# No-regret exploration in goal-oriented reinforcement learning

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, Alessandro Lazaric

# No-Regret Exploration in Goal-Oriented Reinforcement Learning

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

## Abstract

Many popular reinforcement learning problems (e.g., navigation in a maze, some Atari games, mountain car) are instances of the *episodic setting* under its *stochastic shortest path* (SSP) formulation, where an agent has to achieve a goal state while minimizing the cumulative cost. Despite the popularity of this setting, the exploration-exploitation dilemma has been sparsely studied in general SSP problems, with most of the theoretical literature focusing on different problems (i.e., finite-horizon and infinite-horizon) or making the restrictive *loop-free* SSP assumption (i.e., no state can be visited twice during an episode). In this paper, we study the general SSP problem with no assumption on its dynamics (some policies may actually never reach the goal). We introduce UC-SSP, the first no-regret algorithm in this setting, and prove a regret bound scaling as $\widetilde{\mathcal{O}}(DS\sqrt{ADK})$ after $K$ episodes for any unknown SSP with $S$ states, $A$ actions, positive costs and *SSP-diameter* $D$, defined as the smallest expected hitting time from any starting state to the goal. We achieve this result by crafting a novel stopping rule, such that UC-SSP may interrupt the current policy if it is taking *too long* to achieve the goal and switch to alternative policies that are designed to *rapidly* terminate the episode.

## 1. Introduction

We consider the problem of exploration-exploitation in episodic Markov decision processes (MDPs), where the objective is to minimize the expected cost to reach a specific goal state. Several popular reinforcement learning (RL) problems fall into this framework, such as navigation problems, many *Atari games* (e.g., breakout) and Mujoco environments (e.g., reacher). In all these problems, the length of an episode (i.e., the time to reach the goal state) is unknown and depends on the policy executed during the episode. Furthermore, the performance is not directly connected to the length of the episode, as the objective is to minimize the cost over time rather than reaching the goal state as fast as possible. The conditions for the existence and the computation of an optimal policy have been studied in the MDP literature under the name of the *stochastic shortest path* (SSP) problem (Bertsekas, 2012, Sect. 3).

The exploration-exploitation dilemma has been extensively studied in the finite-horizon (see e.g., Azar et al., 2017; Zanette & Brunskill, 2019) and infinite-horizon settings (see e.g., Jaksch et al., 2010; Fruit et al., 2018a;b). In the former, the performance is optimized over a fixed and known horizon of $H$ steps. Typically, this model is used to solve SSP problems by setting $H$ *large enough*. While for $H \to \infty$ the optimal finite-horizon policy converges to the optimal SSP policy, for any finite $H$, this approach may introduce a bias leading exploration algorithms to converge to suboptimal policies and suffer linear regret (see e.g., Toromanoff et al., 2019, for a discussion of this problem in Atari games). In the latter, the performance is optimized for the asymptotic average cost. While this removes any strict "deadline", it does not introduce any incentive to reach the goal state. This may favor policies with small average cost and yet poor performance in the SSP sense, as they may never terminate. Note that SSP forms an important class of MDPs as both infinite-horizon (discounted) and finite-horizon MDPs, two much more extensively researched settings, are a subtype of SSP-MDPs (Bertsekas, 2012; Guillot & Stauffer, 2020).

Prior work on exploration in SSPs can be divided in two cases. The first is the online shortest path routing problem, which has deterministic dynamics and stochastic rewards. In this case, the optimal policy is open-loop (i.e., it is a sequence of actions independent from the states) and it can be solved as an instance of a combinatorial bandit problem (see e.g., György et al., 2007; Talebi et al., 2017). Exploration algorithms know the set of admissible paths of bounded length and regret bounds are available in both the semi- and full-bandit setting. The second case allows for stochastic transitions and mostly considers adversarial problems, but it is restricted to *loop-free* environments (see e.g., Jin et al., 2020; Rosenberg & Mansour, 2019a;b; Neu et al., 2012; 2010; Zimin & Neu, 2013). Under this assumption, the state

[1]Facebook AI Research, Paris, France [2]SequeL team, Inria Lille - Nord Europe, France. Correspondence to: Jean Tarbouriech <jean.tarbouriech@gmail.com>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. 

space can be decomposed into $L$ non-intersecting layers $X_0, \dots, X_L$ such that $X_0 = \{x_0\}$ and $X_L = \{x_L\}$, and transitions are only possible between consecutive layers. In this case, it is possible to derive regret bounds leveraging the fact that *any* episode length is upper bounded by $L$ almost surely. Unfortunately, this requirement is restrictive and fails to hold in many realistic environments.

In this paper, exploration in general SSP problems is investigated for the first time. The solution of an SSP is obtained by computing the policy minimizing the value function, i.e., the expected costs accumulated until reaching the goal state. Studying SSP value functions poses technical difficulties that do not appear in the conventional settings such as loop-free SSP, finite-horizon and infinite-horizon: **1)** it features two possibly conflicting objectives: quickly reaching the goal state while minimizing the costs along the way; **2)** it is unbounded for policies that may never reach the goal state (i.e., non-proper policies); **3)** it is not state-independent (a crucial property of the gain of any optimal policy in infinite-horizon); **4)** its number of summands may differ from one trajectory to another due to variations in the time to reach the goal state (thus making the regret decomposition tricky compared to finite-horizon); **5)** it cannot be computed using backward induction (a crucial technique used in finite-horizon); **6)** it cannot be discounted (since a discount factor would have a undesirable effect of biasing importance towards short-term behavior and thus weakening the incentive to eventually reach the goal state). This last point means that SSP-MDPs do not have a notion of "equivalent horizon", which is $1/(1-\gamma)$ in the special case of infinite-horizon discounted MDPs with known discount factor $\gamma$, thus making the general setting of SSP-MDPs more difficult to analyze.

While we leverage algorithmic and technical tools from both finite- and infinite-horizon settings, tackling the general SSP problem requires introducing novel techniques to manage the challenges highlighted above. Notably, we investigate the properties of *optimistic* policies and their associated *discrete phase-type distributions* (i.e., the hitting time distribution) to design a novel criterion to stop executing the current optimistic SSP policy *during* an episode and switch to alternative policies designed to rapidly reach the goal.

The main contributions of this paper are: **1)** We formalize *exploration-exploitation* in SSP problems by defining an adequate notion of regret (Sect. 2). **2)** We show that the special case of SSP with uniform costs can be cast as an infinite-horizon problem and tackled by UCRL2 (Jaksch et al., 2010) with a regret bound adapting to the complexity of the environment (Sect. 3). **3)** We then introduce UC-SSP, the first algorithm with vanishing regret in general SSP problems (Sect. 4). We also show that not only UC-SSP effectively deals with the general case, but it remains competitive (if not better) even in the limit cases of uniform costs or loop-free SSP, which can be addressed by infinite- and finite-horizon regret minimization algorithms respectively. **4)** Moreover, we demonstrate how our (mild) assumptions (e.g., no dead-end states, positive costs) can be effectively relaxed using variants of UC-SSP (Sect. 5). Finally, we support our theoretical findings with experiments in App. J.

## 2. Stochastic Shortest Path (SSP)

We consider a finite *stochastic shortest path* problem (Bertsekas, 2012, Sect. 3) $M := \langle \mathcal{S}', \mathcal{A}, c, p, s_0 \rangle$, where $\mathcal{S}' := \mathcal{S} \cup \{\overline{s}\}$ is the set of states with $\overline{s}$ being the goal state (also called the terminal state) and $s_0 \in \mathcal{S}$ being the starting state[1], and $\mathcal{A}$ is the set of actions. We denote by $A = |\mathcal{A}|$ and $S = |\mathcal{S}|$ the number of actions and non-goal states. Each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ is characterized by a known, deterministic cost $c(s,a)$ and an unknown transition probability distribution $p(\cdot \mid s,a)$ over next states. The goal state $\overline{s}$ is absorbing (i.e., $p(\overline{s} \mid \overline{s}, a) = 1$ for all $a \in \mathcal{A}$) and cost-free (i.e., $c(\overline{s}, a) = 0$ for all $a \in \mathcal{A}$). We assume the following property of the cost function.

**Assumption 1.** There exist known constants $0 < c_{\min} \leq c_{\max}$ such that $c(s,a) \in [c_{\min}, c_{\max}]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Extending the setting to unknown, stochastic costs poses no major difficulty, as long as the learner knows in advance the range of the costs, i.e., the constants $c_{\min}$ and $c_{\max}$ (see App. I.1). Moreover, in Sect. 5 we derive a variant of our algorithm that can handle zero costs (i.e., $c_{\min} = 0$).

Bertsekas (2012) showed that under Asm. 1 we can restrict the attention to the set of stationary deterministic policies $\Pi^{\text{SD}} := \{\pi : \mathcal{S} \to \mathcal{A}\}$. For any $\pi \in \Pi^{\text{SD}}$ and $(s, s') \in \mathcal{S} \times \mathcal{S}'$, the (possibly unbounded) *hitting time* to $s'$ starting from $s$ is denoted by $\tau_\pi(s \to s') := \inf\{t \geq 0 : s_{t+1} = s' \mid s_1 = s, \pi\}$. We also set $\tau_\pi(s) := \tau_\pi(s \to \overline{s})$.

**Assumption 2.** We define the *SSP-diameter* $D$ as

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

and we assume that $D < +\infty$.

We say that $M$ is *SSP-communicating* when Asm. 2 holds. We defer to Sect. 5 the treatment of the case $D = +\infty$.

The *value function* (also called expected cost-to-go) of any $\pi \in \Pi^{\text{SD}}$ is defined as

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

For any vector $V \in \mathbb{R}^S$, the optimal Bellman operator is

[1] Our algorithm can handle any (possibly unknown) distribution of initial states.

defined as

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

An important role in the definition of the SSP is played by the set $\Pi^{\text{PSD}} \subseteq \Pi^{\text{SD}}$ of proper stationary policies.

**Definition 1.** A stationary policy $\pi$ is *proper* if $\overline{s}$ is reached with probability 1 from any state in $\mathcal{S}$ following $\pi$.[2]

The next lemma shows that the SSP problem is well-posed.

**Lemma 1.** *Under Asm. 1 and 2, there exists an optimal policy $\pi^\star \in \arg\min_{\pi\in\Pi^{PSD}} V^\pi(s_0)$ for which $V^\star = V^{\pi^\star}$ is the unique solution of the optimality equations $V^\star = \mathcal{L}V^\star$ and $V^\star(s) < +\infty$ for any $s \in \mathcal{S}$.*

Similarly to the average-reward case, we can provide a bound on the range of the optimal value function depending on the largest cost and the SSP-diameter.

**Lemma 2.** *Under Asm. 1 and 2, $\|V^\star\|_\infty \le c_{\max} D$.*

For any $\pi \in \Pi^{\text{PSD}}$, its (almost surely finite) hitting time starting from any state in $\mathcal{S}$ follows a *discrete phase-type distribution*, or in short *discrete PH distribution* (see e.g., Latouche & Ramaswami, 1999, Sect. 2.5 for an introduction). Indeed, its induced Markov chain is terminating with a single absorbing state $\overline{s}$ and all the other states are transient. The transition matrix associated to $\pi$, denoted by $P_\pi \in \mathbb{R}^{(S+1)\times(S+1)}$, can thus be arranged in the following canonical form

$$P_\pi = \begin{bmatrix} Q_\pi & R_\pi \\ 0 & 1 \end{bmatrix},$$

where $Q_\pi \in \mathbb{R}^{S\times S}$ is the transition matrix between non-absorbing states (i.e., $\mathcal{S}$) and $R_\pi \in \mathbb{R}^S$ is the transition vector from $\mathcal{S}$ to $\overline{s}$. Note that $Q_\pi$ is strictly substochastic ($Q_\pi \mathbb{1} \le \mathbb{1}$ where $\mathbb{1} := (1,\dots,1)^T \in \mathbb{R}^S$ and $\exists j$ s.t. $(Q_\pi \mathbb{1})_j < 1$). Denoting by $\mathbb{1}_s$ the $S$-sized one-hot vector at the position of state $s \in \mathcal{S}$, we have the following result (see e.g., Latouche & Ramaswami, 1999, Thm. 2.5.3).

**Proposition 1.** *For any $\pi \in \Pi^{PSD}$, $s \in \mathcal{S}$ and $n > 0$,*

$$\mathbb{P}(\tau_\pi(s) > n) = \mathbb{1}_s^\top Q_\pi^n \mathbb{1} = \sum_{s'\in\mathcal{S}} (Q_\pi^n)_{ss'}.$$

Finally, for any $X \in \mathbb{R}^{m\times n}$ we define the $\infty$-matrix-norm $\|X\|_\infty := \max_{1\le i\le m} \sum_{j=1}^n |X_{ij}|$.

**Learning problem.** We consider the learning problem where $\mathcal{S}'$, $\mathcal{A}$, and $c$ are known, while the dynamics $p$ is unknown and can be estimated online. An *environmental episode* starts at $s_0$ and ends *only* when the goal state $\overline{s}$ is reached. We evaluate the performance of an algorithm $\mathfrak{A}$ after $K$ environmental episodes by its cumulative *SSP-regret*

$$\Delta(\mathfrak{A}, K) := \sum_{k=1}^K \Big[ \Big( \sum_{h=1}^{\tau_k(s_0)} c(s_{k,h}, \mu_k(s_{k,h})) \Big) - V^\star(s_0) \Big],$$

where for any $k \in [K]$,[3] $\tau_k(s_0)$ is the length of episode $k$ following a possibly non-stationary policy $\mu_k = (\pi_{k,0}, \pi_{k,1}, \pi_{k,2}, \dots)$, $\pi_{k,i} \in \Pi^{\text{SD}}$, until $\overline{s}$ is reached. Moreover, $s_{k,h}$ denotes the $h$-th state visited during episode $k$. $\Delta(\mathfrak{A}, K)$ also corresponds to the cumulative SSP-regret after $T_K$ steps, where $T_K := \sum_{k=1}^K \tau_k(s_0)$ is the time step at the end of episode $K$. This definition resembles the infinite-horizon regret, where the performance of the algorithm is evaluated by the costs accumulated by executing $\mu_k$. At the same time, it incorporates the episodic nature of finite-horizon problems, where the performance of the optimal policy is evaluated by its value function at the initial state. Nonetheless, notice that we cannot use the finite-horizon regret definition, i.e., $\sum_{k=1}^K V^{\mu_k}(s_0) - V^\star(s_0)$, where a policy $\mu_k$ is chosen at the beginning of the episode and run until its termination. Indeed, as $\mu_k$ may be non-proper and satisfy $V^{\mu_k}(s_0) = +\infty$, the execution of a single non-proper policy would directly lead to an unbounded regret.

## 3. Uniform-cost SSP

In this section we focus on the SSP problems with uniform costs to illustrate a very first case where a sublinear regret can be achieved without any restrictive loop-free assumption. In particular, we show that in this case the SSP problem can be cast as an infinite-horizon problem and that an algorithm such as UCRL2 (Jaksch et al., 2010) can be directly applied and achieve surprisingly good regret guarantees.

**Assumption 3 (only in Sect. 3).** The costs $c(s,a)$ are constant (equal to 1 w.l.o.g.) for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

In this case, solving the SSP problem corresponds to computing the policy minimizing the expected hitting time to the goal $\overline{s}$.

We introduce the infinite-horizon reward-based MDP $M_\infty := \langle \mathcal{S}', \mathcal{A}, r_\infty, p_\infty, s_0 \rangle$, with reward $r_\infty = \mathbb{1}_{\overline{s}}$ and $p_\infty(\cdot \,|\, s,a) = p(\cdot \,|\, s,a)$ for $s \ne \overline{s}$ and $p_\infty(\cdot \,|\, \overline{s}, a) = \mathbb{1}_{s_0}$ for all $a$. In words, the transitions in $M_\infty$ behave as in $M$ and give zero rewards except at $\overline{s}$ where all actions give a reward of 1 and loop back to $s_0$ instead of self-looping with probability 1. We show that the solution of $M_\infty$ coincides with solving the original SSP and we bound the SSP-regret of UCRL2 applied to this problem.

[2] Note that Def. 1 is slightly different from (and is implied by) the conventional definition of Bertsekas (2012, Sect. 3.1), for which a policy is proper if there is a positive probability that $\overline{s}$ will be reached after at most $S$ stages.

[3] For any integer $n$, we denote by $[n]$ the set $\{1,\dots,n\}$.

**Theorem 1.** *For any policy $\pi \in \Pi^{SD}$, let $\rho_\pi := \lim_{T\to+\infty} \mathbb{E}_\pi\big[\sum_{t=1}^T r_t/T\big]$ be the average reward of $\pi$ in the MDP $M_\infty$. Under Asm. 3, we have*

$$\pi^\star = \arg\min_\pi V^\pi(s_0) = \arg\min_\pi \mathbb{E}[\tau_\pi(s_0)] = \arg\max_\pi \rho^\pi.$$

*With probability $1-\delta$,* UCRL2 *run for any $K \geq 1$ episodes suffers a regret*

$$\Delta(\text{UCRL2}, K) \leq 34(V^\star(s_0)+1)DS\sqrt{AT_K \log\Big(\frac{T_K}{\delta}\Big)}, \quad (2)$$

*with*

$$T_K \leq 2(V^\star(s_0)+1)K + \widetilde{O}\big(V^\star(s_0)^2 D^2 S^2 A\big). \quad (3)$$

Up to logarithmic and lower-order terms, the previous bound scales as $\widetilde{O}(V^\star(s_0)DS\sqrt{AT_K})$. This can be contrasted with the infinite-horizon regret $\Delta_\infty := T\rho^\star - \sum_t r_t$ of UCRL2, which in general infinite-horizon problems scales as $\widetilde{O}(D_\infty S\sqrt{AT})$, where $D_\infty := \max_{s\neq s'\in\mathcal{S}'} \min_{\pi\in\Pi^{\text{SD}}} \mathbb{E}[\tau_\pi(s\to s')]$ is the diameter of $M_\infty$ (Jaksch et al., 2010) and measures the longest shortest path between *any* two states. We first notice that the "extra" factor $V^\star(s_0)$ is a direct consequence of the different definition of regret in the two settings. In fact, we have $\Delta = (V^\star(s_0)+1)\Delta_\infty$. As UCRL2 is designed for general infinite-horizon problems, we can only bound the regret $\Delta_\infty$ and use the previous equality to translate it into the corresponding SSP-regret. As such, the factor $V^\star(s_0)$ is the price to pay for adapting UCRL2 to the SSP case. On the other hand, it is easy to see that in general $D \leq D_\infty$. Interestingly, Asm. 2 does not imply that $M_\infty$ is communicating, which is needed for proving regret bounds for UCRL2 in general MDPs. Thm. 1 shows that even when $M_\infty$ is weakly-communicating ($D_\infty = +\infty$) and some states may not be accessible from one another, UCRL2 is able to adapt to the SSP nature of the problem and achieve a bounded regret.

