Abstract

We study the Stochastic Shortest Path (SSP) problem in which an agent has to reach a goal state in minimum total expected cost. In the learning formulation of the problem, the agent has no prior knowledge about the costs and dynamics of the model. She repeatedly interacts with the model for $K$ episodes, and has to minimize her regret. In this work we show that the minimax regret for this setting is $\tilde{O}(\sqrt{B^2_* \cdot |S| \cdot |A| \cdot K})$ where $B_*$ is a bound on the expected cost of the optimal policy from any state, $S$ is the state space, and $A$ is the action space. This matches the $\Omega(\sqrt{B^2_* \cdot |S| \cdot |A| \cdot K})$ lower bound of Rosenberg et al. [2020] for $B_* \geq 1$, and improves their regret bound by a factor of $\sqrt{|S|}$. For $B_* < 1$ we prove a matching lower bound of $\Omega(\sqrt{B_* \cdot |S| \cdot |A| \cdot K})$. Our algorithm is based on a novel reduction from SSP to finite-horizon MDPs. To that end, we provide an algorithm for the finite-horizon setting whose leading term in the regret depends polynomially on the expected cost of the optimal policy and only logarithmically on the horizon.

1 Introduction

We study the stochastic shortest path (SSP) problem in which an agent aims to reach a predefined goal state while minimizing her total expected cost. This is one of the most basic models of reinforcement learning (RL) that includes both finite-horizon and discounted Markov Decision Processes (MDPs) as special cases. In addition, SSP captures a wide variety of realistic scenarios such as car navigation, game playing and drone flying.

We study an online version of SSP in which both the immediate costs and transition distributions of the model are initially unknown to the agent. The agent interacts with the model for $K$ episodes, in each of which she attempts to reach the goal state with minimal cumulative cost. A main challenge in the online model is found when instantaneous costs are small. For example, any learning algorithm that attempts to myopically minimize the accumulated costs might get caught in a cycle with zero cost and never reach the goal state. Nonetheless, even if the costs are not zero, only very small, the agent must be able to trade off the need to minimize costs with that of reaching the goal quickly.

The online setting was originally suggested by Tarbouriech et al. [2020] who gave an algorithm with $\tilde{O}(K^{2/3})$ regret guarantee. In a follow-up work, Rosenberg et al. [2020] improved the previous bound to $\tilde{O}(K^{1/2} \cdot |S| \sqrt{|A|} \cdot K)$, where $S$ is the state space, $A$ is the action space, and $B_*$ is an upper bound on the total expected cost of the optimal policy when initialized at any state. Rosenberg et al. [2020] also provide a lower bound of $\Omega(B_* \sqrt{|S| \cdot |A| \cdot K})$ – leaving a gap of $\sqrt{|S|}$ between the upper and lower bounds. In this work, unlike the previously mentioned works that assume the cost function is deterministic...
and known, we consider the case where the costs are i.i.d. and initially unknown. We prove upper and lower bounds for this case, proving that the optimal regret is of order \( \Theta(\sqrt{B^2 + B^\star N\log K}) \).

The algorithms of both [Tarbouriech et al., 2020], [Rosenberg et al., 2020] were based on a direct application of the “Optimism in the Face of Uncertainty” principle to the SSP model, following the ideas behind the UCRL2 algorithm [Jaksch et al., 2010] for average-reward MDPs. In this work we take a different approach. We propose a novel black-box reduction to finite-horizon MDPs, showing that the SSP problem is not harder than the finite-horizon setting assuming prior knowledge on the expected time it takes for the optimal policy to reach the goal state. While the reduction itself is simple, the analysis is highly nontrivial as one has to show that the goal state is indeed reached in every episode without incurring excessive costs in the process.

The idea of reducing SSP to finite-horizon was previously used by [Chen et al., 2020], [Chen and Luo, 2021] for SSP with adversarially changing costs. However, they run one finite-horizon episode in every SSP episode and then simply try to reach the goal as fast as possible, while we restart a new finite-horizon episode every \( H \) steps. This modification is what enables us to obtain the optimal and improved dependence in the number of states.

In addition, we provide a new algorithm for regret minimization in finite-horizon MDPs called \textsc{ULCVI}. We show that (for large enough number of episodes) its regret depends polynomially on the expected cost of the optimal policy \( B^\star \), and only logarithmically on the horizon length \( H \). This implies that the correct measure for the regret is the expected cost of the optimal policy and not the length of the horizon. We note that regret with logarithmic dependence in the horizon \( H \) was also obtained by [Zhang et al., 2020], yet they make a much stronger assumption: that the cumulative cost of every trajectory is bounded by 1. In contrast, we only assume that the expected cost of the optimal policy is bounded by some constant \( B^\star \), while other policies may suffer a cost of \( H \).

