Reward-Free Exploration beyond Finite-Horizon

Jean Tarbouriech [1 2] **Matteo Pirotta** [1] **Michal Valko** [2] **Alessandro Lazaric** [1]

Abstract

We consider the reward-free exploration framework recently introduced by Jin et al. (2020), where an RL agent interacts with an unknown environment without any explicit reward function to maximize. The objective is to collect enough information during the exploration phase, so that a near-optimal policy can be immediately computed once a specific reward function is provided. In this paper, we move from the finite-horizon setting studied by Jin et al. (2020) to the more general setting of goal-conditioned RL, often referred to as stochastic shortest path (SSP). We first discuss the challenges specific to SSPs and then we study two scenarios: **1)** reward-free goal-free exploration in communicating MDPs, and **2)** reward-free goal-free incremental exploration in non-communicating MDPs where the agent is provided with an action to reset to an initial state. In both cases, we devise novel exploration algorithms and derive sample-complexity bounds.

1. Introduction

In problems where the reward function is sparse or even absent, a reinforcement learning (RL) agent needs to explore the environment driven by objectives other than reward maximization. Recent unsupervised exploration deep RL algorithms successfully solved complex problems such as Montezuma's Revenge (Ecoffet et al., 2020) or real-world robotic manipulation tasks (Pong et al., 2019) solely driven by the objective of *discovering* and *controlling* the environment. Nonetheless, the problem still lacks of a rigorous formalization and algorithms do not have solid theoretical guarantees. A first step in that direction is the reward-free exploration framework introduced by Jin et al. (2020) in finite-horizon Markov decision processes (MDPs). Jin et al. (2020) define an exploration phase where the agent interacts with an unknown environment and collects information about its dynamics. Then in a planning phase, the agent is provided with a reward function and it must return a near-optimal policy without any further learning. The performance of the agent is evaluated by the number of samples collected during the exploration phase.

While the finite-horizon setting is very popular in theoretical RL, it is rarely representative of the type of problems considered in popular benchmarks and real applications in RL. In this paper, we rather focus on the strictly more general and more practical stochastic shortest path (SSP) setting (Bertsekas, 2012) (often referred to as goal-conditioned RL), where the objective is to compute a policy that minimizes the cost accumulated before reaching a specific goal state. We first reformulate the reward-free exploration setting by defining the objective of learning an accurate enough model of the environment so that a near-optimal policy can be computed for *any* SSP problem (i.e., for any initial state, any goal state, and any cost function). We illustrate how this problem may be considerably more difficult than in the finite-horizon setting. We then study two different scenarios (i.e., goal-free cost-free exploration in communicating MDPs and goal-free cost-free incremental exploration in non-communicating MDPs with restart), summarize the sample complexity results that we obtain and contrast them with the guarantees in the finite-horizon case. The details of the algorithms are postponed to the appendix.

2. Preliminaries

A Markov decision process (MDP) is defined as $M := \langle \mathcal{S}, \mathcal{A}, p, c \rangle$, where $\mathcal{S}$ is the state space with $S := |\mathcal{S}|$ states and $\mathcal{A}$ is the action space with $A := |\mathcal{A}|$ actions. Taking action a in state s incurs a cost[1] of $c(s,a) \in [0,1]$ and the next state $s' \in \mathcal{S}$ is selected with probability $p(s'|s,a)$. We denote by $\Gamma := \max_{s,a} \|p(\cdot|s,a)\|_0$ the largest support of the transition model. In the SSP case, for a designated goal state $\overline{s}$, the objective is to compute a policy $\pi : \mathcal{S} \to \mathcal{A}$ minimizing the cumulative cost before reaching $\overline{s}$. Formally, we define the (possibly unbounded) value function

$$V_\pi(\underline{s} \to \overline{s}) := \mathbb{E}\left[\sum_{t=1}^{\tau_\pi(\underline{s} \to \overline{s})} c(s_t, \pi(s_t)) \;\middle|\; s_1 = \underline{s} \right],$$

where $\tau_\pi(\underline{s} \to \overline{s}) := \inf\{t \geq 0 : s_{t+1} = \overline{s} \,|\, s_1 = \underline{s}, \pi\}$ is the (random) number of steps needed to reach $\overline{s}$ from $\underline{s}$

[1]Facebook AI Research, Paris [2]Inria Lille - Nord Europe.

Accepted at the ICML 2020 Workshop on Theoretical Foundations of Reinforcement Learning.

[1]One can translate between costs and rewards by simply taking negation.

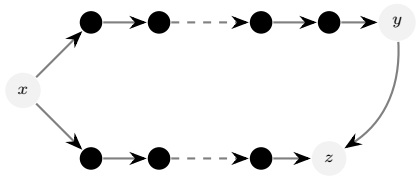

Figure 1. The agent starts at state x and reaches z in H steps with probability $1/2$, and y in $H+1$ steps with probability $1/2$. From state y the agent deterministically transitions to state z in 1 step.

when executing policy π. An optimal policy (if it exists) is denoted by $\pi^\star \in \arg\min_\pi V_\pi(\underline{s} \to \overline{s})$. For more details on the SSP problem we refer to e.g., Bertsekas (2012, Sect. 3).

Jin et al. (2020) introduced the reward-free framework in the finite-horizon case, which is a special case of the SSP problem where each episode terminates after exactly H steps. The agent receives as input an accuracy level $\varepsilon > 0$, a confidence level $\delta \in (0,1)$, the state and action spaces, and the horizon H, while no knowledge is provided about the transition model p. The learning process is decomposed into two phases. ① *Exploration phase:* The agent first collects trajectories from the MDP without a pre-specified reward function and returns an estimate of the transition model $\widehat{p}$. ② *Planning phase:* The agent receives an arbitrary reward function and is tasked with computing an ε-optimal policy with probability at least $1-\delta$, without any additional interaction with the environment. The objective is to minimize the duration of the exploration phase needed to simultaneously enforce any requested planning guarantee.

Jin et al. (2020) study the reward-free exploration problem for any arbitrary MDP, where there may exist states that are difficult or impossible to reach. The core mechanism in their analysis is to partition the states depending on their ease of being reached within H steps. Specifically, they distinguish between *significant* states, that can be sufficiently visited and whose transition probability can thus be accurately estimated, and *insignificant* states that are too difficult to reach within H steps, but therefore have negligible contribution to any reward optimization.

Interestingly, in the goal-conditioned setting this distinction may no longer be meaningful. By way of illustration, consider any fixed horizon H and the toy environment in Fig. 1. Suppose that the objective is to quickly reach state z (i.e., the goal state is z, the starting state is x and all costs are equal to 1). Even though state y is *insignificant* within H steps (in the finite-horizon sense of Jin et al., 2020, for any positive "significance level"), it is actually crucial in solving the objective, as z can be reached deterministically in 1 step from y. Extrapolating this scenario, in the goal-conditioned setting, we may have an effective horizon of $H = +\infty$ for some goals, which implies that the transition model p must be accurately estimated across the state-action space to ensure that a near-optimal policy can be computed.

3. Goal-Free Cost-Free Exploration in Communicating MDPs

In order to guarantee that the environment can be estimated uniformly well, we introduce the following assumption.

Assumption 1 (In Sect. 3)**.** *The MDP M is communicating, with finite and unknown diameter*

$$D = \max_{\underline{s},\overline{s}} D_{\underline{s},\overline{s}} = \max_{\underline{s},\overline{s}} \min_\pi \mathbb{E}[\tau_\pi(\underline{s} \to \overline{s})] < +\infty.$$

We stress that the challenges that emerge in such setting are orthogonal to the ones in (Jin et al., 2020): a *constraint on the environment is added* (all states must now be reachable), allowing the *removal of the constraint on performance* (which is not limited to H steps anymore) and thus enabling to tackle the more general class of goal-oriented problems.