Importantly, notice that no assumption is made about the properness of the policies. The key for UCRL2 to manage policies that may never reach the goal state is the construction of *internal* episodes, where policies are interrupted when the number of samples collected in a state-action pair is doubled. This allows UCRL2 to avoid accumulating too much regret when executing non-proper policies (they are eventually stopped) and, at the same time, perform well when the current policy is near-optimal (it is not stopped too early). Nonetheless, the stopping condition only relies on the number of samples and it is completely agnostic to the episodic nature of the SSP problem.

While the previous analysis suggests that algorithms for infinite-horizon MDPs could be readily executed in SSP problems with strong regret guarantees, this is no longer the case when moving to the general setting of non-uniform costs. Indeed, in order to estimate the performance of a stationary policy w.r.t. its value function, we cannot use the average-cost criterion since it does not capture the incentive to reach the goal state. As an illustrative example, consider the deterministic two-state SSP $M$ from Fig. 1. The optimal SSP policy $\pi^\star$ always selects action $a_2$ since it has minimal value $V^\star(s_0) = c_{\max}$. The optimal infinite-horizon policy always selects action $a_1$ since it has minimal average cost $\rho^\star = c_{\min}$, whereas $\rho_{\pi^\star} = c_{\max}/2$. Consequently, running UCRL2 in general SSP may converge to a suboptimal policy and yield linear SSP-regret.

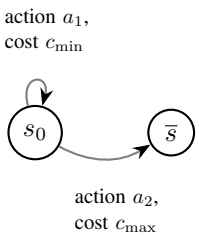


*Figure 1.* Deterministic two-state SSP $M$ with two available actions: $a_1$ which self-loops on $s_0$ with cost $c_{\min}$ and $a_2$ which goes from $s_0$ to $\overline{s}$ with cost $c_{\max} > 2c_{\min}$.

In the next section, we propose a novel algorithm designed to target the general SSP objective function (non-uniform costs) with a two-phase structure and a carefully designed condition to interrupt executing policies.

## 4. General SSP

The general SSP problem requires *(i)* to quickly reach the goal state while *(ii)* at the same time minimizing the cumulative costs. On the one hand, if we constrain the costs to be all equal, objectives *(i)* and *(ii)* coincide and the SSP problem can be addressed using infinite-horizon algorithms as seen in Sect. 3. On the other hand, all previous works in the SSP setting constrain the hitting time of *all* policies (i.e., the loop-free assumption), which means that objective *(i)* is always guaranteed and the algorithm can focus its efforts on objective *(ii)*.

In this section, we tackle head-on the general SSP problem for the first time, where we need *to optimize over the two possibly conflicting objectives (i) and (ii) at the same time.* This poses algorithmic and technical challenges (e.g., non-proper policies may never reach the goal state and have unbounded value function) that require devising a novel optimistic algorithm, specifically designed for SSP problems.

### 4.1. The UC-SSP Algorithm

We present UC-SSP, an algorithm for efficient exploration in general SSP problems (Alg. 1). At a high level, UC-SSP proceeds through each environmental episode $k$ in a *two-phase* fashion. In phase ①, UC-SSP executes a policy trying to

**Algorithm 1** UC-SSP algorithm

**Input:** Confidence $\delta \in (0,1)$, costs, $\mathcal{S}', \mathcal{A}$.
**Initialization:** Set the state-action counter $N_{0,0}(s,a) := 0$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and the time step $t := 1$.
Set $k := 0$. // episode index
Set $G_{0,0} := 0$. // number of attempts in phases ②
**while** $k < K$ **do**
  // New environmental episode
  Increment $k$ += 1.
  Set $j := 0$ // attempts in phase ② of episode $k$
  **while** $s_t \neq \overline{s}$ **do**
    Set $t_{k,j} := t$ and counter $\nu_{k,j}(s,a) := 0$.
    Set $G_{k,j} = G_{k,0} + j$
    Compute $(\widetilde{\pi}_{k,j}, H_{k,j}) := \text{EVI}_{\text{SSP}}(k,j)$.
    **while** $t \leq t_{k,j} + H_{k,j}$ **and** $s_t \neq \overline{s}$ **do**
      Execute action $a_t = \widetilde{\pi}_{k,j}(s_t)$, observe cost $c(s_t,a_t)$ and next state $s_{t+1}$.
      Set $\nu_{k,j}(s_t,a_t)$ += 1.
      Set $t$ += 1.
    **end while**
    **if** $s_t \neq \overline{s}$ **then**
      // Switch to phase ②
      Set $N_{k,j+1}(s,a) := N_{k,j}(s,a) + \nu_{k,j}(s,a)$
      Set $j$ += 1
    **end if**
  **end while**
  Set $N_{k+1,0}(s,a) := N_{k,j}(s,a) + \nu_{k,j}(s,a)$.
  Set $G_{k+1,0} := G_{k,j}$.
**end while**

**Algorithm 2** $\text{EVI}_{\text{SSP}}$

**Input:** Attempt index $(k,j)$ and $N_{k,j}(s,a)$ samples.
**if** $j = 0$ **then**
  $\varepsilon_{k,0} := \frac{c_{\min}}{2t_{k,0}}$, $\gamma_{k,j} := \frac{1}{\sqrt{k}}$.
**else**
  $\varepsilon_{k,j} := \frac{1}{2t_{k,j}}$, $\gamma_{k,j} := \frac{1}{\sqrt{G_{k,j}}}$.
**end if**
Compute estimates $\widehat{p}_{k,j}$ and confidence set $\mathcal{M}_{k,j}$ with the $N_{k,j}$ samples collected so far.
Define the extended optimal Bellman operator $\widetilde{\mathcal{L}}_{k,j}$ as in Eq. (4).
// EVI scheme
Set $m := 0$, $v_0 := \mathbf{0}$ ($S$-sized vector) and $v_1 := \widetilde{\mathcal{L}}_{k,j} v_0$.
**while** $\|v_{m+1} - v_m\|_\infty > \varepsilon_{k,j}$ **do**
  $m$ += 1.
  $v_{m+1} := \widetilde{\mathcal{L}}_{k,j} v_m$.
**end while**
Set $\widetilde{v}_{k,j} := v_m$.
Compute $\widetilde{\pi}_{k,j}$ the optimistic greedy policy w.r.t. $\widetilde{v}_{k,j}$.
Compute $\widetilde{p}_{k,j}$ the corresponding optimistic model.
Compute $\widetilde{Q}_{k,j}$ the transition matrix of $\widetilde{\pi}_{k,j}$ in the optimistic model $\widetilde{p}_{k,j}$ over $\mathcal{S}$, i.e., for any $(s,s') \in \mathcal{S}^2$,

$$\widetilde{Q}_{k,j}(s,s') := \sum_{a \in \mathcal{A}} \widetilde{\pi}_{k,j}(a|s) \widetilde{p}_{k,j}(s'|s,a).$$

Compute $H_{k,j} := \min\Big\{ n > 1 : \|\widetilde{Q}_{k,j}^{n-1}\|_\infty \leq \gamma_{k,j} \Big\}$.
**Output:** policy $\widetilde{\pi}_{k,j}$ and horizon $H_{k,j}$.

solve the SSP problem by tackling both objectives *(i)* and *(ii)* (i.e., reach the goal while minimizing the cumulative costs). We refer to this first policy as an *attempt* in phase ①. As UC-SSP relies on estimates of the true (unknown) SSP, it may select a non-proper policy that would never reach the goal state and incur an unbounded regret. In order to avoid this situation, if the goal state is not reached after a given *pivot* horizon, the algorithm deems the whole episode as a *failure* and it switches to phase ②, whose only objective is to terminate the episode as fast as possible (i.e., it only considers objective *(i)* and disregards the costs). Nonetheless, optimizing an estimate of the hitting time (i.e., objective *(i)*) does not guarantee that the corresponding policy successfully reaches the goal state (i.e., is proper) and multiple *attempts* (i.e., policies) in phase ② may be needed. Similar to phase ①, whenever the goal state is not reached after a certain *pivot* horizon, the current policy is terminated and a new policy is computed. Phase ② and the overall episode ends when the goal state is eventually reached. Notationwise, the $k$-th phase ① is indexed by $(k,0)$ (note that $k$ coincides with the current number of episodes), while the $j$-th attempt in the phase ② of episode $k$ is indexed by $(k,j)$ for $j \geq 1$. Moreover, we denote by $J_k$ the number of attempts performed during the phase ② of episode $k$, and by $G_{k,j}$ the total number of attempts in phases ② up to (and including) attempt $(k,j)$.

**Optimistic policies.** UC-SSP relies on the principle of *optimism in face of uncertainty*. At each attempt, it executes a policy with either lowest optimistic (cost-weighted) value for an attempt in phase ①, or with lowest optimistic expected hitting time for an attempt in phase ②. At the beginning of any attempt $(k,j)$, the algorithm computes a set of plausible MDPs defined as $\mathcal{M}_{k,j} := \{\langle \mathcal{S}, \mathcal{A}, c, \widetilde{p} \rangle \mid \widetilde{p}(\cdot|s,a) \in B_{k,j}(s,a)\}$ where $B_{k,j}(s,a)$ is a high-probability confidence set on the transition probabilities of the true MDP $M$. We set $B_{k,j}(s,a) := \{\widetilde{p} \in \mathcal{C} \mid \widetilde{p}(\cdot \mid \overline{s}, a) = \mathbb{1}_{\overline{s}}, \|\widetilde{p}(\cdot \mid s,a) - \widehat{p}_{k,j}(\cdot \mid s,a)\|_1 \leq \beta_{k,j}(s,a)\}$, with $\mathcal{C}$ the $S'$-dimensional simplex, $\widehat{p}_{k,j}$ the empirical average of transitions prior to attempt $(k,j)$ and

$$\beta_{k,j}(s,a) := \sqrt{\frac{8S \log\left(2AN_{k,j}^+(s,a)\,\delta^{-1}\right)}{N_{k,j}^+(s,a)}},$$

where $N_{k,j}^+(s,a) := \max\{1, N_{k,j}(s,a)\}$ with $N_{k,j}$ being the state-action counts prior to attempt $(k,j)$. The construction of $\beta_{k,j}(s,a)$ guarantees that $M \in \mathcal{M}_{k,j}$ with high probability, as shown in the following lemma.

**Lemma 3.** *Introduce the event* $\mathcal{E} := \bigcap_{k=1}^{+\infty} \bigcap_{j=1}^{J_k} \{M \in \mathcal{M}_{k,j}\}$. *Then* $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\delta}{3}$.

Once $\mathcal{M}_{k,j}$ has been computed, UC-SSP applies an extended value iteration (EVI) scheme (Alg. 2) to compute a policy with lowest optimistic value (if $j = 0$) or lowest optimistic expected hitting time (if $j \geq 1$). Formally, we

define the extended optimal Bellman operator $\widetilde{\mathcal{L}}_{k,j}$ such that for any $v \in \mathbb{R}^{\mathcal{S}}$ and $s \in \mathcal{S}$,

$$\widetilde{\mathcal{L}}_{k,j}v(s) := \min_{a\in\mathcal{A}} \Big\{ c_{k,j}(s,a) + \min_{\widetilde{p}\in B_{k,j}(s,a)} \sum_{y\in\mathcal{S}} \widetilde{p}(y \mid s,a) v(y) \Big\}, \quad (4)$$

where the costs $c_{k,j}$ depend on the phase as follows

$$c_{k,j}(s,a) := \begin{cases} c(s,a) & \text{if } j = 0 \\ 1 & \text{otherwise.} \end{cases}$$

As explained by Jaksch et al. (2010, Sect. 3.1), we can combine all the MDPs in $\mathcal{M}_{k,j}$ into a single MDP $\widetilde{M}$ with extended action set $\mathcal{A}'$. As proved by Bertsekas (2012, Sect. 3.3) about the generalization of the SSP results to a compact action set, the Bellman operator $\widetilde{\mathcal{L}}_{k,j}$ satisfies the contraction property and thus EVI$_{\text{SSP}}$ converges to a vector we denote by $\widetilde{V}^{\star}_{k,j}$. We have the following component-wise inequalities when the stopping condition of Alg. 2 is met.[4]

**Lemma 4.** *For any attempt $(k,j)$, denote by $\widetilde{v}_{k,j}$ the output of* EVI$_{SSP}$ *with operator $\widetilde{\mathcal{L}}_{k,j}$ and accuracy $\varepsilon_{k,j}$. Then $\widetilde{\mathcal{L}}_{k,j}\widetilde{v}_{k,j} \le \widetilde{v}_{k,j} + \varepsilon_{k,j}$. Furthermore, under the event $\mathcal{E}$ we have $\widetilde{v}_{k,j} \le V^{\star}$ if $j = 0$ or $\widetilde{v}_{k,j} \le \min_{\pi} \mathbb{E}(\tau_{\pi})$ otherwise.*

The optimistic policy $\widetilde{\pi}_{k,j}$ executed during attempt $(k,j)$ is the greedy policy w.r.t. $\widetilde{v}_{k,j}$. We also denote by $\widetilde{p}_{k,j}$ the optimistic transition probabilities and by $\widetilde{Q}_{k,j}$ the transition matrix of $\widetilde{\pi}_{k,j}$ in $\widetilde{p}_{k,j}$ over the non-goal states $\mathcal{S}$.

**The pivot horizon.** A crucial aspect for the correct functioning of the algorithm is to carefully select the "pivot" horizon. If the pivot horizon is too small, the algorithm may switch from phase ① to ② too quickly and may perform too many attempts in phase ②. As the policies in phase ② completely disregard the costs, they may lead to suffer large regret. On the other hand, if the pivot horizon is too large and UC-SSP selects a non-proper policy in phase ①, then the regret accumulated during phase ① would be too large.

We select the following length for attempt $(k,j)$

$$H_{k,j} = \min\Big\{ n > 1 : \|(\widetilde{Q}_{k,j})^{n-1}\|_{\infty} \le \frac{\mathbb{1}_{j=0}}{\sqrt{k}} + \frac{\mathbb{1}_{j\ge 1}}{\sqrt{G_{k,j}}} \Big\}. \quad (5)$$

If $\widetilde{\pi}_{k,j}$ is executed for $H_{k,j}$ steps without reaching $\overline{s}$, then attempt $(k,j)$ is said to have *failed* and the next attempt $(k,j+1)$ (necessarily in phase ②) is performed. Otherwise, the attempt is said to have *succeeded*, a new episode begins and the next attempt $(k+1,0)$ (in phase ①) is performed.

[4]Note that the stopping condition is different from the standard one for VI for average reward MDPs (see e.g., Puterman, 2014; Jaksch et al., 2010) that is defined in span seminorm. Also note that as opposed to standard VI, we do not have guarantees of the type $\|v_n - \widetilde{V}^{\star}_{k,j}\|_{\infty} \le \epsilon$ where $\widetilde{V}^{\star}_{k,j} = \widetilde{\mathcal{L}}_{k,j}\widetilde{V}^{\star}_{k,j}$.

Denote by $\widetilde{\tau}_{k,j}$ the hitting time in the model $\widetilde{p}_{k,j}$ of the policy $\widetilde{\pi}_{k,j}$. We first prove that $\widetilde{\pi}_{k,j}$ is proper in $\widetilde{p}_{k,j}$ by connecting its value function to $\widetilde{v}_{k,j}$, which is finite from Lem. 4 (see App. E and Eq. 13). As a result, $\widetilde{\tau}_{k,j}$ follows a *discrete PH distribution* and plugging Prop. 1 into Eq. (5) entails that

$$\max_{s\in\mathcal{S}} \mathbb{P}(\widetilde{\tau}_{k,j}(s) \ge H_{k,j}) \le \frac{\mathbb{1}_{j=0}}{\sqrt{k}} + \frac{\mathbb{1}_{j\ge 1}}{\sqrt{G_{k,j}}}.$$

$H_{k,j}$ is thus selected so that the tail probability of the *optimistic* hitting time is small enough, i.e., there is a high probability that $\widetilde{\pi}_{k,j}$ will *optimistically* reach $\overline{s}$ within $H_{k,j}$ steps. The maximum over $s \in \mathcal{S}$ guarantees this property for any state $s$ from which attempt $(k,j)$ begins (since attempts in phase ② do not necessarily start at $s_0$).

### 4.2. Regret Analysis of UC-SSP

As proved in the following theorem, UC-SSP is the first no-regret learning algorithm in the general SSP setting.

**Theorem 2.** *With overwhelming probability, for any $K \ge 1$, if at each attempt $(k,j)$* EVI$_{SSP}$ *is run with accuracy $\varepsilon_{k,j} := \frac{c_{\min}\mathbb{1}_{j=0} + \mathbb{1}_{j\ge 1}}{2t_{k,j}}$, where $t_{k,j}$ is the time index at the beginning of the attempt, then* UC-SSP *suffers a regret*

$$\Delta(\text{UC-SSP}, K) = \widetilde{O}\Big( c_{\max} DS \sqrt{\frac{c_{\max}}{c_{\min}} ADK} + c_{\max} S^2 A D^2 \Big).$$

**Dependency on $K$ and $D$.** Significantly, UC-SSP achieves an overall rate $\widetilde{O}(\sqrt{K})$ which is optimal w.r.t. the number of episodes $K$. The bound also illustrates how UC-SSP is able to adapt to the complexity of navigating through the MDP as shown by the dependency on the SSP-diameter $D$, which measures the longest shortest path to the goal state from any state. Interestingly, this is achieved without any prior knowledge either on an upper bound of the optimal value function $V^{\star}$ (or of the SSP-diameter itself), or whether the set of policies $\Pi^{\text{SD}}$ contains proper policies or not. We can further inspect the dependency on $D$ by rewriting the regret bound of UC-SSP, which scales as $D^{3/2}\sqrt{K}$ in Thm. 2, as $D\sqrt{T_K}$, where $T_K$ is the total number of steps executed until the end of episode of $K$.[5] As shown in Lem. 2, up to a factor of $c_{\max}$, the SSP-diameter $D$ is an upper bound on the range of the optimal value function and as such it can be (qualitatively) related to the horizon $H$ in the finite-horizon setting and the diameter $D_{\infty}$ in the infinite-horizon setting, which bound the range of the optimal value function and bias function respectively.

**Dependency on cost range.** The multiplicative constant $\frac{c_{\max}}{c_{\min}}$ appearing in the bound quantifies the range of the cost

[5]Even though $T_K$ is a *random* quantity, inspecting the proof (see Sect. 4.3) provides a bound $T_K \lesssim DK$ for $K$ large enough.

function and accounts for the difference from the uniform-cost setting. Interestingly, the presence of the ratio $\frac{c_{\min}}{c_{\max}}$ implies that the regret bound is not invariant w.r.t. a uniform additive perturbation of all costs. This behavior, which does not appear in the finite- or infinite-horizon settings, stems from the fact that an additive offset of costs may alter the optimal policy in the SSP sense (see Lem. 17, App. I).

While the previous discussion shows that UC-SSP successfully tackles general SSP problems, we can also study its behavior in the limit (and much simpler) cases of uniform-cost and loop-free SSP, and compare its regret to infinite- and finite-horizon algorithms respectively.

