arXiv:2205.10729v1 [cs.LG] 22 May 2022

Near-Optimal Algorithms for Autonomous Exploration and Multi-Goal Stochastic Shortest Path

Haoyuan Cai [1] **Tengyu Ma** [2] **Simon Du** [3]

Abstract

We revisit the incremental autonomous exploration problem proposed by Lim & Auer (2012). In this setting, the agent aims to learn a set of near-optimal goal-conditioned policies to reach the L-controllable states: states that are incrementally reachable from an initial state s_0 within L steps in expectation. We introduce a new algorithm with stronger sample complexity bounds than existing ones. Furthermore, we also prove the first lower bound for the autonomous exploration problem. In particular, the lower bound implies that our proposed algorithm, Value-Aware Autonomous Exploration, is nearly minimax-optimal when the number of L-controllable states grows polynomially with respect to L. Key in our algorithm design is a connection between autonomous exploration and multi-goal stochastic shortest path, a new problem that naturally generalizes the classical stochastic shortest path problem. This new problem and its connection to autonomous exploration can be of independent interest.

1. Introduction

Reinforcement learning (RL) with a known state space has been studied in a wide range of settings (e.g., Schmidhuber, 1991; Oudeyer et al., 2007; Oudeyer & Kaplan, 2009; Baranes & Oudeyer, 2009). When the state space is large, it is difficult for a learning agent to discover the whole environment. Instead, the agent can only explore a small portion of the environment. At a high level, we hope that the agent can discover states near the initial state, expand the range of known states by exploration, and learn near-optimal goal-conditioned policies for the known states. Because the agent discovers its known states of the environment incrementally, this learning problem was named Autonomous Exploration (AX) (Lim & Auer, 2012; Tarbouriech et al., 2020).

The autonomous exploration problem generalizes the Stochastic Shortest Path (SSP) problem (Bertsekas et al., 2000) where the agent aims to reach a predefined goal state while minimizing its total expected cost. However, in the autonomous exploration setting, the agent aims to discover a set of reachable states in a large environment and find the optimal policies to reach them. The autonomous exploration formulation is applicable to an increasing number of real-world RL problems, ranging from navigation in mazes (Devo et al., 2020) to game playing (Mnih et al., 2013). For example, in the maze navigation problem, a robot aims to follow a predefined path in an unknown environment, and the robot has to discover and expand the size of regions known to itself autonomously without prior knowledge of the environment. See (Lim & Auer, 2012) for more discussions.

Related Work. The setting of autonomous exploration (AX) was introduced by Lim & Auer (2012), who gave the first algorithm, UcbExplore, with sample complexity $\widetilde{O}(L^3S^2A/\varepsilon^3)$. Here L denotes the distance within which we hope the learning agent to discover, S denotes the number of states we need to explore,[1] A denotes the size of the action space, and ε denotes the error that we can tolerate. Recent work by Tarbouriech et al. (2020) designed the DisCo algorithm with a sample complexity bound $\widetilde{O}(L^3S^2A/\varepsilon^2)$[2], which improves the $1/\varepsilon$ dependency. We will briefly discuss the two algorithms in Sect. 2.1. In this paper, we present a new algorithm, `VALAE` (Alg. 2), to further improve the sample complexity, and we also derive the first lower bound.

[1]Tsinghua University [2]Stanford University [3]University of Washington. Correspondence to: Haoyuan Cai <victorique1929@gmail.com>, Tengyu Ma <tengyuma@stanford.edu>, Simon Du <ssdu@cs.washington.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1]In AX, S is often significantly smaller than the size of the entire state space.

[2]We translate their absolute error ε_{abs} to the relative error ε_{rel}, and $\varepsilon_{\text{abs}} = \varepsilon_{\text{rel}}L$. We will explain the difference of two definitions of ε in Sect. 2.1.

Algorithm	Sample Complexity
UcbExplore (Lim & Auer, 2012)	$\widetilde{O}(L^3S^2A/\varepsilon^3)$
DisCo (Tarbouriech et al., 2020)	$\widetilde{O}(L^3S^2A/\varepsilon^2)$
VALAE	$\widetilde{O}(LSA/\varepsilon^2)$
Lower Bound	$\Omega(LSA/\varepsilon^2)$

Table 1: Comparisons between our results and prior results. Algorithms and results in this paper are in grey cells. L is the exploration radius, A is the number of actions, S is the number of states we need to explore, and ε is the target accuracy. We will define them in Sect. 2. For simplicity, we only display the leading term in terms of the scaling in $1/\varepsilon$.

1.1. Contributions

In this paper, we take important steps toward resolving the autonomous exploration problem. We compare our results with prior ones in Table 1.[3] and we summarize our contributions below:

1. We propose a new algorithm for autonomous exploration problem, `Value-Aware Autonomous Exploration` (`VALAE`), which uses DisCo algorithm (Tarbouriech et al., 2020) and `Re-MG-SSP` (cf. Alg. 1) as initial steps and then uses the estimated value functions to guide our exploration. By doing so, for each state-action pair (s, a), we derive an (s, a)-dependent sample complexity bound, which can exploit the variance information, and yield a sharper sample complexity bound than the bounds for UcbExplore algorithm and DisCo algorithm (cf. Table 1). In particular, `VALAE` improves the dependency on L from cubic to linear, and improves the dependency on S from square to linear.
2. We connect the autonomous exploration problem to a new problem, multi-goal stochastic shortest path, which generalizes classical SSP. And we show that `VALAE` also applies to multi-goal SSP.
3. We give the first lower bound of the autonomous exploration problem. This lower bound shows `VALAE` is nearly minimax-optimal when the number of states we need to explore grows polynomially with respect to L.

[3] In (Lim & Auer, 2012), the cost is 1 uniformly for all state-action pairs. In this paper, we allow non-uniform costs. In Table 1, we consider uniform costs for fair comparisons.

1.2. Main Difficulties and Technique Overview

While our work borrows ideas from prior work on autonomous exploration (Lim & Auer, 2012; Tarbouriech et al., 2020) and recent advances in SSP (Tarbouriech et al., 2021), we develop new techniques to overcome additional difficulties that are unique in autonomous exploration.

Connection between Autonomous Exploration and Multi-Goal SSP. In standard RL setting, it is known that in order to obtain a tight dependency on L, one needs to exploit the variance information in the value function (Azar et al., 2017). However, in autonomous exploration, it is unclear how to exploit the variance information because even which state is reachable is unknown.

To this end, we first consider a simpler problem, multi-goal SSP, and extend the technique for single-goal SSP (Tarbouriech et al., 2021) to this new problem (cf. Alg. 2). We also present a reduction from autonomous exploration to multi-goal SSP (cf. Alg. 1). These two techniques together yield the first tight dependency on L for autonomous exploration.

Using Regret to Bound the Sample Complexity. To estimate the sample complexity of `VALAE`, we need to bound the total number of rounds r. Inspired by (Lim & Auer, 2012), we classify each round into three categories: failure round, success round, and skipped round. Moreover, we adopt the idea of using regret bound.

A failure round has regret larger than $\widetilde{\Omega}(L/\varepsilon)$, but the number of failure rounds r_f is hard to estimate. The number of success rounds and skipped rounds are bounded by $\widetilde{O}(SA)$, but the regret in a success round or skipped round can be negative. Hence, to bound the total number of failure rounds r_f, careful analyses of both the upper bound and the lower bound of regret are required.

For the upper bound, we extend the techniques of variance analysis from classical SSP (cf. (Tarbouriech et al., 2021)) to this problem, and we obtain the upper bound of regret scaling as $\widetilde{O}(\sqrt{r_f})$. For the lower bound, the total regret in all the failure rounds grows linearly with respect to r_f, and we use concentration inequalities to lower bound the total regret in success rounds and skipped rounds (cf. Lem. D.3.) By solving the inequality that the lower bound of regret is no more than the upper bound, we can obtain an upper bound of r_f, and we can finally bound the total number of rounds r.

2. Preliminaries

Notations. For any two vectors $X, Y \in \mathbb{R}^S$, we write their inner product as $XY := \sum_{s\in\mathcal{S}} X(s)Y(s)$. We denote $\|X\|_\infty := \max_{s\in\mathcal{S}} |X(s)|$, and if X is a probability distri-

bution on $\mathcal{S}$, we define $\mathbb{V}(X,Y) := \sum_{s\in\mathcal{S}} X(s)Y(s)^2 - (\sum_{s\in\mathcal{S}} X(s)Y(s))^2$, i.e. the variance of random variable Y over distribution X.

Markov Decision Process. We consider an MDP $M := \langle \mathcal{S}, \mathcal{A}, P, c, s_0\rangle$, where $\mathcal{S}$ is the state space with size S, $\mathcal{A}$ is the action space with size A, and $s_0 \in \mathcal{S}$ is the initial state. In state s, taking action a has a cost drawn i.i.d. from a distribution on $[c_{\min}, 1]$ (where $c_{\min} > 0$) with expectation $c(s,a)$, and transits to the next state s' with probability $P(s'|s,a)$. For convenience, we use $P_{s,a}$ and $P_{s,a,s'}$ to denote $P(\cdot|s,a)$ and $P(s'|s,a)$, respectively. A deterministic and stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping, and the agent following the policy π will take action $\pi(s)$ at state s.

For a fixed state $g \in \mathcal{S}$ we define the random variable $t_g^\pi(s)$ as the number of steps it takes to reach state g starting from state s when executing policy π, i.e. $t_g^\pi(s) := \inf\{t \ge 0 : s_{t+1} = g \mid s_1 = s, \pi\}$. A policy π is a proper policy if for any state $s \in \mathcal{S}$, $t_g^\pi(s) < +\infty$ with probability 1. Then we define the value function of a proper policy π with respect to the goal state g and its corresponding Q-function as follows:

$$V_g^\pi(s) = \mathbb{E}\Big[\sum_{t=1}^{t_g^\pi(s)} c_t(s_t, \pi(s_t)) \mid s_1 = s\Big],$$
$$Q_g^\pi(s,a) = \mathbb{E}\Big[\sum_{t=1}^{t_g^\pi(s)} c_t(s_t, \pi(s_t)) \mid s_1 = s, \pi(s_1) = a\Big],$$

where $c_t \in [c_{\min}, 1]$ is the instantaneous cost at step t incurred by the state-action pair $(s_t, \pi(s_t))$, and the expectation is taken over the random sequence of states generated by executing π starting from state $s \in \mathcal{S}$. Here we have $V_g^\pi(g) = 0$. We use π_Q to denote the greedy policy over a vector $Q \in \mathbb{R}^{S\times A}$, i.e. $\pi_Q(s) := \arg\min_{a\in\mathcal{A}} Q(s,a)$.

For a fixed state $g \in \mathcal{S}$, we denote V_g^* as the value function of the optimal policy on MDP M with respect to goal state g, and here we list some important properties of V_g^*: there exists a stationary, deterministic and proper policy π^*, such that its value function $V_g^* := V_g^{\pi^*}$ and its corresponding Q-function $Q_g^* := Q_g^{\pi^*}$ satisfies the following Bellman optimality equations (cf. Lem. A.1):

$$Q_g^*(s,a) = c(s,a) + P_{s,a}V_g^*, \quad V_g^*(s) = \min_{a\in\mathcal{A}} Q_g^*(s,a).$$

We stress that in our setting, given an MDP M, the agent knows the state space $\mathcal{S}$, the action space $\mathcal{A}$, the constant $c_{\min}$, but the agent has no prior knowledge of the transition model P or the cost function c. In each step t, the agent knows its current state $s_t \in \mathcal{S}$, and taking an action $a_t \in \mathcal{A}$ will transit to another state s_t' with some cost c_t.

Incrementally L-controllable States. Before we introduce the Autonomous Exploration problem, we need to define incrementally L-controllable states, which are the states we need to explore. To formally discuss the setting, we need the following assumption on our MDP M.

Assumption 2.1. The action space contains a RESET action s.t. $P(s_0|s, \text{RESET}) = 1$ for any $s \in \mathcal{S}$. Moreover, taking RESET in any state s will incur a cost c_{RESET} with probability 1, where c_{RESET} is a constant in $[c_{\min}, 1]$.

Given any fixed length $L \ge 1$, the agent needs to learn the set of incrementally controllable states $\mathcal{S}_L^{\rightarrow}$. To introduce the concept of $\mathcal{S}_L^{\rightarrow}$, we first give the definition of policies restricted on a subset:

Definition 2.2 (Policy restricted on a subset)**.** For any $\mathcal{S}' \subseteq \mathcal{S}$, a policy π is restricted on the set $\mathcal{S}'$ if $\pi(s) = \text{RESET}$ for all $s \notin \mathcal{S}'$.

Now we discuss the optimal policy restricted on a set of states $\mathcal{K} \subseteq \mathcal{S}$ with respect to goal state g. We denote $V_{\mathcal{K},g}^* \in \mathbb{R}^S$ as the value function of the optimal policy restricted on $\mathcal{K}$ with goal $g \in \mathcal{S}$, and $Q_{\mathcal{K},g}^*$ as the Q-function corresponding to $V_{\mathcal{K},g}^*$. We consider the case that there exists at least one proper policy restricted on $\mathcal{K}$ with the goal state g. Then, $V_{\mathcal{K},g}^*$ and $Q_{\mathcal{K},g}^*$ are finite, and they satisfy the following Bellman equations:

$$\begin{aligned}
Q_{\mathcal{K},g}^*(s,a) &= c(s,a) + P_{s,a}V_{\mathcal{K},g}^*, && \forall (s,a) \in \mathcal{S}\times\mathcal{A},\\
V_{\mathcal{K},g}^*(s) &= \min_{a\in\mathcal{A}} Q_{\mathcal{K},g}^*(s,a), && \forall s \in \mathcal{K}, s \ne g,\\
V_{\mathcal{K},g}^*(s) &= c_{\text{RESET}} + V_{\mathcal{K},g}^*(s_0), && \forall s \notin \mathcal{K}, s \ne g,\\
V_{\mathcal{K},g}^*(g) &= 0.
\end{aligned}$$

We note that when $\mathcal{K}_1 \subseteq \mathcal{K}_2$, for any $g \in \mathcal{S}$, if $V_{\mathcal{K}_1,g}^*$ is finite, then $V_{\mathcal{K}_2,g}^*$ is also finite, and we have $V_{\mathcal{K}_2,g}^* \le V_{\mathcal{K}_1,g}^*$ component-wise. Also, we have $V_g^* = V_{\mathcal{S},g}^*$ component-wise.

Now we introduce the definition of incrementally controllable states $\mathcal{S}_L^{\rightarrow}$ (see (Tarbouriech et al., 2020) for more intuitions on this definition.):

Definition 2.3 (Incrementally L-controllable states $\mathcal{S}_L^{\rightarrow}$)**.** Let $\prec$ be any partial order on $\mathcal{S}$. We denote $\mathcal{S}_L^{\prec}$ as the set of states reachable from s_0 with expected cost no more than L w.r.t. $\prec$, which is defined as follows:

- $s_0 \in \mathcal{S}_L^{\prec}$,
- if there is a policy π restricted on $\{s' \in \mathcal{S}_L^{\prec} : s' \prec s\}$ such that $V_s^\pi(s_0) \le L$, then $s \in \mathcal{S}_L^{\prec}$.

The set of incrementally L-controllable states $\mathcal{S}_L^{\rightarrow}$ is given by $\mathcal{S}_L^{\rightarrow} = \bigcup_{\prec} \mathcal{S}_L^{\prec}$. And we denote $S_L = |\mathcal{S}_L^{\rightarrow}|$.

Multi-Goal Stochastic Shortest Path. Now we define the multi-goal SSP problem, a natural generalization of the classical SSP problem. In multi-goal SSP, we consider an MDP M that satisfies Asmp. 2.1, and all of its states are incrementally L-controllable, i.e. $\mathcal{S}_L^{\rightarrow} = \mathcal{S}$. Also, we assume that the agent knows L.

A learning algorithm for multi-goal SSP takes the error parameter $\varepsilon \in (0,1)$, confidence $\delta \in (0,1)$, and the goal space $\mathcal{G} \subseteq \mathcal{S}$ as input, and with probability over $1-\delta$, the algorithm outputs a set of policies $\{\pi_s\}_{s\in\mathcal{G}}$, such that

$$\forall s \in \mathcal{G}, V_s^{\pi_s}(s_0) \le V_s^*(s_0) + \varepsilon L,$$

i.e., the algorithm learns near-optimal policies to reach each $s \in \mathcal{G}$. We note that when the goal space $\mathcal{G}$ contains a single element, the problem will reduce to classical SSP.

In multi-goal SSP problem, the learning agent interacts with MDP M in this way: the agent knows its current state s and action space $\mathcal{A}$, but it does not know the model $P(s' \mid s,a)$ and cost function $c(s,a)$. Each time, the agent can choose an action $a \in \mathcal{A}$, and the agent will observe that it transits to a new state s' with a cost c, where s' and c are revealed to the agent. The agent can stop and output the policies anytime when the agent thinks that it has collected enough samples to ensure that it can output near-optimal policies.

The performance of the learning algorithm is measured by the cumulative cost C_T, which is defined as follows. We denote T as the total number of steps the agent uses, and we remark that T is random and chosen by the agent. We denote (s_t, a_t) as the state-action pair at the t-th step. We denote by $c_t(s_t,a_t)$ the instantaneous cost incurred at the t-th step. Then we can define $C_T := \sum_{t=1}^{T} c_t(s_t,a_t)$.

We want to find an algorithm with a probably approximately correct (PAC) bound of C_T, i.e., with probability over $1-\delta$, C_T is bounded by some polynomial of $L, S, A, \varepsilon^{-1}, c_{\min}^{-1}$, and $\log(1/\delta)$.