Without loss of generality, we consider throughout that the maximum $c_{\max}$ of the cost functions that we intend to consider in the planning phase is equal to 1. On the other hand, the minimum value $c_{\min}$ has a more subtle impact on the type of performance guarantees we can obtain. In particular, for any cost function c and any pair of initial and goal states $\underline{s}$ and $\overline{s}$, we introduce a slack parameter $\theta \in [1, +\infty]$ and we say that a policy $\widehat{\pi}$ is (ε, θ)-optimal if [2]

$$V^{\widehat{\pi}}(\underline{s} \to \overline{s}) \le \min_{\pi : \mathbb{E}[\tau_\pi(\underline{s}\to\overline{s})] \le \theta D_{\underline{s},\overline{s}}} V^\pi(\underline{s} \to \overline{s}) + \varepsilon.$$

In the following theorem, we show that depending on the minimum cost $c_{\min}$ in the cost functions of interest and the slack θ, we can solve the goal-free cost-free exploration problem with a bounded sample complexity.

Theorem 1. *Consider any unknown environment satisfying Asm. 1 and the goal-free cost-free exploration problem characterized by an accuracy level $0 < \varepsilon \le 1$, a confidence level $\delta \in (0,1)$, a minimum cost $c_{\min} \in [0,1]$ and a slack parameter $\theta \in [1, +\infty]$. There exists an algorithm $\mathfrak{A}$ whose exploration phase (i.e., number of time steps) is bounded with probability at least $1-\delta$ by*

$$\widetilde{O}\left(\frac{D^4 \Gamma S A}{\omega \varepsilon^2} + \frac{D^3 S^2 A}{\omega \varepsilon} + \frac{D^3 \Gamma S A}{\omega^2}\right),$$

$$\text{where:} \qquad \omega := \max\left\{c_{\min}, \frac{\varepsilon}{\theta D}\right\}.$$

Note that we can have either $c_{\min} = 0$ or $\theta = +\infty$, but not both simultaneously, to guarantee that $\omega > 0$. Following this exploration phase, the algorithm $\mathfrak{A}$ can compute in the planning phase, for any pair of starting and goal state $(\underline{s}, \overline{s}) \in \mathcal{S}^2$, and for any cost function c in $[c_{\min}, 1]$, a policy $\widehat{\pi}$ (depending on $c, \underline{s}, \overline{s}$) that is (ε, θ)-optimal.

[2] This reduces to standard ε-optimality for $\theta \to \infty$.

Algorithmic principle. We first leverage the sample complexity analysis of (Tarbouriech et al., 2020b) to solve SSP problems with a generative model, and we define the number of samples that are needed in each state-action pair to compute an estimated model $\widehat{p}$ that is accurate enough that a near-optimal policy can be computed for any cost function. Then we leverage the online learning algorithm GOSPRL recently introduced in (Tarbouriech et al., 2020c) that explicitly collects the desired amount of samples in any communicating environment. Interestingly, such algorithm is simply defining a sequence of SSP problems, where the goal state is any state for which the required number of samples is not achieved yet.

4. Goal-Free Cost-Free Incremental Exploration

In this section, we seek to provide cost-free guarantees in MDPs with possibly very large state space and diameter (e.g., for non-communicating MDPs where $D = \infty$). In order to make such setting feasible, we need to restrict the type of SSP problems we would like to solve during the planning phase. We propose an alternative approach that builds on the setting of incremental autonomous exploration introduced by Lim & Auer (2012).

Assumption 2 (In Sect. 4)**.** *The MDP M has a finite, possibly large state space $\mathcal{S}$ for which an upper bound S on its cardinality is known, i.e., $|\mathcal{S}| \leq S$.*[3] *It contains a designated initial state $s_0 \in \mathcal{S}$. Since the learner may get stuck in a state without being able to return to s_0, we assume that the action space contains a* RESET *action s.t. $p(s_0|s, \text{RESET}) = 1$ for any $s \in \mathcal{S}$.*

We make explicit the states where a policy π takes action RESET in the following definition.

Definition 1. *For $\mathcal{S}' \subseteq \mathcal{S}$ a policy π is restricted on $\mathcal{S}'$ if $\pi(s) = \text{RESET}$ for any $s \notin \mathcal{S}'$. We denote by $\Pi(\mathcal{S}')$ the set of policies restricted on $\mathcal{S}'$.*

We denote by $\Gamma_{\mathcal{S}'} := \max_{s \in \mathcal{S}', a} \|\{p(s'|s,a)\}_{s' \in \mathcal{S}'}\|_0$ the largest support of the model p restricted to states in $\mathcal{S}' \subseteq \mathcal{S}$.

In (Lim & Auer, 2012), given an input parameter $L \geq 1$ and accuracy $\varepsilon > 0$, the objective of the agent is to identify the set of *incrementally* L-reachable states $\mathcal{S}_L^{\rightarrow}$ (Def. 2), as well as a set of goal-conditioned policies to reach each state in $\mathcal{S}_L^{\rightarrow}$ from s_0 in at most $L + \varepsilon$ steps on average.[4]

[3] Lim & Auer (2012) consider a countable, possibly infinite state space; however this leads to a technical issue in the analysis of UCBEXPLORE (acknowledged by the authors via personal communication), which requires considering finite state spaces.

[4] Lim & Auer (2012) showed that discovering all states in $\mathcal{S}_L := \{s \in \mathcal{S} : \min_{\pi \in \Pi} \mathbb{E}[\tau_\pi(s_0 \to s)] \leq L\}$ may not feasible and require a number of exploration steps that is *exponential* in L or $|\mathcal{S}_L|$, hence the definition of incrementally reachable states.

Definition 2 (Incrementally controllable states $\mathcal{S}_L^{\rightarrow}$)**.** *Let $\prec$ be some partial order on $\mathcal{S}$. The set $\mathcal{S}_L^{\prec}$ of states controllable in L steps w.r.t. $\prec$ is defined inductively as follows. The initial state s_0 belongs to $\mathcal{S}_L^{\prec}$ by definition and if there exists a policy π restricted on $\{s' \in \mathcal{S}_L^{\prec} : s' \prec s\}$ with $\mathbb{E}[\tau_\pi(s_0 \to s)] \leq L$, then $s \in \mathcal{S}_L^{\prec}$. The set $\mathcal{S}_L^{\rightarrow}$ of incrementally L-controllable states is defined as $\mathcal{S}_L^{\rightarrow} := \cup_{\prec} \mathcal{S}_L^{\prec}$, where the union is over all possible partial orders.*

Finally, we introduce $S_L := |\mathcal{S}_L^{\rightarrow}|$ and $\Gamma_L := \Gamma_{\mathcal{S}_L^{\rightarrow}}$.

We extend the formalism of (Lim & Auer, 2012) and define a more challenging cost-free objective. In particular, at the end of the exploration phase, an algorithm should be able to compute a near-optimal policy restricted on $\mathcal{S}_L^{\rightarrow}$ for any SSP problem with initial state s_0, any goal state $s \in \mathcal{S}_L^{\rightarrow}$, and any cost function. As the state space $\mathcal{S}$ may be very large, the set $\mathcal{S}_L^{\rightarrow}$ effectively captures our area of interest, with L being the radius of interest provided as input. The fact that $\mathcal{S}_L^{\rightarrow}$ is unknown in advance and is hard to estimate online poses an additional challenge.

Similar to Thm. 1 we provide an (ε, θ)-optimality guarantee for the planning phase with the additional condition that we consider policies restricted to the initially unknown set $\mathcal{S}_L^{\rightarrow}$.