Our reduction, when combined with our finite-horizon algorithm \textsc{ULCVI}, guarantees SSP regret of \( O(\sqrt{B^2 + B^\star N\log K}) \). This matches the lower bound of [Rosenberg et al., 2020] for \( B^\star \geq 1 \) up to logarithmic factors. However, their lower bound does not hold for \( B^\star < 1 \) suggesting that this is not the correct rate in this case. Indeed, we prove a tighter lower bound of \( \Omega(\sqrt{B^\star N\log K}) \) for \( B^\star < 1 \), showing that our regret guarantees are minimax optimal in all cases.

As a final remark we note that, following our work, [Tarbouriech et al., 2021] were able to obtain a comparable regret bound for SSP without prior knowledge of the optimal policy’s expected time to reach the goal state.

1.1 Additional related work

Planning for stochastic shortest path. Early work by [Bertsekas and Tsitsiklis, 1991] studied planning in SSPs, i.e., computing the optimal strategy efficiently when parameters are known. Under certain assumptions, they established that the optimal strategy is a deterministic stationary policy and can be computed efficiently using standard planning algorithms, e.g., Value Iteration and LP.

Adversarial stochastic shortest path. [Rosenberg and Mansour, 2020] presented stochastic shortest path with adversarially changing costs. Their regret bounds were improved by [Chen et al., 2020], [Chen and Luo, 2021] using a reduction to online loop-free SSP (see next paragraph). As mentioned before, our reduction is different and therefore able to remove the extra \( \sqrt{\log} \) factor in the regret.

Regret minimization in MDPs. There is a vast literature on regret minimization in RL that mostly builds on the optimism principle. Most literature focuses on the tabular setting [Jaksch et al., 2010, Azar et al., 2017, Jin et al., 2018, Fruit et al., 2018, Zanette and Brunskill, 2019, Efroni et al., 2019, Simchowitz and Jamieson, 2019], but recently it was extended to function approximation under various assumptions [Yang and Wang, 2019, Jin et al., 2020b, Zanette et al., 2020a, b].

Online loop-free SSP. A different line of work considers finite-horizon MDPs with adversarially changing costs [Neu et al., 2010, 2012, Zimin and Neu, 2013, Rosenberg and Mansour, 2019b, a, Jin et al., 2020a, Cai et al., 2020, Shani et al., 2020, Lancewicki et al., 2020, Lee et al., 2020, Jin and Luo, 2020]. They refer to finite-horizon adversarial MDPs as online loop-free SSP. This is not to be confused with our setting in which the interaction between the agent and the environment ends only when (and if) the goal state is reached, and not after a fixed number of steps \( H \). See [Rosenberg and Mansour, 2020, Chen et al., 2020] for a discussion on the differences between the models.
2 Preliminaries and main results

An instance of the SSP problem is defined by an MDP $M = (S, A, P, c, s_{init}, g)$ where $S$ is a finite state space and $A$ is a finite action space. The agent begins at an initial state $s_{init} \in S$, and ends her interaction with $M$ by arriving at the goal state $g$ (where $g \notin S$). Whenever she plays action $a$ in state $s$, she pays a cost $C \in [0, 1]$ drawn i.i.d. from a distribution with expectation $c(s, a) \in [0, 1]$ and the next state $s' \in S \cup \{g\}$ is chosen with probability $P(s' \mid s, a)$. Note that the transition function $P$ satisfies $\sum_{s' \in S \cup \{g\}} P(s' \mid s, a) = 1$ for every $(s, a) \in S \times A$.

Proper policies. A stationary and deterministic policy $\pi : S \mapsto A$ is a mapping that selects action $\pi(s)$ whenever the agent is at state $s$. A policy $\pi$ is called proper if playing according to $\pi$ ensures that the goal state is reached with probability 1 when starting from any state (otherwise it is improper). In SSP, the agent has two goals: (a) reach the goal state; (b) minimize the total expected cost. To facilitate the first goal, we make the basic assumption that there exists at least one proper policy. In particular, the goal state is reachable from every state, which is clearly a necessary assumption.

Any policy $\pi$ induces a cost-to-go function $J^\pi : S \mapsto [0, \infty]$. The cost-to-go at state $s$ is defined by $J^\pi(s) = \lim_{T \to \infty} \mathbb{E}_\pi[\sum_{t=1}^T c(s_t, a_t) \mid s_{init} = s]$, where the expectation is taken w.r.t the random sequence of states generated by playing according to $\pi$ when the initial state is $s$. For a proper policy $\pi$, it follows that $J^\pi(s)$ is finite for all $s \in S$. However, note that $J^\pi(s)$ may be finite even if $\pi$ is improper. We additionally denote by $T^\pi(s)$ the expected time it takes for $\pi$ to reach $g$ starting at state $s$; in particular, if $\pi$ is proper then $T^\pi(s)$ is finite for all $s \in S$, and if $\pi$ is improper there must exist some state $s$ such that $T^\pi(s) = \infty$.

