Sample Complexity Bounds for Stochastic Shortest Path with a Generative Model

Jean Tarbouriech[†][§] Matteo Pirotta[†] Michal Valko Alessandro Lazaric[†]

July 2020

Abstract

In this technical note, we consider the problem of computing an ε-optimal policy in a stochastic shortest path (SSP) setting, provided that we can access a generative sampling model. We propose two algorithms for this setting and derive PAC bounds on their sample complexity.

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 literature (Azar et al., 2013; Wang, 2017; Sidford et al., 2018a,b; Zanette et al., 2019; Agarwal et al., 2019; 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 very popular in RL, in practice, many problems are better formalized within the strictly more general stochastic shortest-path (SSP) setting (Bertsekas, 2012), where the objective is to compute a policy that minimizes the cost accumulated before reaching a specific goal state.

Recently, Tarbouriech et al. (2020) and Cohen et al. (2020) studied the SSP problem in the online case and derived the first regret bounds for this setting. In this paper we focus on the generative model case and study the problem of computing a near-optimal policy for a given cost function and goal state. Leveraging tools from (Azar et al., 2013) and (Cohen et al., 2020), we derive two closely related algorithms for this problem 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 adapts 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 and still minimize the cumulative cost). Nonetheless, in this case the objective is to compute an ε-optimal solution in the set of (proper) policies that reach that goal in at most a number of steps that is proportional to the minimum number of steps to the goal (i.e., the SSP-diameter of the problem). Finally, we discuss a simple extension of our algorithms to the cost-free case (i.e., when no cost function is initially provided as input) and we illustrate some directions for future investigation.

[†] Facebook AI Research, Paris

[§] Inria Lille - Nord Europe

2 Preliminaries

Stochastic Shortest Path (SSP). We introduce the notion of MDP with an SSP objective (Bertsekas, 2012, 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 $\tau_\pi(s) := \min\{t \geq 0 : s_{t+1} = g \,|\, s_1 = s, \pi\}$. We define the *SSP-diameter* D as

$$D := \max_{s \in \mathcal{S}} D_s, \qquad \text{with} \quad D_s := \min_{\pi \in \Pi} \mathbb{E}[\tau_\pi(s)], \tag{1}$$

where the expectation is w.r.t. the random trajectory.

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.*

Assumption 1. *There exists at least one proper policy.*

Note that since the number of states S is finite, Asm. 1 implies that $D < +\infty$. The *value function* (also called expected cost-to-go) of a policy $\pi \in \Pi$ is defined as

$$V^\pi(s_0) := \mathbb{E}\Big[\sum_{t=1}^{+\infty} c(s_t, \pi(s_t)) \,\Big|\, s_0\Big] = \mathbb{E}\Big[\sum_{t=1}^{\tau_\pi(s_0)} c(s_t, \pi(s_t)) \,\Big|\, s_0\Big].$$

The 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

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

Lemma 1 (Bertsekas & 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 introduce the following quantities: $B_\star := \|V^\star\|_\infty \leq D$ 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.

Problem Formulation. We assume that the costs c are deterministic and known to the learner, while the transition dynamics p is unknown. We also assume access to a generative model, which for any given state-action pair (s,a) returns a sample drawn from $p(\cdot|s,a)$.

We investigate the following question:

How many calls to the generative model are sufficient to compute a near-optimal policy with high probability?

Since the problem of learning in SSP-MDPs has already been studied in the regret minimization setting by (Tarbouriech et al., 2020; Cohen 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. No guarantee is provided about the value function (i.e., the expected cumulative costs) of one episode. Indeed, 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. Furthermore, the near-optimal guarantees we intend to achieve are for any arbitrary initial state, while regret-style guarantees are only in expectation with respect to a starting state distribution.

3 Main Result

We first illustrate the common structure of our algorithms. We receive as input the 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, similar 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[1]

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

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 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, which 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 it 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, consider a set $\mathcal{N}$ of samples collected so far, a VI precision level $\gamma > 0$ and any strictly positive cost function c' (which entails that the conditions of Lem. 1 hold). $\text{EVI}(\mathcal{N}, c', \gamma)$ outputs an optimistic value vector $\widetilde{v}$ and an optimistic policy $\widetilde{\pi}$ that is greedy w.r.t. $\widetilde{v}$. EVI can be run efficiently as recently shown by Tarbouriech et al. (2020) (here, we consider Bernstein-based concentration inequalities, as done in e.g., (Fruit et al., 2020)). The crucial 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 as shown by the following lemma (which stems from Tarbouriech et al., 2020, Lem. 4 & App. E).

