Sample complexity bounds for stochastic shortest path with a generative model

Jean Tarbouriech, Matteo Pirotta, Michal Valko, Alessandro Lazaric

Sample Complexity Bounds for Stochastic Shortest Path with a Generative Model

Jean Tarbouriech JEAN.TARBOURIECH@GMAIL.COM
Facebook AI Research Paris & Inria Lille

Matteo Pirotta PIROTTA@FB.COM
Facebook AI Research Paris

Michal Valko VALKOM@DEEPMIND.COM
DeepMind Paris

Alessandro Lazaric LAZARIC@FB.COM
Facebook AI Research Paris

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

We consider the objective of computing an ε-optimal policy in a stochastic shortest path (SSP) setting, provided that we can access a generative sampling oracle. We propose two algorithms for this setting and derive PAC bounds on their sample complexity: one for the case of positive costs and the other for the case of non-negative costs under a restricted optimality criterion. While tight sample complexity bounds have been derived for the finite-horizon and discounted MDPs, the SSP problem is a strict generalization of these settings and it poses additional technical challenges due to the fact that no specific time horizon is prescribed and policies may never terminate, i.e., we are possibly facing non-proper policies. As a consequence, we can neither directly apply existing techniques minimizing sample complexity nor rely on a regret-to-PAC conversion leveraging recent regret bounds for SSP. Our analysis instead combines SSP-specific tools and variance reduction techniques to obtain the first sample complexity bounds for this setting.

Keywords: sample complexity, stochastic shortest path, Markov decision process

1. Introduction

A common assumption in approximate dynamic programming and reinforcement learning (RL) is to have access to a generative model of the Markov decision process (MDP), that is, a sampling device which can generate samples of the transition and reward functions at any state-action pair. A large body of prior work (Azar et al., 2013; Wang, 2017; Sidford et al., 2018a,b; Zanette et al., 2019; Agarwal et al., 2020; Li et al., 2020) studied how to compute an ε-optimal policy in the infinite-horizon discounted MDP (DMDP) setting with as few calls to the generative model as possible. While the infinite-horizon discounted setting is common in practical RL, many problems are better formalized within the strictly more general[1] stochastic shortest-path (SSP) setting (Bertsekas, 1995), where the objective is to compute a policy that minimizes the cost accumulated before reaching a specific goal state. Recently, Tarbouriech et al. (2020a) and Rosenberg et al. (2020) studied the SSP problem in the online case and derived the first regret bounds for this setting.

[1] Any DMDP with discount factor γ can indeed be converted into an SSP problem with the same state space augmented by an artificial termination state that is reached with probability $1-\gamma$ at any time step and state-action pair.

In this paper we focus on the generative model setting and study the problem of computing a near-optimal policy in an SSP problem with a given goal state and cost function. We derive two closely related algorithms for this setting and we prove PAC bounds for their sample complexity. The first algorithm is designed to return an ε-optimal policy for any SSP problem with strictly positive cost function and it has a sample complexity that is adaptive to the (unknown) range of the optimal value function. The second can be used for any cost function, including the case when the cost is zero in some states, for which the optimal policy may not even be proper (i.e., it may never reach the goal yet still minimize the cumulative cost). In this case, we re-frame the objective as computing an ε-optimal solution in the set of (proper) policies that in expectation reach the goal in at most a number of steps that is proportional to the minimum number of expected steps to the goal (i.e., the SSP-diameter of the problem). The main technical challenge in deriving these results is due to the fact that in SSP we have no knowledge of the *effective* horizon of the problem, unlike in DMDP (with $1/(1-\gamma)$). As a result, model estimation errors may accumulate indefinitely, thus preventing from achieving any desired level of accuracy. In order to deal with this problem, a first approach is to build on an SSP-specific simulation lemma, which reveals the level of accuracy needed in estimating the MDP to be able to recover proper policies and the role played by the minimum cost. Although this approach yields a bound on the sample complexity, we show that it can in fact be tightened by leveraging and combining SSP-specific tools for regret minimization (Rosenberg et al., 2020) and variance-aware techniques for DMDP sample complexity (Azar et al., 2013).

2. Preliminaries

Stochastic shortest path (SSP) We start by introducing the notion of MDP with an SSP objective as done by Bertsekas (1995, Sect. 3).

Definition 1 (SSP-MDP) *An SSP-MDP is an MDP*

$$M := \langle \mathcal{S}, \mathcal{A}, g, p, c \rangle,$$

where $\mathcal{S}$ is the state space with $S := |\mathcal{S}|$ states and $\mathcal{A}$ is the action space with $A := |\mathcal{A}|$ actions. We denote by $g \notin \mathcal{S}$ the goal state, and we set $\mathcal{S}' := \mathcal{S} \cup \{g\}$. Taking action a in state s incurs a cost of $c(s,a) \in [0,1]$ and the next state $s' \in \mathcal{S}'$ is selected with probability $p(s'|s,a)$. The goal state g is absorbing and zero-cost, i.e., $p(g|g,a) = 1$ and $c(g,a) = 0$ for any action $a \in \mathcal{A}$, which effectively implies that the agent ends its interaction with M when reaching the goal g.

We denote by $\Pi := \{\pi : \mathcal{S} \to \mathcal{A}\}$ the set of stationary deterministic policies. For any $\pi \in \Pi$ and $s \in \mathcal{S}$, the random (possibly unbounded) goal-reaching time starting from s is denoted by $\tau_\pi(s) := \inf\{t \geq 0 : s_{t+1} = g \,|\, s_1 = s, \pi\}$.

Definition 2 (Proper policy) *A policy π is proper if its execution reaches the goal with probability 1 when starting from any state in $\mathcal{S}$. A policy is improper if it is not proper. The set of proper policies is denoted by Π_{proper}.*

Assumption 1 *There exists at least one proper policy, i.e., $\Pi_{\text{proper}} \neq \emptyset$.*

The value function (also called expected cost-to-go) of a policy $\pi \in \Pi$ is defined as

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=1}^{+\infty} c(s_t, \pi(s_t)) \,\middle|\, s_1 = s\right] = \mathbb{E}\left[\sum_{t=1}^{\tau_\pi(s)} c(s_t, \pi(s_t)) \,\middle|\, s_1 = s\right],$$

where the expectation is w.r.t. the random sequence of states generated by executing π starting from state $s \in \mathcal{S}$. Note that for improper policies $\pi \notin \Pi_{\text{proper}}$, V^π has at least one unbounded component. Our objective is to find an optimal policy $\pi^\star$ that minimizes the value function. For any vector $V \in \mathbb{R}^S$, the optimal Bellman operator is defined as

$$\mathcal{L}V(s) := \min_{a \in \mathcal{A}} \Big\{ c(s,a) + \sum_{y \in \mathcal{S}} p(y \,|\, s,a) V(y) \Big\}. \tag{1}$$

Lemma 1 (Bertsekas and Tsitsiklis, 1991, Prop. 2) *Suppose that Asm. 1 holds and that for every improper policy π' there exists at least one state $s \in \mathcal{S}$ such that $V^{\pi'}(s) = +\infty$. Then the optimal policy $\pi^\star$ is stationary, deterministic, and proper. Moreover, $V^\star = V^{\pi^\star}$ is the unique solution of the optimality equations $V^\star = \mathcal{L}V^\star$ and $V^\star(s) < +\infty$ for any $s \in \mathcal{S}$.*

We define the following quantities:

- The *SSP-diameter* D (Tarbouriech et al., 2020a) is defined as

$$D := \max_{s \in \mathcal{S}} D_s, \quad \text{with} \;\; D_s := \min_{\pi \in \Pi} \mathbb{E}[\tau_\pi(s)]. \tag{2}$$

- $B_\star := \max_{s \in \mathcal{S}} \min_{\pi \in \Pi_{\text{proper}}} V^\pi(s)$ is the maximal optimal value function over states.
- $\Gamma := \max_{s,a} \|p(\cdot|s,a)\|_0 \leq S + 1$ is the maximal support of $p(\cdot|s,a)$.
- Finally, $c_{\min} := \min_{(s,a) \in \mathcal{S} \times \mathcal{A}} c(s,a) \in [0,1]$ is the minimum non-goal cost.

Note that since the number of states S is finite, Asm. 1 implies that $B_\star \leq D < +\infty$.

Problem formulation We consider that the costs c are deterministic and known to the learner, while the transition dynamics p is unknown. We assume access to a generative model, which for any state-action pair (s,a) returns a sample drawn from $p(\cdot|s,a)$. We ask the following: *How many calls to the generative model are sufficient to compute a near-optimal policy with high probability?*

On the online-to-batch conversion in the SSP setting Since the problem of learning in SSP has already been studied in the regret-minimization setting (Tarbouriech et al., 2020a; Rosenberg et al., 2020), it may be tempting to leverage a regret-to-PAC conversion to obtain a sample complexity bound and provide a first answer to the question above. For instance, in finite-horizon MDPs, a regret bound can be converted to a PAC guarantee by selecting as a candidate optimal solution any policy chosen at random out of all episodes (Jin et al., 2018). Unfortunately this procedure cannot be applied here. In fact, the SSP-regret differs from the finite-horizon regret, since at each episode it compares the *empirical* costs accumulated along one trajectory with the optimal value function. Indeed, following K episodes with initial state s_0 where for each $k \in [K]$ the (possibly non-stationary) policy executed is denoted by π_k, we recall that the SSP-regret is defined as

$$\Big[\sum_{k=1}^{K} \sum_{h=1}^{\tau_{\pi_k}(s_0)} c(s_{k,h}, \pi_k(s_{k,h})) \Big] - K \cdot \min_{\pi \in \Pi_{\text{proper}}} V^\pi(s_0),$$

where $s_{k,h}$ denotes the h-th state visited during episode k and $\tau_{\pi_k}(s_0)$ is the (possibly infinite) time it takes the agent to complete episode k. As a result, no guarantee is provided for the value

function (i.e., the expected cumulative costs) of one episode. Indeed, it has been shown in the existing regret analyses for SSP that explicitly guaranteeing the properness of the deployed policies is *not* an intermediate step that is required to derive the regret bounds. SSP-regret algorithms may change multiple policies within each episode and none of them may actually be proper (i.e., they may have an unbounded value function, so that there exists a state s such that $V^\pi(s) = +\infty$). As such, it is unclear which policy should be retained as a solution candidate. Finally, the near-optimal guarantees we intend to achieve are for any arbitrary initial state in $\mathcal{S}$, while regret-style guarantees are only in expectation with respect to a starting state distribution.

