Stochastic Shortest Path: Minimax, Parameter-Free and Towards Horizon-Free Regret

Jean Tarbouriech [*1,2] **Runlong Zhou** [*3] **Simon S. Du** [4,1] **Matteo Pirotta** [1] **Michal Valko** [5] **Alessandro Lazaric** [1]

Abstract

[1] We study the problem of learning in the stochastic shortest path (SSP) setting, where an agent seeks to minimize the expected cost accumulated before reaching a goal state. We design a novel model-based algorithm `EB-SSP` and prove that it achieves the minimax regret rate $\widetilde{O}(B_\star\sqrt{SAK})$, where K is the number of episodes, S is the number of states, A is the number of actions, and $B_\star$ bounds the expected cumulative cost of the optimal policy from any state, thus closing the gap with the lower bound. Interestingly, `EB-SSP` obtains this result while being parameter-free, i.e., it does not require any prior knowledge of $B_\star$, nor of $T_\star$, which bounds the expected time-to-goal of the optimal policy from any state. We also show various cases (e.g., positive costs, or general costs when an order-accurate estimate of $T_\star$ is available) where the regret only contains a logarithmic dependence on $T_\star$, thus yielding the first (nearly) horizon-free regret bound beyond the finite-horizon MDP setting.

1. Introduction & Summary of Contributions

Stochastic shortest path (SSP) is a goal-oriented reinforcement learning (RL) setting where the agent aims to reach a predefined goal state while minimizing its total expected cost (Bertsekas, 1995). The interaction between the agent and the environment ends *only* when (and if) the goal state is reached, so the length of an episode is not predetermined (nor bounded) and it is influenced by the agent's behavior. SSP includes both finite-horizon and discounted Markov Decision Processes (MDPs) as special cases. Many common RL problems can be cast under the SSP formulation, such as game playing or navigation.

[*]Equal contribution [1]Facebook AI Research [2]Inria, Scool team [3]Tsinghua University [4]University of Washington [5]DeepMind. Correspondence to: Jean Tarbouriech <jean.tarbouriech@gmail.com>.

ICML 2021 Workshop on Reinforcement Learning Theory.

[1]Extended abstract. Full version available at https://arxiv.org/abs/2104.11186.

We study the online learning problem in the SSP setting (online SSP in short), where both the transition dynamics and the cost function are initially unknown and the agent interacts with the environment through multiple episodes. The learning objective is to achieve a performance as close as possible to the optimal policy $\pi^\star$, that is, the agent should achieve low *regret* (i.e., the cumulative difference between the total cost accumulated across episodes by the agent and by the optimal policy). We identify three desirable properties for a learning algorithm in online SSP.

- **Desired property 1: Minimax.** The lower bound on the regret is $\Omega(B_\star\sqrt{SAK})$ (Rosenberg et al., 2020), where K is the number of episodes, S is the number of states, A is the number of actions, and $B_\star$ bounds the total expected cost of the optimal policy starting from any state.

  An algorithm for online SSP is (nearly) minimax optimal if its regret is bounded by $\widetilde{O}(B_\star\sqrt{SAK})$, up to logarithmic factors and lower-order terms.

- **Desired property 2: Parameter-free.** Another relevant dimension is the amount of prior knowledge required by the algorithm. While knowing S, A and the cost (or reward) range $[0,1]$ is standard across regret-minimization settings (e.g., finite-horizon, discounted, average-reward), the complexity of learning in SSP may be linked to SSP-specific quantities such as $B_\star$ and $T_\star$, which denotes the expected time-to-goal of the optimal policy from any state.

  An algorithm for online SSP is parameter-free if it relies neither on $T_\star$ nor $B_\star$ prior knowledge.

- **Desired property 3: Horizon-free.** A core challenge in SSP is to trade off between minimizing costs and quickly reaching the goal state. This is accentuated when the instantaneous costs are small, i.e., when there is a mismatch between $B_\star$ and $T_\star$. Indeed, while $B_\star \leq T_\star$ always holds since the cost range is $[0,1]$, the gap between the two may be arbitrarily large. The lower bound stipulates that while the regret depends on $B_\star$, the "time horizon" of the problem, i.e., $T_\star$ should a priori not impact the regret, even as a lower-order term.

