Layered State Discovery for Incremental Autonomous Exploration

Liyu Chen [1] **Andrea Tirinzoni** [2] **Alessandro Lazaric** [2] **Matteo Pirotta** [2]

Abstract

We study the autonomous exploration (AX) problem proposed by Lim & Auer (2012). In this setting, the objective is to discover a set of ϵ-optimal policies reaching a set $\mathcal{S}_L^{\rightarrow}$ of incrementally L-controllable states. We introduce a novel layered decomposition of the set of incrementally L-controllable states that is based on the iterative application of a state-expansion operator. We leverage these results to design Layered Autonomous Exploration (LAE), a novel algorithm for AX that attains a sample complexity of $\tilde{\mathcal{O}}(LS^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A\log^{12}(S^{\rightarrow}_{L(1+\epsilon)})/\epsilon^2)$, where $S^{\rightarrow}_{L(1+\epsilon)}$ is the number of states that are incrementally $L(1+\epsilon)$-controllable, A is the number of actions, and $\Gamma_{L(1+\epsilon)}$ is the branching factor of the transitions over such states. LAE improves over the algorithm of Tarbouriech et al. (2020b) by a factor of L^2 and it is the first algorithm for AX that works in a countably-infinite state space. Moreover, we show that, under a certain identifiability assumption, LAE achieves minimax-optimal sample complexity of $\tilde{\mathcal{O}}(LS^{\rightarrow}_{L}A\log^{12}(S^{\rightarrow}_{L})/\epsilon^2)$, outperforming existing algorithms and matching for the first time the lower bound proved by Cai et al. (2022) up to logarithmic factors.

1. Introduction

A distinctive feature of intelligent beings is the ability to explore an unknown environment without any supervision or extrinsic reward while learning skills that solve tasks (e.g., reaching goal states) of increasing difficulty. Lim & Auer (2012) first proposed a formal framework of *autonomous exploration* in reinforcement learning (RL) as the process of progressively discovering states within a certain distance from an initial state s_0 at the same time as learning near-optimal policies to reach them. Lim & Auer (2012) also devised the first sample efficient exploration algorithm (UcbExplore) for this setting, while its sample complexity and optimality guarantees were later improved by DisCo (Tarbouriech et al., 2020b) and VALAE (Cai et al., 2022).

In this paper, we make several contributions to this problem:

- Given an initial state s_0, the autonomous exploration objective is built upon the concept of incrementally L-controllable states, i.e., states that can be reached within L steps from s_0 by only traversing incrementally L-controllable states[1]. While the original definition of the set of incrementally L-controllable states $\mathcal{S}_L^{\rightarrow}$ involves considering all possible partial orders of states in the environment, we derive an equivalent constructive definition that reveals the *layered* structure of $\mathcal{S}_L^{\rightarrow}$, where each layer can be obtained as the set of states that can be reached in L steps by only traversing states in the previous layers (see Section 2.1).
- We then leverage the layered structure of $\mathcal{S}_L^{\rightarrow}$ to design Layered Autonomous Exploration (LAE), a novel algorithm that keeps exploring the environment to learn policies to reach newly discovered states until a new layer can be consolidated and a new step of discovery and learning is started. We prove that the sample complexity of LAE is bounded as $\tilde{\mathcal{O}}(LS^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A/\epsilon^2)$, where L is the exploration radius, $S^{\rightarrow}_{L(1+\epsilon)}$ is the number of states that are incrementally controllable from the initial state within $L(1+\epsilon)$ steps, $\Gamma_{L(1+\epsilon)}$ is the branching factor of the transition function over such states, A is the number actions, and ϵ is target accuracy. As illustrated in Table 1, this improves the sample complexity of DisCo by a factor of L^2 and it avoids the scaling with $S^{\rightarrow}_{2L}$ of VALAE, which in some MDPs may be much larger than $S^{\rightarrow}_{L(1+\epsilon)}$, thus making the bound of LAE preferable. Indeed, in Lemma 43 in appendix we show that $S^{\rightarrow}_{2L}$ may be even exponentially larger than $S^{\rightarrow}_{L(1+\epsilon)}$.
- Under a certain layer identifiability condition (see Assumption 2), we further improve the sample complexity of

[1]University of Southern California [2]Meta. Correspondence to: Liyu Chen <liyuc@usc.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]We say that a state s is L-controllable if there exists a policy that reaches s from s_0 in less than L steps on average. In general an L-controllable state may be reached by policies traversing states that are not L-controllable themselves.

Table 1. Comparison between this work and previous work. Here, L is the exploration radius, S is the number of states, $S^{\rightarrow}_{L(1+\epsilon)}$ is the number of incrementally $L(1+\epsilon)$-controllable states, $\Gamma_{L(1+\epsilon)}$ is the branching factor of transition over such states, A is the number of actions, and ϵ is the target accuracy. The AX objectives are defined in Definition 2 and are such that $\text{AX}^{+} \Rightarrow \text{AX}^{\star} \Rightarrow \text{AX}_L$. We only display the dominating term in $1/\epsilon$. Note that $S^{\rightarrow}_{2L}$ may be much larger (even exponentially) than $S^{\rightarrow}_{L(1+\epsilon)}$ in certain MDPs (Lemma 43).

Algorithm		Sample Complexity	Objective	S dependency
UcbExplore	(Lim & Auer, 2012)	$\tilde{\mathcal{O}}\left(L^3 S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A/\epsilon^3\right)$	AX_L	$\log S$
DisCo	(Tarbouriech et al., 2020b)	$\tilde{\mathcal{O}}\left(L^3 S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A/\epsilon^2\right)$	$\text{AX}^{\star}$	$\log S$
VALAE	(Cai et al., 2022)	$\tilde{\mathcal{O}}\left(L S^{\rightarrow}_{2L}A/\epsilon^2\right)$	$\text{AX}^{\star}$	$\log S$
LAE (Algorithm 3)	Ours	$\tilde{\mathcal{O}}\left(L S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A/\epsilon^2\right)$	AX^{+}	$\log S^{\rightarrow}_{L(1+\epsilon)}$
LAE with Assumption 2	Ours	$\tilde{\mathcal{O}}\left(L S^{\rightarrow}_{L}A/\epsilon^2\right)$	AX^{+}	$\log S^{\rightarrow}_{L}$
Lower Bound ($S^{\rightarrow}_{L} = S^{\rightarrow}_{L(1+\epsilon)}$ by construction)	(Cai et al., 2022)	$\Omega\left(L S^{\rightarrow}_{L}A/\epsilon^2\right)$	AX_L	-

LAE to $\tilde{\mathcal{O}}(LS^{\rightarrow}_{L}A/\epsilon^2)$, which improves w.r.t. VALAE and matches the lower bound in (Cai et al., 2022).

- Similar to existing algorithms, the sample complexity of LAE still depends on the logarithm of the total number of states S. Since in autonomous exploration the state space is unknown and possibly unbounded, such dependency is highly undesirable. We then design an alternative version of LAE, which preserves its original sample complexity but replaces the dependency on $\log S$ with $\log S^{\rightarrow}_{L(1+\epsilon)}$, without requiring any prior knowledge of $S^{\rightarrow}_{L(1+\epsilon)}$ (see Section 4.1).
- LAE also leverages a novel procedure, PolicyConsolidation, that takes a set of states $\mathcal{K}$ as input and returns goal-conditioned policies reaching each state in $\mathcal{K}$ with *multiplicative* ϵ-optimality guarantees, which is stronger than previous algorithms and better suited to the autonomous exploration setting (see Section 4.2).

Related Work In reinforcement learning (RL), several approaches to *unsupervised exploration* have been proposed often grounded in concepts such as curiosity (Schmidhuber, 1991), intrinsic motivation (Singh et al., 2004; Oudeyer et al., 2009; Bellemare et al., 2016; Colas et al., 2020) and with the objective of learning skills in an unsupervised fashion (Gregor et al., 2016; Eysenbach et al., 2019; Pong et al., 2020; Bagaria et al., 2021; Kamienny et al., 2022). On the other hand, a rigorous formalization and theoretical understanding of unsupervised exploration has been rather sparse until recently. Tarbouriech et al. (2020c) studied unsupervised exploration for model estimation, Hazan et al. (2019) formalized the maximum entropy exploration objective, while reward-free RL (e.g., Jin et al., 2020; Kaufmann et al., 2021; Ménard et al., 2021; Zhang et al., 2021; Tarbouriech et al., 2021a; 2022) studies how to efficiently explore an environment to solve any downstream task near-optimally. As autonomous exploration seeks to learn goal-conditioned policies, it also carries strong technical and algorithmic connections with exploration in the stochastic shortest path problem (e.g. Bertsekas & Yu, 2013; Tarbouriech et al., 2020a; 2021b; Chen & Luo, 2021; 2022).

2. Preliminaries

We consider a reward-free Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_0, P)$, where $\mathcal{S}$ is a countable state space, $\mathcal{A}$ is a finite action space, s_0 is the initial state, and $P = \{P_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ with $P_{s,a} \in \Delta_{\mathcal{S}}$ is the transition function, where $\Delta_{\mathcal{S}}$ is the simplex over $\mathcal{S}$. In a general MDP, the learner may get stuck in undesirable states and be unable to return to s_0. To avoid this issue, we make the following assumption.

Assumption 1. *The action space contains a* RESET *action such that* $P_{s,\text{RESET}}(s_0) = 1$ *for all* $s \in \mathcal{S}$.[2]

A deterministic stationary policy $\pi \in \mathcal{A}^{\mathcal{S}}$ is a mapping that assigns an action $\pi(s)$ to each state s, and we define $\Pi = \mathcal{A}^{\mathcal{S}}$ as the set of all policies. To explicitly characterize the behavior of a policy, we say a policy π is *restricted* on $\mathcal{X} \subseteq \mathcal{S}$ if $\pi(s) = \text{RESET}$ for any $s \notin \mathcal{X}$, and we denote by $\Pi(\mathcal{X})$ the set of policies restricted on $\mathcal{X}$.

We measure the performance of a policy in navigating the MDP as follows. For any policy $\pi \in \Pi$ and a pair of states $(s, g) \in \mathcal{S}^2$, let $V^{\pi}_{g}(s) \in [0, +\infty]$ be the expected number of steps it takes to reach g (that is, the *hitting time* of g)

[2] This assumption is also adopted in all previous works (Lim & Auer, 2012; Tarbouriech et al., 2020b; Cai et al., 2022) to our knowledge.

starting from s when executing policy π, that is,

$$V_g^\pi(s) \triangleq \mathbb{E}^\pi \left[\omega_g \,|\, s_1 = s\right], \\ \omega_g \triangleq \inf\left\{i \geq 0 : s_{i+1} = g\right\}.$$

Note that $V_g^\pi(s) = +\infty$ if g is unreachable by playing π starting from s. For any subset $\mathcal{X} \subseteq \mathcal{S}$ and any goal state g, define $V^\star_{\mathcal{X},g}(s) = \min_{\pi\in\Pi(\mathcal{X})} V_g^\pi(s)$ as the minimum hitting time of g following a policy restricted on $\mathcal{X}$. Note that, if $\mathcal{X} \subseteq \mathcal{X}'$, then $V^\star_{\mathcal{X}',g}(s) \leq V^\star_{\mathcal{X},g}(s)$ for any $s, g \in \mathcal{S}$. The objective of the learner is to efficiently navigate in the vicinity of s_0. A state s is *L-controllable* if there exists a policy π such that $V_s^\pi(s_0) \leq L$. While discovering all L-controllable states may be a reasonable objective for exploring the vicinity of s_0 (Tarbouriech et al., 2022), Lim & Auer (2012) showed that this may still require the learner to explore the whole state space, since reaching a L-controllable state may require navigating through non-L-controllable states. To this end, Lim & Auer (2012) propose to only focus on navigating among *incrementally L-controllable states*: states that are L-controllable by policies restricted on other incrementally controllable states.

Definition 1 (Incrementally L-controllable states $\mathcal{S}_L^{\rightarrow}$). *Given a partial order $\prec$ on $\mathcal{S}$, we define $\mathcal{S}_L^{\prec}$ recursively as 1) $s_0 \in \mathcal{S}_L^{\prec}$ and 2) if there exists a policy $\pi \in \Pi\big(\{s' \in \mathcal{S}_L^{\prec} : s' \prec s\}\big)$ with $V_s^\pi(s_0) \leq L$, then $s \in \mathcal{S}_L^{\prec}$. The set $\mathcal{S}_L^{\rightarrow}$ of incrementally L-controllable states is defined as $\mathcal{S}_L^{\rightarrow} \triangleq \cup_{\prec} \mathcal{S}_L^{\prec}$, where the union is over all partial orders.*

Instead of exploring the potentially infinite state space, the objective of the learner is to discover the *finite* set $\mathcal{S}_L^{\rightarrow}$ (Lim & Auer, 2012, Prop. 6) and learn a corresponding set of policies that reliably reach each state in $\mathcal{S}_L^{\rightarrow}$. We introduce three different formulations of the objective.

Definition 2 (AX sample complexity). *For any given length $L \geq 1$, error threshold $\epsilon > 0$, and confidence level $\delta \in (0,1)$, the sample complexities $\mathcal{C}(\mathfrak{A}, L, \epsilon, \delta)$, $\mathcal{C}^\star(\mathfrak{A}, L, \epsilon, \delta)$, and $\mathcal{C}^+(\mathfrak{A}, L, \epsilon, \delta)$ are defined as the number of steps required by a learning algorithm $\mathfrak{A}$ to identify a set of states $\mathcal{K}$ and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ such that, with probability at least $1-\delta$, we have $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}$ and*

(AX_L) $\quad \forall s \in \mathcal{S}_L^{\rightarrow}$, $V_s^{\pi_s}(s_0) \leq L(1+\epsilon)$,

($AX^\star$) $\quad \forall s \in \mathcal{S}_L^{\rightarrow}$, $V_s^{\pi_s}(s_0) \leq V^\star_{\mathcal{S}_L^{\rightarrow},s}(s_0) + L\epsilon$,

(AX^+) $\quad \forall s \in \mathcal{S}_L^{\rightarrow}$, $V_s^{\pi_s}(s_0) \leq V^\star_{\mathcal{S}_L^{\rightarrow},s}(s_0)(1+\epsilon)$.

Note that the three formulations above are increasingly more demanding. AX_L only requires to reach each state in $\mathcal{S}_L^{\rightarrow}$ within $L(1+\epsilon)$ steps, which could correspond to a quite poor performance for a state s with $V^\star_{\mathcal{S}_L^{\rightarrow},s}(s_0) \ll L$. $\text{AX}^\star$ requires to learn a near-optimal policy for reaching each state in $\mathcal{S}_L^{\rightarrow}$. However, the allowed error threshold (i.e., $L\epsilon$) is uniform across all goal states, which again could correspond to a bad performance for a state s with $V^\star_{\mathcal{S}_L^{\rightarrow},s}(s_0) \ll L$. AX^+ solves this issue by requiring a *multiplicative* threshold. This implies that the allowed error for reaching state s (i.e., $V^\star_{\mathcal{S}_L^{\rightarrow},s}(s_0)\epsilon$) scales with the optimal value $V^\star_{\mathcal{S}_L^{\rightarrow},s}(s_0)$ itself, hence making this formulation adaptive to the hardness of reaching each goal state. No existing algorithm is able to achieve AX^+ guarantees, see Table 1.

Note that these conditions cannot be checked at algorithmic time since $\mathcal{S}_L^{\rightarrow}$ is unknown to the algorithm. Existing algorithms verify these conditions directly on the computed set $\mathcal{K}$. Since they guarantee that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}$, $V^\star_{\mathcal{K},g}(s_0) \leq V^\star_{\mathcal{S}_L^{\rightarrow},g}(s_0)$ for any $g \in \mathcal{S}_L^{\rightarrow}$ and thus they satisfy the performance in Definition 2.

Other notation Let $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$. For any $L \geq 1$, define $S_L^{\rightarrow} = |\mathcal{S}_L^{\rightarrow}|$, $\mathcal{N}_L^{s,a} = \{s' \in \mathcal{S}_L^{\rightarrow} : P_{s,a}(s') > 0\}$, $\Gamma_L^{s,a} = |\mathcal{N}_L^{s,a}|$ and $\Gamma_L = \max_{s\in\mathcal{S}_L^{\rightarrow},a} \Gamma_L^{s,a}$. For simplicity, we often write $a = \mathcal{O}(b)$ as $a \lesssim b$. For $n \in \mathbb{N}_+$, define $[n] = \{1, \ldots, n\}$.

2.1. A Constructive Definition of $\mathcal{S}_L^{\rightarrow}$

While Lim & Auer (2012, Proposition 6) showed that there exists a partial order $\prec$ such that $\mathcal{S}_L^{\rightarrow} = \mathcal{S}_L^{\prec}$, no explicit characterization of such partial order is provided. In the following, we develop an alternative definition of $\mathcal{S}_L^{\rightarrow}$ that leads to an explicit constructive procedure to build the set. This alternative definition is the main inspiration for the design of our algorithms.

We introduce an operator $\mathcal{T}_L$ which, given a set $\mathcal{X} \subseteq \mathcal{S}$, selects all the states that are reachable in L steps by a policy restricted on $\mathcal{X}$ and show its connection with $\mathcal{S}_L^{\rightarrow}$.

Lemma 1. *Let $\mathsf{P}(\mathcal{S})$ be the set of all subsets of $\mathcal{S}$. For any $L \geq 1$, define the operator $\mathcal{T}_L : \mathsf{P}(\mathcal{S}) \rightarrow \mathsf{P}(\mathcal{S})$ as follows: for any $\mathcal{X} \subseteq \mathcal{S}$, $\mathcal{T}_L(\mathcal{X}) = \{s \in \mathcal{S} : V^\star_{\mathcal{X},s}(s_0) \leq L\}$. Then,*

1. *$\mathcal{S}_L^{\rightarrow}$ is the fixed-point of $\mathcal{T}_L$ of smallest cardinality, i.e., $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{X}$ if $\mathcal{X} = \mathcal{T}_L(\mathcal{X})$.*

Let us denote by $\{\mathcal{K}_j^\star\}_{j\in\mathbb{N}}$ the unique sequence such that $\mathcal{K}_1^\star = \{s_0\}$, $\mathcal{K}_j^\star = \mathcal{T}_L(\mathcal{K}_{j-1}^\star)$. Then,

2. *For any $j \geq 1$, $\mathcal{K}_j^\star \subseteq \mathcal{K}_{j+1}^\star \subseteq \mathcal{S}_L^{\rightarrow}$;*
3. *There exists $J \leq S_L^{\rightarrow}$ such that $\mathcal{K}_j^\star = \mathcal{S}_L^{\rightarrow}$ for all $j \geq J$ (i.e., $\mathcal{T}_L^J(\mathcal{K}_1^\star) = \lim_{j\rightarrow\infty} \mathcal{T}_L^j(\mathcal{K}_1^\star) = \mathcal{S}_L^{\rightarrow}$).*

Proof. Note that there exists a partial ordering $\prec^\star$ such that $\mathcal{S}_L^{\rightarrow} = \mathcal{S}_L^{\prec^\star}$ (Lim & Auer, 2012, Proposition 6).

Let $\mathcal{X}$ be s.t. $\mathcal{S}_L^{\rightarrow} \not\subseteq \mathcal{X}$. If $\mathcal{S}_L^{\rightarrow} \cap \mathcal{X} = \emptyset$, then $s_0 \notin \mathcal{X}$, which implies that $\mathcal{T}_L(\mathcal{X}) = \{s_0\}$ since $V^\star_{\mathcal{X},s_0}(s_0) = 0 \leq L$ and $V^\star_{\mathcal{X},g}(s_0) = \infty$ for all $g \neq s_0$. Thus, $\mathcal{X}$ cannot be a fixed point of $\mathcal{T}_L$. Then, assume that $\mathcal{S}_L^{\rightarrow} \cap \mathcal{X} \neq \emptyset$. Order the states in $\mathcal{X} \cap \mathcal{S}_L^{\rightarrow}$ according to the ordering $\prec^\star$. Let $s_i \in \mathcal{S}_L^{\prec^\star}$ be the first state s.t. $s \notin \mathcal{X}$ (it exists since $\mathcal{S}_L^{\rightarrow} \not\subseteq \mathcal{X}$). By definition of $\prec^\star$ and $\mathcal{S}_L^{\rightarrow}$, $V^\star_{\{s_0,\dots,s_{i-1}\},s_i}(s_0) \le L$, which implies that $s_i \in \mathcal{T}_L(\mathcal{X})$. As a consequence, $\mathcal{X} \neq \mathcal{T}_L(\mathcal{X})$. Thus, if $\mathcal{X} = \mathcal{T}_L(\mathcal{X})$, we must have $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{X}$. This proves the first point.

Let us prove that $\mathcal{K}_j^\star \subseteq \mathcal{K}_{j+1}^\star$ for all $j \geq 1$. Clearly, $\mathcal{K}_2^\star = \mathcal{T}_L(\mathcal{K}_1^\star) = \{s \in \mathcal{S} : V_{\{s_0\},s}(s_0) \le L\} \supseteq \{s_0\} = \mathcal{K}_1^\star$. Then, suppose that $\mathcal{K}_{j-1}^\star \subseteq \mathcal{K}_j^\star$ for some $j \geq 2$. By definition, for all $s \in \mathcal{K}_j^\star$, $V^\star_{\mathcal{K}_{j-1}^\star,s}(s_0) \le L$, which implies that $V^\star_{\mathcal{K}_j^\star,s}(s_0) \le L$ by the inductive hypothesis. Then, $\mathcal{K}_{j+1}^\star = \mathcal{T}_L(\mathcal{K}_j^\star) = \{s \in \mathcal{S} : V_{\mathcal{K}_j^\star,s}(s_0) \le L\} \supseteq \mathcal{K}_j^\star$.

Now let us prove that $\mathcal{K}_j^\star \subseteq \mathcal{S}_L^{\rightarrow}$ for all $j \geq 1$. Clearly, $\mathcal{K}_1^\star \subseteq \mathcal{S}_L^{\rightarrow}$. Suppose that $\mathcal{K}_j^\star \subseteq \mathcal{S}_L^{\rightarrow}$ for some $j \geq 1$. Then, if $s \in \mathcal{K}_{j+1}^\star$ for some $s \notin \mathcal{S}_L^{\rightarrow}$, it must be that $V_{\mathcal{K}_j^\star,s}(s_0) \le L$. By the inductive hypothesis, this implies that we found an ordering of the states in which s is reachable in L steps by a policy restricted on states of $\mathcal{S}_L^{\rightarrow}$. Hence, $s \in \mathcal{S}_L^{\rightarrow}$, which is a contradiction. This proves point 2.

Let us enumerate over $\mathcal{S}_L^{\rightarrow} = \{s_0, \dots, s_{S_L^{\rightarrow}-1}\}$ in a way that obeys $\prec^\star$. We prove by induction that $s_j \in \mathcal{K}_{j+1}^\star$ for any $0 \le j < S_L^{\rightarrow}$. Given point 2, this implies point 3. Clearly, $s_0 \in \mathcal{K}_1^\star$. Now suppose that $\{s_0, \dots, s_j\} \in \mathcal{K}_{j+1}^\star$ for $0 \le j \le S_L^{\rightarrow} - 2$. Then, we clearly have $s_{j+1} \in \mathcal{K}_{j+2}^\star$ by the definition of $\mathcal{K}_{j+2}^\star$ and the fact that s_{j+1} is L-controllable by a policy restricted on $\{s_0, \dots, s_j\}$. $\square$

This lemma shows that $\mathcal{S}_L^{\rightarrow}$ is a fixed-point solution of $\mathcal{T}_L$. Most importantly, it provides an iterative procedure to construct $\mathcal{S}_L^{\rightarrow}$. Starting from $\{s_0\}$ or $\emptyset$, $\mathcal{T}_L$ acts as an expansive operator over sets (i.e., $T^j(\{s_0\}) \subset T^{j+1}(\{s_0\})$) until the set $\mathcal{S}_L^{\rightarrow}$ is built. From this point, $\mathcal{T}_L$ acts as an identity map since $\mathcal{S}_L^{\rightarrow}$ is a fixed point. In other words, this procedure builds $\mathcal{S}_L^{\rightarrow}$ iteratevely starting from $\mathcal{K}_1^\star$, expanding it to $\mathcal{K}_2^\star = \mathcal{T}_L(\mathcal{K}_1^\star)$, and so on until reaching $\mathcal{S}_L^{\rightarrow}$. For this reason, we shall refer to the sets $(\mathcal{K}_j^\star)_j$ as *layers*. This process is learnable since it evolves only through subsets of $\mathcal{S}_L^{\rightarrow}$ and it is at the core of the design of our algorithm.

It is worth noticing that not all the fixed-point solutions of $\mathcal{T}_L$ are learnable. In fact, Proposition 4 of Lim & Auer (2012) implies that there exist MDPs with fixed points $\mathcal{X} = \mathcal{T}_L(\mathcal{X}) \neq \mathcal{S}_L^{\rightarrow}$ which may require an exponential number of samples to be learned. For example, there exist MDPs where the whole set of states $\mathcal{S}$ is itself a fixed point of $\mathcal{T}_L$ (that is, all states are L-controllable) but $\mathcal{S}$ is exponentially larger than $\mathcal{S}_L^{\rightarrow}$. This reveals an interesting connection between the existence of a *unique* iterative process to reach the fixed-point corresponding to $\mathcal{S}_L^{\rightarrow}$ and its learnability.

3. AX$_L$ through Layer Discovery

Algorithm 1 illustrates Layer-Aware State Discovery (LASD), a novel algorithm for AX$_L$ based on the iterative construction of $\mathcal{S}_L^{\rightarrow}$ introduced in Lemma 1. In Section 4.2, we then introduce a policy consolidation procedure that achieves AX$^+$ when combined with LASD, leading to the LAE algorithm. LASD maintains a set $\mathcal{K}$ of "known" states, i.e., states for which a policy $\widetilde{\pi}_s \in \Pi(\mathcal{K})$ with $V_s^{\widetilde{\pi}_s}(s_0) \le L(1+\epsilon)$ has been learned. These policies are stored in $\Pi_\mathcal{K}$. The set $\mathcal{K}$ is updated only when the algorithm is confident enough to have identified a new layer. To this purpose, $\mathcal{K}'$ is used as a buffer for the new layer, i.e., for states that have been found to be L-controllable by policies restricted on $\mathcal{K}$ and that are waiting to be merged with $\mathcal{K}$. Finally, any other state discovered over time (and potential candidate to be in $\mathcal{S}_L^{\rightarrow}$) is stored in $\mathcal{U}$.

At each round, LASD first uses the samples collected so far to compute an optimistic policy for each state in $\mathcal{U}$ through VISGO (Algorithm 4), a slight variant of the state-of-the-art algorithm for exploration-exploitation in stochastic shortest paths (Tarbouriech et al., 2021b), and it selects the state that is optimistically closer to s_0 as candidate goal $g^\star$.

If the optimistic distance of $g^\star$ from s_0 is larger than L, then no additional state can be confidently added to the current layer $\mathcal{K}'$ and a *set expansion* round is triggered. LASD updates the set of known states by adding the new layer $\mathcal{K}'$ ($\mathcal{K} = \mathcal{K} \cup \mathcal{K}'$) and starts a discovery process where policies in $\Pi_\mathcal{K}$ are used to reach all states in $\mathcal{K}$, then it executes all possible actions in these states, and it adds newly observed states to $\mathcal{U}$. Notice that the samples obtained during this process are not included in the policy improvement of VISGO to avoid statistical dependencies. The sequence of expansion rounds is designed to approximate the sequence $\{\mathcal{K}_j^\star\}_j$. With high probability, every update of $\mathcal{K}$ is not smaller than the application of $\mathcal{T}_L$, i.e., if, for some j, $\mathcal{K}_j^\star \subseteq \mathcal{K} \not\supseteq \mathcal{K}_{j+1}^\star$ before an update (this holds for $\mathcal{K}_1^\star = \{s_0\}$ at the first round), then $\mathcal{K}_{j+1}^\star = \mathcal{T}_L(\mathcal{K}_j^\star) \subseteq \mathcal{K}$ after the update. Thus, $\mathcal{K}'$ is the increment to $\mathcal{K}$ to include the next layer. At the end of the expansion round LASD executes an additional exploration step to ensure that a minimum number of samples is available for each $(s, a) \in \mathcal{K} \times \mathcal{A}$ (see Line 10).

On the other hand, if the optimistic distance of $g^\star$ is smaller than L, LASD performs a *policy evaluation* round by running $\pi_{g^\star}$ to estimate whether the current policy is indeed able to reach $g^\star$ in less than L steps. If the number of visits to some state-action pair is doubled within the current round, then the current round is classified as a *skip round*. If the test on the policy performance fails, then the current round is classified as a *failure round*. In both cases, a new round is started. Otherwise, the current round is classified as a success round and $g^\star$ is added to the new layer $\mathcal{K}'$. The samples collected in policy evaluation rounds are stored and

Algorithm 1: Layer-Aware State Discovery (LASD)

Input: $L \geq 1$, $\epsilon \in (0,1]$, $\delta \in (0,1)$.

1 Let $\mathfrak{N} = \{2^j\}_{j\geq 0}$, $\mathcal{K} \leftarrow \varnothing$, $\mathcal{U} \leftarrow \varnothing$, $\mathcal{K}' \leftarrow \{s_0\}$, $\Pi_{\mathcal{K}} = \{\widetilde{\pi}_{s_0}$ a random policy$\}$, $\mathbf{N}(\cdot,\cdot) \leftarrow 0$, $\mathbf{N}(\cdot,\cdot,\cdot) \leftarrow 0$.
2 **for** *round* $r = 1, \ldots$ **do**
3     $\epsilon_{\mathrm{VI}} \leftarrow 1/\max\{16, \sum_{s,a}\mathbf{N}(s,a)\}$.
    `/* Policy optimisation and goal selection */`
4   Let $g^\star = \mathrm{argmin}_{g\in\mathcal{U}}\left\{V_{\mathcal{K},g}(s_0)\right\}$ where $(Q_{\mathcal{K},g}, V_{\mathcal{K},g}, \pi_g) = \mathrm{VISGO}(\mathcal{K}, g, \epsilon_{\mathrm{VI}}, \mathbf{N}, \frac{\delta}{4r^2S^2})$ (see Algorithm 4).
5   **if** *$g^\star$ does not exist or $V_{\mathcal{K},g^\star}(s_0) > L$* **then**
        `/* Expand or Terminate */`
6   **if** $\mathcal{K}' = \varnothing$ **then return** $\mathcal{K}$ and $\Pi_{\mathcal{K}}$.
7   Set $\mathcal{K} \leftarrow \mathcal{K} \cup \mathcal{K}'$, $\mathcal{K}' = \varnothing$, $\mathcal{U} = \varnothing$.
8         $(\_, \mathcal{U}) \leftarrow \textsc{Explore}(\mathcal{K}, \Pi_{\mathcal{K}}, 0, 2L\log(4SALr^2/\delta))$ (see Algorithm 6).
9   Set $n_{\min} \leftarrow N_0(\mathcal{K}, \frac{\delta}{4r^2S^2}) \lesssim L^2|\mathcal{K}|\log(Sr/\delta)$ (defined in Lemma 3).
10         $(\mathbf{N}, \_) \leftarrow \textsc{Explore}(\mathcal{K}, \Pi_{\mathcal{K}}, \mathbf{N}, n_{\min})$.
11   **else**
        `/* Policy evaluation */`
12   Let $\widehat{\tau} \leftarrow 0$, $\lambda \leftarrow N_{\mathrm{DEV}}(32L, \frac{\epsilon}{256}, \frac{\delta}{4r^2}) \lesssim \frac{1}{\epsilon^2}\log^4\left(\frac{Lr}{\epsilon\delta}\right)$ (defined in Lemma 50).
13   **for** $j = 1, \ldots, \lambda$ **do**
14             $k \xleftarrow{+} 1$, $i \leftarrow 1$, and reset to $s_1^k \leftarrow s_0$ by taking action RESET.
15   **while** $s_i^k \neq g^\star$ **do**
16   Take $a_i^k = \pi_{g^\star}(s_i^k)$, and transits to s_{i+1}^k. Increase $\mathbf{N}(s_i^k, a_i^k)$, $\mathbf{N}(s_i^k, a_i^k, s_{i+1}^k)$, and i by 1.
17   **if** $\sum_{s,a}\mathbf{N}(s,a) \in \mathfrak{N}$ *or* ($s_i^k \in \mathcal{K}$ *and* $\mathbf{N}(s_i^k, a_i^k) \in \mathfrak{N}$) **then** return to Line 2 (skip round).
18   Set $\widehat{\tau} \xleftarrow{+} \frac{c(s_i^k, a_i^k)}{\lambda}$.
19   **if** $\widehat{\tau} > V_{\mathcal{K},g^\star}(s_0) + \epsilon L/2$ **then** return to Line 2 (failure round).
20         $\mathcal{K}' \leftarrow \mathcal{K}' \cup \{g^\star\}$, $\mathcal{U} \leftarrow \mathcal{U} \setminus \{g^\star\}$, $\Pi_{\mathcal{K}} = \Pi_{\mathcal{K}} \cup \{\widetilde{\pi}_{g^\star} := \pi_{g^\star}\}$ (success round).

used in all estimation and planning steps of the algorithm.

LASD terminates whenever the candidate goal $g^\star$ has an optimistic distance larger than L and the new layer is empty, indicating that previous policy evaluation rounds could not identify any good policy and, thus, all states in $\mathcal{S}_L^{\rightarrow}$ have been identified with high probability.

We prove that LASD achieves the following guarantee, the proof can be found in Appendix C.4.

Theorem 1. *Suppose $\mathcal{S}$ is finite. For any $L \geq 1$, $\epsilon \in (0,1]$ and $\delta \in (0,1)$, with probability at least $1-\delta$,* LASD *(Algorithm 1) outputs a set $\mathcal{K}$ such that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{L(1+\epsilon)}^{\rightarrow}$ and $\Pi_{\mathcal{K}}$ such that $V_g^{\pi_g}(s_0) \leq L(1+\epsilon)$ for any $\pi_g \in \Pi_{\mathcal{K}}$, with sample complexity bounded by*

$$\mathcal{O}\left(\frac{S_{L(1+\epsilon)}^{\rightarrow}\Gamma_{L(1+\epsilon)}AL}{\epsilon^2}\iota + \frac{{S_{L(1+\epsilon)}^{\rightarrow}}^2 AL}{\epsilon}\iota + L^3 {S_{L(1+\epsilon)}^{\rightarrow}}^2 A\iota\right)$$

where $\iota = \log^8\left(\frac{SAL}{\epsilon\delta}\right)$.