Here we explain the reason why we need the RESET action (Asmp. 2.1). The classical SSP problem uses an episodic learning protocol, i.e. when the agent reaches the goal state g, the agent can "reset" to initial state s_0 and start a new episode. But in multi-goal SSP, we do not have episode learning protocol because we need to ensure that for each goal $g \in \mathcal{G}$, the agent learns a near-optimal policy to reach g. Therefore, each time when the agent arrives at any of the goal, the agent has to "reset" to s_0. Hence the RESET action is necessary, and the previous works (Lim & Auer, 2012) and (Tarbouriech et al., 2020) also assume the existence of the RESET action.

We also remark that multi-goal SSP is fundamentally different from reward-free RL (Jin et al., 2020). Reward-free RL contains two phases: exploration phase and planning phase. In exploration phase we have no knowledge of reward r, and in planning phase we cannot interact with MDP. But in multi-goal SSP, we can estimate the cost function c, and the agent does not need to separate into two phases.

Autonomous Exploration. Now we introduce the autonomous exploration (AX) problem, which generalizes multi-goal SSP. AX problem was first introduced by (Lim & Auer, 2012), and we use their definition of AX problem.

In AX, we consider an MDP M that satisfies Asmp. 2.1. A learning algorithm of AX problem inputs the exploration radius $L \ge 1$, the error parameter $\varepsilon \in (0,1)$ and confidence $\delta \in (0,1)$, and with probability over $1-\delta$, the algorithm should output a set of "known" states $\mathcal{K} \subseteq \mathcal{S}$ such that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}$, i.e., the algorithm discovers all the states that we want to explore. And the algorithm should also output a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$, such that

$$\forall s \in \mathcal{S}_L^{\rightarrow}, V_s^{\pi_s}(s_0) \le (1+\varepsilon)L,$$

i.e., the algorithm learns a policy to reach each $s \in \mathcal{S}_L^{\rightarrow}$ and the expected cost is no more than $(1+\varepsilon)L$. In AX, we also use cumulative cost C_T to measure the performance, but we hope C_T depends on $|\mathcal{S}_L^{\rightarrow}|$ instead of the global size $|\mathcal{S}|$.

We note that different complexity bounds of C_T may depend on $S_L, S_{2L}, S_{(1+\varepsilon)L}$. But if we assume that S_L grows polynomially with respect to L, i.e., there exist constants C, d independent of L, such that $S_L \le CL^d$ for all $L \ge 1$, we will have $S_{2L} \le C2^dL^d = O(L^d)$, and $S_{(1+\varepsilon)L} = O(L^d)$. Under this assumption, $S_L, S_{2L}, S_{(1+\varepsilon)L}$ are of the same order $O(L^d)$, thus we use S as the abbreviation for all these quantities in Table 1. This assumption is implicitly considered in the literature, because otherwise one may need to consider the logarithmic dependency on S_L.

In AX, the learning agent does not know the set $\mathcal{S}_L^{\rightarrow}$ or the size of $\mathcal{S}_L^{\rightarrow}$, and it needs to discover and explore $\mathcal{S}_L^{\rightarrow}$ by itself and find policies to reach each state in $\mathcal{S}_L^{\rightarrow}$. This is why the problem is called "autonomous exploration".

We remark that in Sect. 3, we will prove that our Alg. 2 outputs a set $\mathcal{K} \supseteq \mathcal{S}_L^{\rightarrow}$ and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ restricted on $\mathcal{K}$, such that

$$\forall s \in \mathcal{K}, V_s^{\pi_s}(s_0) \le V_{\mathcal{K},s}^*(s_0) + \varepsilon L.$$

This implies $\forall s \in \mathcal{S}_L^{\rightarrow}$, $V_s^{\pi_s}(s_0) \le (1+\varepsilon)L$, because when $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}$, we have $V_{\mathcal{K},s}^*(s_0) \le V_{\mathcal{S}_L^{\rightarrow},s}^*(s_0)$, and for any $s \in \mathcal{S}_L^{\rightarrow}$, we have $V_{\mathcal{S}_L^{\rightarrow},s}^*(s_0) \le L$.

In the special case that $\mathcal{S}_L^{\rightarrow} = \mathcal{S}$ (i.e., in the setting of multi-goal SSP), our Alg. 2 will output $\mathcal{K} = \mathcal{S}$, and the inequality above will be reduced to $\forall s \in \mathcal{S}, V_s^{\pi_s}(s_0) \le V_s^*(s_0) + \varepsilon L$. Hence our Alg. 2 for AX problem also solves multi-goal SSP problem with goal space $\mathcal{G} = \mathcal{S}$.

2.1. Review of Prior Algorithms

We review prior algorithms because our algorithm also relies on some components from prior algorithms.

DisCo Algorithm for Autonomous Exploration.

DisCo algorithm was introduced in (Tarbouriech et al., 2020), and we use DisCo algorithm as a burn-in step for Alg. 2. Here we give the lemma of the sample complexity of DisCo algorithm for autonomous exploration.

Lemma 2.4 (Corollary 1, (Tarbouriech et al., 2020)). *Assume that $L \geq 1$, $0 < \varepsilon \leq 1$ and $0 < \delta < 1$. For any MDP $M = \langle \mathcal{S}, \mathcal{A}, P, c, s_0 \rangle$ satisfying Asmp. 2.1, with probability at least $1-\delta$, DisCo algorithm will terminate and output a set of states $\mathcal{K}$ such that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{(1+\varepsilon)L}^{\rightarrow}$, and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ restricted on $\mathcal{K}$, such that $\forall s \in \mathcal{K}, V_s^{\pi_s}(s_0) \leq V_{\mathcal{K},s}^*(s_0) + \varepsilon L$, and the cumulative cost $C_T = \widetilde{O}(L^3 S_{(1+\varepsilon)L}^2 A c_{\min}^{-2} \varepsilon^{-2})$.*

Here we clarify that the definitions of ε in our work and in (Tarbouriech et al., 2020) are different. Tarbouriech et al. (2020) denotes absolute error as ε (i.e., they require that the output policies satisfy $V_s^{\pi_s}(s_0) \leq L+\varepsilon$ and $V_s^{\pi_s}(s_0) \leq V_{\mathcal{K},s}^*(s_0)+\varepsilon$), and our paper denotes relative error as ε (i.e., we require $V_s^{\pi_s}(s_0) \leq (1+\varepsilon)L$ and $V_s^{\pi_s}(s_0) \leq V_{\mathcal{K},s}^*(s_0) + \varepsilon L$). And their absolute error ε_{abs} and our relative error ε_{rel} satisfies the following equation: $\varepsilon_{\text{abs}} = \varepsilon_{\text{rel}} L$.

We also remark that when $c_{\min} = 1$, the original form of sample complexity in Theorem 1, (Tarbouriech et al., 2020) was $\widetilde{O}(L^5 \Gamma_{L+\varepsilon_{\text{abs}}} S_{L+\varepsilon_{\text{abs}}} A \varepsilon_{\text{abs}}^{-2} + L^3 S_{L+\varepsilon_{\text{abs}}}^2 A \varepsilon_{\text{abs}}^{-1})$, where $\Gamma_L := \max_{(s,a)\in\mathcal{S}_L^{\rightarrow}\times\mathcal{A}} \left\| \{P(s' \mid s,a)\}_{s'\in\mathcal{S}_L^{\rightarrow}} \right\|_0$, and $\Gamma_{L+\varepsilon_{\text{abs}}} = S_{L+\varepsilon_{\text{abs}}}$ in the worst case. By setting $\varepsilon_{\text{abs}} = \varepsilon_{\text{rel}} L$ and $\Gamma_{L+\varepsilon_{\text{abs}}} = S_{L+\varepsilon_{\text{abs}}}$, we can obtain the sample complexity bound $\widetilde{O}(L^3 S_{(1+\varepsilon_{\text{rel}})L}^2 A \varepsilon_{\text{rel}}^{-2})$ in Lem. 2.4 when $c_{\min} = 1$. And in Corollary 1, (Tarbouriech et al., 2020), they discussed the case when $c_{\min} \in (0,1)$, which incurs an additional $c_{\min}^{-2}$ in their sample complexity.

3. Algorithms and Sample Complexity Bounds

Now we are ready to describe our main algorithm VALAE (cf. Alg. 2), and currently we focus on autonomous exploration problem. There are three key components in Alg. 2. The first component is running DisCo algorithm (cf. (Tarbouriech et al., 2020)) with $\varepsilon = 1$. Our aim is to discover a set of states $\mathcal{K}$ such that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$, and compute a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ to reach each state $s \in \mathcal{K}$ with expected cost $V_s^{\pi_s}(s_0)$ no more than $2L$. After the first component, we will fix our set $\mathcal{K}$, and to solve the AX problem, we need only learn a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ such that $V_s^{\pi_s}(s_0) \leq V_{\mathcal{K},s}^*(s_0) + \varepsilon L$ for all $s \in \mathcal{K}$.

The second component reduces the autonomous exploration problem to multi-goal SSP (cf. Alg. 1) using the set $\mathcal{K}$ computed from the first component. Alg. 1 first constructs a new MDP $M^\dagger$ by "merging" all the states $s \notin \mathcal{K}$ to a single artificial state x, and to solve AX problem, we need only solve multi-goal SSP problem on MDP $M^\dagger$ with goal space $\mathcal{G} = \mathcal{K}$. Then the algorithm collects fresh samples of the form (s,a,s',c) for all state-action pairs $(s,a) \in \mathcal{K} \times \mathcal{A}$, and the aim is to compute the empirical probability $\widehat{P}(s'|s,a)$ and the average cost $\widehat{c}(s,a)$ with small error.

Algorithm 1 **Re**duce Autonomous Exploration to **M**ulti-Goal **SSP** (Re-MG-SSP)

Input: Confidence $\delta \in (0,1)$, exploration radius $L \geq 1$,
Input: a set of states $\mathcal{K}$, and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$.
Define MDP $M^\dagger = \langle \mathcal{K}^\dagger, \mathcal{A}, P^\dagger, c^\dagger, s_0 \rangle$ where $\mathcal{K}^\dagger$, $P^\dagger$, $c^\dagger$ are defined in Sect. 3.2.
$\forall (s,a,s') \in \mathcal{K}^\dagger \times \mathcal{A} \times \mathcal{K}^\dagger$, set $N(s,a,s') \leftarrow 0$; $\widehat{P}_{s,a,s'} \leftarrow 0$.
$\forall (s,a) \in \mathcal{K}^\dagger \times \mathcal{A}$, set $N(s,a) \leftarrow 0$; $n(s,a) \leftarrow 0$; $\theta(s,a) \leftarrow 0$; $\widehat{c}(s,a) \leftarrow 0$.
Set $\psi \leftarrow 12000 L^2 |\mathcal{K}| c_{\min}^{-2} \ln(\frac{|\mathcal{K}|A}{\delta})$, and $\phi \leftarrow 2^{\lceil \log_2 \psi \rceil}$.
for each $(s,a) \in \mathcal{K} \times \mathcal{A}$ **do**
  while $N(s,a) < \phi$ **do**
    Execute policy π_s on MDP $M^\dagger$ until reaching state s.
    Take action a, incur cost c and observe next state $s' \sim P^\dagger(\cdot \mid s,a)$.
    Set $N(s,a) \leftarrow N(s,a)+1$, $\theta(s,a) \leftarrow \theta(s,a)+c$, $N(s,a,s') \leftarrow N(s,a,s')+1$.
  end while
  Set $\widehat{c}(s,a) \leftarrow \frac{\theta(s,a)}{N(s,a)}$ and $\theta(s,a) \leftarrow 0$.
  For all $s' \in \mathcal{K}^\dagger$, set $n(s,a) \leftarrow N(s,a)$, $\widehat{P}_{s,a,s'} \leftarrow N(s,a,s')/N(s,a)$.
end for
For all $a \in \mathcal{A}$, set $N(x,a) \leftarrow \phi, n(x,a) \leftarrow \phi, \widehat{c}(x,a) \leftarrow c_{\text{RESET}}$, $\widehat{P}_{x,a,s_0} \leftarrow 1$.
For all $a \in \mathcal{A}$, $s' \in \mathcal{S}$, set $\widehat{P}_{x,a,s'} \leftarrow 0$.
Output: $N(), n(), \widehat{P}, \theta(), \widehat{c}$.

In the third component, inspired by recent advances in stochastic shortest path (Tarbouriech et al., 2021), we design a policy evaluation step to obtain near-optimal estimates of the costs of getting to each $s \in \mathcal{S}_L^{\rightarrow}$ (cf. Alg. 2).

Below we give detailed descriptions for each component.

3.1. Running DisCo Algorithm with $\varepsilon = 1$

In the first component of our main algorithm VALAE (cf. Alg. 2), we use DisCo algorithm with (relative) error $\varepsilon = 1$ as a subroutine. By Lem. 2.4, we can obtain a set $\mathcal{K}$ such that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$, and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ such that $\forall s \in \mathcal{K}, V_s^{\pi_s}(s_0) \leq 2L$, and the total cost is bounded by $\widetilde{O}(L^3 S_{(1+\varepsilon)L}^2 A c_{\min}^{-2})$. In the next subsection, we will focus on a fixed set $\mathcal{K}$, and reduce the autonomous exploration problem to multi-goal SSP problem.

3.2. Connection between Autonomous Exploration and Multi-Goal SSP

In our main algorithm VALAE (Alg. 2), after running DisCo with $\varepsilon = 1$, we have obtained a set of known states $\mathcal{K} \supseteq$

$\mathcal{S}_L^{\rightarrow}$ and discovered all the states that we want to explore, and we denote $K = |\mathcal{K}|$. Now we focus on the second component of VALAE (cf. Alg. 1). We will fix our set of known states $\mathcal{K}$, and focus only on the policies restricted on $\mathcal{K}$. Therefore, for all the states $s \notin \mathcal{K}$, we can regard them as one artificial state x, and the only action at state x is RESET. To this purpose, we will construct an MDP $M^\dagger := \langle \mathcal{K}^\dagger, \mathcal{A}, P^\dagger, c^\dagger, s_0 \rangle$ where we first define the artificial state x, and we set $\mathcal{K}^\dagger = \mathcal{K} \cup \{x\}$, and we denote $K' = |\mathcal{K}^\dagger| = K+1$. For any $(s,a) \in \mathcal{K} \times \mathcal{A}$, we define $P^\dagger_{s,a,s'}$ as follows:

$$P^\dagger_{s,a,s'} = P_{s,a,s'},\ \forall s' \in \mathcal{K},\ \text{and}\ P^\dagger_{s,a,x} = \sum_{s' \notin \mathcal{K}} P_{s,a,s'}.$$

We also define $P^\dagger_{x,a,s'} = \mathbb{I}[s' = s_0]$ for any $a \in \mathcal{A}, s' \in \mathcal{K}^\dagger$. Finally, we define $c^\dagger(s,a) = c(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $c^\dagger(x,a) = c_{\text{RESET}}$ for all $a \in \mathcal{A}$. In this way, the AX problem reduces to multi-goal SSP problem on MDP $M^\dagger$ with the set of states being $\mathcal{K}^\dagger$ and goal space $\mathcal{G} = \mathcal{K}$, and all states in $\mathcal{K}$ are incrementally $2L$-controllable from s_0.

Next, we collect $\phi = \widetilde{\Omega}(L^2|\mathcal{K}|/c_{\min}^2)$ fresh samples for each state-action pair $(s,a) \in \mathcal{K} \times \mathcal{A}$. Our aim is that for each state-action pair $(s,a) \in \mathcal{K} \times \mathcal{A}$, we can obtain ϕ samples of the form (s,a,s',c) and compute the empirical probability $\widehat{P}(s'|s,a)$ and the average cost $\widehat{c}(s,a)$, so that our estimation $\widehat{P}(s'|s,a)$ and $\widehat{c}(s,a)$ are close enough to $P^\dagger(s'|s,a)$ and $c^\dagger(s,a)$, respectively. In DisCo algorithm, we have computed a policy π_s for each $s \in \mathcal{K}$, such that we can execute π_s to reach state s from s_0 with expected cost no more than $2L$. Hence, to obtain a sample (s,a,s',c) at any state-action pair $(s,a) \in \mathcal{K} \times \mathcal{A}$, we need only first execute π_s to arrive at state s, then execute action a.

We remark that using fresh samples is essential for Alg. 2 to ensure these samples are independent of $\mathcal{K}$, and we cannot use the samples collected in DisCo algorithm because they are dependent of $\mathcal{K}$. Also, we note that in Alg. 1 and Alg. 2, the estimated transition probability $\widehat{P}(s' \mid s,a)$ and the estimated cost $\widehat{c}(s' \mid s,a)$ are only evaluated for all $(s,a) \in \mathcal{K}^\dagger \times \mathcal{A}$ on MDP $M^\dagger$, rather than for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ on MDP M, hence the computational complexity of Alg. 1 and Alg. 2 does not depend on "global" $|\mathcal{S}|$.

We note that the idea of uniformly connecting ϕ samples for each state-action pair $(s,a) \in \mathcal{K} \times \mathcal{A}$ is similar with DisCo algorithm. The difference is that DisCo algorithm collects $\widetilde{\Omega}(L^2|\mathcal{K}|c_{\min}^{-2}\varepsilon^{-2})$ samples for each state-action pair $(s,a) \in \mathcal{K} \times \mathcal{A}$, but in Alg. 1 our $\phi = \widetilde{\Omega}(L^2|\mathcal{K}|c_{\min}^{-2})$ and is smaller than that in DisCo.