**Uniform-cost SSP.** Under Asm. 3, UC-SSP achieves a regret of $\widetilde{O}(DS\sqrt{ADK})$, in contrast with the bound $\widetilde{O}(V^\star(s_0)DS\sqrt{AV^\star(s_0)K})$ of UCRL2 derived in Sect. 3. While in this restricted setting UCRL2 performs better when $s_0$ is a privileged starting state to reach $\overline{s}$ compared to the rest of states in $\mathcal{S}$, UC-SSP yields an improvement over UCRL2 whenever $V^\star(s_0) \geq D^{1/3}$. Our experiments in App. J illustrate that UC-SSP suffers smaller regret than UCRL2 in a gridworld with uniform costs, showcasing that UC-SSP manages to better adapt to the goal-oriented structure of the problem.

**Loop-free SSP.** Let us assume that there exists a *known* upper bound $H$ on the hitting time of *any* policy. Then a slight variation of the finite-horizon algorithm UCBVI (Azar et al., 2017) can be applied. While its bound would scale as $\widetilde{O}(\sqrt{HSAT})$ and showcase an improved $\sqrt{S}$-dependency, it would regrettably scale with $\sqrt{H}$ which may be much larger than the $D$ factor appearing in Thm. 2 as soon as the hitting times $\tau_\pi$ differ significantly across policies $\pi$. Moreover, UC-SSP does not require the prior knowledge of $H$, as opposed to UCBVI or any other existing algorithm in the finite-horizon or loop-free setting.

The analysis of UC-SSP reveals the crucial role of the pivot horizon in shaping the behavior and performance of the algorithm. In the uniform-cost case, $\text{EVI}_{\text{SSP}}$ and standard EVI used in UCRL2 both converge to the same policy. The main difference between the two algorithms consists in the stopping criterion for the execution of the optimistic policy. While UCRL2 applies a generic doubling scheme (i.e., an internal episode is terminated when the number of samples is doubled in at least a state-action pair), UC-SSP leverages the episodic nature of the SSP problem and sets a pivot horizon such that the current policy should successfully terminate with high (optimistic) probability. In the loop-free setting, UCBVI picks a single policy per episode and waits until termination. While all policies are guaranteed to terminate in finite time, the length of the episode may still be very long. On the other hand, UC-SSP goes through different policies within each episode whenever they are taking *too long* to reach the goal state.

### 4.3. Proof Sketch of Thm. 2

As explained in Sect. 2, tackling the general SSP problem requires introducing the novel notion of SSP-regret. It can neither be managed by a step-by-step comparison between the algorithmic and optimal performances as in infinite-horizon, nor by an episode-by-episode comparison as in finite-horizon. We thus need to derive a new analysis to handle the specificities of the SSP-regret.

Denoting by $T_K$ the total number of steps at the end of episode $K$, we decompose $T_K = T_{K,1} + T_{K,2}$, with $T_{K,1}$ (resp. $T_{K,2}$) the total time during attempts in phase ① (resp. phase ②). We introduce the *truncated* regret

$$\mathcal{W}_K := \sum_{k=1}^{K} \left[ \Big( \sum_{h=1}^{H_{k,0}} c(s_{k,h}, \widetilde{\pi}_{k,0}(s_{k,h})) \Big) - V^\star(s_0) \right], \quad (6)$$

which is obtained by considering the cumulative cost up to $H_{k,0}$ steps rather than for the actual duration of each attempt in phase ①. By assigning a regret of $c_{\max}$ to each step in phase ②, we can then decompose the regret as

$$\Delta(\text{UC-SSP}, K) \leq \mathcal{W}_K + c_{\max} T_{K,2}. \quad (7)$$

This decomposition directly justifies the different nature of the two phases employed by UC-SSP. While phase ① directly tries to minimize $\mathcal{W}_K$, phase ② only needs to keep $T_{K,2}$ under control, which requires executing policies that reach the goal state as quickly as possible.

**Bound on $\mathcal{W}_K$.** We first bound $\mathcal{W}_K$ by drawing inspiration from techniques in the finite-horizon setting (see e.g., Azar et al., 2017), by successively unrolling the Bellman operator to get a telescopic sum which can be bounded using the Azuma-Hoeffding inequality and a pigeonhole principle.

**Lemma 5.** *Introduce* $\Omega_K := \max_{k \in [K]} H_{k,0}$. *With probability at least* $1 - \delta$,

$$\mathcal{W}_K = O\left( c_{\max} DS \sqrt{A \Omega_K K \log\left( \frac{\Omega_K K}{\delta} \right)} \right).$$

**Bound on $\Omega_K$.** On the one hand, since $\mathcal{W}_K$ directly scales with $\sqrt{\Omega_K}$, we must ensure that the lengths of attempts in phase ① are not too long. Ideally, we would set them as relatively tight upper bounds of $V^\star(s_0)$ or $D$, yet these are critically *unknown*. Instead, in Eq. (5) we tune the lengths $H_{k,0}$ depending on optimistic quantities (which can be easily computed at the start of each attempt), and prove in the following lemma that they crucially scale as $\widetilde{O}(D)$.

**Lemma 6.** *Under the event* $\mathcal{E}$,

$$\Omega_K \leq \left\lceil 6 \frac{c_{\max}}{c_{\min}} D \log(2\sqrt{K}) \right\rceil.$$

*Proof sketch.* Consider a state $y \in \mathcal{S}$ such that

$$\|(\widetilde{Q}_{k,0})^{H_{k,0}-2}\|_\infty = \mathbb{1}_y^\top (\widetilde{Q}_{k,0})^{H_{k,0}-2}\mathbb{1}.$$

From Lem. 1, the above is equal to $\mathbb{P}(\widetilde{\tau}_{k,0}(y) \geq H_{k,0}-1)$. To bound it, we apply a corollary of Markov's inequality

$$\mathbb{P}(\widetilde{\tau}_{k,0}(y) \geq H_{k,0}-1) \leq \frac{\mathbb{E}[(\widetilde{\tau}_{k,0})^r]}{(H_{k,0}-1)^r},$$

for a carefully chosen exponent $r := \lceil \log(2\sqrt{k}) \rceil \geq 1$. We then prove that $\widetilde{\tau}_{k,0}$ follows a discrete PH distribution that satisfies $\mathbb{E}[\widetilde{\tau}_{k,0}(s)] \leq \frac{2c_{\max}D}{c_{\min}}$ for all $s \in \mathcal{S}$. This leads us to derive an upper bound on the $r$-th moment of any hitting time distribution with bounded expectation starting from any state (Lem. 15, App. E, which may be of independent interest). Applying this result to $\widetilde{\tau}_{k,0}$ yields

$$\mathbb{E}[(\widetilde{\tau}_{k,0})^r] \leq 2\left(r\frac{2c_{\max}D}{c_{\min}}\right)^r,$$

which gives on the one hand

$$\|(\widetilde{Q}_{k,0})^{H_{k,0}-2}\|_\infty \leq \frac{2\left(r\frac{2c_{\max}D}{c_{\min}}\right)^r}{(H_{k,0}-1)^r}.$$

On the other hand, the choice of $H_{k,0}$ in Eq. (5) entails that

$$\frac{1}{\sqrt{k}} < \|(\widetilde{Q}_{k,0})^{H_{k,0}-2}\|_\infty.$$

Combining the two previous inequalities finally provides the desired upper bound on $H_{k,0}$. □

**Bound on $T_{K,2}$.** On the other hand, since $T_{K,2}$ increases with the number of attempts in phase ②, we must ensure that there are not too many of such attempts and that their lengths can be adequately controlled. In light of this and leveraging the way the length $H_{k,0}$ is constructed (Eq. 5), we bound the number of failed attempts in phase ① up to episode $K$, which we denote by $F_K$.

**Lemma 7.** *With probability at least* $1-\delta$,

$$F_K \leq 2\sqrt{K} + 2\sqrt{2\Omega_K K \log\left(\frac{2(\Omega_K K)^2}{\delta}\right)} + 4S\sqrt{8A\Omega_K K \log\left(\frac{2A\Omega_K K}{\delta}\right)}.$$

*Proof sketch.* We write $F_K = F'_K + F''_K$ with $F'_K := \sum_{k=1}^K \mathbb{P}(\widetilde{\tau}_{k,0}(s_0) > H_{k,0})$ and $F''_K := \sum_{k=1}^K \left[\mathbb{1}_{\{\tau_{k,0}(s_0) > H_{k,0}\}} - \mathbb{P}(\widetilde{\tau}_{k,0}(s_0) > H_{k,0})\right]$. A martingale argument and the pigeonhole principle bound $F''_K$, while the choice of $H_{k,0}$ controls each summand of $F'_K$. □

Equipped with Lem. 7, we proceed in bounding the total duration of the attempts in phase ②.

**Lemma 8.** *With probability at least* $1-\delta$,

$$T_{K,2} = \widetilde{O}\left(DS\sqrt{\frac{c_{\max}}{c_{\min}}ADK} + S^2AD^2\right).$$

Putting everything together, we obtain Thm. 2 by plugging Lem. 5, 6 and 8 into Eq. (7). Note that while the regret decomposition in the two-phase process (Eq. 7) has the advantage of making the analysis intuitive and modular, it renders Bernstein techniques less effective in capturing low-variance deviations, as opposed to the analysis of UCBVI and UCRL2B (Fruit et al., 2020) which shave off a term of $\sqrt{H}$ or $\sqrt{D_\infty}$ for large enough time steps in the finite- and infinite-horizon settings, respectively.

## 5. Relaxation of Assumptions

Although Asm. 1 and 2 seem natural in the SSP problem, we design variants of UC-SSP that can handle dead-end states and/or zero costs. We defer to App. I the complete analysis.

**Relaxation of Asm. 2 ($D = +\infty$).** If $M$ is non-SSP-communicating, there exists at least one (possibly unknown) *dead-end* state from which reaching the goal $\overline{s}$ is impossible. This implies that $\text{EVI}_{\text{SSP}}$, which operates on the entire state space $\mathcal{S}$, fails to converge since the values at dead-end states are infinite. To tackle this problem, we assume that the agent has prior knowledge on an upper bound $J \geq V^\star(s_0)$ and that it has at any time step the "resetting" ability to transition with probability 1 to $s_0$ with a cost of $J$ (to prevent it from getting stuck). Equipped with these two assumptions, by optimizing a value function that is *truncated* at $J$ (Kolobov et al., 2012), we prove that a variant of UC-SSP achieves a regret guarantee identical to Thm. 2 except that the infinite term $D$ is replaced by $J$ (see Lem. 16, App. I.2).

**Relaxation of Asm. 1 ($c_{\min} = 0$).** Under the existence of zero costs, the optimal policy is not even guaranteed to be proper (Bertsekas, 2012). We thus change the definition of SSP-regret and compare to the best *proper* policy, by considering as optimal comparator the quantity $\min_{\pi \in \Pi^{\text{PSD}}} V^\pi$ instead of $\min_{\pi \in \Pi^{\text{SD}}} V^\pi$. We observe that having $c_{\min} = 0$ renders the bound on $\Omega_K$ of Lem. 6 vacuous. To circumvent this issue, we introduce an additive perturbation $\eta_{k,0} > 0$ to the cost of each transition in the *optimistic model* of each attempt $(k, 0)$. Our resulting variant of UC-SSP achieves a $\widetilde{O}(K^{2/3})$ regret bound (see Lem. 18, App. I.3 for the complete bound). The difference in rate ($K^{2/3}$ vs. $\sqrt{K}$) compared to Thm. 2 stems from the fact that our procedure of offsetting the costs introduces a bias, which we minimize with the choice of perturbation $\eta_{k,0} = 1/k^{1/3}$. Note that the later work of (Cohen et al., 2020) devises an algorithm with a Bernstein-based analysis that achieves a $\sqrt{K}$-rate in the case $c_{\min} = 0$.

## 6. Conclusion and Extensions

Although it encompasses numerous goal-oriented RL problems, the setting of episodic RL under its general SSP formulation had until now been neglected by the theoretical literature of RL, or had been studied under the strong, loop-free restriction on the MDP structure. Our key contribution is the design and analysis of UC-SSP, the first no-regret algorithm in the challenging setting of goal-oriented RL. Our analysis carefully combines existing techniques from the related settings of finite-horizon and infinite-horizon RL, as well as introduces refined ingredients to address the novel trade-off between minimizing costs and reaching the goal state. Interesting directions for further investigation include *(1)* designing a model-free algorithm for exploration in SSP, and *(2)* tackling SSP in the setting of linear function approximation.

## References


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

Bertsekas, D. *Dynamic programming and optimal control*, volume 2. Athena scientific Belmont, MA, 2012.

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.

Brémaud, P. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

Canfield, E. R. and Pomerance, C. On the problem of uniqueness for the maximum Stirling number(s) of the second kind. *INTEGERS: Electronic Journal of Combinatorial Number Theory*, 2(A01):2, 2002.

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. Near optimal exploration-exploitation in non-communicating Markov decision processes. In *Advances in Neural Information Processing Systems*, pp. 2994–3004, 2018a.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1573–1581, 2018b.

Fruit, R., Pirotta, M., and Lazaric, A. Improved analysis of UCRL2 with empirical bernstein inequality. *CoRR*, abs/2007.05456, 2020.

Guillot, M. and Stauffer, G. The stochastic shortest path problem: a polyhedral combinatorics perspective. *European Journal of Operational Research*, 285(1):148–158, 2020.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. The online shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(Oct):2369–2403, 2007.

Hansen, E. A. Suboptimality bounds for stochastic shortest path problems. *arXiv preprint arXiv:1202.3729*, 2012.

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

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020.

Joarder, A. H. and Mahmood, M. An inductive derivation of Stirling numbers of the second kind and their applications in statistics. 1997.

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.

Kolobov, A., Mausam, and Weld, D. S. A theory of goal-oriented MDPs with dead ends. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 438–447. AUAI Press, 2012.

Latouche, G. and Ramaswami, V. *Introduction to matrix analytic methods in stochastic modeling*. SIAM, 1999.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pp. 231–243. Citeseer, 2010.

Neu, G., Gyorgy, A., and Szepesvári, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pp. 805–813, 2012.

Puterman, M. L. *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486, 2019a.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019b.

Schweitzer, P. J. On undiscounted Markovian decision processes with compact action spaces. *RAIRO-Operations Research*, 19(1):71–86, 1985.

Talebi, M. S., Zou, Z., Combes, R., Proutiere, A., and Johansson, M. Stochastic online shortest path routing: The value of feedback. *IEEE Transactions on Automatic Control*, 63(4):915–930, 2017.

Teichteil-Königsbuch, F. Stochastic safest and shortest path problems. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

Toromanoff, M., Wirbel, E., and Moutarde, F. Is deep reinforcement learning really superhuman on Atari? *arXiv preprint arXiv:1908.04683*, 2019.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312, 2019.

Zimin, A. and Neu, G. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pp. 1583–1591, 2013.

## A. Proof of Lem. 1, 2 and 4

*Proof of Lem. 1.* Asm. 2 implies that there exists at least one proper policy (i.e., $\Pi^{\text{PSD}} \neq \emptyset$), and Asm. 1 implies that for every non-proper policy $\pi$, the corresponding value function $V^\pi(s)$ is $+\infty$ for at least one state $s \in \mathcal{S}$. The rest follows from Bertsekas (2012, Sect. 3.2). □

*Proof of Lem. 2.* From the definition of the infinity norm and Asm. 1 and 2, we have

$$\|V^\star\|_\infty = \max_{s\in\mathcal{S}} \min_{\pi\in\Pi^{\text{SD}}} \mathbb{E}\left[\sum_{t=1}^{\tau_\pi(s)} c(s_t,\pi(s_t)) \,\middle|\, s\right] \leq c_{\max} \max_{s\in\mathcal{S}} \min_{\pi\in\Pi^{\text{SD}}} \mathbb{E}[\tau_\pi(s)] = c_{\max} D.$$

□

*Proof of Lem. 4.* The first inequality comes from the chosen stopping condition. As for the second, since we consider the initial vector $v^{(0)} = 0$, we know that $v^{(0)} \leq \widetilde{V}^\star_{k,j}$ with $\widetilde{V}^\star_{k,j} = \widetilde{\mathcal{L}}_{k,j}\widetilde{V}^\star_{k,j}$. By monotonicity of the operator $\widetilde{\mathcal{L}}_{k,j}$ (Puterman, 2014; Bertsekas, 2012) we obtain $\widetilde{v}_{k,j} \leq \widetilde{V}^\star_{k,j}$. If $M \in \mathcal{M}_{k,j}$ and $j = 0$, then $\widetilde{V}^\star_{k,j} \leq V^\star$. If $M \in \mathcal{M}_{k,j}$ and $j \geq 1$, then all costs are equal to 1 so the optimal value function is $\min_\pi \mathbb{E}(\tau_\pi)$ and hence $\widetilde{V}^\star_{k,j} \leq \min_\pi \mathbb{E}(\tau_\pi)$. □

## B. Proof of Thm. 1

Recall that we introduce the MDP $M_\infty := \langle \mathcal{S}', \mathcal{A}, r_\infty, p_\infty, s_0\rangle$, with reward $r_\infty = \mathbb{1}_{\overline{s}}$ and $p_\infty(\cdot \,|\, s,a) = p(\cdot \,|\, s,a)$ for $s \neq \overline{s}$ and $p_\infty(\cdot \,|\, \overline{s},a) = \mathbb{1}_{s_0}$ for all $a$. The SSP problem with uniform costs boils down to minimizing the expected hitting time of the goal state, which according to the following lemma is equivalent to maximizing the long-term average reward (or gain) in $M_\infty$. Recall that for any policy $\pi \in \Pi^{\text{SD}}$, its gain $\rho_\pi(s)$ starting from any $s \in \mathcal{S}$ is defined as

$$\rho_\pi(s) := \lim_{T\to+\infty} \mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=1}^{T} r_\infty(s_t,\pi(s_t)) \,\middle|\, s\right].$$

**Lemma 9.** *Let $\pi_\infty \in \arg\max_\pi \rho_\pi(s)$. Then $\pi_\infty$ is optimal in the SSP sense and its constant gain $\rho_\infty$ verifies*

$$\rho_\infty = \frac{1}{V^\star(s_0)+1}.$$

*Proof.* Let $\pi$ be a policy such that $\overline{s}$ is reachable from $s_0$. Denote by $\mathcal{S}_\pi$ the set of communicating states for policy $\pi$ in $M_\infty$. Then the underlying Markov chain (restricted to $\mathcal{S}_\pi$) is irreducible with a finite number of states and is thus recurrent positive (see e.g., Brémaud, 2013, Thm. 3.3). Denoting by $\mu_\pi$ its unique stationary distribution, we have almost surely that

$$\rho_\pi(s) = \lim_{T\to+\infty} \mathbb{E}_\pi\left[\frac{\sum_{t=1}^T r_t}{T}\right] = \lim_{T\to+\infty} \mathbb{E}_\pi\left[\frac{\sum_{t=1}^T \mathbb{1}_{\{s_t=\overline{s}\}}}{T}\right] \overset{\text{(a)}}{=} \sum_{s\in\mathcal{S}_\pi} \mathbb{1}_{\{s=\overline{s}\}}\mu_\pi(s) \overset{\text{(b)}}{=} \frac{1}{1+\mathbb{E}[\tau_\pi(s_0)]},$$

where (a) comes from the Ergodic Theorem for Markov Chains (see e.g., Brémaud, 2013, Thm. 4.1) and (b) uses the fact that $1/\mu_\pi(\overline{s})$ corresponds to the mean return time in state $\overline{s}$, i.e., the expected time to reach $\overline{s}$ starting from $\overline{s}$. We conclude with the fact that $V^\pi(s_0) = \mathbb{E}[\tau_\pi(s_0)]$. □

Hence, we can prove that UCRL2 satisfies the following SSP-regret bound.