Compared to the lower bound (see Table 1), LASD still suffers from an extra $\Gamma_{L(1+\epsilon)}$ dependence. This is because in the analysis we use a Bernstein-like concentration inequality to control the deviation $(P - \bar{P})V$, where $\bar{P}$ are the estimated transitions, for any value function V restricted on $\mathcal{K}$ (i.e., V is constant on all states outside $\mathcal{K}$). Unfortunately, we cannot leverage refined concentration inequalities since $\mathcal{K}$ is random and can take an exponentially large amount of values throughout the execution of LASD.

However, by inspecting the proof of (Cai et al., 2022), we note that the construction of the lower bound leverages a certain separation condition defined as follows.

Assumption 2 (identifiability of $\{\mathcal{K}_j^\star\}_j$)**.** *We say $\{\mathcal{K}_j^\star\}_j$ is ϵ-identifiable, if for any $j \geq 2, g \notin \mathcal{K}_j^\star$, we have $V_{\mathcal{K}_{j-1}^\star, g}^\star(s_0) > L(1+\epsilon)$.*

This means that each layer $\mathcal{K}_j^\star$ can be identified exactly by an algorithm run with accuracy ϵ since states that do not belong to the immediate next layer are clearly separated, i.e., they are more than $L(1+\epsilon)$-steps away. This leads to following remark.

Remark 1. *Assumption 2 implies that $\mathcal{S}_L^{\rightarrow} = \mathcal{S}_{L(1+\epsilon)}^{\rightarrow}$.*

How valid is Assumption 2? One might wonder whether Assumption 2 is a realistic and cover many application scenarios. We have identified two large classes of MDPs that satisfies Assumption 2: 1) deterministic MDPs and 2) MDPs with tree structure. Details are deferred to Appendix A.2.

The fact that states $g \notin \mathcal{K}_j^\star$ are not reachable in $L(1+\epsilon)$ steps from $\mathcal{K}_{j-1}^\star$ allows LASD to uniquely identify the layers. Indeed, under Assumption 2, LASD behaves as the operator $\mathcal{T}_L$ and, after each expansion, we have that $\mathcal{K} = \mathcal{K}_j^\star$ for some $j \in [\mathcal{S}_L^{\rightarrow}]$. Thanks to this property, we can show that LASD is minimax optimal.[3]

Theorem 2. *Suppose that $\mathcal{S}$ is finite. For any $L \geq 1$, $\epsilon \in (0, 1]$ and $\delta \in (0, 1)$, if Assumption 2 holds, with probability at least $1-\delta$,* LASD *(Algorithm 1) outputs $\mathcal{K} = \mathcal{S}^{\rightarrow}_{L(1+\epsilon)} = \mathcal{S}^{\rightarrow}_{L}$ and $\Pi_{\mathcal{K}}$ such that $V^{\pi_g}_g(s_0) \leq L(1+\epsilon)$ for any $\pi_g \in \Pi_{\mathcal{K}}$, with sample complexity bounded by*

$$\mathcal{O}\left(\frac{S^{\rightarrow}_L AL}{\epsilon^2}\iota + \frac{S^{\rightarrow 2}_L AL}{\epsilon}\iota + L^3 S^{\rightarrow 2}_L A\iota\right),$$

where $\iota = \log^8\left(\frac{SAL}{\epsilon\delta}\right)$.

The trick to remove the $\Gamma_{L(1+\epsilon)}$ from Theorem 1 is that, since layers are uniquely identified by the algorithm, we only need to concentrate the term $(P - \bar{P})V$ for any value function in the set $\{V_{\mathcal{K}^\star_j}\}_{j \in [S^{\rightarrow}_L]}$.

Given the result above, one might wonder what is the true sample complexity lower bound of this setting. We include a short discussion in Appendix A.3.

Empirical Evaluations We implemented our LASD algorithm and evaluated it empirically. Implementations can be found in `https://github.com/lchenat/AX_exp`. We manually tune the values of some parameters such as $n_{\min}$ and λ to boost the empirical performance, and then conducted experiments on a 4x4 GridWorld environment. The learner has 5 actions in this environment: moving towards one of the four directions by a grid or reset to s_0 (the upper left corner). When the learner takes a directional action, it has probability 0.9 of moving towards the corresponding direction, and 0.1 probability of randomly moving towards one of the four directions. We run LASD on GridWorld with $L = 4$, $\epsilon = 0.01$, and $\delta = 0.001$. We also identify the ground truth set of $\{\mathcal{K}^\star_j\}_j$ by value iterations. Our experiment results show that LASD is able to exactly identify the layers $\{\mathcal{K}^\star_j\}_j$.

3.1. Proof Sketch

Here we report a sketch of the proof, while the detailed one can be found in Appendix C. All the statements we report here are to be considered to hold with high probability.

The first step of the proof (see Lemma 6) is to show by induction that, at each round, $\mathcal{K} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$. Thanks to the fact that $\tilde{\mathcal{O}}(L^2|\mathcal{K}|)$ samples are always available for each $(s, a) \in \mathcal{K} \times \mathcal{A}$ (Line 10) and the properties of VISGO, it is possible to show that, for the goal $g^\star$ selected at the current round, $\|V^{\pi_{g^\star}}_{g^\star}\| \leq 2\|V^{\pi_{g^\star}}_{\mathcal{K},g^\star}\| \leq 4L$ if Line 5 is passed. Combining this with the properties of policy evaluation and the inductive hypothesis, we have that $\hat{\tau} \geq L(1+\epsilon/2) \geq V^{\pi_{g^\star}}_{\mathcal{K},g^\star}(s_0) - L\epsilon/2$ if $g^\star \in \mathcal{U} \setminus \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$. Thus a failure test is triggered and $g^\star$ is never added to $\mathcal{K}$. This shows that states outside $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ are not added to $\mathcal{K}$. By the same reasoning, we can show that if a goal $g^\star$ is added to $\mathcal{K}'$, the corresponding policy has bounded value function (important prerequisite for policy consolidation) and satisfies AX_L. Furthermore, by properly selecting the number of rollouts in the expansion phase (Line 8), we can show that $\mathcal{U}$ always contains at least those states that are reachable in L steps from $\mathcal{K}$ (see Lemma 7), i.e., $\mathcal{T}_L(\mathcal{K}) \setminus \mathcal{K} \subseteq \mathcal{U}$.

Combining these results with optimism restricted on $\mathcal{K}^\star_j$ (see Lemma 8), we are able to show (see Lemma 9) that $\mathcal{K}$ always expands by at least one layer at each update. Formally, if $\mathcal{K}^\star_j \subseteq \mathcal{K}$ at a certain update, then $\mathcal{K} \cup \mathcal{K}' \supseteq \mathcal{K}^\star_{j+1}$ at the next update in Line 7 (i.e., $\mathcal{K}^\star_{j+1} = \mathcal{T}_L(\mathcal{K}^\star_j) \subseteq \mathcal{K}$), see Lemma 23. If Assumption 2 holds, thanks to the identifiability of the layers, we show that $\mathcal{K} = \mathcal{T}_L(\mathcal{K}^\star_j) = \mathcal{K}^\star_{j+1}$, i.e., the algorithm replicates the $\mathcal{T}_L$ operator (see Lemma 25). In this case, $\mathcal{K}'$ is exactly the set of states needed to move from $\mathcal{K}^\star_j$ to $\mathcal{K}^\star_{j+1}$. By induction, we conclude that $\mathcal{S}^{\rightarrow}_L \subseteq \mathcal{K}$ when the algorithm stops, $\mathcal{K} = \mathcal{S}^{\rightarrow}_L$ with Assumption 2.

These results provide AX_L guarantees when the algorithm stops. For computing the sample complexity we use a reduction to a regret analysis of a stochastic shortest path problem (SSP). We define the SSP regret as $R = \sum_{k=1}^K (I_k - V_k(s_0))$ where K is the total number of episodes done in policy evaluation, I_k is the length of episode k, and V_k is the optimistic value function of the goal selected at episode k. Then, $C_K = \sum_{k=1}^K I_k$ is the sample complexity of policy evaluation. Through the SSP regret analysis we can show that $R \lesssim c_1\sqrt{K} + c_2$ and $C_K \lesssim LK$, where $c_1 = L\sqrt{\Gamma_{L(1+\epsilon)} S^{\rightarrow}_{L(1+\epsilon)} A}$ (resp. $c_1 = L\sqrt{S^{\rightarrow}_{L(1+\epsilon)} A}$ under Assumption 2) and $c_2 = L{S^{\rightarrow}_{L(1+\epsilon)}}^2 A$, see Lemma 11 and Lemma 12. To conclude the analysis of the sample complexity we need to bound K. We note that $K = r_{\mathrm{tot}}\lambda \lesssim r_{\mathrm{tot}}/\epsilon^2$ where r_{tot} is the total number of rounds and λ is the maximum number of episodes per round. Moreover, $r_{\mathrm{tot}} \lesssim \frac{c_1^2}{L^2} + \frac{c_2\epsilon}{L}$ can be controlled since the regret is sublinear (see Lemma 14).

In the expansion phases we execute policies that reach any state $s \in \mathcal{K}$ almost surely since, as mentioned above, $\|V^{\pi_s}_s\| \leq 4L$. By (Rosenberg & Mansour, 2021, Lemma 6) we can bound the number of steps required to reach the goal by $8L$. Then, considering the number of samples that needs to be collected and that there are $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)})$ of such phases, the total sample complexity of the expansion phases is $\tilde{\mathcal{O}}(L^3 {S^{\rightarrow}_{L(1+\epsilon)}}^2 A)$. Summing everything together concludes the proof (see Theorem 6).

[3] Minimax optimality holds for $\epsilon \leq \min\{1/S^{\rightarrow}_L, 1/L\}$, which makes the first term in Theorem 2 dominant (Cai et al., 2022).

4. Improved Algorithms

In this section, we present two improvements to LASD that allow to *i)* replace the $\log(S)$ dependence with a much milder $\log(\mathcal{S}^{\rightarrow}_{L(1+\epsilon)})$; *ii)* move from AX_L to AX^+.

4.1. Log-Adaptivity to $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$

Inspired by intrinsically motivated learning agents, Lim & Auer (2012) originally focused on a learning scenario where the environment is possibly infinite or at least no prior knowledge about it is available. Unfortunately, all the existing algorithms fail in dealing with this scenario since they require prior knowledge of the cardinality of the state space $\mathcal{S}$. While the sample complexity only depends logarithmically on S, this shows that inability of the algorithms to exclusively focus on the portion of environment discovered and consolidated over time and it thus prevents from dealing with arbitrarily large or infinite environments.

In this section, we carefully identify all the aspects of the algorithm causing this problem in LASD, and propose an improved algorithm LASD+ (Algorithm 5 in Appendix D) that replaces the $\log(S)$ dependency by $\log(S^{\rightarrow}_{L(1+\epsilon)})$. This is a much favorable dependency since $S^{\rightarrow}_{L(1+\epsilon)}$ is finite even when $\mathcal{S}$ is countably infinite (Lim & Auer, 2012, Prop. 6). Below we list each source of $\log(S)$ dependency and the corresponding modification to fix it.

A) Limiting the set of candidate goals. In the expansion phase, LASD uses all the newly discovered states to build the set $\mathcal{U}$ of candidates states for $\mathcal{S}^{\rightarrow}_L$. This phase could potentially discover any state $s \in \mathcal{S}$ as long as the transition probability to s from $\mathcal{K}$ is non-zero. This means that any $s \in \mathcal{S}$ can be considered in the goal selection step (Line 4), requiring a union bound over $\mathcal{S}$ when analyzing the concentration of the estimated value functions. To overcome this issue, LASD+ performs a step of state filtering in the construction of $\mathcal{U}$ (Algorithm 5-Line 28).[4] The idea is to include in $\mathcal{U}$ only goal states with estimated hitting time upper bounded by L. To break statistical dependencies we estimate the hitting time of each candidate goal state using fresh samples (i.e., samples that are discarded after this step). It can be showed (see Lemma 24) that using this filtering scheme, $\mathcal{U}$ only includes states that are $\mathcal{O}(L)$-controllable by policies restricted on $\mathcal{K}$, which is a much smaller candidate set of order $S^{\rightarrow}_{L(1+\epsilon)}$.

B) Scaling the confidence bounds. While the state filtering step allows to consider only states in $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ rather than $\mathcal{S}$, the knowledge of $S^{\rightarrow}_{L(1+\epsilon)}$ is required to properly set the confidence level when computing the estimated value functions (Algorithm 5-Line 7). We thus maintain an estimate z of $S^{\rightarrow}_{L(1+\epsilon)}$. Each attempt on a specific value of z is a trial indexed by τ (Algorithm 5-Line 2) that ends when the total number of "known" states ($|\mathcal{K} \cup \mathcal{K}'|$) exceeds the estimated dimension z (Algorithm 5-Line 5). In this case, we double the value of z. We can show (see Lemma 16) that the total number of trials is bounded $\tau \lesssim \log_2(S^{\rightarrow}_{L(1+\epsilon)})$ and $z \lesssim S^{\rightarrow}_{L(1+\epsilon)}$.

C) Controlling the policy quality. An important step in LASD is to gather a minimum number of samples for each "known" state (Line 10) to ensure a reasonable performance of the policy being evaluated. The right number of samples also depends on $S^{\rightarrow}_{L(1+\epsilon)}$. Unfortunately, we cannot leverage z to compute this threshold since z is likely to be smaller than $S^{\rightarrow}_{L(1+\epsilon)}$ throughout the execution of the algorithm. Using z will invalidate the properties of policy evaluation that may lead to halt prematurely, without satisfying the AX properties (e.g., $\mathcal{S}^{\rightarrow}_L \subseteq \mathcal{K}$). This failure mode is not captured by the condition used in Algorithm 5-Line 5 to increase z. We thus introduce a Monte-Carlo reachability test (Algorithm 5-Line 12) before policy evaluation. Intuitively, if the test fails LASD+ gathers new samples to improve the estimate of the MDP, otherwise the test guarantees that $\|V^{\pi_{g^\star}}_{g^\star}\|_\infty \lesssim L$ (see Lemma 29).

Combining these three changes, we are able to obtain the following sample complexity guarantee (see Appendix D.1), which is S-independent.

Theorem 3. *For any $L \geq 1$, $\epsilon \in (0,1]$ and $\delta \in (0,1)$, with probability at least $1-\delta$, LASD+ (Algorithm 5) outputs $\mathcal{S}^{\rightarrow}_L \subseteq \mathcal{K} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ and $\Pi_{\mathcal{K}}$ such that $V^{\pi_g}_g(s_0) \leq L(1+\epsilon)$ for any $\pi_g \in \Pi_{\mathcal{K}}$, with sample complexity bounded by*

$$\mathcal{O}\left(\frac{LMA\iota}{\epsilon^2} + \frac{LS^{\rightarrow}_{L(1+\epsilon)}A\iota}{\epsilon} + L^3 {S^{\rightarrow}_{L(1+\epsilon)}}^3 A\iota\right),$$

where $\iota = \log^{12}(\frac{S^{\rightarrow}_{L(1+\epsilon)}AL}{\epsilon\delta})$ and $M = \Gamma_{L(1+\epsilon)}S^{\rightarrow}_{L(1+\epsilon)}$. If Assumption 2 holds, then $M = S^{\rightarrow}_L$ and $S^{\rightarrow}_{L(1+\epsilon)} = S^{\rightarrow}_L$.

4.2. Policy Consolidation

Both LASD and LASD+ discover a set $\mathcal{K}$ such that $\mathcal{S}^{\rightarrow}_L \subseteq \mathcal{K} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ and a set of goal-conditioned policies satisfying AX_L. We now introduce a procedure that, given a set $\mathcal{K} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ and associated goal-reaching policies $\Pi_{\mathcal{K}}$ with bounded value function, learns a set of goal-condition policies satisfying the AX^+ condition.

POLICYCONSOLIDATION (Algorithm 2) is an algorithm for Multi-Goal Exploration (MGE) (e.g., Tarbouriech et al., 2022) over $\mathcal{K}$. In each round, POLICYCONSOLIDATION randomly selects an "unknown" goal state from $\mathcal{L}$ and computes a policy to reach it (Line 6). It then evaluates the performance of this policy by $\tilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ rollouts, and based

[4] A similar filter is used in DISCO to reduce computational complexity, but as it does not use fresh samples, it still requires a union bound over $\mathcal{S}$ to deal with statistical dependencies.

Algorithm 2: Policy Consolidation (PC)

Input: $L \geq 1$, $\epsilon \in (0, 1]$, $\delta \in (0, 1)$, target state space $\mathcal{K} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, and initial policies $\Pi' = \{\pi'_g\}_{g \in \mathcal{K}}$.

1 Set $k \leftarrow 1$, $\mathfrak{N} = \{2^j\}_{j \geq 0}$, $\mathcal{L} = \mathcal{K}$, $\Pi^+_{\mathcal{K}} = \{\widetilde{\pi}_{s_0}$ a random policy$\}$, $\mathbf{N}(\cdot, \cdot), \mathbf{N}(\cdot, \cdot, \cdot) \leftarrow 0$.

2 $(\mathbf{N}, \_) \leftarrow$ EXPLORE$(\mathcal{K}, \Pi', \mathbf{N}, N_1(|\mathcal{K}| - 1, \frac{\delta}{|\mathcal{K}|}))$ (see Algorithm 6; $N_1 \lesssim L^2 |\mathcal{K}| \log(\frac{|\mathcal{K}|}{\delta})$ is defined in Lemma 4).

3 **for** $r = 1, \ldots$ **do**

4   **if** $\mathcal{L} = \varnothing$ **then return** $\Pi^+_{\mathcal{K}}$.

5     $\epsilon_{\mathrm{VI}} \leftarrow 1 / \max\{16, \sum_{s,a} \mathbf{N}(s, a)\}$.

6   Pick $g^\star \in \mathcal{L}$ arbitrarily and compute $(\widehat{Q}, \widehat{V}, \widehat{\pi}) = \mathrm{VISGO}(\mathcal{K} \setminus \{g\}, g, \epsilon_{\mathrm{VI}}, \mathbf{N}, \frac{\delta}{|\mathcal{K}|})$.

7   Let $\lambda \leftarrow N_{\mathrm{DEV}}(32L, \frac{\epsilon}{256}, \frac{\delta}{2r^2}) \lesssim \frac{1}{\epsilon^2} \log^4\left(\frac{Lr}{\epsilon\delta}\right)$ (defined in Lemma 50) and $\widehat{\tau} \leftarrow 0$.

8   **for** $j = 1, \ldots, \lambda$ **do**

9         $k \xleftarrow{+} 1$, $i \leftarrow 1$, and reset to $s^k_1 \leftarrow s_0$ by taking action RESET.

10   **while** $s^k_i \neq g^\star$ **do**

11   Take $a^k_i = \widehat{\pi}(s^k_i)$, and transits to s^k_{i+1}.

12   Increase $\mathbf{N}(s^k_i, a^k_i)$, $\mathbf{N}(s^k_i, a^k_i, s^k_{i+1})$, and i by 1.

13   **if** $\sum_{s,a} \mathbf{N}(s, a) \in \mathfrak{N}$ *or* ($s^k_i \in \mathcal{K}$ *and* $\mathbf{N}(s^k_i, a^k_i) \in \mathfrak{N}$) **then** return to Line 3 (skip round).

14   Set $\widehat{\tau} \xleftarrow{+} \frac{c(s^k_i, a^k_i)}{\lambda}$.

15   **if** $\widehat{\tau} > \widehat{V}(s_0)(1 + \epsilon/2)$ **then** return to Line 3 (failure round).

16     $\mathcal{L} \leftarrow \mathcal{L} \setminus \{g^\star\}$, $\Pi^+_{\mathcal{K}} \leftarrow \Pi^+_{\mathcal{K}} \cup \{\widetilde{\pi}_{g^\star} = \widehat{\pi}\}$ (success round).

on the evaluation result, the current round is classified into success, skip, or failure round similar to that in Algorithm 1. While it shares a similar structure with VALAE, the crucial difference is the condition of success round (Line 3), which has a form similar to AX$^+$. Thus, one can consider Algorithm 2 as an improved version of VALAE.

Its simplicity and high sample efficiency, allow POLICYCONSOLIDATION to be integrated with any existing algorithm for AX$_L$ or AX* at no cost. As showed in the following lemma, the sample complexity of policy consolidation matches the lower-bound for AX, thus providing a "minor" contribution to the overall sample complexity. Details are deferred to Appendix E.

Theorem 4. *Given a target state space $\mathcal{K} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ for some $\epsilon \in (0, 1)$ and a set of initial policies $\Pi' = \{\pi'_g\}_{g \in \mathcal{K}}$ such that $\left\|V_g^{\pi'_g}\right\|_\infty \lesssim L$, with probability at least $1 - \delta$,* POLICYCONSOLIDATION *(Algorithm 2) outputs a set of policies $\{\widetilde{\pi}_g\}_{g \in \mathcal{K}}$ such that $V_g^{\widetilde{\pi}_g}(s_0) \leq V^\star_{\mathcal{K},g}(s_0)(1 + \epsilon)$ for all $g \in \mathcal{K}$, with sample complexity bounded by*

$$\tilde{\mathcal{O}}\left(\frac{LS^{\rightarrow}_{L(1+\epsilon)}A\iota}{\epsilon^2} + \frac{L{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota}{\epsilon} + L^3{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota\right),$$

where $\iota = \log^{10}(\frac{S^{\rightarrow}_{L(1+\epsilon)}AL}{\epsilon\delta})$.

To achieve this result we developed an improved regret-based analysis. Instead of bounding the total number of rounds as in VALAE, we directly bound the total number of steps in all rounds, which takes varying length of trajectories in different rounds into consideration. This enables POLICYCONSOLIDATION to achieve a better guarantee on the performance of the learned policies compared to VALAE, preserving the same sample complexity.

4.3. AX^+ through Layer Discovery and Consolidation

We combine all these improvement into Layered Autonomous Exploration (LAE) whose pseudo code is reported in Algorithm 3. Combining the previous results, we can state the following guarantee for AX$^+$.

Corollary 5. *For any $L \geq 1$, $\epsilon \in (0, 1]$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$,* LAE *(Algorithm 3) outputs $\mathcal{S}^{\rightarrow}_L \subseteq \mathcal{K} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ and $\Pi_{\mathcal{K}}$ such that $V_g^{\pi_g}(s_0) \leq V^\star_{\mathcal{K},g}(s_0)(1+\epsilon)$, for any $\pi_g \in \Pi_{\mathcal{K}}$, with sample complexity*

$$\mathcal{O}\left(\frac{LMA\iota}{\epsilon^2} + \frac{LS^{\rightarrow}_{L(1+\epsilon)}A\iota}{\epsilon} + L^3{S^{\rightarrow}_{L(1+\epsilon)}}^3A\iota\right)$$

where $\iota = \log^{12}\left(\frac{S^{\rightarrow}_{L(1+\epsilon)}AL}{\epsilon\delta}\right)$ and $M = \Gamma_{L(1+\epsilon)}S^{\rightarrow}_{L(1+\epsilon)}$. If Assumption 2 holds, then $M = S^{\rightarrow}_L$ and $S^{\rightarrow}_{L(1+\epsilon)} = S^{\rightarrow}_L$.

This shows that LAE is the first algorithm able to i) achieve the strongest performance AX$^+ \Rightarrow$ AX$^\star \Rightarrow$ AX$_L$, ii) match the lower-bound under certain settings, and iii) completely remove the dependence on S. In particular, the latter was an open problem since the initial work by Lim & Auer (2012).[5]

Comparisons. LASD/LASD$^+$ shares similarities with both UCBEXPLORE and VALAE. While we leverage the same condition as in VALAE for the failure test of policy evaluation, the policy evaluation in VALAE is only for learning goal-conditioned policies and not for consolidating states. In fact, they first run DISCO for state discovery, and then learn goal-conditioned policies on a potentially much larger set subsuming $\mathcal{S}^{\rightarrow}_{2L}$. However, $\mathcal{S}^{\rightarrow}_{2L}$ can be exponentially larger than $S^{\rightarrow}_{L(1+\epsilon)}$ (see Lemma 43) in general and thus the sample complexity of VALAE is incomparable to other algorithms. Therefore, VALAE only improves the sample complexity of policy learning but not that of state discovery. Similarly to UCBEXPLORE, we perform state and policy identification simultaneously. Our evaluation phase is much more sample efficient compared to UCBEXPLORE, which saves a L^2/ϵ factor in the leading-order term. Compared to DISCO, our algorithm saves a L^2 factor by i) adaptively collecting samples to estimate state values instead of prescribing a fixed number of samples to guarantee a uniformly-accurate

[5] UCBEXPLORE originally considered a countable, possibly infinite state space; however this leads to a technical issue in the analysis (Tarbouriech et al., 2020b, Footnote 2).

Algorithm 3: Layered Autonomous Exploration (LAE)

Input: $L \geq 1$, $\epsilon \in (0, 1]$, and $\delta \in (0, 1)$.
$(\mathcal{K}, \Pi_{\mathcal{K}}^{L}) = \mathrm{LASD}^{+}\big(L, \epsilon, \delta\big)$ see Algorithm 5 in appendix (or LASD for $\log S$). `// AX`$_L$
$\Pi_{\mathcal{K}}^{+} = \mathrm{PC}\big(L, \epsilon, \delta, \mathcal{K}, \Pi_{\mathcal{K}}^{L}\big)$. `// AX`$^{+}$
return $\mathcal{K}$ *and* $\Pi_{\mathcal{K}}^{+}$.

transition estimate over $\mathcal{K}$, and ii) leveraging variance information.

The tool enabling all these improvements is a new Bernstein-type concentration inequality for restricted value functions (see Lemma 46). The key difficulty in our analysis is that the set on which value functions are restricted is random since we learn $\mathcal{K}$ and $\Pi_{\mathcal{K}}$ simultaneously. In comparison, in VALAE the set $\mathcal{K}$ is fixed after the initial phase of state discovery, which makes the analysis much simpler. Specifically, leveraging the fact that the learned goal-conditioned policies are all restricted on $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, we are able to make use of the variance information without incurring a polynomial dependency on S.

5. Conclusion

We introduced a layered decomposition of the set of incrementally L-controllable states. We built on this decomposition and showed that our algorithm LAE attains the strongest performance guarantee AX^{+}, does not need to know S and thus can be used with a countably-infinite state space, and is minimax-optimal when the layers can be uniquely identified. The natural future directions include 1) designing an algorithm with minimax sample complexity without Assumption 2; 2) extending the problem to continuous states and function approximation; 3) identifying benchmarks that can be used to evaluate practical progresses towards the AX capability.

References

Bagaria, A., Senthil, J. K., and Konidaris, G. Skill discovery for exploration and planning using deep skill graphs. In *International Conference on Machine Learning*, pp. 521–531. PMLR, 2021.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

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

Cai, H., Ma, T., and Du, S. S. Near-optimal algorithms for autonomous exploration and multi-goal stochastic shortest path. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2434–2456. PMLR, 2022.

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

Chen, L. and Luo, H. Near-optimal goal-oriented reinforcement learning in non-stationary environments. *arXiv preprint arXiv:2205.13044*, 2022.

Chen, L., Jafarnia-Jahromi, M., Jain, R., and Luo, H. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 2021.

Chen, L., Jain, R., and Luo, H. Improved no-regret algorithms for stochastic shortest path with linear MDP. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3204–3245. PMLR, 2022a.

Chen, L., Luo, H., and Rosenberg, A. Policy optimization for stochastic shortest path. In *COLT*, volume 178 of *Proceedings of Machine Learning Research*, pp. 982–1046. PMLR, 2022b.

Chen, L., Tirinzoni, A., Pirotta, M., and Lazaric, A. Reaching goals is hard: Settling the sample complexity of the stochastic shortest path. In *International Conference on Algorithmic Learning Theory*, 2023.

Cohen, A., Kaplan, H., Mansour, Y., and Rosenberg, A. Near-optimal regret bounds for stochastic shortest path. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 8210–8219. PMLR, 2020.

Colas, C., Karch, T., Sigaud, O., and Oudeyer, P. Intrinsically motivated goal-conditioned reinforcement learning: a short survey. *CoRR*, abs/2012.09830, 2020.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. In *The International Conference on Learning Representations*, 2019.

Gregor, K., Rezende, D. J., and Wierstra, D. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pp. 2681–2691, 2019.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020.

Kamienny, P., Tarbouriech, J., Lamprier, S., Lazaric, A., and Denoyer, L. Direct then diffuse: Incremental unsupervised skill discovery for state covering and goal reaching. In *ICLR*. OpenReview.net, 2022.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021.

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

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

Oudeyer, P.-Y., Baranes, A., and Kaplan, F. *Intrinsically Motivated Exploration for Developmental and Active Sensorimotor Learning*, volume 264, pp. 107–146. 12 2009. ISBN 978-3-642-05180-7. doi: 10.1007/978-3-642-05181-4_6.

Pong, V., Dalal, M., Lin, S., Nair, A., Bahl, S., and Levine, S. Skew-fit: State-covering self-supervised reinforcement learning. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7783–7792. PMLR, 2020.

Rosenberg, A. and Mansour, Y. Stochastic shortest path with adversarially changing costs. In *IJCAI*, pp. 2936–2942. ijcai.org, 2021.

Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In Meyer, J. A. and Wilson, S. W. (eds.), *Proc. of the*

International Conference on Simulation of Adaptive Behavior: From Animals to Animats, pp. 222–227. MIT Press/Bradford Books, 1991.

Singh, S., Barto, A. G., and Chentanez, N. Intrinsically motivated reinforcement learning. In *NIPS*, pp. 1281–1288, 2004.

Tarbouriech, J., Garcelon, E., Valko, M., Pirotta, M., and Lazaric, A. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pp. 9428–9437. PMLR, 2020a.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. Improved sample complexity for incremental autonomous exploration in MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11273–11284. Curran Associates, Inc., 2020b.

Tarbouriech, J., Shekhar, S., Pirotta, M., Ghavamzadeh, M., and Lazaric, A. Active model estimation in markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1019–1028. PMLR, 2020c.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. A provably efficient sample collection strategy for reinforcement learning. In *NeurIPS*, pp. 7611–7624, 2021a.

Tarbouriech, J., Zhou, R., Du, S. S., Pirotta, M., Valko, M., and Lazaric, A. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. In *NeurIPS*, pp. 6843–6855, 2021b.

Tarbouriech, J., Domingues, O. D., Ménard, P., Pirotta, M., Valko, M., and Lazaric, A. Adaptive multi-goal exploration. In *International Conference on Artificial Intelligence and Statistics*, pp. 7349–7383. PMLR, 2022.

Zhang, Z., Du, S., and Ji, X. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, pp. 12402–12412. PMLR, 2021.

Contents

A. Preliminaries

A.1. Notation

Let $(x)_+ = \max\{0, x\}$ and $\mathbb{I}_s(s') = \mathbb{I}\{s' = s\}$. We say that a value function V is **restricted** on a subset $\mathcal{X} \subseteq \mathcal{S}$, if there exists $v > 0$ such that $V(s) = v$ for any $s \notin \mathcal{X}$. When value function V takes the same value within a subset of states y, we define $V(y) = V(s)$ for any $s \in y$. For any subset $y \subseteq \mathcal{S}$ and distribution $P \in \Delta_{\mathcal{S}}$, define $P(y) = \sum_{s' \in y} P(s')$.

Trial In Algorithm 5, a trial is indexed by τ, and each trial corresponds to a value of z estimating $S^{\rightarrow}_{L(1+\epsilon)}$ (Line 1). In Algorithm 1 and Algorithm 2, we assume the whole learning procedure lies in an artificial trial.

Table 2. The notation adopted in this paper.

Symbol	Meaning
$\mathcal{S}$	State Space
$\mathcal{A}$	Action Space (including the RESETaction)
P	Transition function
$\pi : \mathcal{S} \rightarrow \mathcal{A}$	A policy
$\Pi(\mathcal{X})$	Policies restricted to $\mathcal{X}$, RESET is taken outside $\mathcal{X}$
L	Exploration radius
$\mathcal{S}^{\rightarrow}_L$	Incrementally L-controllable states
$\mathcal{N}^{s,a}_L = \{s' \in \mathcal{S}^{\rightarrow}_L : P_{s,a}(s') > 0\}$	States in $\mathcal{S}^{\rightarrow}_L$ reachable from (s, a)
$\Gamma^{s,a}_L = \lvert\mathcal{N}^{s,a}_L\rvert, \Gamma_L = \max_{s \in \mathcal{S}^{\rightarrow}_L, a} \Gamma^{s,a}_L$	Cardinality of $\mathcal{N}^{s,a}_L$ and maximum value
$\mathcal{T}_L(\mathcal{X}) = \{g \in \mathcal{S} : V^\star_{\mathcal{X},g}(s_0) \leq L\}$	Set of L controllable states restricted on $\mathcal{X} \subseteq \mathcal{S}$
$\{\mathcal{K}^\star_j\}_j : \mathcal{K}^\star_1 = \{s_0\}, \mathcal{K}^\star_j = \mathcal{T}_L(\mathcal{K}^\star_{j-1})$	Layers defining $\mathcal{S}^{\rightarrow}_L$
$\mathcal{O}^{\rightarrow}_L = (s_1, \ldots, s_n)$	Ordering of states in $\mathcal{S}^{\rightarrow}_L$ defining the layer $\{\mathcal{K}^\star_j\}$
$\mathcal{K}^\star_{z,j}$	$\mathcal{K}^\star_{z,j} = \mathcal{K}^\star_j$ when $\lvert\mathcal{K}^\star_j\rvert < z$, and $\mathcal{K}^\star_{z,j} = \{s_1, \ldots, s_z\}$ when $\lvert\mathcal{K}^\star_j\rvert \geq z$
$\mathcal{K}^\star_{z,z} = (s_1, \ldots, s_z)$	The first z elements of $\mathcal{O}^{\rightarrow}_L$ or $\mathcal{S}^{\rightarrow}_L$
$\mathcal{U}^\star_z = \mathcal{T}_{2L}(\mathcal{K}^\star_{z,z})$	States reachable in $2L$ steps from $\mathcal{K}^\star_{z,z}$
$\mathcal{N}(\mathcal{X}, p) = \{s' \notin \mathcal{X} : P(s' \mid s, a) \geq p \text{ for some } (s, a) \in \mathcal{X} \times \mathcal{A}\}$	States not in $\mathcal{X}$ reachable with high probability from $\mathcal{X}$
$\bar{\mathcal{U}} = \{s' \in \mathcal{S} : \exists s \in \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}, a \in \mathcal{A}, P(s' \mid s, a) \geq \frac{1}{2L}\}$	States that are reachable from $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ with high probability
Learning Algorithm	
$r \in \mathbb{N}_+$	Round
$\tau \in \mathbb{N}_+$	Trial
z	An estimate of $\lvert S^{\rightarrow}_{L(1+\epsilon)}\rvert$. The value of z is updated at the beginning of each trial.
ϵ	accuracy
$\mathcal{K}$	Set of "known" states, such that $\mathcal{K}^\star_j \subseteq \mathcal{K}$ for some j
$\mathcal{U}$	Set of "unknown" states
$\mathcal{K}'$	Increment to $\mathcal{K}$ leading to include layer $j + 1$
$\mathbf{N}(s, a, s')$	Number of visits to (s, a, s')
λ	Number of episodes for policy evaluation
$\hat{\tau}$	Average number of steps to reach the goal by policy $\pi_{g^\star}$

A.2. How Valid is Assumption 2?