Algorithm 2 **Val**ue-Aware **A**utonomous **E**xploration (VALAE)

Input: Confidence $\delta \in (0,1)$, error $\varepsilon \in (0,1]$, and $L \geq 1$.
Input (for multi-goal SSP only): Goal Space $\mathcal{G} \subseteq \mathcal{S}$.
(For autonomous exploration, set $\mathcal{G} = \emptyset$.)
Specify: Trigger set $\mathcal{N} \leftarrow \{2^{j-1} : j = 1, 2, \ldots\}$.
\\We run DisCo algorithm with $\varepsilon = 1$ and get a set $\mathcal{K}$ such that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$.
Run DisCo algorithm with input $(\delta, \varepsilon = 1, L)$ and we get a set $\mathcal{K}$ and a set of policies $\{\pi_s\}_{s \in \mathcal{K}}$.
Run Alg. 1 with input $(\delta, L, \mathcal{K}, \{\pi_s\}_{s \in \mathcal{K}})$, and we obtain the variables $N(), n(), \widehat{P}, \theta(), \widehat{c}$.
Set time step $t \leftarrow 1$ and trigger index $j \leftarrow 5 + \log_2 \frac{1}{c_{\min}}$.
Set $\epsilon \leftarrow \varepsilon/3$, $B \leftarrow 10L$, $\lambda = \widetilde{O}(1/\epsilon^2)$, and $g \leftarrow s_0$.
Initialize $\mathcal{G} \leftarrow \mathcal{K}$ if $\mathcal{G} = \emptyset$.
\\Solve multi-goal SSP problem on $M^\dagger$ with goal space $\mathcal{G}$.
for round $r = 1, 2, \cdots$ **do**
  \\Phase (a): Compute Optimal Policy
  Compute $(Q, V) := \texttt{VISGO}(g, 2^{-j}/(|\mathcal{K}^\dagger|A))$.
  Set the policy $\tilde{\pi}$ as the greedy policy over Q, and $\hat{\tau} \leftarrow 0$.
  \\Phase (b): Policy Evaluation
  for episode $k = 1, 2, \cdots, \lambda$ **do**
    Set $s_t \leftarrow s_0$ and reset to the initial state s_0, and $\hat{\tau}_k \rightarrow 0$.
    while $s_t \neq g$ **do**
      Take action $a_t = \arg\min_{a \in \mathcal{A}} Q(s_t, a)$ on $M^\dagger$, incur cost c_t and observe next state $s_{t+1} \sim P^\dagger(\cdot \mid s_t, a_t)$.
      Set $(s, a, s', c) \leftarrow (s_t, a_t, s_{t+1}, c_t)$ and $t \leftarrow t + 1$.
      Set $N(s,a) \leftarrow N(s,a) + 1$, $\theta(s,a) \leftarrow \theta(s,a) + c$, $N(s,a,s') \leftarrow N(s,a,s') + 1$.
      if $N(s,a) \in \mathcal{N}$ **then**
        Set $j \leftarrow j + 1$, $\widehat{c}(s,a) \leftarrow \frac{2\theta(s,a)}{N(s,a)}$ and $\theta(s,a) \leftarrow 0$.
        For all $s' \in \mathcal{K}^\dagger$, set $n(s,a) \leftarrow N(s,a)$, $\widehat{P}_{s,a,s'} \leftarrow N(s,a,s')/N(s,a)$.
        Return to line 11, start a new round (the current round has been a *skipped round*).
      end if
      Set $\hat{\tau} \leftarrow \hat{\tau} + \frac{c}{\lambda}$, $\hat{\tau}_k \leftarrow \hat{\tau}_k + c$.
    end while
    if $\hat{\tau} > V(s_0) + \epsilon L$ **then**
      Return to line 11, start a new round. (the current round has been a *failure round*).
    end if
  end for
  Set $\pi_g \leftarrow \tilde{\pi}$. Remove g from $\mathcal{G}$. (The current round has been a *success round*.)
  Choose another state $g \in \mathcal{G}$.
  Stop the algorithm if $\mathcal{G}$ is empty.
end for
Output: The states s in $\mathcal{K}$ and their corresponding policy π_s.

3.3. Value-Aware Algorithms for Autonomous Exploration and Multi-Goal SSP

Finally we describe our main algorithm, Value-Aware Autonomous Exploration (VALAE, cf. Alg. 2). First, VALAE uses DisCo algorithm with $\varepsilon = 1$ as a subroutine, and DisCo algorithm computes a set $\mathcal{K}$ such that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}$. We discard all the samples collected in DisCo algorithm, in order to ensure the independence of $\mathcal{K}$ and $\widehat{P}_{s,a}$. Second, we use Alg. 1 as a burn-in step to collect $\widetilde{\Omega}(L^2|\mathcal{K}|/c_{\min}^2)$ samples for each of the state-action pair

(s,a) so that the empirical model $\widehat{P}$ and the true model $P^\dagger$ are close enough. This guarantees that with high probability, in any round r, the expected cost of the greedy policy $\tilde{\pi}$ in Phase (a) on model $P^\dagger$ is no more than $O(L)$, which is proved in Lem. C.5.

From now on we work on the MDP $M^\dagger$, and we will solve the multi-goal SSP problem on $M^\dagger$ and compute near-optimal policies π_g for all the goal states $g \in \mathcal{K}$. We choose the goal state $g \in \mathcal{K}$ one by one, and we move to another goal state g if the average performance of the policy π_g is close to our estimation of the optimal policy. In each round, we have two phases. In the first phase, we use `VISGO` (cf. Alg. 3 in Appendix C) to estimate the value function of the optimal policy with goal state g (denoted as V), and we set the policy $\tilde{\pi}$ as the greedy policy over its output Q. We note that V is optimistic, i.e., $V(s) \le V^*_{\mathcal{K},g}(s) \le 2L+1$. Since we do not know whether the policy $\tilde{\pi}$ is close enough to the optimal policy, in the second phase, we will execute $\tilde{\pi}$ for $\lambda = \widetilde{O}(1/\epsilon^2)$ times and check whether the average performance is close enough to our estimation of the optimal cost (i.e., check whether $\hat{\tau} \le V(s_0) + \epsilon L$). By setting

$$\lambda = \lceil \frac{2048}{\epsilon^2} \ln^2(\frac{256}{\epsilon}) \ln(\frac{2|\mathcal{K}|}{\delta}) \rceil$$

and using concentration inequalities (Lem. D.1), we can prove that the average performance $\hat{\tau}$ in λ episodes is close enough to the expected cost of $\tilde{\pi}$. In this process, we also collect samples, and use them to help us estimate the value function of the optimal policy.

In the second phase, the current round will be classified into three cases: failure round, skipped round, and success round. This borrows the idea from (Lim & Auer, 2012). If the average performance of the policy $\tilde{\pi}$ is too bad (i.e., $\hat{\tau}$ is larger than $V(s_0) + \epsilon L$), we will consider the current round as a failure round. If the number of samples $N(s,a)$ meets the trigger set (i.e. is a power of 2), we will consider the current round as a skipped round, following the idea in (Jaksch et al., 2010). Otherwise, the current round is a success round. In the case of a failure round or a skipped round, we will not change the goal state g, and in the next round, we compute a new policy by `VISGO` using the samples collected in this round. In the case of a success round, as the average performance of the policy $\tilde{\pi}$ is close to optimal, we can set the $\tilde{\pi}$ as the policy π_g for the goal state g, and choose another goal state g.

Theorem 3.1 (Cumulative Cost for AX)**.** *Assume that $L \ge 1$, $0 < \varepsilon \le 1$ and $0 < \delta < 1$. For any MDP $M = \langle \mathcal{S}, \mathcal{A}, P, c, s_0 \rangle$ satisfying Asmp. 2.1, with probability at least $1-\delta$, our Alg. 2 will terminate and output a set of states $\mathcal{K}$ such that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$, and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ restricted on $\mathcal{K}$, such that $\forall s \in \mathcal{K}, V_s^{\pi_s}(s_0) \le V^*_{\mathcal{K},s}(s_0) + \varepsilon L$, and the cumulative cost $C_T = \widetilde{O}(LS_{2L}A\varepsilon^{-2} + LS_{2L}^2 A\varepsilon^{-1} + L^3 S_{2L}^2 A c_{\min}^{-2})$. And when $\varepsilon \le \min(S_{2L}^{-1}, L^{-1}c_{\min})$, we have $C_T = \widetilde{O}(LS_{2L}A\varepsilon^{-2})$.*

Thm.3.1 shows that Alg.2 solves autonomous exploration problem. Note that in Thm. 3.1, the dependency on L is tight when $\varepsilon \to 0$, because we leverage the variance information in the policy-evaluation phase, which is necessary in RL problems generally. DisCo algorithm does not use the variance information because it collects equal number of samples on each state-action pair (s,a), i.e., the sample collection in DisCo algorithm does not use the estimated value function as the guidance.

We highlight that the leading term of C_T does not have $c_{\min}$. This is because the variance fundamentally does not scale with $c_{\min}$ (cf. Lem. D.2 and Lem. D.3). While we discover a larger set $\mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$ compared with (Lim & Auer, 2012) and (Tarbouriech et al., 2020), we note that if the number of the L-controllable states grows polynomially with respect to L, S_L and S_{2L} will be of the same order. Hence under this assumption, our sample complexity bound strictly improves the existing ones and is nearly minimax optimal.

Lastly, we note that Alg. 2 also solves the multi-goal SSP problem, and it enjoys a near-optimal sample complexity for multi-goal SSP:

Theorem 3.2 (Cumulative Cost for Multi-Goal SSP)**.** *Assume that $L \ge 1$, $0 < \varepsilon \le 1$, $0 < \delta < 1$ and goal space $\mathcal{G} \subseteq \mathcal{S}$. For any MDP $M = \langle \mathcal{S}, \mathcal{A}, P, c, s_0 \rangle$ satisfying Asmp. 2.1 and $\mathcal{S}_L^{\rightarrow} = \mathcal{S}$, with probability at least $1-\delta$, our Alg. 2 will terminate and output a set of policies $\{\pi_s\}_{s\in\mathcal{G}}$ such that $\forall s \in \mathcal{G}, V_s^{\pi_s}(s_0) \le V_s^*(s_0) + \varepsilon L$, and the cumulative cost $C_T = \widetilde{O}(LSA\varepsilon^{-2} + LS^2A\varepsilon^{-1} + L^3S^2Ac_{\min}^{-2})$. And when $\varepsilon \le \min(S^{-1}, L^{-1}c_{\min})$, we have $C_T = \widetilde{O}(LSA\varepsilon^{-2})$.*

4. A Minimax Lower Bound for Autonomous Exploration

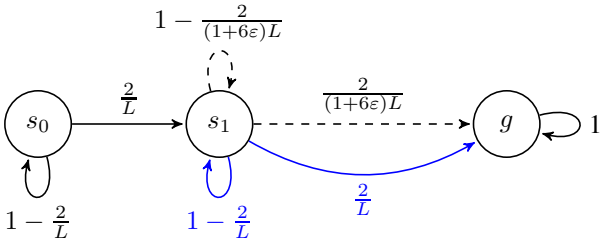

Figure 1: Illustration of our construction of the hard MDP. Here we present our lower bound of sample complexity for the autonomous exploration problem, and we follow the definitions in (Domingues et al., 2021).

We define a learning algorithm as a history-dependent policy $\boldsymbol{\pi}$ used to interact with an MDP M, and the rigorous definition of $\boldsymbol{\pi}$ is in Appendix E. We recall that in AX set-

ting, the algorithm eventually stops and output a set $\mathcal{K} \subseteq \mathcal{S}$ and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$. Hence we define an algorithm for the AX problem as a tuple $(\boldsymbol{\pi}, \tau, \mathcal{K}, \{\pi_s\}_{s\in\mathcal{K}})$, where τ is the stopping time (total number of steps) chosen by the algorithm, $\mathcal{K}$ and $\{\pi_s\}_{s\in\mathcal{K}}$ are the output of the algorithm. Now we formally write the definition of an algorithm for AX problem.

Definition 4.1. An algorithm $(\boldsymbol{\pi}, \tau, \mathcal{K}, \{\pi_s\}_{s\in\mathcal{K}})$ is (ε, δ, L)-PAC for AX problem on MDP M, if with probability over $1-\delta$, the algorithm returns a set of states $\mathcal{K}$ and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ after τ steps, such that $\mathcal{K} \supseteq \mathcal{S}_L^{\rightarrow}$ and $\forall s \in \mathcal{S}_L^{\rightarrow}, V_s^{\pi_s}(s_0) \le (1+\varepsilon)L$.

We note that τ is a random variable over the probability distribution $\mathbb{P}_{\boldsymbol{\pi},M}$, where $\mathbb{P}_{\boldsymbol{\pi},M}$ is determined by algorithm $\boldsymbol{\pi}$ and MDP M, and $\mathbb{P}_{\boldsymbol{\pi},M}$ is defined in Appendix E. Also, we denote the operator $\mathbb{E}_{\boldsymbol{\pi},M}$ as the expectation under $\mathbb{P}_{\boldsymbol{\pi},M}$. Then for any real numbers $L, c_{\min}$ and positive integers S, A, S_L, we define a class of MDPs $\mathfrak{M}(L, S_L)$ as follows: $\mathfrak{M}(L, S_L)$ contains all the MDPs $M = \langle \mathcal{S}, \mathcal{A}, P, c, s_0\rangle$, such that $|\mathcal{S}| \le S$, $|\mathcal{A}| \le A$, $c(s,a) \in [c_{\min}, 1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and M satisfies Asmp. 2.1 and $|\mathcal{S}_L^{\rightarrow}| \le S_L$. Finally, the following theorem states the lower bound for the autonomous exploration problem.

Theorem 4.2. *Assume that $L > 4$, $S > 8$, $A > 4$, $4 \le S_L \le \min\{(A-1)^{\lfloor \frac{L}{2} \rfloor}, S\}$, $0 < \varepsilon < \frac{1}{4}$, $0 < \delta < \frac{1}{16}$, and $0 < c_{\min} \le 1$. Then for any algorithm $(\boldsymbol{\pi}, \tau, \mathcal{K}, \{\pi_s\}_{s\in\mathcal{K}})$ that is (ε, δ, L)-PAC for AX problem on any MDP $M \in \mathfrak{M}(L, S_L)$, there exists an MDP $\mathcal{M} \in \mathfrak{M}(L, S_L)$ such that*

$$\mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}[\tau] = \Omega\Big(\frac{LS_LA}{c_{\min}\varepsilon^2}\log\frac{1}{\delta}\Big).$$

As τ is the total number of steps used in the algorithm $\boldsymbol{\pi}$, the lower bound of cumulative cost C_T is $c_{\min}$ multiplies the lower bound of τ, i.e., $\Omega(LS_LA\varepsilon^{-2}\log\frac{1}{\delta})$. This lower bound further implies our upper bound (Theorem 3.1) is nearly minimax-optimal when S_L and S_{2L} are of the same order. We also have a lower bound for multi-goal SSP (cf. Appendix F).

4.1. Proof Sketch

We briefly sketch our proof of the lower bound. We consider the case $c_{\min} = 1$ and $L > 2$ for convenience, and we first construct our family of hard MDPs for $S = 3$ states (cf. Fig. 1), where s_0 is initial state, s_1 is middle state and g is goal state. In the initial state s_0, taking any action a will transit to state s_1 with probability $\frac{2}{L}$, and stay at state s_0 with probability $1-\frac{2}{L}$. In state s_1, there is only one optimal action a^*. When we take the action a^* in s_1 (the blue edges), the agent will transit to the goal state g with probability $\frac{2}{L}$ and stay at s_1 with probability $1-\frac{2}{L}$. When we take an action $a \ne a^*$ in s_1 (the dashed edges), the agent will transit to the goal state g with smaller probability $\frac{2}{(1+6\varepsilon)L}$. We note that the RESET action is not drawn in Fig. 1.

We can verify that $V_g^*(s_0) = \frac{L}{2} + \frac{L}{2} = L$, and $g \in \mathcal{S}_L^{\rightarrow}$ (hence g should be contained in the learning algorithm's output $\mathcal{K}$). Let π_g be the output policy of the learning algorithm with respect to goal state g. If $\pi_g(s_1) = a^*$, we have $V_g^{\pi_g}(s_0) = L$. Otherwise, we have $V_g^{\pi_g}(s_0) = \frac{L}{2} + \frac{(1+6\varepsilon)L}{2} > (1+\varepsilon)L$, i.e., the policy π_g is not valid output for AX if $\pi_g(s_1) \ne a^*$. Hence, if the algorithm solves AX problem on this MDP, it has to discriminate between two Bernoulli distributions with $p_1 = \frac{2}{L}$ and $p_2 = \frac{2}{(1+6\varepsilon)L}$ among all the A actions, and the KL divergence of the two distributions is $O(\varepsilon^2/L)$. Hence we can prove that we need at least $\widetilde{\Omega}(LA/\varepsilon^2)$ to solve AX on this MDP. The technique of KL divergence is similar with (Domingues et al., 2021). Then we can generalize our hard MDP to larger S_L. We first construct an MDP $\mathcal{M}_0'$ with $S_L - 1$ states, and each middle state s_i can be reached from s_0 in $L/2$ steps in expectation. Then we add a goal states g, and we choose one optimal state-action pair (s_i^*, a^*) among all the middle states and actions. Finally, we set the transition probability $P(g|s_i^*, a^*) = \frac{2}{L}$, and $P(g|s_i, a) = \frac{2}{(1+6\varepsilon)L}$ for other pair of middle state and action (s_i, a). In intuition, this extends the construction in Fig. 1 from A actions to $O(S_LA)$ actions. The full construction is in Appendix E. Under this construction, we can prove that the lower bound scales as $\widetilde{\Omega}(LS_LA/\varepsilon^2)$.

5. Conclusion

We introduced a new algorithm for the autonomous exploration problem, which improves existing ones. Along the way, we also introduced a new problem, multi-goal SSP problem, which can be of independent interest. The natural future directions include designing an algorithm with $\widetilde{O}\big(\frac{LS_LA}{\varepsilon^2}\big)$ sample complexity instead of $\widetilde{O}\big(\frac{LS_{2L}A}{\varepsilon^2}\big)$, and improving the lower order terms in existing bounds.