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

We are now ready to detail our algorithms.

[1] 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 $(\widetilde{v}, \widetilde{\pi}) := \mathrm{EVI}\big(\mathcal{N}, c', \gamma := \frac{\iota\varepsilon}{6\Delta}\big)$ with $\mathcal{N}$ the samples collected so far.
  if $\|\widetilde{v}\|_\infty \leq \Delta$ then
    continue = False.
  end if
end while
Output: the policy $\widetilde{\pi}$.
```

3.1 Strictly Positive Cost Function

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 such that $\|V^\pi - V^\star\|_\infty \leq \varepsilon$ with probability at least $1 - \delta$.*

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. 2, 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. 2) is (ε, δ)-optimal with a sample complexity bounded by*

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

3.2 Any Cost Function and Restricted Optimality

Whenever $c_{\min} = 0$, Algorithm 1 has a possibly unbounded sample complexity. Indeed, the case of zero minimum cost is a complex SSP problem, where the optimal policy is not even guaranteed to be proper Bertsekas (2012). To handle it, we propose to 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 in (e.g., Bertsekas & Yu, 2013; Tarbouriech et al., 2020; Cohen 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

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

```
1: Input: slack parameter $\theta \geq 1$, accuracy $\varepsilon > 0$, confidence level $\delta \in (0,1)$, cost function $c$, allocation function $\phi(\cdot,\cdot)$.
2: First compute $\widehat{D}$ an upper bound estimate of the SSP-diameter (see Alg. 4 of App. B.3).
3: Set cost perturbation $\nu = \frac{\varepsilon}{2\theta\widehat{D}}$.
4: $\widetilde{\pi} = \texttt{SEARCH}(c \vee \nu)$.
5: Output: the policy $\widetilde{\pi}$.
```

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.

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. 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 such that $\|V^\pi - V^\star_\theta\|_\infty \leq \varepsilon$ 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.*

While alternative definitions of restricted set may be introduced, we believe 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. 1). Nonetheless, this poses an additional layer of complexity, since D_s is unknown to the algorithm, which only receives θ as additional parameter. 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. B.3 (Alg. 4) and stress that the amount of samples used in this initial phase is subsumed in the final sample complexity result by the second phase where we compute the final candidate policy.

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 γ. 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. 2) is $(\varepsilon, \delta, \theta)$-optimal with a sample complexity bounded by*

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

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 $\lim_{\varepsilon \to 0} \Pi_{\varepsilon^{-1}} = \Pi$.

4 Discussion

In this section we discuss the bounds obtained in Thm. 1 and 2, and compare those obtained in the DMDP setting, which we recall is a subclass of the SSP-MDP setting (Bertsekas, 2012).

We first observe that the dependency in the state space is in 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 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., 2020; Cohen et al., 2020) scale as $\widetilde{O}(S)$, which contrasts with the lower bound in $\sqrt{S}$ derived in (Cohen et al., 2020) and with the 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.

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 a simulation lemma for SSP (Lem. 4), 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 shares multiple similarities with — and can be seen as a combination of — i) (Cohen et al., 2020) on regret minimization for the SSP problem and ii) (Azar et al., 2013) on sample complexity of DMDPs with a generative model. The latter work removes a factor $1/(1-\gamma)$, with $\gamma < 1$ the discount factor. As such, and as we flesh out in the analysis, a parallel can be made between the SSP-MDP and DMDP settings, by relating $\gamma \sim \exp\left(-c_{\min}/B_\star\right) < 1$. This is not surprising insofar as, more generally, DMDPs are a subclass of SSP-MDPs (Bertsekas, 2012).

On the one hand, 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 to control the trajectory variations of a policy between two models.

Finally, we note that the analysis estimates the transition kernel accurately well across the state-action space, which 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).

5 Conclusion

To the best of our knowledge, this technical note presents the first sample complexity bounds for SSP-MDPs with a generative model. *1)* Improving the bounds is the natural direction for further investigation. In particular, it appears promising to incorporate recent techniques yielding improvements on the sample complexity in the particular case of DMDPs (Sidford et al., 2018a; Agarwal et al., 2019; Li et al., 2020). *2)* There also remains to derive a lower bound for the SSP sample complexity with a generative model. Leveraging the unit-cost lower bound on SSP-regret (Cohen et al., 2020) may be relevant, and it would also be interesting to devise a problem exhibiting the critical role of $c_{\min}$. *3)* Finally, an interesting direction is to derive SSP sample complexity bounds that display the variances of the costs and next-state optimal value function and the gaps in the optimal action-value function, e.g., inspired from recent work in DMDPs (Zanette et al., 2019).

References

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. *arXiv preprint arXiv:1906.03804*, 2019.

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

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

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

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

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

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

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

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *NeurIPS*, pp. 4868–4878, 2018.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning Adversarial MDPs with Bandit Feedback and Unknown Transition. *CoRR*, abs/1912.01192, 2019.