Distinction of cases depending on $c_{\min}$ Our analysis considers two distinct cases, $c_{\min} > 0$ and $c_{\min} = 0$, for which it targets different notions of sample complexity.

First case: sample complexity objective for $c_{\min} > 0$ When the cost function is strictly positive, the conditions of Lem. 1 hold, so the optimal policy is guaranteed to be proper. We seek to achieve the following standard PAC guarantees.

Definition 3 *We say that an algorithm is (ε, δ)-optimal with sample complexity n, if after n calls to the generative model it returns a policy π that verifies $\|V^\pi - V^\star\|_\infty \leq \varepsilon$ with probability at least $1 - \delta$. We denote by* SAMPCOMP(ε) *the corresponding sample complexity n.*

Second case: sample complexity objective for $c_{\min} = 0$ The case of zero minimum cost is a complex SSP problem where the optimal policy may not even be guaranteed to be proper (Bertsekas, 1995). In this case, it is unclear whether the sample complexity of Def. 3 may even be bounded, since estimation errors may propagate indefinitely. On the other hand, we seek ε-optimality guarantees w.r.t. a set of proper policies.

Definition 4 (Restricted set Π_θ) *For any $\theta \in [1, +\infty]$, we define the set*

$$\Pi_\theta := \{\pi \in \Pi : \forall s \in \mathcal{S}, \mathbb{E}[\tau_\pi(s)] \leq \theta D_s\}.$$

Notice that $\Pi_{+\infty} = \Pi$. Moreover, for any $\theta \in [1, +\infty)$, Π_θ only contains proper policies, i.e., $\Pi_\theta \subseteq \Pi_{\text{proper}}$. Similar to Def. 3, we then reformulate the desired notion of optimality.

Definition 5 *We say that an algorithm is $(\varepsilon, \delta, \theta)$-optimal with sample complexity n, if after n calls to the generative model it returns a policy π that verifies $\|V^\pi - V^\star_\theta\|_\infty \leq \varepsilon$ with probability at least $1 - \delta$, where $V^\star_\theta = \min_{\pi \in \Pi_\theta} V^\pi$ is the optimal value function restricted to policies in Π_θ. We denote by* SAMPCOMP(ε, θ) *the corresponding sample complexity n.*

While alternative definitions of restricted set may be introduced, we believe that Def. 4 is well-suited for our problem, as it defines the restriction w.r.t. D_s, a cost-independent quantity describing the difficulty of navigating in the SSP-MDP (Eq. 2). Nonetheless, this poses an additional layer of complexity, since D_s is unknown to the agent, which only receives θ as additional parameter.

3. A first approach: Simulation Lemma for SSP

We begin by stating a general simulation lemma tailored to SSP which is a useful component to derive sample complexity bounds. For any model p and any $\eta > 0$, we introduce the set of models close to p as follows

$$\mathcal{P}^{(p)}_\eta := \left\{p' \in \mathbb{R}^{S' \times A \times S'} : \forall (s,a) \in \mathcal{S}' \times \mathcal{A},\ p'(\cdot|s,a) \in \Delta(\mathcal{S}'),\ \|p(\cdot|s,a) - p'(\cdot|s,a)\|_1 \leq \eta\right\}.$$

Up to a slight difference in the way the set $\mathcal{P}_\eta^{(p)}$ is defined, leveraging the result of Tarbouriech et al. (2020b, App. C) yields the following guarantee (see App. C).

Lemma 2 (Simulation Lemma for SSP) *Consider any $\eta > 0$ and any two models p and $p' \in \mathcal{P}_\eta^{(p)}$ such that, for each model, there exists at least one proper policy w.r.t. the goal state g. Consider a cost function such that $c_{\min} > 0$. Consider any policy π that is proper in p', with value function denoted by V'_π, such that the following condition is verified*

$$\eta \|V'_\pi\|_\infty \leq 2c_{\min}. \tag{3}$$

Then π is proper in p (i.e., its value function verifies $V_\pi < +\infty$ component-wise), and we have

$$\forall s \in \mathcal{S},\ V_\pi(s) \leq \left(1 + \frac{2\eta\|V'_\pi\|_\infty}{c_{\min}}\right) V'_\pi(s),$$

and conversely,

$$\forall s \in \mathcal{S},\ V'_\pi(s) \leq \left(1 + \frac{\eta\|V'_\pi\|_\infty}{c_{\min}}\right) V_\pi(s).$$

Combining the two inequalities above yields

$$\|V_\pi - V'_\pi\|_\infty \leq \frac{7\eta\|V'_\pi\|_\infty^2}{c_{\min}}.$$

For comparison let us now recall the classical simulation lemma for discounted MDPs.

Lemma 3 (Simulation Lemma for DMDP, see e.g., Kearns and Singh, 2002) *Consider any two models p and $p' \in \mathcal{P}_\eta^{(p)}$ for any $\eta > 0$. Consider as value function in p the expected discounted cumulative reward, i.e., for any policy π and state $s \in \mathcal{S}$, $V_\pi(s) := \mathbb{E}\left[\sum_{t=1}^{+\infty} \gamma^t r(s_t, \pi(s_t)) \mid s_1 = s\right]$; and V'_π is the value function in p'. Suppose that the instantaneous rewards are known and bounded in $[0, 1]$. Then for any policy π, we have*

$$\|V_\pi - V'_\pi\|_\infty \leq O\left(\frac{\eta}{(1-\gamma)^2}\right).$$

We spell out the key differences between the simulation lemma in the discounted setting (Lem. 3) and in SSP (Lem. 2), bringing to light the criticalities in the latter setting. First, a guarantee can only be obtained if the condition (3) is verified, which involves both the accuracy η and the value function of π in $p' \in \mathcal{P}_\eta^{(p)}$. We observe that the smaller the minimum cost $c_{\min}$, the smaller the accuracy η needs to be. Importantly, $c_{\min}$ must be positive and the error scales inversely with it. Indeed, the "trajectory length" is captured not by a known hyperparameter $1/(1-\gamma)$ as in DMDPs, but by the ratio between the (a priori unknown) infinity norm of the value function of the policy and the minimum cost $c_{\min}$ (note that this ratio indeed has a time dimension and it upper bounds the expected goal-reaching time of the policy since $\|\mathbb{E}[\tau_\pi]\|_\infty \leq \|V_\pi\|_\infty / c_{\min}$).

Similar to existing approaches in DMDP, we could leverage the simulation lemma to directly derive sample complexity guarantees. More precisely, building on the result of Lem. 2 and plugging

it in the algorithms proposed in Sect. 4 would eventually lead to sample complexities scaling as

$$\textsc{SampComp}(\varepsilon) = \widetilde{O}\left(\frac{B_\star^4 \Gamma S A}{c_{\min}^2 \varepsilon^2} + \frac{B_\star^2 S^2 A}{c_{\min} \varepsilon}\right), \tag{4}$$

$$\textsc{SampComp}(\varepsilon, \theta) = \widetilde{O}\left(\frac{\theta^2 D^2 B_\star^4 \Gamma S A}{\varepsilon^4} + \frac{\theta D B_\star^2 S^2 A}{\varepsilon^2}\right). \tag{5}$$

In the following section, we will show that these two bounds can in fact be improved by refining the guarantee of Lem. 2 thanks to variance-aware arguments (as also leveraged by e.g., Azar et al., 2013; Rosenberg et al., 2020). Notably, it will enable to shave off a $B^\star/c_{\min}$ dependency in the $\widetilde{O}(\varepsilon^{-2})$ main-order term of Eq. 4 for the first case of $c_{\min} > 0$ (see Thm. 1). Furthermore, in the case of $c_{\min} = 0$, the dependency on ε in Eq. 5 will be reduced from $\widetilde{O}(\varepsilon^{-4})$ to $\widetilde{O}(\varepsilon^{-3})$ (see Thm. 2).

4. Main Result

We first illustrate the common structure of our algorithms. As an input, we receive a desired accuracy $\varepsilon \in (0, 1)$, the confidence level $\delta \in (0, 1]$, and the cost function $c \in [0, 1]$. Since the model p is unknown, akin to Azar et al. (2013) for DMDPs, we collect transition samples from the generative model and we use them to compute an estimate $\widehat{p}$ by simply evaluating the frequency of transitions from each state-action pair s, a to any state s'. In particular, we rely on a carefully tuned function to determine the number of transition samples that should be collected for every state-action pair. For some positive values of X and y, we introduce the allocation function[2]

$$\phi(X, y) := \alpha \cdot \left(\frac{X^3 \widehat{\Gamma}}{y\varepsilon^2} \log\left(\frac{XSA}{y\varepsilon\delta}\right) + \frac{X^2 S}{y\varepsilon} \log\left(\frac{XSA}{y\varepsilon\delta}\right) + \frac{X^2 \widehat{\Gamma}}{y^2} \log^2\left(\frac{XSA}{y\delta}\right)\right), \tag{6}$$

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 $\widehat{p}$.