Acknowledgements

The work is supported by JD.com. Simon S. Du gratefully acknowledges the funding from NSF Award's IIS-2110170 and DMS-2134106.

References

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

Baranes, A. and Oudeyer, P.-Y. R-iac: Robust intrinsically motivated exploration and active learning. *IEEE Transactions on Autonomous Mental Development*, 1(3):155–169, 2009.

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

Bertsekas, D. P. et al. *Dynamic programming and optimal control: Vol. 1*. Athena scientific Belmont, 2000.

Cohen, A., Kaplan, H., Mansour, Y., and Rosenberg, A. Near-optimal regret bounds for stochastic shortest path. *arXiv preprint arXiv:2002.09869*, 2020.

Devo, A., Costante, G., and Valigi, P. Deep reinforcement learning for instruction following visual navigation in 3d maze-like environments. *IEEE Robotics and Automation Letters*, 5(2):1175–1182, 2020.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*, 2020.

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.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Oudeyer, P.-Y. and Kaplan, F. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2): 265–286, 2007.

Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. Improved sample complexity for incremental autonomous exploration in mdps. *arXiv preprint arXiv:2012.14755*, 2020.

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. *arXiv preprint arXiv:2104.11186*, 2021.

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

A. Basic Property of the Optimal Policy

Lemma A.1 (Bertsekas & Tsitsiklis, 1991;Yu & Bertsekas, 2013)**.** *Suppose that there exists a proper policy with respect to the goal state g and that for every improper policy π' there exists at least one state $s \in \mathcal{S}$ such that $V_g^{\pi'}(s) = +\infty$. Then the optimal policy π^* is stationary, deterministic, and proper. Moreover, $V_g^* = V_g^{\pi^*}$ is the unique solution of the optimality equations $V_g^* = \mathcal{L}V_g^*$ and $V_g^*(s) < +\infty$ for any $s \in \mathcal{S}$, where for any vector $V \in \mathbb{R}^S$ the optimal Bellman operator $\mathcal{L}$ is defined as*

$$\mathcal{L}V(s) := \min_{a \in \mathcal{A}} \Big\{ c(s,a) + P_{s,a}V \Big\}.$$

Furthermore, the optimal Q-value, denoted by $Q_g^* = Q_g^{\pi^*}$, is related to the optimal value function as follows

$$Q_g^*(s,a) = c(s,a) + P_{s,a}V_g^*, \qquad V_g^*(s) = \min_{a \in \mathcal{A}} Q_g^*(s,a), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

B. High-Probability Event

First we define the high-probability event $\mathcal{E}$ to do concentration on all the samples collected in Alg. 1 and Alg. 2. We note that in Alg. 2, after running DisCo algorithm, the set of known states $\mathcal{K}$ is fixed, and our algorithm focuses on the new MDP $\mathcal{M}^\dagger = \langle \mathcal{K}^\dagger, \mathcal{A}, P^\dagger, c^\dagger, s_0 \rangle$, where $\mathcal{M}^\dagger$ is defined in Sect. 3.2.

We recall that for any two vectors $X, Y \in \mathbb{R}^{K'}$ ($K' = |\mathcal{K}^\dagger|$), we write their inner product as $XY := \sum_{s \in \mathcal{K}^\dagger} X(s)Y(s)$, and we denote $\|X\|_\infty := \max_{s \in \mathcal{K}^\dagger} |X(s)|$. If X is a probability distribution on $\mathcal{K}^\dagger$, we denote $\mathbb{V}(X,Y) := \sum_{s \in \mathcal{K}^\dagger} X(s)Y(s)^2 - (\sum_{s \in \mathcal{K}^\dagger} X(s)Y(s))^2$, i.e. the variance of random varianble Y over distribution X. And we use $P_{s,a}^\dagger$ and $P_{s,a,s'}^\dagger$ to denote $P^\dagger(\cdot|s,a)$ and $P^\dagger(s'|s,a)$, respectively.

Here for any $g \in \mathcal{K}$, we denote the vector $V_g^* \in \mathbb{R}^{K'}$ as the value function of the optimal policy on MDP $\mathcal{M}^\dagger$ with respect to goal g, and we denote $V_g^*(s)$ as the expected cost of the optimal policy to reach state g from s on MDP $M^\dagger$. Also, we define $B_* := \max_{(s,g) \in \mathcal{K}^\dagger \times \mathcal{K}} V_g^*(s)$, and we denote $Q_g^* \in \mathbb{R}^{K' \times A}$ as the Q-function corresponding to V_g^*.

In Alg. 2, we set $B = 10L$. Thus when $\mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$, we have $\forall (s,g) \in \mathcal{K}^\dagger \times \mathcal{K}, V_g^*(s) \le 2L + 1 \le B$, and $B_* \le B$.

Then we define the high-probability event $\mathcal{E}$. We note that our definition of $\mathcal{E}$ is similar with Definition 12 in Sect. D.1, (Tarbouriech et al., 2021).

Definition B.1 (High-probability event $\mathcal{E}$)**.** We define the event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, where

$$\mathcal{E}_1 := \left\{ \forall (s,a) \in \mathcal{K}^\dagger \times \mathcal{A}, \forall n(s,a) \ge 1 \,:\, |\widehat{c}(s,a) - c^\dagger(s,a)| \le 2\sqrt{\frac{2\widehat{c}(s,a)\iota_{s,a}}{n(s,a)}} + \frac{28\iota_{s,a}}{3n(s,a)} \right\},$$

$$\mathcal{E}_2 := \left\{ \forall (s,a,s') \in \mathcal{K}^\dagger \times \mathcal{A} \times \mathcal{K}^\dagger,\ \forall n(s,a) \ge 1 \,:\, |P_{s,a,s'}^\dagger - \widehat{P}_{s,a,s'}| \le \sqrt{\frac{2P_{s,a,s'}^\dagger \iota_{s,a}}{n(s,a)}} + \frac{\iota_{s,a}}{n(s,a)} \right\},$$

$$\mathcal{E}_3 := \left\{ \forall (s,a,g) \in \mathcal{K}^\dagger \times \mathcal{A} \times \mathcal{K}, \forall n(s,a) \ge 1 \,:\, |(\widehat{P}_{s,a} - P_{s,a}^\dagger)V_g^*| \le 2\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a}, V_g^*)\iota_{s,a}}{n(s,a)}} + \frac{14B_*\iota_{s,a}}{3n(s,a)} \right\},$$

where $\iota_{s,a} := 4\ln\Big(\frac{12K'An(s,a)}{\delta}\Big)$.

Lemma B.2. *It holds that $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$.*

Proof. The proof is the same with Lemma 13 in Sect. D.1, (Tarbouriech et al., 2021).

The event $\mathcal{E}_1$ holds with probability $1 - \delta/3$ by Lem. 27 in Sect. F, (Tarbouriech et al., 2021) and by union bound over all $(s,a) \in \mathcal{K}^\dagger \times \mathcal{A}$.

The event $\mathcal{E}_2$ holds with probability $1 - \delta/3$ by Lem. 26 and Lem. 33 in Sect. F, (Tarbouriech et al., 2021) and by union bound over $(s,a,s') \in \mathcal{K}^\dagger \times \mathcal{A} \times \mathcal{K}^\dagger$.

The event $\mathcal{E}_3$ holds with probability $1 - \delta/3$ by Lem. 27 in Sect. F, (Tarbouriech et al., 2021) and by union bound over all $(s,a,g) \in \mathcal{K}^\dagger \times \mathcal{A} \times \mathcal{K}$. □

The lemma above is a direct consequence of concentration inequalities. We note that we do not use the samples collected in DisCo algorithm, and the set $\mathcal{K}$ is fixed after running DisCo algorithm. Hence for any $g \in \mathcal{K}$, the vector V_g^* depends

only on the set $\mathcal{K}$, and V_g^* is fixed after running DisCo algorithm and does not depend on the samples collected in Alg.1 and Alg.2. Thus $\widehat{P}_{s,a}$ and V_g^* are independent for any $(s,a,g) \in \mathcal{K}^\dagger \times \mathcal{A} \times \mathcal{K}$ and $n(s,a) \geq 1$.

Lemma B.3 ((Cohen et al., 2020), Lem. B.5)**.** *Let π be a proper policy such that for some $d > 0$, the expected cost $V_g^\pi(s) \leq d$ for every non-goal state $s \neq g$. Then the probability that the cumulative cost of π to reach the goal state from any state s is more than m, is at most $2e^{-m/(4d)}$ for all $m \geq 0$.*

Lemma B.4. *Let τ be a random variable on $[0,+\infty)$ such that $\Pr(\tau > m) \leq 2e^{-m/4d}$ for any $m \geq 0$, where $d > 0$ is a constant. We define the random variable $\hat{\tau} = \frac{1}{n}\sum_{k=1}^{n} \hat{\tau}_k$, where each $\hat{\tau}_k$ is i.i.d. and has the same distribution with τ. Then for any $\epsilon > 0$, we have $\Pr(E(\tau) > \hat{\tau} + \epsilon d) \leq \exp(-\frac{n\epsilon^2}{128\ln^2(64/\epsilon)})$.*

Proof. We set the constant $\Gamma = \lfloor 8d\ln(64/\epsilon) \rfloor$. Then we define the random variables $\tau_\Gamma = \min(\tau, \Gamma)$, $\check{\tau}_k = \min(\hat{\tau}_k, \Gamma)$, and $\check{\tau} = \frac{1}{n}\sum_{k=1}^{n} \check{\tau}_k$.

As each $\check{\tau}_k$ is a random variable on $[0,\Gamma]$, by Hoeffding's inequality, we have

$$\Pr(E(\tau_\Gamma) > \check{\tau} + \frac{1}{2}\epsilon d) \leq \exp(-\frac{n\epsilon^2 d^2}{2\Gamma^2}) \leq \exp(-\frac{n\epsilon^2}{128\ln^2(64/\epsilon)}).$$

Moreover, we have

$$\begin{aligned} E(\tau) &\leq E(\tau_\Gamma) + \sum_{i=1}^{\infty} i \cdot \Pr(\Gamma + i - 1 < \tau \leq \Gamma + i) = E(\tau_\Gamma) + \sum_{m=\Gamma}^{\infty} \Pr(\tau > m) \\ &\leq E(\tau_\Gamma) + 2\sum_{m=\Gamma}^{\infty} \exp(-m/4d) \leq E(\tau_\Gamma) + \frac{1}{2}\epsilon d. \end{aligned}$$

Therefore, we obtain

$$\Pr(E(\tau) > \hat{\tau} + \epsilon d) \leq \Pr(E(\tau) > \check{\tau} + \epsilon d) \leq \Pr(E(\tau_\Gamma) > \check{\tau} + \frac{1}{2}\epsilon d) \leq \exp(-\frac{n\epsilon^2}{128\ln^2(64/\epsilon)}).$$

□

C. Analysis of a `VISGO` Procedure

In this section, we fix the known states $\mathcal{K}$ and the goal state g and we analysis an execution of the `VISGO` procedure in Alg. 3. We use the value iteration of the form $V^{(i+1)} = \widetilde{\mathcal{L}}V^{(i)}$ to estimate the value funtion of the optimal policy. Here, we define the operator $\widetilde{\mathcal{L}}$ in the following way. For any $U \in \mathbb{R}^{K'}$ such that $U \geq 0$, $U(g) = 0$, and $\|U\|_\infty \leq B$, we first define

$$\widetilde{\mathcal{L}}U(s,a) := \widehat{c}(s,a) + \widetilde{P}_{s,a}U - b(U,s,a),$$

for any $s \in \mathcal{K} \setminus \{g\}$ and $a \in \mathcal{A}$, where we define

$$b(U,s,a) := \max\left\{ c_1\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a},U)\iota_{s,a}}{n(s,a)}},\ c_2\frac{B\iota_{s,a}}{n(s,a)} \right\} + c_3\sqrt{\frac{\widehat{c}(s,a)\iota_{s,a}}{n(s,a)}}, \tag{4}$$

for any $s \in \mathcal{K} \setminus \{g\}$ and $a \in \mathcal{A}$. Here we recall that $B = 10L$, $\iota_{s,a} = 4\ln\left(\frac{12K'An(s,a)}{\delta}\right)$ (cf. Def. B.1), and $\mathbb{V}(X,Y) := \sum_{s\in\mathcal{K}^\dagger} X(s)Y(s)^2 - (\sum_{s\in\mathcal{K}^\dagger} X(s)Y(s))^2$ is the variance of random varianble Y over distribution X. And we define the transition probability $\widetilde{P}_{s,a,s'} = \frac{n(s,a)}{n(s,a)+1}\widehat{P}_{s,a,s'} + \frac{\mathbb{I}[s'=g]}{n(s,a)+1}$ that slightly increases the probability to reach the goal g at each state-action pair.

Then, we set $\widetilde{\mathcal{L}}U(s) := \min_{a\in\mathcal{A}} \widetilde{\mathcal{L}}U(s,a)$ for $s \in \mathcal{K}$ and $s \neq g$, and we set $\widetilde{\mathcal{L}}U(x) := c_{\text{RESET}} + U(s_0)$. Finally, we set $\widetilde{\mathcal{L}}U(g) := 0$.

We note that in Alg. 2, before we executed `VISGO` procedure, we have collected $\phi = \widetilde{O}(L^2|\mathcal{K}|/c_{\min}^2)$ samples for each state-action pair $(s,a) \in \mathcal{K} \times \mathcal{A}$ in Alg. 1. Thus we have $n(s,a) \geq \phi$ for each $(s,a) \in \mathcal{K} \times \mathcal{A}$. We stress that the lemmas of this section are based on the conditions that $n(s,a) \geq \phi$ for all $(s,a) \in \mathcal{K} \times \mathcal{A}$.

Algorithm 3 Subroutine VISGO

Input: Goal state g and ϵ_{VI}.
Global variables: $B,\ L,\ N(),\ n(),\ \widehat{P},\ \theta(),\ \widehat{c}()$.
Specify: Constants $c_1 = 6,\ c_2 = 72,\ c_3 = 2\sqrt{2}$.
For all $(s, a, s') \in \mathcal{K} \times \mathcal{A} \times \mathcal{K}^\dagger$, set

$$\widetilde{P}_{s,a,s'} \leftarrow \frac{n(s,a)}{n(s,a)+1}\widehat{P}_{s,a,s'} + \frac{\mathbb{I}[s'=g]}{n(s,a)+1}.$$

For all $(s, a) \in \mathcal{K} \times \mathcal{A}$, set $\iota_{s,a} \leftarrow 4\ln\left(\frac{12K'An(s,a)}{\delta}\right)$.
Set $i \leftarrow 0$, $V^{(0)} \leftarrow 0$, $V^{(-1)} \leftarrow +\infty$.
while $\|V^{(i)} - V^{(i-1)}\|_\infty > \epsilon_{\text{VI}}$ **do**
  For all $s \in \mathcal{K} \setminus \{g\}$ and $a \in \mathcal{A}$, set

$$b^{(i+1)}(s,a) \leftarrow \max\left\{c_1\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a}, V^{(i)})\iota_{s,a}}{n(s,a)}},\ c_2\frac{B\iota_{s,a}}{n(s,a)}\right\} + c_3\sqrt{\frac{\widehat{c}(s,a)\iota_{s,a}}{n(s,a)}}, \tag{1}$$

$$Q^{(i+1)}(s,a) \leftarrow \widehat{c}(s,a) + \widetilde{P}_{s,a}V^{(i)} - b^{(i+1)}(s,a), \tag{2}$$

$$V^{(i+1)}(s) \leftarrow \min_{a\in\mathcal{A}} Q^{(i+1)}(s,a). \tag{3}$$

  Set $V^{(i+1)}(x) \leftarrow c_{\text{RESET}} + V^{(i)}(s_0)$.
  Set $V^{(i+1)}(g) \leftarrow 0,\ i \leftarrow i+1$.
end while
return $Q^{(i)},\ V^{(i)}$.

We note that the variance $\mathbb{V}(\widehat{P}_{s,a}, U) \le O(L^2)$ when $\|U\|_\infty \le B = 10L$, and $\iota_{s,a}$ contains only logarithmic terms, thus $b(U,s,a) = \widetilde{O}(L/\sqrt{n(s,a)})$. As $n(s,a) \ge \phi = \Omega(L^2Kc_{\min}^{-2})$, we have $b(U,s,a) \le c_{\min}/18 \le c_{\min}$ for any $(s,a) \in \mathcal{K}\times\mathcal{A}$. Therefore, if $U(g) = 0$, $\|U\|_\infty \le B$, and $U \ge 0$ component-wise, when $n(s,a) \ge \phi$ for any $(s,a) \in \mathcal{K} \times \mathcal{A}$, we have $\widetilde{\mathcal{L}}U(s,a) \ge 0$ for any $(s,a) \in \mathcal{K} \times \mathcal{A}$. Hence the output of VISGO (Q, V) satisfies $Q \ge 0$ and $V \ge 0$ component-wise. For convenience, we define $b(U,x,a) := 0$ and $b(U,g,a) := 0$ for any $a \in \mathcal{A}$.

Lemma C.1 ((Tarbouriech et al., 2021), Lemma 12)**.** *For any non-negative vector $U \in \mathbb{R}^{K'}$ such that $U(g) = 0$, for any $(s,a) \in \mathcal{K} \times \mathcal{A}$, it holds that*

$$\widetilde{P}_{s,a}U \le \widehat{P}_{s,a}U \le \widetilde{P}_{s,a}U + \frac{\|U\|_\infty}{n(s,a)+1}.$$

The proof of the following Lem. C.2 is similar with Lem. 16 in (Tarbouriech et al., 2021), but here we have two distributions $\tilde{p}$ and p. We give the whole proof for completeness.

