| &\le \sum_{s,a,h} q_{\widehat{\pi}}(s,a,h)\left(\left|(P_{s,a} - \bar{P}_{s,a})\widehat{V}_{h+1}\right| + b(s,a,\widehat{V}_{h+1})\right) && \text{(Lemma 20)} \\
&\lesssim \sum_{s,a,h} q_{\widehat{\pi}}(s,a,h)\left(\sqrt{\frac{S\mathbb{V}(P_{s,a},\widehat{V}_{h+1})}{N}} + \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a},\widehat{V}_{h+1})}{N}} + \frac{SB}{N}\right) \\
& && \text{(Lemma 25 and } \max\{a,b\} \le (a)_+ + (b)_+) \\
&\lesssim \sum_{s,a,h} q_{\widehat{\pi}}(s,a,h)\left(\sqrt{\frac{S\mathbb{V}(P_{s,a},\widehat{V}_{h+1})}{N}} + \frac{SB}{N}\right) && \text{(Lemma 24)} \\
&\lesssim \sqrt{\frac{SH}{N}\sum_{s,a,h} q_{\widehat{\pi}}(s,a,h)\mathbb{V}(P_{s,a},\widehat{V}_{h+1})} + \frac{SBH}{N}. \\
& && \text{(Cauchy-Schwarz inequality and } \textstyle\sum_{s,a,h} q_{\widehat{\pi}}(s,a,h) \le H)
\end{aligned}$$

Now note that with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{s,a,h} q_{\widehat{\pi}}(s,a,h)\mathbb{V}(P_{s,a},\widehat{V}_{h+1}) = \mathbb{E}_{\widehat{\pi}}\left[\sum_{h=\bar{h}}^{H}\mathbb{V}(P_{s_h,a_h},\widehat{V}_{h+1})\middle| s_{\bar{h}}=\bar{s}\right] \\
&= \mathbb{E}_{\widehat{\pi}}\left[\sum_{h=\bar{h}}^{H}\left(\widehat{V}_{h+1}(s_{h+1})^2-\widehat{V}_h(s_h)^2\right)+\sum_{h=\bar{h}}^{H}\left(\widehat{V}_h(s_h)^2-(P_{s_h,a_h}\widehat{V}_{h+1})^2\right)\middle| s_{\bar{h}}=\bar{s}\right] \\
&\le B^2+3B\mathbb{E}_{\widehat{\pi}}\left[\sum_{h=\bar{h}}^{H}\left(\widehat{Q}_h(s_h,a_h)-P_{s_h,a_h}\widehat{V}_{h+1}\right)_+\middle| s_{\bar{h}}=\bar{s}\right] \\
&\qquad\qquad (a^2-b^2\le(a+b)(a-b)_+ \text{ for } a,b>0 \text{ and } \widehat{V}_h(s_h)=\widehat{Q}_h(s_h,a_h)) \\
&\le B^2+3B\mathbb{E}_{\widehat{\pi}}\left[\sum_{h=\bar{h}}^{H}\left(c(s_h,a_h)+(\bar{P}_{s_h,a_h}-P_{s_h,a_h})\widehat{V}_{h+1}\right)_+\middle| s_{\bar{h}}=\bar{s}\right] \\
&\qquad\qquad (\text{definition of } \widehat{Q}_h \text{ and } (a)_+-(b)_+\le(a-b)_+) \\
&\lesssim B^2+BV_{\bar{h}}^{\widehat{\pi}}(\bar{s})+B\sqrt{\frac{SH}{N}\sum_{s,a,h}q_{\widehat{\pi}}(s,a,h)\mathbb{V}(P_{s,a},\widehat{V}_{h+1})}+\frac{SB^2H}{N},
\end{aligned}$$

where the last step is by $(a+b)_+\le(a)_++(b)_+$, the definition of $V_{\bar{h}}^{\widehat{\pi}}(\bar{s})$, and

$$\begin{aligned}
&\mathbb{E}_{\widehat{\pi}}\left[\sum_{h=\bar{h}}^{H}\left((\bar{P}_{s_h,a_h}-P_{s_h,a_h})\widehat{V}_{h+1}\right)_+\middle| s_{\bar{h}}=\bar{s}\right] \\
&\lesssim \mathbb{E}_{\widehat{\pi}}\left[\sum_{h=\bar{h}}^{H}\left(\sqrt{\frac{S\mathbb{V}(P_{s_h,a_h},\widehat{V}_{h+1})}{N}}+\frac{SB}{N}\right)\middle| s_{\bar{h}}=\bar{s}\right] \qquad \text{(Lemma 25)} \\
&= \sum_{s,a,h}q_{\widehat{\pi}}(s,a,h)\left(\sqrt{\frac{S\mathbb{V}(P_{s,a},\widehat{V}_{h+1})}{N}}+\frac{SB}{N}\right) \le \sqrt{\frac{SH}{N}\sum_{s,a,h}q_{\widehat{\pi}}(s,a,h)\mathbb{V}(P_{s,a},\widehat{V}_{h+1})}+\frac{SBH}{N}. \\
&\qquad\qquad (\text{Cauchy-Schwarz inequality and } \textstyle\sum_{s,a,h}q_{\widehat{\pi}}(s,a,h)\le H)
\end{aligned}$$

Solving a quadratic inequality w.r.t $\sum_{s,a,h}q_{\widehat{\pi}}(s,a,h)\mathbb{V}(P_{s,a},\widehat{V}_{h+1})$, we have

$$\sum_{s,a,h}q_{\widehat{\pi}}(s,a,h)\mathbb{V}(P_{s,a},\widehat{V}_{h+1})\lesssim B^2+BV_{\bar{h}}^{\widehat{\pi}}(\bar{s})+\frac{SB^2H}{N}.$$

Plugging this back, we have

$$\begin{aligned}
\left|V_{\bar{h}}^{\widehat{\pi}}(\bar{s})-\widehat{V}_{\bar{h}}(\bar{s})\right| &\lesssim B\sqrt{\frac{SH}{N}}+\sqrt{\frac{BSHV_{\bar{h}}^{\widehat{\pi}}(\bar{s})}{N}}+\frac{SBH}{N} \\
&\lesssim B\sqrt{\frac{SH}{N}}+\sqrt{\frac{BSH|V_{\bar{h}}^{\widehat{\pi}}(\bar{s})-\widehat{V}_{\bar{h}}(\bar{s})|}{N}}+\frac{SBH}{N}. \qquad (\widehat{V}_{\bar{h}}(\bar{s})\le B)
\end{aligned}$$

Again solving a quadratic inequality w.r.t $|V_{\bar{h}}^{\widehat{\pi}}(\bar{s}) - \widehat{V}_{\bar{h}}(\bar{s})|$ and taking maximum over $(\bar{s}, \bar{h})$ on the left-hand-side, we have

$$\left\|V_{\cdot}^{\widehat{\pi}} - \widehat{V}_{\cdot}\right\|_{\infty} \lesssim B\sqrt{\frac{SH}{N}} + \frac{SBH}{N}. \tag{2}$$

Now define $\widehat{n} = \inf_n\{\text{right-hand-side of Eq. (2)} \leq \epsilon \text{ when } N = n\}$. We have $\widehat{n} \lesssim \frac{B^2SH}{\epsilon^2} + \frac{SBH}{\epsilon}$. This implies that when $\mathbf{N}(s,a) = N \geq \widehat{n}$ for all s, a and $\left\|\widehat{V}_{\cdot}\right\|_{\infty} \leq B$, we have $\left\|V_{\cdot}^{\widehat{\pi}} - \widehat{V}_{\cdot}\right\|_{\infty} \leq \epsilon$ with probability at least $1 - 4\delta$. The proof is then completed by treating $\widehat{n}$ as a function with input B, H, ϵ, δ and replace δ by $\delta/4$ in the arguments above. ■

Lemma 24 *For any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $V \in [-B, B]^{\mathcal{S}_+}$ for some $B > 0$, with probability at least $1 - \delta$, we have $\mathbb{V}(\bar{P}_{s,a}, V) \lesssim \mathbb{V}(P_{s,a}, V) + \frac{SB^2}{\mathbf{N}^+(s,a)}$ for all (s,a), where $\mathbf{N}^+(s,a) = \max\{1, \mathbf{N}(s,a)\}$.*

Proof Note that

$$\begin{aligned}
\mathbb{V}(\bar{P}_{s,a}, V) &\leq \bar{P}_{s,a}(V - P_{s,a}V)^2 && (\tfrac{\sum_i p_i x_i}{\sum_i p_i} = \operatorname{argmin}_z \textstyle\sum_i p_i (x_i - z)^2) \\
&= \mathbb{V}(P_{s,a}, V) + (\bar{P}_{s,a} - P_{s,a})(V - P_{s,a}V)^2 \\
&\lesssim \mathbb{V}(P_{s,a}, V) + B\sqrt{\frac{S\mathbb{V}(P_{s,a}, V)}{\mathbf{N}^+(s,a)}} + \frac{SB^2}{\mathbf{N}^+(s,a)} && \text{(Lemma 25)} \\
&\lesssim \mathbb{V}(P_{s,a}, V) + \frac{SB^2}{\mathbf{N}^+(s,a)}. && \text{(AM-GM inequality)}
\end{aligned}$$

This completes the proof. ■

Lemma 25 *Given any value function $V \in [-B, B]^{\mathcal{S}_+}$, with probability at least $1 - \delta$, $|(P_{s,a} - \bar{P}_{s,a})V| \lesssim \sqrt{\frac{S\mathbb{V}(P_{s,a},V)}{\mathbf{N}^+(s,a)}} + \frac{SB}{\mathbf{N}^+(s,a)}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $\mathbf{N}^+(s,a) = \max\{1, \mathbf{N}(s,a)\}$.*