Let us now consider that the conditions of Lem. 1 hold. A standard approach would then be to execute SSP-value iteration (VI) on the estimated SSP-MDP $\widehat{M} = \langle \mathcal{S}, \mathcal{A}, g, \widehat{p}, c\rangle$ and return the corresponding optimal policy $\widehat{\pi}$. While this approach is effective in DMDPs and finite-horizon problems, it may fail in the SSP setting. In fact, the estimated SSP-MDP may not even admit a proper optimal policy and deriving guarantees on the actual value of $\widehat{\pi}$ (i.e., $V^{\widehat{\pi}}$) may not be possible. As such, instead of solving the estimated SSP-MDP $\widehat{M}$, we rather execute an extended value iteration (EVI) scheme tailored to SSP problems, which can be run efficiently as shown by Tarbouriech et al. (2020a). As detailed in App. B, EVI for SSP constructs confidence intervals for $\widehat{p}$ and builds a suitable SSP-MDP $\widetilde{M} = \langle \mathcal{S}, \mathcal{A}, g, \widetilde{p}, c\rangle$, where $\widetilde{p}$ belongs to the confidence intervals and is chosen so that the corresponding optimal policy $\widetilde{\pi}$ is optimistic w.r.t. to the optimal policy $\pi^\star$ of M. More formally, let us now consider that our SSP-MDP at hand has a strictly positive cost function c (which entails that the conditions of Lem. 1 hold), and consider a set $\mathcal{N}$ of samples collected so far as well as a VI precision level $\mu_{\text{VI}} > 0$. Then $\text{EVI}(\mathcal{N}, c, \mu_{\text{VI}})$ outputs an optimistic value vector $\widetilde{v}$ and an optimistic policy $\widetilde{\pi}$ that is greedy w.r.t. $\widetilde{v}$. Note that here (as opposed to Tarbouriech et al., 2020a), we consider Bernstein-based concentration inequalities for EVI, as it is done by Rosenberg et al. (2020) as well as in average-reward EVI by Fruit et al. (2020). The crucial

[2] The actual choice of X and y is algorithm-specific and it is illustrated later.

Algorithm 1: Algorithm for $c_{\min} > 0$

Input: cost function c with minimum cost $c_{\min} > 0$, accuracy $\varepsilon > 0$, confidence level $\delta \in (0,1)$, allocation function $\phi(\cdot,\cdot)$.

$\widetilde{\pi} :=$ `SEARCH`(c).

Output: the policy $\widetilde{\pi}$.

Algorithm 2: `SEARCH`

Input: A positive cost function c'.

Set $\iota := \min_{s,a} c'(s,a)$ the minimum cost of c'.

Set $\Delta := \frac{1}{2}$ and `continue = True`, sample set $\mathcal{N} = \emptyset$.

while `continue` **do**

  Set $\Delta \leftarrow 2\Delta$.

  Add samples obtained from the generative model to $\mathcal{N}$ until $\phi(\Delta, \iota)$ samples are available at each state-action pair.

  Compute $(\widetilde{v}, \widetilde{\pi}) := \text{EVI}\big(\mathcal{N}, c', \mu_{\text{VI}} := \frac{\iota\varepsilon}{6\Delta}\big)$ with $\mathcal{N}$ the samples collected so far.

  if $\|\widetilde{v}\|_\infty \leq \Delta$ **then**

    `continue = False`.

  end

end

Output: the policy $\widetilde{\pi}$.

advantage of EVI w.r.t. VI run on the estimated SSP-MDP is that $\widetilde{\pi}$ is proper in $\widetilde{M}$. Indeed its value function in the optimistic model $\widetilde{p}$, denoted by $\widetilde{V}^{\widetilde{\pi}}$, is bounded with high-probability. This is shown by the following lemma (which stems from Tarbouriech et al., 2020a, Lem. 4 & App. E), where we denote by $V^\star$ (resp. $\widetilde{V}^\star$) the optimal value function in the true model p (resp. optimistic model $\widetilde{p}$).

Lemma 4 *For any cost function $c \geq c_{\min} > 0$, let $(\widetilde{v}, \widetilde{\pi}) = \text{EVI}(\mathcal{N}, c, \mu_{\text{VI}})$. Then with high probability, we have the component-wise inequalities $\widetilde{v} \leq V^\star$, $\widetilde{v} \leq \widetilde{V}^\star \leq \widetilde{V}^{\widetilde{\pi}}$, and if the* VI *precision level verifies $\mu_{\text{VI}} \leq \frac{c_{\min}}{2}$, then $\widetilde{V}^{\widetilde{\pi}} \leq \Big(1 + \frac{2\mu_{\text{VI}}}{c_{\min}}\Big)\widetilde{v}$.*

We are now ready to detail our algorithms.

4.1. Strictly Positive Cost Function

When $c_{\min} > 0$, we seek to achieve the standard PAC guarantees of Def. 3. The algorithm is reported in Alg. 1. Since no prior knowledge about the optimal policy is available, the algorithm's subroutine `SEARCH` (Alg. 2) relies on a doubling scheme to guess the range of the optimal value function $B_\star$. Starting with $\Delta = 1$, we use the allocation function ϕ to determine a sufficient number of samples to compute an ε-optimal policy *if* the range of the optimal policy was smaller than Δ. We then test whether Δ is indeed a valid upper bound on the range of the optimistic value returned by EVI and, relying on Lem. 4, we stop whenever the test is successful and return $\widetilde{\pi}$. Otherwise, we double the guess Δ and reiterate. Since ϕ is increasing in its first argument, the total number of samples required at iteration is also increasing.

Theorem 1 *For any accuracy $\varepsilon \in (0,1]$, confidence $\delta \in (0,1)$, and cost function c in $[c_{\min}, 1]$, with $c_{\min} > 0$, Algorithm 1 (with the allocation function of Eq. 6) is (ε, δ)-optimal with a sample complexity bounded as follows*

$$\textsc{SampComp}(\varepsilon) = \widetilde{O}\left(\frac{B_\star^3 \Gamma S A}{c_{\min}\varepsilon^2} + \frac{B_\star^2 S^2 A}{c_{\min}\varepsilon} + \frac{B_\star^2 \Gamma S A}{c_{\min}^2}\right).$$

Proof sketch Throughout we assume that the standard high-probability event of satisfied concentration inequalities holds (Lem. 5). The analysis starts by showing that enough samples per state-action pair are available to guarantee that the candidate optimistic policy $\widetilde{\pi}$ is proper not only in the optimistic model $\widetilde{p}$ but also in the true model p. This is proved by applying the simulation lemma for SSP of Lem. 2. Hence, the value functions in the true and optimistic models, denoted by $V^{\widetilde{\pi}}$ and $\widetilde{V}^{\widetilde{\pi}}$ respectively, are each bounded component-wise and close enough up to a multiplicative constant. Moreover, by optimism and choice of VI accuracy μ_{VI}, we obtain that $\widetilde{V}^{\widetilde{\pi}} \leq V^\star + O(\varepsilon)$ component-wise. In addition we can prove that the doubling scheme of Alg. 1 guarantees that $\Delta \leq 2B_\star$. Putting everything together implies two important properties: (i) it holds that $\|V^{\widetilde{\pi}}\|_\infty = O(B_\star)$, and (ii) it is sufficient to bound $V^{\widetilde{\pi}} - \widetilde{V}^{\widetilde{\pi}}$ in order to obtain the sought-after guarantee of Def. 3. To do so, subtracting the two respective Bellman equations yields

$$V^{\widetilde{\pi}}(s) - \widetilde{V}^{\widetilde{\pi}}(s) = \sum_{y \in \mathcal{S}} p(y|s, \widetilde{\pi}(s))(V^{\widetilde{\pi}}(y) - \widetilde{V}^{\widetilde{\pi}}(y)) + W(s),$$

where we introduce

$$W(s) := \sum_{y \in \mathcal{S}} (p(y|s, \widetilde{\pi}(s)) - \widetilde{p}(y|s, \widetilde{\pi}(s)))\widetilde{V}^{\widetilde{\pi}}(y).$$

Let us denote by $Q^{\widetilde{\pi}} \in \mathbb{R}^{S \times S}$ the transition matrix between the non-goal states under policy $\widetilde{\pi}$ in the true model p (i.e., for any $(s, s') \in \mathcal{S}$, $Q^{\widetilde{\pi}}(s, s') := p(s'|s, \widetilde{\pi}(s))$). Since $\widetilde{\pi}$ is proper in p, $Q^{\widetilde{\pi}}$ is strictly substochastic which implies that the matrix $(I - Q^{\widetilde{\pi}})$ is invertible, and therefore we have

$$V^{\widetilde{\pi}}(s) - \widetilde{V}^{\widetilde{\pi}}(s) = \left[(I - Q^{\widetilde{\pi}})^{-1} W\right]_s = \sum_{t=0}^{+\infty} \mathbb{E}_{\widetilde{\pi}, p}\left[\mathbb{1}_{s_t \neq g} W(s_t) \,|s_0 = s\right]. \tag{7}$$