Lemma C.2. *Let $\Upsilon := \{v \in \mathbb{R}^{K'} : v \ge 0,\ v(g) = 0,\ \|v\|_\infty \le B\}$. Let $f : \Delta^{K'} \times \Delta^{K'} \times \Upsilon \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $f(\tilde{p}, p, v, n, B, \iota) := \tilde{p}v - \max\left\{c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}},\ c_2\frac{B\iota}{n}\right\}$, with constants $c_1 = 6$ and $c_2 \ge 2c_1^2$. Then f satisfies, for all $v \in \Upsilon$, $n, \iota > 0$, $\tilde{p}, p \in \Delta^{K'}$ s.t. $\tilde{p}(s) - \frac{1}{2}p(s) \ge 0$ for all $s \ne g$,*

1. *$f(\tilde{p}, p, v, n, B, \iota)$ is non-decreasing in $v(s)$, i.e.*

$$\forall (v, v') \in \Upsilon^2,\ v \le v' \implies f(\tilde{p}, p, v, n, B, \iota) \le f(\tilde{p}, p, v', n, B, \iota);$$

2. *$f(\tilde{p}, p, v, n, B, \iota) \le \tilde{p}v - \frac{c_1}{2}\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} - \frac{c_2}{2}\frac{B\iota}{n} \le \tilde{p}v - 2\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} - 14\frac{B\iota}{n}$;*

3. *If $\tilde{p}(g) > 0$, then $f(\tilde{p}, p, v, n, B, \iota)$ is $\rho_{\tilde{p}}$-contractive in $v(s)$, with $\rho_{\tilde{p}} := 1 - p(g) < 1$, i.e.*

$$\forall (v, v') \in \Upsilon^2,\ |f(\tilde{p}, p, v, n, B, \iota) - f(\tilde{p}, p, v', n, B, \iota)| \le \rho_{\tilde{p}}\|v - v'\|_\infty.$$

Proof. We use the idea in (Tarbouriech et al., 2021), Lemma 14 to finish the proof.

The second claim holds by $\max\{x, y\} \ge (x+y)/2, \forall x, y$, by the choices of c_1, c_2 and because both $\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}}$ and $\frac{B\iota}{n}$ are non-negative. To verify the first and third claims, we fix all other variables but $v(s)$ and view f as a function in $v(s)$.

Because the derivative of f in $v(s)$ does not exist only when $c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} = c_2\frac{B\iota}{n}$, where the condition has at most two solutions, it suffices to prove $\frac{\partial f}{\partial v(s)} \geq 0$ when $c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} \neq c_2\frac{B\iota}{n}$. Direct computation gives that for any $s \in \mathcal{K}^\dagger$ and $s \neq g$,

$$\begin{aligned}
\frac{\partial f}{\partial v(s)} &= \tilde{p}(s) - c_1\mathbb{I}\left[c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} \geq c_2\frac{B\iota}{n}\right]\frac{p(s)(v(s)-pv)\iota}{\sqrt{n\mathbb{V}(p,v)\iota}} \\
&\geq \min\left\{\tilde{p}(s),\ \tilde{p}(s) - \frac{c_1^2}{c_2 B}p(s)\big(v(s)-pv\big)\right\} \\
&\overset{\text{(i)}}{\geq} \min\left\{\tilde{p}(s),\ \tilde{p}(s) - \frac{c_1^2}{c_2}p(s)\right\} \\
&\geq 0.
\end{aligned}$$

Here (i) is by $v(s) - pv \leq v(s) \leq B$. In addition, we have

$$\begin{aligned}
\sum_{s\neq g}\left|\frac{\partial f}{\partial v(s)}\right| &= \sum_{s\neq g}\left[\tilde{p}(s) - c_1\mathbb{I}\left[c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} \geq c_2\frac{B\iota}{n}\right]\frac{p(s)(v(s)-pv)\iota}{\sqrt{n\mathbb{V}(p,v)\iota}}\right] \\
&= 1 - \tilde{p}(g) - c_1\mathbb{I}\left[c_1\sqrt{\frac{\mathbb{V}(p,v)\iota}{n}} \geq c_2\frac{B\iota}{n}\right]\sqrt{\frac{\iota}{n\mathbb{V}(p,v)}}[pv - (1-p(g))\cdot pv] \\
&\leq 1 - \tilde{p}(g).
\end{aligned}$$

Therefore, we obtain that f is $\rho_{\tilde{p}}$-contractive. □

We note that by definition of $\widetilde{P}_{s,a}$, we have $\widetilde{P}_{s,a,s'} - \frac{1}{2}\widehat{P}_{s,a,s'} \geq 0$ for all $(s,a,s') \in \mathcal{K}\times\mathcal{A}\times\mathcal{K}^\dagger$.
The following two lemmas follow the same proof with Lem.18, Lem.19 in (Tarbouriech et al., 2021), respectively.

Lemma C.3. *The sequence $(V^{(i)})_{i\geq 0}$ is non-decreasing.*

Lemma C.4. *$\widetilde{\mathcal{L}}$ is a ρ-contractive operator with modulus $\rho := 1-\nu < 1$, where $\nu = \min_{(s,a)\in\mathcal{K}\times\mathcal{A}}\widetilde{P}_{s,a,g}$, i.e. for any two vectors $U_1, U_2 \in \Upsilon$ (where Υ is defined in Lem. C.2), $\|\widetilde{\mathcal{L}}U_1 - \widetilde{\mathcal{L}}U_2\|_\infty \leq \rho\|U_1 - U_2\|_\infty$. Hence, the `VISGO` procedure will terminate after at most $\lceil\log(1/\epsilon_{\text{VI}})/\log(1/\rho)\rceil$ iterations.*

C.1. The Bounded Error Property of `VISGO`

Now we focus on Alg. 2. We give the following lemma of the bounded error property (Lem. C.5), which indicates that the value function of the policy π_s is close to our estimation. The proof of Lem. C.5 uses the techniques of Lem. 2 in (Tarbouriech et al., 2020). Our Lem. C.5 focuses on a more general operator $\widetilde{\mathcal{L}}$. In our $\widetilde{\mathcal{L}}$, we involve the bonus function $b(U,s,a)$, which is not contained in (Tarbouriech et al., 2020). And we note that in our proof of the following Lem. C.5, we use the condition that $n(s,a) \geq \phi = \widetilde{\Omega}(L^2 K c_{\min}^{-2})$ for each $(s,a) \in \mathcal{K}\times\mathcal{A}$. Also, we have $\epsilon_{\text{VI}} \leq c_{\min}/18$ because we set the initial trigger index $j = 5 + \log_2 c_{\min}^{-1}$ and $\epsilon_{\text{VI}} = 2^{-j}$.
We note that by optimism property (Lem. C.6), when $\mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$, we have $V(s) \leq 2L+1$ for all $s \in \mathcal{K}^\dagger$. Hence the following bounded error property (Lem. C.5) implies that in any round, the expected cost of the greedy policy $\tilde{\pi}$ on model $P^\dagger$ is no more than $2(2L+1) = O(L)$.

Lemma C.5 (Bounded Error Property)**.** *In Alg. 2, under the event $\mathcal{E}$, for any output (Q,V) of the `VISGO` procedure in any round, let π be the greedy policy with respect to Q. Then π is proper on the model $P^\dagger_{s,a,s'}$, and for all $s \in \mathcal{K}^\dagger$, we have $V_g^\pi(s) \leq 2V(s)$, where g is the goal state in that round.*

Proof. We define $\widetilde{V}_g^\pi(s)$ as the value function of π with goal state g on the model $\widetilde{P}_{s,a,s'}$. We will first prove that $\widetilde{V}_g^\pi(s) \leq \frac{4}{3}V(s)$, and then prove that $V_g^\pi(s) \leq \frac{4}{3}\widetilde{V}_g^\pi(s)$ using the simulation lemma on the two models $\widetilde{P}_{s,a,s'}$ and $P^\dagger_{s,a,s'}$. Combining them together yields $V_g^\pi(s) \leq 2V(s)$.
First we focus on model $\widetilde{P}_{s,a,s'}$. We recall that for any $s \in \mathcal{K}$ and $s \neq g$,

$$\widetilde{\mathcal{L}}u(s) := \min_{a\in\mathcal{A}}\left\{\widehat{c}(s,a) - b(u,s,a) + \widetilde{P}_{s,a}u\right\}$$

where $b(u,s,a)$ is defined in Eq. 4, i.e., for any $s \in \mathcal{K} \setminus \{g\}$ and $a \in \mathcal{A}$,

$$b(u,s,a) = \max\left\{ c_1\sqrt{\frac{\mathbb{V}\left(\widehat{P}_{s,a},u\right)\iota_{s,a}}{n(s,a)}}, c_2\frac{B\iota_{s,a}}{n(s,a)} \right\} + c_3\sqrt{\frac{\iota_{s,a}}{n(s,a)}},$$

and we define $b(u,x,a)=0$ and $b(u,g,a)=0$.

We observe that when $\|u\|_\infty \le B = 10L$, the variance $\mathbb{V}\left(\widehat{P}_{s,a},u\right) \le B^2$, and $\iota_{s,a}$ contains only logarithmic terms. Thus we have $b(u,s,a) = \widetilde{O}(L/\sqrt{n(s,a)})$.

As we set

$$\phi = \Theta(\frac{L^2|\mathcal{K}|}{c_{\min}^2}\ln(\frac{|\mathcal{K}|A}{\delta})),$$

and $n(s,a) \ge \phi$, we can obtain $b(u,s,a) \le c_{\min}/18$ when $\|u\|_\infty \le B$.

In addition, under the event $\mathcal{E}_1$, we have $|\widehat{c}(s,a)-c(s,a)| \le \widetilde{O}(1/\sqrt{n(s,a)})$.

Thus when $n(s,a) \ge \phi$, we have $|\widehat{c}(s,a)-c(s,a)| \le c_{\min}/18$ for all $(s,a) \in \mathcal{K}^\dagger \times \mathcal{A}$.

We denote l as the final iteration index of `VISGO`, and $V = V^{(l)}$. In `VISGO`, we have $V^{(i)} = \widetilde{\mathcal{L}}V^{(i-1)}$ for all $i = 1,2,\cdots,l$. As $V^{(l-1)} \le V^{(l)}$ component-wise, we have for any $s \in \mathcal{K}^\dagger$, $V(s) \le V^{(l-1)}(s_0)+1 \le 2L+1$. Thus, $\|V\|_\infty \le 2L+1 \le B$.

We set $\gamma = c_{\min}/6$. As $\epsilon_{\text{VI}} \le c_{\min}/18$, we have $b(u,s,a)+|\widehat{c}(s,a)-c(s,a)|+\epsilon_{\text{VI}} \le \gamma$ when $\|u\|_\infty \le B$.

Given the policy π restricted on $\mathcal{K}$, we introduce the following operators on $\mathbb{R}^{K'}$:

$$\mathcal{L}^\pi u(s) = \widehat{c}(s,\pi(s)) - b(u,s,\pi(s)) + \widetilde{P}_{s,\pi(s)}u,$$

$$\mathcal{T}_\gamma^\pi u(s) := c(s,\pi(s)) - \gamma + \widetilde{P}_{s,\pi(s)}u.$$

We can write component-wise

$$\mathcal{T}_\gamma^\pi V \le \mathcal{L}^\pi V - \epsilon_{\text{VI}} \overset{\text{(a)}}{=} \widetilde{\mathcal{L}}V - \epsilon_{\text{VI}} \overset{\text{(b)}}{\le} V,$$

where (a) uses that π is the greedy policy with respect to V. To prove (b), we recall that $V = V^{(l)} = \widetilde{\mathcal{L}}V^{(l-1)}$. By contraction property of $\widetilde{\mathcal{L}}$ (Lem. C.4), we have $\|\widetilde{\mathcal{L}}V - V\|_\infty \le \|V^{(l)} - V^{(l-1)}\|_\infty$. By stopping condition of `VISGO`, we have $\|V^{(l)} - V^{(l-1)}\|_\infty \le \epsilon_{\text{VI}}$, thus (b) is proved. By monotonicity of the Bellman operator $\mathcal{T}_\gamma^\pi$, we have for all $m > 0$, $(\mathcal{T}_\gamma^\pi)^m V \le (\mathcal{T}_\gamma^\pi)^{m-1}V \le \cdots \le V$.

We observe that the vector $(\mathcal{T}_\gamma^\pi)^m V$ does not increase element-wise when m increases, and $(\mathcal{T}_\gamma^\pi)^m V \ge 0$ element-wise because $V \ge 0$ element-wise. Hence when $m \to \infty$, it will converge to some vector W_γ^π, where W_γ^π is the value function of policy π in the model $\widetilde{P}$ with γ subtracted to all the costs, and we have $W_\gamma^\pi \le V$ component-wise. We define the random variable $\tilde{t}_g^\pi(s)$ as the number of steps it takes to reach g starting from s on model $\widetilde{P}$ when executing policy π. Thus

$$W_\gamma^\pi(s) := \mathbb{E}_{\widetilde{P}}\left[\sum_{t=1}^{\tilde{t}_g^\pi(s)} c(s_t,\pi(s_t)) - \gamma \mid s_1 = s\right] = \widetilde{V}_g^\pi(s) - \gamma\mathbb{E}_{\widetilde{P}}\left[\tilde{t}_g^\pi(s)\right].$$

Moreover, we have $c_{\min}\mathbb{E}\left[\tilde{t}_g^\pi(s)\right] \le \widetilde{V}_g^\pi(s)$. Therefore, we get

$$\widetilde{V}_g^\pi(s) \le \frac{W_\gamma^\pi(s)}{1-\gamma/c_{\min}} \le \frac{V(s)}{1-\gamma/c_{\min}} \le \frac{4}{3}V(s).$$

Under the event $\mathcal{E}_2$, we have $\left|P^\dagger_{s,a,s'} - \widehat{P}_{s,a,s'}\right| = \widetilde{O}(\sqrt{P^\dagger_{s,a,s'}/n(s,a)} + n(s,a)^{-1})$. As $n(s,a) \ge \phi = \widetilde{\Omega}(L^2|\mathcal{K}|)$, and $B = 10L$, we can obtain $\forall (s,a,s') \in \mathcal{K}^\dagger \times \mathcal{A} \times \mathcal{K}^\dagger$,

$$\left|P^\dagger_{s,a,s'} - \widehat{P}_{s,a,s'}\right| \le \frac{c_{\min}}{24B}\sqrt{\frac{P^\dagger_{s,a,s'}}{|\mathcal{K}^\dagger|}} + \frac{c_{\min}}{24B|\mathcal{K}^\dagger|}.$$

By the Cauchy–Schwarz's inequality, $\sum_{s'\in\mathcal{K}^\dagger}\sqrt{P^\dagger_{s,a,s'}}\le\sqrt{|\mathcal{K}^\dagger|}$, hence we obtain

$$\sum_{s'\in\mathcal{K}^\dagger}\left|P^\dagger_{s,a,s'}-\widehat{P}_{s,a,s'}\right|\le\frac{c_{\min}}{12B},\quad\forall(s,a)\in\mathcal{K}^\dagger\times\mathcal{A}.$$

Also, as $|\widetilde{P}_{s,a,s'}-\widehat{P}_{s,a,s'}|\le 1/n(s,a)\le c_{\min}/(12B|\mathcal{K}^\dagger|)$, and $\widetilde{V}^\pi_g(s)\le\frac{4}{3}(2L+1)\le B$ for all $s\in\mathcal{K}^\dagger$, we can obtain that

$$\sum_{s'\in\mathcal{K}^\dagger}\left|P^\dagger_{s,a,s'}-\widetilde{P}_{s,a,s'}\right|\le\frac{c_{\min}}{6\|\widetilde{V}^\pi_g\|_\infty},\quad\forall(s,a)\in\mathcal{K}^\dagger\times\mathcal{A}.$$

Thus by simulation lemma for SSP (Lemma 3 in (Tarbouriech et al., 2020)), π is proper on true model $P^\dagger_{s,a,s'}$, and for all $s\in\mathcal{K}^\dagger$, $V^\pi_g(s)\le(1+\frac{1}{3})\widetilde{V}^\pi_g(s)=\frac{4}{3}\widetilde{V}^\pi_g(s)\le 2V(s)$. The proof is completed. □

C.2. Optimistic Property of `VISGO`

Now we will give the optimistic property. We still focus on Alg. 2, and we will prove that the output of the `VISGO` procedure (Q,V) is optimistic. And we recall that we denote V^*_g as the value function of the optimal policy on MDP $\mathcal{M}^\dagger$ to reach g, and Q^*_g as the Q-function corresponding to V^*_g.

Lemma C.6 (Optimistic Property)**.** *In Alg.2, under the event $\mathcal{E}$, for any output (Q,V) of the `VISGO` procedure, it holds that*

$$\begin{aligned}Q(s,a)&\le Q^*_g(s,a),&&\forall s\in\mathcal{K}\setminus\{g\},a\in\mathcal{A},\\ V(s)&\le V^*_g(s),&&\forall s\in\mathcal{K}^\dagger,\end{aligned}$$

where g is the goal state in `VISGO` procedure.