Proof For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, by Lemma 44, with probability at least $1 - \frac{\delta}{SA}$, we have

$$\begin{aligned}
|(P_{s,a} - \bar{P}_{s,a})V| &\leq \sum_{s'} |P_{s,a}(s') - \bar{P}_{s,a}(s')||V(s') - P_{s,a}V| \quad (\textstyle\sum_{s'}(P_{s,a}(s') - \bar{P}_{s,a}(s')) = 0) \\
&\lesssim \sum_{s'} \left(\sqrt{\frac{P_{s,a}(s')}{\mathbf{N}^+(s,a)}} + \frac{1}{\mathbf{N}^+(s,a)}\right) \left|V(s') - P_{s,a}V\right| \lesssim \sqrt{\frac{S\mathbb{V}(P_{s,a}, V)}{\mathbf{N}^+(s,a)}} + \frac{SB}{\mathbf{N}^+(s,a)},
\end{aligned}$$

where the last step is by Cauchy-Schwarz inequality. Taking a union bound over (s,a) completes the proof. ■

D.3. Proof of Theorem 8

Proof For each index i, define finite-horizon MDP $\mathcal{M}_i = \mathcal{M}_{H_i,c_{f,i}}$. Also define $V^{\pi}_{h,i}$ and $V^{\star}_{h,i}$ as value function V^{π}_h and optimal value function $V^{\star}_h$ in $\mathcal{M}_i$ respectively. We first assume that $T \geq D$ such that $B_{\star,T} < \infty$. In this case, we have $T^{\pi^{\star}_{T,s}}(s) \leq \min\{T_{\ddagger}, T\}$ for any s by $\pi^{\star}_{T,s} = \pi^{\star}$ when $T \geq T_{\ddagger} \geq T_{\star}$. Note that when $B_i \in [20B_{\star,T}, 40B_{\star,T}]$, by $T^{\pi^{\star}_{T,s}}(s) \leq \min\{T_{\ddagger}, T\}$ for any s, definition of H_i, and Lemma 38, we have $V^{\star}_{1,i}(s) \leq V^{\pi^{\star}_{T,s}}_{1,i}(s) \leq V^{\star,T}(s) + 0.6B_i \cdot \frac{\epsilon}{24B_i} \leq 0.1B_i$ and $V^{\star}_{h,i}(s) \leq V^{\pi^{\star}_{T,s}}_{h,i}(s) \leq V^{\star,T}(s) + 0.6B_i \leq 0.7B_i$ for any $s \in \mathcal{S}$ and $h \in [H]$, where applying stationary policy $\pi^{\star}_{T,s}$ in $\mathcal{M}_i$ means executing $\pi^{\star}_{T,s}$ in each step $h \in [H]$. This implies $B_i \geq \left\|V^{\star}_{\cdot,i}\right\|_{\infty}$. Then according to Line 2 and by Lemma 19, with probability at least $1 - \delta_i$, we have $\left\|V^i_1\right\|_{\infty} \leq \left\|V^{\star}_{1,i}\right\|_{\infty} \leq 0.1B_i$, $\left\|V^i_{\cdot}\right\|_{\infty} \leq \left\|V^{\star}_{\cdot,i}\right\|_{\infty} \leq 0.7B_i$, and the while loop should break (Line 3). Let $i^{\star}$ be the value of i when the while loop breaks, we thus have $B_{i^{\star}} \leq 40B_{\star,T}$. Moreover, by Lemma 23 and the definition of N_i, with probability at least $1 - \delta_{i^{\star}}$, we have $V^{\pi^{i^{\star}}}_{1,i^{\star}}(s) \leq (V^{i^{\star}}_1(s) + 0.1B_{i^{\star}})\mathbb{I}\{s \neq g\} \leq c_{f,i^{\star}}(s)$ for any $s \in \mathcal{S}_+$. Thus by Lemma 7, we have $V^{\star}(s) \leq V^{\pi^{i^{\star}}}(s) \leq V^{\pi^{i^{\star}}}_{1,i^{\star}}(s) \leq V^{i^{\star}}_1(s) + 0.1B_{i^{\star}} \leq B_{i^{\star}}$ for any $s \in \mathcal{S}$. This gives $B_{i^{\star}} \geq B_{\star}$. If $T < T_{\ddagger} \leq \frac{B_{i^{\star}}}{c_{\min}}$, then $H_{i^{\star}} \lesssim \min\{\frac{B_{i^{\star}}}{c_{\min}}, T\} = \min\{T_{\ddagger}, T\}$. Otherwise, $T \geq T_{\ddagger}$, $B_{\star,T} = B_{\star}$, and $H_{i^{\star}} \lesssim \min\{\frac{B_{i^{\star}}}{c_{\min}}, T\} \lesssim \min\{T_{\ddagger}, T\}$ by $B_{i^{\star}} \lesssim B_{\star,T}$. Therefore, $H_{i^{\star}} \lesssim \min\{T_{\ddagger}, T\}$. By Lemma 23, $\left\|V^{i^{\star}}_1\right\|_{\infty} \leq 0.1B_{i^{\star}}$, $\left\|V^{i^{\star}}_{\cdot}\right\|_{\infty} \leq 0.7B_{i^{\star}}$ (breaking condition of the while loop), and the definition of N_i, we have with probability at least $1 - \delta_{i^{\star}}$, $\left\|V^{\star}_{1,i^{\star}}\right\|_{\infty} \leq \left\|V^{\pi^{i^{\star}}}_{1,i^{\star}}\right\|_{\infty} \leq \left\|V^{i^{\star}}_1\right\|_{\infty} + \left\|V^{\pi^{i^{\star}}}_{1,i^{\star}} - V^{i^{\star}}_1\right\|_{\infty} \leq 0.2B_{i^{\star}}$ and $\left\|V^{\star}_{\cdot,i^{\star}}\right\|_{\infty} \leq \left\|V^{\pi^{i^{\star}}}_{\cdot,i^{\star}}\right\|_{\infty} \leq \left\|V^{i^{\star}}_{\cdot}\right\|_{\infty} + \left\|V^{\pi^{i^{\star}}}_{\cdot,i^{\star}} - V^{i^{\star}}_{\cdot}\right\|_{\infty} \leq 0.8B_{i^{\star}}$. Therefore, by Lemma 22 and the definition of $N^{\star}_{i^{\star}}$, we have with probability at least $1 - \delta_{i^{\star}}$, $\left\|V^{\widehat{\pi}}_{\cdot,i^{\star}} - V^{\star}_{\cdot,i^{\star}}\right\|_{\infty} \leq \frac{\epsilon}{2}$. Moreover, by Lemma 7 and $V^{\widehat{\pi}}_{1,i^{\star}}(s) \leq (V^{\star}_{1,i^{\star}}(s) + \frac{\epsilon}{2})\mathbb{I}\{s \neq g\} \leq (0.2B_{i^{\star}} + \frac{1}{2})\mathbb{I}\{s \neq g\} \leq c_{f,i^{\star}}(s)$ for all $s \in \mathcal{S}_+$ since $B_{i^{\star}} \geq 2$, we have $V^{\widehat{\pi}}(s) \leq V^{\widehat{\pi}}_{1,i^{\star}}(s)$ for all s. Thus, $V^{\widehat{\pi}}(s) \leq V^{\widehat{\pi}}_{1,i^{\star}}(s) \leq V^{\star}_{1,i^{\star}}(s) + \frac{\epsilon}{2} \leq V^{\star,T}(s) + 0.6B_{i^{\star}}\frac{\epsilon}{24B_{i^{\star}}} + \frac{\epsilon}{2} \leq V^{\star,T}(s) + \epsilon$ by the definition of $H_{i^{\star}}$ and Lemma 38 for any $s \in \mathcal{S}$. Finally, by the definition of N_i and $N^{\star}_i$, the total number of samples spent is of order

$$\begin{aligned}
&\tilde{\mathcal{O}}\left(SA(N_{i^{\star}} + N^{\star}_{i^{\star}})\right) = \tilde{\mathcal{O}}\left(\frac{H_{i^{\star}}B^2_{i^{\star}}SA}{\epsilon^2} + \frac{H_{i^{\star}}B_{i^{\star}}S^2A}{\epsilon} + H^2_{i^{\star}}S^2A\right) \\
&= \tilde{\mathcal{O}}\left(\min\left\{T_{\ddagger}, T\right\}\frac{B^2_{\star,T}SA}{\epsilon^2} + \min\left\{T_{\ddagger}, T\right\}\frac{B_{\star,T}S^2A}{\epsilon} + \min\left\{T_{\ddagger}, T\right\}^2 S^2A\right).
\end{aligned}$$

Moreover, the bound above holds with probability at least $1 - \delta$ since $20\sum_i \delta_i \leq \delta$. Now we consider the case $T < D$. From the arguments above we know that $B_i \leq 40B_{\star,T} \leq 40T$ for all $i \leq i^{\star}$ if $T \geq D$. Thus, we can conclude that $T < D$ if $B_i > 40T$ for some i still in the while loop, and the total number samples used is of order $\tilde{\mathcal{O}}(SAN_i) = \tilde{\mathcal{O}}(S^2AT)$ by the definition of N_i. This completes the proof. ■