Learning formulation. Here, the agent does not have any prior knowledge of the cost function $c$ or transition function $P$. She interacts with the model in episodes: each episode starts at the fixed initial state $s_{init}$ and ends when the agent reaches the goal state $g$ (note that she might never reach the goal state). Success is measured by the agent’s regret over $K$ such episodes, that is the difference between her total cost over the $K$ episodes and the total expected cost of the optimal proper policy:

$$R_K = \sum_{k=1}^K \sum_{i=1}^{I^k} C_i^k - K \min_{\pi \in \Pi_{\text{proper}}} J^\pi(s_{init}),$$

where $I^k$ is the time it takes the agent to complete episode $k$ (which may be infinite), $C_i^k$ is the cost suffered in the $i$-th step of episode $k$ when the agent visited state-action pair $(s_i^k, a_i^k)$, and $\Pi_{\text{proper}}$ is the set of all stationary, deterministic and proper policies (that is not empty by assumption). In the case that $I^k$ is infinite for some $k$, we define $R_K = \infty$.

We denote the optimal proper policy by $\pi^*$, $J^{\pi^*}(s_{init}) = \arg \min_{\pi \in \Pi_{\text{proper}}} J^\pi(s_{init})$. Moreover, let $B_*>0$ be an upper bound on the values of $J^\pi$ and let $T_* > 0$ be an upper bound on the times $T^\pi^*$, i.e., $B_* \geq \max_{s \in S} J^{\pi^*}(s)$ and $T_* \geq \max_{s \in S} T^{\pi^*}(s)$. Finally, let $D = \max_{s \in S} \min_{\pi \in \Pi_{\text{proper}}} T^\pi(s)$ be the SSP-diameter, and note that $B_* \leq D \leq T_*$.

2.1 Summary of our results

In Section 3 we present a novel black-box reduction from SSP to finite-horizon MDPs (Algorithm 1), that yields $\sqrt{K}$ regret bounds when combined with a certain class of optimistic algorithms for regret minimization in finite-horizon MDPs that we call admissible (Definition 1). The regret analysis for the reduction is described in Section 4 and in Section 5 we present an admissible algorithm for regret minimization in finite-horizon MDPs called ULCVI. We show that it guarantees the following optimal regret in the finite-horizon setting (stated formally in Theorem 5.1). Note that (for large enough number of episodes) this bound depends only on the expected cost of the optimal policy and not on the horizon $H$.

**Theorem 2.1.** Running ULCVI (Algorithm 1) in Section 3 in a finite-horizon MDP guarantees, with probability at least $1 - \delta$, a regret bound of

$$O\left(\sqrt{(B_*^2 + B_*|S||A|) \log \frac{MH|S||A|}{\delta}} + H^2 B_*^2 |S|^2 |A| \log \frac{3^2 MH|S||A|}{\delta}\right),$$

The initial state is fixed for simplicity of presentation, but it can be chosen adversarially at the beginning of every episode. Without any change to the algorithm or analysis, the same guarantees hold.
for any number of episodes $M \geq 1$ simultaneously.

Combining ULCVI with our reduction yields the following minimax optimal regret bound for SSP.

**Theorem 2.2.** Running the reduction in Algorithm 1 with the finite-horizon regret minimization algorithm ULCVI ensures, with probability at least $1-\delta$,

$$R_K = O\left(\sqrt{(B_s^2 + B_s)||A||K} \log \frac{KT||A||}{\delta} + T^5||A||^2 \log^{6} \frac{KT||A||^5}{\delta}\right).$$

**Remark 1.** An important observation is that this regret bound is meaningful even for small $K$. Unlike finite-horizon MDPs, where linear regret is trivial, in SSP ensuring finite regret is not easy. Our regret bound also implies that if we play for only one episode, i.e., we are only interested in the time it takes to reach the goal state, then it will take us at most $O(T^5||A||^2)$ time steps to do so.

**Remark 2.** Note that our algorithm needs to know an upper bound on $T_\star$, in advance. However, if all costs are strictly positive (i.e., at least $c_{\min} > 0$), then there is a trivial upper bound of $B_s/c_{\min}$. In this case, our algorithm keeps an optimal regret bound for large enough $K$, since the bound on $T_\star$ only appears in the additive factor. Some previous work used a perturbation argument to generalize their results from the $c_{\min}$ case to general costs [Tarbouriech et al., 2020, Rosenberg et al., 2020, Rosenberg and Mansour [2020]. In our case, it will not work since the dependence on $1/c_{\min}$ in the additive term is too large. This may be an inherent shortcoming of using finite-horizon reduction to solve SSPs, as it also appears in the works of Chen et al. [2020], [Chen and Luo [2021] for the adversarial setting.

