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 γ, 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 π is *proper* if $\overline{s}$ is reached with probability 1 from any state in $\mathcal{S}$ following π.[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 π, 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_π 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 ∞-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 μ_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 μ_k is chosen at the beginning of the episode and run until its termination. Indeed, as μ_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_∞ 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_∞ 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 π in the MDP M_∞. 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_∞ (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 Δ_∞ 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_∞ is communicating, which is needed for proving regret bounds for UCRL2 in general MDPs. Thm. 1 shows that even when M_∞ 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.

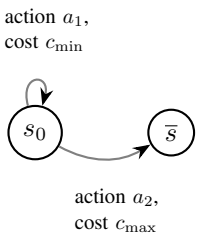

Figure 1. Deterministic two-state SSP M with two available actions: a_1 which self-loops on s_0 with cost $c_{\min}$ and a_2 which goes from s_0 to $\overline{s}$ with cost $c_{\max} > 2c_{\min}$.

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 EVI_{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 Π^{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_{∞} 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 $\frac{c_{\max}}{c_{\min}}$ 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 τ_π differ significantly across policies π. 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, EVI_{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 Ω_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 EVI_{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 Ω_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.