Appendix E. Omitted Details in Section 5

E.1. Proof of Theorem 9

Proof Let $N = \min\{\lfloor B_\star \rfloor, S-3\}$. Consider a family of MDPs $\{\mathcal{M}_j\}_{j\in[A]^N}$ with $\mathcal{S} = \mathcal{S}_N \cup \mathcal{S}'$, $\mathcal{A} = [A]$, and $s_{\text{init}} = s_0$, where $\mathcal{S}_N = \{s_0, s_1, \ldots, s_N\}$ and $\mathcal{S}' = \{s_b, s_c, \ldots\}$. Clearly, $|\mathcal{S}_N| = N+1$ and $|\mathcal{S}'| = S - N - 1$. For each $\mathcal{M}_j$, the cost is 1 for every state-action pair in $\mathcal{S}_N$; at s_0, taking any action transits to g with probability $1-p$, and transits to state s_1 with probability p, where $p = \frac{4\epsilon}{A^N}$; at s_i for $i \in [N]$, taking action j_i transits to s_{i+1} (define $s_{N+1} = g$), while taking any other actions transits to s_1; at s_b, taking any action suffers cost 1 and transits to g with probability $1/B_\star$ (stays at s_b otherwise); at s_c, taking any action suffers cost $c_{\min}$ and directly transits to g; at any of the rest of states in $\mathcal{S}'$, taking any action suffers cost 1 and directly transits to g. Note that all of these MDPs have parameters S, A, $B_\star$, $c_{\min}$, and all these parameters are known to the learner. Also note that states in $\mathcal{S}'$ are unreachable and does not affect the learner. We include them simply to show that we can obtain a hard instance for any values of S, A, $B_\star$, and $c_{\min}$ using dummy states.

Consider a learner that is (ϵ, δ)-correct with $\epsilon \in (0, \frac{1}{4})$, $\delta \in (0, \frac{1}{16})$, and sample complexity $\frac{1}{p}$ on $\{\mathcal{M}_j\}_j$. Denote by $\mathcal{E}_1$ the event that the first $\frac{1}{p}$ steps from s_0 all transit to g, $\mathcal{E}_2$ the event that the learner uses at most $\frac{1}{p}$ samples, and define $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. Also denote by P_j the distribution w.r.t $\mathcal{M}_j$. Note that event $\mathcal{E}$ is agnostic to j, that is, for any interaction history (including the randomness of the learner) $\omega \in \mathcal{E}$, we have $P_j(\omega) = P(\omega)$ for all j. Also note that for any j, we have $P_j(\mathcal{E}_1) = (1-p)^{1/p} \geq \frac{1}{4}$ and $P_j(\mathcal{E}_2) \geq 1 - \delta \geq \frac{7}{8}$. Thus, $P(\mathcal{E}) = P_j(\mathcal{E}) \geq \frac{1}{8}$. Now we show that the failure probability of such a learner is large. Note that when $\mathcal{E}$ is true, the learner outputs $\widehat{\pi}$ before visiting s_1. Moreover, the distribution of $\widehat{\pi}$ under $\mathcal{E}$ is identical for all $\{\mathcal{M}_j\}_j$, that is, $P_j(\widehat{\pi}|\mathcal{E}) = P(\widehat{\pi}|\mathcal{E})$. This is because $P_j(\omega) = P(\omega)$ for any interaction history $\omega \in \mathcal{E}$, and $\widehat{\pi}$ is a function of ω. Denote by $\mathcal{E}'$ the bad event that $\widehat{\pi}$ is not ϵ-optimal. We show that there exists j such that $P_j(\mathcal{E}'|\mathcal{E})$ is sufficiently large.

First, for any given j and any policy π, define $x_j^\pi = \prod_{i=1}^N \pi(j_i|s_i)$ and V_j^π as the value function of π in $\mathcal{M}_j$. Since the learner transits to s_1 if it does not follow the "correct" action sequence, we have $V_j^\pi(s_1) \geq N x_j^\pi + (1 - x_j^\pi)(1 + V_j^\pi(s_1))$, which gives $V_j^\pi(s_1) \geq N + \frac{1}{x_j^\pi} - 1$. Moreover, if π is ϵ-optimal in $\mathcal{M}_j$, then we have $V_j^\pi(s_0) \leq 1 + pN + \epsilon$. Combining with $V_j^\pi(s_0) = 1 + pV_j^\pi(s_1)$ gives $x_j^\pi \geq \frac{1}{1+\epsilon/p} \geq \frac{p}{2\epsilon}$ by $\epsilon/p \geq 1$. Also note that $\sum_j x_j^\pi = 1$. Therefore, each policy π can be ϵ-optimal for at most $\frac{2\epsilon}{p}$ MDPs in $\{\mathcal{M}_j\}_j$.

Denote by $y^\pi(j)$ the indicator of whether policy π is ϵ-optimal in $\mathcal{M}_j$. We have $\sum_j y^\pi(j) \leq \frac{2\epsilon}{p}$ for any π. Therefore, $\sum_j \int_{\widehat{\pi}} P(\widehat{\pi}|\mathcal{E}) y^{\widehat{\pi}}(j) d\widehat{\pi} \leq \frac{2\epsilon}{p}$, which implies that there exist $j^\star$ such that $\int_{\widehat{\pi}} P(\widehat{\pi}|\mathcal{E}) y^{\widehat{\pi}}(j^\star) d\widehat{\pi} \leq \frac{2\epsilon}{pA^N}$. Therefore,

$$P_{j^\star}(\mathcal{E}'|\mathcal{E}) = 1 - \int_{\widehat{\pi}} P(\widehat{\pi}|\mathcal{E}) y^{\widehat{\pi}}(j^\star) d\widehat{\pi} \geq 1 - \frac{2\epsilon}{pA^N} = \frac{1}{2}.$$

The overall failure probability in $\mathcal{M}_{j^\star}$ is thus $P_{j^\star}(\mathcal{E}') \geq P(\mathcal{E}) P_{j^\star}(\mathcal{E}'|\mathcal{E}) \geq \frac{1}{16}$, a contradiction. Therefore, for any (ϵ, δ)-correct learner, there exists $\mathcal{M} \in \{\mathcal{M}_j\}_j$ such that the learner has sample complexity more than $\frac{1}{p} = \Omega(A^N/\epsilon)$ on $\mathcal{M}$. The proof is then completed by the definition of N. ■

E.2. Proof of Theorem 10

Proof If $\min\left\{T_\ddagger, \overline{T}\right\} \frac{B_\star^2 SA}{\epsilon^2} \ln\frac{1}{\delta} > \frac{J}{\epsilon}$, then we simply construct a full $(A-1)$-ary tree following that of Theorem 3, with a remaining action $a_\dagger$. By $J \geq 3B$, we can simply ignore action $a_\dagger$ and the sample complexity lower bound is $\Omega(\min\left\{T_\ddagger, \overline{T}\right\} \frac{B_\star^2 SA}{\epsilon^2} \ln\frac{1}{\delta}) = \Omega(\min\left\{\frac{B_\star}{c_{\min}}, \overline{T}\right\} \frac{B_\star^2 SA}{\epsilon^2} \ln\frac{1}{\delta} + \frac{J}{\epsilon})$.

Otherwise, we have $\min\left\{T_\ddagger, \overline{T}\right\} \frac{B_\star^2 SA}{\epsilon^2} \ln\frac{1}{\delta} \leq \frac{J}{\epsilon}$, and our construction follows that of Theorem 9 except that we have an action $a_\dagger$ at every state. Consider a family of MDPs $\{\mathcal{M}_j\}_{j\in[A-1]^N}$ with $\mathcal{S} = \mathcal{S}_N \cup \mathcal{S}'$, $\mathcal{A} = [A-1] \cup \{a_\dagger\}$, and $s_{\text{init}} = s_0$, where $\mathcal{S}_N = \{s_0, s_1, \ldots, s_N\}$ and $\mathcal{S}' = \{s_b, s_c, \ldots\}$. For each $\mathcal{M}_j$, $c(s,a) = 1$ for all $(s,a) \in \mathcal{S}_N \times [A-1]$; at s_0, taking any action in $[A-1]$ transits to g with probability $1-p$, and transits to state s_1 with probability p, where $p = \frac{4\epsilon}{J}$; at s_i for $i \in [N]$, taking action j_i transits to s_{i+1} (define $s_{N+1} = g$), while taking any other actions in $[A-1]$ transits to s_1; at s_b, taking any action in $[A-1]$ suffers cost 1 and transits to g with probability $1/B$ (stay at s_b otherwise); at s_c, taking any action in $[A-1]$ suffers cost $c_{\min}$ and directly transits to g; at any of the rest of states in $\mathcal{S}'$, taking any action in $[A-1]$ suffers cost 1 and directly transits to g. Note that all of these MDPs have parameters S, A, $c_{\min}$, $\overline{T}$ and satisfy $B_\star = B$. Moreover, all these parameters are known to the learner.