We have identified two large classes of MDPs that satisfy Assumption 2: 1) deterministic MDP. It is clear that when transition is deterministic, we have $\mathcal{K}^\star_j = \{s \in \mathcal{S} : d(s_0, s) = j - 1\}$, where $d(s_0, s)$ is the distance of shortest path from s_0 to s. Moreover, states not in $\mathcal{K}^\star_j$ are unreachable by any policy restricted on $\mathcal{K}^\star_{j-1}$ (any path from s_0 to a state s with $d(s_0, s) = j + 1$ must pass through a state s' with $d(s_0, s') = j$), thus satisfying Assumption 2. 2) MDPs with tree structure, that is, states in the MDP are nodes in a tree; nodes (states) s and s' has an edge if and only if there exist $a \neq$ RESET s.t. $P(s' \mid s, a) > 0$ or $P(s \mid s', a) > 0$. With s_0 being the root of the tree, we have $\mathcal{K}^\star_j \subseteq \mathcal{D}_j$, where $\mathcal{D}_j = \{d'(s_0, s) \leq j - 1\}$ and $d'(s, s')$ is the undirected distance on tree from s to s'. Clearly, this implies $V_{\mathcal{K}^\star_{j-1}, g}(s_0) = \infty$ for any $g \notin \mathcal{K}^\star_j$, satisfying Assumption 2.

A.3. Thoughts on the Lower Bound

We believe that the lower bound should either scale with $S^{\rightarrow}_{L(1+\epsilon)}$ or $S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}$. However, verifying both cases requires brand new ideas. If the lower bound indeed scales with $S^{\rightarrow}_{L(1+\epsilon)}$ in general, then there is room for improvement for existing algorithms and analysis. Unfortunately, due to the exponentially large amount of possible values of $\mathcal{K}$, the standard UCBVI style analysis does not help to remove the $\Gamma_{L(1+\epsilon)}$ dependency. On the other hand, if the lower bound scales with

$S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}$ when Assumption 2 does not hold, then we need to show that having undistinguishable states (states in $S^{\rightarrow}_{L(1+\epsilon)} \setminus S^{\rightarrow}_{L}$) actually worsen the sample complexity. This cannot be handled by the usual lower bound construction and analysis, which counts the number of samples needed to distinguish states in $\mathcal{S}^{\rightarrow}_{L}$ and out of $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$.

Algorithm 4: VISGO

Input: state subset $\mathcal{X}$, goal state $g \notin \mathcal{X}$, precision ϵ_{VI}, counter n, and failure probability δ.
Require: $\left\|V^\star_{\mathcal{X},g}\right\|_\infty \leq 8L$.
Let $c_1 = 3$, $c_2 = 512$, and $\iota_{s,a} = \log\left(\frac{2|\mathcal{X}|An(s,a)}{\delta}\right)$ for all (s,a).
Let $\bar{P}_{s,a}(s') = \frac{n(s,a,s')}{n^+(s,a)}$ and $\widetilde{P}_{s,a}(s') = \frac{n(s,a)}{n(s,a)+1}\bar{P}_{s,a}(s') + \frac{\mathbb{I}\{s'=g\}}{n(s,a)+1}$ for all (s,a,s').
Initialize: $V^{(0)}(\cdot) \leftarrow 0$, $i \leftarrow 0$.
while $i = 0$ *or* $\left\|V^{(i)} - V^{(i-1)}\right\|_\infty > \epsilon_{VI}$ **do**
  if $\left\|V^{(i)}\right\|_\infty > 2L$ **then return** (∞, ∞, π) with π being a random policy.
  $i \leftarrow i + 1$.
  for $s \in \mathcal{X}$ **do**
    $b^{(i)}(s,a) \leftarrow \max\left\{c_1\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^{(i-1)})\iota_{s,a}}{n^+(s,a)}}, \frac{c_2 L\iota_{s,a}}{n^+(s,a)}\right\}$.
    $Q^{(i)}(s,a) \leftarrow \max\left\{0, 1 + \widetilde{P}_{s,a}V^{(i-1)} - b^{(i)}(s,a)\right\}$ for $a \in \mathcal{A}$.
    $V^{(i)}(s) \leftarrow \min_a Q^{(i)}(s,a)$
  $V^{(i)}(s) \leftarrow (1 + V^{(i-1)}(s_0))\mathbb{I}\{s \neq g\}$ for $s \notin \mathcal{X}$.
return $(Q^{(i)}, V^{(i)}, \pi)$ *with* $\pi(s) = \operatorname{argmin}_a Q^{(i)}(s,a)$ *for* $s \in \mathcal{X}$ *and* $\pi_g(s) = \text{RESET}$ *for* $s \notin \mathcal{X}$.

B. Analysis of VISGO

The convergence of VISGO has been proved in (Cai et al., 2022, Lemma C.4). We further introduce some properties of the algorithm.

Lemma 2 (Optimism)**.** *Let $\mathcal{X} \subseteq \mathcal{S}$, $g \in \mathcal{S} \setminus \mathcal{X}$, n be a counter incrementally collecting samples from transition function P, and $\delta \in (0,1)$ be such that $\|V^\star_{\mathcal{X},g}\|_\infty \leq 8L$. For any precision $\xi > 0$, define $(Q_\xi, V_\xi, \_) = \text{VISGO}(\mathcal{X}, g, \xi, n, \delta)$ as the output of Algorithm 4. Let $\mathbb{P}$ be the probability operator on the process generating the counter n and assume that $\mathcal{X}$ and g are independent of n. Then,*

$$\mathbb{P}\Big(\forall \xi > 0, s \in \mathcal{S}, a \in \mathcal{A} : Q_\xi(s,a) \leq Q^\star_{\mathcal{X},g}(s,a), V_\xi(s) \leq V^\star_{\mathcal{X},g}(s)\Big) \geq 1 - \delta.$$

Proof. First, by Lemma 54 and a union bound over $(s,a) \in \mathcal{X} \times \mathcal{A}$, we have with probability at least $1-\delta$, for any $(s,a) \in \mathcal{X} \times \mathcal{A}$,

$$\begin{aligned}
\left|(\bar{P}_{s,a} - P_{s,a})V^\star_{\mathcal{X},g}\right| &\leq 2\sqrt{\frac{2\mathbb{V}(\bar{P}_{s,a}, V^\star_{\mathcal{K},g})\log\frac{2|\mathcal{X}|An(s,a)}{\delta}}{n^+(s,a)}} + \frac{19 \cdot 8L\log\frac{2|\mathcal{X}|An(s,a)}{\delta}}{n^+(s,a)} \\
&\leq \frac{c_1}{2}\sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^\star_{\mathcal{X},g})\iota_{s,a}}{n^+(s,a)}} + \frac{c_2 L\iota_{s,a}}{2n^+(s,a)},
\end{aligned} \tag{1}$$

with $\iota_{s,a}$, c_1, and c_2 are defined in Algorithm 4. We then carry out the proof assuming that such event holds.

Fix a configuration $(\mathcal{X}, g, \xi, n, \delta)$ of the inputs of VISGO and let $(Q^{(i)}, V^{(i)})_{i\geq 0}$ be the iterates of the algorithm. It suffices to show that for any $i \geq 0$, $Q^{(i)}(s,a) \leq Q^\star_{\mathcal{X},g}(s,a)$ for all $(s,a) \in \mathcal{X} \times \mathcal{A}$ and $V^{(i)}(s) \leq V^\star_{\mathcal{X},g}(s)$ for all $s \in \mathcal{S}$. We prove it by induction.

Note that $Q^{(0)}(\cdot) = V^{(0)}(\cdot) = 0$, thus the statement clearly holds for the base case $i = 0$. Suppose it holds at some iteration

$i - 1 \geq 0$. Under event of Eq. (1), for any $i > 0$ and $(s, a) \in \mathcal{X} \times \mathcal{A}$,

$$
\begin{aligned}
&1 + \widetilde{P}_{s,a} V^{(i-1)} - \max\left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^{(i-1)})\iota_{s,a}}{n^+(s,a)}}, \frac{c_2 L \iota_{s,a}}{n^+(s,a)} \right\} \\
&\leq 1 + \widetilde{P}_{s,a} V^\star_{\mathcal{X},g} - \max\left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^\star_{\mathcal{X},g})\iota_{s,a}}{n^+(s,a)}}, \frac{c_2 L \iota_{s,a}}{n^+(s,a)} \right\} && \text{(induction step and Lemma 49)} \\
&\leq 1 + \bar{P}_{s,a} V^\star_{\mathcal{X},g} - \max\left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^\star_{\mathcal{X},g})\iota_{s,a}}{n^+(s,a)}}, \frac{c_2 L \iota_{s,a}}{n^+(s,a)} \right\} && \text{(definition of } \widetilde{P}_{s,a}) \\
&\leq 1 + P_{s,a} V^\star_{\mathcal{X},g} + (\bar{P}_{s,a} - P_{s,a}) V^\star_{\mathcal{X},g} - \frac{c_1}{2} \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^\star_{\mathcal{X},g})\iota_{s,a}}{n^+(s,a)}} - \frac{c_2 L \iota_{s,a}}{2n^+(s,a)} && (\max\{a,b\} \geq \tfrac{a+b}{2}) \\
&\leq Q^\star_{\mathcal{X},g}(s,a). && \text{(Eq. (1))}
\end{aligned}
$$

This also proves that $V^{(i)}(s) \leq V^\star_{\mathcal{X},g}(s)$ for all $s \in \mathcal{X}$. Moreover, for $s \notin \mathcal{X}, s \neq g$, $V^{(i)}(s) = 1 + V^{(i-1)}(s_0) \leq 1 + V^\star_{\mathcal{X},g}(s_0) = V^\star_{\mathcal{X},g}(s)$. Finally, $V^{(i)}(g) = V^\star_{\mathcal{X},g}(g) = 0$. This proves that $V^{(i)}(s) \leq V^\star_{\mathcal{K},g}(s)$ for all $s \in \mathcal{S}$, thus concluding the proof. $\square$

Lemma 3 (Bounded Error). *There exists a function $N_0(z_0, z_0', \delta_0, \delta) \lesssim L^2 z_0 \log \frac{z_0'}{\delta_0 \delta}$ such that, for goal set $\mathcal{G}$ with $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \subseteq \mathcal{G} \subseteq \mathcal{S}$ and $\delta_0 \in (0,1)$, with probability at least $1 - \delta$ over the randomness of a counter n incrementally collecting samples from transition function P, for any $\mathcal{X} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ with $|\mathcal{X}| \leq z_0$, $g \in \mathcal{G} \setminus \mathcal{X}$, precision $\xi \in (0, \frac{1}{8})$, and $\delta' \in [\delta_0, 1)$, if $z_0' \geq |\mathcal{G}|$ and $n(s,a) \geq N_0(z_0, z_0', \delta_0, \delta)$ for all $(s,a) \in \mathcal{X} \times \mathcal{A}$, then $V_g^{\pi_g}(s) \leq 2V(s)$ for all $s \in \mathcal{S}$, where $(\_, V, \pi_g) = \mathrm{VISGO}(\mathcal{X}, g, \xi, n, \delta')$ is the output of Algorithm 4. Also define $N_0(z_0, \delta) = N_0(z_0, S, \delta, \delta)$ and $N_0^{\rightarrow}(\delta) = N_0(S^{\rightarrow}_{L(1+\epsilon)}, |\bar{\mathcal{U}}|, \delta, \delta)$ (recall that $|\bar{\mathcal{U}}| \leq 2LAS^{\rightarrow}_{L(1+\epsilon)}$).*

Proof. Note that the statement clearly holds if VISGO returns a value function $V = \infty$. Otherwise, $\left\|V^{(i)}\right\|_\infty \leq 2L$ for any $i \leq l$, where l is the index of the last iteration in Algorithm 4. By Lemma 46, with probability at least $1 - \delta$[6], for any status of n, $(s,a) \in \mathcal{X} \times \mathcal{A}$, and V s.t. $\|V\|_\infty \leq 2L$,

$$
\begin{aligned}
\left|(P_{s,a} - \widetilde{P}_{s,a})V\right| &\leq \left|(P_{s,a} - \bar{P}_{s,a})V\right| + \left|(\bar{P}_{s,a} - \widetilde{P}_{s,a})V\right| \\
&\lesssim L\sqrt{\frac{z_0 \iota'}{n(s,a)}} + \frac{L z_0 \iota'}{n(s,a)} + \frac{(\bar{P}_{s,a} + \mathbb{I}_g)V}{n(s,a)+1},
\end{aligned}
$$

where $\widetilde{P}_{s,a}$ and $\bar{P}_{s,a}$ are as defined in Algorithm 4 with counter n and $\iota' = \tilde{\mathcal{O}}(\log \frac{z_0'}{\delta})$ by $|\mathcal{G}| \leq z_0'$. Clearly, there exists $n_1 = \tilde{\mathcal{O}}(L^2 z_0 \log(|\mathcal{G}|/\delta))$, such that when $n(s,a) \geq n_1$, we have $|(P_{s,a} - \widetilde{P}_{s,a})V| \leq \frac{1}{8}$. Moreover, we have

$$
b^{(l)}(s,a) \lesssim \max\left\{ \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^{(l-1)})}{n(s,a)}}, \frac{L}{n(s,a)} \right\} \lesssim \frac{L}{\sqrt{n(s,a)}}.
$$

Then there exist $n_2 = \tilde{\mathcal{O}}(L^2 \log(1/\delta_0))$ such that when $n(s,a) \geq n_2$, $b^{(l)}(s,a) \leq \frac{1}{8}$. Thus when $n(s,a) \geq \max\{n_1, n_2\}$ for all $s \in \mathcal{X}, a \in \mathcal{A}$, we can apply the same conclusion as in the proof of Lemma 4 as get the desired result. $\square$

Lemma 4 (Bounded Error with Fresh Samples). *There exists a function $N_1(x, \delta_0, \delta) \lesssim L^2 x \log \frac{x}{\delta_0 \delta}$ (also define $N_1(x, \delta) = N_1(x, \delta, \delta)$) such that for $\mathcal{X} \subseteq \mathcal{S}$, $g \in \mathcal{S} \setminus \mathcal{X}$, $\delta_0 \in (0,1)$, $\delta \in (0,1)$, n a counter incrementally collecting samples from transition function P, and assume that $\mathcal{X}, g, \delta_0$ are independent of n, with probability at least $1 - \delta$, for any precision $\xi \in (0, \frac{1}{8})$ and $\delta' \in [\delta_0, 1)$, if $n(s,a) \geq N_1(|\mathcal{X}|, \delta_0, \delta)$ for all $(s,a) \in \mathcal{X} \times \mathcal{A}$, then $V_g^{\pi_g}(s) \leq 2V(s)$ for all $s \in \mathcal{S}$, where $(\_, V, \pi_g) = \mathrm{VISGO}(\mathcal{X}, g, \xi, n, \delta')$ is the output of Algorithm 4.*

[6] this holds under the same good event of Lemma 46, which does not depend on the chosen $\mathcal{X}, g, \delta', \xi$

Proof. Let $y = \mathcal{S} \setminus (\mathcal{X} \cup \{g\})$ and $\iota_{s,a}^n = \log \frac{4|\mathcal{X}|^2 A n(s,a)}{\delta}$. Consider the following events:

$$E_1 := \left\{ \forall s \in \mathcal{X}, a \in \mathcal{A}, s' \in \mathcal{X}, n(s,a) \geq 1 : |P_{s,a}(s') - \bar{P}_{s,a}(s')| \leq 2\sqrt{\frac{2P_{s,a}(s')\iota_{s,a}^n}{n(s,a)}} + \frac{2\iota_{s,a}^n}{n(s,a)} \right\},$$

$$E_2 := \left\{ \forall s \in \mathcal{X}, a \in \mathcal{A}, n(s,a) \geq 1 : |P_{s,a}(y) - \bar{P}_{s,a}(y)| \leq 2\sqrt{\frac{2P_{s,a}(y)\iota_{s,a}^n}{n(s,a)}} + \frac{2\iota_{s,a}^n}{n(s,a)} \right\}.$$

By Lemma 54 and a union bound, they hold simultaneously with probability at least $1-\delta$. We carry out the proof conditioned on these events holding.

For any $\mathcal{X}, g, \xi, n, \delta'$, the statement clearly holds if $V = \infty$. Otherwise, $\left\|V^{(i)}\right\|_\infty \leq 2L$ for any $i \leq l$, where l is the index of the last iteration in Algorithm 4. Take any status of counter n, precision $\xi \in (0, \frac{1}{8})$, $\delta' \in [\delta_0, 1)$. Let V and π_g be the output of Algorithm 4 with these parameters such that $\|V\|_\infty \leq 2L$. Since V is restricted on $\mathcal{X} \cup \{g\}$, we have $V(s') = 1 + V^{(l-1)}(s_0)$ for any $s' \notin \mathcal{X} \cup \{g\}$. Then, for any $(s,a) \in \mathcal{X} \times \mathcal{A}$,

$$\begin{aligned}
&\left|(P_{s,a} - \widetilde{P}_{s,a})V\right| \leq \left|(P_{s,a} - \bar{P}_{s,a})V\right| + \left|(\bar{P}_{s,a} - \widetilde{P}_{s,a})V\right| \\
&\leq \left|\sum_{s' \in \mathcal{X}} (P_{s,a}(s') - \bar{P}_{s,a}(s'))V(s')\right| + \left|(P_{s,a}(y) - \bar{P}_{s,a}(y))(1 + V^{(l-1)}(s_0))\right| + \left|(\bar{P}_{s,a} - \widetilde{P}_{s,a})V\right| \\
&\leq 2L \sum_{s' \in \mathcal{X}} \left|P_{s,a}(s') - \bar{P}_{s,a}(s')\right| + 2L \left|P_{s,a}(y) - \bar{P}_{s,a}(y)\right| + \left|(\bar{P}_{s,a} - \widetilde{P}_{s,a})V\right| \\
&\lesssim \frac{L\sqrt{|\mathcal{X}|\log(|\mathcal{X}|)}}{\sqrt{n(s,a)}} + \frac{L|\mathcal{X}|\log(|\mathcal{X}|)}{n(s,a)} + \frac{(\bar{P}_{s,a} + \mathbb{I}_g)V}{n(s,a)+1},
\end{aligned}$$

where in the last step we applied Cauchy-Schwarz inequality, the good events, the definition of $\widetilde{P}_{s,a}$, and removed logarithmic terms and constants. Clearly, there exists $n_1 = \tilde{\mathcal{O}}(L^2|\mathcal{X}|\log(|\mathcal{X}|/\delta))$, such that when $n(s,a) \geq n_1$, we have $|(P_{s,a} - \widetilde{P}_{s,a})V| \leq \frac{1}{8}$. Moreover, we have

$$b^{(l)}(s,a) \lesssim \max\left\{ \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V^{(l-1)})}{n(s,a)}}, \frac{L}{n(s,a)} \right\} \lesssim \frac{L}{\sqrt{n(s,a)}}.$$

Then there exist $n_2 = \tilde{\mathcal{O}}(L^2 \log(1/\delta_0))$ such that when $n(s,a) \geq n_2$, $b^{(l)}(s,a) \leq \frac{1}{8}$. Thus when $n(s,a) \geq \max\{n_1, n_2\}$ for all $s \in \mathcal{X}, a \in \mathcal{A}$, for any $s \in \mathcal{X}$,

$$\begin{aligned}
V(s) &= V^{(l)}(s) \geq 1 + \widetilde{P}_{s,\pi_g(s)} V^{(l-1)}(s) - b^{(l)}(s, \pi_g(s)) \\
&\geq 1 - \xi + \widetilde{P}_{s,\pi_g(s)} V^{(l)} - b^{(l)}(s, \pi_g(s)) \\
&\geq 1 - \xi + P_{s,\pi_g(s)} V - \left|(P_{s,\pi_g(s)} - \widetilde{P}_{s,\pi_g(s)})V\right| - b^{(l)}(s, \pi_g(s)) \geq \frac{1}{2} + P_{s,\pi_g(s)} V(s),
\end{aligned}$$

where we used the definition of $V^{(l)}$, the stopping condition of VISGO, and the previously derived bounds. For $s \notin \mathcal{X}$, we have $V(s) = (1 + V^{(l-1)}(s_0))\mathbb{I}\{s \neq g\} \geq (\frac{1}{2} + V(s_0))\mathbb{I}\{s \neq g\}$. Applying this recursively gives $V(s) \geq \frac{1}{2} V_g^{\pi_g}(s)$. This completes the proof. $\square$

Lemma 5. *For any subsets $\mathcal{X}$ and $\mathcal{X}'$ such that $\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{S}$, any $g \in \mathcal{S} \setminus \mathcal{X}'$, $\xi > 0$, counter n, and $\delta \in (0,1)$, we have $V_{\mathcal{X}'}(s) \leq V_{\mathcal{X}}(s)$ for any $s \in \mathcal{S}$, where we define $V_{\mathcal{X}''} = \mathrm{VISGO}(\mathcal{X}'', g, \xi, n, \delta)$ (see Algorithm 4) for any $\mathcal{X}'' \subseteq \mathcal{S}$.*

Proof. For any $\mathcal{X}'' \subseteq \mathcal{S}$, denote by $Q_{\mathcal{X}''}^{(i)}$ and $V_{\mathcal{X}''}^{(i)}$ the values of $Q^{(i)}$ and $V^{(i)}$ in Algorithm 4 respectively when computing $V_{\mathcal{X}''}$. It suffices to prove that $V_{\mathcal{X}'}^{(i)}(s) \leq V_{\mathcal{X}}^{(i)}(s)$ for any $s \in \mathcal{S}$ and $i \geq 0$ by induction. The base case $i = 0$ is clearly true by initialization. When $i > 0$, we consider three disjoint cases: 1) if $s \in \mathcal{X}$, by the induction step and Lemma 49, for any

$a \in \mathcal{A}$,

$$
\begin{aligned}
&1 + \widetilde{P}_{s,a} V_{\mathcal{X}'}^{(i-1)} - \max\left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V_{\mathcal{X}'}^{(i-1)})\iota_{s,a}}{n^+(s,a)}}, \frac{c_2 L \iota_{s,a}}{n^+(s,a)} \right\} \\
&\leq 1 + \widetilde{P}_{s,a} V_{\mathcal{X}}^{(i-1)} - \max\left\{ c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_{s,a}, V_{\mathcal{X}}^{(i-1)})\iota_{s,a}}{n^+(s,a)}}, \frac{c_2 L \iota_{s,a}}{n^+(s,a)} \right\}.
\end{aligned}
$$

This implies that $V_{\mathcal{X}'}^{(i)}(s) \leq V_{\mathcal{X}}^{(i)}(s)$ for $s \in \mathcal{X}$. 2) if $s \in \mathcal{X}' \setminus \mathcal{X}$, we have: $V_{\mathcal{X}'}^{(i)}(s) \leq Q_{\mathcal{X}'}^{(i)}(s, \mathrm{RESET}) \leq 1 + \widetilde{P}_{s,\mathrm{RESET}} V_{\mathcal{X}'}^{(i-1)} \overset{\text{(i)}}{\leq} 1 + V_{\mathcal{X}'}^{(i-1)}(s_0) \overset{\text{(ii)}}{\leq} 1 + V_{\mathcal{X}}^{(i-1)}(s_0) = V_{\mathcal{X}}^{(i)}(s)$, where step (i) is by $P_{s,\mathrm{RESET}}(s_0) = 1$ and step (ii) is by the induction step. 3) if $s \in \mathcal{S} \setminus \mathcal{X}'$, by the induction step we have $V_{\mathcal{X}'}^{(i)}(s) = (1 + V_{\mathcal{X}'}^{(i-1)}(s_0))\mathbb{I}\{s \neq g\} \leq (1 + V_{\mathcal{X}}^{(i-1)}(s_0))\mathbb{I}\{s \neq g\} = V_{\mathcal{X}}^{(i)}(s)$. Combining these three cases completes the proof. $\square$

C. Analysis of Algorithm 1

In this section, we assume the state space is finite (i.e., $S = |\mathcal{S}| < \infty$).

C.1. Properties of the sets built by Algorithm 1

Lemma 6. *Denote by $\mathcal{K}_r$ the set $\mathcal{K}$ at the end of each round r, by $g_r^\star$ the goal selected in such a round, and by $\pi_{g_r^\star,r}$ its corresponding policy (computed by VISGO in Line 4). With probability at least $1-\delta$ over the randomness of Algorithm 1, we have that, for any round r,*

- $\mathcal{K}_r \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$;
- *if Line 5 is False, then* $\|V_{g_r^\star}^{\pi_{g_r^\star,r}}\|_\infty \leq 4L$ *which implies* $\|V^\star_{\mathcal{K}_{r-1},g_r^\star}\|_\infty \leq 4L$;
- *for all* $g \in \mathcal{K}_r$, $\|V_g^{\widetilde{\pi}_g}\|_\infty \leq 4L$ *and* $V_g^{\widetilde{\pi}_g}(s_0) \leq L(1+\epsilon)$.

Proof. Clearly, $\mathcal{K}_1 = \{s_0\} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$. Then, consider a round $r \geq 2$ and suppose $\mathcal{K}_{r-1} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ (inductive hypothesis). If, in this round, the algorithm selects a goal $g_r^\star \in \mathcal{U} \setminus \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, Line 5 is False, and a skip round is not triggered, then Line 19 is reached. We now prove that the "failure test" in that line triggers.

Note that every time $\mathcal{K}$ is updated, the sampling at Line 10 guarantees that for all $(s,a) \in \mathcal{K}_{r-1} \times \mathcal{A}$, $\mathbf{N}_{r-1}(s,a) \geq O(L^2|\mathcal{K}_{r-1}|\log(S/\delta))$. By Lemma 3, since $\mathcal{K}_{r-1} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ (inductive hypothesis), we have that

$$\mathbb{P}\left(\forall g \in \mathcal{S} \setminus \mathcal{K}_{r-1} : V_g^{\pi_g}(s) \leq 2V_{\mathcal{K}_{r-1},g}(s)\right) \geq 1 - \frac{\delta}{4r^2}. \tag{2}$$

where $(\_, V_{\mathcal{K}_{r-1},g}, \_) = \text{VISGO}(\mathcal{K}_{r-1}, g, \xi_r, \mathbf{N}_{r-1}, \frac{\delta}{4r^2S^2})$ and ξ_r is the value of ϵ_{VI} used in round r.

Note that VISGO returns a value function that is either ∞ or bounded by $2L$ for all states (see Alg. 4). Since $g_r^\star$ passes the test of Line 5, then $V_{g_r^\star}^{\pi_{g_r^\star,r}}(s) \leq 2V_{\mathcal{K}_{r-1},g_r^\star}(s) \leq 4L$, for all $s \in \mathcal{S}$. Combining this with Lemma 50 and definition of $\lambda = N_{\text{DEV}}(32L, \frac{\epsilon}{256}, \frac{\delta}{4r^2})$, we have $\widehat{\tau} \geq V_{g^\star}^{\pi_{g^\star}}(s_0) - L\epsilon/2$ with probability at least $1 - \frac{\delta}{4r^2}$. By assumption on $g_r^\star$ and since $\pi_{g_r^\star,r}$ is restricted on $\mathcal{K}_{r-1} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, we have $V_{g_r^\star}^{\pi_{g_r^\star,r}}(s_0) \geq V^\star_{\mathcal{K}_{r-1},g_r^\star}(s_0) \geq V^\star_{\mathcal{S}^{\rightarrow}_{L(1+\epsilon)},g_r^\star}(s_0) > L(1+\epsilon)$, which implies that $\widehat{\tau} \geq L(1+\epsilon/2) \geq V_{\mathcal{K}_{r-1},g_r^\star}(s_0) + \epsilon L/2$ with the same probability, where the last inequality is from the goal-selection rule. Therefore, the failure test of Line 19 triggers and $g_r^\star$ is not added to $\mathcal{K}'_r$ or $\mathcal{K}_r$. Therefore, by the inductive hypothesis $\mathcal{K}_r \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$. A union bound over all $r \geq 1$ yields the first statement with probability at least $1-\delta$.

To prove the second statement, note that we already proved above that $V_{g_r^\star,r}^{\pi_{g_r^\star}}(s) \leq 4L$ at any round r where Line 5 is False (i.e., where $g_r^\star$ reaches the policy evaluation step). Since $\pi_{g_r^\star,r}$ is restricted on $\mathcal{K}_{r-1}$, we clearly have $V^\star_{\mathcal{K}_{r-1},g_r^\star}(s) \leq V_{g_r^\star}^{\pi_{g_r^\star,r}}(s) \leq 4L$. This proves the second statement for any round r, which holds with the same $1-\delta$ probability.

Finally, the third statement is a simple consequence of the fact that any goal $g \in \mathcal{K}_r$ must have reached the policy evaluation step in some round $r' < r$ and the round was successful, and thus $\|V_g^{\widetilde{\pi}_g}\|_\infty \leq 4L$ by the second statement. Moreover, by the definition of success round, value of λ and Lemma 50, we have that, for each $g \in \mathcal{K}_r$, there exists $r' < r$ such that $V_g^{\widetilde{\pi}_g}(s_0) = V_{g_{r'}^\star}^{\pi_{g_{r'}^\star,r'}}(s_0) \leq \widehat{\tau} + \frac{L\epsilon}{2} \leq V_{\mathcal{K}_{r'-1},g_{r'}^\star}(s_0) + L\epsilon \leq L(1+\epsilon)$. This holds with the same $1-\delta$ probability as above since we have already union bounded across the application of Lemma 50 for all $g_r^\star$ at all $r \geq 1$. $\square$

Lemma 7. *With probability at least $1-2\delta$, for any round $r \geq 1$ in which $\mathcal{K}_r$ is updated (i.e., Line 8 is executed), $\mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \subseteq \mathcal{U}_r$.*

Proof. For any round r, let $\mathcal{F}_{r-1}$ denote the sigma-algebra generated by the history up to the previous round. Let H_k denote the event "Line 8 is executed at round k". Note that H_k is $\mathcal{F}_{k-1}$-measurable since no random step happens before Line 8 in round r. Moreover, define the events $E_r := \{\forall g \in \mathcal{K}_r : \|V_g^{\widetilde{\pi}_g}\|_\infty \leq 4L\}$ and $E := \{\forall r \geq 1 : E_r\}$. Note that E holds with probability at least $1-\delta$ by Lemma 6. We have

$$
\begin{aligned}
\mathbb{P}\left(\exists r \geq 1 : H_r, \mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \not\subseteq \mathcal{U}_r\right) &\leq \mathbb{P}\left(\exists r \geq 1 : H_r, \mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \not\subseteq \mathcal{U}_r, E\right) + \mathbb{P}(\neg E) && \text{(union bound)} \\
&\leq \mathbb{P}\left(\exists r \geq 1 : H_r, \mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \not\subseteq \mathcal{U}_r, E_r\right) + \delta && \text{(Lemma 6)} \\
&\leq \sum_{r\geq 1} \mathbb{P}\left(\mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \not\subseteq \mathcal{U}_r, E_r, H_r\right) + \delta. && \text{(union bound)} \\
&\leq \sum_{r\geq 1} \mathbb{P}\left(\mathcal{N}(\mathcal{K}_r, \frac{1}{2L}) \not\subseteq \mathcal{U}_r, E_r, H_r\right) + \delta. && (\mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \subseteq \mathcal{N}(\mathcal{K}_r, \tfrac{1}{2L}))
\end{aligned}
$$

Now take any round $r \geq 1$. Recall that $\mathcal{U}_r$ is built by sampling from each $(s,a) \in \mathcal{K}_r \times \mathcal{A}$ exactly $\mu_r := 2L\log(4SALr^2/\delta)$ times. For each $(s,a) \in \mathcal{K}_r \times \mathcal{A}$, let $s_{i,s,a}$ be the i-th sample (i.e., $s_{i,s,a} \sim P_{s,a}$) for $i \in [\mu_r]$. In order to collect each sample $s_{i,s,a}$, we must play the policy $\tilde{\pi}_s$ from s_0 until reaching s. Note that, under event E_r, $\|V_s^{\tilde{\pi}_s}\|_\infty \leq 4L$ for all $s \in \mathcal{K}_r$, hence all the states in $\mathcal{K}_r$ are reached with probability one (so $s_{i,s,a}$ is well defined for all s, a, i). Then, for any fixed $\mathcal{K}_r$,

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{N}(\mathcal{K}_r, \frac{1}{2L}) \not\subseteq \mathcal{U}_r, E_r, H_r \mid \mathcal{K}_r\right) &\leq \mathbb{P}\left(\exists s' \in \mathcal{N}(\mathcal{K}_r, \frac{1}{2L}), \forall (s,a) \in \mathcal{K}_r \times \mathcal{A}, \forall i \in [\mu_r] : s_{i,s,a} \neq s' \mid \mathcal{K}_r\right) \\
&\leq \sum_{s' \in \mathcal{N}(\mathcal{K}_r, \frac{1}{2L})} \mathbb{P}\left(\forall (s,a) \in \mathcal{K}_r \times \mathcal{A}, \forall i \in [\mu] : s_{i,s,a} \neq s'\right) && \text{(union bound)} \\
&\leq \sum_{s' \in \mathcal{N}(\mathcal{K}_r, \frac{1}{2L})} \max_{(s,a)\in\mathcal{K}_r\times\mathcal{A}} \mathbb{P}\left(\forall i \in [\mu] : s_{i,s,a} \neq s'\right) && \text{(trivial)} \\
&\leq \sum_{s' \in \mathcal{N}(\mathcal{K}_r, \frac{1}{2L})} \max_{(s,a)\in\mathcal{K}_r\times\mathcal{A}} \prod_{i\in[\mu_r]} (1 - P(s'|s,a)) && \text{(all } s_{i,s,a} \text{ are i.i.d.)} \\
&\leq \sum_{s' \in \mathcal{N}(\mathcal{K}_r, \frac{1}{2L})} \left(1 - \frac{1}{2L}\right)^{\mu_r} && \text{(definition of } \mathcal{N}(\mathcal{K}_r, \tfrac{1}{2L})) \\
&\leq \sum_{s' \in \mathcal{N}(\mathcal{K}_r, \frac{1}{2L})} \frac{\delta}{4LASr^2} \leq \frac{\delta}{2r^2}.
\end{aligned}
$$

Now let Ω_{r-1} denote the sample space under which $\mathcal{F}_{r-1}$ is generated, such that $\sum_{\omega\in\Omega_{r-1}} \mathbb{P}(\omega) = 1$. Noting that $\mathcal{K}_r$ is measurable w.r.t. $\mathcal{F}_{r-1}$, define $\mathcal{K}_r(\omega)$ as the set $\mathcal{K}_r$ obtained after history ω. Then,

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{N}(\mathcal{K}_r, \frac{1}{2L}) \not\subseteq \mathcal{U}_r, E_r, H_r\right) &= \sum_{\omega\in\Omega_{r-1}} \mathbb{P}\left(\mathcal{N}(\mathcal{K}_r, \frac{1}{2L}) \not\subseteq \mathcal{U}_r, E_r, H_r \mid \omega\right) \mathbb{P}(\omega) \\
&= \sum_{\omega\in\Omega_{r-1}: E_r, H_r} \mathbb{P}\left(\mathcal{N}(\mathcal{K}_r, \frac{1}{2L}) \not\subseteq \mathcal{U}_r \mid \omega\right) \mathbb{P}(\omega) \\
&= \sum_{\omega\in\Omega_{r-1}: E_r, H_r} \mathbb{P}\left(\mathcal{N}(\mathcal{K}_r, \frac{1}{2L}) \not\subseteq \mathcal{U}_r \mid \mathcal{K}_r(\omega), E_r, H_r\right) \mathbb{P}(\omega) \leq \frac{\delta}{2r^2}.
\end{aligned}
$$

Plugging this into our initial inequality, we get $\mathbb{P}\left(\exists r \geq 1 : H_r, \mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \not\subseteq \mathcal{U}_r\right) \leq 2\delta$. □

Lemma 8 (Restricted Optimism)**.** *With probability at least $1-\delta$ over the randomness of Algorithm 1, for any $j \in [S]$ and any round $r \geq 1$, after executing Line 4, if $\mathcal{K}_j^\star \subseteq \mathcal{K}_r$, then $V_{\mathcal{K}_r,g}(s) \leq V^\star_{\mathcal{K}_j^\star,g}(s)$ for any $s \in \mathcal{S}$ and $g \in \mathcal{K}^\star_{j+1} \setminus \mathcal{K}_r$, where $\mathcal{K}_r$ is the set $\mathcal{K}$ immediately after the execution of Line 4.*

Proof. Let $j \in [S]$ and $g \in \mathcal{K}^\star_{j+1} \setminus \mathcal{K}^\star_j$. Fix some round $r \geq 1$ s.t. $\mathcal{K}^\star_j \subseteq \mathcal{K}_r$. Let $\delta_r = \frac{\delta}{4r^2S^2}$ and $(Q_\xi, V_\xi, \_) =$ VISGO$(\mathcal{K}^\star_j, g, \xi, \mathbf{N}, \delta_r)$. By Lemma 2 [7],

$$
\mathbb{P}\Big(\forall \xi > 0, s \in \mathcal{S} : V_\xi(s) \leq V^\star_{\mathcal{K}^\star_j,g}(s)\Big) \geq 1 - \delta_r. \tag{3}
$$

[7] Note that, by definition, $\|V^\star_{\mathcal{K}^\star_j,g}\|_\infty \leq L + 1 \leq 2L$ for all $g \in \mathcal{K}^\star_{j+1} \setminus \mathcal{K}^\star_j$ (which is a prerequisite of Lemma 2).