Kazerouni, A., Ghavamzadeh, M., Abbasi, Y., and Van Roy, B. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pp. 3910–3919, 2017.

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

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint arXiv:2005.12900*, 2020.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018a.

Sidford, A., Wang, M., Wu, X., and Ye, Y. 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*, pp. 770–787. SIAM, 2018b.

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*, 2020.

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

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

A High-probability Event

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

Lemma 3. *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*

$$|\widehat{p}_t(s'|s,a) - p(s'|s,a)| \leq 4\sqrt{\frac{\widehat{p}_t(s'|s,a)}{N_t^+(s,a)} \log\left(\frac{SAN_t^+(s,a)}{\delta}\right)} + \frac{28\log\left(\frac{SAN_t^+(s,a)}{\delta}\right)}{N_t^+(s,a)}, \tag{3}$$

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. 3 are constructed using the empirical Bernstein inequality, which guarantees that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$, see e.g., Cohen et al. (2020); Fruit et al. (2020). □

B Useful Results

B.1 Simulation Lemma for SSP

For any $\eta > 0$, we introduce the set "around" the model p as follows

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

In the following we state a general simulation lemma in the SSP setting.

Lemma 4 (Simulation Lemma for SSP)**.** *Consider any $p' \in \mathcal{P}_\eta$. Suppose that for each model (p and p'), there exists at least one proper policy w.r.t. the goal state g. Consider any policy π that is proper in p', with value function denoted by V'_π, such that the following condition is verified*

$$\eta \|V'_\pi\|_\infty \leq 2c_{\min}. \tag{4}$$

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

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

Proof. The proof of Lem. 4, which is a straightforward generalization of (Cohen et al., 2020, Lem. B.4), requires the following property.

Lemma 5 (Bertsekas & Tsitsiklis, 1991, Lem. 1)**.** *Consider an SSP instance with terminal state g, non-terminal states $\mathcal{S}$, transition dynamics p, positive non-terminal costs c, such that there exists at least one proper policy. 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 $s \in \mathcal{S}$.*

First, let us assume that π is proper in the model p'. This implies that $V' < +\infty$ component-wise. Moreover, for any non-terminal state $s \in \mathcal{S}$, the Bellman equation holds as follows

$$\begin{aligned} V'(s) &= c(s, \pi(s)) + \sum_{y \in \mathcal{S}} p'(y|s, \pi(s))V'(y) \\ &= c(s, \pi(s)) + \sum_{y \in \mathcal{S}} p(y|s, \pi(s))V'(y) + \sum_{y \in \mathcal{S}} (p'(y|s, \pi(s)) - p(y|s, \pi(s)))V'(y). \end{aligned} \tag{5}$$

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

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

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

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

Hence, from Lem. 5, π is proper in p (i.e., $V < +\infty$), and we have

$$V \leq V'' \leq \left(1 + 2\frac{\eta \|V'\|_\infty}{c_{\min}}\right) V', \tag{6}$$

where the last inequality stems from condition (4) and the fact that $\frac{1}{1-x} \leq 1 + 2x$ holds for any $0 \leq x \leq \frac{1}{2}$. Conversely, analyzing Eq. 5 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 $V'' := \left(1 + \frac{\eta \|V'\|_\infty}{c_{\min}}\right)^{-1} V'$. Then

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

We then obtain in the same vein as Lem. 5 (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

$$V' \leq \left(1 + \frac{\eta \|V'\|_\infty}{c_{\min}}\right) V. \tag{7}$$

Combining Eq. 6 and 7 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 (4) into Eq. 6.

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'. □

Finally, for comparison purposes, let us recall the classical simulation lemma in finite-horizon RL.

Lemma 6 (Simulation Lemma in Finite-Horizon, see e.g., (Kearns & Singh, 2002))**.** *Consider any $p' \in \mathcal{P}_\eta$ for any $\eta > 0$. Consider as value function in p the expected cumulative reward over H steps, i.e., for any policy π and starting state $s \in \mathcal{S}$, $V_\pi(s) := \mathbb{E}\Big[\sum_{t=1}^{H} r(s_t, \pi(s_t)) \mid s_1 = s\Big]$ (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*

$$\|V_\pi - V'_\pi\|_\infty \leq \eta H^2.$$

We spell out the key differences between the simulation lemma in the finite-horizon setting (Lem. 6) and in the SSP setting (Lem. 4), bringing to light the criticalities in the latter setting. First, a guarantee can only be obtained if the condition (4) is verified, which involves both the accuracy η and the value function of π in $p' \in \mathcal{P}_\eta$. We observe that the smaller the minimum cost $c_{\min} := \min_{s,a} c(s,a)$, 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 hyperparameter H, but by the ratio between the (a priori unknown) infinity norm of the value functions of the considered policy and the minimum cost $c_{\min}$.

B.2 Other Useful Results

We also 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 (Cohen 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 $\frac{m}{c_{\min}}$.*

We now 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. 2 and by the choice of the VI precision $\gamma := \frac{\varepsilon c_{\min}}{6\Delta}$, we have component-wise that $\widetilde{V} \leq \Big(1 + \frac{2\gamma}{c_{\min}}\Big)\widetilde{v} \leq V^\star + \frac{\varepsilon}{3\Delta}\widetilde{v} \leq V^\star + \frac{\varepsilon}{3}$ since $\widetilde{v} \leq \Delta$ by construction of Alg. 1. □

B.3 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 $D \leq \widehat{D} \leq (1 + 2\varepsilon(1+\varepsilon))(1+\varepsilon)D$.*

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 $\delta$ (we verify this using the Bernstein upper
    bound of Eq. 3)
    Compute $(\widetilde{v}, \_) := \mathrm{EVI}\big(\mathcal{N}, c = 1, \gamma := \frac{\varepsilon}{2}\big)$.
end while
Output: the optimistic quantity $\widehat{D} := (1 + 2\eta\|\widetilde{v}\|_\infty)\|\widetilde{v}\|_\infty$.
```