Consider a learner that is (ϵ,δ)-correct with $\epsilon \in (0, \frac{1}{4})$, $\delta \in (0, \frac{1}{16})$, and sample complexity $\frac{1}{p}$ on $\{\mathcal{M}_j\}_j$. Denote by $\mathcal{E}_1$ the event that the first $\frac{1}{p}$ steps from (s_0, a) for some $a \in [A-1]$ all transit to g, $\mathcal{E}_2$ the event that the learner uses at most $\frac{1}{p}$ samples, and define $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. Also denote by P_j the distribution w.r.t $\mathcal{M}_j$. Note that event $\mathcal{E}$ is agnostic to j, that is, for any interaction history (including the randomness of the learner) $\omega \in \mathcal{E}$, we have $P_j(\omega) = P(\omega)$ for all j. Also note that for any j, we have $P_j(\mathcal{E}_1) = (1-p)^{1/p} \geq \frac{1}{4}$ and $P_j(\mathcal{E}_2) \geq 1 - \delta \geq \frac{7}{8}$. Thus, $P(\mathcal{E}) = P_j(\mathcal{E}) \geq \frac{1}{8}$. Now we show that the failure probability of such a learner is large. Note that when $\mathcal{E}$ is true, the learner outputs $\widehat{\pi}$ before ever visiting s_1. Moreover, the distribution of $\widehat{\pi}$ under $\mathcal{E}$ is identical for all $\{\mathcal{M}_j\}_j$, that is, $P_j(\widehat{\pi}|\mathcal{E}) = P(\widehat{\pi}|\mathcal{E})$. This is because $P_j(\omega) = P(\omega)$ for any interaction history $\omega \in \mathcal{E}$, and $\widehat{\pi}$ is a function of ω. Denote by $\mathcal{E}'$ the bad event that $\widehat{\pi}$ is not ϵ-optimal. We show that there exists j such that $P_j(\mathcal{E}'|\mathcal{E})$ is sufficiently large.

First, for any given j and any policy π, define $x_j^\pi = \prod_{i=1}^N \pi(j_i|s_i)$, V_j^π as the value function of π in $\mathcal{M}_j$, and $y^\pi = \pi(a_\dagger|s_1)$. If π is ϵ-optimal in $\mathcal{M}_j$, then we have $V_j^\pi(s_0) \leq 1 + pN + \epsilon < J$. Combining with $V_j^\pi(s_0) \geq \min\{J, 1 + pV_j^\pi(s_1)\}$, we have $V_j^\pi(s_1) \leq N + \frac{\epsilon}{p} = N + \frac{J}{4} < J - 1$. Moreover, the learner suffers cost N if it follows the "correct" action sequence, and suffers at least cost $J > 1 + V_j^\pi(s_1)$ if it ever takes action $a_\dagger$. Thus, we have $V_j^\pi(s_1) \geq Nx_j^\pi + Jy^\pi + (1 - x_j^\pi - y^\pi)(1 + V_j^\pi(s_1))$, which gives $V_j^\pi(s_1) \geq \frac{Nx_j^\pi + Jy^\pi}{x_j^\pi + y^\pi} + \frac{1}{x_j^\pi + y^\pi} - 1$. Combining with $V_j^\pi(s_1) \leq N + \frac{\epsilon}{p}$, we have

$$(1 + \epsilon/p)(x_j^\pi + y^\pi) \geq 1 + (J - N)y^\pi. \tag{3}$$

Now note that $\sum_j x_j^\pi \leq 1 - y^\pi$ for any π. Define $\mathcal{X}^\pi$ as the set of $j \in [A-1]^N$ where π is ϵ-optimal in $\mathcal{M}_j$. Summing over $j \in \mathcal{X}^\pi$ for Eq. (3), we have $(1 + \epsilon/p)(1 - y^\pi + |\mathcal{X}^\pi| y^\pi) \geq |\mathcal{X}^\pi| + (J - N)y^\pi|\mathcal{X}^\pi|$. Reorganizing terms and assuming $|\mathcal{X}^\pi| \geq 1$ gives

$$1 - \frac{1 + \epsilon/p}{|\mathcal{X}^\pi|} \leq y^\pi \left(\left(1 - \frac{1}{|\mathcal{X}^\pi|}\right)\left(1 + \frac{\epsilon}{p}\right) - (J - N)\right) \leq y^\pi \left(1 + \frac{\epsilon}{p} - (J - N)\right).$$

Note that $1 + \frac{\epsilon}{p} \leq J - N$ by $p \geq \frac{\epsilon}{J-N-1}$. Thus, the right-hand-side ≤ 0, which gives $|\mathcal{X}^\pi| \leq 1 + \epsilon/p \leq J - N$. Therefore, each policy can be ϵ-optimal for at most $J - N$ MDPs in $\{\mathcal{M}_j\}_j$.

Denote by $z^\pi(j)$ the indicator of whether π is ϵ-optimal in $\mathcal{M}_j$. We have $\sum_j z^\pi(j) \le J - N$ for any policy π. Therefore, $\sum_j \int_{\widehat{\pi}} P(\widehat{\pi}|\mathcal{E}) z^{\widehat{\pi}}(j) d\widehat{\pi} \le J - N$, which implies that there exist $j^\star$ such that $\int_{\widehat{\pi}} P(\widehat{\pi}|\mathcal{E}) z^{\widehat{\pi}}(j^\star) d\widehat{\pi} \le \frac{J-N}{(A-1)^N}$. Therefore,

$$P_{j^\star}(\mathcal{E}'|\mathcal{E}) = 1 - \int_{\widehat{\pi}} P(\widehat{\pi}|\mathcal{E}) z^{\widehat{\pi}}(j^\star) d\widehat{\pi} \ge 1 - \frac{J-N}{(A-1)^N} \ge \frac{1}{2}.$$

The overall failure probability in $\mathcal{M}_{j^\star}$ is thus $P_{j^\star}(\mathcal{E}') \ge P(\mathcal{E}) P_{j^\star}(\mathcal{E}'|\mathcal{E}) \ge \frac{1}{16}$, a contradiction. Therefore, for any (ϵ, δ)-correct learner, there exists $\mathcal{M} \in \{\mathcal{M}_j\}_j$ such that the learner has sample complexity more than $\frac{1}{p} = \Omega(J/\epsilon) = \Omega(\min\{T_\ddagger, \overline{T}\} \frac{B_\star^2 SA}{\epsilon^2} \ln \frac{1}{\delta} + J/\epsilon)$ on $\mathcal{M}$. This completes the proof. ■

Appendix F. Omitted Details in Section 6

In this section, we present the proof of Theorem 11.

Notations Denote by $V_\cdot^\star$, $Q_\cdot^\star$ the optimal value function and action-value function of $\mathcal{M}_{H,c_f}$, where $H = \frac{32J}{c_{\min}} \ln \frac{8J}{\epsilon}$ and $c_f(s) = J\mathbb{I}\{s \ne g\}$. Clearly, we have $V_h^\star(s) = \operatorname{argmin}_a Q_h^\star(s,a)$, $Q_h^\star(s,a) = c(s,a) + P_{s,a} V_{h+1}^\star$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, and $V_{H+1}^\star = c_f$. The whole learning process is divided into episodes indexed by m. Define $H_m = \inf_h \{s_{h+1}^m = g \text{ or } a_{h+1}^m = a_\dagger\}$ as the length of episode m and $R_{[m',M']} = \sum_{m=m'}^{M'} (\sum_{h=1}^{H_m+1} c_h^m - V_1^m(s_1^m))$ the regret w.r.t estimated value functions of episodes in $[m', M']$, where $c_h^m = c(s_h^m, a_h^m)$ and $c_{H_m+1}^m = c_f(s_{H_m+1}^m)$. We also write $R_{[1,M']}$ as $R_{M'}$. Define $s_h^m = g$ for all $h > H_m + 1$. Denote by $\bar{P}^m$, Q^m, V^m, b_h^m, $\mathbf{N}_h^m$ the value of $\bar{P}$, $\widehat{Q}$, $\widehat{V}$, $b(s_h^m, a_h^m, V_{h+1}^m)$, $\mathbf{N}^+(s_h^m, a_h^m)$ from Algorithm 3 executed in Line 2 in episode m. For any episode m in round r, define $\pi_m = \pi^r$. Define $P_h^m = P_{s_h^m, a_h^m}$ and $\bar{P}_h^m = \bar{P}_{s_h^m, a_h^m}^m$. Define $C^m = \sum_{h=1}^{H_m+1} c_h^m$ and $C_{M'} = \sum_{m=1}^{M'} C^m$. Denote by λ_r the value of λ in round r (computed in Line 4) and B_m the value of B in episode m.

Lemma 26 *With probability at least* $1 - \delta$, $Q_h^m(s,a) \le Q_h^\star(s,a)$ *for any* $m \ge 1$ *and* $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

Proof This is simply by $\max_{h \in [H+1]} \|V_h^\star\|_\infty \le \|V^\star\|_\infty + \|c_f\|_\infty \le 2J$ and Lemma 19. ■

We are now ready to prove Theorem 11.