We now apply variance-aware arguments, similar to e.g., Azar et al. (2013, 2017); Rosenberg et al. (2020); Fruit et al. (2020), in order to decompose $W(s_t)$ and thus obtain

$$V^{\widetilde{\pi}}(s) - \widetilde{V}^{\widetilde{\pi}}(s) \leq ❶ + ❷ + ❸,$$

with

$$❶ = \widetilde{O}\left(\sqrt{\frac{\widehat{\Gamma}}{n}} \sum_{t=0}^{+\infty} \mathbb{E}_{\widetilde{\pi}, p}\left[\mathbb{1}_{s_t \neq g} \sqrt{\mathbb{V}(s_t)}\right]\right),$$

$$❷ = \widetilde{O}\left(\sqrt{\frac{\widehat{\Gamma}}{n}} \sqrt{c_{\min}\Delta} \sum_{t=0}^{+\infty} \mathbb{P}_{\widetilde{\pi}, p}(s_t \neq g)\right),$$

$$❸ = \widetilde{O}\left(\frac{\Delta S}{n} \sum_{t=0}^{+\infty} \mathbb{P}_{\widetilde{\pi}, p}(s_t \neq g)\right),$$

where n denotes the minimum number of samples collected at each state-action pair and where we define the empirical branching factor $\widehat{\Gamma} := \max_{s,a} \|\widehat{p}(\cdot|s,a)\|_0 \leq \Gamma$ as well as the following variance quantity

$$\mathbb{V}(s_t) := \sum_{s' \in \mathcal{S}'} p(s'|s_t, \widetilde{\pi}(s_t)) \left(\widetilde{V}^{\widetilde{\pi}}(s') - \sum_{s'' \in \mathcal{S}} p(s''|s_t, , \widetilde{\pi}(s_t)) \widetilde{V}^{\widetilde{\pi}}(s'') \right)^2 .$$

The series that appears in the terms ❷ and ❸ is bounded by leveraging the exponential decay of the probability of not reaching the goal w.r.t. the time step t, i.e.,

$$\mathbb{P}_{\widetilde{\pi},p}(s_t \neq g) = O\left(\exp\left(-\frac{c_{\min} t}{\|V^{\widetilde{\pi}}\|_\infty} \right) \right). \tag{8}$$

Furthermore, the series appearing in term ❶ is split by the Cauchy-Schwarz inequality and by the time step decomposition of Rosenberg et al. (2020) into *intervals* that are carefully constructed so that the expected variance $\mathbb{V}$ accumulated over a whole interval is adequately bounded. Ultimately, there remains to invert the equation ❶ + ❷ + ❸ $\leq \varepsilon$ w.r.t. n in order to obtain a lower bound on n.

High-level connection to DMDP analysis Here we point out a high-level parallel with the DMDP analysis of Azar et al. (2013). Recall informally that the latter analysis handles sums of the sort $\left[(I - \gamma P^{\widehat{\pi}})^{-1} W\right]_s = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}_{\widehat{\pi},p}\left[W(s_t) \mid s_0 = s\right]$, for different yet related quantities W. In contrast, the SSP setting does not display a natural discount factor in the Bellman equations ($\gamma = 1$ in Eq. 1). Instead, the "shrinking" of summands in Eq. 7 is captured by the indicator $\mathbb{1}_{s_t \neq g}$, which is an *unknown, time-dependent, state-dependent and policy-dependent* quantity. Crucially, we obtain that its *expectation* displays an exponential decay similar to the γ^t-phenomenom observed in DMDPs, specifically Eq. 8. As such, we can argue that any proper policy π in SSP displays a pseudo-discounting property with rate $\gamma_\pi \sim \exp\left(-c_{\min}/\|V^\pi\|_\infty\right) < 1$. Despite this interesting connection with the DMDP analysis, note that one cannot simply plug in the DMDP analysis for the SSP setting considered here (recall that DMDPs are a subclass of SSP-MDPs, not the other way around). In fact, we need to consider SSP-specific analytical tools to handle the upper bounding of Eq. 7 (e.g., interval decomposition) as alluded to above and detailed in App. D.

4.2. Any Cost Function and Restricted Optimality

Whenever $c_{\min} = 0$, Alg. 1 has a possibly unbounded sample complexity. To handle this, we add a small perturbation to all the costs (denoted by ν) during the computation of the optimistic policies. Note that this perturbation technique is also employed by e.g., Bertsekas and Yu (2013); Tarbouriech et al. (2020a); Rosenberg et al. (2020). Executing Alg. 1 with the modified cost function would directly return a policy that is ε-optimal w.r.t. the optimal policy of the SSP-MDP with perturbed cost. Nonetheless, this is not a significant guarantee, since it does not say anything about the performance of the policy in the original SSP-MDP. For this reason, we rather derive ε-optimality guarantees w.r.t. a set of restricted policies, as discussed in Sect. 2. Note that Def. 5 involves the unknown quantities $\{D_s\}_{s \in \mathcal{S}}$. In fact, in order to properly tune the cost perturbation ν and return an ε-optimal policy, we need to compute an upper bound $\widehat{D}$ the SSP-diameter. This requires an additional initial phase to perform such estimation step. We explain the procedure in App. C.1 (Alg. 4) and show that the amount of samples used in this initial phase is subsumed in the final sample complexity bound by the second phase where we compute the final candidate policy.

Algorithm 3: Algorithm for $\theta < +\infty$

Input: slack parameter $\theta \geq 1$, accuracy $\varepsilon > 0$, confidence level $\delta \in (0,1)$, cost function c, allocation function $\phi(\cdot,\cdot)$.

First compute $\widehat{D}$ an upper bound estimate of the SSP-diameter (see Alg. 4 of App. C.1).

Set cost perturbation $\nu = \frac{\varepsilon}{2\theta\widehat{D}}$.

$\widetilde{\pi} = \texttt{SEARCH}(c \vee \nu)$.

Output: the policy $\widetilde{\pi}$.

The second phase is basically the same as in Alg. 1 except for the perturbation on the original cost function by ν and a slightly different precision level μ_{VI}. The analysis simply applies Thm. 1 in the *perturbed model* and leverages the optimality restriction of Def. 4 to control the bias induced by the cost perturbation (see App. E). We obtain the following sample complexity guarantees.

Theorem 2 *For any accuracy $\varepsilon \in (0,1]$, confidence $\delta \in (0,1)$, cost function c in $[0,1]$, and slack parameter $\theta \geq 1$, Algorithm 3 (with the allocation function of Eq. 6) is $(\varepsilon,\delta,\theta)$-optimal with a sample complexity bounded as follows*

$$\textsc{SampComp}(\varepsilon,\theta) = \widetilde{O}\left(\frac{\theta D B_\star^3 \Gamma S A}{\varepsilon^3} + \frac{\theta D B_\star S^2 A}{\varepsilon^2} + \frac{\theta^2 D^2 B_\star^2 \Gamma S A}{\varepsilon^2}\right).$$

In Thm. 2 the slack parameter $\theta \geq 1$ is implicitly considered as a bounded constant which implies that $\Pi_\theta \subsetneq \Pi$. It is possible to link θ to the accuracy ε, by for example instantiating $\theta = \varepsilon^{-1}$. In this special case, at the cost of a worse dependency on ε for Thm. 2 (namely, in $\widetilde{O}(\varepsilon^{-4})$), we obtain an ε-accurate guarantee w.r.t. the optimal proper policy as ε tends to 0, since $\lim_{\varepsilon\to 0} \Pi_{\varepsilon^{-1}} = \Pi$.

5. Discussion

In this section we discuss the bounds obtained in Thm. 1 and 2 and compare them with existing bounds in related settings.

First, let us consider the unit-cost case ($c_{\min} = 1$). In this case, it is easy to show that the sample complexity is lower-bounded as $\Omega(SA(B_\star)^3/\varepsilon^2)$. In fact, the inclusion DMDP $\subset$ SSP-MDP, and the mapping $1/(1-\gamma) = B_\star$ in the unit-cost case (see footnote[1]), allows us to directly inherit the lower bound of Azar et al. (2013) in DMDPs, which scales as $\Omega(SA/(1-\gamma)^3\varepsilon^2)$. This shows that in this case, the sample complexity in Thm. 1 matches the lower bound in the ε, A and $B_\star$ terms.

As for the dependency on the state space, our bound scales as ΓS with $\Gamma \in [1, S+1]$ the maximal branching factor. While in many environments $\Gamma = O(1)$ as long as the dynamics are not too chaotic, ΓS may scale with S^2 in the worst case. This possibly quadratic dependency in S is worse than the linear dependency for sample complexity in DMDPs with a generative model (Azar et al., 2013). This bound mismatch is also present between SSP-MDP and finite-horizon in the regret minimization framework, where no-regret algorithms for SSP (Tarbouriech et al., 2020a; Rosenberg et al., 2020) scale as $\widetilde{O}(S)$, which contrasts with the lower bound in $\sqrt{S}$ derived by Rosenberg et al. (2020) and with regret bounds in finite-horizon (e.g., Azar et al., 2017) which match the $\sqrt{S}$ lower bound. How to improve the state dependency for the SSP setting remains an open question, whether it be in the regret minimization or sample complexity setting. Note that

the ΓS dependency stems from the analysis estimating the transition kernel accurately well across the state-action space. As an immediate byproduct, this implies that after its sample collection phase, each algorithm Alg. 1 or 3 can actually guarantee ε-optimal planning for *any* cost function in $[c_{\min}, 1]$, respectively where $c_{\min} > 0$ (Thm. 1) and where $c_{\min} = 0$ with $\theta < +\infty$ (Thm. 2).

The role of the effective horizon H in finite-horizon or $1/(1-\gamma)$ in DMDPs is captured in the SSP setting by the ratio $B_\star/c_{\min}$ (when $c_{\min} > 0$). Compared to the application of the simulation lemma for SSP (Lem. 2), the use of variance-aware techniques succeeds in shaving off a term $B_\star/c_{\min}$ in the main order term of the sample complexity. Our analysis combines techniques on regret minimization for the SSP problem (Rosenberg et al., 2020) and on sample complexity of DMDPs with a generative model (Azar et al., 2013). The latter work removes a factor $1/(1-\gamma)$, with $\gamma < 1$ the discount factor. As we fleshed out in our analysis, the role of the discount factor γ in DMDPs is implicitly captured in SSP by the policy-dependent indicator $\mathbb{1}_{\{s_t \neq g\}}$, whose expectation decays exponentially w.r.t. the time t with rate scaling as $\exp(-c_{\min}/B_\star) < 1$. Recall that this high-level parallel between the SSP-MDP and DMDP settings is not surprising insofar as, more generally, DMDPs are a subclass of SSP-MDPs (Bertsekas, 1995).