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, for any given goal state, denote by $\widetilde{v}$ the vector computed using* EVI *for SSP. Then under the event $\mathcal{E}$, we have component-wise (i.e., starting from any non-goal state): $\widetilde{v} \leq \min_\pi V_p^\pi \leq D$.*

Proof. The first inequality stems from Lem. 2 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. 3 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

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

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

C Proof of Thm. 1

Leveraging the results of App. B, we are ready to establish the proof of Thm. 1. Recall that when Alg. 1 terminates, it is aware of a quantity $\Delta > 0$ such that $\|\widetilde{v}\|_\infty \leq \Delta$. Moreover, since at any stage of the algorithm we have $\|\widetilde{v}\|_\infty \leq B_\star$ and given the way Δ is doubled at each iteration, we get that $\Delta \leq 2B_\star$.

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); Cohen 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 $\widetilde{p} \in \mathcal{P}_\eta$ 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. 3), and the first term is bounded the same way by construction of EVI. Hence, by inverting Eq. 3 to extract n and after some algebraic manipulations (i.e., by applying technical lemma (Kazerouni et al., 2017, Lem. 8)), is it sufficient to require

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

The simulation lemma (Lem. 4) 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 between the non-goal states of policy $\widetilde{\pi}$ in the true model p (i.e., for any $(s, s') \in \mathcal{S}, Q(s, s') := p(s'|s, \widetilde{\pi}(s))$). Since $\widetilde{\pi}$ is proper in p, Q is strictly substochastic which implies that the matrix $(I - Q)$ is invertible, and therefore we have

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

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

$$V(s) - \widetilde{V}(s) \leq \sum_{t=0}^{+\infty} \mathbb{E}_{\widetilde{\pi}, p} \Big[\mathbb{1}_{s_t \neq g} |W(s_t)| \quad | s_0 = s \Big]. \qquad (8)$$

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 in e.g., (Fruit et al., 2020; Jin et al., 2019; Cohen 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