Theorem 2. *Consider any unknown environment satisfying Asm. 2 and the goal-free cost-free incremental exploration problem characterized by an accuracy level $0 < \varepsilon \leq 1$, a confidence level $\delta \in (0,1)$, a minimum cost $c_{\min} \in [0,1]$ and a slack parameter $\theta \in [1, +\infty]$. There exists an algorithm $\mathfrak{A}$ whose exploration phase (i.e., number of time steps) is bounded with probability at least $1 - \delta$ by*

$$\widetilde{O}\left(\frac{L^5 \Gamma_{L+\varepsilon} S_{L+\varepsilon} A}{\omega^2 \varepsilon^2} + \frac{L^3 S_{L+\varepsilon}^2 A}{\omega \varepsilon} \right),$$

$$\text{where:} \qquad \omega := \max\left\{ c_{\min}, \frac{\varepsilon}{\theta L} \right\}.$$

Note that we can have either $c_{\min} = 0$ or $\theta = +\infty$, but not both simultaneously, to guarantee that $\omega > 0$. Following this exploration phase, the algorithm $\mathfrak{A}$ has confidently identified a set $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K} \subseteq \mathcal{S}_{L+\varepsilon}^{\rightarrow}$, and has collected enough information such that for any goal state $\overline{s} \in \mathcal{S}_L^{\rightarrow}$ and any cost function c in $[c_{\min}, 1]$, it can compute in the planning phase a policy $\widehat{\pi}$ (depending on $c, \overline{s}$) that verifies

$$V^{\widehat{\pi}}(s_0 \to \overline{s}) \leq \min_{\pi \in \Pi(\mathcal{S}_L^{\rightarrow}) : \mathbb{E}[\tau_\pi(s_0 \to \overline{s})] \leq \theta L} V^{\pi}(s_0 \to \overline{s}) + \varepsilon.$$

Algorithmic principle. Despite the difference in the setting, we leverage similar algorithmic principles as in Sect. 3. In this case, we define the sample requirements *limited* to the states that have currently been discovered and for which a shortest-path policy is available. Such policies are then used to collect new samples and a one-step random exploration is used to expand the set of controllable states until all incrementally L-controllable states have been identified.

Reference	RF-FH — (Jin et al., 2020)	RF-COMM — Sect. 3 of this paper (Thm. 1)		RF-INC — Sect. 4 of this paper (Thm. 2)	
Setting	Finite-horizon RL	Goal-conditioned RL (i.e., SSP)		Goal-conditioned RL (i.e., SSP)	
Feedback	Any rewards $r \in [0,1]$	Any goal state, any costs $c \in [c_{\min}, 1]$ with $c_{\min} \geq 0$		Any goal state in $S_L^{\rightarrow}$, any costs $c \in [c_{\min}, 1]$ with $c_{\min} \geq 0$	
MDP	👍 Non-communicating and resetting after H steps	👎 Communicating with diameter D		👍 Non-communicating and reset action	
Optimality	👎 Restricted to H steps	👍 Arbitrary* length to goal		👍 Arbitrary* length to goal + 👎 Incremental Optimality	
State dep.	👎 Total state space S	👎 Total state space S		👍 State space of interest $S_L^{\rightarrow} \ll S$	
Sample comp.	$\widetilde{O}\left(\frac{S^2 A\, poly(H)}{\varepsilon^2} + \frac{S^4 A\, poly(H)}{\varepsilon}\right)$	$\widetilde{O}\left(\frac{S^2 A\, poly(D)}{\varepsilon^2}\right)$ for not too small $c_{\min} > 0$	$\widetilde{O}\left(\frac{S^2 A\, poly(D)}{\varepsilon^3}\right)$ for very small $c_{\min} \simeq 0$	$\widetilde{O}\left(\frac{S_L^2 A\, poly(L)}{\varepsilon^2}\right)$ for not too small $c_{\min} > 0$	$\widetilde{O}\left(\frac{S_L^2 A\, poly(L)}{\varepsilon^4}\right)$ for very small $c_{\min} \simeq 0$

Table 1. High-level comparison between (Jin et al., 2020) and this paper. Asterisk* introduces the subtlety that, only in the case of $c_{\min} \simeq 0$ (the second sub-column), the length to goal targeted by the candidate policy is restricted if it is "too long".

5. Discussion

We stress that RF-FH (Jin et al., 2020), RF-COMM (Sect. 3) and RF-INC (Sect. 4) tackle orthogonal settings, each posing different challenges. That notwithstanding, we believe that it is insightful to compare the three settings in terms of algorithmic approach and resulting bound (see Table 1).

Similarities in the three algorithmic designs. All three approaches construct accurate estimates of the transitions. RF-FH (Jin et al., 2020) restrict their attention to "significant" states within H steps. As previously explained, such a reasoning cannot be directly extended to general SSP problems, as there is no more notion of fixed horizon, with some states possibly becoming non-negligible for value optimization at some random point before the goal state is reached. This is why RF-COMM enforces to visit uniformly enough the state-action space, which explains the need for the communicating assumption (Asm. 1). By focusing on incremental exploration, RF-INC can effectively restrict its attention to the (unknown) state space of interest $\mathcal{S}_L^{\rightarrow}$, which removes the need for the communicating assumption. Finally, note that to collect the sought-after samples, Jin et al. (2020) deploy a finite-horizon algorithm for regret minimization, whereas our algorithms leverage SSP policies.

Comparison between RF-FH and RF-COMM. In the main order term w.r.t. ε, the dependencies in S^2 and A are equivalent, matching the lower bound derived in the finite-horizon case (Jin et al., 2020, Thm. 4.1). Moreover, the role of the horizon H in RF-FH is captured by the ratio $D/c_{\min}$ in RF-COMM (when $c_{\min} > 0$). Note that this ratio is not a strict horizon (as the performance may last longer, as opposed to finite-horizon which always truncates it at H steps), and it is environment-dependent and thus crucially *unknown*, which introduces a significant layer of complexity to the problem. RF-COMM (Thm. 1) is the first result tackling the reward-free framework beyond finite-horizon, for goal-conditioned RL in communicating MDPs. The resulting exploration bound scales polynomially with D, which is somewhat unavoidable. Finally, the bound of RF-COMM inherits a $\widetilde{O}(\varepsilon^{-2})$ dependency (as in RF-FH of Jin et al., 2020) whenever $c_{\min}$ is not too small (and can therefore be considered as a constant). Otherwise, RF-COMM can cope with very small (or even zero-valued) $c_{\min}$, yet the bound worsens to $\widetilde{O}(\varepsilon^{-3})$, and the performance becomes *restricted* to policies with not too large expected goal-reaching time (via the slack parameter θ). This interesting behavior does not appear in the finite-horizon case (where the range of rewards has no influence on the rate in ε), and it captures the key role of the minimum cost played in the behavior of the optimal goal-reaching policy.

Specificity of RF-INC. The incremental focus of RF-INC enables to tackle goal-conditioned tasks while removing the communicating assumption of RF-COMM, where the dependency on the diameter D is replaced by the parameter L, which may be designed to be much smaller than D. In fact, while L defines the horizon of interest, resetting after every L steps (as in finite-horizon) would prevent the agent to identify incrementally L-reachable states and lead to poor performance. Another interesting element of comparison is the dependency on the size of the state space. While the RF-FH algorithm of (Jin et al., 2020) is robust w.r.t. states that can be reached with very low probability, it still displays a polynomial dependency on the global state space S. On the other hand, in virtue of its incremental focus, RF-INC (Thm. 2) depends polynomially on the number of $(L+\varepsilon)$-controllable states and only *logarithmically* on S. This result is significant since not only $S_{L+\varepsilon}$ can be arbitrarily smaller than S, but also because the set $\mathcal{S}_{L+\varepsilon}^{\rightarrow}$ itself is initially unknown to the learner.

REFERENCES

Bertsekas, D. *Dynamic programming and optimal control*, volume 2. 2012.

Cohen, A., Kaplan, H., Mansour, Y., and Rosenberg, A. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, 2020.

Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O., and Clune, J. First return then explore. *arXiv preprint arXiv:2004.12919*, 2020.