Proof [of Theorem 11] Denote by $\mathcal{I}_r$ the set of episodes in round r. First note that in the last round (success round) R, we have with probability at least $1 - 35\delta$,

$$\begin{aligned}
V^{\widehat{\pi}}(s_{\text{init}}) - V^\star(s_{\text{init}}) &\le V^{\widehat{\pi}}(s_{\text{init}}) - V_1^\star(s_{\text{init}}) + \frac{\epsilon}{4} && \text{(definition of } H \text{ and Lemma 38)} \\
&\le \frac{1}{\lambda_R} \sum_{m \in \mathcal{I}_R} (V_1^{\pi_m}(s_1^m) - C^m) + \frac{1}{\lambda_R} \sum_{m \in \mathcal{I}_R} (C^m - V_1^m(s_1^m)) + \frac{\epsilon}{4} \\
& && (\pi_m = \widehat{\pi} \text{ for } m \in \mathcal{I}_R,\ s_1^m = s_{\text{init}}, \text{ and Lemma 26}) \\
&\le \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} \le \epsilon,
\end{aligned}$$

where the last step is by Lemma 27, the definition of λ, the condition of success round, and $V_1^m = V^R$ for $m \in \mathcal{I}_R$. Thus, the output policy $\widehat{\pi}$ is indeed ϵ-optimal. Now we bound the sample complexity of the algorithm. Note that by Lemma 38 and the definition of H, for any $h \le H/2+1$ and $s \in \mathcal{S}$, we have $V^\star(s) \le V_h^\star(s) \le V^\star(s) + \frac{\epsilon J}{2J} \le \frac{3}{2}B_\star$. Then by Lemma 26, we have $B_{M'} \le 3B_\star$ for any $M' \ge 1$ with probability at least $1-\delta$. Note that there are at most $\mathcal{O}(SA\log_2(R\lambda_R))$ skip rounds, and one success round. Thus, it suffices to bound the number of failure rounds R_f. In each failure round, we have at least $\frac{\lambda_R \epsilon}{2}$ regret by the condition in Line 7. Moreover, in each skip or success round r, we have with probability at least $1-35\delta$,

$$\sum_{m\in\mathcal{I}_r}(C^m - V_1^m(s_1^m)) \ge \sum_{m\in\mathcal{I}_r}(C^m - V_1^{\pi_m}(s_1^m)) \gtrsim -\lambda_R\epsilon.$$

where the last step is by Eq. (4) and the value of λ_r. Thus, the total regret is lower bounded as follows: $R_M \gtrsim (R_f - SA)\lambda_R\epsilon$. Denote by M the total number episodes. By Lemma 28, Lemma 32 and $M \le R\lambda_R \lesssim (SA+R_f)\lambda_R$, we have with probability at least $1-38\delta$,

$$R_M \lesssim B_M\sqrt{SAM} + J^2H^{1.25}S^2A \lesssim B_M\sqrt{SA(SA+R_f)\lambda_R} + J^2H^{1.25}S^2A.$$

Solving a quadratic inequality w.r.t R_f, we have

$$R_f \lesssim SA + \frac{B_M SA}{\epsilon\sqrt{\lambda_R}} + \frac{B_M^2 SA}{\epsilon^2\lambda_R} + \frac{J^2H^{1.25}S^2A}{\lambda_R\epsilon} \lesssim SA.$$

Therefore, $M \le R\lambda_R \lesssim \frac{B_\star^2 SA}{\epsilon^2} + \frac{J^2H^2S^2A^2}{\epsilon}$ by $B_M \le 3B_\star$. Moreover, by Eq. (6) with $M' = M$ and $B_M \le 3B_\star$, we know that $C_M \lesssim B_\star M + JS^2A$ and thus the total number of samples is of order

$$\tilde{\mathcal{O}}\left(\frac{B_\star M}{c_{\min}} + \frac{JS^2A}{c_{\min}}\right) = \tilde{\mathcal{O}}\left(\frac{T_\ddagger B_\star^2 SA}{\epsilon^2} + \frac{B_\star J^4S^2A^2}{c_{\min}^3\epsilon}\right).$$

This completes the proof. ■

Lemma 27 *There exists function $N_{\mathrm{DEV}}(B,\epsilon,\delta) \lesssim \frac{B^2}{\epsilon^2} + \frac{(H^{1.5}+J^2H^{1.25})SA}{\epsilon}$, such that for any $m' \ge 1$ and $n \ge N_{\mathrm{DEV}}(B,\epsilon,\delta)$, if $B \ge B_{m'+n-1}$, then $\frac{1}{n}\left|\sum_{m=m'}^{m'+n-1}(V_1^{\pi_m}(s_1^m) - C^m)\right| \le \epsilon$ with probability at least $1-35\delta$.*

Proof By Lemma 30 and Lemma 32, for any $m', n \ge 1$, we have with probability at least $1-35\delta$,

$$\left|\frac{1}{n}\sum_{m=m'}^{m'+n-1}(V_1^{\pi_m}(s_1^m) - C^m)\right| \lesssim B_{m'+n-1}\sqrt{\frac{1}{n}} + \frac{(H^{1.5}+J^2H^{1.25})SA}{n}. \tag{4}$$

Thus, for any given B, ϵ, δ, there exists $N \lesssim \frac{B^2}{\epsilon^2} + \frac{(H^{1.5}+J^2H^{1.25})SA}{\epsilon}$ such that if $n \ge N$ and $B \ge B_{m'+n-1}$, then $\frac{1}{n}\left|\sum_{m=m'}^{m'+n-1}(V_1^{\pi_m}(s_1^m) - C^m)\right| \le \epsilon$ with probability at least $1-35\delta$. Treating N as function of (B,ϵ,δ) completes the proof. ■

Below we state the regret guarantee of LCBVI. The main idea is to bound the regret w.r.t B instead of $B_\star$, which is useful in deciding the number of episodes needed.

Lemma 28 *For any $m' \geq 1$, with probability at least $1 - 9\delta$, we have for all $M' \geq m'$ simultaneously*

$$R_{[m',M']} \lesssim \sqrt{SA \sum_{m=m'}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + JS^2A.$$

Proof Without loss of generality, we assume $m' = 1$. Note that

$$\begin{aligned} R_{M'} &= \sum_{m=1}^{M'} \left(\sum_{h=1}^{H_m+1} c_h^m - V_1^m(s_1^m) \right) \overset{\text{(i)}}{\lesssim} \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(c_h^m + V_{h+1}^m(s_{h+1}^m) - V_h^m(s_h^m) \right) + JSA \\ &\leq \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left((\mathbb{I}_{s_{h+1}^m} - P_h^m) V_{h+1}^m + (P_h^m - \bar{P}_h^m) V_{h+1}^m + b_h^m \right) + JSA, \quad \text{(definition of } V_h^m\text{)} \end{aligned}$$

where (i) is by the fact that $V_{H_m+1}^m(s_{H_m+1}^m) \neq c_{H_m+1}^m \leq J$ only if m is the last episode of a skip round, and there are at most $\tilde{\mathcal{O}}(SA)$ skip rounds. We bound the three sums above separately. For the first sum, by Lemma 42, we have with probability at least $1 - \delta$,

$$\sum_{m=1}^{M'} \sum_{h=1}^{H_m} (\mathbb{I}_{s_{h+1}^m} - P_h^m) V_{h+1}^m \lesssim \sqrt{\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + J.$$

For the second sum, we have with probability at least $1 - 7\delta$,

$$\begin{aligned} &\sum_{m=1}^{M'} \sum_{h=1}^{H_m} (P_h^m - \bar{P}_h^m) V_{h+1}^m = \sum_{m=1}^{M'} \sum_{h=1}^{H_m} (P_h^m - \bar{P}_h^m) V_{h+1}^\star + \sum_{m=1}^{M'} \sum_{h=1}^{H_m} (P_h^m - \bar{P}_h^m)(V_{h+1}^m - V_{h+1}^\star) \\ &\lesssim \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(\sqrt{\frac{\mathbb{V}(P_h^m, V_{h+1}^\star)}{\mathbf{N}_h^m}} + \sqrt{\frac{S\mathbb{V}(P_h^m, V_{h+1}^m - V_{h+1}^\star)}{\mathbf{N}_h^m}} + \frac{SJ}{\mathbf{N}_h^m} \right) \\ &\qquad \text{(Lemma 44 and Lemma 25)} \\ &\lesssim \sqrt{SA \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^\star)} + \sqrt{S^2A \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m - V_{h+1}^\star)} + JS^2A. \\ &\qquad \text{(Cauchy-Schwarz inequality and Lemma 35)} \\ &\lesssim \sqrt{SA \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + \sqrt{S^2A \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m - V_{h+1}^\star)} + JS^2A. \\ &\qquad (\textsc{Var}[X+Y] \leq \textsc{Var}[X] + \textsc{Var}[Y]) \\ &\lesssim \sqrt{SA \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + JS^2A. \qquad \text{(Lemma 29 and AM-GM inequality)} \end{aligned}$$

Plugging these back and applying Lemma 34 to bound $\sum_{m=1}^{M'} \sum_{h=1}^{H_m} b_h^m$ completes the proof. ■

Lemma 29 *For any $m' \geq 1$, with probability at least $1-5\delta$, we have $\sum_{m=m'}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^\star - V_{h+1}^m) \lesssim J\sqrt{SA \sum_{m=m'}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + J^2S^2A$ for all $M' \geq m'$.*

Proof Without loss of generality, we assume $m' = 1$. First note that with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left((V_h^\star(s_h^m) - V_h^m(s_h^m))^2 - (P_h^m(V_{h+1}^\star - V_{h+1}^m))^2 \right) \\
&\lesssim J \sum_{m=1}^{M'} \sum_{h=1}^{H_m} (V_h^\star(s_h^m) - V_h^m(s_h^m) - P_h^m V_{h+1}^\star + P_h^m V_{h+1}^m)_+ \\
&\qquad \text{(Lemma 26 and } a^2 - b^2 \leq (a+b)(a-b)_+ \text{ for } a, b > 0) \\
&\lesssim J \sum_{m=1}^{M'} \sum_{h=1}^{H_m} (c_h^m + P_h^m V_{h+1}^m - V_h^m(s_h^m))_+. \qquad (V_h^\star(s_h^m) \leq Q_h^\star(s_h^m, a_h^m) = c_h^m + P_h^m V_{h+1}^\star)
\end{aligned}$$