**Lemma 10.** *Under Asm. 3, with probability at least $1-\delta$, for any $K \geq 1$,*

$$\Delta(\text{UCRL2}, K) \leq 34(V^\star(s_0)+1)DS\sqrt{AT_K \log\left(\frac{T_K}{\delta}\right)},$$

*where $T_K = \sum_{k=1}^K \tau_k(s_0)$.*

*Proof.* Using the fact that $K = \sum_{t=1}^{T_K} \mathbb{1}_{\{s_t=\overline{s}\}}$, the SSP-regret can be written as

$$\Delta(\mathfrak{A}, K) = \sum_{k=1}^{K} \left[ \sum_{t=1}^{\tau_k} \mathbb{1}_{\{s_t \neq \overline{s}\}} - V^\star(s_0) \right] = T_K - K - V^\star(s_0)K = T_K - (V^\star(s_0)+1)K.$$

For any $T \geq 1$ denote by $\Delta_\infty(\mathfrak{A}, T, M_\infty)$ the (reward-based) infinite-horizon total regret of algorithm $\mathfrak{A}$ after $T$ steps in $M_\infty$, i.e., $\Delta_\infty(\mathfrak{A}, T, M_\infty) = T\rho^\dagger - \sum_{t=1}^{T} r_t$ where $\rho^\dagger := \max_\pi \rho_\pi(s)$ for all $s \in \mathcal{S}$. From Lem. 9 we have $\rho^\dagger = \rho_\infty$. Moreover, since the rewards satisfy $r_\infty = \mathbb{1}_{\overline{s}}$, we have $\sum_{t=1}^{T_K} r_t = K$. Putting everything together yields

$$\Delta(\mathfrak{A}, K) = T_K - (V^\star(s_0)+1)K = (V^\star(s_0)+1)(T_K\rho^\dagger - K) = (V^\star(s_0)+1)\Delta_\infty(\mathfrak{A}, T_K, M_\infty).$$

Note that $M_\infty$ is weakly-communicating, where its communicating set of states corresponds to all the states in $\mathcal{S}'$ that are accessible from $s_0$ with non-zero probability. Although it is weakly-communicating, the specific reward structure, combined with the fact that rewards are necessarily known (since we consider the uniform-cost SSP setting and since the goal state $\overline{s}$ is assumed to be known), allows to run UCRL2 on this problem (see the Remark at the end of App. B for more detail).

Technically, EVI is guaranteed to converge since the associated extended MDP is weakly-communicating and by Puterman (2014) it is sufficient for convergence of value iteration, see e.g., Puterman (2014, Chap. 9) for finite action space or Schweitzer (1985, Thm. 1) for compact spaces.

From Jaksch et al. (2010, Thm. 2) and using the anytime nature of UCRL2, we have with probability at least $1-\delta$ for any $T > 1$,

$$\Delta_\infty(\text{UCRL2}, T, M_\infty) \leq 34 D_\infty S \sqrt{AT \log(\frac{T}{\delta})},$$

where $D_\infty := \max_{s \neq s' \in \mathcal{S}'} \min_{\pi \in \Pi^{SD}(M_\infty)} \mathbb{E}[\tau_\pi(s \to s')]$ is the diameter of $M_\infty$. However, this bound may be vacuous since it depends on $D_\infty$ which may be equal to $+\infty$. By slightly changing the analysis of this result we can obtain an improved dependency on the SSP-diameter $D$. In particular it is sufficient to prove that for any UCRL2 episode $k$ and for any iteration $i$ of the optimal extended Bellman operator $L_{\mathcal{M}_k}$ (with $h_0 = 0$ and $h_i = (L_{\mathcal{M}_k})^i h_0$), we have that $\text{sp}(h_i) \leq D$ instead of the conventional upper bound $D_\infty$. The remainder of the proof shows this result. It is straightforward that $h_i(\overline{s}) \geq h_i(s)$ for any $s \in \mathcal{S}$ (this can be proved by recurrence on $i$ using the definition of $h_i = L_{\mathcal{M}_k} h_{i-1}$ and the fact that the reward in $\mathcal{M}_k$ is equal to $\mathbb{1}_{\overline{s}}$). Introduce $\underline{s} \in \arg\min_s h_i(s)$ and $\varphi_{\widetilde{M}}(\underline{s} \to \overline{s})$ the minimum expected shortest path from $\underline{s}$ to $\overline{s}$ in any MDP $\widetilde{M}$. Then from Lem. 12 we have $\text{sp}(h_i) = h_i(\overline{s}) - h_i(\underline{s}) \leq \varphi_{\mathcal{M}_k}(\underline{s} \to \overline{s})$. Since the "true" MDP $M_\infty \in \mathcal{M}_k$, we have $\varphi_{\mathcal{M}_k}(\underline{s} \to \overline{s}) \leq \varphi_{M_\infty}(\underline{s} \to \overline{s})$. Furthermore, $\varphi_{M_\infty}(\underline{s} \to \overline{s}) = \varphi_M(\underline{s} \to \overline{s}) \leq D$. Putting everything together, we obtain that $\text{sp}(h_i) \leq D$. We thus have with probability at least $1-\delta$ for any $T > 1$,

$$\Delta_\infty(\text{UCRL2}, T, M_\infty) \leq 34 D S \sqrt{AT \log(\frac{T}{\delta})}.$$

□

While we would like to assess the dependency of the regret on the number of episodes $K$ (as in the finite-horizon case), the bound in Lem. 10 contains the random total number of steps $T_K$ needed to reach $K$ episodes. In light of this, we derive in the following lemma an upper bound of $T_K$ that depends on the quantity of interest $K$. Plugging it in Lem. 10 finally yields the result of Thm. 1.

**Lemma 11.** *Under the same event for which Lem. 10 holds with probability at least $1-\delta$, we have*

$$T_K \leq 2(V^\star(s_0)+1)K + \widetilde{O}\left( V^\star(s_0)^2 D^2 S^2 A \log\left(\frac{1}{\delta}\right)\right).$$

*Proof.* With probability at least $1-\delta$, we have from the proof of Lem. 10 that

$$T_K - (V^\star(s_0)+1)K \leq 34(V^\star(s_0)+1)DS\sqrt{AT_K \log\left(\frac{T_K}{\delta}\right)}.$$

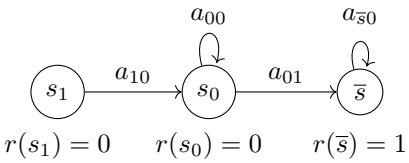


*Figure 2.* A toy example of SSP-communicating ($D = 2$) reward-based MDP.

This implies that

$$T_K \leq 2(V^\star(s_0)+1)K \underbrace{-T_K + 68(V^\star(s_0)+1)DS\sqrt{AT_K \log\left(\frac{T_K}{\delta}\right)}}_{:=(y)},$$

where $(y)$ can be bounded using Lem. 13 (with the constants $a_1 = 68(V^\star(s_0)+1)DS\sqrt{A}$, $a_2 = \frac{1}{\delta}$ and $a_3 = 1$) as follows

$$(y) \leq \frac{16}{9}\Big(68(V^\star(s_0)+1)DS\sqrt{A}\Big)^2 \left[\log\left(\frac{136(V^\star(s_0)+1)DS\sqrt{Ae}}{\sqrt{\delta}}\right)\right]^2.$$

□

**Lemma 12.** *Consider an (extended) MDP $\widetilde{M}$ and define $L_{\widetilde{M}}$ as the associated optimal (extended) Bellman operator (of undiscounted value iteration). Given $h_0 = 0$ and $h_i = (L_{\widetilde{M}})^i h_0$ we have that*

$$\forall s_1, s_2 \in \mathcal{S}', h_i(s_2) - h_i(s_1) \leq r_{\max}\, \varphi_{\widetilde{M}}(s_1 \to s_2),$$

*where $\varphi_{\widetilde{M}}(s_1 \to s_2)$ is the minimum expected shortest path from $s_1$ to $s_2$ in $\widetilde{M}$ and $r_{\max}$ is the maximal state-action reward.*

*Proof.* The proof follows from the application of the argument of Jaksch et al. (2010, Sect. 4.3.1). □

**Lemma 13** (Kazerouni et al., 2017, Lem. 8)**.** *For any $x \geq 2$ and $a_1, a_2, a_3 > 0$, the following holds*

$$-a_3 x + a_1\sqrt{x}\log(a_2 x) \leq \frac{16a_1^2}{9a_3}\left[\log\left(\frac{2a_1\sqrt{a_2}e}{a_3}\right)\right]^2.$$

**Remark.** Consider the reward-based SSP $M$ in Fig. 2. $M$ is SSP-communicating while the associated MDP $M_\infty$ is weakly-communicating since $s_1$ is transient under every policy. There are just two possible deterministic policies: $\pi_0(s_0) = a_{00}$ and $\pi_1(s_0) = a_{01}$. If rewards are unknown, UCRL2 will periodically alternate between policy $\pi_0$ and $\pi_1$ without converging to any of the two. This is due to the fact that, in the set of plausible MDPs $\mathcal{M}_k$ there will always be (i.e., $\forall k > 0$) an MDP with arbitrarily small but non-zero transition probability $\tilde{p}$ to state $s_1$, where, due to maximum uncertainty, there will be a self loop with probability 1 and reward $r_{\max}$ (since $N_k(s_1, a_{10}) \in \{0, 1\}$ depending on the initial state for any $k$). The probability $\tilde{p}$ will be sometimes higher for action $a_{00}$ and sometimes for $a_{01}$ depending on the counter $N_k$. This is why UCRL2 will never converge. However, if the rewards are known (which is always the case under Asm. 3 and as long as the goal state $\overline{s}$ is known), after a burn-in phase, it will be clear to UCRL2 that action $a_{00}$ is suboptimal. Even if there is probability $\tilde{p} > 0$ to go to $s_1$, in $s_1$ the optimistic behaviour will be to go to $\overline{s}$ since it is the only one to provide reward. However, this imagined policy is suboptimal since it has an additional step and thus UCRL2 will select $\pi_1$. Note that while it is possible to make the MDP stochastic, this will lead to a longer burn-in phase but will not change the behaviour of UCRL2 in the long run.

## C. Proof of Lem. 3

The proof is almost identical to the proof of Fruit et al. (2020, Thm. 10) and we report it below for completeness.

Recall that we define $\mathcal{M}_{k,j} := \{\langle \mathcal{S}, \mathcal{A}, c, \widetilde{p}\rangle \mid \widetilde{p} \in B_{k,j}\}$ to be the extended MDP defined by the confidence interval $B_{k,j} := \{\widetilde{p} \in \mathcal{C} \mid \widetilde{p}(\cdot|\overline{s}, a) = \mathbb{1}_{\overline{s}}$ and $\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \|\widetilde{p}(\cdot|s,a) - \widehat{p}_{k,j}(\cdot|s,a)\|_1 \le \beta_{k,j}(s,a)\}$, with $\mathcal{C}$ the $S'$-dimensional simplex and

$$\beta_{k,j}(s,a) := \sqrt{\frac{8S\log\left(\frac{2AN^+_{k,j}(s,a)}{\delta}\right)}{N^+_{k,j}(s,a)}}.$$

Furthermore we introduce $B_{k,j}(s,a) := \{\widetilde{p} \in \mathcal{C} : \|\widetilde{p}(\cdot|s,a) - \widehat{p}_{k,j}(\cdot|s,a)\|_1 \le \beta_{k,j}(s,a)\}$ (and similarly for $B_{k,j}(s,a,s')$). We want to bound the probability of event $\mathcal{E}^{\mathcal{C}} := \bigcup_{k=1}^{+\infty}\bigcup_{j=1}^{J_k}\{M \notin \mathcal{M}_{k,j}\}$. As explained by Lattimore & Szepesvári (2020, Chap. 5), when $(s,a)$ is visited for the $n$-th times, the next state that we observe is the $n$-th element of an infinite sequence of i.i.d. r.v. lying in $\mathcal{S}'$ with probability density function $p(\cdot|s,a)$. In UCRL2 (Jaksch et al., 2010), the sample means $\widehat{p}_{k,j}$ and the confidence intervals $B_{k,j}$ are defined as depending on $(k,j)$. Actually, these quantities depend only on the first $N_{k,j}(s,a)$ elements of the infinite i.i.d. sequences that we just mentioned. For the rest of the proof, we will therefore slightly change our notations and denote by $\widehat{p}_n(s'|s,a)$ and $B_n(s'|s,a)$ the sample means and confidence intervals after the first $n$ visits in $(s,a)$. Thus, the r.v. that we denoted by $\widehat{p}_{k,j}$ actually corresponds to $\widehat{p}_{N_{k,j}(s,a)}$ with our new notation (and similarly for $B_{k,j}$). This change of notation will make the proof easier.

If $M \notin \mathcal{M}_{k,j}$, then there exists a $k \ge 1$ and $j \ge 0$ s.t. $p(\cdot|s,a) \notin B_{N_{k,j}(s,a)}(s,a)$ for at least one $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}'$. This means that there exists at least one value $n \ge 0$ s.t. $p(s'|s,a) \notin B_n(s,a,s')$. Consequently we have the following inclusion

$$\mathcal{E}^{\mathcal{C}} \subseteq \bigcup_{s,a}\bigcup_{n=0}^{+\infty}\{p(\cdot|s,a) \notin B_n(s,a)\}.$$

Using Boole's inequality we have

$$\mathbb{P}(\mathcal{E}^{\mathcal{C}}) \le \sum_{s,a}\sum_{n=0}^{+\infty}\mathbb{P}(p(\cdot|s,a) \notin B_n(s,a)).$$

Let us fix a tuple $(s,a) \in \mathcal{S}\times\mathcal{A}$ and define for all $n \ge 0$

$$\epsilon_n(s,a) := \sqrt{\frac{2\log((2^{S'}-2)5SA(n^+)^2/\delta)}{n^+}},$$

where $n^+ := \max\{n,1\}$. Since $S' = S+1 \le 2S$, it is immediate to verify that almost surely, $\epsilon_n(s,a) \le \beta_n(s,a)$. Using Weissman's inequality (Weissman et al., 2003; Jaksch et al., 2010) we have that for all $n \ge 1$

$$\mathbb{P}(\|p(\cdot|s,a) - \widehat{p}_n(\cdot|s,a)\|_1 \ge \beta_n(s,a)) \le \mathbb{P}(\|p(\cdot|s,a) - \widehat{p}_n(\cdot|s,a)\|_1 \ge \epsilon_n(s,a)) \le \frac{\delta}{5n^2SA}.$$

Note that when $n = 0$ (i.e., when there has not been any observation of $(s,a)$), $\epsilon_0(s,a) \ge 2$ so $\mathbb{P}(\|p(\cdot|s,a) - \widehat{p}_0(\cdot|s,a)\|_1 \ge \epsilon_0(s,a)) = 0$ by definition. As a result, we have that for all $n \ge 1$

$$\mathbb{P}(p(\cdot|s,a) \notin B_n(s,a)) \le \frac{\delta}{5n^2SA},$$

and this probability is equal to 0 if $n = 0$. Finally we obtain

$$\mathbb{P}(\exists k \ge 1, \exists j \in [0, J_k], \text{ s.t. } M \notin \mathcal{M}_{k,j}) \le \sum_{s,a}\left(0 + \sum_{n=1}^{+\infty}\frac{\delta}{5n^2SA}\right) = \frac{\pi^2\delta}{30} \le \frac{\delta}{3},$$

which concludes the proof.

## D. Proof of Lem. 5

For notational ease, in Sect. D we adopt the notation $H_k := H_{k,0}$, $\widetilde{\pi}_k := \widetilde{\pi}_{k,0}$, $\varepsilon_k := \varepsilon_{k,0}$ (i.e., we remove the subscript 0). Furthermore, for any $k \in [K]$ and $h \in [H_k]$, we denote by $s_{k,h}$ the state visited in the $h$-th step of episode $k$.