Then, from a union bound and $|\mathcal{K}^\star_{j+1} \setminus \mathcal{K}^\star_j| \leq S$, the event above holds simultaneously across all $j \in [S]$, and $g \in \mathcal{K}^\star_{j+1} \setminus \mathcal{K}^\star_j$ with probability at least $1 - \frac{\delta}{4r^2}$. This implies that the same result holds for all $g \in \mathcal{K}^\star_{j+1} \setminus \mathcal{K}_r$ since $\mathcal{K}^\star_{j+1} \setminus \mathcal{K}_r \subseteq \mathcal{K}^\star_{j+1} \setminus \mathcal{K}^\star_j$. A union bound implies that this holds at all rounds simultaneously with probability at least $1 - \delta$.

Now consider the execution of Line 4 and let $\mathcal{K}_r, \delta_r, \xi_r, \mathbf{N}_r$ be the values of the parameters used by VISGO in such a round, such that $\mathcal{K}^\star_j \subseteq \mathcal{K}_r$ for some $j \in [S]$. For any $g \in \mathcal{K}^\star_{j+1} \setminus \mathcal{K}_r$, let $(\_, V_{\mathcal{K}_r,g}, \_) = \text{VISGO}(\mathcal{K}_r, g, \xi_r, \mathbf{N}_r, \delta_r)$ and $(\_, V_{\mathcal{K}^\star_j,g}, \_) = \text{VISGO}(\mathcal{K}^\star_j, g, \xi_r, \mathbf{N}_r, \delta_r)$. Then, Eq. 3 implies that, for any $s \in \mathcal{S}$, $V_{\mathcal{K}^\star_j,g}(s) \leq V^\star_{\mathcal{K}^\star_j,g}(s)$. If $\mathcal{K}^\star_j \subseteq \mathcal{K}_r$, by the update rule of Algorithm 4 and Lemma 5, we also have $V_{\mathcal{K}_r,g}(s) \leq V_{\mathcal{K}^\star_j,g}(s) \leq V^\star_{\mathcal{K}^\star_j,g}(s)$. □

The following lemma shows that if a set $\mathcal{K}^\star_j \subseteq \mathcal{K}$ at some round, at the next update of $\mathcal{K}$ it must be that $\mathcal{K}^\star_{j+1} \subseteq \mathcal{K}$ (if the algorithm does not terminate) and ensures correctness, in the sense that the algorithm returns a set of states including $\mathcal{S}^{\rightarrow}_L$ with high probability.

Lemma 9 (Correctness)**.** *Denote by $\mathcal{K}_r$ (resp $\mathcal{U}_r$) the set $\mathcal{K}$ (resp. $\mathcal{U}$) at the end of each round r. With probability at least $1 - 3\delta$, for any $j \geq 1$ and round $r \geq 1$ in which $\mathcal{K}_r$ is updated or returned (i.e., Line 8 is executed) and $\mathcal{K}_{r-1} \supseteq \mathcal{K}^\star_j$, we have $\mathcal{K}^\star_{j+1} \subseteq \mathcal{K}_r$. Moreover, under the same probability, we have that, for any $r \geq 1$, $\mathcal{S}^{\rightarrow}_L \subseteq \mathcal{K}_r$ if the algorithm terminates at round r.*

Proof. Define the event $E := \{\forall r \geq 1 \text{ in which } \mathcal{K}_r \text{ is updated} : \mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \subseteq \mathcal{U}_r\}$. By Lemma 7, it holds with probability at least $1 - 2\delta$. Let us carry out the proof conditioned on E holding.

Take some round r such that Line 8 is executed and $\mathcal{K}_{r-1} \supseteq \mathcal{K}^\star_j$. Let r' be the last round where $\mathcal{K}_{r'}$ was updated (and thus $\mathcal{U}_{r'}$ was created). Note that $\mathcal{K}_{r'} = \mathcal{K}_{r-1} \supseteq \mathcal{K}^\star_j$. Then, event E and the definition of the sets $(\mathcal{K}^\star_j)_j$ directly imply that $\mathcal{K}^\star_{j+1} := \mathcal{T}_L(\mathcal{K}^\star_j) \subseteq \mathcal{T}_L(\mathcal{K}_{r'}) \subseteq \mathcal{U}_{r'} \cup \mathcal{K}_{r'}$. Since $\mathcal{K}_r$ can only be formed by adding states in $\mathcal{U}_{r'}$ to $\mathcal{K}_{r'}$, and the union of these sets contains $\mathcal{K}^\star_{j+1}$, if $\mathcal{K}^\star_{j+1} \not\subseteq \mathcal{K}_r$, it must be that there exists $g \in \mathcal{U}_{r-1} \cap \mathcal{K}^\star_{j+1}$ s.t. $V_{\mathcal{K}_{r-1},g}(s_0) > L$. However, Lemma 8, which holds with probability $1 - \delta$, implies that, at any round $r \geq 1$, if $\mathcal{K}^\star_j \subseteq \mathcal{K}_{r-1}$, then $V_{\mathcal{K}_{r-1},g}(s_0) \leq V^\star_{\mathcal{K}^\star_j,g}(s_0) \leq L$ for any $g \in \mathcal{K}^\star_{j+1} \setminus \mathcal{K}_{r-1}$. This is a contradiction, which implies that $\mathcal{U}_{r-1} \cap \mathcal{K}^\star_{j+1} = \emptyset$ and, thus, all states in $\mathcal{K}^\star_{j+1}$ must have been added to $\mathcal{K}_r$. A union bound over the application of Lemma 7 and Lemma 8 yields the statement.

To prove the second statement, let us use the same events as above. First note that, since $\mathcal{K}_1 = \mathcal{K}^\star_1 = \{s_0\}$, it must be that, at any round r, $\mathcal{K}_r \supseteq \mathcal{K}^\star_j$ for some $j \geq 1$. Now take any round r in which the algorithm terminates and suppose $\mathcal{K}_{r-1} \not\supseteq \mathcal{S}^{\rightarrow}_L$. Let $j^\star$ be the largest j s.t. $\mathcal{K}_r \supseteq \mathcal{K}^\star_j$. By Lemma 1, it must be that $j < J$, hence $\mathcal{K}^\star_{j^\star+1} \supset \mathcal{K}^\star_{j^\star}$. Let r' be the last round at which $\mathcal{K}_{r'}$ was updated. Since the algorithm terminates at round r it must be that $\mathcal{K}'_{r-1} = \emptyset$, i.e., no state in $\mathcal{U}_{r-1} = \mathcal{U}_{r'}$ has been found to be added to $\mathcal{K}_r$. From the same argument as above, under E it must be that $\mathcal{K}^\star_{j^\star+1} \subseteq \mathcal{U}_{r'} \cup \mathcal{K}_{r'}$. Since $\mathcal{K}_{r-1} \not\supseteq \mathcal{S}^{\rightarrow}_L$, and no addition to $\mathcal{K}_{r-1}$ is performed as the algorithm stops at r, it must be that there exists $g \in \mathcal{U}_{r-1} \cap \mathcal{K}^\star_{j^\star+1}$ s.t. $V_{\mathcal{K}_{r-1},g}(s_0) > L$. However, in the first part of the proof, we already found a contradiction for this case under the event of Lemma 8. This implies that the algorithm cannot stop at r since some state must be added. Hence, whenever the algorithm stops it must be that $\mathcal{K}_r \supseteq \mathcal{S}^{\rightarrow}_L$. This completes the proof. □

Lemma 10 (Correctness under Assumption 2)**.** *Denote by $\mathcal{K}_r$ the set $\mathcal{K}$ at the end of each round r. With Assumption 2, with probability at least $1 - 5\delta$ over the randomness of Algorithm 1, for any round $r \geq 1$, we have that $\mathcal{K}_r = \mathcal{K}^\star_j$ for some $j \in [S^{\rightarrow}_L]$ and $\mathcal{K}_r = \mathcal{S}^{\rightarrow}_L$ if the algorithm terminates at round r.*

Proof. By Lemma 6 and Lemma 9, with probability at least $1 - 4\delta$, we have $\mathcal{S}^{\rightarrow}_L \subseteq \mathcal{K}_r \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ if the algorithm terminates at round r. By Remark 1, $\mathcal{K} = \mathcal{S}^{\rightarrow}_L$. Thus, it suffices to show that, at any round r, $\mathcal{K}_r = \mathcal{K}^\star_j$ for some $j \leq |\mathcal{S}^{\rightarrow}_L|$.

The algorithm is such that $\mathcal{K}_1 = \mathcal{K}^\star_1 = \{s_0\}$. Suppose at, in some round $r \geq 1$, we have that $\mathcal{K}_r = \mathcal{K}^\star_j$ for some $j \geq 1$. By Lemma 9, with the same probability as above, if the condition of Line 7 becomes True for the first time in some round $r' > r$ (i.e., the set $\mathcal{K}$ is updated in such round), then we must have $\mathcal{K}^\star_{j+1} \subseteq \mathcal{K}_{r'}$ at then end of round r'. We shall prove that we also have $\mathcal{K}_{r'} \subseteq \mathcal{K}^\star_{j+1}$, which implies the statement.

Take any round r such that $\mathcal{K}_{r-1} = \mathcal{K}^\star_j$ and $g^\star_r \in \mathcal{U} \setminus \mathcal{K}^\star_{j+1}$. Since, the last time $\mathcal{K}$ was updated Line 10 was called, we must have $\mathbf{N}_{r-1}(s,a) \geq O(L^2|\mathcal{K}^\star_j|\log(S/\delta))$ for all $(s,a) \in \mathcal{K}^\star_j \times \mathcal{A}$. Then, by Lemma 3, with probability at least $1 - \frac{\delta}{4r^2}$, for all $s \in \mathcal{S}$, $V^{\pi_{g^\star_r}}_{g^\star_r}(s) \leq 2V_{\mathcal{K}_{r-1},g^\star_r}(s) \leq 4L$ due to properties of VISGO if Line 5 is False. If a skip round is not triggered, combining this with Lemma 50 and definition of λ, we have $\widehat{\tau} \geq V^{\pi_{g^\star_r}}_{g^\star_r}(s_0) - L\epsilon/2$ with probability at least $1 - \frac{\delta}{4r^2}$.

By Assumption 2, assumption on $g^\star_r$, and since $\pi_{g^\star_r}$ is restricted on $\mathcal{K}_{r-1} = \mathcal{K}^\star_j$, we have $V^{\pi_{g^\star_r}}_{g^\star_r}(s_0) \geq V^\star_{\mathcal{K}^\star_j, g^\star_r}(s_0) > L(1+\epsilon)$, which implies that $\widehat{\tau} \geq L(1+\epsilon/2) \geq V_{\mathcal{K}_{r-1}, g^\star_r}(s_0) + \epsilon L/2$ with the same probability, where the last inequality is from the fact that Line 5 is False. Therefore, the failure test triggers and $g^\star_r$ is not added to $\mathcal{K}'_r$ or $\mathcal{K}_r$ since a failure round is triggered. This holds with probability at least $1-\delta$ across all rounds by a union bound. Therefore, for any round r in which $\mathcal{K}$ is updated and $\mathcal{K}_{r-1} = \mathcal{K}^\star_j$, we must have $\mathcal{K}_r \subseteq \mathcal{K}^\star_{j+1}$. This concludes the proof, and the statement holds with probability at least $1-5\delta$ by a union bound. □

C.2. Analysis of Policy Evaluation

We consider the regret over the trajectories generated in the policy evaluation phase. We concatenate all policy evaluation episodes in all rounds and index them with $k \geq 1$. To make the notation consistent with Algorithm 5, we treat the whole learning procedure as an artificial trial. Let $\mathcal{K}_k$, V_k, and Q_k be the $\mathcal{K}$, $V_{\mathcal{K},g^\star}$, and $Q_{\mathcal{K},g^\star}$ in episode k. Let π_k and g_k be the corresponding policy $\pi_{g^\star}$ and goal $g^\star$. Denote by $\mathcal{F}_k$ the σ-algebra of events up to episode k. Let K be the total number of episodes throughout the execution of Algorithm 1. For any sequence of indicators $\mathcal{I} = \{\mathbf{1}_k\}_k$ with $\mathbf{1}_k \in \mathcal{F}_{k-1}$, define $R_{K',\mathcal{I}} = \sum_{k=1}^{K'}(I_k - V_k(s_0))\mathbf{1}_k$ and $C_{K'} = \sum_{k=1}^{K'} I_k$ for $K' \in [K]$. Define $P^k_i = P_{s^k_i, a^k_i}$. In episode k, when $s^k_i \in \mathcal{K}$, denote by $\bar{P}^k_i$, $\widetilde{P}^k_i$, $\mathbf{N}^k_i$, b^k_i the values of $\bar{P}_{s^k_i,a^k_i}$, $\widetilde{P}_{s^k_i,a^k_i}$, $n^+(s^k_i, a^k_i)$, and $b^{(l)}(s^k_i, a^k_i)$, where $\bar{P}$, n^+, $b^{(l)}$ are used in Algorithm 4 to compute V_k and l is the final value of i in Algorithm 4; when $s^k_i \notin \mathcal{K}$, define $\bar{P}^k_i = \mathbb{I}_{s_0}$, $\mathbf{N}^k_i = \infty$, and $b^k_i = 0$. Also define ϵ_k, δ_k as the value of $\epsilon_{\text{VI}}, \delta$ used in Algorithm 4 to compute V_k. Note that $I_k < \infty$ with probability 1 by Line 17, and $s^k_{I_k+1} \neq g$ only when a skip round is triggered in episode k.

C.2.1. Regret bound without Assumption 2

Lemma 11. *For any sequence of indicators $\mathcal{I} = \{\mathbf{1}_k\}_k$ with $\mathbf{1}_k \in \mathcal{F}_{k-1}$, we have, with probability at least $1-6\delta$, for any $K' \in [K]$,*

$$R_{K',\mathcal{I}} \lesssim L\log(SAL/\delta)^2\log(K)\sqrt{S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}AK'} + L{S^{\rightarrow}_{L(1+\epsilon)}}^2 A(\log K')^2\log(SAL/\delta)^3.$$

Moreover, $C_{K'} \lesssim LK' + L{S^{\rightarrow}_{L(1+\epsilon)}}^2 A(\log K')^2 \log(SAL/\delta)^3$.

Proof. We start by decomposing the regret as

$$\begin{aligned}\sum_{k=1}^{K'}(I_k - V_k(s_0))\,\mathbf{1}_k &\leq \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(1 + V_k(s^k_{i+1}) - V_k(s^k_i)\right)\mathbf{1}_k && (\pm \textstyle\sum_{i=1}^{I_k} V_k(s^k_{i+1}))\\ &\leq \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((\mathbb{I}_{s^k_{i+1}} - P^k_i)V_k + (P^k_i - \bar{P}^k_i)V_k + (\bar{P}^k_i - \widetilde{P}^k_i)V_k + b^k_i + \epsilon_k\right)\mathbf{1}_k, && (\text{definition of } V_k)\end{aligned}$$

where the last inequality uses that $V^{(l)}_k(s) = 1 + \widetilde{P}^k_{s,a}V^{(l-1)}_k - b^k_{s,a}$ for any $s \in \mathcal{K}_k, a \in \mathcal{A}$, where l is the index of the last iteration of VISGO when called with $(\_, V_k, \pi_g) = \text{VISGO}(\mathcal{K}_k, g_k, \epsilon_k, \mathbf{N}_k, \delta_k)$, and $\|V^{(l)}_k - V^{(l-1)}_k\|_\infty \leq \epsilon_k$ by definition of its termination condition (recall that V_k is bounded since Line 5 was passed). Note that, if $s^k_i \notin \mathcal{K}_k$, then the i,k term in the sum of the second line is clearly an upper bound to the corresponding term in the first line. We bound the terms above separately.

First term By Lemma 55 and $\|V_k\|_\infty \leq 2L$ (by VISGO and since Line 5 was passed), with probability at least $1-\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(\mathbb{I}_{s^k_{i+1}} - P^k_i)V_k\mathbf{1}_k \leq \sqrt{\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbf{1}_k\mathbb{V}(P^k_i, V_k)\iota} + L\iota,$$

where $\iota = 9\log(16L^2C^3_{K'}/\delta)$.

Second term Note that, by the event of Lemma 6, $\mathcal{K}_k \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ in all episodes k. Moreover, when $s^k_i \notin \mathcal{K}_k$, the k,i term in the sum is zero by definition of P^k_i and $\bar{P}^k_i$. Therefore, we have all the preconditions to apply Lemma 46 on terms

$(P_i^k - \bar{P}_i^k)V_k$ for all i, k s.t. $s_i^k \in \mathcal{K}_k$, which yields, with probability $1-\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(P_i^k - \bar{P}_i^k)V_k\mathbf{1}_k \lesssim \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(\sqrt{\frac{\Gamma_{L(1+\epsilon)}\mathbb{V}(P_i^k, V_k)\iota'}{\mathbf{N}_i^k}} + \frac{LS^{\rightarrow}_{L(1+\epsilon)}\iota'}{\mathbf{N}_i^k}\right),$$

where $\iota' = O(\log \frac{SALC_{K'}}{\delta})$. Note that Lemma 46 already union bounds across all possible counts, value functions and state-action pair, so we do not need an extra union bound over episodes and steps here.

Then, by Lemma 40 and Cauchy-Schwarz inequality, with the same probability,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(P_i^k - \bar{P}_i^k)V_k\mathbf{1}_k \lesssim \sqrt{S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k)\iota''} + LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota'',$$

where $\iota'' = O(\log(SALC_{K'}/\delta)\log(C_{K'}))$.

Third term By the expressions of $\widetilde{P}_i^k$ and $\bar{P}_i^k$ (cf. Algorithm 4) and Lemma 40,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(\bar{P}_i^k - \widetilde{P}_i^k)V_k\mathbf{1}_k \le \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbf{1}_k\frac{(\bar{P}_i^k + \mathbb{I}_g)V_k}{\mathbf{N}_i^k + 1} \lesssim LS^{\rightarrow}_{L(1+\epsilon)}A\log(C_{K'}). \qquad (\mathbb{I}_g(s') \triangleq \mathbb{I}\{s' = g\})$$

Fourth and fifth term By Lemma 39 and Lemma 41, with probability at least $1-\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(b_i^k + \epsilon_k)\mathbf{1}_k \lesssim \sqrt{S^{\rightarrow}_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k)\iota'} + LS^{\rightarrow}_{L(1+\epsilon)}{}^{1.5}A\iota'.$$

Combining all terms Note that all the derived bounds can be absorbed into the one of the second term. Plugging everything back to our initial expression of the regret,

$$\begin{aligned}\sum_{k=1}^{K'}(I_k - V_k(s_0))\,\mathbf{1}_k &\lesssim \sqrt{S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k)\iota''} + LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota'' \\ &\lesssim \sqrt{LS^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}AC_{K'}\iota''} + LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota''. \qquad \text{(Lemma 36)}\end{aligned}$$

Note that $\iota'' \lesssim \log(SAL/\delta)(\log C_{K'})^2$. Now assuming $\mathbf{1}_k = 1$ for all k, we can solve an inequality to find C_K. First, using that $\log(x) \le x^\alpha/\alpha$ for any $x, \alpha > 0$ together with the derived regret bound, we can find the crude bound on C_K,

$$C_{K'} \lesssim \left(\sum_{k=1}^{K}V_k(s_0) + LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\log(SAL/\delta)\right)^4 \le \left(K'L + LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\log(SAL/\delta)\right)^4.$$

This implies that $\iota'' \lesssim (\log K')^2\log(SAL/\delta)^3$. Plugging this into the regret bound, we get a quadratic inequality in $C_{K'}$. Solving it yields

$$C_{K'} \lesssim \sum_{k=1}^{K'}V_k(s_0) + LS^{\rightarrow}_{L(1+\epsilon)}{}^2A(\log K')^2\log(SAL/\delta)^3 \le LK' + LS^{\rightarrow}_{L(1+\epsilon)}{}^2A(\log K')^2\log(SAL/\delta)^3.$$

Plugging this back into the regret bound gives the stated bound. Throughout the proof we used following events with the corresponding probabilities:

- Lemma 55: $1-\delta$
- Lemma 6: $1-\delta$

- Lemma 46: $1-\delta$
- Lemma 39: $1-\delta$
- Lemma 36: $1-2\delta$

A union bound concludes the proof. $\square$

C.2.2. Regret bound under Assumption 2

Lemma 12. *Under Assumption 2, for any sequence of indicators $\mathcal{I}=\{\mathbf{1}_k\}_k$ with $\mathbf{1}_k \in \mathcal{F}_{k-1}$, we have, with probability at least $1-14\delta$, for any $K' \in [K]$,*

$$R_{K',\mathcal{I}} \lesssim L\log(SAL/\delta)^2\log(K')\sqrt{S^{\rightarrow}_{L(1+\epsilon)}AK'} + LS^{\rightarrow}_{L(1+\epsilon)}{}^2 A(\log K')^2\log(SAL/\delta)^3.$$

Moreover, $C_{K'} \lesssim LK' + LS^{\rightarrow}_{L(1+\epsilon)}{}^2 A(\log K')^2\log(SAL/\delta)^3$.

Proof. Note that, under Assumption 2 and by Lemma 10, in any episode, $\mathcal{K}=\mathcal{K}^\star_j$ for some $j \le J \le |\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}| \le S$ (cf. Lemma 1). Moreover, by Lemma 6, for any round in which $g^\star$ reaches the policy evaluation step, $\|V^\star_{\mathcal{K},g^\star}\|_\infty \le 4L$, which implies that $\|V^\star_{\mathcal{K}^\star_j,g^\star}\|_\infty \le 4L$ for some j in that round. Let $\mathcal{G}_j := \{g \in \mathcal{S} : \|V^\star_{\mathcal{K}^\star_j,g}\|_\infty \le 4L\}$. Consider the event

$$E := \left\{ \forall s \in \mathcal{S}, a \in \mathcal{A}, j \in [S], g \in \mathcal{G}_j, \forall n(s,a) \ge 1 : |(\bar{P}^n_{s,a} - P_{s,a})V^\star_{\mathcal{K}^\star_j,g}| \le \sqrt{\frac{\mathbb{V}(P_{s,a}, V^\star_{\mathcal{K}^\star_j,g})\iota'_{s,a}}{n(s,a)}} + \frac{L\iota'_{s,a}}{n(s,a)} \right\},$$

where $\iota'_{s,a} = 8\log(2S^3An(s,a)/\delta)$. Clearly, by Lemma 54 and a union bound, E holds with probability at least $1-\delta$. Then, assuming E and the events of Lemma 10 and Lemma 6 hold, we clearly have, for all episodes k and steps i,

$$(P^k_i - \bar{P}^k_i)V^\star_k \lesssim \sqrt{\frac{\mathbb{V}(P^k_i, V^\star_k)\iota'}{\mathbf{N}^k_i}} + \frac{L\iota'}{\mathbf{N}^k_i}, \tag{4}$$

where $\iota' = O(\log(SALC_{K'}/\delta))$. Note that we inflated the ι' term with an extra L since it will simplify the bounds later. Now we split the regret as

$$\begin{aligned}
\sum_{k=1}^{K'} (I_k - V_k(s_0))\,\mathbf{1}_k &\le \sum_{k=1}^{K'}\sum_{i=1}^{I_k} \left(1 + V_k(s^k_{i+1}) - V_k(s^k_i)\right)\mathbf{1}_k && (\pm \textstyle\sum_{i=1}^{I_k} V_k(s^k_{i+1})) \\
&\le \sum_{k=1}^{K'}\sum_{i=1}^{I_k} \left((\mathbb{I}_{s^k_{i+1}} - P^k_i)V_k + (P^k_i - \bar{P}^k_i)V_k + (\bar{P}^k_i - \widetilde{P}^k_i)V_k + b^k_i + \epsilon_k \right)\mathbf{1}_k, && (\text{definition of } V_k)
\end{aligned}$$

where the last inequality uses that $V^{(l)}_k(s) = 1 + \widetilde{P}^k_{s,a}V^{(l-1)}_k - b^k_{s,a}$ for any $s \in \mathcal{K}_k, a \in \mathcal{A}$, where l is the index of the last iteration of VISGO when called with $(\_, V_k, \pi_g) = \text{VISGO}(\mathcal{K}_k, g_k, \epsilon_k, \mathbf{N}_k, \delta_k)$, and $\|V^{(l)}_k - V^{(l-1)}_k\|_\infty \le \epsilon_k$ by definition of its termination condition (recall that V_k is bounded since Line 5 was passed). Note that, if $s^k_i \notin \mathcal{K}_k$, then the i,k term in the sum of the second line is clearly an upper bound to the corresponding term in the first line.

We bound the terms above separately.

First term By Lemma 55 and $\|V_k\|_\infty \le 2L$ (by VISGO and since Line 5 was passed), with probability at least $1-\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k} (\mathbb{I}_{s^k_{i+1}} - P^k_i)V_k\mathbf{1}_k \le \sqrt{\sum_{k=1}^{K'}\sum_{i=1}^{I_k} \mathbf{1}_k\mathbb{V}(P^k_i, V_k)\iota} + L\iota,$$

where $\iota = 9\log(16L^2C^3_{K'}/\delta)$.

Second term Note that, from (4),

$$\begin{aligned}\sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k-\bar{P}_i^k)V_k|\mathbf{1}_k &\leq \sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k-\bar{P}_i^k)V_k| \\ &= \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(|(P_i^k-\bar{P}_i^k)V_k^\star| + |(P_i^k-\bar{P}_i^k)(V_k-V_k^\star)|\right) \\ &\leq \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(\sqrt{\frac{\mathbb{V}(P_i^k,V_k^\star)\iota'}{\mathbf{N}_i^k}} + \frac{L\iota'}{\mathbf{N}_i^k} + |(P_i^k-\bar{P}_i^k)(V_k-V_k^\star)|\right).\end{aligned}$$

Note that, by the event of Lemma 6, $\mathcal{K}_k \subseteq \mathcal{S}_{L(1+\epsilon)}^{\rightarrow}$ in all episodes k. Moreover, for all k,i, either $(s_i^k,a_i^k)\in\mathcal{K}_k\times\mathcal{A}$ or the second term above is zero. Since $\|V_k-V_k^\star\|_\infty \leq 6L$, we have all the preconditions to apply Lemma 46 on the terms $|(P_i^k-\bar{P}_i^k)(V_k-V_k^\star)|$, which yields, with probability $1-\delta$, for all i,k,

$$|(P_i^k-\bar{P}_i^k)(V_k-V_k^\star)| \lesssim \sqrt{\frac{S_{L(1+\epsilon)}^{\rightarrow}\mathbb{V}(P_i^k,V_k-V_k^\star)\iota'}{\mathbf{N}_i^k}} + \frac{LS_{L(1+\epsilon)}^{\rightarrow}\iota'}{\mathbf{N}_i^k},$$

where ι' was defined above. Note that Lemma 46 already union bounds across all possible counts, value functions and state-action pair, so we do not need an extra union bound over episodes and steps here. By $\mathrm{VAR}[X+Y]\leq 2(\mathrm{VAR}[X]+\mathrm{VAR}[Y])$, we have that $\mathbb{V}(P_i^k,V_k^\star)\leq 2\mathbb{V}(P_i^k,V_k-V_k^\star)+2\mathbb{V}(P_i^k,V_k)$ and thus

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k-\bar{P}_i^k)V_k| \leq \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(\sqrt{\frac{\mathbb{V}(P_i^k,V_k)\iota'}{\mathbf{N}_i^k}} + \sqrt{\frac{S_{L(1+\epsilon)}^{\rightarrow}\mathbb{V}(P_i^k,V_k-V_k^\star)\iota'}{\mathbf{N}_i^k}} + \frac{LS_{L(1+\epsilon)}^{\rightarrow}\iota'}{\mathbf{N}_i^k}\right).$$

Then, by Cauchy-Schwarz inequality, with the same probability and Lemma 40,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k-\bar{P}_i^k)V_k| \lesssim \sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota''} + \sqrt{{S_{L(1+\epsilon)}^{\rightarrow}}^2A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k-V_k^\star)\iota''} + L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota'',$$

where $\iota''=O(\log(SALC_{K'}/\delta)\log(C_{K'}))$. Now by Lemma 13, with probability at least $1-2\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k^\star-V_k) \lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k-\bar{P}_i^k)V_k| + L\sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota'} + L^2{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota',$$

where ι' was defined above. Let $Z_K := \sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k-\bar{P}_i^k)V_k|$. Plugging this into the previous inequality, using $\sqrt{xy}\leq x+y$ and $\iota'\leq\iota''$, we get

$$Z_{K'} \lesssim \sqrt{{S_{L(1+\epsilon)}^{\rightarrow}}^2AL\iota''Z_{K'}} + \sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota''} + L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota''.$$

Solving thi quadratic inequality for $Z_{K'}$, we conclude with

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k-\bar{P}_i^k)V_k| \lesssim \sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota''} + L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota''.$$

Third term By the expressions of $\widetilde{P}_i^k$ and $\bar{P}_i^k$ (cf. Algorithm 4) and Lemma 40,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(\bar{P}_i^k-\widetilde{P}_i^k)V_k\mathbf{1}_k \leq \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbf{1}_k\frac{(\bar{P}_i+\mathbb{I}_g)V_k}{\mathbf{N}_i^k+1} \lesssim LS_{L(1+\epsilon)}^{\rightarrow}A\log(C_{K'}). \tag{5}$$

Fourth and fifth term By Lemma 39 and Lemma 41, with probability at least $1-\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(b_i^k+\epsilon_k)\mathbf{1}_k \lesssim \sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota'} + LS_{L(1+\epsilon)}^{\rightarrow}{}^{1.5}A\iota'. \tag{6}$$

Combining all terms Note that all the derived bounds can be absorbed into the one of the second term. Plugging everything back to our initial expression of the regret,

$$\begin{aligned}\sum_{k=1}^{K'}\left(I_k - V_k(s_0)\right)\mathbf{1}_k &\lesssim \sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota''} + LS_{L(1+\epsilon)}^{\rightarrow}{}^{2}A\iota'' \\ &\lesssim \sqrt{LS_{L(1+\epsilon)}^{\rightarrow}AC_{K'}\iota''} + LS_{L(1+\epsilon)}^{\rightarrow}{}^{2}A\iota''. \qquad \text{(Lemma 36)}\end{aligned}$$

Note that $\iota'' \lesssim \log(SAL/\delta)(\log C_{K'})^2$. Now assuming $\mathbf{1}_k = 1$ for all k, we can solve an inequality to find $C_{K'}$. First, using that $\log(x) \le x^\alpha/\alpha$ for any $x,\alpha>0$ together with the derived regret bound, we can find the crude bound on $C_{K'}$,

$$C_{K'} \lesssim \left(\sum_{k=1}^{K'}V_k(s_0) + LS_{L(1+\epsilon)}^{\rightarrow}{}^{2}A\log(SAL/\delta)\right)^4 \le \left(K'L + LS_{L(1+\epsilon)}^{\rightarrow}{}^{2}A\log(SAL/\delta)\right)^4.$$

This implies that $\iota'' \lesssim (\log K')^2\log(SAL/\delta)^3$. Plugging this into the regret bound, we get a quadratic inequality in $C_{K'}$. Solving it yields

$$C_{K'} \lesssim \sum_{k=1}^{K'}V_k(s_0) + LS_{L(1+\epsilon)}^{\rightarrow}{}^{2}A(\log K')^2\log(SAL/\delta)^3 \le LK' + LS_{L(1+\epsilon)}^{\rightarrow}{}^{2}A(\log K')^2\log(SAL/\delta)^3.$$

Plugging this back into the regret bound gives the stated bound. Throughout the proof we used following events with the corresponding probabilities:

- Lemma 10: $1-5\delta$
- Lemma 6: $1-\delta$
- Event E in this proof: $1-\delta$
- Lemma 55: $1-\delta$
- Lemma 46: $1-\delta$
- Lemma 39: $1-\delta$
- Lemma 13: $1-2\delta$
- Lemma 36: $1-2\delta$

A union bound concludes the proof. □

C.3. Auxiliary results for policy evaluation

Lemma 13. *With probability at least $1-2\delta$, for any $K' \in [K]$, if 1) $\|V_k\|_\infty = \mathcal{O}(L)$ for any $k \in [K']$, and 2) $V_k(s) \le V_k^\star(s)$ for any $k \in [K']$ and $s \in \mathcal{S}$, then*

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k^\star - V_k) \lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k - \bar{P}_i^k)V_k| + L\sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota'} + L^2S_{L(1+\epsilon)}^{\rightarrow}{}^{2}A\iota',$$

where $\iota' = O(\log(SALC_{K'}/\delta))$.