**Remark 3.** In practice, one can think of $T_\star$ as a parameter of the algorithm that controls computational complexity and the number of episodes to complete $K$ episodes. By choosing the parameter $T_\star = x$ for example, we can guarantee that the regret bound of Theorem 2.2 holds against the best proper policy with expected time to the goal of at most $x$ (assuming there exists one), and we can also guarantee that the total computational complexity of the algorithm is $O(x \log K)$ (see Remark 5). Furthermore, the algorithm will take at most $O(xK + poly\left(x, ||A||, ||A||\right)$ steps to complete $K$ episodes.

**Remark 4.** While the additive term in our regret bound is standard for most cases, it becomes large when $B_s$ is extremely small because of the dependence in $B_s^{-1}$. This was not an issue in previous work [Tarbouriech et al., 2020, Rosenberg et al., 2020] since they assumed that the costs are deterministic and known. We believe that this dependence is an artifact of our analysis that may be avoided with a more careful definition of $\omega_A$ (see Definition 1) that depends on the actual cost in each state-action pair and not just $B_s$. Nevertheless, the main focus of this paper is on establishing that the minimax optimal regret for SSP is $\tilde{\Theta}(\sqrt{B_s^2 + B_s}||A||K}$, and not on optimizing lower order terms. By that we also show that this is the minimax optimal regret for finite-horizon which is independent of the horizon $H$ (up to logarithmic factors). Tightening the additive term and eliminating its dependence in $B_s^{-1}$ is left as an interesting future direction.

In Appendix D we prove that our regret bound is indeed minimax optimal. To complement the $\Omega(B_s\sqrt{||A||K})$ lower bound of Rosenberg et al. [2020] that assumes $B_s \geq 1$, we provide the following tighter lower bound for the case that $B_s < 1$.

**Theorem 2.3.** Let $B_s \leq \frac{1}{2}$. There exists an SSP problem instance $\mathcal{M} = (S, A, P, c, s_{init}, g)$ in which $J^*(s) \leq B_s$, for all $s \in S$, $||A|| \geq 2$, $K \geq B_s||A||$, such the expected regret of any learner after $K$ episodes satisfies

$$\mathbb{E}[R_K] \geq \frac{1}{32} \sqrt{B_s||A||K}.$$
\( \hat{c}(s) = 8B_s \mathbb{I}\{s \neq g\} \), where \( \mathbb{I}\{s \neq g\} \) is the indicator function that equals 1 if \( s \neq g \) and 0 otherwise. Moreover, \( P, \hat{c} \) are the natural extensions of \( P, c \) to the goal state. That is, \( \hat{c}(s, a) = c(s, a)\mathbb{I}\{s \neq g\} \) and
\[
\hat{P}(s' | s, a) = \begin{cases} 
P(s' | s, a), & s \neq g; \\
1, & s = g, s' = g; \\
0, & s = g, s' \neq g.
\end{cases}
\]
The horizon \( H \) (which we will set to be roughly \( T_s \)) is chosen such that the optimal SSP policy will reach the goal state in \( H \) time steps with high probability (recall that the expected hitting time of the optimal policy is bounded by \( T_s \)). The additional terminal cost is there to encourage the agent to reach the goal state within \( H \) steps, which otherwise is not necessarily optimal with respect to the planning horizon.

**Algorithm 1** REDUCTION FROM SSP TO FINITE-HORIZON MDP

1: **input:** state space \( S \), action space \( A \), initial state \( s_{init} \), goal state \( g \), confidence parameter \( \delta \), number of episodes \( K \), bound on the expected time of the optimal policy \( B_\ast \), bound on the expected time of the optimal policy \( T_s \), and algorithm \( \mathcal{A} \) for regret minimization in finite-horizon MDPs.
2: **initialize** \( \mathcal{A} \) with state space \( \bar{S} = S \cup \{g\} \), action space \( A \), horizon \( H = 8T_s \log(8K) \), confidence parameter \( \delta/4 \), terminal costs \( \hat{c}(s) = 8B_s \mathbb{I}\{s \neq g\} \) and bound on the expected cost of the optimal policy \( 9B_\ast \).
3: **initialize** intervals counter \( m \leftarrow 0 \) and time steps counter \( t \leftarrow 1 \).
4: for \( k = 1, \ldots, K \) do
5: set \( s_t \leftarrow s_{init} \).
6: while \( s_t \neq g \) do
7: set \( m \leftarrow m + 1 \), feed initial state \( s_t \) to \( \mathcal{A} \) and obtain policy \( \pi^m = \{ \pi^m_h : \bar{S} \rightarrow A \}_{h=1}^H \).
8: for \( h = 1, \ldots, H \) do
9: play action \( a_t = \pi^m_h(s_t) \), suffer cost \( C_t \sim c(s_t, a_t) \), and set \( s^m_h = s_t, a^m_h = a_t, C^m_h = C_t \).
10: observe next state \( s_{t+1} \sim P(\cdot | s_t, a_t) \) and set \( t \leftarrow t + 1 \).
11: if \( s_t = g \) then
12: pad trajectory to be of length \( H \) and BREAK.
13: end if
14: end for
15: set \( s^m_{H+1} = s_t \).
16: feed trajectory \( U^m = (s^m_1, a^m_1, \ldots, s^m_H, a^m_H, s^m_{H+1}) \) and costs \( \{C^m_h\}_{h=1}^H \) to \( \mathcal{A} \).
17: end while
18: end for