By the definition of V_h^m and $(a)_+ - (b)_+ \leq (a-b)_+$, with probability at least $1-3\delta$, we continue with

$$\begin{aligned}
&\lesssim J \sum_{m=1}^{M'} \sum_{h=1}^{H_m} ((P_h^m - \bar{P}_h^m) V_{h+1}^\star + (P_h^m - \bar{P}_h^m)(V_{h+1}^m - V_{h+1}^\star) + b_h^m)_+ \\
&\lesssim J \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(\sqrt{\frac{\mathbb{V}(P_h^m, V_{h+1}^\star)}{\mathbf{N}_h^m}} + \sqrt{\frac{S\mathbb{V}(P_h^m, V_{h+1}^m - V_{h+1}^\star)}{\mathbf{N}_h^m}} + \frac{SJ}{\mathbf{N}_h^m} + b_h^m \right) \\
&\qquad \text{(Lemma 44 and Cauchy-Schwarz inequality)} \\
&\lesssim J \left(\sqrt{SA \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + \sqrt{S^2 A \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m - V_{h+1}^\star)} \right) + J^2 S^2 A,
\end{aligned}$$

where in the last step we apply $\text{VAR}[X+Y] \leq \text{VAR}[X] + \text{VAR}[Y]$, Cauchy-Schwarz inequality, Lemma 35 and Lemma 34. Then applying Lemma 33 with $\left\| V_{h+1}^\star - V_h^m \right\|_\infty \leq J$ and solving a quadratic inequality w.r.t $\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^\star - V_{h+1}^m)$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^\star - V_{h+1}^m) \\
&\lesssim \sum_{m=1}^{M'} (V_{H_m+1}^\star(s_{H_m+1}^m) - V_{H_m+1}^m(s_{H_m+1}^m))^2 + J\sqrt{SA \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + J^2 S^2 A.
\end{aligned}$$

Now note that $\sum_{m=1}^{M'} (V_{H_m+1}^\star(s_{H_m+1}^m) - V_{H_m+1}^m(s_{H_m+1}^m))^2 \lesssim J^2 SA$ since $V_{H_m+1}^\star(s_{H_m+1}^m) \neq V_{H_m+1}^m(s_{H_m+1}^m)$ only when m is the last episode of a skip round, and the number of skip rounds is of order $\tilde{\mathcal{O}}(SA)$. Plugging this back completes the proof. ■

Lemma 30 *For any $m' \geq 1$, with probability at least $1-6\delta$, we have for all $M' \geq m'$ simultaneously,*

$$\left|\sum_{m=m'}^{M'}(C^m - V_1^{\pi_m}(s_1^m))\right| \lesssim \sqrt{\sum_{m=m'}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^m)} + H^{1.5}SA.$$

Proof Without loss of generality, we assume $m'=1$. Denote by $\{(\widetilde{s}_h^m, \widetilde{a}_h^m, \widetilde{s}_{h+1}^m)\}_{h=1}^H$ the visited state-action-next-state triplets in episode m if the learner follows π^m till the end, that is, it does not stop the current episode immediately if the number of visits to some state-action pair is doubled. Note that $(\widetilde{s}_h^m, \widetilde{a}_h^m) \neq (s_h^m, a_h^m)$ only if m is the last episode of a skip round and $H_m < h$. Also define $\widetilde{C}^m = \sum_{h=1}^H c(\widetilde{s}_h^m, \widetilde{a}_h^m) + c_f(\widetilde{s}_{H+1}^m)$ the corresponding total cost in episode m, and $\widetilde{P}_h^m = P_{\widetilde{s}_h^m, \widetilde{a}_h^m}$. By (Chen et al., 2022, Lemma 26) and the fact that π^m is a deterministic policy for any m, we have $\textsc{Var}_{\pi_m}[\widetilde{C}^m] = \mathbb{E}_{\pi^m}[\sum_{h=1}^H \mathbb{V}(\widetilde{P}_h^m, V_{h+1}^{\pi_m})]$. Moreover, there are at most $\widetilde{\mathcal{O}}(SA)$ skip rounds. Then with probability at least $1-2\delta$,

$$\left|\sum_{m=1}^{M'}(C^m - V_1^{\pi_m}(s_1^m))\right| \lesssim \left|\sum_{m=1}^{M'}(\widetilde{C}^m - V_1^{\pi_m}(s_1^m))\right| + HSA \lesssim \sqrt{\sum_{m=1}^{M'}\textsc{Var}_{\pi_m}[\widetilde{C}^m]} + HSA$$

(Lemma 42)

$$\lesssim \sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H}\mathbb{V}(\widetilde{P}_h^m, V_{h+1}^{\pi_m})} + H^{1.5}SA \lesssim \sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\pi_m})} + H^{1.5}SA$$

(Lemma 43, and $\sum_{h=1}^H \mathbb{V}(\widetilde{P}_h^m, V_{h+1}^{\pi_m}) \leq H^3$)

$$\lesssim \sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^m)} + \sqrt{\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m, V_{h+1}^{\pi_m} - V_{h+1}^m)} + H^{1.5}SA.$$

($\textsc{Var}[X+Y] \leq 2(\textsc{Var}[X] + \textsc{Var}[Y])$)

For the second term above, note that with probability at least $1-3\delta$,

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left((V_h^{\pi_m}(s_h^m)-V_h^m(s_h^m))^2-(P_h^m(V_{h+1}^{\pi_m}-V_{h+1}^m))^2\right)$$

$$\lesssim H\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(V_h^{\pi_m}(s_h^m)-V_h^m(s_h^m)-P_h^mV_{h+1}^{\pi_m}+P_h^mV_{h+1}^m\right)_+$$

(Lemma 26 and $a^2-b^2\le(a+b)(a-b)_+$ for $a,b>0$)

$$\lesssim H\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(c_h^m+P_h^mV_{h+1}^m-V_h^m(s_h^m)\right)_+ \qquad (V_h^{\pi_m}(s_h^m)=c(s_h^m,a_h^m)+P_h^mV_{h+1}^{\pi_m})$$

$$\lesssim H\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left((P_h^m-\bar{P}_h^m)V_{h+1}^m+b_h^m\right)_+. \qquad \text{(definition of } V_h^m)$$

$$\lesssim H\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{S\mathbb{V}(P_h^m,V_{h+1}^m)}{\mathbf{N}_h^m}}+\frac{SJ}{\mathbf{N}_h^m}\right)+H\sqrt{SA\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+HJS^{1.5}A$$

(Lemma 44, Cauchy-Schwarz inequality, and Lemma 34)

$$\lesssim H\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+HJS^2A. \quad \text{(Cauchy-Schwarz inequality and Lemma 35)}$$

Thus, by Lemma 33 with $\left\|V_h^{\pi_m}-V_h^m\right\|_\infty\le H$ and $\sum_{m=1}^{M'}(V_{H_m+1}^{\pi_m}(s_{H_m+1}^m)-V_{H_m+1}^m(s_{H_m+1}^m))^2\lesssim H^2SA$ since $V_{H_m+1}^{\pi_m}(s_{H_m+1}^m)\neq V_{H_m+1}^m(s_{H_m+1}^m)$ only when the number of visits to some state-action pair is doubled, we have with probability at least $1-\delta$,

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^{\pi_m}-V_{h+1}^m)\lesssim H\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+H^2S^2A.$$

Plugging this back and applying AM-GM inequality completes the proof. ■

Lemma 31 (Coarse Bound) *For any $m'\ge 1$, with probability at least $1-2\delta$, we have for all $M'\ge m'$, $\sum_{m=m'}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)\lesssim JC_{M'}+J^2S^2A$.*

Proof Without loss of generality, we assume $m'=1$. By the definition of V_h^m and $(a)_+-(b)_+\le(a-b)_+$, we have with probability at least $1-\delta$,

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_h^m(s_h^m)-P_h^mV_{h+1}^m)_+\lesssim\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(c_h^m+(\bar{P}_h^m-P_h^m)V_{h+1}^m)_+$$

$$\lesssim\sum_{m=1}^{M'}\sum_{h=1}^{H_m}c_h^m+\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\left(\sqrt{\frac{S\mathbb{V}(P_h^m,V_{h+1}^m)}{\mathbf{N}_h^m}}+\frac{SJ}{\mathbf{N}_h^m}\right) \qquad \text{(Lemma 25)}$$

$$\lesssim\sum_{m=1}^{M'}\sum_{h=1}^{H_m}c_h^m+\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+JS^2A, \qquad (5)$$

where the last step is by Lemma 35. Therefore,

$$\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_h^m(s_h^m)^2-(P_h^mV_{h+1}^m)^2)\lesssim J\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_h^m(s_h^m)-P_h^mV_{h+1}^m)_+$$
$$(a^2-b^2\le(a+b)(a-b)_+ \text{ for } a,b>0)$$
$$\lesssim J\sum_{m=1}^{M'}\sum_{h=1}^{H_m}c_h^m+J\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+J^2S^2A.$$

Thus by Lemma 33, we have with probability at least $1-\delta$,

$$\sum_{m=m'}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)$$
$$\lesssim\sum_{m=1}^{M'}\sum_{h=1}^{H_m}V_{H_m+1}^m(s_{H_m+1}^m)^2+J\sum_{m=1}^{M'}\sum_{h=1}^{H_m}c_h^m+J\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+J^2S^2A$$
$$\lesssim JC_{M'}+J\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+J^2S^2A. \qquad (V_{H_m+1}^m(s_{H_m+1}^m)\le 2c_{H_m+1}^m)$$

Solving a quadratic inequality w.r.t $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)$ completes the proof. ■

Lemma 32 (Refined Bound) *For any $m'\ge 1$, with probability at least $1-29\delta$, for all $M'\ge m'$, we have $\sum_{m=m'}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)\lesssim B_{M'}^2(M'-m'+1)+J^4H^{2.5}S^2A$.*