Proof. First note that, by Condition 1) and 2), for any $s \in \mathcal{S}$, $V_k^\star(s) - V_k(s) \geq 0$ and $V_k^\star(s) - V_k(s) \leq O(L)$. Thus, by Lemma 38, with probability at least $1-\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k^\star - V_k) \lesssim \underbrace{\sum_{k=1}^{K'}(V_k^\star(s_{I_k+1}^k) - V_k(s_{I_k+1}^k))^2}_{(a)} + \underbrace{\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((V_k^\star(s_i^k) - V_k(s_i^k))^2 - (P_i^k(V_k^\star - V_k))^2\right)}_{(b)} + L^2\iota,$$

where $\iota = O(\log(LC_{K'}/\delta))$.

Bounding (a) Note that, since $V_k^\star(g_k) = V_k(g_k) = 0$, we must have $(a) \leq \sum_{k=1}^{K'}\mathbb{I}\{s_{I_k+1}^k \neq g\}$. Since the event $\{s_{I_k+1}^k \neq g\}$ happens only in skip rounds, it must be that $(a) \lesssim S_{L(1+\epsilon)}^{\rightarrow}A$.

Bounding (b) Using that $V_k(s) \leq V_k^\star(s)$ for all $s \in \mathcal{S}$ (Condition 2), $(a+b)(a-b)_+$ for $a, b \geq 0$,

$$\begin{aligned}\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((V_k^\star(s_i^k) - V_k(s_i^k))^2 - (P_i^k(V_k^\star - V_k))^2\right) &\lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(V_k^\star(s_i^k) - V_k(s_i^k) - P_i^kV_k^\star + P_i^kV_k)_+ \\ &\lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(1 + P_i^kV_k - V_k(s_i^k))_+,\end{aligned}$$

where in the second inequality we used $V_k^\star(s_i^k) \leq 1 + P_i^kV_k^\star$ by definition of $V_k^\star$. Since, for all i, k, $V_k(s_i^k) \geq 1 + \widetilde{P}_i^kV_k - b_i^k - \epsilon_k$ (cf. Algorithm 4), we also have

$$\begin{aligned}\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((V_k^\star(s_i^k) - V_k(s_i^k))^2 - (P_i^k(V_k^\star - V_k))^2\right) &\lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}((P_i^k - \widetilde{P}_i^k)V_k + b_i^k + \epsilon_k)_+ \\ &= L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}((P_i^k - \bar{P}_i^k)V_k + (\bar{P}_i^k - \widetilde{P}_i^k)V_k + b_i^k + \epsilon_k)_+ \\ &\leq L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(|(P_i^k - \bar{P}_i^k)V_k| + |(\bar{P}_i^k - \widetilde{P}_i^k)V_k| + b_i^k + \epsilon_k)\end{aligned}$$

All terms but the first one are bounded in (5) and (6), which gives the following bound on (b) holding with probability at least $1-2\delta$,

$$(b) \lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}|(P_i^k - \bar{P}_i^k)V_k| + L\sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k)\iota'} + L^2{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota',$$

where $\iota' = O(\log(SALC_{K'}/\delta))$. Combining the bounds on (a) and (b) concludes the proof. $\square$

Lemma 14. *Assume that for any sequence of indicators $\mathcal{I} = \{\mathbf{1}_k\}_k$ such that $\mathbf{1}_k \in \mathcal{F}_{k-1}$, we have $R_{K',\mathcal{I}} \lesssim c_1\sqrt{K'}\log^p(K') + c_2\log^p(K')$ and $C_{K'} \lesssim c_3K' + \log^p(K')c_4$ for any $K' \in [K]$, where $c_1 \geq L$ and $c_4 \gtrsim S_{L(1+\epsilon)}^{\rightarrow}A/\epsilon$. Then, the total number rounds r_{tot} with at least one episode is of order*

$$\frac{c_1^2}{L^2}\log^{2p}\left(\frac{c_1c_4}{\epsilon}\right) + \left(\frac{c_2\epsilon}{L} + S_{L(1+\epsilon)}^{\rightarrow}A + \frac{c_1}{L}\sqrt{S_{L(1+\epsilon)}^{\rightarrow}A}\right)\log^p\left(\frac{c_1c_2c_4}{\epsilon}S_{L(1+\epsilon)}^{\rightarrow}A\right).$$

Moreover, $C_K \lesssim \frac{c_3r_{tot}}{\epsilon^2} + c_4\log^p(r_{tot}/\epsilon)$ with probability at least $1-4\delta$.

Proof. Denote by $\bar{V}_r$, $\bar{\pi}_r$ and $\bar{g}_r$ the values of $V_{\mathcal{K},g^\star}$, $\pi_{g^\star}$, and $g^\star$ used for policy evaluation in round r respectively. For any $R' \geq 1$, let K' be the total number of episodes in the first R' rounds. Denote by r'_{tot} the total number of rounds with at least one episode and r_f the number of failure rounds within the first K' episodes. The number of success rounds is at most

$S^{\rightarrow}_{L(1+\epsilon)}$ by Lemma 6 (which holds with probability $1-\delta$), and the number of skip rounds is at most $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)} A \log(C_{K'}))$ since we have a skip round only when the total number of steps or the number of visits of some state-action pair in $\mathcal{K} \times \mathcal{A}$ is doubled. Therefore, $r'_{\text{tot}} \lesssim r_f + S^{\rightarrow}_{L(1+\epsilon)} A \log(C_{K'}) \lesssim r_f + S^{\rightarrow}_{L(1+\epsilon)} A \log(K') + S^{\rightarrow}_{L(1+\epsilon)} A \log(c_4)$, where the last inequality is by assumption on $C_{K'}$.

Define $\mathcal{W} = \{r : V^{\bar{\pi}_r}_{\bar{g}_r}(s_0) > \bar{V}_r(s_0)\}$. Note that $\mathcal{W}$ includes all failure rounds with probability at least $1-\delta$. This is because, for any round $r \geq 1$ in which $V^{\bar{\pi}_r}_{\bar{g}_r}(s_0) \leq \bar{V}_r(s_0)$ and the skip round condition is not triggered, by Lemma 50 and the value of λ in Algorithm 1 in round r, we have $\widehat{\tau} \leq \bar{V}_r(s_0) + \epsilon L/2$ with probability at least $1 - \frac{\delta}{2r^2}$. This implies that a success round is triggered. A union bound over all rounds proves that all failure rounds are indeed included in $\mathcal{W} = \{r : V^{\bar{\pi}_r}_{\bar{g}_r}(s_0) > \bar{V}_r(s_0)\}$ with probability at least $1-\delta$.

Define $\mathcal{I} = \{\mathbf{1}_k\}_k$ such that $\mathbf{1}_k = \mathbb{I}\{r \in \mathcal{W}\} \in \mathcal{F}_{k-1}$ for any episode k in round r, the regret within these rounds satisfies

$$\begin{aligned} R_{K,\mathcal{I}} &\lesssim \left(\frac{c_1}{\epsilon}\sqrt{r_f + S^{\rightarrow}_{L(1+\epsilon)} A \log(K') + S^{\rightarrow}_{L(1+\epsilon)} A \log(c_4)} + c_2\right) \log^p(K') \\ &\lesssim \left(\frac{c_1}{\epsilon}\sqrt{r_f + S^{\rightarrow}_{L(1+\epsilon)} A \log(r_f/\epsilon) + S^{\rightarrow}_{L(1+\epsilon)} A \log(c_4)} + c_2\right) (\log(r_f/\epsilon) + \log(c_4))^p \end{aligned}$$

by $K = r'_{\text{tot}}\lambda \lesssim \frac{r'_{\text{tot}}}{\epsilon^2}$ (since $\lambda \lesssim 1/\epsilon^2$) and $\log(K') \lesssim \log(r_f/\epsilon) + \log(S^{\rightarrow}_{L(1+\epsilon)} A/\epsilon) \lesssim \log(r_f/\epsilon) + \log(c_4)$ by assumption on c_4. This shows that if we bound r'_{tot} we can also control $C_{K'}$.

Now we build a lower bound to $R_{K',\mathcal{I}}$. For each failure round r, let C be the total number of steps within this round and m the number of episodes within this round. By definition, the regret within this round satisfies $C - m\bar{V}_r(s_0) \geq C - \lambda\bar{V}_r(s_0) = \lambda(\widehat{\tau} - \bar{V}_r(s_0)) > \frac{\lambda\epsilon L}{2} = \Omega(L/\epsilon)$ (since $C/\lambda = \widehat{\tau} > \bar{V}_r(s_0) + \epsilon L/2$ in a failure round).

For any round $r \geq 1$, let m be its number of episodes and C be the total number of steps. By Lemma 51, $mV^{\bar{\pi}_r}_{\bar{g}_r}(s_0) \leq C + L\sqrt{m}\log^2 \frac{mLr}{\delta}$ with probability at least $1 - \frac{\delta}{2r^2}$. By a union bound, this holds simultaneously across all rounds with probability at least $1-\delta$. Then, with such probability, for each success and skip round r in $\mathcal{W}$,

$$\sum_{j=u_r}^{u'_r} \left(I_j - \bar{V}_r(s_0)\right) \geq \sum_{j=u_r}^{u'_r - 1} I_j - mV^{\bar{\pi}_r}_{\bar{g}_r}(s_0) - L \gtrsim -L\sqrt{\lambda}\log^2\left(\frac{\lambda r L}{\delta}\right) \gtrsim -\frac{L}{\epsilon},$$

where $\{u_r, \ldots, u'_r\}$ are the episodes in round r, and we lower bound the regret in the last episode by $\Omega(-L)$ since the last trajectory in a skipped round is truncated. Note that the first inequality holds since $r \in \mathcal{W}$.

Since there are at most $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)} A \log(C_{K'})) = \mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)} A (\log(r_f/\epsilon) + \log(c_4)))$ of these rounds, we have

$$\begin{aligned} \frac{Lr_f}{\epsilon} - \frac{LS^{\rightarrow}_{L(1+\epsilon)} A(\log(r_f/\epsilon) + \log(c_4))}{\epsilon} &\lesssim R_{K',\mathcal{I}} \\ &\lesssim \left(\frac{c_1}{\epsilon}\sqrt{r_f + S^{\rightarrow}_{L(1+\epsilon)} A \log(r_f/\epsilon) + S^{\rightarrow}_{L(1+\epsilon)} A \log(c_4)} + c_2\right) (\log(r_f/\epsilon) + \log(c_4))^p . \end{aligned}$$

This implies,

$$\begin{aligned} r_f &\lesssim \left(\frac{c_1}{L}\sqrt{r_f} + \frac{c_2\epsilon}{L} + S^{\rightarrow}_{L(1+\epsilon)} A + \frac{c_1}{L}\sqrt{S^{\rightarrow}_{L(1+\epsilon)} A}\right) (\log(r_f/\epsilon) + \log(c_4))^p. \\ &\lesssim \left(\underbrace{\frac{c_1}{L}}_{:=a}\sqrt{r_f} + \underbrace{\frac{c_2\epsilon}{L} + S^{\rightarrow}_{L(1+\epsilon)} A + \frac{c_1}{L}\sqrt{S^{\rightarrow}_{L(1+\epsilon)} A}}_{:=b}\right) \log(r_f \underbrace{c_4/\epsilon}_{:=c})^p. \end{aligned}$$

By Lemma 28 of (Chen et al., 2022a), a, b, c as defined above,

$$r_f \lesssim \frac{c_1^2}{L^2}\log^{2p}\left(\frac{c_1 c_4}{\epsilon}\right) + \left(\frac{c_2\epsilon}{L} + S^{\rightarrow}_{L(1+\epsilon)} A + \frac{c_1}{L}\sqrt{S^{\rightarrow}_{L(1+\epsilon)} A}\right) \log^p\left(\frac{c_1 c_2 c_4}{\epsilon} S^{\rightarrow}_{L(1+\epsilon)} A\right).$$

The proof is concluded by $r'_{\text{tot}} \lesssim r_f + S^{\rightarrow}_{L(1+\epsilon)} A \log(r_f/\epsilon) + S^{\rightarrow}_{L(1+\epsilon)} A \log(c_4)$ as showed above and setting $K' = K$ (that is, $r'_{\text{tot}} = r_{\text{tot}}$). $\square$

C.4. Proof of Theorem 1 and Theorem 2

We restate and prove the two theorems together.

Theorem 6 (Unified statement of Theorem 1 and Theorem 2)**.** *With probability at least* $1-23\delta$*, after collecting* N_{tot} *samples, Algorithm 1 outputs* $\mathcal{K}$ *and* $\{\widetilde{\pi}_g\}_{g\in\mathcal{K}}$ *such that* $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{L(1+\epsilon)}^{\rightarrow}$ *and* $V_g^{\widetilde{\pi}_g}(s_0) \le L(1+\epsilon)$ *for all* $g\in\mathcal{K}$*, where*

- $N_{tot} = \mathcal{O}\left(\frac{S_{L(1+\epsilon)}^{\rightarrow}\Gamma_{L(1+\epsilon)}AL}{\epsilon^2}\iota + \frac{{S_{L(1+\epsilon)}^{\rightarrow}}^2AL}{\epsilon}\iota + L^3{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota\right)$ *in the general case;*
- $N_{tot} = \mathcal{O}\left(\frac{S_{L(1+\epsilon)}^{\rightarrow}AL}{\epsilon^2}\iota + \frac{{S_{L(1+\epsilon)}^{\rightarrow}}^2AL}{\epsilon}\iota + L^3{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota\right)$ *with Assumption 2.*

Here $\iota = \log^8\left(\frac{SAL}{\epsilon\delta}\right)$.

Proof. By Lemma 6 and Lemma 9, with probability $1-4\delta$, the output $\mathcal{K}$ and $\{\widetilde{\pi}_g\}_{g\in\mathcal{K}}$ clearly satisfy the first statement.

Let us bound the sample complexity. Each round can be classified into one of the following cases: 1) expansion of the sets (Line 5 is true), and 2) policy evaluation is performed (from Line 12, so Line 5 is false). Note that the sample complexity of case 2 is given by C_K. We shall bound it later.

In case 1), the algorithm terminates or at least one state is added into $\mathcal{K}$. Thus, the number of rounds satisfying case 1) in each trial is at most $1+S_{L(1+\epsilon)}^{\rightarrow}$ by Lemma 6. In a round satisfying case 1), if the algorithm terminates, then no samples are collected. Otherwise, Line 8 and Line 10 are executed. Take any round r in which this happens and denote by $\mathcal{K}_r$ the set $\mathcal{K}$ at the end of round r. Note that Line 10 collects at most $O(L^2|\mathcal{K}_r|\log(Sr/\delta))$ for each $s\in\mathcal{K}_r$ and $a\in\mathcal{A}$, while Line 8 collects $O(L\log(SALr/\delta))$ samples from each state $s\in\mathcal{K}_r$ and $a\in\mathcal{A}$, so the total number of samples collected from each $s\in\mathcal{K}_r$ and $a\in\mathcal{A}$ is at most $n_r = O(L^2|\mathcal{K}_r|\log(SALr/\delta))$.

Since, by Lemma 6, at any round r, $\|V_g^{\widetilde{\pi}_g}\|_\infty \le 4L$ for each $g\in\mathcal{K}_r$, by Lemma 52, with probability $1-\delta'$ it takes no more than $8L\log(2/\delta')$ steps to reach the goal state g following $\tilde{\pi}_g$. Therefore, by setting $\delta' = \frac{\delta}{2r^2|\mathcal{K}_r||\mathcal{A}|n_r}$, with probability $1-\frac{\delta}{2r^2}$, all trajectories in round r reach the goal within $8L\log(2/\delta')$ steps. Then, by a union bound over all rounds, with probability at least $1-\delta$, the total sample complexity is $\tilde{\mathcal{O}}(L^3|\mathcal{K}_r|^2|\mathcal{A}|\log^2(SALr/\delta))$ at any round r.

Note that, among these samples, only $\tilde{\mathcal{O}}(L|\mathcal{K}_r||\mathcal{A}|\log^2(SALr/\delta))$ cumulate over rounds. This is because the sampling of Line 10 is performed only if the current counters are below the sampling requirement. Since the number of rounds in case 1) is at most $1+S_{L(1+\epsilon)}^{\rightarrow}$ and the total number of rounds R performed by the algorithm satisfies $R\le r_{\text{tot}} + S_{L(1+\epsilon)}^{\rightarrow}+1$ (by summing the rounds in both cases) and $|\mathcal{K}_r|\le S_{L(1+\epsilon)}^{\rightarrow}$ by Lemma 6, we have that Line 10 contributes to at most $\tilde{\mathcal{O}}(L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\log^2(SALr_{\text{tot}}/\delta))$ sample complexity and the total sample complexity of Case 1) is thus $\tilde{\mathcal{O}}(L^3{S_{L(1+\epsilon)}^{\rightarrow}}^2A\log^2(SALr_{\text{tot}}/\delta))$.

We now conclude the sample complexity proof depending on whether Assumption 2 is considered or not.

Without Assumption 2 Plugging the regret bound of Lemma 11 into Lemma 14, using $p=2$, $c_1 = L\log(SAL/\delta)^2\sqrt{S_{L(1+\epsilon)}^{\rightarrow}\Gamma_{L(1+\epsilon)}A}$, $c_2 = L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\log(SAL/\delta)^3$, $c_3 = L$, $c_4 = L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\log(SAL/\delta)^3/\epsilon$,

$$
\begin{aligned}
r_{\text{tot}} &\lesssim \left(\log(SAL/\delta)^4S_{L(1+\epsilon)}^{\rightarrow}\Gamma_{L(1+\epsilon)}A + {S_{L(1+\epsilon)}^{\rightarrow}}^2A\log(SAL/\delta)^3\epsilon + \log(SAL/\delta)^2S_{L(1+\epsilon)}^{\rightarrow}\sqrt{\Gamma_{L(1+\epsilon)}}A\right)\log^4\left(\frac{SAL}{\epsilon}\right) \\
&\lesssim \left(S_{L(1+\epsilon)}^{\rightarrow}\Gamma_{L(1+\epsilon)}A + {S_{L(1+\epsilon)}^{\rightarrow}}^2A\epsilon\right)\log^8\left(\frac{SAL}{\epsilon\delta}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
C_K &\lesssim \frac{L}{\epsilon^2}\left(S_{L(1+\epsilon)}^{\rightarrow}\Gamma_{L(1+\epsilon)}A + {S_{L(1+\epsilon)}^{\rightarrow}}^2A\epsilon\right)\log^8\left(\frac{SAL}{\epsilon\delta}\right) + \frac{L{S_{L(1+\epsilon)}^{\rightarrow}}^2A}{\epsilon}\log^5\left(\frac{SAL}{\epsilon\delta}\right), \\
&\lesssim \left(\frac{S_{L(1+\epsilon)}^{\rightarrow}\Gamma_{L(1+\epsilon)}AL}{\epsilon^2} + \frac{{S_{L(1+\epsilon)}^{\rightarrow}}^2AL}{\epsilon}\right)\log^8\left(\frac{SAL}{\epsilon\delta}\right).
\end{aligned}
$$

Thus, the total sample complexity of the algorithm (which is given by C_K plus the sample complexity of case 1) is

$$\left(\frac{S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}AL}{\epsilon^2} + \frac{{S^{\rightarrow}_{L(1+\epsilon)}}^2 AL}{\epsilon} + L^3 {S^{\rightarrow}_{L(1+\epsilon)}}^2 |\mathcal{A}|\right) \log^8\left(\frac{SAL}{\epsilon\delta}\right).$$

With Assumption 2 Plugging the regret bound of Lemma 12 into Lemma 14, using $p = 2$, $c_1 = L\log(SAL/\delta)^2\sqrt{S^{\rightarrow}_{L(1+\epsilon)}A}$, $c_2 = L{S^{\rightarrow}_{L(1+\epsilon)}}^2 A\log(SAL/\delta)^3$, $c_3 = L$, $c_4 = L{S^{\rightarrow}_{L(1+\epsilon)}}^2 A\log(SAL/\delta)^3/\epsilon$,

$$\begin{aligned} r_{\text{tot}} &\lesssim \left(\log(SAL/\delta)^4 S^{\rightarrow}_{L(1+\epsilon)}A + {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\log(SAL/\delta)^3\epsilon + \log(SAL/\delta)^2 S^{\rightarrow}_{L(1+\epsilon)}\sqrt{\Gamma_{L(1+\epsilon)}}A\right)\log^4\left(\frac{SAL}{\epsilon}\right) \\ &\lesssim \left(S^{\rightarrow}_{L(1+\epsilon)}A + {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\epsilon\right)\log^8\left(\frac{SAL}{\epsilon\delta}\right) \end{aligned}$$

and

$$\begin{aligned} C_K &\lesssim \frac{L}{\epsilon^2}\left(S^{\rightarrow}_{L(1+\epsilon)}A + {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\epsilon\right)\log^8\left(\frac{SAL}{\epsilon\delta}\right) + \frac{L{S^{\rightarrow}_{L(1+\epsilon)}}^2 A}{\epsilon}\log^5\left(\frac{SAL}{\epsilon\delta}\right), \\ &\lesssim \left(\frac{S^{\rightarrow}_{L(1+\epsilon)}AL}{\epsilon^2} + \frac{{S^{\rightarrow}_{L(1+\epsilon)}}^2 AL}{\epsilon}\right)\log^8\left(\frac{SAL}{\epsilon\delta}\right). \end{aligned}$$

Thus, the total sample complexity of the algorithm (which is given by C_K plus the sample complexity of case 1) is

$$\left(\frac{S^{\rightarrow}_{L(1+\epsilon)}AL}{\epsilon^2} + \frac{{S^{\rightarrow}_{L(1+\epsilon)}}^2 AL}{\epsilon} + L^3 {S^{\rightarrow}_{L(1+\epsilon)}}^2 |\mathcal{A}|\right) \log^8\left(\frac{SAL}{\epsilon\delta}\right).$$

A union bound over the events of adopted lemmas (Lemma 6, Lemma 9, Lemma 6 of (Rosenberg & Mansour, 2021), Lemma 14, and Lemma 11 without Assumption 2 or Lemma 12 with Assumption 2) yields the result with probability at least $1 - 23\delta$. $\square$

Algorithm 5: Improved Layer-Aware State Discovery (LASD$^+$)

Input: $L \geq 1$, $\epsilon \in (0, 1]$, and $\delta \in (0, 1)$.

1 Let $\tau \leftarrow 1$, $\mathfrak{N} = \{2^j\}_{j \geq 0}$, $z \leftarrow 2$.
2 **while** *True* **do**
3 Let $\mathcal{K} \leftarrow \varnothing, \mathcal{U} \leftarrow \varnothing, \mathcal{K}' \leftarrow \{s_0\}$, $\Pi_{\mathcal{K}} = \{\widetilde{\pi}_{s_0}$ a random policy$\}$, $\mathbf{N}(\cdot, \cdot) \leftarrow 0, \mathbf{N}(\cdot, \cdot, \cdot) \leftarrow 0$, $n_{\min} \leftarrow 1$, $k \leftarrow 0$.
4 **for** *round* $r = 1, \ldots$ **do**
5 **if** $|\mathcal{K} \cup \mathcal{K}'| \geq z$ **then** $z \leftarrow 2|\mathcal{K} \cup \mathcal{K}'|$, $\tau \overset{+}{\leftarrow} 1$, and return to Line 2.
6 $\epsilon_{\mathrm{VI}} \leftarrow 1/\max\{16, \sum_{s,a} \mathbf{N}(s, a)\}$.
7 Let $g^\star = \operatorname{argmin}_{g \in \mathcal{U}} \left\{V_{\mathcal{K},g}(s_0)\right\}$ where $(Q_{\mathcal{K},g}, V_{\mathcal{K},g}, \pi_g) = \mathrm{VISGO}(\mathcal{K}, g, \epsilon_{\mathrm{VI}}, \mathbf{N}, \frac{\delta}{4\tau^2 z^4 AL})$ (see Algorithm 4).
8 **if** *$g^\star$ does not exist or $V_{\mathcal{K},g^\star}(s_0) > L$* **then**
`/* Expand or Terminate */`
9 **if** $\mathcal{K}' = \varnothing$ **then return** $\mathcal{K}$ and $\Pi_{\mathcal{K}}$.
10 Set $\mathcal{K} \leftarrow \mathcal{K} \cup \mathcal{K}'$, $\mathcal{K}' = \varnothing, \mathcal{U} = \varnothing$.
11 $\mathcal{U} \leftarrow$`ComputeU(`$\mathcal{K}$, $\Pi_{\mathcal{K}}$, $\frac{\delta}{4\tau^2 r^2}$`)`.
12 **else if** $\mathrm{RTEST}(\Pi_{\mathcal{K}}, \pi_{g^\star}, g^\star, \frac{\delta}{4(\tau r)^2}) = $ ***False*** *(see Algorithm 7)* **then**
13 $n_{\min} \leftarrow 2n_{\min}$.
14 $(\mathbf{N}, \_) \leftarrow \mathrm{EXPLORE}(\mathcal{K}, \Pi_{\mathcal{K}}, \mathbf{N}, n_{\min})$ (see Algorithm 6).
15 **else**
`/* Policy evaluation */`
16 Let $\widehat{\tau} \leftarrow 0$, $\lambda \leftarrow N_{\mathrm{DEV}}(32L, \frac{\epsilon}{256}, \frac{\delta}{2r^2}) \lesssim \frac{1}{\epsilon^2}\log^4(\frac{Lr}{\epsilon\delta})$ (defined in Lemma 50).
17 **for** $j = 1, \ldots, \lambda$ **do**
18 $k \overset{+}{\leftarrow} 1$, $i \leftarrow 1$, and reset to $s_1^k \leftarrow s_0$ by taking action RESET.
19 **while** $s_i^k \neq g^\star$ **do**
20 Take $a_i^k = \pi_{g^\star}(s_i^k)$, and transits to s_{i+1}^k. Increase $\mathbf{N}(s_i^k, a_i^k)$, $\mathbf{N}(s_i^k, a_i^k, s_{i+1}^k)$, and i by 1.
21 **if** *$\sum_{s,a} \mathbf{N}(s, a) \in \mathfrak{N}$ or ($s_i^k \in \mathcal{K}$ and $\mathbf{N}(s_i^k, a_i^k) \in \mathfrak{N}$)* **then** return to Line 4 (skip round).
22 Set $\widehat{\tau} \overset{+}{\leftarrow} \frac{c(s_i^k, a_i^k)}{\lambda}$.
23 **if** $\widehat{\tau} > V_{\mathcal{K},g^\star}(s_0) + \epsilon L/2$ **then** return to Line 4 (failure round).
24 $\mathcal{K}' \leftarrow \mathcal{K}' \cup \{g^\star\}$, $\mathcal{U} \leftarrow \mathcal{U} \setminus \{g^\star\}$, $\Pi_{\mathcal{K}} = \Pi_{\mathcal{K}} \cup \{\widetilde{\pi}_{g^\star} := \pi_{g^\star}\}$ (success round).
25 **Procedure** `ComputeU(`$\mathcal{X}$, $\Pi_{\mathcal{X}}$, δ`)`
26 $(\_, \mathcal{U}') \leftarrow \mathrm{EXPLORE}(\mathcal{X}, \Pi_{\mathcal{X}}, 0, 2L\log\frac{4LA|\mathcal{X}|}{\delta})$ (see Algorithm 6).
27 $(\mathbf{N}', \_) \leftarrow \mathrm{EXPLORE}(\mathcal{X}, \Pi_{\mathcal{X}}, 0, N_1(|\mathcal{X}|, \frac{\delta}{4|\mathcal{U}'|}))$ where N_1 is defined in Lemma 4.
28 Let $\mathcal{U} = \{g \in \mathcal{U}' : V'_{\mathcal{X},g}(s_0) \leq L\}$ where $(\_, V'_{\mathcal{X},g}, \pi'_g) = \mathrm{VISGO}(\mathcal{X}, g, \frac{1}{16}, \mathbf{N}', \frac{\delta}{4|\mathcal{U}'|})$.
29 **return** $\mathcal{U}$

D. Analysis of Algorithm 5

Notation Define $\mathcal{N}(\mathcal{K}, p) = \{s' \notin \mathcal{K} : P(s'|s, a) \geq p$ for some $(s, a) \in \mathcal{K} \times \mathcal{A}\}$. Fix any ordering $\mathcal{O}_L^{\rightarrow} = (s_1, \ldots, s_n)$ of states in $\mathcal{S}_L^{\rightarrow}$ such that it can be partitioned into J (defined in Lemma 1) segments with states in the j-th segment belonging to $\mathcal{K}_j^\star \setminus \mathcal{K}_{j-1}^\star$. For an arbitrary $z \in \mathbb{N}_+$, also define $\{\mathcal{K}_{z,j}^\star\}_j$, such that $\mathcal{K}_{z,j}^\star = \mathcal{K}_j^\star$ when $|\mathcal{K}_j^\star| < z$, and $\mathcal{K}_{z,j}^\star = \{s_1, \ldots, s_z\}$ when $|\mathcal{K}_j^\star| \geq z$. Therefore, $\mathcal{K}_{z,z}^\star = (s_1, \ldots, s_z)$ (the first z elements of $\mathcal{O}_L^{\rightarrow}$) or $\mathcal{S}_L^{\rightarrow}$ by definition. Define $\mathcal{U}_z^\star = \mathcal{T}_{2L}(\mathcal{K}_{z,z}^\star)$. Clearly, $\mathcal{U}_z^\star \subseteq \{s' \in \mathcal{S} : \exists s \in \mathcal{K}_{z,z}^\star, a \in \mathcal{A}, P(s'|s, a) \geq \frac{1}{2L}\}$, and thus $|\mathcal{U}_z^\star| \leq 2zAL$.

D.1. Proof of Theorem 3

Proof. We condition on the events of Lemma 20, Lemma 28, and Lemma 23, which happen with probability at least $1 - 7\delta$. By the events of Lemma 23 and Lemma 20, the output $\mathcal{K}$ and $\Pi_{\mathcal{K}} = \{\widetilde{\pi}_g\}_{g \in \mathcal{K}}$ clearly satisfy the statement. By Lemma 16, there are at most $\mathcal{O}(\log S_{L(1+\epsilon)}^{\rightarrow})$ trials. Thus, it suffices to bound the number of samples used in each trial. Define $\iota = \log\frac{LS_{L(1+\epsilon)}^{\rightarrow}A}{\delta\epsilon}$. Each round in a trial can be classified into one of the following cases: 1) Line 8 is verified, 2) Line 12 is verified, and 3) policy evaluation is performed (Line 16). In case 1), the algorithm terminates or at least one state is added into $\mathcal{K}$ (Line 9). Thus, the number of rounds satisfying case 1) in each trial is at most

$1 + S^{\rightarrow}_{L(1+\epsilon)}$ by Lemma 23. By Lemma 15 and the update rule of $n_{\min}$, the number of rounds satisfying case 2) is of order $\mathcal{O}(\log(LS^{\rightarrow}_{L(1+\epsilon)}))$. By Lemma 19 and Lemma 17, with probability at least $1-8\delta$, the total number of rounds satisfying case 3) is of order $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A\iota^6 + {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\epsilon\iota^6)$. So the total number of rounds in each trial is at most $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A\iota^6 + {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\epsilon\iota^6)$.