The algorithm \( \mathcal{A} \) is initialized with the state and action spaces as in the original SSP instance, the horizon length \( H \), a confidence parameter \( \delta/4 \), a set of terminal costs \( \hat{c} \) and a bound on the expected cost of the optimal policy in the finite-horizon model \( 9B_\ast \). At the beginning of each interval, it takes as input an initial state and outputs a policy to be used throughout the interval. In the end of the interval it receives the trajectory and costs observed through the interval.

Note that while Algorithm 1 may run any finite-horizon regret minimization algorithm, in the analysis we require that \( \mathcal{A} \) possesses some properties (that most optimistic algorithms already have) in order to establish our regret bound. We specifically require \( \mathcal{A} \) to be an *admissible* algorithm—a model-based optimistic algorithm for regret minimization in finite-horizon MDPs, e.g., UCBVI [Azar et al. 2017] and EULER [Zanette and Brunskill 2019]. Admissible algorithms are defined formally as follows.

**Definition 1.** A model-based algorithm \( \mathcal{A} \) for regret minimization in finite-horizon MDPs is called *admissible* if, when running \( \mathcal{A} \) with confidence parameter \( \delta \), there is a good event that holds with probability at least \( 1 - \delta \), under which the following hold:

(i) \( \mathcal{A} \) provides anytime regret guarantees without prior knowledge of the number of episodes, and when the initial state of each episode is arbitrary. The regret bound that \( \mathcal{A} \) guarantees for \( M \) episodes is denoted by \( \tilde{R}_M(M) \), for some non-decreasing function \( \tilde{R}_M \).
(ii) The policy \( \pi^m \) that \( \mathcal{A} \) picks in episode \( m \) is greedy with respect to an estimate of the optimal policy’s \( Q \)-function.
(iii) The algorithm’s estimate \( \hat{J}^m \) of \( J^* \) (the cost-to-go function associated with the optimal finite-horizon policy) is optimistic, i.e., \( \hat{J}_h^m(s) \leq \hat{J}_h^m(s) \) for every \( s \in S \) and \( h = 1, \ldots, H + 1 \).
We operate in

We can obtain a reasonable estimate (up to a constant multiplicative factor) of the cost-to-go from state

Let

Theorem 3.1. The computational complexity of

is

visited in interval

for interval

and the deviation of the actual cost in each interval from its expected value. To that end, we use

regret can be bounded by the sum of two terms: the expected regret of the finite-horizon algorithm,

the details to Appendix A. We start the analysis with a regret decomposition that states that the SSP

3.1 Unknown expected optimal cost

Inspired by techniques for estimation of the SSP-diameter in the adversarial SSP literature [Rosenberg and Mansour, 2020, Chen and Luo, 2021], in Appendix C, we show that our reduction does not need to know

B

in advance, but can instead estimate it on the fly.

We can obtain a reasonable estimate (up to a constant multiplicative factor) of the cost-to-go from state

s

by running the Bernstein-SSP algorithm of [Rosenberg et al., 2020] for regret minimization in SSPs (that does not need to know

B

= 0) with initial state

s

for roughly

T

= \frac{1}{9H}

episodes, where

T

is

2

log \left( \frac{T \cdot |S| \cdot |A| \cdot \omega_{\mathcal{A}}} {\delta} \right)

The computational complexity of \textsc{ULCVI} is \( O(H|S|^3 |A|^2 \log(MH)) \), and therefore our optimal regret for SSP is achieved in total computational complexity of \( O(T^6 |S|^3 |A|^2 \log^2 K \omega_{\mathcal{A}}^2 \delta) \) which is only logarithmic in the number of episodes.