Assume from now on that the event $\mathcal{E}$ holds. From Lem. 4 we have

$$\mathcal{W}_K = \sum_{k=1}^{K}\left[\left(\sum_{h=1}^{H_k} c(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\right) - V^\star(s_0)\right] \le \sum_{k=1}^{K}\left[\left(\sum_{h=1}^{H_k} c(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\right) - \widetilde{v}_k(s_0)\right] = \sum_{k=1}^{K}\Theta_{k,1}(s_{k,1}),$$

where $s_{k,1} := s_0$, and for any $k \in [K]$ and $h \in [H_k]$, we introduce

$$\Theta_{k,h}(s_{k,h}) := \sum_{t=h}^{H_k} c(s_{k,t}, \widetilde{\pi}_k(s_{k,t})) - \widetilde{v}_k(s_{k,h}).$$

For any $h \in [H_k - 1]$, we introduce

$$\Phi_{k,h} := \widetilde{v}_k(s_{k,h+1}) - \sum_{y\in\mathcal{S}} p(y \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\widetilde{v}_k(y).$$

We then have

$$\begin{aligned}
\Theta_{k,h}(s_{k,h}) &= \sum_{t=h}^{H_k} c(s_{k,t}, \widetilde{\pi}_k(s_{k,t})) - \widetilde{v}_k(s_{k,h}) \\
&\le \sum_{t=h}^{H_k} c(s_{k,t}, \widetilde{\pi}_k(s_{k,t})) - \widetilde{\mathcal{L}}_k\widetilde{v}_k(s_{k,h}) + \varepsilon_k \\
&\overset{(a)}{=} \sum_{t=h}^{H_k} c(s_{k,t}, \widetilde{\pi}_k(s_{k,t})) - c(s_{k,h}, \widetilde{\pi}_k(s_{k,h})) - \sum_{y\in\mathcal{S}} \widetilde{p}_k(y \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\widetilde{v}_k(y) + \varepsilon_k \\
&= \sum_{t=h+1}^{H_k} c(s_{k,t}, \widetilde{\pi}_k(s_{k,t})) - \sum_{y\in\mathcal{S}}[\widetilde{p}_k(y \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h})) - p(y \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h})) + p(y \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h}))]\widetilde{v}_k(y) + \varepsilon_k \\
&\overset{(b)}{\le} \sum_{t=h+1}^{H_k} c(s_{k,t}, \widetilde{\pi}_k(s_{k,t})) - \sum_{y\in\mathcal{S}} p(y \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\widetilde{v}_k(y) \\
&\quad + \|p(\cdot \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h})) - \widetilde{p}_k(\cdot \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\|_1 \|\widetilde{v}_k\|_\infty + \varepsilon_k \\
&\overset{(c)}{\le} \sum_{t=h+1}^{H_k} c(s_{k,t}, \widetilde{\pi}_k(s_{k,t})) - \sum_{y\in\mathcal{S}} p(y \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\widetilde{v}_k(y) + 2\beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))c_{\max}D + \varepsilon_k \\
&= \Theta_{k,h+1}(s_{k,h+1}) + \widetilde{v}_k(s_{k,h+1}) - \sum_{y\in\mathcal{S}} p(y \mid s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\widetilde{v}_k(y) + 2\beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))c_{\max}D + \varepsilon_k \\
&= \Theta_{k,h+1}(s_{k,h+1}) + \Phi_{k,h} + 2\beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))c_{\max}D + \varepsilon_k, \qquad (8)
\end{aligned}$$

where (a) stems from the fact that $\widetilde{\pi}_k$ is the greedy policy with respect to $(\widetilde{v}_k, \varepsilon_k)$, (b) leverages that $\widetilde{v}_k \ge 0$ component-wise and (c) combines Lem. 4 and 2. Furthermore, whatever the value of $s_{k,H_k}$ we have

$$\begin{aligned}
\Theta_{k,H_k}(s_{k,H_k}) &= c(s_{k,H_k}, \widetilde{\pi}_k(s_{k,H_k})) - \widetilde{v}_k(s_{k,H_k}) \\
&\le c(s_{k,H_k}, \widetilde{\pi}_k(s_{k,H_k})) - \widetilde{\mathcal{L}}_k\widetilde{v}_k(s_{k,H_k}) + \varepsilon_k \\
&= c(s_{k,H_k}, \widetilde{\pi}_k(s_{k,H_k})) - c(s_{k,H_k}, \widetilde{\pi}_k(s_{k,H_k})) - \sum_{y\in\mathcal{S}} \widetilde{p}_k(y \mid s_{k,H_k}, \widetilde{\pi}_k(s_{k,H_k}))\underbrace{\widetilde{v}_k(y)}_{\ge 0} + \varepsilon_k \\
&\le \varepsilon_k.
\end{aligned}$$

By telescopic sum we get (using Eq. 8)

$$\begin{aligned}\Theta_{k,1}(s_{k,1}) &= \sum_{h=1}^{H_k-1} \left(\Theta_{k,h}(s_{k,h}) - \Theta_{k,h+1}(s_{k,h+1})\right) + \Theta_{k,H_k}(s_{k,H_k}) \\ &\leq \sum_{h=1}^{H_k-1} \Phi_{k,h} + 2c_{\max} D \sum_{h=1}^{H_k-1} \beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h})) + (H_k-1)\varepsilon_k + \Theta_{k,H_k}(s_{k,H_k}) \\ &\leq \sum_{h=1}^{H_k-1} \Phi_{k,h} + 2c_{\max} D \sum_{h=1}^{H_k-1} \beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h})) + H_k\varepsilon_k.\end{aligned}$$

Summing over the episode index $k$ yields

$$\sum_{k=1}^{K} \Theta_{k,1}(s_{k,1}) \leq \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H_k-1} \Phi_{k,h}}_{:=X_K} + 2c_{\max} D \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H_k-1} \beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))}_{:=Y_K} + \underbrace{\sum_{k=1}^{K} H_k\varepsilon_k}_{:=Z_K}.$$

In order to bound $X_K$, we can write

$$\begin{aligned}&\mathbb{P}\left(\sum_{k=1}^{K}\sum_{h=1}^{H_k-1} \Phi_{k,h} \geq 2c_{\max} D \sqrt{2\left(\sum_{k=1}^{K} H_k\right)\log\left(\frac{2\left(\sum_{k=1}^{K} H_k\right)^2}{\delta}\right)}\right) \\ &\leq \sum_{n=1}^{+\infty} \mathbb{P}\left(\sum_{k=1}^{K}\sum_{h=1}^{H_k} \Phi_{k,h} \geq 2c_{\max} D\sqrt{2n\log\left(\frac{2n^2}{\delta}\right)} \bigcap \sum_{k=1}^{K} H_k = n\right) \\ &\leq \sum_{n=1}^{+\infty} \mathbb{P}\left(\sum_{t=1}^{n} \widetilde{\Phi}_t \geq 2c_{\max} D\sqrt{2n\log\left(\frac{2n^2}{\delta}\right)}\right),\end{aligned}$$

where we introduce for any $t > 0$,

$$\widetilde{\Phi}_t = \begin{cases} \Phi_{\widetilde{k}_t, t-Z_t} & \text{if } t > Z_t, \\ \Phi_{\widetilde{k}_t+1,1} & \text{otherwise,}\end{cases}$$

where $\widetilde{k}_t = \max\{k \mid \sum_{k'=1}^{k} H_{k'} \leq t\}$ and $Z_t = \sum_{k'=1}^{\widetilde{k}_t-1} H_{k'} + 1$, i.e., we map a value $t$ to the double index $(k,h)$. Denote by $\mathcal{G}_q$ the history of all random events up to (and including) step $h$ of episode $k$ (i.e., $q = \sum_{k'=1}^{k-1} H_k + h$). We have $\mathbb{E}[\Phi_{k,h}|\mathcal{G}_q] = 0$ (since $\widetilde{v}_k(\bar{s}) = 0$), and furthermore the stopping time $H_k$ is selected at the beginning of episode $k$ so it is adapted w.r.t. $\mathcal{G}_q$. Hence, $(\widetilde{\Phi}_t)$ is a martingale difference sequence, such that $|\widetilde{\Phi}_t| \leq 2c_{\max} D$. For any fixed $n > 0$, we thus have from Azuma-Hoeffding's inequality that

$$\mathbb{P}\left(\sum_{t=1}^{n} \widetilde{\Phi}_t \geq 2c_{\max} D\sqrt{2n\log\left(\frac{2n^2}{\delta}\right)}\right) \leq \frac{\delta}{2n^2}.$$

As a result, from a union bound over all possible values of $n > 0$, we have with probability at least $1 - \frac{2\delta}{3}$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H_k-1} \Phi_{k,h} \leq 2c_{\max} D \sqrt{2\left(\sum_{k=1}^{K} H_k\right)\log\left(\frac{3\left(\sum_{k=1}^{K} H_k\right)^2}{\delta}\right)}. \tag{9}$$

We now proceed in bounding $Y_K$ using a pigeonhole principle. Denoting by $N^{(1)}$ the counter of samples *only* collected during attempts in phase ①, we get

$$\sum_{k=1}^{K}\sum_{h=1}^{H_k-1} \sqrt{\frac{1}{N_k^{(1)}(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))}} \leq \sum_{s,a} \sum_{n=1}^{N_K^{(1)}(s,a)} \sqrt{\frac{1}{n}} \leq \sum_{s,a} 2\sqrt{N_K^{(1)}(s,a)} \leq 2\sqrt{SA}\sqrt{\sum_{s,a} N_K^{(1)}(s,a)} \leq 2\sqrt{SAT_{K,1}}.$$

We have $N_k^+(s,a) \geq N_k^{(1)+}(s,a)$ so by applying the technical Lem. 14 (and considering that $A \geq 2$ since if $A = 1$ there is no learning problem), we get

$$\beta_k(s,a) = \sqrt{\frac{8S\log\left(\frac{2AN_k^+(s,a)}{\delta}\right)}{N_k^+(s,a)}} \leq \sqrt{\frac{8S\log\left(\frac{2AN_k^{(1)+}(s,a)}{\delta}\right)}{N_k^{(1)+}(s,a)}}.$$

Therefore we obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H_k-1}\beta_k(s_{k,h},\widetilde{\pi}_k(s_{k,h})) \leq 2S\sqrt{8AT_{K,1}\log\left(\frac{2AT_{K,1}}{\delta}\right)}. \tag{10}$$

We finally bound $Z_K$. We have for any $k \in [K]$, $H_k \leq \Omega_K$ and we select $\varepsilon_k = \frac{c_{min}}{2t_{k,0}}$, hence we have $T_{K,1} \leq \Omega_K K$ and

$$\sum_{k=1}^{K} H_k\varepsilon_k \leq \frac{c_{\min}}{2}\sum_{t=1}^{T_{K,1}}\frac{\Omega_K}{t} \leq \frac{c_{\min}}{2}\Omega_K(1+\log(\Omega_K K)).$$

Putting everything together, a union bound and Lem. 3 yields with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\left[\left(\sum_{h=1}^{H_k}c(s_{k,h},\widetilde{\pi}_k(s_{k,h}))\right) - \widetilde{v}_k(s_0)\right] \leq 4c_{\max}DS\sqrt{8AT_{K,1}\log\left(\frac{2AT_{K,1}}{\delta}\right)} + 2c_{\max}D\sqrt{2T_{K,1}\log\left(\frac{3T_{K,1}^2}{\delta}\right)} + \frac{c_{\min}}{2}\Omega_K(1+\log(\Omega_K K)).$$

**Lemma 14.** *For any constant $c \geq 4$, the function $f(x) := \sqrt{\frac{\log(cx)}{x}}$ is a non-increasing function for $x \geq 1$.*

*Proof.* Introduce the function $g(x) := f(x)^2$. We have $g'(x) = \frac{1-\log(cx)}{x^2} \leq 0$ since $x \geq 1 \geq \frac{e}{c}$. So $g$ is non-increasing, hence by composition of functions, $f = \sqrt{g}$ is also non-increasing. □

Interestingly, the bound of Lem. 5 resembles a combination of finite- and infinite-horizon guarantees. On the one hand, we have the standard dependency of finite-horizon problems on the horizon $H$ and number of episodes $K$. On the other hand, $H$ is no longer bounding the range of the value functions, which is replaced by $c_{\max}D$ as in infinite-horizon problems.

## E. Proof of Lem. 6

We start the proof of Lem. 6 by deriving a general result — which may be of independent interest — that *upper bounds the moments of any discrete PH distribution*.[6]

**Lemma 15.** *Consider an absorbing Markov Chain with state space $\mathcal{Y} \cup \{\overline{y}\}$, a single absorbing state $\overline{y}$ and $|\mathcal{Y}|$ transient states. Denote by $Q \in \mathbb{R}^{Y\times Y}$ the transition matrix within the states in $\mathcal{Y}$ and by $\tau(y) := \tau(y \to \overline{y})$ the first hitting time of state $\overline{y}$ starting from state $y$. Suppose that there exists a constant $\lambda \geq 2$ such that for any state $y \in \mathcal{Y}$, we have $\mathbb{E}[\tau(y \to \overline{y})] \leq \lambda$. Then for any $r \geq 1$ and any state $y \in \mathcal{Y}$, we have*

$$\mathbb{E}[\tau(y)^r] \leq 2(r\lambda)^r.$$

[6] Note that while there actually exists a *closed-form* expression of the moments of a continuous PH distribution (see e.g., Latouche & Ramaswami, 1999, Eq. 2.13), it does not extend to the discrete case.

*Proof.* We first leverage a closed-form expression of the *factorial moments* of discrete PH distributions. For any $r \geq 1$, denoting by $(\tau)_r$ the $r$-th factorial moment of $\tau$, i.e., $(\tau)_r := \tau(\tau-1)...(\tau-r+1)$, we have (see e.g., Latouche & Ramaswami, 1999, Eq. 2.15) that for any starting state $y \in \mathcal{Y}$,

$$\mathbb{E}[(\tau)_r(y)] = r! \mathbb{1}_y^\top (I-Q)^{-r} Q^{r-1} \mathbb{1}.$$

Recalling that the $\|\cdot\|_\infty$ (resp. $\|\cdot\|_1$) norm of a matrix is equal to its maximum absolute row (resp. column) sum, we have by Hölder's inequality, for any $j \in [r]$,

$$\begin{aligned}
\mathbb{E}[(\tau)_j(y)] &= j! \langle (\mathbb{1}_y^\top (I-Q)^{-j})^\top, Q^{j-1}\mathbb{1} \rangle \\
&\leq j! \|(\mathbb{1}_y^\top (I-Q)^{-j})^\top\|_1 \|Q^{j-1}\mathbb{1}\|_\infty \\
&= j! \|((I-Q)^{-j})^\top \mathbb{1}_y\|_1 \|Q^{j-1}\mathbb{1}\|_\infty \\
&\leq j! \|((I-Q)^{-j})^\top\|_1 \|\mathbb{1}_y\|_1 \|Q^{j-1}\|_\infty \|\mathbb{1}\|_\infty \\
&\leq j! \|(I-Q)^{-j}\|_\infty \|Q^{j-1}\|_\infty \\
&\leq j! \|(I-Q)^{-1}\|_\infty^j, \qquad (11)
\end{aligned}$$

where the last inequality uses the fact that $\|Q^{j-1}\|_\infty \leq 1$ since the matrix $Q^{j-1}$ is substochastic. There remains to upper bound the quantity $\|(I-Q)^{-1}\|_\infty$. Consider a state

$$z \in \arg\max_{y \in \mathcal{Y}} \sum_{y' \in \mathcal{Y}} (I-Q)^{-1}_{yy'}.$$

By choice of $z$ and non-negativity of the matrix $(I-Q)^{-1}$, we have

$$\begin{aligned}
\|(I-Q)^{-1}\|_\infty &= \sum_{y' \in \mathcal{Y}} |(I-Q)^{-1}_{zy'}| \\
&= \sum_{y' \in \mathcal{Y}} (I-Q)^{-1}_{zy'} \\
&= \mathbb{1}_z^\top (I-Q)^{-1} \mathbb{1} \\
&= \sum_{n=0}^{\infty} \mathbb{1}_z^\top Q^n \mathbb{1}.
\end{aligned}$$

Since $\tau(z)$ follows a discrete PH distribution, we have from Lem. 1 that

$$\mathbb{1}_z^\top Q^n \mathbb{1} = \mathbb{P}(\tau(z) > n).$$

Consequently,

$$\|(I-Q)^{-1}\|_\infty = \sum_{n=0}^{\infty} \mathbb{P}(\tau(z) > n) = \mathbb{E}[\tau(z)] \leq \lambda. \qquad (12)$$

Plugging Eq. (12) into Eq. (11) thus yields for any $y \in \mathcal{Y}$,

$$\mathbb{E}[(\tau)_j(y)] \leq j! \lambda^j.$$

Furthermore, the (raw) moment of a random variable can be expressed in terms of its factorial moments by the following formula (see e.g., Joarder & Mahmood, 1997, Eq. 3.1)

$$\mathbb{E}[\tau(y)^r] = \sum_{j=1}^{r} \left\{ {r \atop j} \right\} \mathbb{E}[(\tau)_j(y)],$$

where the curly braces denote Stirling numbers of the second kind, i.e.,

$$\left\{ {r \atop j} \right\} := \frac{1}{j!} \sum_{i=0}^{j} (-1)^{j-i} \binom{j}{i} i^r.$$

Using the upper bound (see e.g., Canfield & Pomerance, 2002, Eq. 9)

$$\left\{ r \atop j \right\} \leq \frac{j^r}{j!},$$

we obtain

$$\mathbb{E}[\tau(y)^r] \leq \sum_{j=1}^{r} j^r \lambda^j.$$

We conclude the proof of Lem. 15 with the fact that

$$\sum_{j=1}^{r} j^r \lambda^j \leq r^r \sum_{j=1}^{r} \lambda^j \leq r^r \lambda \frac{\lambda^r - 1}{\lambda - 1} \leq r^r 2\lambda^r,$$

where the last inequality holds since $\lambda \geq 2$. □

We are now ready to prove Lem. 6.

For notational ease, in Sect. E we adopt the notation $H_k := H_{k,0}$, $\widetilde{\pi}_k := \widetilde{\pi}_{k,0}$, $\varepsilon_k := \varepsilon_{k,0}$ (i.e., we remove the subscript 0).

Denote by $\mathcal{G}_{k-1}$ the history of all random events up to (and including) episode $k-1$. In this section as well as in Sect. F, we will write $\mathbb{E}\left[\mathbb{1}_{\{\tau_\pi^p(s) > H_k - 1\}} \,|\, \mathcal{G}_{k-1}\right] = \mathbb{P}(\tau_\pi^p(s) > H_k - 1)$, i.e. the probability $\mathbb{P}$ is only over the randomization of the sequence of states generated by the policy $\pi$ in the model $p$ starting from state $s$ (i.e., it is conditioned on $\mathcal{G}_{k-1}$, the policy $\pi$, the model $p$ and the starting state $s$).

Suppose that the event $\mathcal{E}$ holds and fix an episode $k \in [K]$. Denote by $\widetilde{Q} := Q_{\widetilde{\pi}_k}^{\widetilde{p}_k}$ the optimistic transition matrix within $\mathcal{S}$ of policy $\widetilde{\pi}_k$ in the transition model $\widetilde{p}_k$. Also, for any state $s \in \mathcal{S}$, denote by $\widetilde{\tau}(s) := \tau_{\widetilde{\pi}_k}^{\widetilde{p}_k}(s)$ the hitting time of $\overline{s}$ starting from $s$ following policy $\widetilde{\pi}_k$ in the transition model $\widetilde{p}_k$.