  An algorithm for online SSP is (nearly) horizon-free if its regret depends only logarithmically on $T_\star$.

Algorithm	Regret	Minimax	Parameters	Horizon-Free
(Tarbouriech et al., 2020)	$\widetilde{O}_K(K^{2/3})$	No	None	No
(Rosenberg et al., 2020)	$\widetilde{O}\left(B_\star S\sqrt{AK} + T_\star^{3/2}S^2A\right)$	No	$B_\star$	No
	$\widetilde{O}\left(B_\star^{3/2} S\sqrt{AK} + T_\star B_\star S^2A\right)$	No	None	No
(Cohen et al., 2021) *(concurrent work)*	$\widetilde{O}\left(B_\star\sqrt{SAK} + T_\star^4S^2A\right)$	Yes	$B_\star, T_\star$	No
This work	① $\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star S^2A\right)$	**Yes**	$B_\star, T_\star$	**Yes**
	② $\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star S^2A + \frac{T_\star}{\text{poly}(K)}\right)$	**Yes**	$B_\star$	No*
	③ $\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star^3S^3A\right)$	**Yes**	$T_\star$	**Yes**
	④ $\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star^3S^3A + \frac{T_\star}{\text{poly}(K)}\right)$	**Yes**	**None**	No*
Lower Bound	$\Omega(B_\star\sqrt{SAK})$	-	-	-

Figure 1: Regret comparisons of algorithms for online SSP. $\widetilde{O}$ omits logarithmic factors and $\widetilde{O}_K$ only reports the dependence in K. **Regret** is the performance metric of Eq. 1. **Minimax**: Whether the regret matches the $\Omega(B_\star\sqrt{SAK})$ lower bound (Rosenberg et al., 2020), up to logarithmic and lower-order terms. **Parameters**: The parameters that the algorithm requires as input: either both $B_\star$ and $T_\star$, or one of them, or none (i.e., parameter-free). **Horizon-Free**: Whether the regret bound depends only logarithmically on $T_\star$. *If K is known in advance, the additive term $T_\star/\text{poly}(K)$ has a denominator that is polynomial in K, so it becomes negligible for large values of K (if K is unknown, the additive term is $T_\star$). See the full version[1] for the complete statements of our bounds.

Our definition extends the property of so-called horizon-free bounds recently uncovered in finite-horizon MDPs with total reward bounded by 1 (Wang et al., 2020; Zhang et al., 2020; 2021). These bounds depend only logarithmically on the horizon H, which is the number of time steps by which *any* policy terminates. Such notion of horizon would clearly be too strong in the more general class of SSP, where some (even most) policies may never reach the goal, thus having unbounded time horizon. A more adequate notion of horizon in SSP is $T_\star$, which bounds the *expected* time of the *optimal* policy to terminate the episode starting from any state.

Finally, while the previous properties focus on the learning aspects of the algorithm, another important consideration is computational efficiency. It is desirable that a learning algorithm has run-time complexity polynomial in $K, S, A, B_\star$, and $T_\star$. All existing algorithms for online SSP, including the one proposed in this paper, meet such requirement.

Table 1 reviews the existing work on online learning in SSP.

Contributions. We now summarize our main contributions:

- We propose `EB-SSP` (`Exploration Bonus for SSP`), a novel algorithm for online SSP. It carefully skews the empirical transitions and perturbs the empirical costs with an exploration bonus to induce an optimistic SSP problem whose associated value iteration scheme is guaranteed to converge. In this optimistic model, the goal can be reached from *each* state-action pair with positive probability, thus *all* policies are in fact proper (i.e., they eventually reach the goal with probability 1 starting from any state). We decay the bias over time in a way that it only contributes to a lower-order regret term. See Sect. 3 for an overview of our algorithm and analysis. Note that `EB-SSP` is *not* based on a model-optimistic approach (Tarbouriech et al., 2020; Rosenberg et al., 2020), and it does *not* rely on a reduction from SSP to finite-horizon (Cohen et al., 2021) (i.e., we operate at the level of the non-truncated SSP model);
- `EB-SSP` is the first algorithm to achieve the **minimax** regret rate of $\widetilde{O}(B_\star\sqrt{SAK})$ while simultaneously being **parameter-free**: it does not require to know nor estimate $T_\star$, and it is able to bypass the knowledge of $B_\star$ at the cost of only logarithmic and lower-order terms in the regret;
- `EB-SSP` is the first algorithm to achieve **horizon-free** regret for SSP in various cases: i) positive costs, ii) no almost-sure zero-cost cycles, and iii) the general cost case when an order-accurate estimate of $T_\star$ is available (i.e., a value $\overline{T}_\star$ such that $\frac{T_\star}{\upsilon} \le \overline{T}_\star \le \lambda T_\star^\zeta$ for some unknown constants $\upsilon, \lambda, \zeta \ge 1$ is available). This property is especially relevant if $T_\star$ is much larger than $B_\star$, which can occur in SSP models with very small instantaneous costs. Moreover, `EB-SSP` achieves its horizon-free guarantees while maintaining the minimax rate. For instance, un-

der general costs when relying on $T_\star$ and $B_\star$, its regret is $\widetilde{O}(B_\star\sqrt{SAK} + B_\star S^2 A)$. To the best of our knowledge, EB-SSP yields the first set of (nearly) horizon-free bounds beyond the setting of finite-horizon MDPs.

2. Preliminaries

An SSP problem is an MDP $M := \langle \mathcal{S}, \mathcal{A}, P, c, s_0, g\rangle$, where $\mathcal{S}$ is the finite state space with cardinality S, $\mathcal{A}$ is the finite action space with cardinality A, and $s_0 \in \mathcal{S}$ is the initial state. We denote by $g \notin \mathcal{S}$ the goal state, and we set $\mathcal{S}' := \mathcal{S} \cup \{g\}$ (thus $S' := S+1$). Taking action a in state s incurs a cost drawn i.i.d. from a distribution on $[0,1]$ with expectation $c(s,a)$, and the next state $s' \in \mathcal{S}'$ is selected with probability $P(s'|s,a)$ (where $\sum_{s'\in\mathcal{S}'} P(s'|s,a) = 1$). The goal state g is absorbing and zero-cost, i.e., $P(g|g,a) = 1$ and $c(g,a) = 0$ for any action a.[2]

A stationary and deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping from state s to action $\pi(s)$. A policy π is said to be proper if it reaches the goal with probability 1 when starting from any state in $\mathcal{S}$ (otherwise it is improper). We denote by Π_{proper} the set of proper, stationary and deterministic policies. We make the following basic assumption which ensures that the SSP problem is well-posed.

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

The value function (also called cost-to-go) of a policy π and its associated Q-function are defined as

$$V^\pi(s) := \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^{T} c_t(s_t,\pi(s_t)) \,\middle|\, s_1 = s\right],$$

$$Q^\pi(s,a) := \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^{T} c_t(s_t,\pi(s_t)) \,\middle|\, s_1 = s, \pi(s_1) = a\right],$$

where $c_t \in [0,1]$ is the (instantaneous) cost incurred at time t at state-action pair $(s_t, \pi(s_t))$, and the expectation is w.r.t. the random sequence of states generated by executing π starting from state $s \in \mathcal{S}$ (and taking action $a \in \mathcal{A}$ in the second case). Note that V^π may have unbounded components if π never reaches the goal. For a proper policy π, $V^\pi(s)$ and $Q^\pi(s,a)$ are finite for any s, a. By definition of the goal, we set $V^\pi(g) = Q^\pi(g,a) = 0$ for all policies π and actions a. Finally, we denote by $T^\pi(s)$ the expected time that π takes to reach g starting at state s; in particular, if π is proper then $T^\pi(s)$ is finite for all s, yet if π is improper there must exist at least one s such that $T^\pi(s) = \infty$.