Proof Without loss of generality, we assume $m'=1$ and write $B_{M'}$ as B for simplicity. By Lemma 28 and Lemma 31, we have with probability at least $1-11\delta$,

$$R_{M'}=C_{M'}-\sum_{m=1}^{M'}V_1^m(s_1^m)\lesssim\sqrt{JSAC_{M'}}+JS^2A.$$

Solving a quadratic inequality w.r.t $C_{M'}$ gives

$$C_{M'}\lesssim\sum_{m=1}^{M'}V_1^m(s_1^m)+JS^2A\lesssim BM'+JS^2A. \tag{6}$$

Thus, with probability at least $1-16\delta$,

$$\sum_{m=1}^{M'}(V_1^{\pi_m}(s_1^m)-V_1^\star(s_1^m))\lesssim\sum_{m=1}^{M'}(V_1^{\pi_m}(s_1^m)-C^m)+\sum_{m=1}^{M'}(C^m-V_1^\star(s_1^m))$$
$$\lesssim\sqrt{SA\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+H^{1.5}S^2A \qquad \text{(Lemma 30, Lemma 28, and Lemma 26)}$$
$$\lesssim\sqrt{JSAC_{M'}}+H^{1.5}S^2A\lesssim\sqrt{JBSAM}+H^{1.5}S^2A, \tag{7}$$

where the last two steps are by Lemma 31 and Eq. (6) respectively. Now note that

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_h^m(s_h^m)^2-(P_h^mV_{h+1}^m)^2)\\
&\le\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_h^m(s_h^m)+P_h^mV_h^m)(V_h^m(s_h^m)-P_h^mV_h^m)_+\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad(a^2-b^2\le(a+b)(a-b)_+\text{ for }a,b>0)\\
&=2B\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_h^m(s_h^m)-P_h^mV_h^m)_+ +\sum_{m=1}^{M'}\sum_{h=1}^{H_m}(V_h^m(s_h^m)+P_h^mV_h^m-2B)(V_h^m(s_h^m)-P_h^mV_h^m)_+\\
&\lesssim BC_{M'}+B\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+JBS^2A+J^2\sum_{m=1}^{M'}\sum_{h=H/2+1}^{H_m}\mathbb{I}\{s_{H/2+1}^m\neq g\},\qquad(8)
\end{aligned}$$

where the last step is by Eq. (5) and $\|V_h^m\|_\infty\le B$ when $h\le H/2+1$. Similarly, we have

$$\sum_{m=1}^{M'}V_{H_m+1}^m(s_{H_m+1}^m)^2=B^2M'+\sum_{m=1}^{M'}(V_{H_m+1}^m(s_{H_m+1}^m)^2-B^2)\le B^2M'+J^2\sum_{m=1}^{M'}\mathbb{I}\{s_{H/2+1}^m\neq g\}.\tag{9}$$

It suffices to bound $\sum_{m=1}^{M'}\mathbb{I}\{s_{H/2+1}^m\neq g\}$. Let $\widetilde{V}_1^\pi$ and $\widetilde{V}_1^\star$ be the value function and optimal value function of $\mathcal{M}_{H/4,\mathbf{0}}$, where $\mathbf{0}$ represents constant function with value 0. By the value of H and (Chen et al., 2021a, Lemma 1), we have $V^\star(s)\le\widetilde{V}_1^\star(s)+\frac{\epsilon}{4J}$ for all $s\in\mathcal{S}_+$. Moreover, when $s_{H/2+1}^m\neq g$, we have $\sum_{h=H/4+1}^{H/2}c_h^m\ge 2J$. Denote by $P_m(\cdot)$ the probability distribution conditioned on the events before episode m. We have

$$\begin{aligned}
&2J\sum_{m=1}^{M'}P_m(s_{H/2+1}^m\neq g)+\sum_{m=1}^{M'}(\widetilde{V}_1^{\pi_m}(s_1^m)-\widetilde{V}_1^\star(s_1^m))\\
&\le\sum_{m=1}^{M'}(V_1^{\pi_m}(s_1^m)-V^\star(s_1^m))+\frac{M'\epsilon}{J}\lesssim\sqrt{JBSAM}+H^{1.5}S^2A+\frac{M'\epsilon}{J}.\qquad\text{(Eq. (7))}
\end{aligned}$$

Therefore, $\sum_{m=1}^{M'}P_m(s_{H/2+1}^m\neq g)\lesssim\sqrt{SAM'}+\frac{H^{1.5}S^2A}{J}+\frac{M'\epsilon}{J^2}$ by $\widetilde{V}_1^{\pi_m}(s_1^m)\ge\widetilde{V}_1^\star(s_1^m)$, and so does $\sum_{m=1}^{M'}\mathbb{I}\{s_{H/2+1}^m\neq g\}$ with probability at least $1-\delta$ by Lemma 43. Plugging this back to Eq. (8), Eq. (9), and by Lemma 33, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)\\
&\lesssim BC_{M'}+B^2M'+B\sqrt{S^2A\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)}+JH^{2.5}S^2A+J^2H\sqrt{SAM'}+M'H\epsilon.
\end{aligned}$$

Solving a quadratic inequality w.r.t $\sum_{m=1}^{M'}\sum_{h=1}^{H_m}\mathbb{V}(P_h^m,V_{h+1}^m)$, applying AM-GM inequality on $J^2H\sqrt{SAM'}$, and applying Eq. (6) completes the proof. ■

Lemma 33 *(Chen and Luo, 2022, Lemma 9) For any sequence of value functions $\{V_h^m\}_{m,h}$ with $V_h^m \in [0,B]^{\mathcal{S}_+}$ for some $B > 0$, we have $\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m) \lesssim \sum_{m=1}^{M'} V_{H_m+1}^m(s_{H_m+1}^m)^2 + \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(V_h^m(s_h^m)^2 - (P_h^m V_{h+1}^m)^2\right) + B^2$ for all $M' \geq 1$ with probability at least $1-\delta$.*

Proof For any $M' \geq 1$, we decompose the sum as follows:

$$\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m) = \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(P_h^m (V_{h+1}^m)^2 - V_{h+1}^m(s_{h+1}^m)^2\right) + \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(V_{h+1}^m(s_{h+1}^m)^2 - V_h^m(s_h^m)^2\right) + \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(V_h^m(s_h^m)^2 - (P_h^m V_{h+1}^m)^2\right).$$

For the first term, by Lemma 42 and Lemma 40, with prbability at least $1-\delta$,

$$\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(P_h^m (V_{h+1}^m)^2 - V_{h+1}^m(s_{h+1}^m)^2\right) \lesssim \sqrt{\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, (V_{h+1}^m)^2)} + B^2 \lesssim B\sqrt{\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + B^2.$$

The second term is bounded by $\sum_{m=1}^{M'} V_{H_m+1}^m(s_{H_m+1}^m)^2$. Plugging these back and solving a quadratic inequality w.r.t $\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)$ completes the proof. ■

Lemma 34 *(Chen and Luo, 2022, Lemma 10) With probability at least $1-\delta$, for all $M' \geq 1$, $\sum_{m=1}^{M'} \sum_{h=1}^{H_m} b_h^m \lesssim \sqrt{SA \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + JS^{1.5}A$.*

Proof Note that

$$\begin{aligned}
\sum_{m=1}^{M'} \sum_{h=1}^{H_m} b_h^m &\lesssim \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(\sqrt{\frac{\mathbb{V}(\bar{P}_h^m, V_{h+1}^m)}{\mathbf{N}_h^m}} + \frac{J}{\mathbf{N}_h^m}\right) && (\max\{a,b\} \leq a+b) \\
&\lesssim \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \left(\sqrt{\frac{\mathbb{V}(P_h^m, V_{h+1}^m)}{\mathbf{N}_h^m}} + \frac{J\sqrt{S}}{\mathbf{N}_h^m}\right) && \text{(Lemma 24)} \\
&\lesssim \sqrt{SA \sum_{m=1}^{M'} \sum_{h=1}^{H_m} \mathbb{V}(P_h^m, V_{h+1}^m)} + JS^{1.5}A. && \text{(Cauchy-Schwarz inequality and Lemma 35)}
\end{aligned}$$

This completes the proof. ■

Lemma 35 $\sum_{m=1}^{M'} \sum_{h=1}^{H_m} \frac{1}{\mathbf{N}_h^m} \lesssim SA$.

Appendix G. Horizon-free Regret is Impossible in SSP under general costs

Recently Zhang et al. (2022) show that in finite-horizon MDPs it is possible to achieve a horizon-free regret bound with no horizon dependency even in logarithmic terms. For SSPs, Tarbouriech et al. (2021c) achieves a nearly horizon-free regret bound $R_K \lesssim B_\star\sqrt{SAK}\ln\frac{1}{\lambda}+B_\star S^2A+\lambda T_\star K$ for any given $\lambda > 0$ in K episodes without knowledge of $T_\star$, where regret $R_K = \sum_{k=1}^{K}(\sum_{i=1}^{I_k} c(s_i^k, a_i^k) - V^\star(s_{\text{init}}))$, and $R_K = \infty$ if $I_k = \infty$ for some k. If a prior knowledge $\overline{T} = T_\star$ is available, their result is nearly horizon-free with logarithmic dependency on $T_\star$. A natural question to ask is whether (completely) horizon-free regret is possible in SSPs without prior knowledge. We show that this is actually impossible.