We introduce the Bellman operator $\mathcal{T}_{\varepsilon_k}^{\widetilde{\pi}_k}$ for policy $\widetilde{\pi}_k$, that verifies for any vector $v \in \mathbb{R}^S$ and state $s \in \mathcal{S}$,

$$\mathcal{T}_{\varepsilon_k}^{\widetilde{\pi}_k} v(s) := c(s, \widetilde{\pi}_k(s)) - \varepsilon_k + \sum_{y \in \mathcal{S}} \widetilde{p}_k(y \,|\, s, \widetilde{\pi}_k(s)) v(y),$$

i.e., it corresponds to the operator $\widetilde{\mathcal{L}}_k^{\widetilde{\pi}_k}$ with $\varepsilon_k$ subtracted to all the costs. Note that its costs are all positive by choice of $\varepsilon_k = \frac{c_{\min}}{2t_k}$. Combining Lem. 4 and the fact that $\widetilde{\pi}_k$ is the greedy policy w.r.t. $\widetilde{v}_k$ yields that $\widetilde{\mathcal{L}}_k^{\widetilde{\pi}_k} \widetilde{v}_k = \widetilde{\mathcal{L}}_k \widetilde{v}_k \leq \widetilde{v}_k + \varepsilon_k$. Consequently, we have the following component-wise inequality

$$\mathcal{T}_{\varepsilon_k}^{\widetilde{\pi}_k} \widetilde{v}_k \leq \widetilde{v}_k.$$

By monotonicity of the operator $\mathcal{T}_{\varepsilon_k}^{\widetilde{\pi}_k}$ (Puterman, 2014; Bertsekas, 2012), we have for all $m > 0$,

$$(\mathcal{T}_{\varepsilon_k}^{\widetilde{\pi}_k})^m \widetilde{v}_k \leq \widetilde{v}_k,$$

and hence taking the limit $m \to +\infty$ yields $\widetilde{U}_{\widetilde{\pi}_k, \varepsilon_k} \leq \widetilde{v}_k$, where $\widetilde{U}_{\widetilde{\pi}_k, \varepsilon_k}$ is defined as the value function of policy $\widetilde{\pi}_k$ in the model $\widetilde{p}_k$ with $\varepsilon_k$ subtracted to all the costs, i.e.,

$$\widetilde{U}_{\widetilde{\pi}_k, \varepsilon_k}(s) := \mathbb{E}_{\widetilde{p}_k}\left[ \sum_{t=1}^{\widetilde{\tau}(s)} (c(s_t, \widetilde{\pi}_k(s_t)) - \varepsilon_k) \;\middle|\; s_1 = s \right] = \widetilde{V}_{\widetilde{\pi}_k}(s) - \varepsilon_k \mathbb{E}[\widetilde{\tau}(s)],$$

where $\widetilde{V}_{\widetilde{\pi}_k}(s) := \mathbb{E}_{\widetilde{p}_k}\left[ \sum_{t=1}^{\widetilde{\tau}(s)} c(s_t, \widetilde{\pi}_k(s_t)) \;\middle|\; s_1 = s \right]$ is the value function of policy $\widetilde{\pi}_k$ in the model $\widetilde{p}_k$. Since $\widetilde{V}_{\widetilde{\pi}_k}(s) \geq c_{\min} \mathbb{E}[\widetilde{\tau}(s)]$, we have

$$(c_{\min} - \varepsilon_k) \mathbb{E}[\widetilde{\tau}(s)] \leq \widetilde{V}_{\widetilde{\pi}_k}(s) - \varepsilon_k \mathbb{E}[\widetilde{\tau}(s)] = \widetilde{U}_{\widetilde{\pi}_k, \varepsilon_k}(s) \leq \widetilde{v}_k(s).$$

Using successively the above inequality, the fact that $\varepsilon_k \leq \frac{c_{\min}}{2}$, Lem. 4 and 2, we obtain for any $s \in \mathcal{S}$,

$$\mathbb{E}[\widetilde{\tau}(s)] \leq \frac{\widetilde{v}_k(s)}{c_{\min} - \varepsilon_k} \leq \frac{2V^\star(s)}{c_{\min}} \leq \frac{2c_{\max}D}{c_{\min}}. \tag{13}$$

Fix any $r \geq 1$ and $s \in \mathcal{S}$. According to a corollary of Markov's inequality (since $x \mapsto x^r$ is a monotonically increasing non-negative function for the non-negative reals), we have

$$\mathbb{P}(\widetilde{\tau}(s) \geq H_k - 1) \leq \frac{\mathbb{E}[\widetilde{\tau}(s)^r]}{(H_k-1)^r}.$$

We can apply Lem. 15 to the discrete PH distribution $\widetilde{\tau}$ with the choice of $\lambda := \frac{2c_{\max}D}{c_{\min}}$ guaranteed by Eq. (13). This yields

$$\mathbb{E}[\widetilde{\tau}(s)^r] \leq 2\left(r\frac{2c_{\max}D}{c_{\min}}\right)^r.$$

Hence, we have

$$\mathbb{P}(\widetilde{\tau}(s) \geq H_k - 1) \leq \frac{2\left(r\frac{2c_{\max}D}{c_{\min}}\right)^r}{(H_k-1)^r}. \tag{14}$$

There exists $y \in \mathcal{S}$ such that

$$\|\widetilde{Q}^{H_k-2}\|_\infty = \mathbb{1}_y^\top \widetilde{Q}^{H_k-2}\mathbb{1} = \mathbb{P}(\widetilde{\tau}(y) > H_k - 2) = \mathbb{P}(\widetilde{\tau}(y) \geq H_k - 1), \tag{15}$$

where the before-last equality uses Lem. 1 applied to $\widetilde{\pi}_k \in \Pi^{PSD}(\langle \mathcal{S}', \mathcal{A}, c, \widetilde{p}_k, y\rangle)$ (the fact that $\widetilde{\pi}_k$ is proper in $\widetilde{p}_k$ stems from Eq. 13), while the last equality uses that the hitting time $\widetilde{\tau}(y)$ is an integer. By definition of $H_k := \min\left\{n > 1 : \|\widetilde{Q}^{n-1}\|_\infty \leq \frac{1}{\sqrt{k}}\right\}$, we have $\|\widetilde{Q}^{H_k-2}\|_\infty > \frac{1}{\sqrt{k}}$. Combining this with Eq. (14) and (15) yields

$$\frac{2\left(r\frac{2c_{\max}D}{c_{\min}}\right)^r}{(H_k-1)^r} > \frac{1}{\sqrt{k}},$$

which implies that

$$H_k - 1 < r\frac{2c_{\max}D}{c_{\min}}\left(2\sqrt{k}\right)^{\frac{1}{r}}.$$

In particular, selecting $r := \lceil \log(2\sqrt{k})\rceil$ yields

$$\begin{aligned} H_k - 1 &< \frac{2c_{\max}D}{c_{\min}}\lceil \log(2\sqrt{k})\rceil (2\sqrt{k})^{\frac{1}{\lceil \log(2\sqrt{k})\rceil}} \\ &\leq \frac{2c_{\max}D}{c_{\min}}\lceil \log(2\sqrt{k})\rceil \underbrace{(2\sqrt{k})^{\frac{1}{\log(2\sqrt{k})}}}_{=e}. \end{aligned}$$

Hence,

$$\Omega_K \leq \left\lceil 6\frac{c_{\max}}{c_{\min}} D \log(2\sqrt{K})\right\rceil.$$

## F. Proof of Lem. 7

For notational ease, in Sect. F we adopt the notation $H_k := H_{k,0}$, $\widetilde{\pi}_k := \widetilde{\pi}_{k,0}$, $\varepsilon_k := \varepsilon_{k,0}$ (i.e., we remove the subscript 0).

We denote by $\tau_k$ (resp. $\widetilde{\tau}_k$) the hitting time of policy $\pi_k$ in the true model $p$ (resp. in the optimistic model $\widetilde{p}_k$). For any $h \in [H_k]$ we define

$$\Gamma_{k,h}(s_{k,h}) = \mathbb{1}_{\{\tau_k(s_{k,h}) > H_k - h\}} - \mathbb{P}(\widetilde{\tau}_k(s_{k,h}) > H_k - h).$$

Since $F_K = \sum_{k=1}^K \mathbb{1}_{\{\tau_k(s_{k,1}) > H_k - 1\}}$, we have

$$F_K = \sum_{k=1}^K \Gamma_{k,1}(s_{k,1}) + \sum_{k=1}^K \mathbb{P}(\widetilde{\tau}_k(s_0) > H_k - 1).$$

We have for $h \in [H_k - 1]$, $\mathbb{1}_{\{\tau_k(s_{k,h}) > H_k - h\}} = \mathbb{1}_{\{\tau_k(s_{k,h+1}) > H_k - h - 1\}}$ and therefore

$$\begin{aligned}
\Gamma_{k,h}(s_{k,h}) &= \mathbb{1}_{\{\tau_k(s_{k,h+1}) > H_k - h - 1\}} - \sum_{y \in \mathcal{S}'} \widetilde{p}_k(y \,|\, s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\mathbb{P}(\widetilde{\tau}_k(y) > H_k - h - 1) \\
&\leq \mathbb{1}_{\{\tau_k(s_{k,h+1}) > H_k - h - 1\}} - \sum_{y \in \mathcal{S}'} p(y \,|\, s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\mathbb{P}(\widetilde{\tau}_k(y) > H_k - h - 1) + 2\beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h})) \\
&= \Gamma_{k,h+1}(s_{k,h+1}) + \Psi_{k,h} + 2\beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h})),
\end{aligned}$$

where we define

$$\Psi_{k,h} = \mathbb{P}(\widetilde{\tau}_k(s_{k,h+1}) > H_k - h - 1) - \sum_{y \in \mathcal{S}'} p(y \,|\, s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\mathbb{P}(\widetilde{\tau}_k(y) > H_k - h - 1).$$

Furthermore, whatever the value of $s_{k,H_k}$ we have

$$\Gamma_{k,H_k}(s_{k,H_k}) = \mathbb{1}_{\{\tau_k(s_{k,H_k}) > 0\}} - \mathbb{P}(\widetilde{\tau}_k(s_{k,H_k}) > 0) = \mathbb{1}_{\{s_{k,H_k} \neq \overline{s}\}} - \mathbb{1}_{\{s_{k,H_k} \neq \overline{s}\}} = 0.$$

By telescopic sum we thus get

$$\begin{aligned}
\Gamma_{k,1}(s_{k,1}) &= \sum_{h=1}^{H_k - 1} (\Gamma_{k,h}(s_{k,h}) - \Gamma_{k,h+1}(s_{k,h+1})) + \Gamma_{k,H_k}(s_{k,H_k}) \\
&\leq \sum_{h=1}^{H_k - 1} \Psi_{k,h} + 2 \sum_{h=1}^{H_k - 1} \beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h})).
\end{aligned}$$

Summing over the episode index $k$ yields

$$F_K \leq \sum_{k=1}^K \sum_{h=1}^{H_k - 1} \Psi_{k,h} + 2 \sum_{k=1}^K \sum_{h=1}^{H_k - 1} \beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h})) + \sum_{k=1}^K \mathbb{P}(\widetilde{\tau}_k(s_0) > H_k - 1).$$

$(\Psi_{k,h})$ is a martingale difference sequence with $|\Psi_{k,h}| \leq 2$, so from Azuma-Hoeffding's inequality, in the same vein as in Eq. (9), we have with probability at least $1 - \frac{2\delta}{3}$

$$\sum_{k=1}^K \sum_{h=1}^{H_k - 1} \Psi_{k,h} \leq 2\sqrt{2\left(\sum_{k=1}^K H_k\right) \log\left(\frac{3\left(\sum_{k=1}^K H_k\right)^2}{\delta}\right)} \leq 2\sqrt{2\Omega_K K \log\left(\frac{3(\Omega_K K)^2}{\delta}\right)}.$$

By the pigeonhole principle (Eq. 10), we have

$$\sum_{k=1}^K \sum_{h=1}^{H_k - 1} \beta_k(s_{k,h}, \widetilde{\pi}_k(s_{k,h})) \leq 2S\sqrt{8A\Omega_K K \log\left(\frac{2A\Omega_K K}{\delta}\right)}.$$

From Lem. 1 and Hölder's inequality, we have

$$\sum_{k=1}^{K} \mathbb{P}(\widetilde{\tau}_k(s_0) > H_k - 1) = \sum_{k=1}^{K} \mathbb{1}_{s_0} (Q_{\widetilde{\pi}_k}^{\widetilde{p}_k})^{H_k - 1} \mathbb{1} \leq \sum_{k=1}^{K} \|\mathbb{1}_{s_0}\|_1 \|(Q_{\widetilde{\pi}_k}^{\widetilde{p}_k})^{H_k - 1} \mathbb{1}\|_\infty \leq \sum_{k=1}^{K} \|(Q_{\widetilde{\pi}_k}^{\widetilde{p}_k})^{H_k - 1}\|_\infty.$$

Consequently, by choice of $H_k := \min\{n > 1 \mid \|(Q_{\widetilde{\pi}_k}^{\widetilde{p}_k})^{n-1}\|_\infty \leq \frac{1}{\sqrt{k}}\}$, we get

$$\sum_{k=1}^{K} \mathbb{P}(\widetilde{\tau}_k(s_0) > H_k - 1) \leq \sum_{k=1}^{K} \frac{1}{\sqrt{k}} \leq 2\sqrt{K}.$$

## G. Proof of Lem. 8

Recall that $T_{K,2}$ is the number of time steps during attempts in phase ② up to the end of environmental episode $K$. We introduce $\Omega'_K := \max_{k \in [K]} \max_{j \in [J_k]} H_{k,j}$ and $G_K := \sum_{k=1}^{K} J_k$ which is the total number of attempts in phase ② up to episode $K$. This means that $T_{K,2} \leq \Omega'_K G_K$.

First, by adapting Lem. 6 and using that in attempts in phase ② we have $c_{\max} = c_{\min} = 1$, we have under the event $\mathcal{E}$,

$$\Omega'_K \leq \left\lceil 6D \log(2\sqrt{G_K}) \right\rceil. \tag{16}$$

We can decompose $G_K$ as the sum of attempts that succeed in reaching $\overline{s}$ (equal to $F_K$ which is upper bounded by Lem. 7) and of those that fail in reaching $\overline{s}$, whose number we denote by $F_K^\dagger$. We then have

$$G_K \leq F_K + F_K^\dagger. \tag{17}$$

By adapting Lem. 7, we have the following high-probability bound, for any value of $G_K$,

$$F_K^\dagger = O\left( S \sqrt{A \Omega'_K G_K \log\left( \frac{A \Omega'_K G_K}{\delta} \right)} \right). \tag{18}$$

Plugging Eq. (16) and (17) into Eq. (18) yields

$$G_K \leq F_K + O\left( S \sqrt{AD G_K} \log\left( \frac{AD G_K}{\delta} \right) \right).$$

Hence we get

$$G_K \leq 2F_K \underbrace{- G_K + O\left( S \sqrt{AD G_K} \log\left( \frac{AD G_K}{\delta} \right) \right)}_{:= (y)},$$

where $(y)$ can be bounded using the technical Lem. 13 as follows

$$(y) \leq O\left( S^2 AD \left[ \log\left( \frac{SAD}{\sqrt{\delta}} \right) \right]^2 \right).$$

Plugging in the result of Lem. 7 yields

$$G_K = \widetilde{O}\left( S \sqrt{\frac{c_{\max}}{c_{\min}} ADK \log\left( \frac{K}{\delta} \right)} + S^2 AD \log\left( \frac{1}{\delta} \right) \right).$$

This bound can be translated in a bound on $T_{K,2}$ using Eq. (16) as follows

$$T_{K,2} = O\Big( DG_K \log(S\sqrt{G_K}) \Big) = \widetilde{O}\left( DS \sqrt{\frac{c_{\max}}{c_{\min}} ADK \log\left( \frac{K}{\delta} \right)} \log(K) + S^2 AD^2 \log\left( \frac{1}{\delta} \right) \log(K) \right).$$

## H. Proof of Thm. 2

The (possibly non-stationary) policy $\mu_k$ executed at each episode $k$ can be written as $(\widetilde{\pi}_{k,0}, \widetilde{\pi}_{k,1}, \dots, \widetilde{\pi}_{k,J_k})$. As explained in Sect. 4.3, by assigning a regret of $c_{\max}$ to each time step during attempts in phase ② (i.e., during the executions of the policies $\widetilde{\pi}_{k,1}, \dots, \widetilde{\pi}_{k,J_k}$), we can decompose the regret as

$$\Delta(\text{UC-SSP}, K) = \sum_{k=1}^{K}\left[\left(\sum_{h=1}^{\tau_{k,0}} c(s_{k,h}, \mu_k(s_{k,h}))\right) - V^\star(s_0)\right] \le \sum_{k=1}^{K}\left[\left(\sum_{h=1}^{H_{k,0}} c(s_{k,h}, \widetilde{\pi}_{k,0}(s_{k,h}))\right) - V^\star(s_0)\right] + c_{\max} T_{K,2}.$$

Suppose from now on that the event $\mathcal{E}$ is true (this holds with probability at least $1 - \frac{\delta}{3}$). Lem. 5 yields that with probability at least $1 - \frac{2\delta}{3}$,

$$\sum_{k=1}^{K}\left[\left(\sum_{h=1}^{H_{k,0}} c(s_{k,h}, \widetilde{\pi}_{k,0}(s_{k,h}))\right) - V^\star(s_0)\right] \le 4c_{\max} DS\sqrt{8A\Omega_K K \log\left(\frac{2A\Omega_K K}{\delta}\right)} + 2c_{\max} D\sqrt{2\Omega_K K \log\left(\frac{3(\Omega_K K)^2}{\delta}\right)} + \frac{c_{\min}}{2}\Omega_K(1 + \log(\Omega_K K)),$$

where according to Lem. 6,

$$\Omega_K \le \left\lceil 6\frac{c_{\max}}{c_{\min}} D \log(2\sqrt{K})\right\rceil.$$

On the other hand, Lem. 8 yields

$$T_{K,2} = \widetilde{O}\left(DS\sqrt{\frac{c_{\max}}{c_{\min}} ADK \log\left(\frac{K}{\delta}\right)}\log(K) + S^2AD^2\log\left(\frac{1}{\delta}\right)\log(K)\right).$$

Putting everything together finally yields that with probability at least $1 - \delta$, for any $K \ge 1$,

$$\Delta(\text{UC-SSP}, K) = \widetilde{O}\left(c_{\max} DS\sqrt{\frac{c_{\max}}{c_{\min}} ADK \log\left(\frac{K}{\delta}\right)}\log(K) + c_{\max} S^2AD^2\log\left(\frac{1}{\delta}\right)\log(K)\right).$$

## I. Relaxation of Assumptions

### I.1. Straightforward extension to unknown, stochastic costs

Although we assume (as in e.g., Azar et al., 2017) that the costs are known and deterministic for ease of exposition, we emphasize that extending the setting to unknown stochastic costs poses no major difficulty. The only requirement is that the learner needs to know in advance the range of the non-goal costs, i.e., the constants $c_{\min}$ and $c_{\max}$. In that case, at the beginning of each attempt $(k, 0)$ (i.e., in phase ①), the confidence set $\mathcal{M}_{k,0}$ is not only defined with the confidence interval on the transition probabilities but also with a confidence interval on the costs. Namely, we consider

$$\mathcal{M}_{k,0} := \{\langle \mathcal{S}, \mathcal{A}, \widetilde{c}, \widetilde{p}\rangle \mid \widetilde{p}(\cdot|s,a) \in B_{k,0}(s,a), \widetilde{c}(s,a) \in B'_{k,0}(s,a)\},$$

where $B_{k,0}(s,a)$ is defined as in Sect. 4.1, and where for any $a \in \mathcal{A}$, $\widetilde{c}(\bar{s}, a) = 0$ while for any $s \in \mathcal{S}$,

$$B'_{k,0}(s,a) := [\widehat{c}_{k,0}(s,a) - \beta'_{k,0}(s,a), \widehat{c}_{k,0}(s,a) + \beta'_{k,0}(s,a)] \cap [c_{\min}, c_{\max}],$$

with $\widehat{c}_{k,0}(s,a)$ the empirical costs and

$$\beta'_{k,0}(s,a) := 2\sqrt{\frac{\log\left(\frac{6SAN^+_{k,0}(s,a)}{\delta}\right)}{N^+_{k,0}(s,a)}}.$$

The analysis on the regret bound of UC-SSP then only adds an additional error term on estimating the transition costs, which is subsumed by the other terms. Consequently, we obtain exactly the same regret bound as in Thm. 2.