Equipped with Asm. 1 and an additional condition on improper policies defined below, one can derive important properties on the optimal policy $\pi^\star$ that minimizes the value function component-wise.

Lemma 2 (Bertsekas & Tsitsiklis, 1991; Yu & Bertsekas, 2013)**.** *Suppose that Asm. 1 holds and that for every improper policy π' there exists at least one state $s \in \mathcal{S}$ such that $V^{\pi'}(s) = +\infty$. Then the optimal policy $\pi^\star$ is stationary, deterministic, and proper. Moreover, $V^\star = V^{\pi^\star}$ is the unique solution of the optimality equations $V^\star = \mathcal{L}V^\star$ and $V^\star(s) < +\infty$ for any $s \in \mathcal{S}$, where for any vector $V \in \mathbb{R}^S$ the optimal Bellman operator $\mathcal{L}$ is defined as $\mathcal{L}V(s) := \min_{a\in\mathcal{A}} \{c(s,a) + P_{s,a}V\}$. Also, the optimal Q-value, denoted by $Q^\star = Q^{\pi^\star}$, is related to the optimal value function as follows: $Q^\star(s,a) = c(s,a) + P_{s,a}V^\star$ and $V^\star(s) = \min_{a\in\mathcal{A}} Q^\star(s,a)$, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.*

Learning formulation. The agent does not have any prior knowledge of the cost function c or transition function P. Each episode starts at the initial state s_0 (the extension to any possibly unknown distribution of initial states is straightforward), and ends *only* when the goal state g is reached (note that this may never happen if the agent does not reach the goal). We evaluate the performance of the agent after K episodes by its *regret*, which is defined as

$$R_K := \sum_{k=1}^{K}\sum_{h=1}^{I^k} c_h^k - K \cdot \min_{\pi\in\Pi_{\text{proper}}} V^\pi(s_0), \qquad (1)$$

where I^k is the time needed to complete episode k and c_h^k is the cost incurred in the h-th step of episode k when visiting (s_h^k, a_h^k). If there exists k such that I^k is infinite, then we define $R_K = \infty$. Throughout we denote the optimal proper policy by $\pi^\star$ and $V^\star(s) := V^{\pi^\star}(s) = \min_{\pi\in\Pi_{\text{proper}}} V^\pi(s)$ and $Q^\star(s,a) := Q^{\pi^\star}(s,a) = \min_{\pi\in\Pi_{\text{proper}}} Q^\pi(s,a)$ for all (s,a). Let $B_\star > 0$ bound the values of $V^\star$, i.e., $B_\star := \max_{s\in\mathcal{S}} V^\star(s)$. Note that $Q^\star(s,a) \le 1 + B_\star$. Let $T_\star > 0$ bound the expected time-to-goal of the optimal policy, i.e., $T_\star := \max_{s\in\mathcal{S}} T^{\pi^\star}(s)$. We see that $B_\star \le T_\star < +\infty$.

3. Main Algorithm

We introduce our algorithm EB-SSP (Exploration Bonus for SSP) in Alg. 1. It takes as input the state-action space $\mathcal{S}\times\mathcal{A}$ and confidence level $\delta \in (0,1)$. For now it considers that an estimate B such that $B \ge B_\star$ is available, and we later handle the case of unknown $B_\star$. It enforces the conditions of Lem. 2 to hold by adding a small cost perturbation $\eta \in [0,1]$ (cf. lines 3, 12 in Alg. 1): either $\eta = 0$ if the agent is aware that costs are already positive, otherwise a careful choice of $\eta > 0$ will be specified.

EB-SSP sequentially constructs optimistic lower bounds on the optimal Q-function and executes the policy that greedily minimizes them. Similar to the MVP algorithm in Zhang et al. (2020) designed for finite-horizon RL, we adopt the

[2] For notational convenience, we will write $P_{s,a} := P(\cdot|s,a)$, $P_{s,a,s'} := P(s'|s,a)$. For any two vectors X, Y of size S', we write their inner product as $XY := \sum_{s\in\mathcal{S}'} X(s)Y(s)$, we denote by X^2 the vector $[X(1)^2, X(2)^2, \dots, X(S')^2]^\top$, let $\|X\|_\infty := \max_{s\in\mathcal{S}'} |X(s)|$, and if X is a probability distribution on $\mathcal{S}'$, then $\mathbb{V}(X,Y) := \sum_{s\in\mathcal{S}'} X(s)Y(s)^2 - (\sum_{s\in\mathcal{S}'} X(s)Y(s))^2$.