Now it suffices to bound the number of samples collected in a round satisfying each of the cases above in a trial. In a round satisfying case 1), if the algorithm terminates, then no samples are collected. Otherwise, `ComputeU` is called, and $\mathcal{O}(L^3 {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\iota^2)$ samples are collected with probability at least $1-\delta$ by Lemma 27 (Line 11 and a union bound over all trials and rounds). In a round satisfying case 2), with probability at least $1-4\delta$, $\mathcal{O}(LS^{\rightarrow}_{L(1+\epsilon)}\iota^2)$ samples are collected in performing RTEST by Lemma 20 and Lemma 29 (Line 12 and a union bound over all trials and rounds), and $\mathcal{O}(L^3 {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\iota^2)$ samples are collected in executing EXPLORE by Lemma 15 and Lemma 30. In a round satisfying case 3), with probability at leat $1-\delta$, $\mathcal{O}(LS^{\rightarrow}_{L(1+\epsilon)}\iota^2)$ samples are collected in performing RTEST similar to that of case 2), and $\mathcal{O}(L\iota^5/\epsilon^2)$ samples are collected by the value of λ and the fact that $\pi_{g^\star}$ passes the test in Line 12 (Lemma 29 and a union bound over all trials and rounds). Thus, the total sample complexity is

$$
\begin{aligned}
&\sum_{i=1}^{3} [\text{\#rounds satisfying case } i]\cdot[\text{\#samples in a round satisfying case } i]\cdot\iota \\
&\lesssim S^{\rightarrow}_{L(1+\epsilon)}\cdot L^3 {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\iota^3 + L^3{S^{\rightarrow}_{L(1+\epsilon)}}^2 A\iota^4 + (S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A + {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\epsilon)\cdot\left(\frac{L}{\epsilon^2} + LS^{\rightarrow}_{L(1+\epsilon)}\right)\iota^{12} \\
&\lesssim \left(\frac{LS^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A}{\epsilon^2} + \frac{L{S^{\rightarrow}_{L(1+\epsilon)}}^2 A\epsilon}{\epsilon} + L^3{S^{\rightarrow}_{L(1+\epsilon)}}^3 A\right)\iota^{12}.
\end{aligned}
$$

This completes the proof. To prove the second statement, we can simply follow the proof above except that we involve Lemma 18 instead of Lemma 17 when applying Lemma 19 to bound the total number of rounds satisfying case 3), which holds with probability at least $1-20\delta$. □

Lemma 15. *With probability at least $1-2\delta$, if the events of Lemma 23 and Lemma 24 hold, then $n_{\min} \lesssim L^2 S^{\rightarrow}_{L(1+\epsilon)}\log S^{\rightarrow}_{L(1+\epsilon)}$ throughout the execution of Algorithm 5.*

Proof. In any trial τ, when $n_{\min} \geq N_0^{\rightarrow}(\frac{\delta}{4\tau^2 z^4 AL})$ (defined in Lemma 3), we have with probability at least $1-\frac{\delta}{2\tau^2}$, $\left\|V^{\pi_{g^\star}}_{g^\star}\right\|_\infty \leq 2\left\|V_{\mathcal{K},g^\star}\right\|_\infty \leq 2(1+V_{\mathcal{K},g^\star}(s_0)) \leq 4L$ in any round such that $g^\star$ exists and $V_{\mathcal{K},g^\star}(s_0)\leq L$. This implies that with probability at least $1-\sum_{r=1}^{\infty}\frac{\delta}{4\tau^2 r^2} \geq 1-\frac{\delta}{2\tau^2}$, the condition of Line 12 is always false by Lemma 29, and the value of $n_{\min}$ will no longer change within this trial. A union bound over all trials and noting the update rule of $n_{\min}$ completes the proof. □

Lemma 16. *Conditioned on the event of Lemma 23, we have $z \leq 2S^{\rightarrow}_{L(1+\epsilon)}+2$ and $\tau \leq 1+\log_2(S^{\rightarrow}_{L(1+\epsilon)}+1)$ throughout the execution of Algorithm 5.*

Proof. The proof of Lemma 23 shows that $s \notin \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ will never be added to $\mathcal{K}'$, which implies $\mathcal{K}\cup\mathcal{K}' \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ throughtout the execution of Algorithm 5. Thus, when $z \geq S^{\rightarrow}_{L(1+\epsilon)}+1$, z will not be updated again. Then, the statement is proved by the update rule of z and τ. □

D.2. Lemmas for Policy Evaluation

Notation Let g_k, $\mathcal{K}_k$, V_k, Q_k, $V^\star_k$ be the values of $g^\star$, $\mathcal{K}$, $V_{\mathcal{K},g^\star}$, $Q_{\mathcal{K},g^\star}$, and $V^\star_{\mathcal{K},g^\star}$ in episode k respectively. Denote by I_k the number of steps in episode k. Note that $I_k<\infty$ with probability 1 by Line 21, and $s^k_{I_k+1}\neq g_k$ only when a skip round is triggered in episode k. Denote by $\mathcal{F}_k$ the σ-algebra of events up to episode k. Define K as the total number of episodes throughout the execution of Algorithm 5. For any sequence of indicators $\mathcal{I}=\{\mathbf{1}_k\}_k$ and $K'\leq K$, define $R_{K',\mathcal{I}}=\sum_{k=1}^{K'}(I_k - V_k(s_0))\mathbf{1}_k$ and $C_{K'}=\sum_{k=1}^{K'} I_k$. Define $P^k_i = P_{s^k_i,a^k_i}$. In episode k, when $s^k_i\in\mathcal{K}$, denote by $\bar{P}^k_i$, $\widetilde{P}^k_i$, $\mathbf{N}^k_i$, b^k_i the values of $\bar{P}_{s^k_i,a^k_i}$, $\widetilde{P}_{s^k_i,a^k_i}$, $n^+(s^k_i,a^k_i)$, and $b^{(l)}(s^k_i,a^k_i)$, where $\bar{P}$, n^+, $b^{(l)}$ are used in Algorithm 4 to compute V_k and l is the final value of i in Algorithm 4; when $s^k_i\notin\mathcal{K}$, define $\bar{P}^k_i=\mathbb{I}_{s_0}$, $\mathbf{N}^k_i=\infty$, and $b^k_i=0$. Also define ϵ_k as the value of ϵ_{VI} used in Algorithm 4 to compute V_k.

Lemma 17. *With probability at least $1-5\delta$, if the events of Lemma 23 and Lemma 24 hold, then in any trial, for any sequence of indicators $\mathcal{I}=\{\mathbf{1}_k\}_k$ with $\mathbf{1}_k\in\mathcal{F}_{k-1}$, we have $R_{K',\mathcal{I}}\lesssim\sqrt{S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}AL^2K'\iota}+LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota$ for any $K'\leq K$, where $\iota=\log^2\frac{LS^{\rightarrow}_{L(1+\epsilon)}AK'}{\delta}$.*

Proof. Note that by Lemma 42,

$$\begin{aligned}\sum_{k=1}^{K'}\left(I_k-V_k(s_0)\right)\mathbf{1}_k&\leq\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(1+V_k(s^k_{i+1})-V_k(s^k_i)\right)\mathbf{1}_k\\&\lesssim\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((\mathbb{I}_{s^k_{i+1}}-P^k_i)V_k+(P^k_i-\bar{P}^k_i)V_k+b^k_i+\epsilon_k\right)\mathbf{1}_k.\end{aligned}$$

We bound the sums above separately. By Lemma 55 and $\|V_k\|_\infty\leq 2L$, with probability at least $1-\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(\mathbb{I}_{s^k_{i+1}}-P^k_i)V_k\mathbf{1}_k\lesssim\sqrt{\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P^k_i,V_k)\log\frac{LC_{K'}}{\delta}}+L\log\frac{LC_{K'}}{\delta}.$$

By Lemma 46, $\mathcal{K}_k\in\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ (Lemma 23), $g_k\in\bar{\mathcal{U}}\setminus\mathcal{K}_k$ (Lemma 24), Cauchy-Schwarz inequality, and Lemma 40, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(P^k_i-\bar{P}^k_i)V_k\mathbf{1}_k&\lesssim\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbf{1}_k\sqrt{\frac{\Gamma_{L(1+\epsilon)}\mathbb{V}(P^k_i,V_k)\iota'}{\mathbf{N}^k_i}}+\frac{LS^{\rightarrow}_{L(1+\epsilon)}\iota'}{\mathbf{N}^k_i}\\&\qquad(\mathbf{N}^k_i=\infty\text{ when }s^k_i\notin\mathcal{K}_k\text{ and }\iota'=\log\tfrac{S^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta})\\&\lesssim\sqrt{S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P^k_i,V_k)\iota'}+LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota'.\\&\qquad(\iota'=\log\tfrac{S^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta}\log(C_{K'}))\end{aligned}$$

Finally, by Lemma 39 and Lemma 41, with probability at least $1-\delta$,

$$\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(b^k_i+\epsilon_k)\mathbf{1}_k\lesssim\sqrt{S^{\rightarrow}_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P^k_i,V_k)\iota'}+LS^{\rightarrow}_{L(1+\epsilon)}{}^{1.5}A\iota'.\qquad(\iota'=\log\tfrac{S^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta})$$

Plugging these back, we have with probability at least $1-2\delta$,

$$\begin{aligned}\sum_{k=1}^{K'}\left(I_k-V_k(s_0)\right)\mathbf{1}_k&\lesssim\sqrt{S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P^k_i,V_k)\iota'}+LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota'\\&\lesssim\sqrt{S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)}ALC_{K'}\iota'}+LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota',\end{aligned}\tag{7}$$

where $\iota'=\log\frac{LS^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta}\log(C_{K'})$ and in the last step we apply Lemma 36. Now assuming $\mathbf{1}_k=1$ for all k and solving a "quadratic" inequality (Lemma 47) w.r.t. $C_{K'}$, we have

$$C_{K'}\lesssim\sum_{k=1}^{K'}V_k(s_0)+LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota'\lesssim LK'+LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota'.\qquad(\iota'=\log^2\tfrac{LS^{\rightarrow}_{L(1+\epsilon)}AK'}{\delta})$$

Plugging this back to Eq. (7) completes the proof. □

Lemma 18. *With Assumption 2, with probability at least $1-12\delta$, if the events of Lemma 28, Lemma 16, Lemma 25, and Lemma 26 hold, in any trial, for any sequence of indicators $\mathcal{I}=\{\mathbf{1}_k\}_k$ with $\mathbf{1}_k\in\mathcal{F}_{k-1}$, we have $R_{K',\mathcal{I}}\lesssim L\sqrt{S^{\rightarrow}_{L(1+\epsilon)}AK'\iota}+LS^{\rightarrow}_{L(1+\epsilon)}{}^2A\iota$ for any $K'\leq K$, where $\iota=\log^2\frac{LS^{\rightarrow}_{L(1+\epsilon)}AK'}{\delta}$.*

Proof. Note that with Assumption 2 and by Lemma 25 and Lemma 26, in any episode, $\mathcal{K} = \mathcal{K}^\star_j$ for some $j \le z$ and $g^\star \in \mathcal{U}^\star_z$. Thus by Lemma 54 and a union bound over $\{V^\star_{\mathcal{K}^\star_{z,j},g}\}_{j\in[z],g\in\mathcal{U}^\star_z}$ and $(s,a) \in \mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \times \mathcal{A}$, we have with probability at least $1-\delta$,

$$(P^k_i - \bar{P}^k_i)V^\star_k \lesssim \sqrt{\frac{\mathbb{V}(P^k_i, V^\star_k)\iota'}{\mathbf{N}^k_i}} + \frac{L\iota'}{\mathbf{N}^k_i}, \tag{8}$$

where $\iota' = \log \frac{S^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta}$. Thus, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{k=1}^{K'}(I_k - V_k(s_0))\,\mathbf{1}_k \le \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(1 + V_k(s^k_{i+1}) - V_k(s^k_i)\right)\mathbf{1}_k \\
&\lesssim \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((\mathbb{I}_{s^k_{i+1}} - P^k_i)V_k + (P^k_i - \bar{P}^k_i)V_k + b^k_i + \epsilon_k\right)\mathbf{1}_k && \text{(Lemma 42)} \\
&\lesssim \sqrt{\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P^k_i, V_k)\log\frac{LC_{K'}}{\delta}} + L\log\frac{LC_{K'}}{\delta} + \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((P^k_i - \bar{P}^k_i)V^\star_k\mathbf{1}_k + (P^k_i - \bar{P}^k_i)(V_k - V^\star_k)\mathbf{1}_k + b^k_i\right),
\end{aligned}$$

where the last step is by Lemma 55 and Lemma 41. Note that by Eq. (8), Lemma 46, and $\|V^\star_k\|_\infty \le 2L+1$ by Lemma 28 and Lemma 44, with probability at least $1-2\delta$,

$$\begin{aligned}
&\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((P^k_i - \bar{P}^k_i)V^\star_k\mathbf{1}_k + (P^k_i - \bar{P}^k_i)(V_k - V^\star_k)\mathbf{1}_k + b^k_i\right) \\
&\lesssim \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(\sqrt{\frac{\mathbb{V}(P^k_i, V^\star_k)\iota'}{\mathbf{N}^k_i}} + \sqrt{\frac{\Gamma_{L(1+\epsilon)}\mathbb{V}(P^k_i, V_k - V^\star_k)\iota'}{\mathbf{N}^k_i}} + \frac{L\Gamma_{L(1+\epsilon)}\iota'}{\mathbf{N}^k_i} + b^k_i\right) && \left(\iota' = \log\frac{S^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta}\right) \\
&\lesssim \sqrt{S^{\rightarrow}_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P^k_i, V_k)\iota'} + \sqrt{{S^{\rightarrow}_{L(1+\epsilon)}}^2A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P^k_i, V_k - V^\star_k)\iota'} + L{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota'. \\
& && \left(\iota' = \log^2\frac{S^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta}\right)
\end{aligned}$$

where the last step is by Lemma 40, Cauchy-Schwarz inequality, $\mathrm{VAR}[X+Y] \le 2(\mathrm{VAR}[X] + \mathrm{VAR}[Y])$, and Lemma 39. Plugging this back, applying Lemma 37 with Lemma 2 on $\{V^\star_{\mathcal{K}^\star_j,g}\}_{j\in[z],g\in\mathcal{U}^\star_z\setminus\mathcal{K}^\star_j}$ (where all $V^\star_k$ lies in), Lemma 25, and Lemma 26, and then applying AM-GM inequality, we have with probability at least $1-8\delta$,

$$\begin{aligned}
\sum_{k=1}^{K'}(I_k - V_k(s_0))\,\mathbf{1}_k &\lesssim \sqrt{S^{\rightarrow}_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P^k_i, V_k)\iota'} + L{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota' \\
&\lesssim \sqrt{LS^{\rightarrow}_{L(1+\epsilon)}AC_{K'}\iota'} + L{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota', && \text{(Lemma 36)}
\end{aligned}$$

where $\iota' = \log^2\frac{LS^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta}$. Now assuming $\mathbf{1}_k = 1$ for all k and solving a "quadratic" inequality (Lemma 47), we have

$$C_{K'} \lesssim \sum_{k=1}^{K'}V_k(s_0) + L{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota' \le LK' + L{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota'. \qquad \left(\iota' = \log^2\frac{LS^{\rightarrow}_{L(1+\epsilon)}AK'}{\delta}\right)$$

Plugging this back completes the proof. ☐

Lemma 19. *In any trial, with probability at least $1-8\delta$, if for any sequence of indicators $\mathcal{I} = \{\mathbf{1}_k\}_k$ with $\mathbf{1}_k \in \mathcal{F}_{k-1}$, we have $R_{K',\mathcal{I}} \lesssim c_1\sqrt{K'\log^p(c_3K')} + c_2\log^p(c_3K')$ with $c_1, c_2 \ge 1$, and $c_3 = \frac{LS^{\rightarrow}_{L(1+\epsilon)}A}{\delta}$ for any $K' \le K$, then the total number of rounds with at least one epsiode is of order $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)}A\iota^4 + \frac{c_1^2}{L^2}\iota^{p+4} + c_2\epsilon\iota^p/L)$, where $\iota = \log\frac{c_1c_2c_3}{\epsilon\delta}$.*

Proof. For any $R' \geq 1$, let K' be the total number of episodes in the first R' rounds. Denote by r_{tot} the total number of rounds with at least one episode, and r_f the number of failure rounds in the first R' rounds. First note that by $V_k(s_0) \leq L$ (Line 8) and setting $\mathbf{1}_k = 1$, the regret guarantee in the assumption gives $C_{K'} \lesssim LK' + c_1\sqrt{K' \log^p(c_3K')} + c_2 \log^p(c_3K')$, which gives $\log(C_{K'}) \lesssim \log(c_1c_2c_3K')$. Moreover, $K' \lesssim \frac{r_{\text{tot}}}{\epsilon^2} \log^4 \frac{Lr_{\text{tot}}}{\epsilon\delta}$ by the value of λ in each round (Line 16). Thus, $\log(C_{K'}) \lesssim \log \frac{c_1c_2c_3r_{\text{tot}}}{\epsilon\delta}$ and $\log(c_3K') \lesssim \log \frac{c_1c_2c_3r_{\text{tot}}}{\epsilon\delta}$.

Fixed a trial, denote by $\bar{V}_r$, $\bar{\pi}_r$ and $\bar{g}_r$ the values of $V_{\mathcal{K},g^\star}$, $\pi_{g^\star}$, and $g^\star$ used for policy evaluation in round r respectively. It is clear that in the first R' rounds, the number of success round is at most $S^{\rightarrow}_{L(1+\epsilon)}$ by Lemma 23, and the number of skip rounds is at most $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)}A\log(C_{K'}))$ since we have a skip round only when the total number of steps or the number of visits of some state-action pair in $\mathcal{K} \times \mathcal{A}$ is doubled. Therefore, $r_{\text{tot}} \lesssim r_f + S^{\rightarrow}_{L(1+\epsilon)}A\log(C_{K'}) \lesssim r_f + S^{\rightarrow}_{L(1+\epsilon)}A\log \frac{c_1c_2c_3r_{\text{tot}}}{\epsilon\delta}$. By Lemma 47, we have $r_{\text{tot}} \lesssim r_f + S^{\rightarrow}_{L(1+\epsilon)}A\log \frac{c_1c_2c_3r_f}{\epsilon\delta}$. Now define $\iota(r_f) = \log \frac{c_1c_2c_3r_f}{\epsilon\delta}$. It remains to bound r_f. Define $\mathcal{W} = \{r : V^{\bar{\pi}_r}_{\bar{g}_r}(s_0) > \bar{V}_r(s_0)\}$. Note that $\mathcal{W}$ includes all failure rounds with probability at least $1-\delta$, since when $V^{\bar{\pi}_r}_{\bar{g}_r}(s_0) \leq \bar{V}_r(s_0)$ and r is not a skip round, by Lemma 50 and the value of λ in round r we have $\widehat{\tau} \leq \bar{V}_r(s_0) + \epsilon L/2$ in round r. Define $\mathcal{I} = \{\mathbf{1}_k\}_k$ such that $\mathbf{1}_k = \mathbb{I}\{r \in \mathcal{W}\} \in \mathcal{F}_{k-1}$ for any episode k in round r, the regret within these rounds satisfies $R_{K',\mathcal{I}} \lesssim \frac{c_1}{\epsilon}\sqrt{r_f + S^{\rightarrow}_{L(1+\epsilon)}A} + c_2$.

$$\begin{aligned}
R_{K',\mathcal{I}} &\lesssim c_1\sqrt{K'\log^p(c_3K')} + c_2\log^p(c_3K') \lesssim \frac{c_1}{\epsilon}\sqrt{(r_f + S^{\rightarrow}_{L(1+\epsilon)}A\iota(r_f))\iota(r_f)^{p+4}} + c_2\iota(r_f)^p \\
&\lesssim \frac{c_1}{\epsilon}\sqrt{r_f\iota(r_f)^{p+4}} + \frac{c_1^2\iota(r_f)^{p+4}}{L\epsilon} + \frac{LS^{\rightarrow}_{L(1+\epsilon)}A\iota(r_f)}{\epsilon} + c_2\iota(r_f)^p. && \text{(AM-GM inequality)}
\end{aligned}$$

For each failure round r, let C be the total cost within this round and m the number of episodes within this round. By definition, regret within this round satisfies $C - mV_{\mathcal{K},g^\star}(s_0) \geq C - \lambda V_{\mathcal{K},g^\star}(s_0) = \lambda(\widehat{\tau} - V_{\mathcal{K},g^\star}(s_0)) > \frac{\lambda\epsilon L}{2} = \Omega(L/\epsilon)$. By Lemma 51, with probability at least $1-\delta$, for each success and skip round r in $\mathcal{W}$ ($V^{\bar{\pi}_r}_{\bar{g}_r}(s_0) > \bar{V}_r(s_0)$),

$$\sum_{j=u_r}^{u'_r}\left(I_j - \bar{V}_r(s_0)\right) \gtrsim \sum_{j=u_r}^{u'_r-1}\left(I_j - V^{\bar{\pi}_r}_{\bar{g}_r}(s_0)\right) - L \gtrsim -L\sqrt{\lambda}\log^2\frac{L\lambda}{\delta} = -\frac{L}{\epsilon}\log^4\frac{Lr}{\delta\epsilon},$$

where $\{u_r, \ldots, u'_r\}$ are the episodes in round r, and we lower bound the regret in the last episode by $\Omega(-L)$ since the last trajectory in a skipped round is truncated. Since there are at most $\tilde{\mathcal{O}}(S^{\rightarrow}_{L(1+\epsilon)}A)$ these rounds, we have

$$\frac{Lr_f}{\epsilon} - \frac{LS^{\rightarrow}_{L(1+\epsilon)}A}{\epsilon}\log^4\frac{Lr_f}{\epsilon\delta} \lesssim \frac{c_1}{\epsilon}\sqrt{r_f\iota(r_f)^{p+4}} + \frac{c_1^2\iota(r_f)^{p+4}}{L\epsilon} + \frac{LS^{\rightarrow}_{L(1+\epsilon)}A\iota(r_f)}{\epsilon} + c_2\iota(r_f)^p.$$

This gives $r_f \lesssim S^{\rightarrow}_{L(1+\epsilon)}A\iota^4 + \frac{c_1^2}{L^2}\iota^{p+4} + c_2\epsilon\iota^p/L$, where $\iota = \log\frac{c_1c_2c_3}{\epsilon\delta}$. Setting R' to be the total number rounds completes the proof. ☐

Lemma 20. *With probability at least $1-2\delta$, throughout the execution of Algorithm 5, for each $g \in \mathcal{K}$ we have $V^{\widetilde{\pi}_g}_g(s_0) \leq L(1+\epsilon)$ and $\left\|V^{\widetilde{\pi}_g}_g\right\|_\infty \leq 32L$.*

Proof. By Lemma 29 and a union bound over all trials and rounds, with probability at least $1-\delta$, we have $\left\|V^{\widetilde{\pi}_g}_g\right\|_\infty \leq 32L$ for each $g \in \mathcal{K}$, since $\widetilde{\pi}_g$ passes the test in Line 12. Moreover, by the definition of success round, value of λ, and Lemma 50, with probability at least $1-\delta$, for each $g \in \mathcal{K}$, in the round that g is added to $\mathcal{K}$, we have $V^{\widetilde{\pi}_g}_g(s_0) = V^{\pi_g}_g(s_0) \leq \widehat{\tau} + \frac{L\epsilon}{2} \leq V_{\mathcal{K},g}(s_0) + L\epsilon \leq L(1+\epsilon)$. ☐

D.3. Properties of the sets built by Algorithm 5

Lemma 21 (Restricted Optimism)**.** *With probability at least $1-\delta$ over the randomness of Algorithm 5, at any trial and any round, after executing Line 7, if $\mathcal{K}^\star_{z,j} \subseteq \mathcal{K}$ for some $j \in [z]$, then $V_{\mathcal{K},g}(s) \leq V^\star_{\mathcal{K}^\star_{z,j},g}(s)$ for any $s \in \mathcal{S}$ and $g \in \mathcal{K}^\star_{z,j+1} \setminus \mathcal{K}$.*

Proof. For any $\tau' \geq 1$, $z' \geq 1$, $j \in [z']$, $g \in \mathcal{K}^\star_{z',j+1} \setminus \mathcal{K}^\star_{z',j}$, by Lemma 2 and $\left\|V^\star_{\mathcal{K}^\star_{z',j},g}\right\|_\infty \leq L+1$ (Lemma 44), with probability at least $1 - \frac{\delta}{4(z')^4(\tau')^2}$, for any status of $\mathbf{N}$ and $\xi > 0$, we have $V(s) \leq V^\star_{\mathcal{K}^\star_{z',j},g}(s)$ for all $s \in \mathcal{S}$ where

$(\_, V, \_) = \text{VISGO}(\mathcal{K}^\star_{z',j}, g, \xi, \mathbf{N}, \frac{\delta}{4(\tau')^2(z')^4AL})$. By a union bound, all events above hold simultaneously with probability at least $1-\delta$.

At any trial τ and round, after executing Line 7, let $(\_, V_{\mathcal{K}^\star_{z,j},g}, \_) = \text{VISGO}(\mathcal{K}^\star_{z,j}, g, \epsilon_{\text{VI}}, \mathbf{N}, \delta')$ (no need to compute explicitly) for any $j \in [z]$, and $g \in \mathcal{K}^\star_{z,j+1} \setminus \mathcal{K}^\star_{z,j}$, where $\delta' = \frac{\delta}{4\tau^2 z^4 AL}$. The union bound above implies that $V_{\mathcal{K}^\star_{z,j},g}(s) \le V^\star_{\mathcal{K}^\star_{z,j},g}(s)$ for any $s \in \mathcal{S}$. Then by Lemma 5, we also have $V_{\mathcal{K},g}(s) \le V^\star_{\mathcal{K}^\star_{z,j},g}(s)$ if $\mathcal{K}^\star_{z,j} \subseteq \mathcal{K}$ ($V_{\mathcal{K},g}$ is computed in Line 7). □

Lemma 22. *For a given trial (τ, z), denote by $\mathcal{K}_r$ the set $\mathcal{K}$ at the end of each round r. With probability at least $1-2\delta$, for any $j \ge 1$ and round $r \ge 1$ in any trial in which $\mathcal{K}_r$ is updated or returned (i.e., Line 8 is executed) and $\mathcal{K}_{r-1} \supseteq \mathcal{K}^\star_j$, we have $\mathcal{K}^\star_{j+1} \subseteq \mathcal{K}_r$.*

Proof. In this lemma we denote by $\mathcal{U}_r$ the value of $\mathcal{U}$ *at the end of* round r. Define the event $E := \{$for any trial, $\forall r \ge 1$ in which $\mathcal{K}_r$ is updated : $\mathcal{T}_L(\mathcal{K}_r) \setminus \mathcal{K}_r \subseteq \mathcal{U}_r\}$. By Lemma 28, it holds with probability at least $1-\delta$. Let us carry out the proof conditioned on E holding.

In any trial, take some round r such that Line 8 is executed and $\mathcal{K}_{r-1} \supseteq \mathcal{K}^\star_j$. Let $r' < r$ be the last round where $\mathcal{K}_{r'}$ was updated (and thus $\mathcal{U}_{r'}$ was created). Note that $\mathcal{K}_{r'} = \mathcal{K}_{r-1} \supseteq \mathcal{K}^\star_j$. Then, event E and the definition of the sets $(\mathcal{K}^\star_j)_j$ directly imply that $\mathcal{K}^\star_{j+1} := \mathcal{T}_L(\mathcal{K}^\star_j) \subseteq \mathcal{T}_L(\mathcal{K}_{r'}) \subseteq \mathcal{U}_{r'} \cup \mathcal{K}_{r'}$. Since $\mathcal{K}_r$ can only be formed by adding states in $\mathcal{U}_{r'}$ to $\mathcal{K}_{r'}$, and the union of these sets contains $\mathcal{K}^\star_{j+1}$, if $\mathcal{K}^\star_{z,j+1} \not\subseteq \mathcal{K}_r$, it must be that there exists $g \in \mathcal{U}_{r-1} \cap \mathcal{K}^\star_{z,j+1}$ s.t. $V_{\mathcal{K}_{r-1},g}(s_0) > L$. However, Lemma 21, which holds with probability $1-\delta$, implies that, at any round $r \ge 1$, if $\mathcal{K}^\star_j \subseteq \mathcal{K}_{r-1}$ (which implies that $z > |\mathcal{K}^\star_j|$ and $\mathcal{K}^\star_j = \mathcal{K}^\star_{z,j}$ by Line 5), then $V_{\mathcal{K}_{r-1},g}(s_0) \le V^\star_{\mathcal{K}^\star_j,g}(s_0) \le L$ for any $g \in \mathcal{K}^\star_{z,j+1} \setminus \mathcal{K}_{r-1}$. This is a contradiction, which implies that $\mathcal{U}_{r-1} \cap \mathcal{K}^\star_{z,j+1} = \emptyset$ and, thus, all states in $\mathcal{K}^\star_{z,j+1}$ must have been added to $\mathcal{K}_r$. Moreover, since a new trial is not triggered in round r, by Line 5, we have $z > |\mathcal{K}^\star_{z,j+1}|$ and $\mathcal{K}^\star_{z,j+1} = \mathcal{K}^\star_{j+1}$. This completes the proof. □

Lemma 23. *For a given trial (τ, z), denote by $\mathcal{K}_r$ the set $\mathcal{K}$ at the end of each round r inside the trial. With probability at least $1-4\delta$, at any trial (τ, z), we have $\mathcal{K}_r \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ for any round r, and $\mathcal{S}^{\rightarrow}_L \subseteq \mathcal{K}_r$ if the algorithm terminates at round r.*

Proof. Fix any trial (τ, z). Clearly, $\mathcal{K}_1 \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$. To prove the first statement, consider a round $r \ge 1$ and suppose $\mathcal{K}_r \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$. If, in this round, the algorithm selects a goal $g^\star \in \mathcal{U} \setminus \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, $\pi_{g^\star}$ passes the test of Line 12, and a skip round is not triggered, then we show that the "failure test" in Line 23 is triggered.

Since $\pi_{g^\star}$ passed the test of Line 12, we have $\|V^{\pi_{g^\star}}_{g^\star}\|_\infty \le 32L$ with probability at least $1-\delta$ by Lemma 29 and a union bound over all trials and rounds. Combining this with Lemma 50 and the value of λ (Line 16) (again by a union bound over all trials and rounds), we have $\widehat{\tau} \ge V^{\pi_{g^\star}}_{g^\star}(s_0) - L\epsilon/2$ with probability at least $1-2\delta$. By assumption on $g^\star$ and since $\pi_{g^\star}$ is restricted on $\mathcal{K}_r \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, we have $V^{\pi_{g^\star}}_{g^\star}(s_0) \ge V^\star_{\mathcal{K}_r,g^\star}(s_0) \ge V^\star_{\mathcal{S}^{\rightarrow}_{L(1+\epsilon)},g^\star}(s_0) > L(1+\epsilon)$, which implies that $\widehat{\tau} \ge L(1+\epsilon/2) \ge V_{\mathcal{K}_r,g^\star}(s_0) + \epsilon L/2$, where the last inequality is from the goal-selection rule. Therefore, the failure test triggers and $g^\star$ is not added to $\mathcal{K}'$. Overall, any $g \notin \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ will never be added to $\mathcal{K}$ or $\mathcal{K}'$ throughout the execution of Algorithm 5.

To prove the second statement, let us consider any trial (τ, z) where the algorithm stops. Clearly, $\mathcal{K}^\star_1 \subseteq \mathcal{K}_1$ at the end of round $r = 1$ in this last trial. Then, if r is the round where the algorithm terminates, and $\mathcal{K}^\star_j \subseteq \mathcal{K}_{r-1}$ for some $j \ge 1$, we have $\mathcal{K}^\star_{j+1} \subseteq \mathcal{K}_r$ with probability at least $1-2\delta$ by Lemma 22. Moreover, since $\mathcal{K}' = \varnothing$ in round r, we have $\mathcal{K}^\star_{j+1} \subseteq \mathcal{K}_{r-1} = \mathcal{K}_r$. By a recursive application of Lemma 22, we have $\mathcal{K}^\star_j \subseteq \mathcal{K}_r$ for any $j \ge 1$ (note that $\mathcal{K}' = \varnothing$ at the beginning of round r). Lemma 1 then implies the statement. □

Lemma 24. *Conditioned on the events of Lemma 28 and Lemma 23, $\mathcal{U} \subseteq \bar{\mathcal{U}}$ at the beginning of any round in any trial.*

Proof. This is clearly true at the beginning of the first round of any trial since $\mathcal{U} = \varnothing$. Then by the events of Lemma 28 and Lemma 23, $\mathcal{U} \subseteq \mathcal{T}_{2L}(\mathcal{K}) \setminus \mathcal{K} \subseteq \bar{\mathcal{U}}$ every time after executing Line 11. Moreover, we only remove elements from $\mathcal{U}$ except when executing Line 11. This completes the proof. □

Lemma 25. *Denote by $\mathcal{K}_r$ the set $\mathcal{K}$ at the end of each round r. With Assumption 2, with probability at least $1-8\delta$ over the randomness of Algorithm 5, we have that $\mathcal{K}_r = \mathcal{K}_j^\star$ for some $j \in [S_L^{\rightarrow}]$ at any round r and, $\mathcal{K}_r = \mathcal{S}_L^{\rightarrow}$ if the algorithm terminates at round r.*

Proof. By Lemma 23, with probability at least $1-4\delta$, we have $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{L(1+\epsilon)}^{\rightarrow}$ if the algorithm terminates. By Remark 1, $\mathcal{K} = \mathcal{S}_L^{\rightarrow}$. Thus, it suffices to show that at any trial $\mathcal{K} = \mathcal{K}_j^\star$ for some $j \leq S_L^{\rightarrow}$.

The algorithm is such that $\mathcal{K}_1^\star = \mathcal{K}_1 = \{s_0\}$. Suppose at the end of a round r we have that $\mathcal{K}_r = \mathcal{K}_j^\star$ for some $j \geq 1$. By Lemma 22, with probability at least $1-2\delta$, if the condition of Line 8 is verified the first time in some round $r' > r$, then we must have $\mathcal{K}_{j+1}^\star \subseteq \mathcal{K}_{r'}$. If we also have $\mathcal{K}_{r'} \subseteq \mathcal{K}_{j+1}^\star$, then the statement is proved.

In any round r such that $\mathcal{K} = \mathcal{K}_j^\star$, $g^\star \in \mathcal{U} \setminus \mathcal{K}_{j+1}^\star$, $\pi_{g^\star}$ passes the test of Line 12, and a skip round is not triggered, by Lemma 50, the value of λ, and Lemma 29 (applying a union bound over all trials and rounds), we have $\widehat{\tau} \geq V_{g^\star}^{\pi_{g^\star}}(s_0) - L\epsilon/2$ with probability at least $1-2\delta$. By assumption on $g^\star$ and since $\pi_{g^\star}$ is restricted on $\mathcal{K} \subseteq \mathcal{K}_j^\star$, we have $V_{g^\star}^{\pi_{g^\star}}(s_0) \geq V_{\mathcal{K},g^\star}^\star(s_0) \geq V_{\mathcal{K}_j^\star,g^\star}^\star(s_0) > L(1+\epsilon)$, which implies that $\widehat{\tau} \geq L(1+\epsilon/2) \geq V_{\mathcal{K},g^\star}(s_0) + \epsilon L/2$, where the last inequality is from the goal-selection rule. Therefore, the failure test triggers and $g^\star$ is not added to $\mathcal{K}'$ or $\mathcal{K}$. This proves $\mathcal{K} \subseteq \mathcal{K}_{j+1}^\star$ in round r'. □

Lemma 26. *With Assumption 2, conditioned on the events of Lemma 28 and Lemma 25, in any trial, $\mathcal{U} \subseteq \mathcal{U}_z^\star$ at the beginning of any round.*

Proof. By Lemma 25, in any trial, we have $\mathcal{K} = \mathcal{K}_j^\star \subseteq \mathcal{K}_{z,z}^\star$ for some $j \leq z$ at the end of any round. Then by Lemma 28, we have $\mathcal{U} \subseteq \mathcal{T}_{2L}(\mathcal{K}) \setminus \mathcal{K} \subseteq \mathcal{U}_z^\star$ every time Line 11 is executed. □

D.4. Properties of $\mathcal{U}$

Given $\mathcal{X}$, $\Pi_{\mathcal{X}} = \{\pi_g\}_{g\in\mathcal{X}}$ and δ as input of `ComputeU`, let $\mathcal{D}_0$ and $\mathcal{D}_1$ be the random samples collected respectively in Line 26 and Line 27. Define

$$\mathcal{E}_0(\mathcal{D}_0) = \left\{\mathcal{N}(\mathcal{X}, \frac{1}{2L}) \not\subseteq \mathcal{U}'\right\},$$
$$\mathcal{E}_1(\mathcal{D}_0, \mathcal{D}_1) = \left\{\exists g \in \mathcal{U}', V_{\mathcal{X},g}'(s_0) > V_{\mathcal{X},g}^\star(s_0)\right\},$$
$$\mathcal{E}_2(\mathcal{D}_0, \mathcal{D}_1) = \left\{\exists g \in \mathcal{U}', V_g^{\pi_g}(s) > 2V_{\mathcal{X},g}'(s)\right\}.$$

In this section we use $\mathbb{E}$ and $\mathbb{P}$ to denote expectation and probability w.r.t. these two random generation processes.