We can wonder whether the $c_{\min}$ dependency is unavoidable or not in the sample complexity result of Thm. 1 for the case $c_{\min} > 0$. While we do not have a definite answer, we can investigate the question by drawing a high-level analogy with the finite-horizon setting. Taking the regret of UCBVI (Azar et al., 2017) in the stationary case $\sum_k (V^{\pi_k} - V^\star) \leq \sqrt{H^2SAK}$, and performing a regret-to-PAC conversion, we notice that to obtain $V^\pi - V^\star \leq \varepsilon$, we require $K' \approx H^2SA/\varepsilon^2$ episodes and hence HK' time steps of sample complexity, since each episode accounts for H interactions with the environment. We thus see that the dependency in $H^3 = H^2H$ can be decomposed as H^2 capturing the range of the value function, and another H capturing the length of an episode. Now by analogy, pretending such a regret-to-PAC conversion works in SSP (which we recall from Sect. 2 is not the case), we would obtain the dependency $B_\star^2(B_\star/c_{\min})$, because the range of the optimal value function is $B_\star$ while the characteristic length of an optimal episode scales as $B_\star/c_{\min}$ in the worst case. This reasoning provides an intuitive support to our conjecture that the lower bound must contain a dependency on the characteristic length of an optimal episode which in the worst case scales as $B_\star/c_{\min}$. It remains an open question whether it is possible or not to construct a lower bound problem explicitly displaying the critical role of $c_{\min}$. Finally, the bound of Thm. 1 inherits a $\widetilde{O}(\varepsilon^{-2})$ dependency, when $c_{\min}$ is considered as a positive constant. On the other hand, Thm. 2 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 or discounted 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: the more the minimum cost is allowed to be small, the longer the duration of the trajectory to reach the goal may be, thus the harder it is analysis-wise to control the trajectory variations of a policy between two models.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

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

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

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved analysis of UCRL2 with empirical bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *NeurIPS*, pages 4868–4878, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning Adversarial MDPs with Bandit Feedback and Unknown Transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 3910–3919, 2017.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872, 2020.

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

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196, 2018a.

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

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

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

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

Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems*, pages 5626–5635, 2019.

Appendix A. High-probability Event

Here we characterize the high-probability event, denoted by $\mathcal{E}$. Throughout the remainder of the analysis, we will assume that $\mathcal{E}$ holds.

Lemma 5 *Denote by $\mathcal{E}$ the event under which for any time step $t \geq 1$ and for any state-action pair (s,a) and next state s', it holds that*

$$|\widehat{p}_t(s'|s,a) - p(s'|s,a)| \leq 4\sqrt{\frac{\widehat{p}_t(s'|s,a)}{N_t^+(s,a)} \log\left(\frac{SAN_t^+(s,a)}{\delta}\right)} + \frac{28\log\left(\frac{SAN_t^+(s,a)}{\delta}\right)}{N_t^+(s,a)}, \quad (9)$$

where $N_t^+(s,a) := \max\{1, N_t(s,a)\}$ with N_t the state-action counts accumulated up to (and including) time t. Then we have $\mathbb{P}(\mathcal{E}) \geq 1-\delta$.

Proof The confidence intervals in Eq. 9 are constructed using the empirical Bernstein inequality, which guarantees that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$, see e.g., Rosenberg et al. (2020); Fruit et al. (2020). ■

Appendix B. Extended Value Iteration for SSP

Here we briefly recall how to perform an extended value iteration (EVI) scheme tailored to SSP, as explained by Tarbouriech et al. (2020a). Note that we leverage a Bernstein-based construction of confidence intervals, as done by e.g., Fruit et al. (2020); Rosenberg et al. (2020).

Formally, we consider as input any SSP-MDP instance M (cf. Def. 1) with cost function c lower bounded by $c_{\min} > 0$, a set $\mathcal{N}$ of samples collected so far (with corresponding state-action counters denoted by N) and a VI precision level $\mu_{\text{VI}} > 0$. We now detail what the scheme $\text{EVI}(\mathcal{N}, c, \mu_{\text{VI}})$ does and outputs.

First it computes a set of plausible SSP-MDPs defined as

$$\mathcal{M} := \{\langle \mathcal{S}, \mathcal{A}, g, \widetilde{p}, c\rangle \mid \widetilde{p}(g|g,a) = 1,\ \widetilde{p}(s'|s,a) \in \mathcal{B}(s,a,s'),\ \sum_{s' \in \mathcal{S}'} \widetilde{p}(s'|s,a) = 1\},$$

where for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\mathcal{B}(s,a,s')$ is a high-probability confidence set on the dynamics of the true SSP-MDP M. Specifically, we define the compact sets $\mathcal{B}(s,a,s') := [\widehat{p}(s'|s,a) - \beta(s,a,s'), \widehat{p}(s'|s,a) + \beta(s,a,s')] \cap [0,1]$, where $\beta(s,a,s')$ denotes the right hand side of Eq. 9. From Lem. 5 the choice of $\beta(s,a,s')$ guarantees that $M \in \mathcal{M}$ with probability at least $1-\delta$.

Once $\mathcal{M}$ has been computed, the scheme applies extended value iteration (EVI) to compute a policy with lowest optimistic value. Formally, it defines the extended optimal Bellman operator $\widetilde{\mathcal{L}}$ such that for any vector $\widetilde{v} \in \mathbb{R}^S$ and non-goal state $s \in \mathcal{S}$,

$$\widetilde{\mathcal{L}}\widetilde{v}(s) := \min_{a \in \mathcal{A}} \left\{ c(s,a) + \min_{\widetilde{p} \in \mathcal{B}(s,a)} \sum_{s' \in \mathcal{S}} \widetilde{p}(s'|s,a)\widetilde{v}(s') \right\}.$$

We consider an initial vector $\widetilde{v}_0 := 0$ and set iteratively $\widetilde{v}_{i+1} := \widetilde{\mathcal{L}}\widetilde{v}_i$. For the predefined VI precision $\mu_{\text{VI}} > 0$, the stopping condition is reached for the first iteration j such that $\|\widetilde{v}_{j+1} - \widetilde{v}_j\|_\infty \leq \mu_{\text{VI}}$. The policy $\widetilde{\pi}$ is then selected to be the optimistic greedy policy w.r.t. the vector $\widetilde{v}_j$. While $\widetilde{v}_j$ is not the value function of $\widetilde{\pi}$ in the optimistic model $\widetilde{p}$, which we denote by $\widetilde{V}^{\widetilde{\pi}}$, both quantities can be related according to Lem. 4.

Appendix C. Useful Results

Proof of Lem. 2 Although we consider here a slightly different $\mathcal{P}_\eta^{(p)}$ set, the result is almost identical to Tarbouriech et al. (2020b, App. C). For completeness, we report the full derivation here. The analysis follows the proof of Rosenberg et al. (2020, Lem. B.4) whose result can be seen as a special case of Lem. 2. First, let us assume that π is proper in the model p'. This implies that its value function, denoted by V', is bounded component-wise. Moreover, for any state $s \in \mathcal{S}$, the Bellman equation holds as follows

$$\begin{aligned} V'(s) &= c(s,\pi(s)) + \sum_{y\in\mathcal{S}} p'(y|s,\pi(s))V'(y) \\ &= c(s,\pi(s)) + \sum_{y\in\mathcal{S}} p(y|s,\pi(s))V'(y) + \sum_{y\in\mathcal{S}} \big(p'(y|s,\pi(s)) - p(y|s,\pi(s))\big)V'(y). \end{aligned} \tag{10}$$

By successively using Hölder's inequality and that $p' \in \mathcal{P}_\eta^{(p)}$ and $c(s,\pi(s)) \geq c_{\min}$, we get

$$V'(s) \geq c(s,\pi(s)) - \eta\|V'\|_\infty + p(\cdot|s,\pi(s))^\top V' \geq c(s,\pi(s))\Big(1 - \frac{\eta\|V'\|_\infty}{c_{\min}}\Big) + p(\cdot|s,\pi(s))^\top V'.$$

Let us now introduce the vector $V'' := \Big(1 - \frac{\eta\|V'\|_\infty}{c_{\min}}\Big)^{-1} V'$. Then for all $s \in \mathcal{S}$,

$$V''(s) \geq c(s,\pi(s)) + p(\cdot|s,\pi(s))^\top V''.$$

Hence, from Lem. 6, π is proper in p (i.e., $V < +\infty$), and we have

$$V \leq V'' \leq \left(1 + 2\frac{\eta\|V'\|_\infty}{c_{\min}}\right) V', \tag{11}$$

where the last inequality stems from condition (3) and the fact that $\frac{1}{1-x} \leq 1 + 2x$ holds for any $0 \leq x \leq \frac{1}{2}$. Conversely, analyzing Eq. 10 from the other side, we get

$$V'(s) \leq c(s,\pi(s))\left(1 + \frac{\eta\|V'\|_\infty}{c_{\min}}\right) + p(\cdot|s,\pi(s))^\top V'.$$

Let us now introduce the vector $V'' := \Big(1 + \frac{\eta\|V'\|_\infty}{c_{\min}}\Big)^{-1} V'$. Then

$$V''(s) \leq c(s,\pi(s)) + p(\cdot|s,\pi(s))^\top V''.$$

We then obtain in the same vein as Lem. 6 (by leveraging the monotonicity of the Bellman operator $\mathcal{L}^\pi U(s) := c(s,\pi(s)) + p(\cdot|s,\pi(s))^\top U$) that $V'' \leq V$, and therefore

$$V' \leq \left(1 + \frac{\eta\|V'\|_\infty}{c_{\min}}\right) V. \tag{12}$$

Combining Eq. 11 and 12 yields component-wise

$$\|V - V'\|_\infty \leq 2\frac{\eta\|V'\|_\infty}{c_{\min}}\|V'\|_\infty + \frac{\eta\|V'\|_\infty}{c_{\min}}\|V\|_\infty \leq 7\frac{\eta\|V'\|_\infty^2}{c_{\min}},$$

where the last inequality stems from plugging condition (3) into Eq. 11.