Proof. We prove by induction that for any inner iteration i of `VISGO`, $Q^{(i)}(s,a)\le Q^*_g(s,a)$ for any $(s,a)\in\mathcal{K}\times\mathcal{A}$, and $V^{(i)}(s)\le V^*_g(s)$ for any $s\in\mathcal{K}^\dagger$. By definition we have $Q^{(0)}=0\le Q^*_g$, and $V^{(0)}=0\le V^*_g$. Assume that the optimistic property holds for iteration i, then for any $(s,a)\in\mathcal{K}\times\mathcal{A}$ and $s\ne g$,

$$Q^{(i+1)}(s,a)=\widehat{c}(s,a)+\widetilde{P}_{s,a}V^{(i)}-b^{(i+1)}(s,a),$$

where

$$\begin{aligned}&\widehat{c}(s,a)+\widetilde{P}_{s,a}V^{(i)}-b^{(i+1)}(s,a)\\
&=\widehat{c}(s,a)+\widetilde{P}_{s,a}V^{(i)}-\max\left\{c_1\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a},V^{(i)})\iota_{s,a}}{n(s,a)}},\,c_2\frac{B\iota_{s,a}}{n(s,a)}\right\}-c_3\sqrt{\frac{\widehat{c}(s,a)\iota_{s,a}}{n(s,a)}}\\
&\overset{\text{(i)}}{\le}c(s,a)+\widetilde{P}_{s,a}V^{(i)}-\max\left\{c_1\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a},V^{(i)})\iota_{s,a}}{n(s,a)}},\,c_2\frac{B\iota_{s,a}}{n(s,a)}\right\}+\frac{28\iota_{s,a}}{3n(s,a)}\\
&=c(s,a)+f(\widetilde{P}_{s,a},\widehat{P}_{s,a},V^{(i)},n(s,a),B,\iota_{s,a})+\frac{28\iota_{s,a}}{3n(s,a)}\\
&\overset{\text{(ii)}}{\le}c(s,a)+f(\widetilde{P}_{s,a},\widehat{P}_{s,a},V^*_g,n(s,a),B,\iota_{s,a})+\frac{28\iota_{s,a}}{3n(s,a)}\\
&\overset{\text{(iii)}}{\le}c(s,a)+\widetilde{P}_{s,a}V^*_g-2\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a},V^*_g)\iota_{s,a}}{n(s,a)}}-\frac{14B\iota_{s,a}}{3n(s,a)}\\
&\overset{\text{(iv)}}{\le}c(s,a)+\widehat{P}_{s,a}V^*_g-2\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a},V^*_g)\iota_{s,a}}{n(s,a)}}-\frac{14B\iota_{s,a}}{3n(s,a)}\\
&\overset{\text{(v)}}{\le}\underbrace{c(s,a)+P_{s,a}V^*_g}_{=Q^*_g(s,a)}-(B-B_*)\frac{14\iota_{s,a}}{3n(s,a)}\\
&\le Q^*_g(s,a),\end{aligned}$$

where (i) is by definition of $\mathcal{G}_1$ and choice of c_3, (ii) uses the first property of Lem. C.2 and the induction hypothesis that $V^{(i)} \le V_g^*$, (iii) uses the second property of Lem. C.2 and assumption $B \ge \max\{B_*, 1\}$, (iv) uses Lem. C.1, (v) is by definition of $\mathcal{G}_3$. Ultimately, for any $(s,a) \in \mathcal{K} \times \mathcal{A}$ and $s \neq g$,

$$Q^{(i+1)}(s,a) \le Q_g^*(s,a).$$

Then for any $s \in \mathcal{K}$ and $s \neq g$, we have $V^{(i+1)}(s) = \min_{a \in \mathcal{A}} Q^{(i+1)}(s,a) \le \min_{a \in \mathcal{A}} Q_g^*(s,a) = V_g^*(s)$.
In addition, $V^{(i+1)}(g) = 0 = V_g^*(g)$, and we have

$$V^{(i+1)}(x) = c_{\text{RESET}} + V^{(i)}(s_0) \le c_{\text{RESET}} + V_g^*(s_0) = V_g^*(x).$$

This completes the proof of this lemma. □

D. Proof of Thm. 3.1

Here we give a proof of Thm. 3.1 and we focus on the fixed set $\mathcal{K}^\dagger$ and our constructed MDP $M^\dagger = \langle \mathcal{K}^\dagger, \mathcal{A}, P^\dagger, c^\dagger, s_0 \rangle$. We denote $K = |\mathcal{K}|$, and $K' = |\mathcal{K}^\dagger| = K + 1$.

Proof idea. First we prove that Alg.2 solves AX problem (cf. Lem. D.1), i.e., $\mathcal{K} \supseteq \mathcal{S}_L^{\rightarrow}$, and $\forall s \in \mathcal{K}, V_s^{\pi_s}(s_0) \le V_{\mathcal{K},s}^*(s_0) + \varepsilon L$. By running DisCo algorithm with $\varepsilon = 1$, we have $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$. To prove that each output policy π_s is near-optimal, we observe that in the success round with respect to goal s, the average cost of executing π_s (denoted as $\hat{\tau}$) in λ episodes is no more than $V(s_0) + \epsilon L$. As we set $\lambda = \widetilde{O}(1/\epsilon^2)$, by concentration inequalities (cf. Lem. B.4), we can obtain that the expected cost of π_s is close to the average cost $\hat{\tau}$, i.e., $V_s^{\pi_s}(s_0) \le \hat{\tau} + \epsilon L \le V(s_0) + 2\epsilon L$. By optimistic property of VISGO (Lem. C.6), we have $V(s_0) \le V_{\mathcal{K},s}^*(s_0)$, i.e., our estimation of the value function $V(s_0)$ is no more than the optimal cost. Hence we obtain $V_s^{\pi_s}(s_0) \le V_{\mathcal{K},s}^*(s_0) + 2\epsilon L \le V_{\mathcal{K},s}^*(s_0) + \varepsilon L$ for all $s \in \mathcal{K}$.
Then we bound the cumulative cost C_T. We first bound the total cost in DisCo algorithm and Alg. 1. By Lem. 2.4, with probability over $1-\delta$, DisCo algorithm with $\varepsilon = 1$ uses no more than $\widetilde{O}(L^3 S_{2L}^2 A / c_{\min}^2)$ samples. In Alg. 1, for each state-action pair $(s,a) \in \mathcal{K} \times \mathcal{A}$, we collected $\widetilde{O}(L^2 K / c_{\min}^2)$ samples. And to reach each $s \in \mathcal{K}$, we executed the policy π_s, and the cost to reach s is no larger than $\widetilde{O}(L)$ with high probability. Thus the total cost in Alg. 1 can be bounded by $\widetilde{O}(L^3 S_{2L}^2 A / c_{\min}^2)$.
Now we will bound the cumulative cost of Alg. 2 after running DisCo algorithm and Alg. 1. It's straightforward to show that the total cost in each round is bounded by $\widetilde{O}(L\lambda) = \widetilde{O}(L/\varepsilon^2)$. Hence to bound the cumulative cost C_T, we need only to bound the total number of rounds r. The number of success rounds is at most K. As the trigger condition holds for at most $\log_2(2T)$ times for each state-action pair (s,a), the number of skipped rounds can be bounded by $K'A\log_2(2T)$ (where T is the total number of samples we collected in Alg. 2). Now we need only to bound the number of failure rounds r_f.
To bound r_f, we borrow the idea from (Lim & Auer, 2012). We first define the regret of an episode as the total cost in this episode minus our estimation of the optimal cost $V(s_0)$, and define the total regret as the sum of the regret in each episode (cf. Eq. 5). Then we will give both the upper bound and lower bound of the regret, where the upper bound scales as $\widetilde{O}(\sqrt{r_f})$ and the lower bound scales as $\widetilde{\Omega}(r_f)$. For the upper bound, we extend the techniques in (Tarbouriech et al., 2021) from classical SSP to multi-goal SSP, and we obtain an upper bound that scales as $\widetilde{O}(L\epsilon^{-1}\sqrt{KAr_f} + LKA\epsilon^{-1} + LK^2A)$ (cf. Eq. 6).
For the lower bound, as the regret in each failure round is at least $\lambda\epsilon L = \widetilde{\Omega}(L/\epsilon)$, we need only to give a lower bound for the regret of all the success rounds and skipped rounds (which can be negative). We observe that the regret in any round is larger than the total cost of executing policy $\tilde{\pi}$ in this round minus the expected cost of $\tilde{\pi}$, hence we can use concentration inequalities to bound the regret in all the success rounds and skipped rounds (which scales as $-\widetilde{O}(LKA\epsilon^{-1})$). The lower bound of the total regret scales as $\widetilde{\Omega}(L\epsilon^{-1}r_f - LKA\epsilon^{-1})$ (cf. Lem. D.3). Hence we can prove that the number of failure rounds $r_f = \widetilde{O}(KA + \varepsilon K^2 A)$, and the total number of rounds $r = \widetilde{O}(S_{2L}A + \varepsilon S_{2L}^2 A)$ (here we used $K \le S_{2L}$).
As the cost in each round is bounded by $\widetilde{O}(L/\varepsilon^2)$, the cumulative cost after running Alg. 1 is bounded by $\widetilde{O}(LS_{2L}A\varepsilon^{-2} + LS_{2L}^2 A\varepsilon^{-1})$. And the cumulative cost in DisCo algorithm and Alg. 1 is bounded by $\widetilde{O}(L^3 S_{2L}^2 A c_{\min}^{-2})$, hence we complete the proof of Thm. 3.1. □

Now we give the full proof. We recall that we denote $V_g^*(s)$ as the expected cost of the optimal policy on MDP $\mathcal{M}^\dagger$ to reach goal g from state s, which equals to $V_{\mathcal{K},g}^*(s)$ for all $s \in \mathcal{K}$.

First we prove the correctness of Alg. 2, i.e., with probability over $1-\delta$, Alg. 2 outputs a set $\mathcal{K} \supseteq \mathcal{S}_L^{\rightarrow}$ and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$, such that

$$\forall s \in \mathcal{K}, V_s^{\pi_s}(s_0) \le V_{\mathcal{K},s}^*(s_0) + \varepsilon L.$$

The main intuition is that each policy π_s has been tested for λ times in a success round, and the average cost it takes to reach s from s_0 is less than our estimate for optimal cost $V(s_0)$ plus ϵL. Thus by concentration inequalities, the expected cost of π_s is close to optimal.

Lemma D.1 (VALAE Solves AX Problem)**.** *Let $\{\pi_s\}_{s\in\mathcal{K}}$ be the set of policies output by Alg.2. With probability at least $1-\delta$, $V_s^{\pi_s}(s_0) \le V_s^*(s_0) + \varepsilon L$ for all $s \in \mathcal{K}$.*

Proof. We fix any state $s \in \mathcal{K}$. In any given round where the chosen target is s, let $\hat{\tau}_k$ be the total cost in the k-th episode of that round. Recall that for the algorithm to output a policy π_s, its empirical performance after λ episodes must satisfy that $\hat{\tau} \le V(s_0) + \epsilon L$, where $\hat{\tau} = \frac{\sum_{k=1}^{\lambda}\hat{\tau}_k}{\lambda}$ and V is the output of VISGO in that round. By optimism property (Lem. C.6), when $\mathcal{K} \subseteq \mathcal{S}_{2L}^{\rightarrow}$, we have $V(s) \le 2L+1$ for all $s \in \mathcal{K}^\dagger$.

We define the random variable τ as the total cost it takes to reach the goal state s from the start state s_0 when executing policy π_s, and we have $E(\tau) = V_s^{\pi_s}(s_0)$ by definition. We note that we have collected $\phi = \widetilde{\Omega}(L^2K/c_{\min}^2)$ samples for each of the state-action pair (s,a) (cf. Alg. 1). By Lem. C.5, under event $\mathcal{E}$, the policy π_s is proper, and we have $E(\tau) \le 2V(s_0) \le 4L+2$. Moreover, we have $d := \|V_s^{\pi_s}\|_\infty \le 4L+2$. By Lem B.3, we obtain $\Pr(\tau > m) \le 2\exp(-m/4d)$ for any $m > 0$. As we set

$$\lambda = \lceil \frac{2048}{\epsilon^2}\ln^2(\frac{256}{\epsilon})\ln(\frac{2|\mathcal{K}|}{\delta})\rceil,$$

by Lem. B.4, we obtain that with probability at least $1-\delta/(2K')$, we have

$$V_s^{\pi_s}(s_0) = E(\tau) \le \hat{\tau} + \epsilon L.$$

We note that $\hat{\tau} \le V(s_0) + \epsilon L \le V_s^*(s_0) + \epsilon L$ by the optimistic property (Lem. C.6). Hence, as we set $\epsilon = \varepsilon/3$ in initial, we obtain that with probability at least $1-\delta/(2K)$,

$$V_s^{\pi_s}(s_0) \le V_s^*(s_0) + \varepsilon L.$$

Finally, as there are at most K states in total, and the event $\mathcal{E}$ holds with probability at least $1-\delta$, by the union bound and setting $\delta \to \delta/C$ in initial (C is a large constant), the total success probability is at least $1-\delta$. □

Now we focus on bounding the cumulative cost of Alg.2.

First, we bound the total cost in DisCo algorithm and Alg. 1. Disco algorithm has sample complexity $\widetilde{O}(L^3S_{(1+\varepsilon)L}^2A/(c_{\min}^2\varepsilon^2))$, and when $\varepsilon = 1$, the total cost is bounded by $\widetilde{O}(L^3S_{2L}^2A/c_{\min}^2)$. In Alg. 1, we collect $\phi = \widetilde{O}(L^2K/c_{\min}^2)$ samples for each state-action pair $(s,a) \in \mathcal{K}\times\mathcal{A}$. To collect each sample (s,a,s',c), we executed a policy π_s to reach the state s, and the expected cost $V_s^{\pi_s}(s_0) \le 2L$. By Lem. B.3, we obtain that with probability at least $1-\delta$, for any state s, each time when the policy π_s is executed, the total cost to reach s from s_0 is no larger than $O(L\log(K/\delta))$. Therefore, the total cost of Alg. 1 can be bounded by $O(KA\phi L\log(K/\delta)) = \widetilde{O}(L^3S_{2L}^2A/c_{\min}^2)$. We note that we used Lem. B.3 no more than ϕKA times, the total failure probability is no more than $\phi KA\delta$. Substituting δ by $\delta/(2\phi KA)$ in the proof, we can obtain that the total cost of DisCo and Alg. 1 is bounded by $\widetilde{O}(L^3S_{2L}^2A/c_{\min}^2)$ with probability $1-\delta$.

Then we bound the total cost of Alg.2 after running Alg. 1.

The key idea lies in bounding the "regret". We will use the regret to bound the total number of rounds. We first define the regret in the k-th episode of the j-th round. We denote $H^{j,k}$ as the number of steps it takes in the k-th episode of the j-th round. The regret in an episode k is defined as

$$(\sum_{h=1}^{H^{j,k}} c_h^{j,k}) - V^j(s_0),$$

where $c_h^{j,k}$ is the empirical cost in the h-th step in the k-th episode in the j-th round, and $V^j(s_0)$ is the value of $V(s_0)$ in the j-th round. Let n_j be the total number of episodes executed in the j-th round, we define the regret in the j-th round as follows:

$$\sum_{k=1}^{n_j}((\sum_{h=1}^{H^{j,k}} c_h^{j,k}) - V^j(s_0)).$$

Then we will define the total regret of Alg.2. Let r be the total number of rounds, n_j be the total number of episodes executed in the j-th round, and $0 \le n_j \le \lambda$. Then we know that the total number of episodes in the whole process of Alg.2 is $M = \sum_{j=1}^{r} n_j$. For notation convenience, we define H^m as the number of steps it takes in the m-th episode of the whole process of Alg.2, and denote c_h^m as the empirical cost in the h-step of episode m. Finally we define the total regret of all the rounds as

$$R := \sum_{j=1}^{r}\sum_{k=1}^{n_j}((\sum_{h=1}^{H^{j,k}} c_h^{j,k}) - V^j(s_0)) = \sum_{m=1}^{M}((\sum_{h=1}^{H^m} c_h^m) - V^m(s_0)). \tag{5}$$

We will give both the upper bound and the lower bound of the regret. Here we give the upper bound.

Lemma D.2 (Upper Bound of Regret)**.** *Under event $\mathcal{E}$, the total regret in M episodes is at most*

$$R = \widetilde{O}(L\sqrt{KAM} + LK^2A).$$

This upper bound comes from the regret bound of the EB-SSP algorithm (cf. (Tarbouriech et al., 2021)), which solves the classical SSP problem with a single goal state g. To extend their result to multi-goal SSP, instead of only concentrating on $\widehat{P}_{s,a}V_g^*$ for a single goal g and one vector V_g^*, in our high probability event $\mathcal{E}_3$, we use concentration over $(\widehat{P}_{s,a} - P_{s,a}^{\dagger})V_g^*$ for all the goal states $g \in \mathcal{K}$. Then following similar proof with Thm.3, (Tarbouriech et al., 2021), we can obtain the regret upper bound in Lem. D.2.

We note that the original form of the regret upper bound in Thm.3, (Tarbouriech et al., 2021) was $\widetilde{O}(B_*\sqrt{SAM} + BS^2A)$, where $B_* := \max_{s\in\mathcal{S}} V_g^*(s)$ in their work, B is an upper bound of B_* which is used in `VISGO`, and M is the number of episodes. In our Alg. 2, we work on MDP $\mathcal{M}^{\dagger}$, and all the states in $\mathcal{K}$ are incrementally $2L$-controllable from s_0. Hence in our settings, $B_* := \max_{(s,g)\in\mathcal{K}^{\dagger}\times\mathcal{K}} V_g^*(s) \le 2L+1$, and the number of states in MDP $\mathcal{M}^{\dagger}$ is $K' = K+1$. And in our Alg. 2, we set $B = 10L$. Therefore, by setting $B_* = O(L), B = O(L), S = K+1$ in their regret bound $\widetilde{O}(B_*\sqrt{SAM} + BS^2A)$, we can obtain the regret bound $\widetilde{O}(L\sqrt{KAM} + LK^2A)$.

We observe that there are at most $\widetilde{O}(KA)$ skipped rounds and K success rounds. We denote by r_f the number of failure rounds, and we have the total number of episodes $M = \widetilde{O}((KA + r_f)\lambda) = \widetilde{O}((KA + r_f)/\epsilon^2)$. Thus the total regret in r rounds can be bounded by r_f sublinearly:

$$R = \widetilde{O}(\frac{L}{\epsilon}\sqrt{KAr_f} + \frac{LKA}{\epsilon} + LK^2A). \tag{6}$$

Then we gives the lower bound of the total regret in terms of the number of failure rounds r_f.