Lemma 27. *With any $\mathcal{X}$, $\{\pi_g \in \Pi(\mathcal{X})\}_{g\in\mathcal{X}}$ such that $\left\|V_g^{\pi_g}\right\|_\infty = \mathcal{O}(L)$, and $\delta \in (0,1)$ as input,* `ComputeU` *ensures*

$$\mathbb{P}\left(\mathcal{T}_L(\mathcal{X}) \setminus \mathcal{X} \subseteq \mathcal{U} \subseteq \mathcal{T}_{2L}(\mathcal{X}) \setminus \mathcal{X}\right) \geq 1-\delta.$$

With the same probability, the sample complexity of `ComputeU` *is bounded by $\mathcal{O}(L^3|\mathcal{X}|^2 A \log^2 \frac{L|\mathcal{X}|A}{\delta})$.*

Proof. Denote by $\{s_{i,s,a}\}_{i,s,a}$ the set of next state samples collected in Line 26 for each (s,a). Let $\mu = 2L\log(4LA|\mathcal{X}|/\delta)$, then

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_0(\mathcal{D}_0)\right) &= P\left(\exists s' \in \mathcal{N}(\mathcal{X}, \frac{1}{2L}), \forall (s,a) \in \mathcal{X}\times\mathcal{A}, \forall i \in [\mu] : s_{i,s,a} \neq s'\right) \\
&\leq \sum_{s'\in\mathcal{N}(\mathcal{X},\frac{1}{2L})} P\left(\forall (s,a) \in \mathcal{X}\times\mathcal{A}, \forall i \in [\mu] : s_{i,s,a} \neq s'\right) \\
&\leq \sum_{s'\in\mathcal{N}(\mathcal{X},\frac{1}{2L})} \prod_{(s,a)\in\mathcal{X}\times\mathcal{A}} \prod_{i\in[\mu]} (1 - P(s'|s,a)) \leq \sum_{s'\in\mathcal{N}(\mathcal{X},\frac{1}{2L})} (1 - P(s'|\bar{s},\bar{a}))^\mu \\
&\qquad\qquad (\bar{s}, \bar{a} \text{ such that } P(s'|\bar{s},\bar{a}) \geq \tfrac{1}{2L}) \\
&\leq \sum_{s'\in\mathcal{N}(\mathcal{X},\frac{1}{2L})} \left(1 - \frac{1}{2L}\right)^\mu \leq \sum_{s'\in\mathcal{N}(\mathcal{X},\frac{1}{2L})} \frac{\delta}{4LA|\mathcal{X}|} \leq \delta/2. \qquad (|\mathcal{N}(\mathcal{X}, \tfrac{1}{2L})| \leq 2LA|\mathcal{X}|)
\end{aligned}$$

Algorithm 6: EXPLORE

Input: States $\mathcal{X}$, policies $\Pi = \{\pi_x\}_{x\in\mathcal{X}}$ such that $\|V_x^{\pi_x}\|_\infty = \mathcal{O}(L)$, counters n, target value $\bar{n}$.
$\mathcal{S}_{\text{next}} \leftarrow \varnothing$.
for $(x, a) \in \mathcal{X} \times \mathcal{A}$ **do**
  while $n(x, a) < \bar{n}$ **do**
    Reset to s_0 and execute π_x until reaching x.
    Execute action a, observe $x' \sim P_{x,a}$, and update $n(x, a, x') \xleftarrow{+} 1$.
    if $x' \notin \mathcal{X}$ **then** $\mathcal{S}_{\text{next}} \leftarrow \mathcal{S}_{\text{next}} \cup \{x'\}$.
return n *and* $\mathcal{S}_{next}$.

Let N_1 be defined as in Lemma 4. Then, from Lemma 2 and Lemma 4, by using $\delta/(4|\mathcal{U}'|)$, we have that $\mathbb{P}\left(\mathcal{E}_1(\mathcal{D}_0, \mathcal{D}_1)|\mathcal{D}_0\right) \leq \delta/4$ and $\mathbb{P}\left(\mathcal{E}_2(\mathcal{D}_0, \mathcal{D}_1)|\mathcal{D}_0\right) \leq \delta/4$. Then, we can write that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_0(\mathcal{D}_0) \cup \mathcal{E}_1(\mathcal{D}_0, \mathcal{D}_1) \cup \mathcal{E}_2(\mathcal{D}_0, \mathcal{D}_1)) &\leq \mathbb{P}(\mathcal{E}_0(\mathcal{D}_0)) + \mathbb{P}(\mathcal{E}_1(\mathcal{D}_0, \mathcal{D}_1) \cup \mathcal{E}_2(\mathcal{D}_0, \mathcal{D}_1)) \\ &\leq \delta/2 + \sum_{\mathcal{D}_0} \mathbb{P}(\mathcal{D}_0) \underbrace{\mathbb{P}(\mathcal{E}_1(\mathcal{D}_0, \mathcal{D}_1) \cup \mathcal{E}_2(\mathcal{D}_0, \mathcal{D}_1)|\mathcal{D}_0)}_{\leq \delta/2, \forall \mathcal{D}_0} = \delta\end{aligned}$$

We then carry out the proof under event $E = \neg(\mathcal{E}_1(\mathcal{D}_0) \cup \mathcal{E}_1(\mathcal{D}_0, \mathcal{D}_1) \cup \mathcal{E}_2(\mathcal{D}_0, \mathcal{D}_1))$ which hold with probability $1 - \delta$.

Since π'_g is restricted on $\mathcal{X}$, we have that $V^\star_{\mathcal{X},g}(s_0) \leq V_g^{\pi'_g}(s_0)$ by the definition of optimal policy. We have that, for any $g \in \mathcal{U}$, $V^\star_{\mathcal{X},g}(s_0) \leq V_g^{\pi'_g}(s_0) \leq 2V'_{\mathcal{X},g}(s_0) \leq 2L$ by the definition of $\mathcal{U}$. This implies that $\mathcal{U} \subseteq \mathcal{T}_{2L}(\mathcal{X}) \cap \mathcal{U}' \subseteq \mathcal{T}_{2L}(\mathcal{X}) \setminus \mathcal{X}$ since $\mathcal{U}' \cap \mathcal{X} = \emptyset$ by definition.

Finally, note that, by the definition of $\mathcal{T}_L(\mathcal{X})$ and the event $\neg\mathcal{E}_0$, $\mathcal{T}_L(\mathcal{X}) \setminus \mathcal{X} \subseteq \mathcal{N}(\mathcal{X}, \frac{1}{2L}) \subseteq \mathcal{U}'$ w.h.p. Furthermore, under the event $\neg\mathcal{E}_1(\mathcal{D}_0, \mathcal{D}_1)$, we have that for any $g \in \mathcal{U}'$, if $V^\star_{\mathcal{X},g}(s_0) \leq L$, then $V'_{\mathcal{X},g}(s_0) \leq V^\star_{\mathcal{X},g}(s_0) \leq L$. Thus, $\mathcal{T}_L(\mathcal{X}) \setminus \mathcal{X} \subseteq \mathcal{U}$.

Sample complexity. Since $\|V_g^{\pi_g}\|_\infty = \mathcal{O}(L)$, by Lemma 30 with $\bar{n} = \mu$ and $N_1(|\mathcal{X}|, \frac{\delta}{4|\mathcal{U}'|})$, with probability at least $1 - \delta$, the sample complexity is $\mathcal{O}(L|\mathcal{X}|An' \log \frac{|\mathcal{X}|An'}{\delta})$, where $n' = \mu + N_1(|\mathcal{X}|, \delta/(4|\mathcal{U}'|)$. Given that $N_1(|\mathcal{X}|, \frac{\delta}{4|\mathcal{U}'|}) = \mathcal{O}(L^2|\mathcal{X}| \log(|\mathcal{U}'||\mathcal{X}|/\delta))$ (see Lemma 4), we have $n' = \mathcal{O}(L^2|\mathcal{X}| \log(L|\mathcal{X}|A/\delta))$. Plugging this back, the sample complexity is $\mathcal{O}(L^3|\mathcal{X}|^2 A \log^2 \frac{L|\mathcal{X}|A}{\delta})$. ☐

Lemma 28. *With probability at least $1 - \delta$ over the randomness of Algorithm 5, at any trial and round, $\mathcal{T}_L(\mathcal{K}) \setminus \mathcal{K} \subseteq \mathcal{U} \subseteq \mathcal{T}_{2L}(\mathcal{K}) \setminus \mathcal{K}$ after executing Line 11 (if it is executed).*

Proof. This is simply by Lemma 27 and the choice of confidence level in Line 11 in each trial and round. ☐

D.5. RTEST and EXPLORE

Here we show auxiliary algorithms and related lemmas used in Algorithm 5.

Lemma 29. *For any $\mathcal{X} \subseteq \mathcal{S}$, $\{\pi_g\}_{g\in\mathcal{X}}$, policy $\overline{\pi} \in \Pi(\mathcal{X})$, goal state $g \in \mathcal{S}$, and $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left(\text{RTEST}(\mathcal{X}, \{\pi_g\}_{g\in\mathcal{X}}, \pi, g, \delta) = \text{TRUE} \middle| \left\|V_g^{\overline{\pi}}\right\|_\infty \leq 4L\right) \geq 1 - \delta,$$

$$\mathbb{P}\left(\text{RTEST}(\mathcal{X}, \{\pi_g\}_{g\in\mathcal{X}}, \pi, g, \delta) = \text{TRUE} \implies \left\|V_g^{\overline{\pi}}\right\|_\infty \leq 32L\right) \geq 1 - \delta.$$

Moreover, if $\left\|V_g^{\pi_g}\right\|_\infty = \mathcal{O}(L)$ for any $g \in \mathcal{X}$, then with probability at least $1 - \delta$, the sample complexity is $\tilde{\mathcal{O}}(L|\mathcal{X}| \log^2 \frac{|\mathcal{X}|}{\delta})$.

Proof. Let $\{\eta_i\}_{i\in[n]}$ be rollouts of length at most $\bar{l}$ generated running $\overline{\pi}$ from state s, and denote by $p^{\overline{\pi}}_{\bar{l},g}(s)$ the probability of reaching the goal g in at most $\bar{l}$ steps by following policy $\overline{\pi}$ starting from s. Let $\mathbf{1}(\eta) = 1$ if the goal has been reached in

Algorithm 7: RTEST

Input: reaching policy $\{\pi_s\}_{s\in\mathcal{X}}$, test policy $\pi \in \Pi(\mathcal{X})$, goal state g, and failure probability δ.
Let $n = 2^{10}\log\frac{2|\mathcal{X}|}{\delta}$.
for $s \in \mathcal{X}$ **do**
    $i_s \leftarrow 0$.
    for $j = 1, \ldots, n$ **do**
        Reset to s_0 and execute π_s until s is reached.
        Execute π until g is reached or $8L$ steps is taken.
        if *g is reached* **then** $i_s \overset{+}{\leftarrow} 1$
    if $i_s/n < \frac{7}{16}$ **then return** FALSE.
return TRUE.

rollout η, zero otherwise. $X_i = \mathbf{1}_g(\eta_i) - p_g^{\pi}(s)$ is a martingale difference sequence ($|X_i| \le 1$) and by Azuma's inequality (see Lemma 53), setting $n = 2^{10}\log(\frac{2|\mathcal{X}|}{\delta})$, we have

$$\mathbb{P}\left(\forall s \in \mathcal{X}, \frac{1}{n}\left|\sum_{i=1}^{n} X_i\right| \le \frac{1}{16}\right) \ge 1 - \delta. \tag{9}$$

1) If $\left\|V_g^{\overline{\pi}}\right\|_\infty \le 4L$, by Markov's inequality, $p_{\bar{l},g}^{\overline{\pi}}(s) \ge 1/2$ when $\bar{l} = 8L$. This gives $\frac{i_s}{n} = \sum_i \frac{\mathbf{1}_g(\eta_i)}{n} \ge p_g^{\overline{\pi}}(s) - \frac{1}{16} \ge \frac{7}{16}$ for any $s \in \mathcal{X}$, and thus the algorithm returns TRUE on termination.

2) If the output is TRUE, then $\frac{i_s}{n} \ge \frac{7}{16}$ for all $s \in \mathcal{X}$. By (9), we have that $p_g^{\overline{\pi}}(s) \ge \frac{i_s}{n} - \frac{1}{16} \ge \frac{3}{8}$. Thus for any $s \in \mathcal{X}$, $V_g^{\pi}(s) \le 8L + \frac{5}{8}\left\|V_g^{\pi}\right\|_\infty$, which gives $\left\|V_g^{\pi}\right\|_\infty \le 1 + 8L + \frac{5}{8}\left\|V_g^{\pi}\right\|_\infty$ by $\pi \in \Pi(\mathcal{X})$. This implies $\left\|V_g^{\pi}\right\|_\infty \le 32L$.

Sample complexity. If $\|V_s^{\pi_s}\|_\infty = \mathcal{O}(L)$ for any $s \in \mathcal{X}$, by Lemma 52, with probability $1-\delta$, all trajectories generated by π_s for some $s \in \mathcal{X}$ reaches state s in $\mathcal{O}(L\log(2n|\mathcal{X}|/\delta))$ steps. Noting that we generate n trajectories for each $s \in \mathcal{X}$ completes the proof. □

Lemma 30. *For any $\mathcal{X} \subseteq \mathcal{S}$, $\Pi = \{\pi_x\}_{x\in\mathcal{X}}$, counter n, threshold $\bar{n} \ge 1$, and $\delta \in (0,1)$, with probability at least $1-\delta$, the sample complexity of* EXPLORE$(\mathcal{X}, \Pi, n, \bar{n})$ *is* $\mathcal{O}(L|\mathcal{X}|A\bar{n}\log\frac{|\mathcal{X}|A\bar{n}}{\delta})$.

Proof. For any $x \in \mathcal{X}$, since $\|V_x^{\pi_x}\|_\infty = \mathcal{O}(L)$, by Lemma 52, with probability $1-\delta'$ it takes $\mathcal{O}(L\log(1/\delta'))$ steps to reach the goal state following π_x from any $s \in \mathcal{X}$. Therefore, by setting $\delta' = \frac{\delta}{|\mathcal{X}|A\bar{n}}$, with probability $1-\delta$, all trajectories reach the desired goal state within $\mathcal{O}(L\log(1/\delta'))$ steps. Given that there are at most $|\mathcal{X}|A\bar{n}$ trajectories, with probability at least $1-\delta$, the total sample complexity is $\mathcal{O}(L|\mathcal{X}|A\bar{n}\log\frac{|\mathcal{X}|A\bar{n}}{\delta})$. □

E. Analysis of Policy Consolidation

In this section, we bound the sample complexity of Algorithm 2.

Notation We assume that all episodes lie in one (artificial) trial. Let g_k, $\mathcal{K}_k$, V_k $V^\star_k$ be the values of $g^\star$, $\mathcal{K}\setminus\{g^\star\}$, $\widehat{V}$, and $V^\star_{\mathcal{K},g^\star}$ in episode k respectively. Denote by I_k the number of steps in episode k. Note that $I_k<\infty$ with probability 1 by Line 13, and $s^k_{I_k+1}\neq g_k$ only when a skip round is triggered in episode k. Denote by $\mathcal{F}_k$ the σ-algebra of events up to episode k. Define K as the total number of episodes throughout the execution of Algorithm 2. For any $K'\leq K$, define $R_{K'}=\sum_{k=1}^{K'}(I_k-V_k(s_0))$ and $C_{K'}=\sum_{k=1}^{K'}I_k$. Define $P^k_i=P_{s^k_i,a^k_i}$. In episode k, when $s^k_i\in\mathcal{K}$, denote by $\bar{P}^k_i$, $\widetilde{P}^k_i$, $\mathbf{N}^k_i$, b^k_i the values of $\bar{P}_{s^k_i,a^k_i}$, $\widetilde{P}_{s^k_i,a^k_i}$, $n^+(s^k_i,a^k_i)$, and $b^{(l)}(s^k_i,a^k_i)$, where $\bar{P}$, n^+, $b^{(l)}$ are used in Algorithm 4 to compute V_k and l is the final value of i in Algorithm 4; when $s^k_i\notin\mathcal{K}$, define $\bar{P}^k_i=\mathbb{I}_{s_0}$, $\mathbf{N}^k_i=\infty$, and $b^k_i=0$. Also define ϵ_k as the value of ϵ_{VI} used in Algorithm 4 to compute V_k. In this section, $\mathcal{K}\subseteq\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ is an input of Algorithm 2 and thus does not have randomness.

Proof of Theorem 4. By Lemma 32, the output policies $\{\widetilde{\pi}_g\}_g$ clearly satisfies the statement. Define $\iota=\log\left(\frac{LS^{\rightarrow}_{L(1+\epsilon)}A}{\delta\epsilon}\right)$. It suffices to bound the number of samples collected in Line 2 and policy evaluation. With probability at least $1-\delta$, the number of samples collected in Line 2 is of order $\mathcal{O}(L^3{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota^2)$ by Lemma 30 and Lemma 4. With probability at least $1-16\delta$, by Lemma 31 and Lemma 33 ($c_1=\sqrt{LS^{\rightarrow}_{L(1+\epsilon)}A}$, $c_2=L{S^{\rightarrow}_{L(1+\epsilon)}}^2A$, and $p=2$), the number of samples collected in policy evaluation is of order $\tilde{\mathcal{O}}\left(\frac{LS^{\rightarrow}_{L(1+\epsilon)}A\iota^{10}}{\epsilon^2}+\frac{L{S^{\rightarrow}_{L(1+\epsilon)}}^2A\iota^{10}}{\epsilon}\right)$. Combining all cases completes the proof. ☐

Lemma 31. *With probability at least $1-4\delta$, if $R_{K'}\lesssim c_1\sqrt{\sum_{k=1}^{K'}V_k(s_0)\log^p(c_3K')}+c_2\log^p(c_3K')$ for any $K'\geq 1$ with $c_1,c_2\geq 1$ and $c_3=\frac{LS^{\rightarrow}_{L(1+\epsilon)}A}{\delta}$, then $C_K\lesssim\frac{LS^{\rightarrow}_{L(1+\epsilon)}A\iota^8}{\epsilon^2}+\frac{c_1^2\iota^{p+8}}{\epsilon^2}+\frac{c_2\iota^{p+4}}{\epsilon}$, where $\iota=\log\frac{c_1c_2c_3}{\epsilon\delta}$.*

Proof. For any $R'\geq 1$, let K' be the total number of episodes in the first R' rounds. Let $Z_{K'}=\sum_{k=1}^{K'}V_k(s_0)$. First note that the regret gives $C_{K'}\lesssim Z_{K'}+c_1\sqrt{Z_{K'}\log^p(c_3K')}+c_2\log^p(c_3K')$ and thus $\log(C_{K'})\lesssim\log(c_1c_2c_3Z_{K'})$. By $K'\lesssim C_{K'}$ and solving a "quadratic" inequality (Lemma 47), we have $C_{K'}\lesssim Z_{K'}+(c_1^2+c_2)\log^p(c_1c_2c_3Z_{K'})$. Denote by $\bar{g}_r$, $\bar{V}_r$, $\bar{\pi}_r$ the value of $g^\star$, $\widehat{V}$, and $\widehat{\pi}$ in round r respectively. For each failure round r, let C be the total cost within this round and m the number of episodes within this round. By definition, regret within this round satisfies $C-m\bar{V}_r(s_0)\geq C-\lambda\bar{V}_r(s_0)=\lambda(\widehat{\tau}-\bar{V}_r(s_0))>\frac{\lambda\epsilon\bar{V}_r(s_0)}{2}=\Omega(\bar{V}_r(s_0)/\epsilon)$. For each success and skip round r, by Lemma 35, Lemma 34, Lemma 51, and the value of λ, we have

$$\sum_{j=u_r}^{u'_r}\left(I_j-\bar{V}_r(s_0)\right)\gtrsim\sum_{j=u_r}^{u'_r-1}\left(I_j-V^{\bar{\pi}_r}_{\bar{g}_r}(s_0)\right)-L\gtrsim -L\sqrt{\lambda}\log^2\frac{L\lambda}{\delta}\gtrsim-\frac{L}{\epsilon}\log^4\frac{Lr}{\delta\epsilon}\gtrsim-\frac{L}{\epsilon}\log^4\frac{LC_{K'}}{\delta\epsilon},$$

where $\{u_r,\ldots,u'_r\}$ are the episodes in round r, and we lower bound the regret in the last episode by $\Omega(-L)$ since the last trajectory in a skipped round is truncated. Denote by $\mathcal{R}_f$ the total number of failure rounds within the first R' rounds. By the assumption in Algorithm 2 that $\mathcal{K}\subseteq\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, in the first R' rounds, the number of success round is at most $S^{\rightarrow}_{L(1+\epsilon)}$ and the number of skip rounds is at most $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)}A\log(C_{K'}))$. Since there are at most $\mathcal{O}(S^{\rightarrow}_{L(1+\epsilon)}A\log(C_{K'}))$ these rounds, in each round there are at most $\tilde{\mathcal{O}}(\frac{\log^4\frac{LC_{K'}}{\delta\epsilon}}{\epsilon^2})$ episodes (Line 7), and $\bar{V}_r(s_0)\leq 2L$ in any round r by Lemma 35, we have

$$\begin{aligned}Z_{K'}&\lesssim\frac{\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)\log^4\frac{LC_{K'}}{\delta\epsilon}}{\epsilon^2}+\frac{LS^{\rightarrow}_{L(1+\epsilon)}A\log^5\frac{c_1c_2c_3Z_{K'}}{\delta\epsilon}}{\epsilon^2}\\&\lesssim\frac{\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)\log^4\frac{c_1c_2c_3Z_{K'}}{\delta\epsilon}}{\epsilon^2}+\frac{LS^{\rightarrow}_{L(1+\epsilon)}A\log^5\frac{c_1c_2c_3Z_{K'}}{\delta\epsilon}}{\epsilon^2}.\end{aligned}$$

By Lemma 47, this gives

$$Z_{K'}\lesssim\frac{\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)\log^4(c_4\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0))}{\epsilon^2}+\frac{LS^{\rightarrow}_{L(1+\epsilon)}A\log^5(c_4\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0))}{\epsilon},$$

and $\log(Z_{K'}) \lesssim \log(\frac{c_1c_2c_3\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)}{\delta\epsilon}) \triangleq \log(c_4\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0))$, where $c_4=\frac{c_1c_2c_3}{\delta\epsilon}$. Therefore, the regret upper and lower bound and $\log(K')\le\log(C_{K'})\lesssim\log(c_1c_2c_3Z_{K'})\lesssim\log(c_4\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0))$ give

$$\frac{\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)}{\epsilon}-\frac{LS^{\rightarrow}_{L(1+\epsilon)}A}{\epsilon}\log^4\frac{LC_{K'}}{\delta\epsilon}\lesssim c_1\sqrt{Z_{K'}\log^p(c_3K')}+c_2\log^p(c_3K')$$
$$\lesssim\frac{c_1}{\epsilon}\sqrt{\left(\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)+LS^{\rightarrow}_{L(1+\epsilon)}A\log\left(c_4\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)\right)\right)\log^{p+4}\left(c_4\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)\right)}+c_2\log^p\left(c_4\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)\right).$$

Applying Lemma 47 gives $\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)\lesssim LS^{\rightarrow}_{L(1+\epsilon)}A\log^4(c_4)+c_1^2\log^{p+4}(c_4)+c_2\epsilon\log^p(c_4)$ and $\log(\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0))\lesssim\log(c_4)$. Now by the regret bound and AM-GM inequality, we have

$$\begin{aligned}C_{K'}&\lesssim Z_{K'}+c_1\sqrt{Z_{K'}\log^p(c_3K')}+c_2\log^p(c_3K')\lesssim Z_{K'}+(c_1^2+c_2)\log^p(c_4)\\&\lesssim\frac{\sum_{r\in\mathcal{R}_f}\bar{V}_r(s_0)\log^4(c_4Z_{K'})}{\epsilon^2}+\frac{LS^{\rightarrow}_{L(1+\epsilon)}A\log^5(c_4Z_{K'})}{\epsilon^2}+(c_1^2+c_2)\log^p(c_4)\\&\lesssim\frac{LS^{\rightarrow}_{L(1+\epsilon)}A\log^8(c_4)}{\epsilon^2}+\frac{c_1^2\log^{p+8}(c_4)}{\epsilon^2}+\frac{c_2\log^{p+4}(c_4)}{\epsilon}.\end{aligned}$$

Setting R' to be the total number of rounds, we have $K'=K$ and the proof completes. □

Lemma 32. *With probability at least $1-4\delta$, we have $V_g^{\widetilde{\pi}_g}(s_0)\le V^\star_{\mathcal{K},g}(s_0)(1+\epsilon)$ for $g\in\mathcal{K}$ throughout the execution of Algorithm 2.*

Proof. By Lemma 34 and Lemma 44, with probability at least $1-2\delta$, we have $V_{g^\star}^{\widehat{\pi}}(s)\le 2V^\star_{\mathcal{K},g^\star}(s)\le 4V^\star_{\mathcal{K},g^\star}(s_0)\le\min\{8L,4V_{g^\star}^{\widehat{\pi}}(s_0)\}$ for any $s\in\mathcal{S}$ throughout the execution. For any $g\in\mathcal{K}$, at the round that $\widetilde{\pi}_g$ is determined (where $g^\star=g$), by Lemma 50, value of λ and definition of success round, $V_g^{\widetilde{\pi}_g}(s_0)=V_g^{\widehat{\pi}}(s_0)\le\widehat{\tau}+\frac{\epsilon}{256}\left\|V_g^{\widehat{\pi}}\right\|_\infty\le\widehat{\tau}+\frac{\epsilon}{4}V_g^{\widehat{\pi}}(s_0)\le\widehat{V}(s_0)(1+\frac{\epsilon}{2})+\frac{\epsilon}{4}V_g^{\widehat{\pi}}(s_0)$. This gives $V_g^{\widetilde{\pi}_g}(s_0)\le\frac{1+\frac{\epsilon}{2}}{1-\frac{\epsilon}{4}}\widehat{V}(s_0)\le(1+\epsilon)V^\star_{\mathcal{K},g}(s_0)$ by $\widehat{V}(s_0)\le V^\star_{\mathcal{K},g}(s_0)$ (Lemma 35) and $\epsilon\in(0,1]$. □

Lemma 33. *With probability at least $1-12\delta$, for any $K'\le K$, we have $R_{K'}\lesssim\sqrt{LS^{\rightarrow}_{L(1+\epsilon)}A\sum_{k=1}^{K'}V_k(s_0)\iota}+{LS^{\rightarrow}_{L(1+\epsilon)}}^2A\iota$, where $\iota=\log^2\frac{LS^{\rightarrow}_{L(1+\epsilon)}AK'}{\delta}$.*

Proof. By Lemma 54 and a union bound on $\{V^\star_{\mathcal{K},g}\}_{g\in\mathcal{K}}$ and $(s,a)\in\mathcal{K}\times\mathcal{A}$, with probability at least $1-\delta$, $(P_i^k-\bar{P}_i^k)V_k^\star\lesssim\sqrt{\frac{\mathbb{V}(P_i^k,V_k^\star)\iota'}{\mathbf{N}_i^k}}+\frac{L\iota'}{\mathbf{N}_i^k}$ for any $k\in[K']$ and $i\in[I_k]$ (note that this holds even if $s_i^k\notin\mathcal{K}$), where $\iota'=\log\frac{S^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta}$. Moreover, with probability at least $1-\delta$,

$$\begin{aligned}&\sum_{k=1}^{K'}(I_k-V_k(s_0))\le\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(1+V_k(s_{i+1}^k)-V_k(s_i^k)\right)\\&\lesssim\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((\mathbb{I}_{s_{i+1}^k}-P_i^k)V_k+(P_i^k-\bar{P}_i^k)V_k+b_i^k+\epsilon_k\right) &&\text{(Lemma 42)}\\&\lesssim\sqrt{\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\log\frac{LC_{K'}}{\delta}}+\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((P_i^k-\bar{P}_i^k)V_k^\star+(P_i^k-\bar{P}_i^k)(V_k-V_k^\star)+b_i^k\right)+L\log\frac{LC_{K'}}{\delta}.\end{aligned}$$

where the last step is by Lemma 41 and Lemma 55. Now note that with probability at least $1-2\delta$,

$$\begin{aligned}
&\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((P_i^k-\bar{P}_i^k)V_k^\star+(P_i^k-\bar{P}_i^k)(V_k-V_k^\star)+b_i^k\right)\\
&\lesssim \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(\sqrt{\frac{\mathbb{V}(P_i^k,V_k^\star)\iota'}{\mathbf{N}_i^k}}+\sqrt{\frac{\Gamma_{L(1+\epsilon)}\mathbb{V}(P_i^k,V_k-V_k^\star)\iota'}{\mathbf{N}_i^k}}+\frac{\Gamma_{L(1+\epsilon)}L\iota'}{\mathbf{N}_i^k}+b_i^k\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left(\text{Lemma 46}, \|V_k^\star\|_\infty\leq 2L+1, \iota'=\log\frac{S_{L(1+\epsilon)}^{\rightarrow}AC_{K'}}{\delta}\right)\\
&\lesssim \sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota'}+\sqrt{S_{L(1+\epsilon)}^{\rightarrow}\Gamma_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k-V_k^\star)\iota'}+L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota',
\end{aligned}$$

where in the last step $\iota'=\log^2\frac{S_{L(1+\epsilon)}^{\rightarrow}AC_{K'}}{\delta}$ and we apply Lemma 40, Cauchy-Schwarz inequality, Lemma 39, and $\textsc{Var}[X+Y]\leq 2(\textsc{Var}[X]+\textsc{Var}[Y])$. Thus, by Lemma 37 with Lemma 35 and AM-GM inquality, with probability at least $1-8\delta$, we continue with

$$\begin{aligned}
C_{K'}-\sum_{k=1}^{K'}V_k(s_0)&\lesssim\sqrt{S_{L(1+\epsilon)}^{\rightarrow}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\iota'}+L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota'\\
&\lesssim\sqrt{LS_{L(1+\epsilon)}^{\rightarrow}AC_{K'}\iota'}+L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\iota', \qquad\qquad \text{(Lemma 36)}
\end{aligned}$$

where $\iota'=\log^2\frac{LS_{L(1+\epsilon)}^{\rightarrow}AC_{K'}}{\delta}$. Solving a "quadratic" inequality w.r.t $C_{K'}$ (Lemma 47), we have $C_{K'}\lesssim\sum_{k=1}^{K'}V_k(s_0)+L{S_{L(1+\epsilon)}^{\rightarrow}}^2A\log^2\frac{LS_{L(1+\epsilon)}^{\rightarrow}AK'}{\delta}$. Plugging this back to the last inequality above completes the proof. □

Lemma 34. *With probability at least $1-2\delta$, throughout the execution of Algorithm 2, $V_{g^\star}^{\widehat{\pi}}(s)\leq 2V_{\mathcal{K},g^\star}^\star(s)$ for any $s\in\mathcal{S}$.*

Proof. By Lemma 35, value of ν (Line 2), and applying Lemma 4 with $\mathcal{X}=\mathcal{K}\setminus\{g\}$ for each $g\in\mathcal{K}$, we have $V_{g^\star}^{\widehat{\pi}}(s)\leq 2\widehat{V}(s)\leq 2V_{\mathcal{K},g^\star}^\star(s)$ for all $s\in\mathcal{S}$. □

Lemma 35. *With probability at least $1-\delta$, throughout the execution of Algorithm 2, $\widehat{V}(s)\leq V_{\mathcal{K},g^\star}^\star(s)$ for any $s\in\mathcal{S}$.*

Proof. This is simply by the value of $\widehat{V}$ in each round and applying Lemma 2 on $\{V_{\mathcal{K},g}^\star\}_{g\in\mathcal{K}}$. □

F. Lemmas for Policy Evaluation

In this section, we present a set of lemmas related to regret analysis shared among Algorithm 1, Algorithm 5, and Algorithm 2. In Algorithm 5, a trial is indexed by τ, and each trial corresponds to a value of z estimating $S_{L(1+\epsilon)}^{\rightarrow}$ (Line 1). In Algorithm 1 and Algorithm 2, we assume the whole learning procedure lies in an artificial trial. Note that when lemmas below are involved, we have $b_i^k=0$, $\mathbf{N}_i^k=\infty$, and $\bar{P}_i^k=\mathbb{I}_{s_0}$ when $s_i^k\notin\mathcal{K}_k$.