Algorithm 1: Algorithm EB-SSP

Input: $\mathcal{S},\ s_0 \in \mathcal{S},\ g \notin \mathcal{S},\ \mathcal{A},\ \delta$.
Input: an estimate B guaranteeing $B \geq \max\{B_\star, 1\}$ (see Sect. 4 if not available).
Optional input: cost perturbation $\eta \in [0, 1]$.
Specify: Trigger set $\mathcal{N} \leftarrow \{2^{j-1} : j = 1, 2, \ldots\}$.
Constants $c_1 = 6,\ c_2 = 36,\ c_3 = 2\sqrt{2},\ c_4 = 2\sqrt{2}$.
For $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}'$, set
$N(s,a) \leftarrow 0;\ n(s,a) \leftarrow 0;\ N(s,a,s') \leftarrow 0;\ \widehat{P}_{s,a,s'} \leftarrow 0;$
$\theta(s,a) \leftarrow 0;\ \widehat{c}(s,a) \leftarrow 0;\ Q(s,a) \leftarrow 0;\ V(s) \leftarrow 0$.
Set initial time step $t \leftarrow 1$ and trigger index $j \leftarrow 0$.
for episode $k = 1, 2, \ldots$ **do**
  Set $s_t \leftarrow s_0$
  while $s_t \neq g$ **do**
    Take action $a_t = \arg\min_{a \in \mathcal{A}} Q(s_t, a)$, incur cost c_t and observe next state $s_{t+1} \sim P(\cdot|s_t, a_t)$.
    Set $(s, a, s', c) \leftarrow (s_t, a_t, s_{t+1}, \max\{c_t, \eta\})$ and $t \leftarrow t + 1$.
    Set $N(s,a) \leftarrow N(s,a) + 1$, $\theta(s,a) \leftarrow \theta(s,a) + c$, $N(s,a,s') \leftarrow N(s,a,s') + 1$.
    if $N(s,a) \in \mathcal{N}$ **then**
      \\ *Update triggered:* VISGO *procedure*
      Set $\widehat{c}(s,a) \leftarrow \mathbb{I}[N(s,a) \geq 2]\frac{2\theta(s,a)}{N(s,a)} + \mathbb{I}[N(s,a) = 1]\theta(s,a)$ and $\theta(s,a) \leftarrow 0$.
      For $s' \in \mathcal{S}'$, set $\widehat{P}_{s,a,s'} \leftarrow N(s,a,s')/N(s,a)$, $n(s,a) \leftarrow N(s,a)$, and $\widetilde{P}_{s,a,s'}$ as in Eq. 2.
      Set $j \leftarrow j + 1$, $\epsilon_{\text{VI}} \leftarrow 2^{-j}/(SA)$ and $i \leftarrow 0$, $V^{(0)} \leftarrow 0$, $V^{(-1)} \leftarrow +\infty$.
      For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, set $n^+(s,a) \leftarrow \max\{n(s,a), 1\}$ and $\iota_{s,a} \leftarrow \ln\big(12SAS'[n^+(s,a)]^2\delta^{-1}\big)$.
      while $\|V^{(i)} - V^{(i-1)}\|_\infty > \epsilon_{\text{VI}}$ **do**
        For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, set
        $b^{(i+1)}(s,a) \leftarrow b(V^{(i)}, s, a)$, \\ *see Eq. 3*
        $Q^{(i+1)}(s,a) \leftarrow \max\big\{\widehat{c}(s,a) + \widetilde{P}_{s,a}V^{(i)} - b^{(i+1)}(s,a),\ 0\big\}$,
        $V^{(i+1)}(s) \leftarrow \min_a Q^{(i+1)}(s,a)$.
        Set $V^{(i+1)}(g) = 0$ and $i \leftarrow i + 1$.
      Set $Q \leftarrow Q^{(i)}$, $V \leftarrow V^{(i)}$.

doubling update framework (first proposed in Jaksch et al., 2010): whenever the number of visits of a state-action pair is doubled, the algorithm updates the empirical cost and transition probability of this state-action pair, and computes a new optimistic Q-estimate and optimistic greedy policy.