### I.2. Relaxation of Asm. 2 (i.e., if $M$ is non-SSP-communicating, i.e., $D = +\infty$)

The requirement that the goal is reachable from any state (Asm. 2) is a natural and inherent assumption of the SSP problem as introduced in Bertsekas (2012). However, a reasonable extension is to allow for the existence of (potentially unknown) *dead-end* states, i.e., states from which reaching the goal is impossible. In that case, EVI$_{\text{SSP}}$, which operates of the entire state space $\mathcal{S}$, fails to converge since the values at dead-end states are infinite. Kolobov et al. (2012) propose to put a "cap" on any state's cost by optimizing the *truncated value function*, or Finite-Penalty criterion,

$$V_J^\pi(s) := \min\{J,\ V^\pi(s)\},$$

where $J > 0$ corresponds to a penalty incurred if a dead-end state is visited. From Kolobov et al. (2012), there exists an optimal policy $\pi_J^\star(s)$ that minimizes $V_J^\pi(s)$ and the optimal truncated value function $V_J^\star$ is a fixed point of the modified Bellman operator $\mathcal{L}_J$ defined as

$$\mathcal{L}_J V(s) := \min\left\{J,\ \min_{a\in\mathcal{A}}\left[c(s,a) + \sum_{y\in\mathcal{S}} p(y|s,a)V(y)\right]\right\}.$$

Denote by $\mathcal{S}^{DE} \subsetneq \mathcal{S}$ the set of dead-end states. We replace Asm. 2 with the following assumptions.

**Assumption 4.** 1) $s_0 \notin \mathcal{S}^{DE}$. 2) $V^\star(s_0) < +\infty$ and an upper bound $J$ on $V^\star(s_0)$ is known. 3) We augment the action space $\mathcal{A}$ with an action $\overline{a}$ that causes a transition from any state in $\mathcal{S}$ to the target state with probability 1 and cost $J$ (i.e., we place ourselves in a resetting environment).

Note that 1) and 3) of Asm. 4 are required to make the learning problem and the definition of regret sensible (i.e., we have $V^\star(s_0) < +\infty$ and we have the possibility to reset whenever we are stuck in a dead-end state). Moreover, 2) guarantees that $V^\star(s_0) = V_J^\star(s_0)$ and that if we run EVI$_{\text{SSP}}$ on $\mathcal{L}_J$ instead of $\mathcal{L}$, then $J$ is an upper bound on the optimistic value function output by EVI$_{\text{SSP}}$(instead of $c_{\max}D$ which is vacuous when $D = +\infty$). Note that 2) is tightly related to the requirement of Fruit et al. (2018b) of prior knowledge on an upper bound of the span of the optimal bias function, and that 1) is similar to the assumption of a starting state belonging to the set of communicating states in TUCRL (Fruit et al., 2018a).

With those assumptions at hand, we consider the algorithm UC-SSP-$\mathcal{L}_J$, which differs from UC-SSP in 3 ways: it iterates EVI$_{\text{SSP}}$ on the operator $\mathcal{L}_J$, the length of the $k$-th phase ① is set to $H_k^{(J)} := 6\frac{J}{c_{\min}}\log(2\sqrt{k})$, and it executes action $\overline{a}$ at the end of each attempt ① (this means that there is no more phase ②, and the $k$-th attempt ① exactly corresponds to the $k$-th environmental episode).

**Lemma 16.** *Under Asm. 4 and 1, with probability at least* $1-\delta$,

$$\Delta(\text{UC-SSP-}\mathcal{L}_J, K) = O\left(JS\sqrt{A\Omega_K^{(J)}K}\log\left(\frac{\Omega_K^{(J)}K}{\delta}\right)\right),$$

*where* $\Omega_K^{(J)} := 6\frac{J}{c_{\min}}\log(2\sqrt{K})$.

*Proof.* We have

$$\Delta(\text{UC-SSP-}\mathcal{L}_J, K) = \sum_{k=1}^{K}\left[\left(\sum_{h=1}^{\tau_k(s_0)} c(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\right) - V_J^\star(s_0)\right] \le \sum_{k=1}^{K}\left[\left(\sum_{h=1}^{H_k^{(J)}} c(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\right) - V_J^\star(s_0)\right] + JF_K,$$

where the double sum can be bounded by

$$O\left(JS\sqrt{A\Omega_K K\log\left(\frac{\Omega_K K}{\delta}\right)}\right)$$

by adapting the proof of Lem. 5, since $\widetilde{\pi}_k$ is the greedy policy w.r.t. the optimistic value function $\widetilde{v}_k^{(J)}$ which satisfies both $\widetilde{v}_k^{(J)}(s_0) \le V_J^\star(s_0)$ and $\|\widetilde{v}_k^{(J)}\|_\infty \le J$.

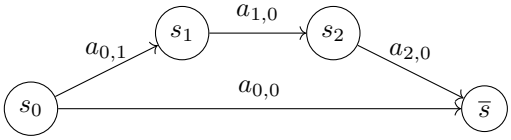


*Figure 3.* SSP instance used in the proof of Lem. 17.

Note that the optimistic hitting time $\tau^{\widetilde{p}_k}_{\widetilde{\pi}_k}$ starting from any state in $\mathcal{S} \setminus \mathcal{S}^{DE}$ still follows a discrete PH distribution with $|\mathcal{S}^{DE}| + 1$ absorbing states (which can be reduced to a discrete PH distribution with a single absorbing state and with the same distribution of the time to absorption). Consequently, using the same reasoning as in the proof of Lem. 6, we can prove that under the event $\mathcal{E}$,

$$\mathbb{P}(\tau^{\widetilde{p}_k}_{\widetilde{\pi}_k}(s_0) \geq H_k^{(J)}) \leq \frac{1}{\sqrt{k}}.$$

Hence we can bound $F_K$ exactly as in Lem. 7. We obtain the desired regret bound by using that $\Omega_K^{(J)} := \max_{k \in [K]} H_K^{(J)} = 6\frac{J}{c_{\min}} \log(2\sqrt{K})$ by choice of $H_k^{(J)}$. □

Interesting future directions in the setting where $D = +\infty$ could be to attempt to remove the need for the prior knowledge $J$ (i.e., weaken Asm. 4), or to focus on the related problem of maximizing the probability of reaching the goal state while keeping cumulative costs low (see e.g., Kolobov et al., 2012, Sect. 6).

### I.3. Relaxation of Asm. 1 (i.e., if $c_{\min} = 0$)

While the assumption of positive costs seems natural in numerous episodic problems and is commonly used in the SSP literature (see e.g., Hansen, 2012; Teichteil-Königsbuch, 2012), we now consider the case where zero non-goal costs may exist, i.e., $c_{\min} = 0$. In such case, the optimal policy is not guaranteed to be proper anymore (Bertsekas, 2012). We thus change the definition of SSP-regret and compare to the best proper policy, that is,

$$\Delta(\mathfrak{A}, K) := \sum_{k=1}^{K} \left[ \Big( \sum_{h=1}^{\tau_k(s_0)} c(s_{k,h}, \mu_k(s_{k,h})) \Big) - V^\star(s_0) \right], \qquad \text{with} \quad V^\star := \min_{\pi \in \Pi^{\text{PSD}}} V^\pi, \quad \pi^\star \in \arg\min_{\pi \in \Pi^{\text{PSD}}} V^\pi. \tag{19}$$

The existence of $c_{\min} > 0$ is leveraged in our analysis to bound $\Omega_K$, more specifically in Eq. (13), which uses that the property of optimism w.r.t. the value functions (i.e., $\widetilde{v}_{k,0} \leq V^\star$ component-wise) yields a "cost-weighted optimism" w.r.t. the expected hitting times, i.e., $\mathbb{E}(\widetilde{\tau}_{k,0}) \leq \frac{2c_{\max}}{c_{\min}} \mathbb{E}(\tau_{\pi^\star})$ component-wise. Yet if zero costs are possible (i.e., $c_{\min} = 0$), then this implication fails to hold.

To circumvent this problem a natural idea is to introduce an additive perturbation $\eta_{k,0} > 0$ to the cost of each transition in the true SSP (note that a small offset of costs to avoid to tricky case of zero costs is also performed by Bertsekas & Yu, 2013). One may hope that this would not affect the behavior of the optimal policy, yet whereas in finite- and infinite-horizon this is indeed the case (i.e., offsetting the costs by a positive constant does not affect the behavior of the optimal policy), Lem. 17 shows that this property does not hold in the SSP setting.

**Lemma 17.** *For any $\eta > 0$, there exists an SSP instance whose optimal policy is different from the one of an identical SSP with all of its transition costs offset by $\eta$.*

*Proof.* Let us consider the SSP from Fig. 3, whose costs are $c(s_0, a_{0,0}) = 4\eta$ and $c(s_0, a_{0,1}) = c(s_1, a_{1,0}) = c(s_2, a_{2,0}) = \eta$. The optimal policy executes action $a_{0,0}$ in state $s_0$. Yet if the costs are all offset by $\eta$, the optimal policy executes action $a_{0,1}$ in state $s_0$. □

Offsetting the costs thus introduces a bias which should be adequately controlled by the choice of $\eta_{k,0}$. We consider the algorithm UC-SSP-$\mathcal{L}_\eta$, which differs from UC-SSP by introducing an additive perturbation $\eta_{k,0} > 0$ to the cost of each

transition in the *optimistic model* for each attempt $(k,0)$ (i.e., in phase ①), i.e., the algorithm iterates $\text{EVI}_{\text{SSP}}$ up to an accuracy of $\varepsilon_{k,0} := \frac{c_{\max}}{t_{k,0}}$ on the operator $\mathcal{L}_\eta$ defined as

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

where $\eta > 0$ depends on the episode $k \in [K]$.

**Lemma 18.** *If $c_{\min} = 0$, under Asm. 2 and the regret definition of Eq. (19) , by selecting $\eta_{k,0} = \frac{1}{k^{1/3}}$, we get with overwhelming probability that*

$$\begin{aligned}\Delta(\text{UC-SSP-}\mathcal{L}_\eta, K) = \widetilde{O}\Big(&c_{\max}DS\sqrt{c_{\max}DA}K^{2/3} + \Upsilon^\star K^{2/3} + c_{\max}DS\sqrt{\Upsilon^\star AK} \\ &+\Upsilon^\star S\sqrt{c_{\max}DA}K^{1/3} + \Upsilon^\star S\sqrt{\Upsilon^\star A}K^{1/6} + S^2AD^2\Big),\end{aligned}$$

*where $\Upsilon^\star := \|\mathbb{E}[\tau_{\pi^\star}]\|_\infty$ is the worst-case (in terms of starting state) expected hitting time of the optimal policy $\pi^\star$ in the original SSP (i.e., without any cost offset).*

*Proof.* For notational ease, throughout the proof of Lem. 18 we adopt the notation $\eta_k := \eta_{k,0}$, $H_k := H_{k,0}$, $\widetilde{\pi}_k := \widetilde{\pi}_{k,0}$, $\varepsilon_k := \varepsilon_{k,0}$ (i.e., we remove the subscript 0).

UC-SSP-$\mathcal{L}_\eta$ modifies the EVI procedure so that it selects a pair $(\widetilde{\pi}_k, \widetilde{p}_k)$ that satisfies for any $s \in \mathcal{S}$,

$$(\widetilde{\pi}_k, \widetilde{p}_k) \in \arg\min_{\widetilde{\pi},\widetilde{p}} \widetilde{v}^{(\eta)}_{\widetilde{\pi},\widetilde{p}}(s), \tag{20}$$

where

$$\widetilde{v}^{(\eta)}_{\widetilde{\pi},\widetilde{p}}(s) := \mathbb{E}_{\widetilde{p}}\left[\sum_{t=1}^{\tau_{\widetilde{\pi}}(s)} c(s_t, \widetilde{\pi}(s_t)) + \eta_k \,\middle|\, s\right] = \mathbb{E}_{\widetilde{p}}\left[\sum_{t=1}^{\tau_{\widetilde{\pi}}(s)} c(s_t, \widetilde{\pi}(s_t)) \,\middle|\, s\right] + \eta_k \mathbb{E}_{\widetilde{p}}[\tau_{\widetilde{\pi}}(s)],$$

and we introduce for ease of notation $\widetilde{v}^{(\eta)}_k(s) := \widetilde{v}^{(\eta)}_{\widetilde{\pi}_k,\widetilde{p}_k}(s)$ and $\widetilde{v}_k(s) := \mathbb{E}_{\widetilde{p}_k}\left[\sum_{t=1}^{\tau_{\widetilde{\pi}_k}(s)} c(s_t, \widetilde{\pi}_k(s_t)) \,\middle|\, s\right]$.

From Eq. (20) we have that under the event $\mathcal{E}$, $\widetilde{v}^{(\eta)}_k(s) \le \widetilde{v}^{(\eta)}_{\pi^\star,p}(s)$, or equivalently by expanding,

$$\widetilde{v}^{(\eta)}_k(s) = \widetilde{v}_k(s) + \eta_k \mathbb{E}_{\widetilde{p}_k}[\tau_{\widetilde{\pi}_k}(s)] \le \mathbb{E}_p\left[\sum_{t=1}^{\tau_{\pi^\star}} c(s_t, \pi^\star(s_t)) + \eta_k \,\middle|\, s\right] = V^\star(s) + \eta_k \mathbb{E}[\tau_{\pi^\star}(s)]. \tag{21}$$

Plugging into Eq. (21) that $\widetilde{v}_k(s) \ge 0$ and $\|V^\star\|_\infty \le c_{\max}D$ from Lem. 2 (which does not require $c_{\min} > 0$) yields

$$\|\mathbb{E}_{\widetilde{p}_k}[\tau_{\widetilde{\pi}_k}]\|_\infty \le \frac{c_{\max}D}{\eta_k} + \Upsilon^\star. \tag{22}$$

Hence the term $\frac{c_{\max}D}{c_{\min}}$ in Eq. (13) (and thus in Lem. 6) can be replaced by the upper bound in Eq. (22), which implies that under the event $\mathcal{E}$,

$$\Omega_K \le 6\left(\frac{c_{\max}D}{\eta_K} + \Upsilon^\star\right)\log(S\sqrt{K}).$$

Furthermore, using Eq. (21) the regret can be decomposed as

$$\sum_{k=1}^K \left[\left(\sum_{h=1}^{\tau_k(s_0)} c(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\right) - V^\star(s_0)\right] \le \sum_{k=1}^K \left[\left(\sum_{h=1}^{H_k} c(s_{k,h}, \widetilde{\pi}_k(s_{k,h}))\right) - \widetilde{v}^{(\eta)}_k(s_0)\right] + \Upsilon^\star \sum_{k=1}^K \eta_k + c_{\max}T_{K,2},$$

where the double sum can be bounded by (excluding lower-order terms)

$$O\left((c_{\max}D + \eta_K \Upsilon^\star)S\sqrt{A\Omega_K K \log\left(\frac{\Omega_K K}{\delta}\right)}\right),$$

by adapting the proof of Lem. 5, since $\widetilde{\pi}_k$ is the greedy policy w.r.t. the optimistic value function $\widetilde{v}_k^{(\eta)}$ which satisfies $\|\widetilde{v}_k^{(\eta)}\|_\infty \leq c_{\max}D + \eta_k \Upsilon^\star$ from Eq. (21). Moreover, we can bound $T_{K,2}$ as in Sect. 4 by using Lem. 8.

Hence selecting $\eta_k = \frac{1}{k^{1/3}}$ and plugging in the bound on $\Omega_K$ yields the desired bound. □

An interesting future direction could be to allow for negative costs yet this extension is outside the scope of the paper.

### I.4. Summary

We report in Table 1 the regret guarantees of UC-SSP (by isolating the dependencies on $K$ and on $D$ or $J$), depending on the assumptions made (and the corresponding choices of Bellman operator for $\text{EVI}_{\text{SSP}}$). We notice that if $D = +\infty$ and under Asm. 4, UC-SSP-$\mathcal{L}_J$ satisfies a regret bound where the infinite term $D$ is replaced with the known upper bound $J \geq V^\star(s_0)$. Moreover, UC-SSP-$\mathcal{L}_\eta$ can deal with the existence of zero costs, however the rate worsens from $\sqrt{K}$ (in Thm. 2 which requires $c_{\min} > 0$) to $K^{2/3}$, due to the bias introduced by offsetting the costs in the optimistic model. Finally, it is straightforward to combine the two aforementioned variants and derive UC-SSP-$\mathcal{L}_{J,\eta}$ which can handle both $D = +\infty$ (under Asm. 4) and $c_{\min} = 0$.

| **Assumptions** | **Regret bound** |
|---|---|
| $c_{\min} > 0$ **(Asm. 1) and** $D < \infty$ **(Asm. 2)** | $\widetilde{O}(D^{3/2}\sqrt{K})$ |
| $c_{\min} > 0$ **(Asm. 1) and** $V^\star(s_0) \leq J$ **w/ RESET (Asm. 4)** | $\widetilde{O}(J^{3/2}\sqrt{K})$ |
| $c_{\min} = 0$ **and** $D < \infty$ **(Asm. 2)** | $\widetilde{O}(D^{3/2}K^{2/3})$ |
| $c_{\min} = 0$ **and** $V^\star(s_0) \leq J$ **w/ RESET (Asm. 4)** | $\widetilde{O}(J^{3/2}K^{2/3})$ |

*Table 1.* Regret guarantees of UC-SSP depending on the assumptions made.

## J. Experiments

In this section, we empirically validate our theoretical findings and perform an ablation study of the algorithms. We consider 3 scenarios: 1) uniform-cost SSP; 2) SSP with $c_{\min} > 0$ and 3) SSP with $c_{\min} = 0$. In all the experiments, we consider the same $(3 \times 4)$ gridworld but we modify the cost function. The agent can move using the cardinal actions (Right, Down, Left, Up). An action fails with probability $p_f = 0.05$. In this case (failure), the agent uniformly follows one of the other directions. Walls are absorbing, i.e., if the action leads against the wall, the agent stays in the current position with probability 1. For example, $p((0,0)|(0,0), right) = \frac{2p_f}{3}$, $p((1,0)|(0,0), right) = \frac{p_f}{3}$ and $p((0,1)|(0,0), right) = 1 - p_f$. If we consider *Up*, we have $p((0,0)|(0,0), Up) = 1$. For the experiments we used the theoretical confidence intervals without constants, i.e., $\beta_{k,j}(s,a) = \sqrt{\frac{SL}{N_{k,j}^+(s,a)}}$ with $L = \log(SAN_{k,j}^+(s,a)/0.1)$.The remaining parameters are set as prescribed by the theory. All the results are averaged over 200 runs.