Fruit, R., Pirotta, M., and Lazaric, A. Improved analysis of UCRL2 with empirical bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

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

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

Pong, V. H., Dalal, M., Lin, S., Nair, A., Bahl, S., and Levine, S. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019.

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

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. On the sample complexity of stochastic shortest path with a generative model, 2020b. URL https://jtarbouriech.github.io/docs/ssp_genmodel.pdf.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. A provably efficient sample collection strategy for reinforcement learning. *arXiv preprint arXiv:2007.06437*, 2020c.

A. Cost-Free Goal-Free Exploration in Communicating MDPs (Sect. 3)

We leverage the GOSPRL algorithm of (Tarbouriech et al., 2020c), an algorithm that mimics the behavior of a generative model in communicating MDPs. Specifically, in any unknown communicating environment with diameter D and for any arbitrary (possibly time-varying) requirement of samples $b_t(s,a)$ (where the sequence is bounded from above by $\bar{b}(s,a)$), GOSPRL requires (with high probability) at most $\widetilde{O}(BD + D^{3/2}S^2A)$ times steps to collected the sought-after samples for each state-action pair (s,a), where $B \leq \sum_{s,a} \bar{b}(s,a)$.

We now show that instantiating GOSPRL for carefully selected sampling requirements $b_t(s,a)$ enables to obtain the guarantee of Thm. 1. To do so, we build on the sample complexity analysis of solving SSP problems with a generative model derived in (Tarbouriech et al., 2020b, Thm. 1). As such, we introduce the following sampling requirement function

$$\phi(X,y) := \alpha \cdot \left(\frac{X^3 \widehat{\Gamma}}{y\varepsilon^2} \log\left(\frac{XSA}{y\varepsilon\delta}\right) + \frac{X^2 S}{y\varepsilon} \log\left(\frac{XSA}{y\varepsilon\delta}\right) + \frac{X^2\widehat{\Gamma}}{y^2} \log^2\left(\frac{XSA}{y\delta}\right) \right), \tag{1}$$

where $\alpha > 0$ is a numerical constant and $\widehat{\Gamma} := \max_{s,a} \|\widehat{p}(\cdot|s,a)\|_0 \leq \Gamma$ is the largest support of $\widehat{p}$.

This sampling requirement function for carefully selected values of X and y is used to guide the GOSPRL algorithm. Specifically, we set y to be equal to the minimum cost (in either the true or cost-perturbed model), i.e., $y := \omega^{-1}$. As for the value of X, let us perform the following distinction of cases.

① First let us assume that the learning agent has prior knowledge of the diameter D. Then we set $X = D$. From (Tarbouriech et al., 2020b), collecting at least $\phi(D, \omega^{-1})$ samples from each state-action pair enables to guarantee the ε-optimality cost-free planning guarantee of Thm. 1. The total time required to collect such samples is upper bounded by $DSA\phi(D,\omega^{-1})$, which directly yields the sample complexity guarantee stated in Thm. 1.

② Second we show that we can relax the assumption of knowing the diameter D without altering the sample complexity guarantee. To do so, we begin the algorithm by a procedure with computes a quantity $\widehat{D}$ such that $D \leq \widehat{D} \leq D(1+\varepsilon)$ with high probability. From (Tarbouriech et al., 2020c, App. H), this can be done in $\widetilde{O}(D^3S^2A/\varepsilon^2)$ time steps by leveraging GOSPRL. We thus begin the algorithm by running such diameter-estimation subroutine. Crucially, we note that its sample complexity is *subsumed* in the total sample complexity of Thm. 1. Then we simply apply the reasoning in case ① by considering $X = \widehat{D}$ in the allocation of Eq. 1 instead of $X = D$. Since $\widehat{D}$ is a sufficiently tight upper bound on D (i.e., $\widehat{D} = O(D)$), we ultimately obtain the same sample complexity guarantee as in case ①.

B. Goal-Free Cost-Free Incremental Exploration (Sect. 4)

We now provide the algorithm DISCO, a novel algorithm for incremental exploration, which yields the guarantee of Thm. 2.

Unit-cost case. We first focus on the unit-cost setting, where the objective is defined as follows.

Definition 3 (AX* sample complexity)**.** *Fix any length $L \geq 1$, error threshold $\varepsilon > 0$ and confidence level $\delta \in (0,1)$. The sample complexities $\mathcal{C}_{\mathrm{AX}^\star}(\mathfrak{A}, L, \varepsilon, \delta)$ is defined as the number of time steps required by a learning algorithm $\mathfrak{A}$ to identify a set $\mathcal{K} \supseteq \mathcal{S}_L^{\rightarrow}$ such that with probability at least $1-\delta$, it has learned a set of policies $\{\pi_s\}_{s\in\mathcal{K}}$ that verifies the following requirement: $\forall s \in \mathcal{K}, v_{\pi_s}(s_0 \to s) \leq V^\star_{\mathcal{S}_L^{\rightarrow}}(s_0 \to s) + \varepsilon$, where we set $v_{\pi_s}(s_0 \to s) := \mathbb{E}[\tau_{\pi_s}(s_0 \to s)]$.*

The algorithm DISCO (*Discover and Control*) is detailed in Alg. 1. It maintains a set $\mathcal{K}$ of "controllable" states and a set $\mathcal{U}$ of states that are considered "uncontrollable" *so far*. A state s is tagged as "controllable" when a policy to reach s in at most $L+\varepsilon$ steps has been found (with high confidence), and we denote by π_s such policy. The states in $\mathcal{U}$ are states that have been discovered as potential members of $\mathcal{S}_L^{\rightarrow}$, but the algorithm has yet to produce a policy to control any of them in less than $L+\varepsilon$ steps. All observed samples are used to estimate the underlying transition model denoted by $\widehat{p}(s'|s,a) = N(s,a,s')/N(s,a)$. The algorithm proceeds through rounds, which are indexed by k and incremented whenever a state in $\mathcal{U}$ gets transferred to the set $\mathcal{K}$, i.e., when the transition model reaches a level of accuracy sufficient to compute a policy to control one of the states encountered before. We denote by $\mathcal{K}_k$ (resp. $\mathcal{U}_k$) the set of controllable (resp. uncontrollable) states at the beginning of round k. DISCO stops when it can confidently claim that all remaining states are not L-controllable.

Each round k is divided into two phases. The first is a *sample collection* phase. At the beginning of round k, the agent collects additional samples until n_k samples are available at each state-action pair in $\mathcal{K}_k \times \mathcal{A}$ (step ①). A key challenge lies in the careful (and adaptive) choice of the minimal required value for n_k (see Thm. 3 for its exact definition). Importantly, the incremental construction of $\mathcal{K}_k$ entails that sampling at each state $s \in \mathcal{K}_k$ can be done efficiently. In fact, for all $s \in \mathcal{K}_k$ the agent has already confidently learned a policy π_s to reach s in at most $L+\varepsilon$ steps on average (see how such policy is computed in the second phase). The sample collection phase actually achieves two objectives at once. First, it serves as a discovery step: when generating new samples from a state-action pair $(s,a) \in \mathcal{K}_k \times \mathcal{A}$, all the observed states not in $\mathcal{U}_k$ are added to it. Second, it improves the accuracy of the model p on $\mathcal{K}_k$, which is essential in computing near-optimal policies.

The second phase does not require interacting with the environment and it focuses on the *computation of optimistic policies*. The agent begins by significantly "pruning" the set of candidate states to alleviate the computational complexity of the algorithm. Namely, among all the states in $\mathcal{U}_k$, it discards those that do not have a high probability of belonging to $\mathcal{S}_L^{\rightarrow}$ by considering a restricted set $\mathcal{W}_k \subseteq \mathcal{U}_k$ (step ②). In fact, if the estimated probability $\widehat{p}_k$ of reaching a state $s \in \mathcal{U}_k$ from any of the controllable states in $\mathcal{W}_k$ is lower than $(1-\varepsilon/2)/L$, then no shortest-path policy restricted on $\mathcal{K}_k$ could get to s from s_0 in less than $L+\varepsilon$ steps on average. Then DISCO proceeds with computing for all states in $\mathcal{W}_k$ an optimistic policy restricted on $\mathcal{K}_k$. For any candidate state $s' \in \mathcal{W}_k$, we define the induced stochastic shortest path (SSP) MDP M'_k as follows.