3.1 Unknown expected optimal cost

4 Regret analysis

In this section we prove Theorem 3.1. Below we give a high-level overview of the proofs and defer the details to Appendix A. We start the analysis with a regret decomposition that states that the SSP

regret can be bounded by the sum of two terms: the expected regret of the finite-horizon algorithm, and the deviation of the actual cost in each interval from its expected value. To that end, we use the notations: \( M \) for the total number of intervals, \( U^m = (u^m_1, u^m_2, \ldots, u^m_m) \) for the trajectory visited in interval \( m \), \( C^m_h \) for the cost suffered in step \( h \) of interval \( m \), \( \pi^m \) for the policy chosen by \( \mathcal{A} \) for interval \( m \), and \( J^\pi_h(s) \) for the expected finite-horizon cost when playing policy \( \pi \) starting from state \( s \) in time step \( h \).
Lemma 4.1. For $H = 8T$, $\log(8K)$, we have the following bound on the regret of Algorithm 1:

$$R_K \leq \hat{R}_A(M) + \sum_{m=1}^{M} \left( \sum_{h=1}^{H} C_m^h + \hat{c}_f(s^m_{h+1}) - \hat{J}_1^m(s^m_0) \right) + B_\star.$$  (1)

The bound in Eq. (1) is comprised of two summands and an additional constant. The first summand is an upper bound on the expected cost of the optimal policy in the SSP model $M$ (notice that $M$ (Definition 1). Note that this bound is in terms of the number of intervals $M$ which is a random variable and not necessarily bounded. In what follows we show, using the admissibility of $A$, we can actually bound $M$ by the number of SSP episodes $K$ plus a constant that depends on $\omega_A, I, |A|, T_\star$ (but not on $K$). The second summand in Eq. (1) relates to the deviation of the total finite-horizon cost from its expected value.

The proof of Lemma 4.1 builds on two key ideas. The first is that, by setting $H$ to be $O(T_\star, \log K)$, we ensure that the expected cost of the optimal policy in the SSP model $M$ is close to that in the finite-horizon model $\hat{M}$. The second idea is that if the agent does not reach the goal state in a certain interval, then she must suffer the terminal cost in the finite-horizon model. Therefore, although in a single episode there may be many intervals in which the agent does not reach the goal state, we can upper bound the cost in these extra intervals in $M$ by the corresponding terminal costs in $\hat{M}$.

Next, we bound the deviation of the actual cost in each interval from its expected value which appears as the second summand in Eq. (1). The bound is due to the following lemma.

Lemma 4.2. Assume that the reduction is performed using an admissible algorithm $A$. Then, the following holds with probability at least $1 - 3^H$, $\delta$,

$$\sum_{m=1}^{M} \left( \sum_{h=1}^{H} C_m^h + \hat{c}_f(s^m_{h+1}) - \hat{J}_1^m(s^m_0) \right) = O\left( \sqrt{(B_\star^2 + B_\star) M \log \frac{M}{\delta} + H \omega_A |I| |A| \log \frac{M K T_\star |I| |A|}{\delta}} \right).$$

The key observation here relies on the notion of unknown state-action pairs – pairs that were not visited at least $\omega_A$ times. After $\omega_A$ visits to some state-action pair $s, a$, we have a reasonable estimate of the next-state distribution $P(\cdot | s, a)$ therefore we can show that the expected accumulated cost in an interval until reaching an unknown state-action pair or the goal state is of order $B_\star$. Moreover, the second moment of this cost is of order $B_\star^2 + B_\star$. Thus, using Freedman inequality, we bound the deviation by $O(\sqrt{(B_\star^2 + B_\star) M})$, plus a cost of $O(H)$ for each “bad” interval in which we do not reach an unknown state-action pair or the goal state (there are roughly $\omega_A |I| |A|$ such intervals).

Lastly, we need to bound the number of intervals $M$ to obtain a regret bound in terms of $K$ and not $M$ (notice that $M$ is a random variable that is not bounded a-priori).

Lemma 4.3. Assume that the reduction is performed using an admissible algorithm $A$. Then, with probability at least $1 - 3^H$, $M \leq 4K + 4 \cdot 10^4 |I| |A| \omega_A \log(KT_\star |I| |A| / \delta)$.

The proof shows that in every interval there is a constant probability to reach the goal state or an unknown state-action pair. Leveraging this observation with a concentration inequality, we can bound the number of intervals by $O(K + \omega_A |I| |A| H)$.

We can now prove a bound on the regret of Algorithm 1 using any admissible algorithm $A$.

Proof of Theorem 3.1. The regret bound of $A$, Lemmas 4.2 and 4.3 all hold with probability at least $1 - \delta$, via a union bound. Using Lemmas 4.1 and 4.2, we can write

$$R_K \leq \hat{R}_A(M) + O\left( \sqrt{(B_\star^2 + B_\star) M \log \frac{M}{\delta} + H \omega_A |I| |A| \log \frac{M K T_\star |I| |A|}{\delta}} \right) + B_\star.$$  

Finally, we use Lemma 4.3 to bound $M$ by $4K + 4 \cdot 10^4 |I| |A| \omega_A \log(KT_\star |I| |A| / \delta)$.