Note that here p and p' play symmetric roles; we can perform the same reasoning in the case where π is proper in the model p and it would yield an equivalent result by switching the dependencies on V and V'. ■

Lemma 6 (Bertsekas and Tsitsiklis, 1991, Lem. 1) *Consider an SSP instance under the conditions of Lem. 1. Let π be any policy, then*

- *If there exists a vector $U : \mathcal{S} \to \mathbb{R}$ such that $U(s) \geq c(s, \pi(s)) + \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s))U(s')$ for all $s \in \mathcal{S}$, then π is proper, and V^π the value function of π is upper bounded by U component-wise, i.e., $V^\pi(s) \leq U(s)$ for all $s \in \mathcal{S}$.*
- *If π is proper, then its value function V^π is the unique solution to the Bellman equations $V^\pi(s) = c(s, \pi(s)) + \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s))V^\pi(s')$ for all $s \in \mathcal{S}$.*

We now state a useful result which showcases the exponential decay of the goal-reaching probability of a proper policy with component-wise bounded value function.

Lemma 7 (Rosenberg et al., 2020, Lem. B.5) *Let π be a proper policy such that for some $d > 0$, $V^\pi(s) \leq d$ for every non-goal state s. 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 $m \geq 0$. Note that a cost of at most m implies that the number of steps is at most $m/c_{\min}$.*

We finally spell out an important property stemming from optimism.

Lemma 8 *Under the event $\mathcal{E}$, we have $\widetilde{V} \leq V^\star + \frac{\varepsilon}{3}$ component-wise.*

Proof Denote by $\widetilde{v}$ the VI vector output by the computation of the candidate policy $\widetilde{\pi}$ via EVI. From Lem. 4 and by the choice of the VI precision $\mu_{\text{VI}} := \frac{\varepsilon c_{\min}}{6\Delta}$, we have component-wise that $\widetilde{V} \leq \left(1 + \frac{2\mu_{\text{VI}}}{c_{\min}}\right)\widetilde{v} \leq V^\star + \frac{\varepsilon}{3\Delta}\widetilde{v} \leq V^\star + \frac{\varepsilon}{3}$ since $\widetilde{v} \leq \Delta$ by construction of Alg. 2. ■

C.1. Procedure to Estimate an Upper Bound of the SSP-Diamater

Lemma 9 (D-SUBROUTINE) *With probability at least $1 - \delta$, the D-SUBROUTINE (Alg. 4):*

- *has a sample complexity bounded by $\widetilde{O}\big(D^2\Gamma SA/\varepsilon^2 + DS^2A/\varepsilon\big)$,*
- *requires at most $\log_2(D(1+\varepsilon)) + 1$ inner iterations,*
- *outputs a quantity $\widehat{D}$ that verifies $D \leq \widehat{D} \leq (1 + 2\varepsilon(1+\varepsilon))(1+\varepsilon)D$.*

We now delve into the analysis of the $\widehat{D}$-SUBROUTINE. Throughout the remainder of the proof, we will assume that the event $\mathcal{E}$ holds. We now give a useful statement stemming from optimism.

Lemma 10 *At any stage of the $\widehat{D}$-SUBROUTINE, denote by $\widetilde{v}$ the vector computed using* EVI *for SSP (App. B). Then under the event $\mathcal{E}$, we have component-wise (i.e., starting from any non-goal state) that $\widetilde{v} \leq \min_\pi V_p^\pi \leq D$.*

Algorithm 4: D-SUBROUTINE

Input: accuracy $\varepsilon > 0$, confidence level $\delta \in (0,1)$.
Set $W := \frac{1}{2}$ and $\|\widetilde{v}\|_\infty := 1$.
while $\|\widetilde{v}\|_\infty > W$ **do**
  Set $W \leftarrow 2W$.
  Set the accuracy $\eta := \frac{\varepsilon}{W}$.
  Collect additional samples until $\widehat{p} \in \mathcal{P}_{\eta/2}$ with confidence level δ (we verify this using the Bernstein upper bound of Eq. 9)
  Compute $(\widetilde{v}, \_) := \text{EVI}\big(\mathcal{N}, c=1, \mu_{\text{VI}} := \frac{\varepsilon}{2}\big)$.
end
Output: the optimistic quantity $\widehat{D} := (1 + 2\eta\|\widetilde{v}\|_\infty)\|\widetilde{v}\|_\infty$.

Proof The first inequality stems from Lem. 4 while the second inequality uses the definition of the SSP-diameter D and the fact that the considered costs are equal to 1. ■

We now prove Lem. 9. Denote by i the iteration index of the subroutine (starting at $i = 1$), so that $W_i = 2^i$. Introduce $j := \min\{i \geq 1 : \|\widetilde{v}_i\|_\infty \leq W_i\}$. By choice of each optimistic model, we have $\|\widetilde{v}_i\|_\infty \leq D$ at any iteration $i \geq 1$ from Lem. 10. Since $(W_i)_{i\geq 1}$ is a strictly increasing sequence, the subroutine is bound to end in a finite number of iterations (i.e., $j < +\infty$), and given that $W_{j-1} \leq \|\widetilde{v}_{j-1}\|_\infty \leq D$, we get $j \leq \log_2(D) + 1$. Moreover, we have $\|\widetilde{v}_j\|_\infty \leq W_j$ and $\eta_j = \frac{\varepsilon}{W_j}$, which implies that $\eta_j \leq \frac{\varepsilon}{\|\widetilde{v}_j\|_\infty}$. Moreover, combining $W_{j-1} \leq D$ and $W_{j-1} = \frac{W_j}{2} = \frac{\varepsilon}{2\eta_j}$ yields that $\frac{\varepsilon}{2D} \leq \eta_j$. The Bernstein upper bound of Eq. 9 entails that the total sample complexity is bounded by $\widetilde{O}\big(D^2\Gamma SA/\varepsilon^2 + DS^2A/\varepsilon\big)$. Now, denote by $\widetilde{v}$ the optimistic matrix output by the $\widehat{D}$-SUBROUTINE. Let us consider $s_1 \in \arg\max_s \min_\pi \mathbb{E}[\tau_\pi(s)]$. Denote by $\widetilde{\pi}$ the greedy policy w.r.t. the vector $\widetilde{v}$ in the optimistic model. Then we have

$$
\begin{aligned}
D = \min_\pi \mathbb{E}[\tau_\pi(s_1)] \leq \mathbb{E}[\tau_{\widetilde{\pi}}(s_1)] &\overset{(a)}{\leq} (1 + 2\eta\|\mathbb{E}[\widetilde{\tau}_{\widetilde{\pi}}]\|_\infty)\mathbb{E}[\widetilde{\tau}_{\widetilde{\pi}}(s_1)] \\
&\overset{(b)}{\leq} (1 + 2\eta(1+\varepsilon)\|\widetilde{v}\|_\infty)(1+\varepsilon)\widetilde{v}(s_1) \\
&\leq (1 + 2\eta(1+\varepsilon)\|\widetilde{v}\|_\infty)(1+\varepsilon)\|\widetilde{v}\|_\infty := \widehat{D} \\
&\overset{(c)}{\leq} (1 + 2\eta(1+\varepsilon)\|\widetilde{v}\|_\infty)(1+\varepsilon)D \\
&\overset{(d)}{\leq} (1 + 2\varepsilon(1+\varepsilon))(1+\varepsilon)D,
\end{aligned}
$$

where (a) corresponds to the simulation lemma for SSP (Lem. 2), (b) comes from the value iteration precision $\mu_{\text{VI}} := \frac{\varepsilon}{2}$ which implies that $\mathbb{E}[\widetilde{\tau}_{\widetilde{\pi}}] \leq (1 + 2\mu_{\text{VI}})\widetilde{v} \leq (1+\varepsilon)\widetilde{v}$ component-wise according to Lem. 4, (c) is implied by Lem. 10, and finally (d) uses that $\eta\|\widetilde{v}\|_\infty \leq \varepsilon$ as proved above.

Appendix D. Proof of Thm. 1

Here we provide the proof of Thm. 1. Denoting by the subscript i the quantities considered at any iteration i of Alg. 2, recall that the algorithm terminates at the first iteration i such that $\|\widetilde{v}_i\|_\infty \leq \Delta_i$. This implies that at the previous iteration $i-1$, $\Delta_{i-1} < \|\widetilde{v}_{i-1}\|_\infty$. Also, $\Delta_i = 2\Delta_{i-1}$ and $\|\widetilde{v}_{i-1}\|_\infty \leq B_\star$ from optimism. Combining everything gives $\Delta_i \leq 2B_\star$. Therefore when Alg. 3 terminates it is aware of a quantity $\Delta := \Delta_i$ such that $\|\widetilde{v}\|_\infty \leq \Delta \leq 2B_\star$.