Definition 4. *We define the SSP-MDP $M'_k := \langle \mathcal{S}, \mathcal{A}'_k(\cdot), c'_k, p'_k \rangle$, where the action space is such that $\mathcal{A}'_k(s) = \mathcal{A}$ for all $s \in \mathcal{K}_k$ and $\mathcal{A}'_k(s) = \{\mathrm{RESET}\}$ otherwise (i.e., we restrict our attention to policies restricted on $\mathcal{K}_k$). The cost function is such that for all $a \in \mathcal{A}$, $c'_k(s',a) = 0$ and for any $s \neq s'$, $c'_k(s,a) = 1$. The transition model is $p'_k(s'|s',a) = 1$ and $p'_k(\cdot|s,a) = p(\cdot|s,a)$ otherwise.*[5]

The solution of M'_k is the shortest-path policy from s_0 to s' restricted on $\mathcal{K}_k$. Since p'_k is unknown, DISCO cannot compute the exact solution of M'_k, but instead, it executes optimistic value iteration EVI for SSP (Tarbouriech et al., 2020a) to obtain a value function $\widetilde{u}_{s'}$ and its associated greedy policy $\widetilde{\pi}_{s'}$ restricted on $\mathcal{K}_k$.

The agent then chooses a candidate goal state $s^\dagger$ for which the value $\widetilde{u}^\dagger := \widetilde{u}_{s^\dagger}(s_0)$ is the smallest. This step can be interpreted as selecting the optimistically most promising new state to control. Two cases are possible. If $\widetilde{u}^\dagger \leq L$, then $s^\dagger$ is added to $\mathcal{K}_k$ (step ④), since the policy $\widetilde{\pi}_{s^\dagger}$ is able to "reach" the state $s^\dagger$ in less than L steps with high probability (i.e., $s^\dagger$ is L-controllable). Otherwise, we can guarantee that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}_k$ with high probability. In this case, the algorithm terminates and, using the current estimates of the model, it recomputes an optimistic shortest-path policy π_s restricted on the final set $\mathcal{K}_k$ for each state $s \in \mathcal{K}_k$ (step ⑤).

[5] In words, all actions in states in $\mathcal{K}_k$ behave exactly as in M and suffer a cost of 1, in all states outside $\mathcal{K}_k$ only the reset action to s_0 is available with a cost of 1, and all actions in s' self-loop with a cost of 0. Notice also that in practice we do not need to define M'_k over the whole state space $\mathcal{S}$ and we can limit it to the set of states observed so far, thus greatly reducing the complexity of the value iteration algorithm used to compute the optimistic policy (see App. B.2.1 for more details).

Algorithm 1: Algorithm DISCO

Input: Actions $\mathcal{A}$, initial state s_0, confidence parameter $\delta \in (0,1)$, error threshold $\varepsilon > 0$, $L \geq 1$ and (possibly adaptive) allocation function $\phi : \mathcal{P}(\mathcal{S}) \to \mathbb{N}$.

Initialize $k := 0$, $\mathcal{K}_0 := \{s_0\}, \mathcal{U}_0 := \{\}$ and a restricted policy $\pi_{s_0} \in \Pi(\mathcal{K}_0)$.
Set $\varepsilon := \min\{\varepsilon, 1\}$ and `continue` := `True`.
while `continue` **do**
  Set $k \mathrel{+}= 1$. `//new round`
  `//` ① **Sample collection on** $\mathcal{K}$
  For each $(s,a) \in \mathcal{K}_k \times \mathcal{A}$, execute policy π_s until the total number of visits $N_k(s,a)$ to (s,a) satisfies $N_k(s,a) \geq n_k := \phi(\mathcal{K}_k)$. For each $(s,a) \in \mathcal{K}_k \times \mathcal{A}$, add $s' \sim p(\cdot|s,a)$ to $\mathcal{U}_k$ if $s' \notin \mathcal{K}_k$.
  `//` ② **Restriction of candidate states** $\mathcal{U}$
  Compute transitions $\widehat{p}_k(s'|s,a)$ and $\mathcal{W}_k := \left\{ s' \in \mathcal{U}_k : \exists (s,a) \in \mathcal{K}_k \times \mathcal{A}, \widehat{p}_k(s'|s,a) \geq \frac{1-\varepsilon/2}{L} \right\}$.
  if $\mathcal{W}_k$ is empty **then**
    Set `continue` := `False`. `//condition STOP1`
  else
    `//` ③ **Computation of the optimistic policies on** $\mathcal{K}$
    for each state $s' \in \mathcal{W}_k$ **do**
      Compute $(\widetilde{u}_{s'}, \widetilde{\pi}_{s'}) := \mathrm{EVI}(\mathcal{K}_k, \mathcal{A}, s', N_k, \frac{\varepsilon}{6L})$.
    Let $s^\dagger := \arg\min_{s \in \mathcal{W}_k} \widetilde{u}_s(s_0)$ and $\widetilde{u}^\dagger := \widetilde{u}_{s^\dagger}(s_0)$.
    if $\widetilde{u}^\dagger > L$ **then**
      Set `continue` := `False`. `//condition STOP2`
    else
      `//` ④ **State transfer from** $\mathcal{U}$ **to** $\mathcal{K}$
      Set $\mathcal{K}_{k+1} := \mathcal{K}_k \cup \{s^\dagger\}$, $\mathcal{U}_{k+1} := \mathcal{U}_k \setminus \{s^\dagger\}$ and $\pi_{s^\dagger} := \widetilde{\pi}_{s^\dagger}$.
`//` ⑤ **Final policy computation on** $\mathcal{K}$
for each state $s \in \mathcal{K}_k$ **do**
  Compute $(\widetilde{u}_s, \widetilde{\pi}_s) := \mathrm{EVI}(\mathcal{K}_k, \mathcal{A}, s, N_k, \frac{\varepsilon}{6L})$.

Output: the states s in $\mathcal{K}_k$ and their corresponding policy $\pi_s := \widetilde{\pi}_s$.

B.1. Sample Complexity Analysis of DISCO

We now present our main result: a sample complexity guarantee for DISCO for the problem AX*.

Theorem 3. *There exists an absolute constant $\alpha > 0$ such that for any $L \geq 1$, $\varepsilon \in (0,1]$, and $\delta \in (0,1)$, if we set the allocation function ϕ as*

$$\phi : \mathcal{X} \to \alpha \cdot \left(\frac{L^4 \widehat{\Theta}(\mathcal{X})}{\varepsilon^2} \log^2\left(\frac{LSA}{\varepsilon\delta}\right) + \frac{L^2|\mathcal{X}|}{\varepsilon} \log\left(\frac{LSA}{\varepsilon\delta}\right) \right), \tag{2}$$

with $\widehat{\Theta}(\mathcal{X}) := \max_{(s,a)\in\mathcal{X}\times\mathcal{A}} \|\{\widehat{p}(s'|s,a)\}_{s'\in\mathcal{X}}\|_0$, then the algorithm DISCO *(Alg. 1) satisfies the following sample complexity bound for* AX*

$$\mathcal{C}_{\mathrm{AX}^\star}(\textsc{DisCo}, L, \varepsilon, \delta) = \widetilde{O}\left(\frac{L^5 \Gamma_{L+\varepsilon} S_{L+\varepsilon} A}{\varepsilon^2} + \frac{L^3 S_{L+\varepsilon}^2 A}{\varepsilon} \right), \tag{3}$$

where $S_{L+\varepsilon} := |\mathcal{S}^{\rightarrow}_{L+\varepsilon}|$ and

$$\Gamma_{L+\varepsilon} := \max_{(s,a)\in\mathcal{S}^{\rightarrow}_{L+\varepsilon}\times\mathcal{A}} \|\{p(s'|s,a)\}_{s'\in\mathcal{S}^{\rightarrow}_{L+\varepsilon}}\|_0 \leq S_{L+\varepsilon}$$

is the maximal support of the transition probabilities $p(\cdot|s,a)$ restricted to the set $\mathcal{S}^{\rightarrow}_{L+\varepsilon}$.