**1)** The first experiment aims to compare UCRL2 (Jaksch et al., 2010) and UC-SSP in the case of uniform-cost SSP studied in Sect. 3 (see Fig. 4). Thus we set $c(s,a) = 1$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $c(\overline{s}, a) = 0$ for all $a \in \mathcal{A}$. We evaluate the algorithms at $K = 3000$ episodes. Fig. 4(top left) shows that the regret of both algorithms is sublinear, as stated by the theoretical analysis. Interestingly, the regret of UCRL is higher than the one incurred by UC-SSP. This is possibly due to algorithmic structure of UCRL, which behaves in epochs (or algorithmic episodes) and each epoch ends when the number of visits to some state-action pair is doubled. UCRL computes the policy only at the beginning of an epoch. As shown by the vertical lines in Fig. 4(top left), between each planning step, the agent may reach the goal multiple times. While this can be computationally efficient, the drawback is that UCRL may execute sub-optimal policies for long time. On the other hand, we believe that by planning more often, UC-SSP is able to execute better policies than UCRL. In fact, Fig. 4(bottom left) shows

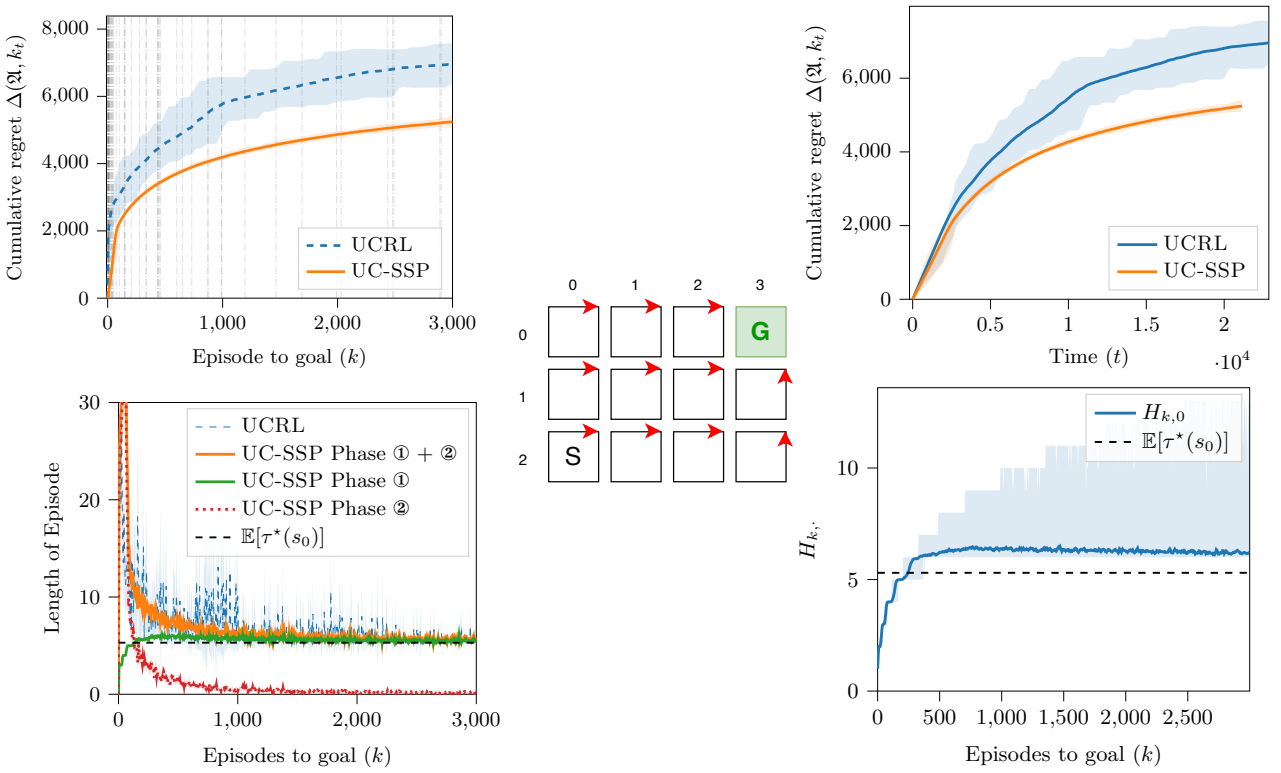


*Figure 4.* Comparison of UC-SSP and UCRL in the case of uniform-cost SSP. The plots are averaged over 200 repetitions. We report the mean and the maximum and minimum value for top line and figure bottom right. For the bottom-left figure, we report the standard deviation of the mean at 96% to simplify the visualization.

that the time required by UCRL to reach the goal $\overline{s}$ is often higher than the one of UC-SSP. It also shows that the length of phase ② in UC-SSP quickly goes to zero, meaning that policy executed by UC-SSP is able to quickly reach the goal. Fig. 4(top right) shows that UCRL requires more time (i.e., steps) than UC-SSP to successfully complete 2000 episodes. This test sheds light on the relationship between UCRL and UC-SSP and shows that, despite the good regret guarantees, UCRL may not exploit the specific structure of the SSP problem and poorly performs compared to UC-SSP. Finally, we also plot the estimate of the hitting time computed by UC-SSP (see Fig. 4(bottom right)). As expected, it is a "tight" upper-bound to the expected hitting time of the optimal SSP policy ($\mathbb{E}[\tau_{\pi^\star}(s_0)] = 5.3$), except in the initial episodes where the optimistic model is far away from the true one. In the latter case, the imagined SSP problem has high probability of reaching $\overline{s}$ from any other state due to the high uncertainty.

**2)** The second experiment focuses on non-uniform cost. At each step, the agent incurs a cost of $\beta > 0$ except when in $\widetilde{s} = (1,1) = P$ where the cost is 1. The state $\widetilde{s}$ is considered to be a sand pit and has the effect of slowing down the agent (i.e., higher cost). Formally, $c(s,a) = \beta$ for all $(s,a) \in (\mathcal{S} \setminus \{\widetilde{s}\}) \times \mathcal{A}$, $c(\widetilde{s},a) = 1$ for all $a \in \mathcal{A}$, and $c(\overline{s},a) = 0$ for all $a \in \mathcal{A}$. Clearly, $c_{\min} = \beta > 0$. Note that the optimal SSP policy is the same for all the selected values of $\beta$. As before, we evaluate the algorithms at $K = 3000$ episodes. In Fig. 5(right) we show the impact of $c_{\min}$ on the regret of UC-SSP. First of all, we show how $c_{\min}$ affects the true solution of the SSP problem. To do so, we run VI on the true model with $\epsilon = 1.e-10$ and obtain

$$V^\star(s_0|\beta = 0.5) = 2.66, \;\; V^\star(s_0|\beta = 0.1) = 0.55, \;\; V^\star(s_0|\beta = 0.01) = 0.07, \;\; V^\star(s_0|\beta = 0.001) = 0.02.$$

To remove the impact of the different magnitude of the cost, we consider the normalized regret $\overline{\Delta}(\mathfrak{A}, K) := \frac{\Delta(\mathfrak{A},K)}{V^\star(s_0)}$. Fig. 5(right) shows that the complexity of the learning problem scales inversely with $c_{\min}$, in the sense that the smaller $c_{\min}$ the higher the regret (i.e., the higher the learning complexity). This supports our theoretical result.

**3)** The final experiment deals with the case $c_{\min} = 0$. We consider the states $(0,0), (0,1), (1,1)$ and $(1,0)$ to have zero cost, see Fig. 6(left). All the other states have cost defined as in experiment 2) with $\beta = 0.4$. Note that there exists loops

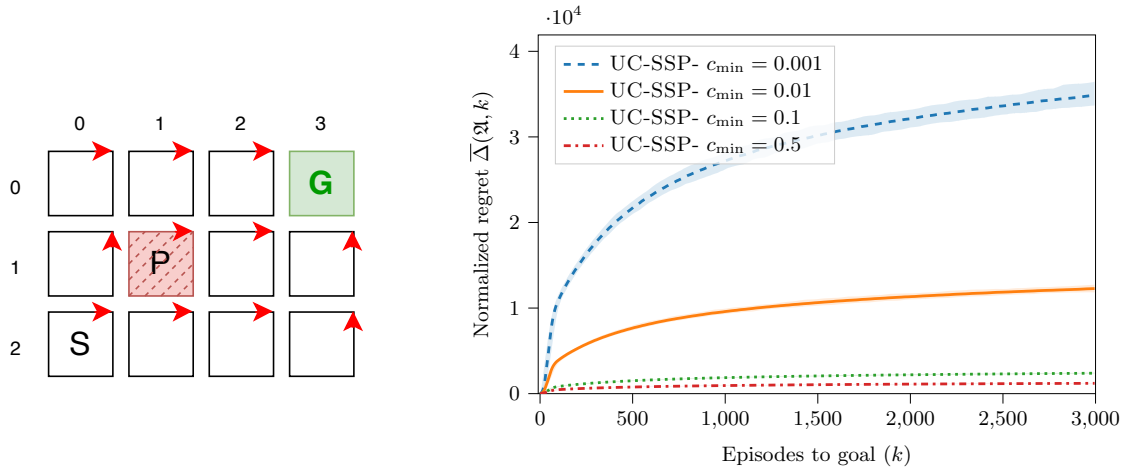


*Figure 5.* Evaluation of the effect of $c_{\min} > 0$ on the regret of UC-SSP. Results are averaged over 200 runs. We report mean value and maximum and minimum observed values.

with zero costs, which means that there exist improper policies with finite $V$-values. As mentioned in App. I.3, in this case we compete against the optimal proper policy (see Fig. 6(top left)). To compute the optimal proper policy and its value $V$, we use VI with perturbation of $1e-10$ (Bertsekas & Yu, 2013). We evaluate the algorithms at $K = 3000$ episodes. We notice that UC-SSP has sublinear regret as expected. The perturbation of the costs has a large impact on the initial phase of UC-SSP when both uncertainty and perturbation are high. In this case, UC-SSP highly overestimates the hitting time of the optimal policy, leading to the execution of suboptimal policies for a long time (due to Phase ①). Once the perturbation and/or the uncertainty decreases, we notice that the estimated hitting time drops rapidly and approaches the true value. It is also interesting to notice that the estimated hitting time of phase ② is never too high. This is due to the fact that phase ② aims to find the policy reaching the goal state in the smallest time.

### J.1. Bernstein Inequalities

In this section, we provide an evaluation of the proposed algorithm with Bernstein inequalities and perform empirical comparison with later work (Cohen et al., 2020). Similarly to (e.g., Azar et al., 2017; Fruit et al., 2020), we consider the following concentration inequality of the transition probabilities: $\forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}'$,

$$|\widetilde{p}(s'|s,a) - \widehat{p}_{k,j}(s'|s,a)| \leq \beta_{k,j}(s,a,s') \approx \sqrt{\frac{\sigma_p^2(s,a,s')L}{N_{k,j}^+(s,a)}} + \frac{L}{N_{k,j}^+(s,a)} \tag{23}$$

where $L = \log(SAN_{k,j}^+(s,a)/0.1)$ and $\sigma_p^2(s,a,s') = \widehat{p}_{k,j}(s'|s,a)(1-\widehat{p}_{k,j}(s'|s,a))$. Optimistic SSP planning can be performed using extended value iteration (as in Alg. 2). We thus use the optimistic Bellman operator defined in Eq. (4) with $B_{k,j}(s,a) := \{\widetilde{p} \in \mathcal{C} \mid \widetilde{p}(\cdot \mid \overline{s},a) = \mathbb{1}_{\overline{s}}, |\widetilde{p}(s' \mid s,a) - \widehat{p}_{k,j}(s' \mid s,a)| \leq \beta_{k,j}(s,a,s')\}$.

We compare with UCRL-SSP (Cohen et al., 2020). UCRL-SSP is a variant of UCRL2B (Fruit et al., 2020) where the average reward planning is replaced with the SSP planning. When $c_{\min} = 0$, UCRL-SSP leverages the same perturbation idea used by UC-SSP. The cost is then defined as $c(s,a) = \max\{c(s,a), \epsilon\}$ with $\epsilon = \frac{S^2A}{K}$.

The main goal of this section is to empirically show that, despite the $K^{2/3}$ regret bound when $c_{\min} = 0$, UC-SSP is competitive with UCRL-SSP whose regret bound scales as $\sqrt{K}$. We also show the role of the pivot horizon used by UC-SSP.

As done in the previous section, we start considering the uniform cost case. Fig. 7 shows that UC-SSP outperforms UCRL-SSP. From Fig. 7 we can see that the lower regret of UC-SSP comes from the use of the pivot horizon. Indeed, when we integrate the pivot horizon idea in UCRL-SSP[7] the algorithms behave similarly. In Fig. 7 we can see that UCRL-SSP

[7] UCRL-SSP uses the same condition of UCRL2B to terminate an algorithmic episode, i.e., when the number of visits to a state-action pair is doubled, the algorithmic episode ends. When using the pivot horizon, we simply limit the number of steps in the algorithmic episode to be at most the pivot horizon (as done for UC-SSP). We also integrated the condition of planning every time the goal state is reached but we didn't observe any significant change in this domain.

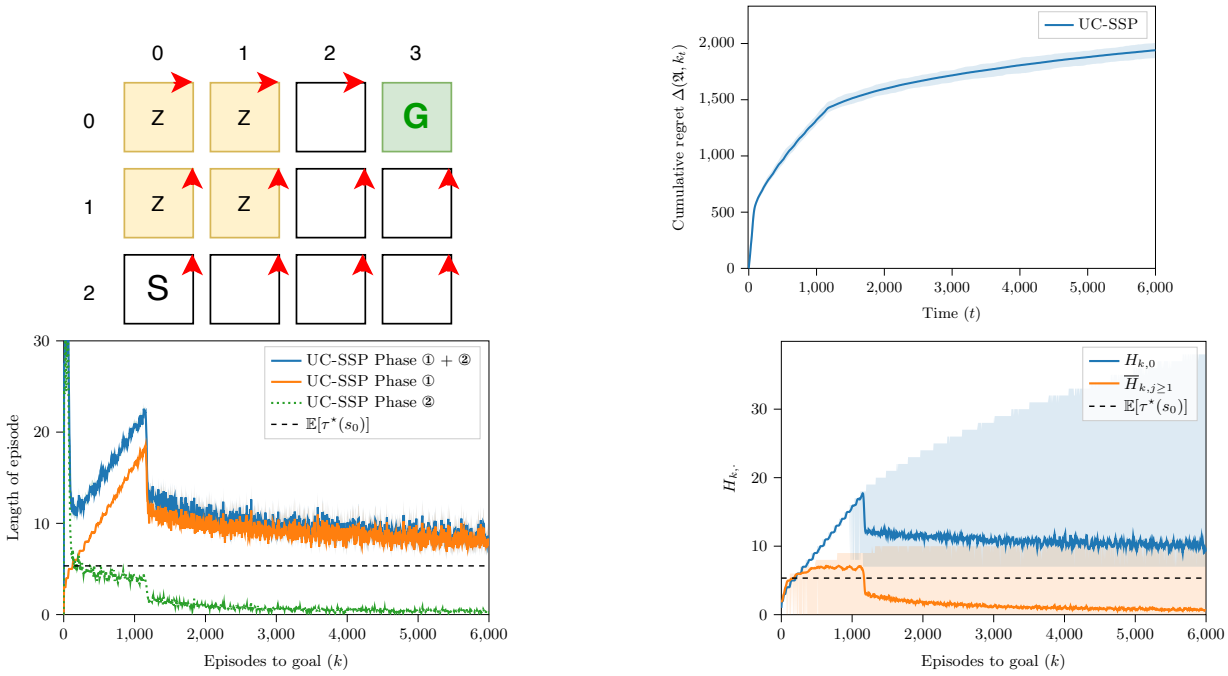


*Figure 6.* Evaluation of UC-SSP for $c_{\min} = 0$. See Fig. 4 for details.

behaves as UCRL2B. This is due to the fact that SSP planning is equivalent to average reward planning in this setting (i.e., uniform cost). Furthermore, it shows that, in this domain, UCRL-SSP is not able to leverage the structure of the SSP problem. In contrast, UC-SSP adapts to the SSP problem thanks to the pivot horizon.

The second experiment focuses on the case when $c_{\min} = 0$. As shown in Fig. 8*(left)*, UC-SSP has a low regret even in this case. UCRL-SSP achieves the same performance of UC-SSP only when using the pivot horizon as a stopping condition of the algorithmic episode. This shows again that the stopping condition based on pivot horizon allows the algorithms to better adapt to the the SSP structure of this problem. Finally, Fig. 8*(right)* shows that phase ② happens only at the early stages of the learning process. As a consequence, UC-SSP does not suffer additional regret due to phase ② in this domain.

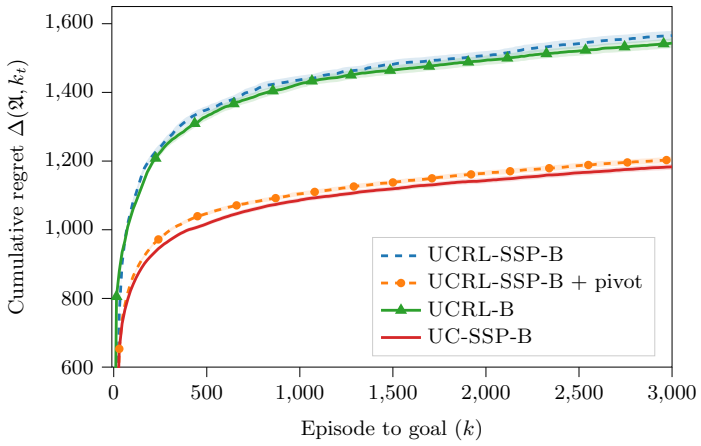


*Figure 7.* Evaluation of the algorithms with Bernstein inequalities and uniform cost. See Fig. 4 for details. We average the results over 200 runs and report the standard deviation of the mean at 96%.

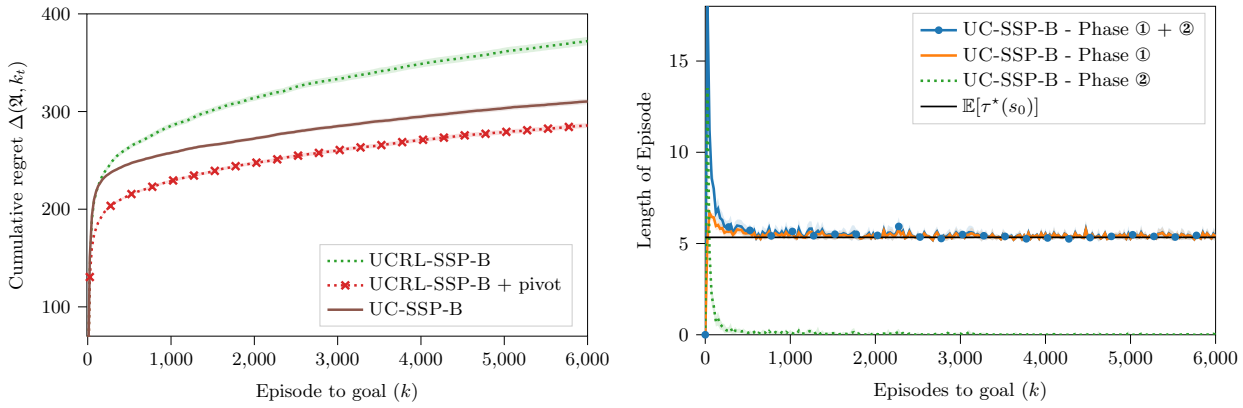


*Figure 8.* Evaluation of the algorithms with Bernstein inequalities and $c_{\min} = 0$. See Fig. 6 for details. Right figure shows the average length of Phase ① and ② for UC-SSP with Bernstein inequalities.