We denote by n the minimum number of samples collected at each state-action pair. We denote by $\widetilde{\pi}$ the candidate policy output by Alg. 1. Let us denote by V and $\widetilde{V}$ the value functions of policy $\widetilde{\pi}$ in the true model p and the optimistic model $\widetilde{p}$, respectively (note that we may have $V = +\infty$ for some components if $\widetilde{\pi}$ is not proper in p).

Note that $p := p(\cdot|\cdot, \widetilde{\pi}(\cdot))$, $\widehat{p} := \widehat{p}(\cdot|\cdot, \widetilde{\pi}(\cdot))$ and $\widetilde{p} := \widetilde{p}(\cdot|\cdot, \widetilde{\pi}(\cdot))$ can be seen as matrices. Our analysis draws inspiration from variance-aware techniques, see e.g., Azar et al. (2013, 2017); Fruit et al. (2020); Rosenberg et al. (2020). We will make multiple use of the Cauchy-Schwartz inequality, for which we will use the symbol $\overset{\text{(C-S)}}{\leq}$. We assume throughout that the event $\mathcal{E}$ holds. Finally, we introduce the (unknown) quantity $\Gamma := \max_{s,a}\|p(\cdot|s,a)\|_0$, and its empirical counterpart $\widehat{\Gamma} := \max_{s,a}\|\widehat{p}(\cdot|s,a)\|_0$ (note that we always have $\widehat{\Gamma} \leq \Gamma$).

We first require to have $\widetilde{p} \in \mathcal{P}_\eta^{(p)}$ with accuracy $\eta = \frac{c_{\min}}{6\Delta}$. To do so, we use the triangle inequality to write $|\widetilde{p} - p| \leq |\widetilde{p} - \widehat{p}| + |\widehat{p} - p|$. The second term is bounded by the empirical Bernstein inequality (Eq. 9), and the first term is bounded the same way by construction of EVI. Hence, by inverting Eq. 9 to extract n and after some algebraic manipulations (i.e., by applying the technical lemma of Kazerouni et al., 2017, Lem. 8), is it sufficient to require

$$\boxed{n = \Omega\left(\frac{\Delta^2\widehat{\Gamma}}{c_{\min}^2}\log^2\left(\frac{\Delta SA}{\delta c_{\min}}\right) + \frac{\Delta}{c_{\min}}\log\left(\frac{\Delta SA}{\delta c_{\min}}\right)\right).} \qquad (\alpha)$$

The simulation lemma (Lem. 2) then ensures that $\widetilde{\pi}$ is proper in p, and moreover that its value function verifies $V \leq 2\Delta$ component-wise by virtue of Lem. 8. Since $\widetilde{\pi}$ is proper in both p and $\widetilde{p}$, the associated Bellman equations hold, thus entailing the following for any non-goal state s

$$\begin{aligned} V(s) - \widetilde{V}(s) &= \sum_{y\in\mathcal{S}} p(y|s)V(y) - \sum_{y\in\mathcal{S}} \widetilde{p}(y|s)\widetilde{V}(y) \\ &= \sum_{y\in\mathcal{S}} p(y|s)(V(y) - \widetilde{V}(y)) + \sum_{y\in\mathcal{S}}(p(y|s) - \widetilde{p}(y|s))\widetilde{V}(y). \end{aligned}$$

Let us define

$$W(s) := \sum_{y\in\mathcal{S}}(p(y|s) - \widetilde{p}(y|s))\widetilde{V}(y).$$

Note that $W(g) = 0$. Denote by $Q \in \mathbb{R}^{S\times S}$ the transition matrix restricted between the non-goal states of policy $\widetilde{\pi}$ in the true model p, i.e., for any $(s,s') \in \mathcal{S}^2, Q(s,s') := p(s'|s, \widetilde{\pi}(s))$. Since $\widetilde{\pi}$ is proper in p, the matrix Q is strictly substochastic which implies that the matrix $(I - Q)$ is invertible,

and therefore we have

$$V(s) - \widetilde{V}(s) = \left[(I - Q)^{-1} W\right]_s$$
$$= \sum_{t=0}^{+\infty} \mathbb{E}_{\widetilde{\pi},p}\Big[\mathbb{1}_{s_t \neq g} W(s_t) \quad |s_0 = s\Big].$$

First, let us consider that $V(s) \leq \widetilde{V}(s)$. Then from Lem. 8 we immediately have that $V(s) \leq V^\star(s) + \frac{\varepsilon}{3}$. From now on, we thus consider that $V(s) \geq \widetilde{V}(s)$. Hence we have

$$V(s) - \widetilde{V}(s) \leq \sum_{t=0}^{+\infty} \mathbb{E}_{\widetilde{\pi},p}\Big[\mathbb{1}_{s_t \neq g} |W(s_t)| \quad |s_0 = s\Big]. \tag{13}$$

From now on, for notational simplicity, we will omit the (implicit) dependency $s_0 = s$ for the expectations. We bound each term $|W(s_t)|$. Given that $\widetilde{V}(g) = 0$ and both $p(\cdot|s)$ and $\widetilde{p}(\cdot|s)$ are probability distributions over $\mathcal{S}'$, the "shifting" trick (also performed by e.g., Fruit et al., 2020; Jin et al., 2020; Rosenberg et al., 2020) yields

$$W(s_t) = \sum_{y \in \mathcal{S}'} (p(y|s_t) - \widetilde{p}(y|s_t)) \left(\widetilde{V}(y) - \sum_{z \in \mathcal{S}} p(z|s_t) \widetilde{V}(z)\right).$$

In addition the empirical Bernstein inequality entails that there exist two absolute positive constants c_1 and c_2 such that $|p(s'|s_t) - \widetilde{p}(s'|s_t)| \leq c_1 \sqrt{\frac{\widehat{p}(s'|s_t) \log(S' A \delta^{-1} n)}{n}} + c_2 \frac{\log(S' A \delta^{-1} n)}{n}$ (see e.g., Fruit et al., 2020, Thm. 10). Recall that $S' = S + 1$ amounts to the total number of states (i.e., the S non-goal states plus the goal state g). Setting $Z(s_t) := \sum_{z \in \mathcal{S}} p(z|s_t) \widetilde{V}(z)$, we have

$$\begin{aligned}
|W(s_t)| &\leq \sum_{s' \in \mathcal{S}'} |\widetilde{V}(s') - Z(s_t)| \cdot |p(s'|s_t) - \widetilde{p}(s'|s_t)| \\
&\leq c_1 \sum_{s' \in \mathcal{S}'} \sqrt{\frac{\widehat{p}(s'|s_t) \Big(|\widetilde{V}(s') - Z(s_t)|\Big)^2 \log(S' A \delta^{-1} n)}{n}} + 2 c_2 \sum_{s' \in \mathcal{S}'} \frac{\Delta \log(S' A \delta^{-1} n)}{n} \\
&\overset{\text{(C-S)}}{\leq} c_1 \sqrt{\frac{\widehat{\Gamma} \log(S' A \delta^{-1} n)}{n}} \sqrt{\sum_{s' \in \mathcal{S}'} \widehat{p}(s'|s_t) \Big(|\widetilde{V}(s') - Z(s_t)|\Big)^2} + 2 c_2 \sum_{s' \in \mathcal{S}'} \frac{\Delta \log(S' A \delta^{-1} n)}{n} \\
&\leq c_1 \sqrt{\frac{\log(S' A \delta^{-1} n)}{n}} \sqrt{\left|\sum_{s' \in \mathcal{S}'} (\widehat{p}(s'|s_t) - p(s'|s_t)) 4 \Delta^2\right|} \\
&\quad + c_1 \sqrt{\frac{\widehat{\Gamma} \log(S' A \delta^{-1} n) \mathbb{V}(s_t)}{n}} + 2 c_2 \frac{\Delta S' \log(S' A \delta^{-1} n)}{n},
\end{aligned} \tag{14}$$

where we use the subadditivity of the square root and define the following variance

$$\mathbb{V}(s_t) := \sum_{s' \in \mathcal{S}'} p(s'|s_t) \left(\widetilde{V}(s') - \sum_{s'' \in \mathcal{S}} p(s''|s_t) \widetilde{V}(s'')\right)^2.$$

Leveraging (α) which guarantees that $\widehat{p} \in \mathcal{P}_\eta^{(p)}$ with accuracy $\eta = \frac{c_{\min}}{6\Delta}$, the first term in Eq. 14 can be bounded as $c_1\sqrt{\frac{\widehat{\Gamma}\log(S'A\delta^{-1}n)}{n}}\Delta\sqrt{\frac{c_{\min}}{6\Delta}}$. Consequently, plugging the bound of Eq. 14 into Eq. 13 yields

$$V(s) - \widetilde{V}(s) \leq \text{❶} + \text{❷} + \text{❸},$$

where

$$\begin{aligned}
\text{❶} &:= c_1\sqrt{\frac{\widehat{\Gamma}\log(S'A\delta^{-1}n)}{n}}\sum_{t=0}^{+\infty}\mathbb{E}_{\widetilde{\pi},p}\Big[\mathbb{1}_{s_t\neq g}\sqrt{\mathbb{V}(s_t)}\Big],\\
\text{❷} &:= c_1\sqrt{\frac{\widehat{\Gamma}\log(S'A\delta^{-1}n)}{n}}\Delta\sqrt{\frac{c_{\min}}{6\Delta}}\sum_{t=0}^{+\infty}\mathbb{P}_{\widetilde{\pi},p}(s_t\neq g),\\
\text{❸} &:= c_2\frac{\Delta S'\log(S'A\delta^{-1}n)}{n}\sum_{t=0}^{+\infty}\mathbb{P}_{\widetilde{\pi},p}(s_t\neq g).
\end{aligned}$$