Given the definition of AX*, Thm. 3 implies that DISCO **1)** terminates after $\mathcal{C}_{\mathrm{AX}^\star}(\textsc{DisCo}, L, \varepsilon, \delta)$ time steps, **2)** discovers a set of states $\mathcal{K} \supseteq \mathcal{S}^{\rightarrow}_L$ with $|\mathcal{K}| \leq S_{L+\varepsilon}$, **3)** and for each $s \in \mathcal{K}$ outputs a policy π_s which is ε-optimal w.r.t. policies restricted on $\mathcal{S}^{\rightarrow}_L$ (i.e., $v_{\pi_s}(s_0 \to s) \leq V^\star_{\mathcal{S}^{\rightarrow}_L}(s_0 \to s) + \varepsilon$).

General cost case of Thm. 2. The extension to the general cost case is straightforward insofar as that the model is accurately estimated over the state space of interest $\mathcal{S}_L^{\rightarrow}$. It follows from adapting the allocation function of DisCo in Eq. 2 (as stipulated by the simulation lemma for SSP of Tarbouriech et al., 2020b, Lem. 4) and extending the proof of Thm. 3 beyond the unit-cost case (specifically, Lem. 5 and 8).

B.2. Proof of Thm. 3

B.2.1. Computation of the Optimistic Policies

At each round k, for each target state $s^\dagger \in \mathcal{W}_k$, DisCo computes an optimistic goal-oriented policy associated to the MDP $M'_k(s^\dagger)$ constructed as in Def. 4. This MDP is defined over the entire state space $\mathcal{S}$ and restricts the action to the only action RESET outside $\mathcal{K}_k$. We can build an equivalent MDP by restricting the focus on $\mathcal{K}_k$. To this end, we define the following SSP-MDP.

Definition 5. *Define* $M_k^\dagger(s^\dagger) := \langle \mathcal{S}_k^\dagger, \mathcal{A}_k^\dagger(\cdot), c_k^\dagger, p_k^\dagger \rangle$ *where* $\mathcal{S}_k^\dagger := \mathcal{K}_k \cup \{s^\dagger, x\}$ *and* $S_k^\dagger = |\mathcal{S}_k^\dagger| = |\mathcal{K}_k| + 2$. *State* x *is a meta-state that encapsulates all the states that have been observed so far and are not in* $\mathcal{K}_k$. *The action space* $\mathcal{A}_k^\dagger(\cdot)$ *is such that* $\mathcal{A}_k^\dagger(s) = \mathcal{A}$ *for all states* $s \in \mathcal{K}_k$ *and* $\mathcal{A}_k^\dagger(s) = \{\text{RESET}\}$ *for* $s \in \{s^\dagger, x\}$. *The cost function is* $c_k^\dagger(x,a) = 0$ *for any* $a \in \mathcal{A}_k^\dagger(x)$ *and* $c_k^\dagger(s,a) = 1$ *everywhere else. The transition function is defined as* $p_k^\dagger(s^\dagger|s^\dagger,a) = p_k^\dagger(s_0|x,a) = 1$ *for any* a, $p_k^\dagger(y|s,a) = p(y|s,a)$ *for any* $(s,a,y) \in \mathcal{K}_k \times \mathcal{A} \times (\mathcal{K}_k \cup \{s^\dagger\})$ *and* $p_k^\dagger(x|s,a) = 1 - \sum_{y \in \mathcal{K}_k \cup \{s^\dagger\}} p_k^\dagger(y|s,a)$.

Note that solving $M_k^\dagger$ we get a policy effectively restricted to the set $\mathcal{K}_k$ since we can interpret the meta-state x as $\mathcal{S} \setminus \{\mathcal{K}_k \cup \{s^\dagger\}\}$. Since p is unknown, we cannot construct $M_k^\dagger(s^\dagger)$. Let N_k the state-action counts accumulated up until now. We denote by $\widehat{p}_k$ the "global" empirical estimates, i.e., $\widehat{p}_k(y|s,a) = N_k(s,a,y)/N_k(s,a)$. Given them, we define the "restricted" empirical estimates $\widehat{p}_k^\dagger$ as follows: $\widehat{p}_k^\dagger(y|s,a) := \widehat{p}_k(y|s,a)$ for any $(s,a,y) \in \mathcal{K}_k \times \mathcal{A} \times (\mathcal{K}_k \cup \{s^\dagger\})$ and $\widehat{p}_k^\dagger(x|s,a) := 1 - \sum_{y \in \mathcal{K}_k \cup \{s^\dagger\}} \widehat{p}_k^\dagger(y|s,a)$. Denoting $N_k^+(s,a) := \max\{1, N_k(s,a)\}$, we then define the following bonuses for any $(s,a,y) \in \mathcal{K}_k \times \mathcal{A} \times (\mathcal{K}_k \cup \{s^\dagger\})$,

$$\beta_k(s,a,y) := 4\sqrt{\frac{\widehat{p}_k(y|s,a)}{N_k^+(s,a)} \log\left(\frac{3SAN_k^+(s,a)}{\delta}\right)} + \frac{28 \log\left(\frac{3SAN_k^+(s,a)}{\delta}\right)}{N_k^+(s,a)}, \tag{4}$$

$$\beta_k(s,a,x) := \sum_{y \in \mathcal{K}_k \cup \{s^\dagger\}} \beta_k(s,a,y). \tag{5}$$

Moreover, we set the uncertainty about the MDP at the meta-state x and at the goal state $s^\dagger$ to 0 by construction (since there outgoing transitions are deterministic, respectively to s_0 and $s^\dagger$).

We now leverage the extended value iteration (EVI) scheme for SSP proposed in (Tarbouriech et al., 2020a), with the confidence intervals selected according to Eqs. 4, 5. Specifically, we compute the optimistic value function $\widetilde{u}_k^\dagger$ and policy $\widetilde{\pi}_k^\dagger$ using EVI for SSP: $(\widetilde{u}_k^\dagger, \widetilde{\pi}_k^\dagger) = \text{EVI}(\mathcal{S}_k^\dagger, \mathcal{A}_k^\dagger, s^\dagger, N_k, \frac{\varepsilon}{4L})$, with $\mathcal{S}_k^\dagger$ the state space, $\mathcal{A}_k^\dagger$ the action space, $s^\dagger$ the goal state, N_k the samples collected so far and $\frac{\varepsilon}{4L}$ the VI precision level (see Tarbouriech et al., 2020a).

B.2.2. High-Probability Events

We now spell out two high probability events. Define the set of plausible transition probabilities as

$$C_k^\dagger := \bigcap_{(s,a) \in \mathcal{S}_k^\dagger \times \mathcal{A}} C_k^\dagger(s,a),$$

where

$$C_k^\dagger(s,a) := \{\widetilde{p} \in \mathcal{C} \mid \widetilde{p}(\cdot \mid s^\dagger, a) = \mathbb{1}_{s^\dagger}, \widetilde{p}(\cdot \mid x, a) = \mathbb{1}_{s_0}, |\widetilde{p}(s'|s,a) - \widehat{p}_k(s'|s,a)| \le \beta_k(s,a,s')\},$$

with $\mathcal{C}$ the $S_k^\dagger$-dimensional simplex and $\widehat{p}_k$ the empirical average of transitions.