The main algorithmic component lies in how to compute the Q-values (w.r.t. which the policy is greedy) when a doubling condition is met. To this purpose, we introduce a procedure called VISGO, for Value Iteration with Slight Goal Optimism. Starting with optimistic values $V^{(0)} = 0$, it iteratively computes $V^{(i+1)} = \widetilde{\mathcal{L}}V^{(i)}$ for a carefully defined operator $\widetilde{\mathcal{L}}$. It ends when a stopping condition is met, specifically once $\|V^{(i+1)} - V^{(i)}\|_\infty \leq \epsilon_{\text{VI}}$ for a precision level $\epsilon_{\text{VI}} > 0$ (specified later), and it outputs the values $V^{(i+1)}$ (and Q-values $Q^{(i+1)}$). Let $\widehat{P}$ and $\widehat{c}$ be the current empirical transition probabilities and costs, and let $n(s,a)$ be the current number of visits to state-action pair (s,a) (and $n^+(s,a) = \max\{n(s,a), 1\}$). We first define transition probabilities $\widetilde{P}$ that are slightly skewed towards the goal w.r.t. $\widehat{P}$, as follows

$$\widetilde{P}_{s,a,s'} := \frac{n(s,a)}{n(s,a)+1}\widehat{P}_{s,a,s'} + \frac{\mathbb{I}[s' = g]}{n(s,a)+1}. \tag{2}$$

We then define the bonus for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and vector $V \in \mathbb{R}^{S'}$ such that $V(g) = 0$, as follows

$$b(V,s,a) := \max\left\{c_1\sqrt{\frac{\mathbb{V}(\widetilde{P}_{s,a}, V)\iota_{s,a}}{n^+(s,a)}},\ c_2\frac{B\iota_{s,a}}{n^+(s,a)}\right\} + c_3\sqrt{\frac{\widehat{c}(s,a)\iota_{s,a}}{n^+(s,a)}} + c_4\frac{B\sqrt{S'\iota_{s,a}}}{n^+(s,a)}, \tag{3}$$

given the estimate B, specific positive constants c_1, c_2, c_3, c_4 and a state-action dependent logarithmic term $\iota_{s,a}$. Given the transitions $\widetilde{P}$ and exploration bonus b, we are ready to define the operator $\widetilde{\mathcal{L}}$ as

$$\widetilde{\mathcal{L}}V(s) := \max\left\{\min_{a \in \mathcal{A}}\left\{\widehat{c}(s,a) + \widetilde{P}_{s,a}V - b(V,s,a)\right\},\ 0\right\}.$$

We see that $\widetilde{\mathcal{L}}$ promotes optimism in two different ways:

(i) On the empirical cost function $\widehat{c}$, via the bonus b (Eq. 3) that intuitively lowers the costs to $\widehat{c} - b$;

(ii) On the empirical transition function $\widehat{P}$, via the transitions $\widetilde{P}$ (Eq. 2) that slightly bias $\widehat{P}$ with the addition of a non-zero probability of reaching the goal from *every* state-action pair.

While the first feature **(i)** is standard in finite-horizon approaches, the second **(ii)** is SSP-specific, and is required to cope with the fact that the empirical model $\widehat{P}$ may *not* admit any proper policy, meaning that executing value iteration for SSP on $\widehat{P}$ may diverge. Our simple transition skewing actually guarantees that *all* policies are proper in $\widetilde{P}$, for any fixed and bounded cost function.[3] By decaying the extra goal-reaching probability inversely with $n(s,a)$, we can tightly control the gap between $\widetilde{P}$ and $\widehat{P}$ and ensure that it only accounts for a lower-order regret term.