5 ULCVI: an admissible algorithm for finite-horizon MDPs

In this section we present the Upper Lower Confidence Value Iteration algorithm (ULCVI; Algorithm 2) for regret minimization in finite-horizon MDPs. This result holds independently of our SSP
Algorithm 2 Upper Lower Confidence Value Iteration (ULCVI)

1: input: state space $S$, action space $A$, horizon $H$, confidence parameter $\delta$, terminal costs $\hat{c}_f$ and upper bound on the expected cost of the optimal policy $B_\star$.
2: initialize: $n^0(s, a) = 0, N^0(s, a, s') = 0, N^0(s, a, s') = 0 \forall (s, a, s') \in S \times A \times S$.
3: initialize: $C^0(s, a) = 0, \bar{C}^0(s, a) = 0, \bar{P}^0(s'|s, a) = I(s' = s) \forall (s, a, s') \in S \times A \times S$.
4: initialize: PlanningTrigger = true.
5: for $m = 1, 2, \ldots$ do
6: observe initial state $s^m_1$.
7: if PlanningTrigger = true then
8: set $n^{m-1}(s, a) \leftarrow N^{m-1}(s, a), n^{m-1}(s, a, s') \leftarrow N^{m-1}(s, a, s') \forall (s, a, s')$.
9: set $\bar{P}^m(s'|s, a) \leftarrow \frac{e^{m-2}(s'|s, a)}{\max \{1, e^{m-2}(s'|s, a)\}} \max \{1, e^{m-1}(s, a)\} \forall (s, a, s')$.
10: compute $\{\pi^m_\pi(s)\}_{s \in S}$ via Optimistic-Pessimistic Value Iteration (Algorithm 3).
11: set PlanningTrigger $\leftarrow$ false.
12: else
13: set $n^{m-1}(s, a) \leftarrow n^{m-2}(s, a, s') \leftarrow n^{m-2}(s, a, s') \forall (s, a, s')$.
14: set $\bar{P}^m(s'|s, a) \leftarrow \bar{P}^m(s'|s, a), \bar{C}^m(s, a) \leftarrow \bar{C}^m(s, a) \forall (s, a, s')$.
15: set $\pi^m_\pi(s) \leftarrow \max \{1, \pi^m_\pi(s)\}$ for all $s \in S$ and $h = 1, \ldots, H$.
16: end if
17: set $N^m(s, a) \leftarrow N^{m-1}(s, a, s'), C^m(s, a) \leftarrow C^{m-1}(s, a) \forall (s, a, s')$.
18: for $h = 1, \ldots, H$ do
19: pick action $a^m_h = \pi^m_\pi(s^m_h)$.
20: suffer cost $C^m_h \sim \bar{C}(s^m_h, a^m_h)$ and observe next state $s^{m+1}_h \sim \bar{P}(s^m_h, a^m_h)$.
21: update visits counters $n^{m+1}(s^m_h, a^m_h) \leftarrow n^m(s^m_h, a^m_h) + 1, n^m(s^m_h, a^m_h, s^{m+1}_h) \leftarrow n^m(s^m_h, a^m_h, s^{m+1}_h) + 1$.
22: update accumulated cost $C^m_{h+1} \leftarrow C^m(s^m_h, a^m_h) + C^m_{h+1}$.
23: if $N^m(s^m_h, a^m_h) \geq 2n^{m-1}(s^m_h, a^m_h)$ then
24: set PlanningTrigger $\leftarrow$ true.
25: end if
26: end for
27: Suffer terminal cost $\hat{c}_f(s^m_{H+1})$.
28: end for

algorithm. Since the algorithm is similar to previous optimistic algorithms for the finite-horizon setting, e.g., UCBVI [Azar et al., 2017] and ORLC [Dann et al., 2019], we defer the analysis to Appendix B and focus on our technical novelty – bounding the regret in terms of the optimal value function and not the horizon.

In each episode $m$, the ULCVI algorithm maintains an optimistic lower bound $L^m(s)$ and a pessimistic upper bound $J^m(s)$ on the cost-to-go function of the optimal policy $J^*_f(s)$, and acts greedily with respect to the optimistic estimates. These optimistic and pessimistic estimates are computed based on the empirical transition function $\bar{P}^{m-1}(s'|s, a)$ and the empirical cost function $\bar{C}^{m-1}(s, a)$ to which we add an exploration bonus $b^m_c(s, a) + b^m_p(s, a)$, where $b^m_c$ handles the approximation error in the transitions and $b^m_p$ handles the approximation error in the costs. The bonuses are defined as follows,