Lemma 1. *Introduce the event* $\Theta := \bigcap_{k=1}^{+\infty} \bigcap_{s^\dagger \in \mathcal{W}_k} \{p_k^\dagger \in C_k^\dagger\}$. *Then* $\mathbb{P}(\Theta) \ge 1 - \frac{\delta}{3}$.

Proof. We have with probability at least $1 - \frac{\delta}{3}$ that, for any $y \neq x$, $|p_k^\dagger(y|s,a) - \widehat{p}_k^\dagger(y|s,a)| \leq \beta_k(s,a,y)$ from the empirical Bernstein inequality (see e.g., Fruit et al., 2020, Thm. 10), and moreover $|\widehat{p}_k^\dagger(x|s,a) - p_k^\dagger(x|s,a)| = \left|1 - \sum_{y \in \mathcal{K}_k \cup \{s^\dagger\}} p_k^\dagger(y|s,a) - \left(1 - \sum_{y \in \mathcal{K}_k \cup \{s^\dagger\}} \widehat{p}_k^\dagger(y|s,a)\right)\right| \leq \sum_{y \in \mathcal{K}_k \cup \{s^\dagger\}} |p_k^\dagger(y|s,a) - \widehat{p}_k^\dagger(y|s,a)| \leq \beta_k(s,a,x)$. □

Lemma 2 (Variant of Lem. 17 of (Lim & Auer, 2012))**.** *Suppose that for every state $s \in \mathcal{S}$, each action $a \in \mathcal{A}$ is executed $b \geq \lceil L \log\left(\frac{3ALS}{\delta}\right) \rceil$ times. Let $\mathcal{S}'_{s,a}$ be the set of all next states visited during the b executions of (s,a). Denote by Λ the complementary of the event*

$$\left\{ \exists (s', s, a) \in \mathcal{S}^2 \times \mathcal{A} : p(s'|s,a) \geq \frac{1}{L} \wedge s' \notin \mathcal{S}'_{s,a} \right\}.$$

Then $\mathbb{P}(\Lambda) \geq 1 - \frac{\delta}{3}$.

Throughout the remainder of the proof, we will assume that the event $\Theta \cap \Lambda$ holds.

B.2.3. Properties of the Optimistic Policies and Value Vectors

We recall notation. Let us fix any round k and any target state $s^\dagger \in \mathcal{W}_k$. We denote by $\widetilde{\pi}_k^\dagger$ the greedy policy w.r.t. $\widetilde{u}_k^\dagger(\cdot \to s^\dagger)$ in the optimistic model $\widetilde{p}_k^\dagger$. Let $\widetilde{v}_k^\dagger(s \to s^\dagger)$ be the value function of policy $\widetilde{\pi}_k^\dagger$ starting from state s in the model $\widetilde{p}_k^\dagger$. From the optimistic construction of $\widetilde{p}$ and given the choice of VI accuracy γ, we get the two following important properties.

Lemma 3. *For any round k, goal state $s^\dagger \in \mathcal{W}_k$ and state $s \in \mathcal{K}_k \cup \{x\}$, we have under the event Θ,*

$$\widetilde{u}_k^\dagger(s \to s^\dagger) \leq V_{\mathcal{K}_k}^\star(s \to s^\dagger).$$

Lemma 4. *For any round k, goal state $s^\dagger \in \mathcal{W}_k$ and state $s \in \mathcal{K}_k \cup \{x\}$, we have*

$$\widetilde{v}_k^\dagger(s \to s^\dagger) \leq (1 + 2\gamma)\widetilde{u}_k^\dagger(s \to s^\dagger).$$

B.2.4. State Transfer from $\mathcal{U}$ to $\mathcal{K}$ (step ④)

We fix any round k and any target state $s^\dagger \in \mathcal{W}_k$ that is added to the set of "controllable" states $\mathcal{K}$, i.e., for which $\widetilde{u}_k^\dagger(s_0 \to s^\dagger) \leq L$.

Lemma 5. *Under the event Θ, we have both following inequalities*

$$\begin{cases} v_k^\dagger(s_0 \to s^\dagger) \leq L + \varepsilon, \\ v_k^\dagger(s_0 \to s^\dagger) \leq V_{\mathcal{K}_k}^\star(s_0 \to s^\dagger) + \varepsilon. \end{cases}$$

In particular, the first inequality entails that $s^\dagger \in \mathcal{S}_{L+\varepsilon}^{\to}$, which justifies the validity of the state transfer from $\mathcal{U}$ to $\mathcal{K}$.

Proof. The proof comes from applying (Tarbouriech et al., 2020b, Lem. 4) on the simulation lemma for SSP, by leveraging the specific choice of allocation function ϕ in Alg. 1 which means that the prescribed sampling requirements are met by DisCo. □

B.2.5. Termination of the Algorithm

Lemma 6. *Under the event $\Theta \cap \Lambda$, for any round k, either $\mathcal{S}_L^{\to} \subseteq \mathcal{K}_k$, or there exists a state $s^\dagger \in \mathcal{S}_L^{\to} \setminus \mathcal{K}_k$ such that $s^\dagger \in \mathcal{W}_k$ and is L-reachable with a policy restricted to $\mathcal{K}_k$. Moreover, $|\mathcal{W}_k| \leq 2LA|\mathcal{K}_k|$.*

Proof of Lem. 6. Take a round k such that $\mathcal{S}_L^{\to} \setminus \mathcal{K}_k$ is non-empty. Due to the incremental construction of the set $\mathcal{S}_L^{\to}$ (Def. 2), there exists a state $s^\dagger \in \mathcal{S}_L^{\to}$ and a policy restricted to $\mathcal{K}_k$ that can reach $s^\dagger$ in L steps. Hence there exists a state-action pair $(s,a) \in \mathcal{K}_k \times \mathcal{A}$ such that $p(s^\dagger|s,a) \geq \frac{1}{L}$. Since B_k samples are available with $B_k \geq \lceil L \log\left(\frac{3ALS}{\delta}\right) \rceil$, according to Lem. 2, we get that, under the event Λ, $s^\dagger$ is found during the sample collection procedure for the state-action pair (s,a) (step ①), which implies that $s^\dagger \in \mathcal{U}_k$.

Moreover, the choice of allocation function ϕ guarantees in particular that there are more than $\Omega(\frac{4L^2}{\varepsilon^2}\log(\frac{2LSA}{\delta\varepsilon}))$ samples available at each state-action pair $(s,a) \in \mathcal{K}_k \times \mathcal{A}$. From the empirical Bernstein inequality, we thus have that $|p(s^\dagger|s,a) - \widehat{p}_k(s^\dagger|s,a)| \leq \frac{\varepsilon}{2L}$ under the event Θ. Consequently we have

$$\widehat{p}_k(s^\dagger|s,a) \geq \frac{1}{L} - |p(s^\dagger|s,a) - \widehat{p}_k(s^\dagger|s,a)| \geq \frac{1-\frac{\varepsilon}{2}}{L},$$

which implies that $s^\dagger \in \mathcal{W}_k$. Furthermore, we can decompose $\mathcal{W}_k$ the following way

$$\mathcal{W}_k = \bigcup_{(s,a)\in\mathcal{K}_k\times\mathcal{A}} \mathcal{Y}_k(s,a),$$

where we introduce the subset

$$\mathcal{Y}_k(s,a) := \left\{ s' \in \mathcal{U}_k : \widehat{p}_k(s'|s,a) \geq \frac{1-\frac{\varepsilon}{2}}{L} \right\}.$$

We then have

$$1 = \sum_{s'\in\mathcal{S}} \widehat{p}_k(s'|s,a) \geq \sum_{s'\in\mathcal{Y}_k(s,a)} \widehat{p}_k(s'|s,a) \geq \frac{1-\frac{\varepsilon}{2}}{L} |\mathcal{Y}_k(s,a)|.$$