Lemma D.3 (Lower Bound of Regret)**.** *With probability $1-\delta$, when $r = \widetilde{O}((KA)^2)$, the total regret in the first r rounds is at least*

$$R = \widetilde{\Omega}(\frac{Lr_f}{\epsilon} - \frac{LKA}{\epsilon}),$$

where r_f is the number of failure rounds in the r rounds.

Proof. By the criterion of our performance check, in any failure round, we have $\hat{\tau} > V(s_0) + \epsilon L$, and in round j, we have $\hat{\tau} = \frac{1}{\lambda}\sum_{k=1}^{n_j}(\sum_{h=1}^{H^{j,k}} c_h^{j,k})$ by definition. Hence, in any failure round j, the regret is $\lambda\hat{\tau} - n_j V^j(s_0) \ge \lambda(\hat{\tau} - V^j(s_0)) \ge \lambda\epsilon L = \widetilde{\Omega}(\frac{Lr_f}{\epsilon})$.

Then we focus on skipped rounds and success rounds. We denote g^j as the goal state in the j-th round, and π_j as the policy $\tilde{\pi}$ in the j-th round, which is the greedy policy over the Q-function in the j-th round. We observe that the regret in any round j satisfies

$$\sum_{k=1}^{n_j}((\sum_{h=1}^{H^{j,k}} c_h^{j,k}) - V^j(s_0)) \ge -L + \sum_{k=1}^{n_j-1}((\sum_{h=1}^{H^{j,k}} c_h^{j,k}) - V_{g^j}^*(s_0)) \ge -L + \sum_{k=1}^{n_j-1}((\sum_{h=1}^{H^{j,k}} c_h^{j,k}) - V_{g^j}^{\pi_j}(s_0)),$$

where we used the optimism property in Lem. C.6. We note that $\sum_{h=1}^{H^{j,k}} c_h^{j,k}$ is the empirical cost of policy π_j in episode k, and we will use the concentration inequality to give a lower bound of the regret in round j. As the last episode in a skipped

round can terminate before reaching the goal, we should take special considerations the last episode of each round. We directly use $-L$ to lower bound the regret of the last episode in round j. Then we denote $n = n_j - 1$, and focus on the previous n episodes in round j.

Now we fix the round index j. We denote the random variable τ as the cost to reach g^j from s_0, and we recall that $\hat{\tau}_k = \sum_{h=1}^{H^{j,k}} c_h^{j,k}$. By Lem. B.4 with $d = 4L$, with probability at least $1 - \frac{\delta}{(KA)^2}$, we have

$$\sum_{k=1}^{n}(\hat{\tau}_k - E(\tau)) \geq -2\Gamma\sqrt{n\ln(\frac{KA}{\delta})} \geq -2\Gamma\sqrt{\lambda\ln(\frac{KA}{\delta})} \geq -\widetilde{O}(\frac{L}{\epsilon}),$$

where $\Gamma = \lfloor 8d\ln(64/\epsilon)\rfloor$. Thus the regret in any round j is larger than $-\widetilde{O}(\frac{L}{\epsilon})$. As there are at most $\widetilde{O}(KA)$ skipped rounds and K success rounds, we obtain that the total regret R has the lower bound

$$R = \widetilde{\Omega}(\frac{Lr_f}{\epsilon} - \frac{LKA}{\epsilon}).$$

Now we bound the total failure probability. The number of rounds $r = \widetilde{O}((KA)^2)$, in each round the failure probability is at most $\frac{\delta}{(KA)^2}$, and the events $\mathcal{E}$ fails with probability δ. By replacing δ by δ/C throughout the proof (C is a large constant), we obtain that the total failure probability is at most δ. □

As the lower bound is linear in r_f, and the upper bound is sublinear in r_f, we can solve it and obtain that $r_f = \widetilde{O}(KA + \epsilon K^2 A)$, thus the total number of rounds can be bounded by $\widetilde{O}(KA + \epsilon K^2 A)$.

To get the cumulative cost bound in Thm. 3.1, we need only to bound the cost in a round. In any round, we observe that except for the last episode, the average cost $\hat{\tau}$ for all the other episodes is no larger than $V(s_0) + \epsilon L \leq 2L$, thus the total cost in these episodes is no larger than $2L\lambda = \widetilde{O}(L/\epsilon^2)$. Also, we know that in the any episode, the expected cost of the policy $\tilde{\pi}$ to reach the goal from s_0 is no larger than $2L$. Thus by Lem. B.3, in any round, with probability at least $1 - \frac{\delta}{(KA)^2}$, the cost in the last episode is no larger than $\widetilde{O}(L)$. Hence, the total cost in each round is no larger than $\widetilde{O}(L/\epsilon^2)$. By multiplying it with $\widetilde{O}(KA + \epsilon K^2 A)$ and using $K \leq S_{2L}$, the cumulative cost in Alg.2 can be bounded by $\widetilde{O}(LS_{2L}A/\varepsilon^2 + LS_{2L}^2 A/\varepsilon + L^3 S_{2L}A/c_{\min}^2)$, where the term $L^3 S_{2L}A/c_{\min}^2$ comes from the subroutine of DisCo algorithm and Alg. 1. Hence we obtain the bound in Thm. 3.1.

Now we count the total failure probability. First, DisCo algorithm fails with probability δ, the event $\mathcal{E}$ fails with probability δ, and the lower bound of the total regret R fails with probability δ. And in the previous paragraph, to bound the cost in the last episode of each round using Lem. B.3, the failure probability in each episode is at most $\frac{\delta}{(KA)^2}$. We observe that the total number of these failures is no larger than the total number of rounds, and the total number of rounds can be bounded by $\widetilde{O}(KA + \epsilon K^2 A)$, where we omit the logarithmic factors. Thus by setting $\delta \to \delta/C$ in the proof (C is a large constant), we can bound the total failure probability by δ, and the proof of Thm. 3.1 is completed.

Here we briefly discuss the time complexity and space complexity of our algorithm `VALAE`. The time complexity scales as $\widetilde{O}(TK^3A^2(KA + \epsilon K^2 A))$, where T is the total number of samples collected, and $\widetilde{O}(KA + \epsilon K^2 A)$ is the total number of rounds. The bottleneck is on the `VISGO` procedure, and the time complexity of a `VISGO` procedure is analyzed in Appendix G, (Tarbouriech et al., 2021), which scales as $\widetilde{O}(TK^3A^2)$. The space complexity scales as $\tilde{O}(T + K^2 A) = \tilde{O}(T)$, and the bottleneck is on storing the samples and the empirical model $\widehat{P}$.

E. Analysis of the Lower Bound

Here we discuss the lower bound of the autonomous exploration problem. We recall that our algorithm needs output a set $\mathcal{K} \supseteq \mathcal{S}_L^{\rightarrow}$ and a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$, and when $s \in \mathcal{S}_L^{\rightarrow}$, the policy π_s satisfies $V_s^{\pi_s}(s_0) \leq (1+\varepsilon)L$. Moreover, we note that in our proof of the lower bound, we allow the algorithm to output Markov policies, i.e. non-stationary and non-deterministic policies, which is defined in the next paragraph.

We recall the some basic concepts about the definition of a learning algorithm, and we use the notations in (Domingues et al., 2021). Let $\mathcal{I}^t = (\mathcal{S}\times\mathcal{A})^{t-1}\times\mathcal{S}$ be the set of all possible histories up to t steps, i.e., be the set of tuples of the form $\left(s^1, a^1, s^2, a^2, \ldots, s^t\right) \in \mathcal{I}^t$. Let $\Delta(\mathcal{A})$ be the set of probability distributions over the action space $\mathcal{A}$, and $\mathbb{N}^*$ be the set of positive integers. A Markov policy is a function $\pi : \mathcal{S}\times\mathbb{N}^* \to \Delta(\mathcal{A})$ such that $\pi(a \mid s, h)$ denotes the probability of taking action a in state s at step h. And we note that the Markov policy π is history-independent.

A history-dependent policy is a family of functions denoted as $\boldsymbol{\pi} \triangleq (\pi^t)_{t\geq 1}$, where $\pi^t : \mathcal{I}^t \to \Delta(\mathcal{A})$ describes the probability of taking action $a \in \mathcal{A}$ after observing some history $i^t \in \mathcal{I}^t$.

Given an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, c, s_0 \rangle$, a policy $\boldsymbol{\pi}$ interacting with the MDP $\mathcal{M}$ defines a stochastic process denoted by $(S^t, A^t)_{t\geq 1}$, where (S^t, A^t) is the state-action pair at time t. The Ionescu-Tulcea theorem ensures the existence of the probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\mathcal{M}})$ such that

$$\mathbb{P}_{\mathcal{M}}\big[S^1 = s\big] = \mathbb{I}[s = s_0], \mathbb{P}_{\mathcal{M}}\big[S^{t+1} = s \mid A^t, I^t\big] = p\big(s \mid S^t, A^t\big), \text{ and } \mathbb{P}_{\mathcal{M}}\big[A^t = a \mid I^t\big] = \pi^t\big(a \mid I^t\big),$$

where $\boldsymbol{\pi} = (\pi^t)_{t\geq 1}$ and for any t, $I^t \triangleq \big(S^1, A^1, S^2, A^2, \dots S^t\big)$ is the random vector in $\mathcal{I}^t$ containing all state-action pairs observed up to step t. We denote the σ-algebra generated by I^t as $\mathcal{F}^t$. And we denote by $\mathbb{P}_{\mathcal{M}}^{I^T}$ the measure of I^T under $\mathbb{P}_{\mathcal{M}}$ as follows:

$$\mathbb{P}_{\mathcal{M}}^{I^T}\big[i^T\big] \triangleq \mathbb{P}_{\mathcal{M}}\big[I^T = i^T\big] = \mathbb{I}(s^1 = s_0) \prod_{t=1}^{T-1} \pi^t\big(a^t \mid i^t\big) p\big(s^{t+1} \mid s^t, a^t\big).$$

Then we denote $\mathbb{E}_{\mathcal{M}}$ as the expectation under $\mathbb{P}_{\mathcal{M}}$. Note that the dependence of $\mathbb{P}_{\mathcal{M}}$ and $\mathbb{E}_{\mathcal{M}}$ on the policy $\boldsymbol{\pi}$ is denoted implicitly in the definition of $\mathbb{P}_{\mathcal{M}}$. We will denote them explicitly as $\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}$ and $\mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}$ respectively when we need to stress $\boldsymbol{\pi}$.

We recall that we define an algorithm for the AX problem as a tuple $(\boldsymbol{\pi}, \tau, \mathcal{K}, \{\pi_s\}_{s\in\mathcal{K}})$, where $\boldsymbol{\pi}$ is a history-dependent policy, τ is the stopping time chosen by the algorithm, $\mathcal{K}$ and $\{\pi_s\}_{s\in\mathcal{K}}$ are the output of the algorithm. And given the algorithm $\boldsymbol{\pi}$ and the MDP $\mathcal{M}$ for AX problem, we can regard the number τ, the set of states $\mathcal{K}$, and the set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ as random variables on distribution $\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}$.

Moreover, for any the MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, c, s_0 \rangle$, given a Markov policy π with goal state g, we denote $V_{\mathcal{M},g}^{\pi}(s)$ as the expected cost of policy π to reach state g from state s in MDP $\mathcal{M}$. Formally, $V_{\mathcal{M},g}^{\pi}(s) = \mathbb{E}_{\pi,\mathcal{M}}\Big[\sum_{t=1}^{t_g^{\pi}(s)} c_t(s_t, \pi(s_t))\Big]$, where $t_g^{\pi}(s) := \inf\{t \geq 0 : s_{t+1} = g\}$. And we denote $V_{\mathcal{M},g}^{*}(s)$ as the expected cost of the optimal policy π to reach the goal state g from the state s on MDP $\mathcal{M}$.

Here we introduce the basic definitions and the technical lemmas used in our proof.

Definition E.1 (KL divergence)**.** The Kullback-Leibler divergence between two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on a measurable space $(\Omega, \mathcal{G})$ is defined as

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) \triangleq \int_{\Omega} \log\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_2}(\omega)\right) \mathrm{d}\mathbb{P}_1(\omega),$$

if $\mathbb{P}_1 \ll \mathbb{P}_2$ and $+\infty$ otherwise. For Bernoulli distributions, we define $\forall (p, q) \in [0, 1]^2$,

$$\mathrm{kl}(p, q) \triangleq \mathrm{KL}(\mathcal{B}(p), \mathcal{B}(q)) = p \log\left(\frac{p}{q}\right) + (1 - p) \log\left(\frac{1 - p}{1 - q}\right).$$

Lemma E.2 (Lemma 5, (Domingues et al., 2021), modified)**.** *Let $\mathcal{M}$ and $\mathcal{M}'$ be two MDPs that are identical except for their transition probabilities, denoted by p and p', respectively. Assume that we have $\forall (s, a)$, $p(\cdot \mid s, a) \ll p'(\cdot \mid s, a)$. Then, for any stopping time τ with respect to $(\mathcal{F}^t)_{t\geq 1}$ that satisfies $\mathbb{P}_{\mathcal{M}}[\tau < \infty] = 1$,*

$$\mathrm{KL}\Big(\mathbb{P}_{\mathcal{M}}^{I^\tau}, \mathbb{P}_{\mathcal{M}'}^{I^\tau}\Big) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mathbb{E}_{\mathcal{M}}\big[N_{s,a}^{\tau}\big] \mathrm{KL}(p(\cdot \mid s, a), p'(\cdot \mid s, a)),$$

where $N_{s,a}^{\tau} \triangleq \sum_{t=1}^{\tau} \mathbb{1}\{(S^t, A^t) = (s, a)\}$ and I^τ is the random vector representing the history of τ samples.

Lemma E.3 (Lemma 1, (Garivier et al., 2019))**.** *Consider a measurable space $(\Omega, \mathcal{F})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $\mathcal{F}$-measurable function $Z : \Omega \to [0, 1]$, we have*

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) \geq \mathrm{kl}(\mathbb{E}_1[Z], \mathbb{E}_2[Z]),$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively.

Lemma E.4. *For any $p, q \in (0, \frac{1}{2}]$, $\mathrm{kl}(p, q) \leq 2(p - q)^2/q$.*

Lemma E.5 (Lemma 15, (Domingues et al., 2021))**.** *For any $p, q \in [0, 1]$, $\mathrm{kl}(p, q) \geq -(1 - p)\log(1 - q) - \log(2)$.*

Now we construct a family of adversarial MDPs to obtain the lower bound of sample complexity.

The construction of hard MDPs with general S_L. Now we fix $L, S, A, S_L, \varepsilon, c_{\min}$ such that $L > 4$, $S > 8$, $A > 4$, $4 \leq S_L \leq \min\{(A - 1)^{\lfloor \frac{L}{2} \rfloor}, S\}$, $0 < \varepsilon < \frac{1}{4}$, and $0 < c_{\min} \leq 1$.

We first construct an MDP $\mathcal{M}_0' = \langle \mathcal{S}, \mathcal{A}, p_0', c', s_0 \rangle$ with $|\mathcal{S}| = S_L - 1$ states and $|\mathcal{A}| = A$ actions ($\mathcal{M}_0'$ does not contain the goal state g in Fig. 2). As is illustrated in Fig. 2, the construction on $\mathcal{M}_0'$ follows a tree structure. This is inspired from (Domingues et al., 2021). We denote $\mathcal{S}'$ as all the leaf states, and for any $s \notin \mathcal{S}'$ and $a \in \mathcal{A}$ ($a \neq$ RESET), we set $c(s,a) = 1$ with probability 1, and the transition $p_0'(\cdot|s,a)$ is deterministic, i.e., taking any action a at a non-leaf node s will transit to one of its son $s' \in \mathcal{S}$ with probability 1.
As $S_L \leq (A-1)^{\lfloor \frac{L}{2} \rfloor}$, there exists a tree structure with depth $d_0 \leq L/2$ ($d_0 \in \mathbb{N}$), such that the number of leaves $|\mathcal{S}'| \geq \frac{S_L}{2}$, and $V^*_{\mathcal{M}_0',s}(s_0) = d_0$ for all $s \in \mathcal{S}'$, i.e., all the leaf nodes can be reached within d_0 steps from s_0. Hence in MDP $\mathcal{M}_0'$, all the states in $\mathcal{S}$ are incrementally d_0-controllable. And we denote $d_1 = L - d_0$.

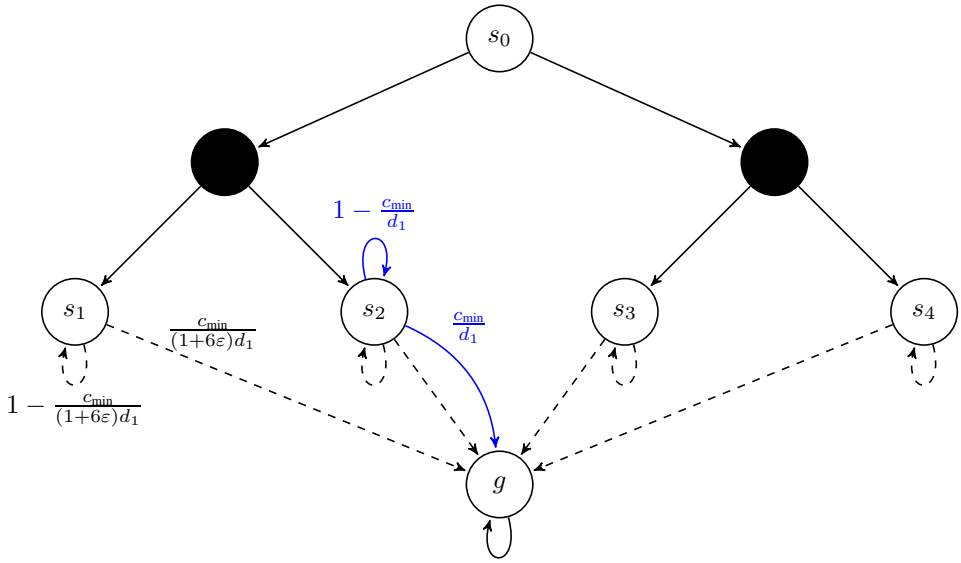

Figure 2: Illustration of the hard MDP with general S_L.