Lemma 36. *Let $\mathcal{G}$ be the goal set such that $\mathcal{S}_{L(1+\epsilon)}^{\rightarrow}\subseteq\mathcal{G}\subseteq\mathcal{S}$. In any trial, with probability at least $1-2\delta$, for any $K'\in[K]$, if $\mathcal{K}_k\subseteq\mathcal{S}_{L(1+\epsilon)}^{\rightarrow}$ and $g_k\in\mathcal{G}\setminus\mathcal{K}_k$ for any $k\in[K']$, then $\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k,V_k)\lesssim LC_{K'}+L^2\Gamma_{L(1+\epsilon)}S_{L(1+\epsilon)}^{\rightarrow}A\iota$, where $\iota=\mathcal{O}(\log(|\mathcal{G}|ALC_{K'}/\delta)\log(C_{K'}))$.*

Proof. Note that $\|V_k\|_\infty \leq 2L$ by the stopping condition (Line 1) of Algorithm 4, and with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(V_k(s_i^k)^2 - (P_i^k V_k)^2\right) \lesssim L\sum_{k=1}^{K}\sum_{i=1}^{I_k}(V_k(s_i^k) - P_i^k V_k)_+ \qquad (a^2-b^2 \leq (a+b)(a-b)_+ \text{ for } a,b\geq 0)\\
&\lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(1 + (\bar{P}_i^k - P_i^k)V_k + \frac{1}{\mathbf{N}_i^k} + \epsilon_k\right)_+ \qquad \text{(Lemma 42)}\\
&\lesssim LC_{K'} + L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(\sqrt{\frac{\Gamma_{L(1+\epsilon)}\mathbb{V}(P_i^k, V_k)\iota'}{\mathbf{N}_i^k}} + \frac{L\Gamma_{L(1+\epsilon)}\iota'}{\mathbf{N}_i^k} + \epsilon_k\right) \qquad (\text{Lemma 46 and } \mathbf{N}_i^k = \infty \text{ when } s_i^k \notin \mathcal{K}_k)\\
&\lesssim LC_{K'} + L\sqrt{\Gamma_{L(1+\epsilon)} S^{\rightarrow}_{L(1+\epsilon)} A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k)\iota'\log(C_{K'})} + L^2\Gamma_{L(1+\epsilon)}S^{\rightarrow}_{L(1+\epsilon)}A\iota'\log(C_{K'}),
\end{aligned}$$

where $\iota' = \log(|\mathcal{G}|AC_{K'}/\delta)$, and the last step is by Cauchy-Schwarz inequality, Lemma 40, and Lemma 41. Now let $Z_{K'} = \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k)$. Applying Lemma 38 and $\sum_{k=1}^{K'} V_k(s^k_{I_k+1})^2 \lesssim L^2 S^{\rightarrow}_{L(1+\epsilon)}A\iota'$ (this is because $V_k(s^k_{I_k+1})$ is non-zero only in skip rounds), we have with probability a least $1-\delta$,

$$Z_{K'} \lesssim LC_{K'} + L\sqrt{\Gamma_{L(1+\epsilon)}S^{\rightarrow}_{L(1+\epsilon)}AZ_{K'}\iota} + L^2\Gamma_{L(1+\epsilon)}S^{\rightarrow}_{L(1+\epsilon)}A\iota,$$

where $\iota = \mathcal{O}(\log(|\mathcal{G}|ALC_{K'}/\delta)\log(C_{K'}))$. Solving a quadratic inequality completes w.r.t. $Z_{K'}$ the proof. $\square$

Lemma 37. *In any trial, with probability at least $1-5\delta$, for any $K' \in [K]$ if 1) $\{V_k^\star\}_{k\in[K']} \subseteq \mathcal{V}$ where $\mathcal{V}$ is determined at the beginning of the trial, $|\mathcal{V}|$ is upper bounded by polynomials of $S^{\rightarrow}_{L(1+\epsilon)}$, and $\|V\|_\infty = \mathcal{O}(L)$ for any $V \in \mathcal{V}$, 2) $V_k(s) \leq V_k^\star(s)$ for any $k \in [K']$ and $s \in \mathcal{S}$, 3) $\mathcal{K}_k \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ for any $k \in [K']$, and 4) $g_k \in \bar{\mathcal{U}}\setminus\mathcal{K}_k$ for any $k \in [K']$, then $\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k^\star - V_k) \lesssim L\sqrt{S^{\rightarrow}_{L(1+\epsilon)}A\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\mathbb{V}(P_i^k, V_k)\iota'} + L^2 {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\iota'$, where $\iota' = \log^2\frac{LS^{\rightarrow}_{L(1+\epsilon)}AC_{K'}}{\delta}$.*

Proof. First note that

$$\begin{aligned}
&\sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left((V_k^\star(s_i^k) - V_k(s_i^k))^2 - (P_i^k(V_k^\star - V_k))^2\right)\\
&\lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(V_k^\star(s_i^k) - V_k(s_i^k) - P_i^k V_k^\star + P_i^k V_k)_+\\
&\qquad (V_k(s) \leq V_k^\star(s) \text{ for all } s \text{ and } a^2-b^2 \leq (a+b)(a-b)_+ \text{ for } a,b\geq 0)\\
&\lesssim L\sum_{k=1}^{K'}\sum_{i=1}^{I_k}(1 + P_i^k V_k - V_k(s_i^k))_+. \qquad (V_k^\star(s_i^k) \leq 1 + P_i^k V_k^\star)
\end{aligned}$$

Let $\bar{P}_{s,a}(s') = \frac{\mathbf{N}(s,a,s')}{\mathbf{N}^+(s,a)}$. By Lemma 54, with probability at least $1-\delta$, for any $(s,a) \in \mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \times \mathcal{A}$, $V \in \mathcal{V}$, and status of counter $\mathbf{N}$:

$$(P_{s,a} - \bar{P}_{s,a})V \lesssim \sqrt{\frac{\mathbb{V}(P_{s,a}, V)\iota'}{\mathbf{N}(s,a)}} + \frac{L\iota'}{\mathbf{N}(s,a)}, \tag{10}$$

where $\iota' = \log \frac{S^{\rightarrow}_{L(1+\epsilon)} AC_{K'}}{\delta}$. By Lemma 42, with probability at least $1 - 2\delta$, we continue with

$$\lesssim L \sum_{k=1}^{K'} \sum_{i=1}^{I_k} ((P_i^k - \bar{P}_i^k)V_k^\star + (P_i^k - \bar{P}_i^k)(V_k - V_k^\star) + b_i^k + \epsilon_k)_+$$

$$\lesssim L \sum_{k=1}^{K'} \sum_{i=1}^{I_k} \left(\sqrt{\frac{\mathbb{V}(P_i^k, V_k^\star)\iota'}{\mathbf{N}_i^k}} + \sqrt{\frac{\Gamma_{L(1+\epsilon)}\mathbb{V}(P_i^k, V_k - V_k^\star)\iota'}{\mathbf{N}_i^k}} + \frac{\Gamma_{L(1+\epsilon)} L\iota'}{\mathbf{N}_i^k} + b_i^k + \epsilon_k \right)$$

(Eq. (10), Lemma 46, conditions 3) and 4), $\iota' = \log \frac{S^{\rightarrow}_{L(1+\epsilon)} AC_{K'}}{\delta}$)

$$\lesssim L \left(\sqrt{S^{\rightarrow}_{L(1+\epsilon)} A \sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k)\iota'} + \sqrt{{S^{\rightarrow}_{L(1+\epsilon)}}^2 A \sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k - V_k^\star)\iota'} \right) + L^2 {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\iota',$$

where in the last step $\iota' = \log^2 \frac{S^{\rightarrow}_{L(1+\epsilon)} AC_{K'}}{\delta}$ and we apply $\mathrm{VAR}[X_1 + X_2] \leq \mathrm{VAR}[X_1] + \mathrm{VAR}[X_2]$, Cauchy-Schwarz inequality, Lemma 40, Lemma 41, and Lemma 39. Then applying Lemma 38 with $\|V_k^\star - V_k\|_\infty \lesssim L$ and solving a quadratic inequality w.r.t. $\sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k^\star - V_k)$, we have with probability at least $1 - \delta$,

$$\sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k^\star - V_k)$$

$$\lesssim \sum_{k=1}^{K'} (V_k^\star(s_{I_k+1}^k) - V_k(s_{I_k+1}^k))^2 + L\sqrt{S^{\rightarrow}_{L(1+\epsilon)} A \sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k)\iota'} + L^2 {S^{\rightarrow}_{L(1+\epsilon)}}^2 A\iota'. \quad (\iota' = \log^2 \frac{LS^{\rightarrow}_{L(1+\epsilon)} AC_{K'}}{\delta})$$

The proof is completed by noting that $V_k^\star(g) = V_k(g) = 0$ and $\sum_{k=1}^{K'} \mathbb{I}\{s_{I_k+1}^k \neq g\} \lesssim S^{\rightarrow}_{L(1+\epsilon)} A$. $\square$

Lemma 38. *Let $K \in \mathbb{N}$ and $\{V_k\}_{k\in[K]}$ be a sequence of value functions with $V_k \in [0, B]^{\mathcal{S}}$ for $B > 0$. With probability at least $1 - \delta$, for any $K' \in [K]$,*

$$\sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k) \lesssim \sum_{k=1}^{K'} V_k(s_{I_k+1}^k)^2 + \sum_{k=1}^{K'} \sum_{i=1}^{I_k} \left(V_k(s_i^k)^2 - (P_i^k V_k)^2\right) + B^2\iota,$$

where $\iota = \log(BC_{K'}/\delta)$.

Proof. We decompose the sum as follows:

$$\sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k) = \sum_{k=1}^{K'} \sum_{i=1}^{I_k} \left(P_i^k (V_k)^2 - V_k(s_{i+1}^k)^2\right) + \sum_{k=1}^{K'} \sum_{i=1}^{I_k} \left(V_k(s_{i+1}^k)^2 - V_k(s_i^k)^2\right) + \sum_{k=1}^{K'} \sum_{i=1}^{I_k} \left(V_k(s_i^k)^2 - (P_i^k V_k)^2\right).$$

For the first term, by Lemma 55, Lemma 48, and $I_k < \infty$ for any $k \in [K]$ by the skip-round condition, with probability at least $1 - \delta$, for all $K' \in [K]$,

$$\sum_{k=1}^{K'} \sum_{i=1}^{I_k} \left(P_i^k (V_k)^2 - V_k(s_{i+1}^k)^2\right) \lesssim \sqrt{\sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, (V_k)^2)\iota} + B^2\iota$$

$$\lesssim B\sqrt{\sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k)\iota} + B^2\iota,$$

where $\iota = \mathcal{O}(\log(BC_{K'}/\delta))$. The second term is clearly upper bounded by $\sum_{k=1}^{K'} V_k(s_{I_k+1}^k)^2$. Putting everything together and solving a quadratic inequality w.r.t. $\sum_{k=1}^{K'} \sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k)$ completes the proof. $\square$

Lemma 39. *Let $\mathcal{G}$ be the goal set such that $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \subseteq \mathcal{G} \subseteq \mathcal{S}$. In any trial, with probability at least $1-\delta$, for any $K' \in [K]$, if $\mathcal{K}_k \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ and $g_k \in \mathcal{G} \setminus \mathcal{K}_k$ for any $k \in [K']$, then $\sum_{k=1}^{K'}\sum_{i=1}^{I_k} b_i^k \lesssim \sqrt{S^{\rightarrow}_{L(1+\epsilon)} A \sum_{k=1}^{K'}\sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k)\iota} + L{S^{\rightarrow}_{L(1+\epsilon)}}^{1.5} A\iota$, where $\iota = \log(|\mathcal{G}|AC_{K'}/\delta)$.*

Proof. Note that with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=1}^{K'}\sum_{i=1}^{I_k} b_i^k &\lesssim \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(\sqrt{\frac{\mathbb{V}(\bar{P}_i^k, V_k)\iota}{\mathbf{N}_i^k}} + \frac{L\iota}{\mathbf{N}_i^k}\right) && \text{(definition of } b_i^k \text{ and } \max\{a,b\} \le a+b) \\
&\lesssim \sum_{k=1}^{K'}\sum_{i=1}^{I_k}\left(\sqrt{\frac{\mathbb{V}(P_i^k, V_k)\iota}{\mathbf{N}_i^k}} + \frac{L\sqrt{S^{\rightarrow}_{L(1+\epsilon)}}\iota}{\mathbf{N}_i^k}\right) && \text{(Lemma 45)} \\
&\lesssim \sqrt{S^{\rightarrow}_{L(1+\epsilon)} A \sum_{k=1}^{K'}\sum_{i=1}^{I_k} \mathbb{V}(P_i^k, V_k)\iota} + L{S^{\rightarrow}_{L(1+\epsilon)}}^{1.5} A\iota. && \text{(Cauchy-Schwarz inequality and Lemma 40)}
\end{aligned}$$

This completes the proof. □

Lemma 40. *In any trial, for any $K' \in [K]$, if $\mathcal{K}_k \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ for any $k \in [K']$, we have $\sum_{k=1}^{K'}\sum_{i=1}^{I_k} \frac{1}{\mathbf{N}_i^k} \lesssim S^{\rightarrow}_{L(1+\epsilon)} A \log_2(C_{K'})$.*

Proof. Note that, for any i, k, if $s_i^k \notin \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$ we must have $s_i^k \notin \mathcal{K}_k$, which implies that the corresponding count N_i^k is ∞. Then,

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{i=1}^{I_k} \frac{1}{\mathbf{N}_i^k} &\le \sum_{s \in \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}, a \in \mathcal{A}} \; \sum_{0 \le h \le \log_2(C_K)} \sum_{k=1}^{K}\sum_{i=1}^{I_k} \mathbb{I}\left[(s_i^k, a_i^k) = (s,a), \mathbf{N}_i^k(s,a) = 2^h\right]\frac{1}{2^h} \\
&\le |\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}| A \log_2(C_k).
\end{aligned}$$

□

Lemma 41. *In any trial, for any $K' \in [K]$, $\sum_{k=1}^{K'}\sum_{i=1}^{I_k} \epsilon_k = \mathcal{O}(\log C_{K'})$.*

Lemma 42. *In any trial, $1 + \bar{P}_i^k V_k - 2b_i^k - \epsilon_k \le V_k(s_i^k) \le 1 + \bar{P}_i^k V_k + \epsilon_k$ for any $k \in [K], i \in [I_k]$.*

Proof. When $s_i^k \notin \mathcal{K}_k$, we have $b_i^k = \frac{1}{\mathbf{N}_i^k} = 0$ and $\bar{P}_i^k V_k = V_k(s_0)$. Thus, the statement holds. When $s_i^k \in \mathcal{K}_k$, by the definition of V_k and the stopping rule of Algorithm 4, we have

$$\begin{aligned}
V_k(s_i^k) &\ge 1 + \widetilde{P}_i^k V_k - b_i^k - \epsilon_k \ge 1 + \bar{P}_i^k V_k - b_i^k - \epsilon_k - \frac{\bar{P}_i^k V_k}{\mathbf{N}_i^k} && \text{(definition of } \widetilde{P}_i^k) \\
&\ge 1 + \bar{P}_i^k V_k - 2b_i^k - \epsilon_k,
\end{aligned}$$

where the last step is by $\frac{\bar{P}_i^k V_k}{\mathbf{N}_i^k} \le \frac{2L}{\mathbf{N}_i^k} \le b_i^k$. Moreover, $V_k(s_i^k) \le 1 + \widetilde{P}_i^k V_k + \epsilon_k \le 1 + \bar{P}_i^k V_k + \epsilon_k$. This completes the proof. □

G. Auxiliary Results

Lemma 43. *For any $S \geq 1$, $A \geq 2$, $\frac{3}{2} \leq L \leq \frac{1}{2} + \frac{\log(S/2)}{2\log(A)}$, and $0 < \epsilon < \frac{L-1}{L}$, there exists an MDP with S states and A actions (including action* RESET*) such that $S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)} = 1$ while $S^{\rightarrow}_{2L} \geq A^{2(L-1)}$.*

Proof. Consider an MDP with the following structure. At s_0, taking any action transits to one of $\{s_1, \ldots, s_L\}$ with probability $\frac{1}{L}$. At any state in $\{s_1, \ldots, s_L\}$, taking any action transits to state $s^\star$. States reachable from $s^\star$ form a full A-ary tree with depth $2(L-1)$. The rest of the states are ignored (note that $S \geq 2A^{2L-1} \geq 1 + L + \sum_{i=0}^{2(L-1)} A^i$). It is not hard to see that it takes $2L-1$ steps to reach any s_i for $i \in [L]$ by a policy restricted on $\{s_0\}$. Therefore, all unignored states are $2L$ incrementally controllable and thus $S^{\rightarrow}_{2L} \geq A^{2(L-1)}$ states. On the other hand, by $L(1+\epsilon) < 2L-1$, $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)} = \{s_0\}$ and $\Gamma_{L(1+\epsilon)} = 1$ (note that the agent can reach s_0 from s_0 by taking RESET). $\square$

Remark 2. *The construction in Lemma 43 also have $S^{\rightarrow}_{2L} = \Omega(S)$ while $S^{\rightarrow}_{L(1+\epsilon)}\Gamma_{L(1+\epsilon)} = \mathcal{O}(1)$.*

Lemma 44. *For any $\mathcal{X} \subseteq \mathcal{S}$ and $g \in \mathcal{S}$, we have $\left\|V^\star_{\mathcal{X},g}\right\|_\infty \leq 1 + V^\star_{\mathcal{X},g}(s_0)$.*

Proof. Clearly $V^\star_{\mathcal{X},g}(g) = 0 \leq 1 + V^\star_{\mathcal{X},g}(s_0)$ and $V^\star_{\mathcal{X},g}(s) = 1 + V^\star_{\mathcal{X},g}(s_0)$ for any $s \in \mathcal{S} \setminus (\mathcal{X} \cup \{g\})$. For any $s \in \mathcal{X} \setminus \{g\}$, by Bellman optimality and RESET $\in \mathcal{A}$ we have $V^\star_{\mathcal{X},g}(s) \leq 1 + V^\star_{\mathcal{X},g}(s_0)$. $\square$

Lemma 45. *Let n be a counter incrementally collecting samples from transition function P, and define $\bar{P}^n_{s,a}(s') := \frac{n(s,a,s')}{n^+(s,a)}$. Let $\mathcal{G}$ be the goal set such that $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \subseteq \mathcal{G} \subseteq \mathcal{S}$. With probability at least $1-\delta$, for any status of n, $(s,a) \in \mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \times \mathcal{A}$, $\mathcal{X} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, $g \in \mathcal{G} \setminus \mathcal{X}$, and value function V restricted on $\mathcal{X} \cup \{g\}$ with $\|V\|_\infty \leq B$ for some $B > 0$, we have $\mathbb{V}(\bar{P}^n_{s,a}, V) \lesssim \mathbb{V}(P_{s,a}, V) + \frac{\Gamma_{L(1+\epsilon)}B^2\iota'_{s,a}}{n^+(s,a)}$, where $\iota'_{s,a} = \mathcal{O}(\log\frac{|\mathcal{G}|An^+(s,a)}{\delta})$.*

Proof. Note that

$$\begin{aligned}
\mathbb{V}(\bar{P}_{s,a}, V) &\leq \bar{P}_{s,a}(V - P_{s,a}V)^2 && \left(\tfrac{\sum_i p_i x_i}{\sum_i p_i} = \operatorname{argmin}_z \textstyle\sum_i p_i(x_i - z)^2\right)\\
&= \mathbb{V}(P_{s,a}, V) + (\bar{P}_{s,a} - P_{s,a})(V - P_{s,a}V)^2\\
&\lesssim \mathbb{V}(P_{s,a}, V) + B\sqrt{\frac{\Gamma_{L(1+\epsilon)}\mathbb{V}(P_{s,a},V)\iota'_{s,a}}{n^+(s,a)}} + \frac{\Gamma_{L(1+\epsilon)}B^2\iota'_{s,a}}{n^+(s,a)} && \text{(Lemma 46 and Lemma 48)}\\
&\lesssim \mathbb{V}(P_{s,a}, V) + \frac{\Gamma_{L(1+\epsilon)}B^2\iota'_{s,a}}{n^+(s,a)}. && \text{(AM-GM inequality)}
\end{aligned}$$

This completes the proof. $\square$

Lemma 46. *Let n be a counter incrementally collecting samples from transition function P, and define $\bar{P}^n_{s,a}(s') := \frac{n(s,a,s')}{n^+(s,a)}$. Let $\mathcal{G}$ be the goal set such that $\mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \subseteq \mathcal{G} \subseteq \mathcal{S}$.[8] With probability at least $1-\delta$, for any status of n, $(s,a) \in \mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \times \mathcal{A}$, $\mathcal{X} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$, $g \in \mathcal{G} \setminus \mathcal{X}$, and value function V restricted on $\mathcal{X} \cup \{g\}$ with $\|V\|_\infty \leq B$ for some $B > 0$, we have*

$$|(P_{s,a} - \bar{P}^n_{s,a})V| \lesssim \sqrt{\frac{\min\{|\mathcal{X}|, \Gamma^{s,a}_{L(1+\epsilon)}\}\mathbb{V}(P_{s,a},V)\iota'_{s,a}}{n^+(s,a)}} + \frac{B\min\{|\mathcal{X}|, \Gamma^{s,a}_{L(1+\epsilon)}\}\iota'_{s,a}}{n^+(s,a)},$$

where $\iota'_{s,a} = \mathcal{O}(\log\frac{S^{\rightarrow}_{L(1+\epsilon)}A\Gamma^2_{L(1+\epsilon)}|\mathcal{G}|n^+(s,a)}{\delta})$.

Proof. By Lemma 54 and a union bound, for any $\delta' \in (0,1)$, with probability at least $1 - \frac{\delta'}{S^{\rightarrow}_{L(1+\epsilon)}A\Gamma_{L(1+\epsilon)}\binom{\Gamma^{s,a}_{L(1+\epsilon)}}{i}|\mathcal{G}|}$, for each status of n, $(s,a) \in \mathcal{S}^{\rightarrow}_{L(1+\epsilon)} \times \mathcal{A}$, size $i \in [\Gamma^{s,a}_{L(1+\epsilon)}]$, subset $y' \subseteq \mathcal{N}^{s,a}_{L(1+\epsilon)}$ with $|y'| = i$, and $g \in \mathcal{G} \setminus y'$,

$$|P_{s,a}(y) - \bar{P}^n_{s,a}(y)| \leq 2\sqrt{2\frac{P_{s,a}(y)(1-P_{s,a}(y))\log(2n^+(s,a)/\delta')}{n^+(s,a)}} + \frac{\log(2n^+(s,a)/\delta')}{n^+(s,a)},$$

[8] In most cases, we apply this lemma with $\mathcal{G} \in \{\mathcal{S}^{\rightarrow}_{L(1+\epsilon)}, \mathcal{S}\}$.

where $y = \mathcal{S} \setminus (y' \cup \{g\})$. Let $y' = \mathcal{X}' \triangleq \mathcal{X} \cap \mathcal{N}^{s,a}_{L(1+\epsilon)}$ such that $y = \mathcal{S} \setminus (\mathcal{X}' \cup \{g\})$. By another application of Lemma 54 and a union bound, for any $\delta' \in (0, 1)$, with probability at least $1 - \frac{\delta'}{|\mathcal{G}|}$, for all $s' \in \mathcal{X}' \cup \{g\} \subseteq \mathcal{G}$,

$$|P_{s,a}(s') - \bar{P}^n_{s,a}(s')| \leq 2\sqrt{2\frac{P_{s,a}(s')(1 - P_{s,a}(s'))\log(2n^+(s,a)/\delta')}{n^+(s,a)}} + \frac{\log(2n^+(s,a)/\delta')}{n^+(s,a)}.$$

Thus, setting $\delta' = \delta/2S^{\rightarrow}_{L(1+\epsilon)}A\Gamma_{L(1+\epsilon)}\binom{\Gamma^{s,a}_{L(1+\epsilon)}}{i}|\mathcal{G}|$ and using $\binom{n}{i} \leq n^{\min\{i,n-i\}}$, the two inequalities above simplify as

$$|P_{s,a}(y) - \bar{P}^n_{s,a}(y)| \lesssim \sqrt{\frac{i \cdot P_{s,a}(y)(1 - P_{s,a}(y))\iota'_{s,a}}{n^+(s,a)}} + \frac{i\iota'_{s,a}}{n^+(s,a)}, \tag{11}$$

$$|P_{s,a}(s') - \bar{P}^n_{s,a}(s')| \lesssim \sqrt{\frac{P_{s,a}(s')(1 - P_{s,a}(s'))\iota'_{s,a}}{n^+(s,a)}} + \frac{\iota'_{s,a}}{n^+(s,a)}. \tag{12}$$

These hold with probability at least $1 - \delta$. Now define, for all $s' \in \mathcal{S}$,

$$V'(s') = \begin{cases} V(s'), & s' \in \mathcal{X}' \cup \{g\} \\ V(\mathcal{S} \setminus (\mathcal{X} \cup \{g\})), & \text{otherwise} \end{cases}$$

and $V_\dagger(s') = V'(s') - P_{s,a}V'$ for all s'. Clearly, V' and $V_\dagger$ are restricted on $\mathcal{X}' \cup \{g\}$. Moreover, $V(s') \neq V'(s') \implies s' \in \mathcal{X} \setminus y' \implies s' \in \mathcal{X} \setminus \mathcal{N}^{s,a}_{L(1+\epsilon)} \implies P_{s,a}(s') = 0$ by $\mathcal{X} \subseteq \mathcal{S}^{\rightarrow}_{L(1+\epsilon)}$. Thus, $P_{s,a}V = P_{s,a}V'$, and

$$\begin{aligned}
&(P_{s,a} - \bar{P}^n_{s,a})V = (P_{s,a} - \bar{P}^n_{s,a})V' = (P_{s,a} - \bar{P}^n_{s,a})V_\dagger \\
&= \sum_{s' \in \mathcal{X}'} (P_{s,a}(s') - \bar{P}^n_{s,a}(s'))V_\dagger(s') + (P_{s,a}(g) - \bar{P}^n_{s,a}(g))V_\dagger(g) + (P_{s,a}(y) - \bar{P}^n_{s,a}(y))V_\dagger(y) \\
&\lesssim \sum_{s' \in \mathcal{X}' \cup \{g\}} \sqrt{\frac{P_{s,a}(s')\iota'_{s,a}}{n^+(s,a)}}|V_\dagger(s')| + \sqrt{\frac{|\mathcal{X}'|P_{s,a}(y)\iota'_{s,a}}{n^+(s,a)}}|V_\dagger(y)| + \frac{B|\mathcal{X}'|\iota'_{s,a}}{n^+(s,a)} && \text{(Eq. (11) and Eq. (12))} \\
&\lesssim \sqrt{\frac{|\mathcal{X}'|\mathbb{V}(P_{s,a}, V)\iota'_{s,a}}{n^+(s,a)}} + \frac{B|\mathcal{X}'|\iota'_{s,a}}{n^+(s,a)}.
\end{aligned}$$

where in the last step we apply Cauchy-Schwarz inequality and

$$\begin{aligned}
\sum_{s'} P_{s,a}(s')V_\dagger(s')^2 &= \sum_{s'} P_{s,a}(s')(V'(s') - P_{s,a}V)^2 && (P_{s,a}V = P_{s,a}V') \\
&= \sum_{s'} P_{s,a}(s')(V(s') - P_{s,a}V)^2 && (P_{s,a}(s') = 0 \text{ when } V'(s') \neq V(s')) \\
&= \mathbb{V}(P_{s,a}, V).
\end{aligned}$$

This completes the proof. □

Lemma 47. *If $x \leq a\sqrt{x\log^p(dx)} + b\log^p(dx) + c$ for some $a, b, c \geq 0$, $d > 0$ and some absolute constant $p \geq 1$, then $x = \mathcal{O}((a^2 + b)\log^p((a + b + c)d) + c)$.*

Proof. By AM-GM inequality and $\log x < x$ for $x > 0$, we have

$$x \leq a\sqrt{x\log^p(dx)} + b\log^p(dx) + c \leq \frac{x}{2} + (a^2/2 + b)\log^p(dx) + c \leq \frac{x}{2} + (a^2/2 + b)(2p)^p\sqrt{dx} + c.$$

Solving a quadratic inequality w.r.t. x gives $x = \mathcal{O}((a^2 + b)^2 d + c)$. Plugging this back to the original inequality gives $x \leq a\sqrt{x\iota} + b\iota + c$, where $\iota = \log^p((a + b + c)d)$. Further solving a quadratic inequality w.r.t x completes the proof. □

Lemma 48. *(Chen et al., 2023, Lemma 40) For any random variable $X \in [-B, B]$, for some $B > 0$, we have* $\textsc{Var}[X^2] \leq 4B^2\textsc{Var}[X]$.

Lemma 49. *(Cai et al., 2022, Lemma C.2) For some $B > 0$, let $\Upsilon = \{v \in \mathbb{R}^{\mathcal{S}}_{\geq 0} : v(g) = 0, \|v\|_\infty \leq B\}$ and $f : \Delta_{\mathcal{S}} \times \Delta_{\mathcal{S}} \times \Upsilon \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ with $f(\widetilde{p}, p, v, n, \iota) = \widetilde{p}v - \max\left\{c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}}, c_2\frac{B\iota}{n}\right\}$ with some constants $c_1 \geq 0$ and $c_2 \geq 2c_1^2$. Then f ensures for all v, n, ι, and $\widetilde{p}$, p s.t. $\widetilde{p}(s) - \frac{1}{2}p(s) \geq 0$ for all $s \neq g$,*

1. *$f(\widetilde{p}, p, v, n, \iota)$ is non-decreasing in $v(s)$, that is,*

$$\forall v, v' \in \Upsilon, v \leq v' \implies f(\widetilde{p}, p, v, n, \iota) \leq f(\widetilde{p}, p, v', n, \iota);$$

2. *if $\widetilde{p}(g) > 0$, then $f(\widetilde{p}, p, v, n, \iota)$ is $\rho_{\widetilde{p}}$-contractive in $v(s)$, with $\rho_{\widetilde{p}} = 1 - \widetilde{p}(g) < 1$, that is,*

$$\forall v, v' \in \Upsilon, |f(\widetilde{p}, p, v, n, \iota) - f(\widetilde{p}, p, v', n, \iota)| \leq \rho_{\widetilde{p}} \|v - v'\|_\infty .$$

Lemma 50. *There exist a function $N_{\text{DEV}}(L_0, \epsilon, \delta) = \mathcal{O}(\log^4 \frac{L_0}{\epsilon\delta}/\epsilon^2)$, such that for any $g \in \mathcal{S}$ and policy π with $\left\|V_g^\pi\right\|_\infty \leq L_0$ for some $L_0 > 0$, we have with probability at least $1 - \delta$, for all $n \geq N_{\text{DEV}}(L_0, \epsilon, \delta)$ simultaneously, $|\widehat{\tau}_n - V_g^\pi(s_0)| \leq \left\|V_g^\pi\right\|_\infty \epsilon$, where $\widehat{\tau}_n = \frac{1}{n}\sum_{i=1}^n C_i$ and each C_i is a realization of the total cost incurred by following π starting from s_0 with goal state g.*

Proof. By Lemma 51, with probability at least $1 - \delta$, $\left|\widehat{\tau}_n - V_g^\pi(s_0)\right| \leq \frac{8\left\|V_g^\pi\right\|_\infty}{\sqrt{n}} \log^2 \frac{8n^2\left\|V_g^\pi\right\|_\infty}{\delta}$ for all $n \geq 1$. Solving the range of n for the inequality $\frac{8\left\|V_g^\pi\right\|_\infty}{\sqrt{n}} \log^2 \frac{8n^2 L_0}{\delta} \leq \left\|V_g^\pi\right\|_\infty \epsilon$ (Lemma 47) completes the proof. □

Lemma 51. *For any $g \in \mathcal{S}$ and policy π with $\left\|V_g^\pi\right\|_\infty \leq L_0$ for some $L_0 \geq 1$, we have with probability at least $1 - \delta$, for all $n \geq 1$ simultaneously, $|\widehat{\tau}_n - V_g^\pi(s_0)| \leq \frac{8L_0}{\sqrt{n}} \log^2 \frac{8n^2 L_0}{\delta}$, where $\widehat{\tau}_n = \frac{1}{n}\sum_{i=1}^n C_i$ and each C_i is a realization of the total cost incurred by following π starting from s_0 with goal state g.*

Proof. By Lemma 52 and a union bound,

$$\mathbb{P}\left(\exists i \geq 1 : C_i > 4L_0 \log \frac{8i^2 L_0}{\delta}\right) \leq \sum_{i \geq 1} \mathbb{P}\left(C_i > 4L_0 \log \frac{8i^2 L_0}{\delta}\right) \leq \sum_{i \geq 1} \frac{\delta}{4i^2 L_0} \leq \frac{\delta}{2}.$$

Then, under the complement of the event above (which holds with probability at least $1 - \frac{\delta}{2}$), we have $\bar{\tau}_n = \widehat{\tau}_n$ for all $n \geq 1$, where $\bar{\tau}_n = \frac{1}{n}\sum_{i=1}^n C_i \mathbb{I}\{C_i \leq 4L_0 \log \frac{8n^2 L_0}{\delta}\}$. Moreover, by Lemma 53 and a union bound,

$$\mathbb{P}\left(\exists n \geq 1 : |\bar{\tau}_n - \mathbb{E}[\bar{\tau}_n]| > 4L_0 \log \frac{8n^2 L_0}{\delta}\sqrt{\frac{2\log\frac{8n^2}{\delta}}{n}}\right) \leq \sum_{n \geq 1} \frac{\delta}{4n^2} \leq \frac{\delta}{2}.$$

A union bound on the complement of the two events above yields that, with probability at least $1 - \delta$, for all $n \geq 1$ simultaneously,

$$\widehat{\tau}_n - V_g^\pi(s_0) = \bar{\tau}_n - V_g^\pi(s_0) \leq \bar{\tau}_n - \mathbb{E}[\bar{\tau}_n] \leq 4L_0 \log \frac{8n^2 L_0}{\delta}\sqrt{\frac{2\log\frac{8n^2}{\delta}}{n}},$$

and by Lemma 52,

$$V_g^\pi(s_0) - \widehat{\tau}_n \leq \mathbb{E}[\bar{\tau}_n] - \bar{\tau}_n + L_0 \cdot \frac{1}{2nL_0} \leq 4L_0 \log \frac{8n^2 L_0}{\delta}\sqrt{\frac{2\log\frac{8n^2}{\delta}}{n}} + \frac{1}{2n}.$$

Combining these two cases gives $\left|\widehat{\tau}_n - V_g^\pi(s_0)\right| \leq \frac{8L_0}{\sqrt{n}} \log^2 \frac{8n^2 L_0}{\delta}$. □

Lemma 52. *(Cohen et al., 2020, Lemma B.5) For a given $g \in \mathcal{S}$, let π be a policy such that $\left\|V_g^\pi\right\|_\infty \leq \tau$. Then, for any $n \in \mathbb{N}$, the probability that the cost of π to reach the goal state starting from any state is more than n, is at most $2e^{-\frac{n}{4\tau}}$.*

Lemma 53 (Azuma's inequality)**.** *Let $\{X_t\}_{t=1}^n$ be a martingale difference sequence with $|X_t| \leq B$. Then with probability at least $1-\delta$, $|\sum_{t=1}^n X_i| \leq B\sqrt{2n\log\frac{2}{\delta}}$.*

Lemma 54. *(Chen et al., 2021, Lemma 34) Let $\{X_t\}_t$ be a sequence of i.i.d random variables with mean μ, variance σ^2, and $0 \leq X_t \leq B$. Then with probability at least $1-\delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\left|\sum_{t=1}^n (X_t - \mu)\right| \leq 2\sqrt{2\sigma^2 n \log\frac{2n}{\delta}} + 2B\log\frac{2n}{\delta}.$$

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

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 55. *(Chen et al., 2022b, Lemma 50) Let $\{X_i\}_{i=1}^\infty$ be a martingale difference sequence adapted to the filtration $\{\mathcal{F}_i\}_{i=0}^\infty$ and $|X_i| \leq B$ for some $B > 0$. Then with probability at least $1-\delta$, for all $n \geq 1$ simultaneously,*

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