Leveraging that $V \leq 2\Delta$ component-wise, we obtain that $\mathbb{P}_{\widetilde{\pi},p}(s_t \neq g) \leq 2\exp\big(-\frac{c_{\min}t}{8\Delta}\big)$ by applying Lem. 7 with $m = c_{\min}t$. To make an analogy to the infinite-horizon discounted setting studied by Azar et al. (2013), we can observe that we have $\mathbb{P}_{\widetilde{\pi},p}(s_t \neq g) \sim \gamma^t$ where $\gamma \sim \exp\big(-\frac{c_{\min}}{\Delta}\big) < 1$.

$$\sum_{t=0}^{+\infty}\mathbb{P}_{\widetilde{\pi},p}(s_t\neq g) \leq \frac{2}{1-\exp\big(-\frac{c_{\min}}{8\Delta}\big)} = \frac{2\exp\big(\frac{c_{\min}}{8\Delta}\big)}{\exp\big(\frac{c_{\min}}{8\Delta}\big)-1} \leq \frac{19\Delta}{c_{\min}},$$

where the last inequality uses that $e^x \geq 1 + x$ holds for any real x. Consequently, we get

$$\text{❸} \leq \frac{19c_2\Delta^2 S'\log(S'A\delta^{-1}n)}{c_{\min}n}.$$

We seek to ensure that $\text{❸} \leq \frac{2\varepsilon}{9}$. There simply remains to invert the inequality above to extract n and do some algebraic manipulations (see e.g., Kazerouni et al., 2017, Lem. 9). We thus require that:

$$\boxed{n = \Omega\left(\frac{\Delta^2 S}{c_{\min}\varepsilon}\log\left(\frac{\Delta SA}{c_{\min}\varepsilon\delta}\right)\right)} \qquad (\beta)$$

Furthermore, we have

$$\text{❷} \leq 19c_1\frac{\Delta}{c_{\min}}\sqrt{\frac{\widehat{\Gamma}\log(S'A\delta^{-1}n)}{n}}\Delta\sqrt{\frac{c_{\min}}{6\Delta}}.$$

We seek to ensure that $\text{❷} \leq \frac{2\varepsilon}{9}$. There simply remains to invert the inequality above to extract n and do some algebraic manipulations (see e.g., Kazerouni et al., 2017, Lem. 9). We thus require that:

$$\boxed{n = \Omega\left(\frac{\Delta^3\widehat{\Gamma}}{c_{\min}\varepsilon^2}\log\left(\frac{\Delta SA}{c_{\min}\varepsilon\delta}\right)\right).} \qquad (\gamma)$$

We now proceed in bounding ❶. To do so, we split the time into *intervals*, similar to Rosenberg et al. (2020). The first interval begins at the first time step, and each interval ends when its total cost accumulates to at least Δ (or when the goal state g is reached). Denote by t_m the time step at the beginning of the m-th interval, and by H_m the length of the m-th interval. An important property is that $H_m \leq 2\Delta/c_{\min}$. Denote by $\mathbb{I}_m$ the boolean equal to 1 if the goal g is not reached by the end of the m-th interval, and denote by $s_{(m)}$ the state at the end of the m-th interval. Note that $\mathbb{I}_m = 1 \iff s_{(m)} \neq g$, implying that $\mathbb{E}_{\widetilde{\pi},p}[\mathbb{I}_m] = \mathbb{P}_{\widetilde{\pi},p}(s_{(m)} \neq g)$. We introduce a change of variable in the sums, from the time index t to the interval index m. Formally, for any time index t, there exists an interval $m+1$ (during which it occurs) and an integer $h \in [H_{m+1}]$ such that $t = \sum_{i=0}^{m} H_i + h$. The change of variable yields the following

$$\begin{aligned}
\sum_{t=0}^{+\infty} \mathbb{E}_{\widetilde{\pi},p}\left[\mathbb{1}_{s_t \neq g}\sqrt{\mathbb{V}(s_t)}\right] &\leq \sum_{m=0}^{+\infty} \mathbb{E}_{\widetilde{\pi},p}\left[\mathbb{I}_m \sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}} \sqrt{\mathbb{V}(s_h)}\right] \\
&\overset{\text{(C-S)}}{\leq} \sum_{m=0}^{+\infty} \mathbb{P}_{\widetilde{\pi},p}(s_{(m)} \neq g) \sqrt{\mathbb{E}_{\widetilde{\pi},p}\left[\left(\sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}} \sqrt{\mathbb{V}(s_h)}\right)^2\right]} \\
&\overset{\text{(C-S)}}{\leq} \sum_{m=0}^{+\infty} \mathbb{P}_{\widetilde{\pi},p}(s_{(m)} \neq g) \sqrt{\mathbb{E}_{\widetilde{\pi},p}\left[H_{m+1} \sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}} \mathbb{V}(s_h)\right]} \\
&\leq \sqrt{\frac{2\Delta}{c_{\min}}} \sum_{m=0}^{+\infty} \mathbb{P}_{\widetilde{\pi},p}(s_{(m)} \neq g) \sqrt{\mathbb{E}_{\widetilde{\pi},p}\left[\sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}} \mathbb{V}(s_h)\right]}. \qquad (15)
\end{aligned}$$

To bound the expression above, we first use the property shown by Rosenberg et al. (2020, Lem. 4.7) that whenever every state-action pair has been sampled sufficiently many times (specifically, at least $\alpha B_\star S c_{\min}^{-1} \log(B_\star S A c_{\min}^{-1} \delta^{-1})$ times for some constant $\alpha > 0$, which is the case here), then the expected variance $\mathbb{V}$ accumulated over a whole interval m can be bounded as $O(B_\star^2)$, i.e., there exists an absolute constant $c_3 > 0$ such that

$$\mathbb{E}_{\widetilde{\pi},p}\left[\sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}} \mathbb{V}(s_h)\right] \leq c_3 \Delta^2. \qquad (16)$$

Second, we bound the series of the probabilities. The construction of the intervals entails that if the m-th interval does not end in the goal state, then the cumulative cost to reach the goal state is more than Δm. Furthermore, the probability of the latter event can be bounded by Lem. 7 leveraging the component-wise inequality $V \leq 2\Delta$. As a result, we get

$$\mathbb{P}_{\widetilde{\pi},p}(s_{(m)} \neq g) \leq 2\exp\left(-\frac{\Delta m}{8\Delta}\right) = 2\exp\left(-\frac{1}{8}\right)^m,$$

which implies that

$$\sum_{m=0}^{+\infty} \mathbb{P}_{\widetilde{\pi},p}(s_{(m)} \neq g) \leq \frac{2}{1-\exp(-\frac{1}{8})}. \qquad (17)$$

Plugging Eq. 16 and 17 into Eq. 15 gives

$$❶ \leq 25 c_1 \sqrt{c_3} \sqrt{\frac{\Delta^{3/2} \widehat{\Gamma} \log(S' A \delta^{-1} n)}{c_{\min} n}}.$$

We seek to ensure that $❶ \leq \frac{2\varepsilon}{9}$. There simply remains to invert the inequality above to extract n and do some algebraic manipulations (see e.g., Kazerouni et al., 2017, Lem. 9). We thus require (once again) that:

$$\boxed{n = \Omega\left(\frac{\Delta^3 \widehat{\Gamma}}{c_{\min} \varepsilon^2} \log\left(\frac{\Delta S A}{c_{\min} \varepsilon \delta}\right)\right).} \tag{γ}$$

Overall, combining the requirements of Eq. (α), (β) and (γ) means that we get the component-wise guarantee that $V \leq \widetilde{V} + \frac{2\varepsilon}{3}$, and therefore from Lem. 8 that $V \leq V^\star + \varepsilon$, as soon as:

$$\boxed{n = \Omega\left(\frac{\Delta^3 \widehat{\Gamma}}{c_{\min} \varepsilon^2} \log\left(\frac{\Delta S A}{c_{\min} \varepsilon \delta}\right) + \frac{\Delta^2 S}{c_{\min} \varepsilon} \log\left(\frac{\Delta S A}{c_{\min} \varepsilon \delta}\right) + \frac{\Delta^2 \widehat{\Gamma}}{c_{\min}^2} \log^2\left(\frac{\Delta S A}{c_{\min} \delta}\right)\right).}$$

Appendix E. Proof of Thm. 2

Here we provide the proof of Thm. 2 by establishing that the output policy $\widetilde{\pi}$ of Alg. 3 is ε-optimal w.r.t. the restricted set Π_θ. We assume that the event $\mathcal{E}$ holds. Here we offset all the costs with the additive perturbation $\nu = \frac{\varepsilon}{2\theta \widehat{D}}$. We use the subscript ν to denote quantities considered in the *perturbed model*. In the perturbed model, the costs are set to $c'_\nu(s,a) := \max\{c(s,a), \nu\}$, which in particular implies that the minimum cost verifies $\min_{s,a} c'_\nu(s,a) \geq \nu$.

The application of Thm. 1 in the perturbed model immediately yields the component-wise inequality $V_\nu \leq V_\nu^\star + \frac{\varepsilon}{2}$. Moreover, let $\pi^\S \in \min_{\pi \in \Pi_\theta} V^\pi$, $V^\S := V^{\pi^\S}$ and $T^\S := \mathbb{E}[\tau_{\pi^\S}]$. In particular, we have $V_\nu^\star \leq V_\nu^\S$ and $T^\S(s) \leq \theta D_s \leq \theta \widehat{D}$. Furthermore, given the choice of ν and the fact that $c'_\nu(s,a) \leq c(s,a) + \nu$, we have $V_\nu^\S \leq V^\S + \nu T^\S \leq V^\S + \frac{\varepsilon}{2}$. Lastly, we have $V \leq V_\nu$. Putting everything together yields the sought-after inequality $V \leq V^\S + \varepsilon$.