Definition 36 *We say an algorithm is (c_1, c_2)-horizon-free if when it takes number of episodes $K \geq 1$, failure probability $\delta \in (0, 1)$, and an SSP instance $\mathcal{M}$ with parameters $B_\star$, S, A as input, it achieves $R_K \leq c_1(B_\star, S, A, K, \delta)\sqrt{K} + c_2(B_\star, S, A, K, \delta)$ on $\mathcal{M}$ with probability at least $1-\delta$, where c_1, c_2 are functions of $B_\star$, S, A, K, δ that have poly-logarithmic dependency on K (no dependency on $T_\star$ and $\frac{1}{c_{\min}}$).*

Theorem 37 *For any c_1, c_2 that are functions of $B_\star$, S, A, K, δ, and have poly-logarithmic dependency on K, there is no (c_1, c_2)-horizon-free algorithm.*

Note that in regret minimization the regret bound can scale with $T_\star$ even without knowledge of $T_\star$, while in sample complexity we cannot (Theorem 3). Therefore, PAC learning in SSP is in some sense more difficult than regret minimization.

Proof [of Theorem 37] Consider an SSP $\mathcal{M}_0$ with $\mathcal{S} = \{s_0, s_1\}$, $\mathcal{A} = \{a_0, a_g\}$ and $s_{\text{init}} = s_0$. The cost function satisfies $c(s_0, a_0) = 0$, $c(s_0, a_g) = 1$, and $c(s_1, a) = \frac{1}{2}$ for $a \in \mathcal{A}$. The transition function satisfies $P(g|s_0, a_g) = 1$, $P(s_0|s_0, a_0) = 1$, and $P(g|s_1, a) = 1$ for $a \in \mathcal{A}$. Clearly, $c_{\min} = 0$ and $B_\star = 1$ in $\mathcal{M}_0$. Suppose the learner is a (c_1, c_2)-horizon-free algorithm for some functions c_1, c_2 as described in Definition 36. Pick $\delta \in (0, \frac{1}{8})$ and K large enough as input to the learner, such that $c_1^0\sqrt{K} + c_2^0 < \frac{K}{2}$ and $c_1^1\sqrt{K} + c_2^1 < \frac{K}{2}$, where $c_i^0 = c_i(1, 2, 2, K, \delta)$ and $c_i^1 = c_i(\frac{1}{2}, 2, 2, K, \delta)$. Let $\mathcal{E}_1$ be the event that the learner reaches the goal state through (s_0, a_g) in all K episodes. Since the learner ensures finite regret with high probability, we have $P(\mathcal{E}_1) \geq 1-\delta$ in $\mathcal{M}_0$. Denote by t the number of times the learner visits (s_0, a_0). By $P(t\mathbb{I}_{\mathcal{E}_1} < \infty) = 1$ in $\mathcal{M}_0$, there exists an integer $n \geq 2$ such that $P(t\mathbb{I}_{\mathcal{E}_1} \leq n) \geq \frac{7}{8}$ in $\mathcal{M}_0$. Define $\mathcal{E}_2 = \{t \leq n\}$ and $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. We have $P(\mathcal{E}) = P(\mathcal{E}_1 \cap \{t\mathbb{I}_{\mathcal{E}_1} \leq n\}) \geq \frac{3}{4}$ in $\mathcal{M}_0$ by $\delta \in (0, \frac{1}{8})$.

Now consider another MDP $\mathcal{M}_1$ that is the same as $\mathcal{M}_0$ except that $P(s_1|s_0, a_0) = \frac{1}{n}$ and $P(s_0|s_0, a_0) = 1 - P(s_1|s_0, a_0)$. Clearly, $B_\star = V^\star(s_{\text{init}}) = \frac{1}{2}$ in $\mathcal{M}_1$. Let W be the interaction history between the learner and the environment, and $L_j(w) = P_j(W = w)$, where P_j is the distribution w.r.t $\mathcal{M}_j$. Also define $\gamma(w) = \mathbb{I}\{L_0(w) > 0\}$. Note that $\frac{L_1(W)}{L_0(W)}\mathbb{I}_{\mathcal{E}}(W)\gamma(W) = (1-\frac{1}{n})^t\mathbb{I}_{\mathcal{E}}(W)\gamma(W) \geq (1-\frac{1}{n})^n\mathbb{I}_{\mathcal{E}}(W)\gamma(W) \geq \frac{\mathbb{I}_{\mathcal{E}}(W)\gamma(W)}{4}$. Therefore,

$$P_1(\mathcal{E}_1) \geq P_1(\mathcal{E}) \geq \mathbb{E}_1[\mathbb{I}_{\mathcal{E}}(W)\gamma(W)] = E_0\left[\frac{L_1(W)}{L_0(W)}\mathbb{I}_{\mathcal{E}}(W)\gamma(W)\right] \geq \frac{P_0(\mathcal{E})}{4} \geq \frac{3}{16} > \frac{1}{8}.$$

Note that the learner ensures $R_K \leq c_1^i\sqrt{K} + c_2^i$ with probability at least $1-\delta$ in $\mathcal{M}_i$ for $i \in \{0, 1\}$. Moreover, when $\mathcal{E}_1$ is true, $R_K = \frac{K}{2} > c_1^1\sqrt{K} + c_2^1$ in $\mathcal{M}_1$. Therefore, the learner must ensure $P_1(\mathcal{E}_1) < \delta < \frac{1}{8}$, a contradiction. This completes the proof. ■

Appendix H. Auxiliary Lemmas

Lemma 38 *(Rosenberg and Mansour, 2020, Lemma 6) Let π be a policy whose expected hitting time starting from any state is at most τ. Then the probability that π takes more than n steps to reach the goal state is at most $2e^{-\frac{n}{4\tau}}$.*

Lemma 39 *(Zhang et al., 2020b, Lemma 14) Define $\Upsilon = \{v \in [0,B]^{\mathcal{S}_+} : v(g) = 0\}$. Let $f : \Delta_{\mathcal{S}_+} \times \Upsilon \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ with $f(p,v,n,B,\iota) = pv - \max\left\{c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}}, c_2\frac{B\iota}{n}\right\}$ with $c_1^2 \le c_2$. Then, f is non-increasing in v, that is, for all $p \in \Delta_{\mathcal{S}_+}, v, v' \in \Upsilon$ and $n, \iota > 0$,*

$$v(s) \le v'(s), \forall s \in \mathcal{S}^+ \implies f(p,v,n,B,\iota) \le f(p,v',n,B,\iota).$$

Lemma 40 *(Chen and Luo, 2022, Lemma 16) For any random variable $X \in [-C,C]$ for some $C > 0$, we have $\textsc{Var}[X^2] \le 4C^2\textsc{Var}[X]$.*

Lemma 41 *(Chen et al., 2021a, Lemma 34) Let $\{X_t\}_t$ be a sequence of i.i.d random variables with mean μ, variance σ^2, and $0 \le X_t \le B$. Then with probability at least $1-\delta$, the following holds for all $n \ge 1$ simultaneously:*

$$\left|\sum_{t=1}^{n}(X_t - \mu)\right| \le 2\sqrt{2\sigma^2 n \ln\frac{2n}{\delta}} + 2B\ln\frac{2n}{\delta}.$$
$$\left|\sum_{t=1}^{n}(X_t - \mu)\right| \le 2\sqrt{2\hat{\sigma}_n^2 n \ln\frac{2n}{\delta}} + 19B\ln\frac{2n}{\delta}.$$

where $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{t=1}^{n}X_t^2 - (\frac{1}{n}\sum_{t=1}^{n}X_t)^2$.

Lemma 42 *(Chen et al., 2021b, Lemma 38) Let $\{X_i\}_{i=1}^{\infty}$ be a martingale difference sequence adapted to the filtration $\{\mathcal{F}_i\}_{i=0}^{\infty}$ and $|X_i| \le B$ for some $B > 0$. Then with probability at least $1-\delta$, for all $n \ge 1$ simultaneously,*

$$\left|\sum_{i=1}^{n}X_i\right| \le 3\sqrt{\sum_{i=1}^{n}\mathbb{E}[X_i^2|\mathcal{F}_{i-1}]\ln\frac{4B^2n^3}{\delta}} + 2B\ln\frac{4B^2n^3}{\delta}.$$

Lemma 43 *(Chen et al., 2022, Lemma 51) Let $\{X_i\}_{i=1}^{\infty}$ with $X_i \in [0,B]$ be a martingale sequence w.r.t the filtration $\{\mathcal{F}_i\}_{i=0}^{\infty}$. Then with probability at least $1-\delta$, for all $n \ge 1$,*

$$\sum_{i=1}^{n}\mathbb{E}[X_i|\mathcal{F}_{i-1}] \le 2\sum_{i=1}^{n}X_i + 4B\ln\frac{4n}{\delta}.$$

Lemma 44 *(Chen et al., 2021a, Lemma 34) Let $\{X_t\}_t$ be a sequence of i.i.d random variables with mean μ, variance σ^2, and $0 \le X_t \le B$. Then with probability at least $1-\delta$, the following holds for all $n \ge 1$ simultaneously:*

$$\left|\sum_{t=1}^{n}(X_t - \mu)\right| \le 2\sqrt{2\sigma^2 n \ln\frac{2n}{\delta}} + 2B\ln\frac{2n}{\delta}.$$
$$\left|\sum_{t=1}^{n}(X_t - \mu)\right| \le 2\sqrt{2\hat{\sigma}_n^2 n \ln\frac{2n}{\delta}} + 19B\ln\frac{2n}{\delta}.$$

where $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{t=1}^{n}X_t^2 - (\frac{1}{n}\sum_{t=1}^{n}X_t)^2$.