We conclude the proof by writing that

$$|\mathcal{W}_k| \leq \sum_{(s,a)\in\mathcal{K}_k\times\mathcal{A}} |\mathcal{Y}_k(s,a)| \leq \frac{L}{1-\frac{\varepsilon}{2}} A|\mathcal{K}_k| \leq 2LA|\mathcal{K}_k|,$$

where the last inequality uses that $\varepsilon \leq 1$ (from line 1 of Alg. 1). □

Lemma 7. *Under the event $\Theta \cap \Lambda$, when either condition* STOP1 *or* STOP2 *is triggered (at a round indexed by K), we have $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}_K$.*

Proof. If condition STOP1 is triggered, Lem. 6 immediately guarantees that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}_K$ under the event Λ. If condition STOP2 is triggered, we have for all $s \in \mathcal{W}_K$, $\widetilde{u}_s(s_0 \to s) > L$. From Lem. 3 this means that, under the event Θ, for all $s \in \mathcal{W}_K$, $V^\star_{\mathcal{K}_K}(s_0 \to s) > L$. Hence none of the states in $\mathcal{W}_K$ is reachable in L steps with a policy restricted to $\mathcal{K}_K$. We conclude the proof using Lem. 6. □

Lemma 8. *Under the event $\Theta \cap \Lambda$, when* DISCO *terminates at round K, for any state $s \in \mathcal{K}_K$, the policy π_s computed during step ⑤ verifies*

$$v_{\pi_s}(s_0 \to s) \leq \min_{\pi\in\Pi(\mathcal{S}_L^{\rightarrow})} v_\pi(s_0 \to s) + \varepsilon.$$

Moreover, we have that $\mathcal{S}_L^{\rightarrow} \subseteq \mathcal{K}_K \subseteq \mathcal{S}_{L+\varepsilon}^{\rightarrow}$.

Proof. Assume that the event $\Theta \cap \Lambda$ holds. Then when the final set $\mathcal{K}_K$ is considered and the new policies are computed using all the samples, Lem. 5 yields for all $s \in \mathcal{K}_K$,

$$v_{\pi_s}(s_0 \to s) \leq \min_{\pi\in\Pi(\mathcal{K}_K)} v_\pi(s_0 \to s) + \varepsilon.$$

Moreover Lem. 7 entails that $\mathcal{K}_K \supseteq \mathcal{S}_L^{\rightarrow}$, which implies that

$$\min_{\pi\in\Pi(\mathcal{K}_K)} v_\pi(s_0 \to s) \leq \min_{\pi\in\Pi(\mathcal{S}_L^{\rightarrow})} v_\pi(s_0 \to s),$$

which means that $\mathcal{K}_K \subseteq \mathcal{S}_{L+\varepsilon}^{\rightarrow}$. □

B.2.6. High Probability Bound on the Sample Collection Phase (step ①)

Denote by K the (random) index of the last round during which the algorithm terminates. We focus on the sample collection procedure for any state $s \in \mathcal{K}_K$. We denote by k_s the index of the round during which s was added to the set of "known" states $\mathcal{K}$. To collect samples at state s, the learner uses the goal-conditioned policy π_s. We say that an attempt to collect a specific sample is a *rollout*. We denote by $Z_K := |\mathcal{K}_K| A \phi(\mathcal{K}_K)$ the total number of samples that the learner needs to collect. As such, at most Z_K rollouts must take place. Assume that the event Θ holds. Then from Lem. 8, we have $\mathcal{K}_K \subseteq \mathcal{S}^{\rightarrow}_{L+\varepsilon}$. Hence, denoting $S_{L+\varepsilon} := |\mathcal{S}^{\rightarrow}_{L+\varepsilon}|$, we have $Z_K \leq Z_{L+\varepsilon} := S_{L+\varepsilon} A \Phi(\mathcal{S}^{\rightarrow}_{L+\varepsilon})$. The following lemma provides a high-probability upper bound on the time steps required to meet the sampling requirements.

Lemma 9. *Assume that the event Θ holds. Set*

$$T := 4(L+\varepsilon+1)\log\left(\frac{6Z_{L+\varepsilon}}{\delta}\right),$$

and introduce the following event

$$\mathcal{T} := \left\{\exists \text{ one rollout (with goal state } s) \text{ s.t. } \tau_{\pi_s}(s_0 \to s) > T\right\}.$$

We have $\mathbb{P}(\mathcal{T}) \leq \frac{\delta}{3}$.

Proof. Assume that the event Θ holds. Leveraging a union bound argument and applying (Cohen et al., 2020, Lem. B.5) to policy π_s which verifies $v_{\pi_s}(s' \to s) \leq L + \varepsilon + 1$ for any $s' \in K_{k_s}$, we get

$$\mathbb{P}(\mathcal{T}) \leq \sum_{rollouts} 2\exp\left(-\frac{T}{4(L+\varepsilon+1)}\right) \leq Z_{L+\varepsilon} 2\exp\left(-\frac{T}{4(L+\varepsilon+1)}\right) \leq \frac{\delta}{3},$$

where the last inequality stems from the choice of T. □

B.2.7. Putting Everything Together: Sample Complexity Bound

The sample complexity of the algorithm is solely induced by the sample collection procedure (step ①). Recall that we denote by K the index of the round at which the algorithm terminates. With probability at least $1 - \frac{2\delta}{3}$, Lem. 7 holds, and so does the event Θ. Hence the algorithm discovers a set of states $\mathcal{K}_K \supseteq \mathcal{S}^{\rightarrow}_{L}$. Moreover, from Lem. 8, the algorithm outputs for each $s \in \mathcal{K}_K$ a policy π_s with $\mathbb{E}[\tau_{\pi_s}(s_0 \to s)] \leq V^{\star}_{\mathcal{S}^{\rightarrow}_{L}}(s) + \varepsilon$. Hence we also have $|\mathcal{K}_K| \leq S_{L+\varepsilon} := |\mathcal{S}^{\rightarrow}_{L+\varepsilon}|$.

We denote by $Z_K := |\mathcal{K}_K| A \phi(\mathcal{K}_K)$ the total number of samples that the learner needs to collect. From Lem. 9, with probability at least $1 - \frac{\delta}{3}$, the total sample complexity of the algorithm is at most TZ_K, where $T := 4(L+\varepsilon+1)\log\left(\frac{6Z_{L+\varepsilon}}{\delta}\right)$.

The total requirement is $\phi(\mathcal{K}_K)$, where ϕ is the allocation function selected by DisCo. Note that we have

$$\widehat{\Theta}(\mathcal{K}_K) \leq \Gamma_K := \max_{(s,a)\in\mathcal{K}_K\times\mathcal{A}} \|\{p(s'|s,a)\}_{s'\in\mathcal{K}_K}\|_0 \leq |\mathcal{K}_K|.$$

Combining everything yields with probability at least $1-\delta$

$$TZ_K = \widetilde{O}\left(\frac{L^5 \Gamma_K |\mathcal{K}_K| A}{\varepsilon^2} + \frac{L^3 |\mathcal{K}_K|^2 A}{\varepsilon}\right).$$

We now use that $\mathcal{K}_K \subset \mathcal{S}^{\rightarrow}_{L+\varepsilon}$ from Lem. 8, which implies that

$$\mathcal{C}_{\mathrm{AX}^\star}(\textsc{DisCo}, L, \varepsilon, \delta) = \widetilde{O}\left(\frac{L^5 \Gamma_{L+\varepsilon} S_{L+\varepsilon} A}{\varepsilon^2} + \frac{L^3 S^2_{L+\varepsilon} A}{\varepsilon}\right),$$

where $\Gamma_{L+\varepsilon} := \max_{(s,a)\in\mathcal{S}^{\rightarrow}_{L+\varepsilon}\times\mathcal{A}} \|\{p(s'|s,a)\}_{s'\in\mathcal{S}^{\rightarrow}_{L+\varepsilon}}\|_0$. This concludes the proof of Thm. 3.