Equipped with these two sources of optimism, as long as $B \geq B_\star$, we are able to prove that a VISGO procedure verifies the following two key properties:

(1) ***Optimism:*** VISGO outputs an optimistic estimator of the optimal Q-function at each iteration step, i.e., $Q^{(i)}(s,a) \leq Q^\star(s,a), \forall i \geq 0$,

[3] Interestingly this transition skewing implies that an SSP problem defined on $\widetilde{P}$ is equivalent to a discounted RL problem, with a varying state-action dependent discount factor.

(2) ***Finite-time near-convergence:*** VISGO terminates within a finite number of iteration steps (note that the final iterate $V^{(j)}$ approximates the fixed point of $\widetilde{\mathcal{L}}$ up to an error scaling with ϵ_{VI}).

To satisfy **(1)**, we derive similarly to MVP (Zhang et al., 2020) a *monotonicity* property for $\widetilde{\mathcal{L}}$, achieved by carefully tuning the constants c_1, c_2, c_3, c_4 in the bonus of Eq. 3. On the other hand, the requirement **(2)** is SSP-specific, as it is not needed in finite-horizon where value iteration requires exactly H backward induction steps. *Without* bonuses, the design of $\widetilde{P}$ would have directly entailed that $\widetilde{\mathcal{L}}$ is contractive and convergent (Bertsekas, 1995). However, our variance-aware exploration bonuses introduce a subtle correlation between value iterates (i.e., b depends on V in Eq. 3), which leads to a cost function that varies across iterates. By directly analyzing $\widetilde{\mathcal{L}}$, we establish that it is contractive with modulus $\rho := 1 - \nu < 1$, where $\nu := \min_{s,a} \widetilde{P}_{s,a,g} > 0$. This *contraction* property guarantees a polynomially bounded number of iterations before terminating, i.e., **(2)**.

4. Main Results

Besides ensuring the computational efficiency of EB-SSP, the properties of VISGO lay the foundations for our regret analysis (proof in the full version[1]). For simplicity we state here the results in the case of $B_\star \geq 1$.[4]

Theorem 3. *Assume that $B \geq B_\star$ and that the conditions of Lem. 2 hold. Then with probability at least $1-\delta$ the regret of* EB-SSP *(Alg. 1 with $\eta = 0$) can be bounded by*

$$R_K = O\left(B_\star \sqrt{SAK} \log \frac{B_\star SAT}{\delta} + BS^2 A \log^2 \frac{B_\star SAT}{\delta}\right),$$

with T the accumulated time within the K episodes.

Thm. 3 is an intermediate result for the regret of EB-SSP, as it depends on the *random and possibly unbounded* total number of steps T executed over K episodes, it requires the possibly restrictive second condition of Lem. 2, and it relies on the parameter B being properly tuned. Nonetheless, it already displays interesting properties: **1)** The dependence on T is limited to logarithmic terms; **2)** The parameter B only affects the lower order term, while the main order term naturally scales with the exact range $B_\star$; **3)** Up to dependence on T, the main order term displays minimax optimal dependencies on $B_\star$, S, A, and K.

Our analysis proceeds as follows (details in full version[1]):

Known $B_\star$. First we assume that $B = B_\star$ (i.e., the agent has prior knowledge of $B_\star$) and we focus on bounding T. We instantiate the regret of EB-SSP under various conditions on the SSP model:

(C1) positive costs, i.e., costs lower bounded by an unknown constant $c_{\min} > 0$;

(C2) no almost-sure zero-cost cycles;

(C3) general costs (i.e., non-negative), with no assumption other than Asm. 1; see row ② in Table 1;

(C4) general costs when an order-accurate estimate of $T_\star$ is available (i.e., a value $\overline{T}_\star$ such that $\frac{T_\star}{\upsilon} \leq \overline{T}_\star \leq \lambda T_\star^\zeta$ for some unknown constants $\upsilon, \lambda, \zeta \geq 1$ is available); see row ① in Table 1;

In these four cases, the regret achieved by EB-SSP is always **minimax-optimal**, and moreover in cases (C1, C2, C4) it is also **horizon-free**.