$$\begin{aligned}
|W(s_t)| &\leq \sum_{s' \in \mathcal{S}'} |\widetilde{V}(s') - Z(s_t)| \cdot |p(s'|s_t) - \widetilde{p}(s'|s_t)| \\
&\leq c_1 \sum_{s' \in \mathcal{S}'} \sqrt{\frac{\widehat{p}(s'|s_t)\Big(|\widetilde{V}(s') - Z(s_t)|\Big)^2 \log(S'A\delta^{-1}n)}{n}} + 2c_2 \sum_{s' \in \mathcal{S}'} \frac{\Delta \log(S'A\delta^{-1}n)}{n} \\
&\overset{\text{(C-S)}}{\leq} c_1 \sqrt{\frac{\widehat{\Gamma} \log(S'A\delta^{-1}n)}{n}} \sqrt{\sum_{s' \in \mathcal{S}'} \widehat{p}(s'|s_t)\Big(|\widetilde{V}(s') - Z(s_t)|\Big)^2} + 2c_2 \sum_{s' \in \mathcal{S}'} \frac{\Delta \log(S'A\delta^{-1}n)}{n} \\
&\leq c_1 \sqrt{\frac{\log(S'A\delta^{-1}n)}{n}} \sqrt{\left| \sum_{s' \in \mathcal{S}'} (\widehat{p}(s'|s_t) - p(s'|s_t)) 4\Delta^2 \right|} \\
&\quad + c_1 \sqrt{\frac{\widehat{\Gamma} \log(S'A\delta^{-1}n)\mathbb{V}(s_t)}{n}} + 2c_2 \frac{\Delta S' \log(S'A\delta^{-1}n)}{n},
\end{aligned} \tag{9}$$

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 $\widehat{p} \in \mathcal{P}_\eta$ with accuracy $\eta = \frac{c_{\min}}{6\Delta}$, the first term in Eq. 9 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. 9 into Eq. 8 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 in (Azar et al., 2013), we can observe that we have $\mathbb{P}_{\widetilde{\pi},p}(s_t \neq g) \sim \gamma^t$ where $\gamma \sim \exp\left(-\frac{c_{\min}}{\Delta}\right) < 1$.

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

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

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

We seek to ensure that 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:

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

Furthermore, we have

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

We seek to ensure that 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:

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

We now proceed in bounding ❶. To do so, we split the time into *intervals*, similar to (Cohen 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 $H_m \leq \frac{2\Delta}{c_{\min}}$. 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

$$\begin{aligned}
\sum_{t=0}^{+\infty}\mathbb{E}_{\widetilde{\pi},p}\left[\mathbb{1}_{s_t\neq g}\sqrt{\mathbb{V}(s_t)}\right] &\leq \sum_{m=0}^{+\infty}\mathbb{E}_{\widetilde{\pi},p}\left[\mathbb{I}_m\sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}}\sqrt{\mathbb{V}(s_h)}\right] \\
&\overset{\text{(C-S)}}{\leq}\sum_{m=0}^{+\infty}\mathbb{P}_{\widetilde{\pi},p}(s_{(m)}\neq g)\sqrt{\mathbb{E}_{\widetilde{\pi},p}\left[\left(\sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}}\sqrt{\mathbb{V}(s_h)}\right)^2\right]} \\
&\overset{\text{(C-S)}}{\leq}\sum_{m=0}^{+\infty}\mathbb{P}_{\widetilde{\pi},p}(s_{(m)}\neq g)\sqrt{\mathbb{E}_{\widetilde{\pi},p}\left[H_{m+1}\sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}}\mathbb{V}(s_h)\right]} \\
&\leq \sqrt{\frac{2\Delta}{c_{\min}}}\sum_{m=0}^{+\infty}\mathbb{P}_{\widetilde{\pi},p}(s_{(m)}\neq g)\sqrt{\underbrace{\mathbb{E}_{\widetilde{\pi},p}\left[\sum_{h=t_{m+1}}^{t_{m+1}+H_{m+1}}\mathbb{V}(s_h)\right]}_{\leq c_3\Delta^2}} \\
&\leq \sqrt{2c_3}\Delta\sqrt{\frac{\Delta}{c_{\min}}}\sum_{m=0}^{+\infty}\mathbb{P}_{\widetilde{\pi},p}(s_{(m)}\neq g),
\end{aligned}$$

where we used that the expected variance $\mathbb{V}$ accumulated over a whole interval can be bounded by $c_3\Delta^2$ with c_3 an absolute constant, i.e., $\mathbb{E}[\sum_{interval}\mathbb{V}] \leq c_3\Delta^2$, as shown in (Cohen et al., 2020, Lem. 4.7). There remains to bound the series above. 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

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

Consequently, we get

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

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:

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

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:

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

D Proof of Thm. 2

We prove that the output policy $\widetilde{\pi}$ 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 critically 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 $V_\nu \leq V^\star_\nu + \frac{\varepsilon}{2}$. 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 $V^\star_\nu \leq V^\S_\nu$ 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 $V^\S_\nu \leq V^\S + \nu T^\S \leq V^\S + \frac{\varepsilon}{2}$. Lastly, we have $V \leq V_\nu$. Putting everything together yields the sought-after inequality $V \leq V^\S + \varepsilon$.