Then we construct the MDP $\mathcal{M}_0 = \langle \mathcal{S} \cup \{g\}, \mathcal{A}, p_0, c, s_0 \rangle$ based on $\mathcal{M}_0'$ by adding a new state g. The state g can only be reached from all the leaf nodes $s \in \mathcal{S}'$. And for any leaf state $s \in \mathcal{S}'$ and any action $a \in \mathcal{A}$ ($a \neq$ RESET), we set $c(s,a) = c_{\min}$, and

$$p_0(g|s,a) = \frac{c_{\min}}{(1+6\varepsilon)d_1}, \quad p_0(s|s,a) = 1 - \frac{c_{\min}}{(1+6\varepsilon)d_1}.$$

Finally, we set $p_0(g|g,a) = 1$ for any action $a \neq$ RESET.
In this way, we have $V^*_{\mathcal{M}_0,s}(s_0) = d_0$ for any $s \in \mathcal{S}'$, and $V^*_{\mathcal{M}_0,g}(s_0) = d_0 + (1+6\varepsilon)d_1 > (1+\varepsilon)L$. Hence $g \notin \mathcal{S}_L^{\rightarrow}$ for MDP $\mathcal{M}_0$. Also, we note that in MDP $\mathcal{M}_0$, with probability 1, the goal state g cannot be reached from s_0 within d_0 steps.
Now we will construct other adversarial MDPs based on $\mathcal{M}_0$. We choose any $(s^*, a^*) \in \mathcal{S}' \times \mathcal{A}$ ($a^* \neq$ RESET), and we define the MDP $\mathcal{M}_{(s^*,a^*)} = \langle \mathcal{S} \cup \{g\}, \mathcal{A}, p_{(s^*,a^*)}, c, s_0 \rangle$ by slightly increasing $p_0(g|s^*,a^*)$, i.e., we set

$$p_{(s^*,a^*)}(g|s^*,a^*) = \frac{c_{\min}}{d_1}, \quad p_{(s^*,a^*)}(s^*|s^*,a^*) = 1 - \frac{c_{\min}}{d_1}.$$

In this way, we have $V^*_{\mathcal{M}_0,g}(s_0) = d_0 + d_1 = L$. Hence $g \in \mathcal{S}_L^{\rightarrow}$ for MDP $\mathcal{M}_{(s^*,a^*)}$.
Finally, we define the family of our adversarial MDPs as $\{\mathcal{M}_0\} \cup \{\mathcal{M}_{(s,a)}\}_{(s,a) \in \mathcal{S}' \times \mathcal{A}}$. We note that for each MDP $\mathcal{M}_{(s,a)}$, its $|\mathcal{S}_L^{\rightarrow}| = S_L$, and it satisfies Asmp. 2.1. Also, for the MDP $\mathcal{M}_0$, its $|\mathcal{S}_L^{\rightarrow}| = S_L - 1$, and it also satisfies Asmp. 2.1. Thus the family is valid for the AX problem.
We note that in MDP $\mathcal{M}_{(s^*,a^*)}$, for any Markov policy π,

$$V^{\pi}_{\mathcal{M}_{(s^*,a^*)},g}(s_0) = \mathbb{E}_{\pi,\mathcal{M}_{(s,a)}}\left[\sum_{t=1}^{t_g^{\pi}(s)} c_t(s_t, \pi(s_t))\right]$$

$$= \mathbb{E}_{\pi,\mathcal{M}_{(s^*,a^*)}}\left[\sum_{t=1}^{t_g^\pi(s)} c_t(s_t,\pi(s_t)) \mid (s_{d_0},a_{d_0})=(s^*,a^*)\right]\mathbb{P}_{\pi,\mathcal{M}_{(s^*,a^*)}}[(s_{d_0},a_{d_0})=(s^*,a^*)]$$

$$+ \mathbb{E}_{\pi,\mathcal{M}_{(s^*,a^*)}}\left[\sum_{t=1}^{t_g^\pi(s)} c_t(s_t,\pi(s_t)) \mid (s_{d_0},a_{d_0})\neq(s^*,a^*)\right]\mathbb{P}_{\pi,\mathcal{M}_{(s^*,a^*)}}[(s_{d_0},a_{d_0})\neq(s^*,a^*)]$$

With probability 1, we need at least d_0 steps to reach any of the leaf state. And the expected cost to reach g from state-action pair (s^*,a^*) is d_1.
Hence we have $\mathbb{E}_{\pi,\mathcal{M}_{(s^*,a^*)}}\left[\sum_{t=1}^{t_g^\pi(s)} c_t(s_t,\pi(s_t)) \mid (s_{d_0},a_{d_0})=(s^*,a^*)\right] \geq d_0+d_1 = L$.
Also, when $(s_{d_0},a_{d_0})\neq(s^*,a^*)$, the expected cost to reach g from state-action pair (s_{d_0},a_{d_0}) is at least $(1+6\varepsilon)d_1$.
Hence we have $\mathbb{E}_{\pi,\mathcal{M}_{(s^*,a^*)}}\left[\sum_{t=1}^{t_g^\pi(s)} c_t(s_t,\pi(s_t)) \mid (s_{d_0},a_{d_0})\neq(s^*,a^*)\right] \geq d_0+(1+6\varepsilon)d_1 \geq (1+3\varepsilon)L$.
Therefore, if $V^\pi_{\mathcal{M}_{(s^*,a^*)},g}(s_0)\leq(1+\varepsilon)L$, we have $\mathbb{P}_{\pi,\mathcal{M}_{(s^*,a^*)}}[(s_{d_0},a_{d_0})=(s^*,a^*)]\geq 2/3$. We will use it in our proof of Thm. 4.2.
Now we give our proof of Thm. 4.2 through the adversarial family of MDPs. Here we use the techniques of Thm. 7 in (Tarbouriech et al., 2020).

Proof. We denote by $\mathbb{P}_{(s^*,a^*)} \triangleq \mathbb{P}_{\boldsymbol{\pi},\mathcal{M}_{(s^*,a^*)}}$ and $\mathbb{E}_{(s^*,a^*)} \triangleq \mathbb{E}_{\boldsymbol{\pi},\mathcal{M}_{(s^*,a^*)}}$ the probability measure and expectation in the MDP $\mathcal{M}_{(s^*,a^*)}$ by following $\boldsymbol{\pi}$ and by $\mathbb{P}_0$ and $\mathbb{E}_0$ the corresponding operators in the MDP $\mathcal{M}_0$. We fix any algorithm $(\boldsymbol{\pi},\tau,\mathcal{K},\{\pi_s\}_{s\in\mathcal{K}})$ that solves the AX problem. We will prove that when working on the MDP $\mathcal{M}_0$, the algorithm will cost at least $\Omega(\frac{LS_LA}{c_{\min}\varepsilon^2}\log\frac{1}{\delta})$ samples in expectation, i.e.

$$\mathbb{E}_0[\tau] = \Omega(\frac{LS_LA}{c_{\min}\varepsilon^2}\log\frac{1}{\delta}),$$

which yields that the lower bound of the total cost is $\Omega(\frac{LS_LA}{\varepsilon^2}\log\frac{1}{\delta})$.
Now we fix the state-action pair $(s^*,a^*)\in\mathcal{S}'\times\mathcal{A}$ ($a^*\neq$ RESET). Also, we denote the random variable $N^\tau_{(s,a)}$ as the number of samples that the algorithm takes at the state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$. For any $\mathcal{F}^\tau$-measurable random variable Z taking values in $[0,1]$, we have

$$\begin{aligned}
&\mathbb{E}_0\left[N^\tau_{(s^*,a^*)}\right]\frac{144c_{\min}\varepsilon^2}{L}\\
\overset{(a)}{\geq}\ &\mathbb{E}_0\left[N^\tau_{(s^*,a^*)}\right]\mathrm{kl}\left(\frac{c_{\min}}{(1+6\varepsilon)d_1},\frac{c_{\min}}{d_1}\right)\\
\overset{(b)}{=}\ &\mathrm{KL}\left(\mathbb{P}_0^{I^\tau},\mathbb{P}_{(s^*,a^*)}^{I^\tau}\right)\\
\overset{(c)}{\geq}\ &\mathrm{kl}\big(\mathbb{E}_0[Z],\mathbb{E}_{(s^*,a^*)}[Z]\big),
\end{aligned}$$

where (a) uses Lemma E.4 and $d_1\geq L/2$; (b) uses Lemma E.2; (c) uses Lemma E.3.
For any $(s,a)\in\mathcal{S}'\times\mathcal{A}$, we define the event $Z_{s,a}=\mathbb{1}\{$The algorithm's output satisfies $g\in\mathcal{K}$ and $V^{\pi_g}_{\mathcal{M}_{(s,a)},g}(s_0)\leq(1+\varepsilon)L\}$. And we set the event $Z=Z_{s^*,a^*}$. We note that $Z_{s,a}$ can be viewed as a random event on distribution $\mathbb{P}_{(s,a)}$, and can also be viewed as a random event on distribution $\mathbb{P}_0$ (i.e., $\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}_0}$).
First we focus on distribution $\mathbb{P}_{(s^*,a^*)}$. We observe that as the algorithm $(\boldsymbol{\pi},\tau,\mathcal{K},\{\pi_s\}_{s\in\mathcal{K}})$ solves the AX problem, when working on the MDP $\mathcal{M}_{(s^*,a^*)}$, with probability at least $1-\delta$, its output should satisfy $g\in\mathcal{K}$ and the expected cost of the policy π_g to reach g from state s_0 is no more than $(1+\varepsilon)L$. Therefore, for any $(s^*,a^*)\in\mathcal{S}'\times\mathcal{A}$ ($a^*\neq$ RESET), we have

$$\mathbb{P}_{(s^*,a^*)}[Z_{s^*,a^*}]\geq 1-\delta.$$

Then we focus on probability distribution $\mathbb{P}_0$ (i.e., $\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}_0}$). We recall that the event $Z_{s,a}$ implies $\mathbb{P}_{\pi_g,\mathcal{M}_0}[(s_{d_0},a_{d_0})=(s,a)]\geq 2/3$. And for any two distinct state-action pairs (s,a) and (s',a'), the event $\mathbb{P}_{\pi_g,\mathcal{M}_0}[(s_{d_0},a_{d_0})=(s,a)]\geq 2/3$ and the event $\mathbb{P}_{\pi_g,\mathcal{M}_0}[(s_{d_0},a_{d_0})=(s',a')]\geq 2/3$ are mutually exclusive. Hence $Z_{s,a}$ and $Z_{s',a'}$ are mutually exclusive on $\mathbb{P}_0$, and we have

$$\sum_{(s,a)\in\mathcal{S}'\times\mathcal{A}}\mathbb{P}_0[Z_{s,a}]\leq 1.$$

We recall that we set $Z = Z_{s^*,a^*}$, and we can obtain

$$\begin{aligned}\mathrm{kl}\big(\mathbb{E}_0[Z], \mathbb{E}_{(s^*,a^*)}[Z]\big) &= \mathrm{kl}\big(\mathbb{P}_0[Z_{s^*,a^*}], \mathbb{P}_{(s^*,a^*)}[Z_{s^*,a^*}]\big) \\ &\overset{\text{(a)}}{\geq} (1-\mathbb{P}_0[Z_{s^*,a^*}])\log\left(\frac{1}{1-\mathbb{P}_{(s^*,a^*)}[Z_{s^*,a^*}]}\right) - \log(2) \\ &\overset{\text{(b)}}{\geq} (1-\mathbb{P}_0[Z_{s^*,a^*}])\log\left(\frac{1}{\delta}\right) - \log(2),\end{aligned}$$

where (a) uses Lem. E.5; (b) uses that $\mathbb{P}_{(s^*,a^*)}[Z_{s^*,a^*}] \geq 1-\delta$. Therefore, we have

$$\mathbb{E}_0\left[N^{\tau}_{(s^*,a^*)}\right] \geq \frac{L}{144 c_{\min}\varepsilon^2}((1-\mathbb{P}_0[Z_{s^*,a^*}])\log\left(\frac{1}{\delta}\right) - \log(2)).$$

We recall that $\sum_{(s,a)\in\mathcal{S}'\times\mathcal{A}} \mathbb{P}_0[Z_{s,a}] \leq 1$. Thus summing up all the state-action pairs $(s^*,a^*) \in \mathcal{S}'\times\mathcal{A}$, we can obtain that

$$\sum_{(s^*,a^*)\in\mathcal{S}'\times\mathcal{A}} \mathbb{E}_0\left[N^{\tau}_{(s^*,a^*)}\right] \geq \frac{L}{144 c_{\min}\varepsilon^2}((|\mathcal{S}'||\mathcal{A}|-1)\log\left(\frac{1}{\delta}\right) - \log(2)|\mathcal{S}'||\mathcal{A}|).$$

Hence provided that $|\mathcal{S}'| \geq \frac{S_L}{2}$, $L > 4$, $S > 8$, $A > 4$, $4 \leq S_L \leq \min\{(A-1)^{\lfloor \frac{L}{2} \rfloor}, S\}$, $0 < \varepsilon < \frac{1}{4}$, and $0 < \delta < \frac{1}{16}$, we can eventually obtain the lower bound of the total number of steps τ,

$$\mathbb{E}_0[\tau] = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{E}_0[N^{\tau}_{(s,a)}] \geq \sum_{(s^*,a^*)\in\mathcal{S}'\times\mathcal{A}} \mathbb{E}_0\left[N^{\tau}_{(s^*,a^*)}\right] \geq \Omega(\frac{LS_LA}{c_{\min}\varepsilon^2}\log\frac{1}{\delta}).$$

□

F. Lower Bounds for Multi-goal SSP

Here we formulize the lower bound for the multi-goal SSP problem. First we define an algorithm for the multi-goal SSP problem with goal space $\mathcal{G}$ as a triple $(\boldsymbol{\pi}, \tau, \{\pi_s\}_{s\in\mathcal{G}})$, which means the algorithm executes a history-dependent policy $\boldsymbol{\pi}$, and returns a set of policies $\{\pi_s\}_{s\in\mathcal{G}}$ after sampling τ times. Also, we allow π_s to be Markov policies. And we release the multi-goal SSP problem in this way: we only require the algorithm output policies π_s such that $V_s^{\pi_s}(s_0) \leq (1+\varepsilon)L$.

Definition F.1. An algorithm $(\boldsymbol{\pi}, \tau, \{\pi_s\}_{s\in\mathcal{S}})$ is (ε,δ,L)-PAC for multi-goal SSP problem on MDP M with goal space $\mathcal{G} \subseteq \mathcal{S}$, if with probability over $1-\delta$, the algorithm returns a set of policies $\{\pi_s\}_{s\in\mathcal{G}}$ after τ steps, such that $\forall s \in \mathcal{G}, V_s^{\pi_s}(s_0) \leq (1+\varepsilon)L$.

Then for any real numbers $L, c_{\min}$ and positive integers S, A, we define a class of MDPs $\mathfrak{M}_{\text{MSSP}}(L,S)$ as follows: $\mathfrak{M}_{\text{MSSP}}(L,S)$ contains all the MDPs $M = \langle \mathcal{S}, \mathcal{A}, P, c, s_0\rangle$, such that $|\mathcal{S}| \leq S$, $|\mathcal{A}| \leq A$, $c(s,a) \in [c_{\min}, 1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and M satisfies Asmp. 2.1 and $\mathcal{S}_L^{\rightarrow} = \mathcal{S}$.

We remark that our constructed adversarial examples (cf. Fig. 2) for the autonomous exploration problem can also be applied to multi-goal SSP using the similar proof with Thm. 4.2. Thus we obtain the following lower bound for multi-goal SSP, which implies that our Alg. 2 is also minimax for multi-goal SSP problem. See Appendix E for more details.

Theorem F.2. *Assume that $L > 4$, $A > 4$, $8 < S \leq (A-1)^{\lfloor \frac{L}{2} \rfloor}$, $0 < \varepsilon < \frac{1}{4}$, $0 < \delta < \frac{1}{16}$, and $0 < c_{\min} \leq 1$. Then for any algorithm $(\boldsymbol{\pi}, \tau, \mathcal{K}, \{\pi_s\}_{s\in\mathcal{K}})$ that is (ε,δ,L)-PAC for multi-goal SSP problem on any MDP $M \in \mathfrak{M}_{MSSP}(L,S)$ with any goal space $\mathcal{G} \subseteq \mathcal{S}$, there exists an MDP $\mathcal{M} \in \mathfrak{M}_{MSSP}(L,S)$ such that*

$$\mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}[\tau] = \Omega(\frac{LSA}{c_{\min}\varepsilon^2}\log\frac{1}{\delta}).$$

We note that in our construction of adversarial examples (cf. Fig. 2) and in our proof of Thm. 4.2, we only involved one goal state g. Hence we can also prove that for the classical single-goal SSP problem with $\mathcal{G} = \{g\}$, learning a policy π_g such that $V_g^{\pi_g}(s_0) \leq (1+\varepsilon)L$ also requires $\Omega(\frac{LSA}{c_{\min}\varepsilon^2}\log\frac{1}{\delta})$ samples, and the lower bound of cumulative cost scales as $\Omega(LSA\varepsilon^{-2}\log\frac{1}{\delta})$.