Unknown $B_\star$. We now derive our regret bounds for unknown $B_\star$. Note that the challenge of not knowing the range of the optimal value function does not appear in finite-horizon MDPs, where the bound H (or 1 in Zhang et al., 2020) is assumed to be known to the agent. Without a valid estimate $B \geq B_\star$, one may design an under-specified exploration bonus which cannot guarantee optimism. The unknown $B_\star$ case is non-trivial: it appears impossible to properly estimate $B_\star$ (as some states may never be visited) and it is unclear how a standard doubling trick may be used.[5]

We devise parameter-free EB-SSP as follows (pseudo-code and analysis in the full version[1]). At a high level, it initializes an estimate $\widetilde{B} = 1$ and increments it with two different speeds. On the one hand, it tracks both the cumulative cost and the range of the value function computed by VISGO, and it doubles $\widetilde{B}$ whenever either exceeds thresholds (that depend on $\widetilde{B}$, the number of episodes elapsed so far and other computable terms). We ensure that this happens at most once when $\widetilde{B} \geq B_\star$ and we bound the regret by the cumulative cost threshold. On the other hand, the $\widetilde{B}$ estimate is also increased when a new episode begins by setting $\widetilde{B} \leftarrow \max\{\widetilde{B}, \sqrt{k}/(S^{3/2}A^{1/2})\}$, with k the current episode index. This ensures that for large enough k we have $\widetilde{B} \geq B_\star$, at which point we can bound the regret using Thm. 3.

Theorem 4. *Assume the conditions of Lem. 2 hold. Then with probability at least $1-\delta$ the regret of parameter-free* EB-SSP *can be bounded by*

$$R_K = O\left(R_K^\star \log \frac{B_\star SAT}{\delta} + B_\star^3 S^3 A \log^3 \frac{B_\star SAT}{\delta}\right),$$

where T is the cumulative time within the K episodes and $R_K^\star$ denotes the regret after K episodes of EB-SSP *in the case of known $B_\star$ (i.e., the bound of Thm. 3 with $B = B_\star$).*

Thus all our regret bounds can be made **parameter-free** up to additional logarithmic and lower-order regret terms.

[4] Otherwise, the bounds hold by replacing $B_\star$ with $\max\{B_\star, 1\}$, except for the $B_\star$ factor in the leading term that becomes $\sqrt{B_\star}$.

[5] Note that Qian et al. (2019) raised an open question whether it is possible to design an exploration bonus strategy in a setting where no prior knowledge of the "optimal range" is available. Indeed their approach in average-reward MDPs relies on prior knowledge of an upper bound on the optimal bias span.

REFERENCES

Bertsekas, D. *Dynamic programming and optimal control*, volume 2. 1995.

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

Cohen, A., Efroni, Y., Mansour, Y., and Rosenberg, A. Minimax regret for stochastic shortest path. *arXiv preprint arXiv:2103.13056*, 2021.

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

Qian, J., Fruit, R., Pirotta, M., and Lazaric, A. Exploration bonus for regret minimization in discrete and continuous average reward mdps. In *Advances in Neural Information Processing Systems*, pp. 4891–4900, 2019.

Rosenberg, A., Cohen, A., Mansour, Y., and Kaplan, H. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Tarbouriech, J., Garcelon, E., Valko, M., Pirotta, M., and Lazaric, A. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pp. 9428–9437. PMLR, 2020.

Wang, R., Du, S. S., Yang, L. F., and Kakade, S. M. Is long horizon RL more difficult than short horizon RL? In *Advances in Neural Information Processing Systems*, 2020.

Yu, H. and Bertsekas, D. P. On boundedness of q-learning iterates for stochastic shortest path problems. *Mathematics of Operations Research*, 38(2):209–227, 2013.

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020.

Zhang, Z., Yang, J., Ji, X., and Du, S. S. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. *arXiv preprint arXiv:2101.12745*, 2021.