\[
b^m_c(s, a) = \sqrt{\frac{2 \text{Var}_{s,a}^m(C)L^m}{\max \{1, n^{m-1}(s, a)\}} + \frac{5L^m}{\max \{1, n^{m-1}(s, a)\}}} \\
b^m_p(s, a) = \sqrt{\frac{2 \text{Var}_{s,a}^m(C)L^m}{\max \{1, n^{m-1}(s, a)\}} + \frac{62H^2B^m\bar{C}^{m-1}(s')L^m}{\max \{1, n^{m-1}(s, a)\}} + \frac{B^m}{16H^2}E_{C^{m-1}(s', a)}[J^m_{h+1}(s') - J^m_{h+1}(s')]^2},
\]

where $L^m = 3 \log(3\text{Var}_{s,a}^m(C)/Hm/\delta)$ is a logarithmic factor and $n^{m-1}(s, a)$ is the number of visits to $(s, a)$ in the first $m - 1$ episodes. Furthermore, $\text{Var}_{s,a}^m(C)$ is the empirical variance of the observed costs in $(s, a)$ in the first $m - 1$ episodes $[8]$. Lastly, the term $\text{Var}_{C^{m-1}(s', a)}^m(J^m_{h+1})$ is the variance of the next state value $J^m_{h+1}$ from state-action pair $(s, a)$, calculated via the empirical transition model, i.e., $\text{Var}_{C^{m-1}(s', a)}^m(J^m_{h+1}) = E_{C^{m-1}(s', a)}[J^m_{h+1}(s')^2] - E_{C^{m-1}(s', a)}[J^m_{h+1}(s')]^2$.

The empirical variance of $n$ numbers $a_1, \ldots, a_n$ is defined by $\frac{1}{n} \sum_{i=1}^{n} (a_i - \frac{1}{n} \sum_{j=1}^{n} a_j)^2$. 

8
We prove the following high probability regret bound.

**Algorithm 3** Optimistic-Pessimistic Value Iteration

1. input: \( n \), \( p \), \( c \), \( \hat{c} \), \( B \).
2. initialize \( J^m_{H+1} = J^m_H = \hat{c}_f \) for all \( s \in S \).
3. for \( h = H, H-1, \ldots, 1 \) do
4. \hspace{1em} for \( s \in S \) do
5. \hspace{2em} set the bonus \( b_h^m(s, a) = b^m_c(s, a) + b^m_h(s, a) \) defined in Eq. (2).
6. \hspace{2em} compute optimistic and pessimistic Q-functions:
   \[
   Q^m_h(s, a) = \hat{c}^{m-1}(s, a) - b^m_h(s, a) + \mathbb{E}_{p \cdot (\cdot, a)}[J^m_{h+1}(s')],
   \]
   \[
   \hat{Q}^m_h(s, a) = \hat{c}^{m-1}(s, a) + b^m_h(s, a) + \mathbb{E}_{p \cdot (\cdot, a)}[J^m_{h+1}(s')].
   \]
7. \hspace{2em} end for
8. \hspace{1em} \( \pi^m_h(s) = \arg\min_{a \in A} Q^m_h(s, a) \).
9. \hspace{1em} \( J^m_h(s) = \max \{ Q^m_h(s, \pi^m_h(s)), 0 \} \), \( J^m_h(s) = \min \{ \hat{Q}^m_h(s, \pi^m_h(s)), H \} \).
10. end for
11. end for

For improved computational complexity, we compute the optimistic policy only in episodes in which the number of visits to some state-action pair was doubled. This ensures that the number of optimistic policy computations grows only logarithmically with the number of episodes, i.e., it is bounded by \( 3\|S\|\|A\| \log \left( MH \right) \). Since each optimal policy computation costs \( O(H\|S\|^2\|A\|) \) in the finite-horizon MDP model, our algorithm enjoys a total computational complexity of \( O(H\|S\|^3\|A\|^2 \log \left( MH \right) \).

For clarity, we keep the notation of the finite-horizon MDP as \( \tilde{M} = (S, A, \tilde{P}, \tilde{c}, \tilde{c}_s) \), and let \( B_* = \max_{s, a} J^\pi(s) \) where \( J^\pi \) is the value function of policy \( \pi \) (in the case of our SSP reduction this parameter is simply \( 9B_* \) by Lemma [A.1].) This implies that \( \hat{c}_f(s) \leq B_* \) for every \( s \), and for simplicity, we assume that \( B_* \leq H \). Thus, the maximal total cost in an episode is bounded by \( H + B_* \leq 2H \). In Appendix [B] we prove the following high probability regret bound.

**Theorem 5.1.** ULCVI (Algorithm 2) is admissible with the following guarantees:

(i) With probability at least \( 1 - \delta \), the regret bound of ULCVI is

\[
\hat{R}_{ULCVI}(M) = O \left( \sqrt{(B_*^2 + B_*)\|S\|\|A\| M \log \frac{M\|S\|\|A\|}{\delta}} + H^4 B_*^{-1}\|S\|^2\|A\| \log^{3/2} \frac{M\|S\|\|A\|}{\delta} \right)
\]

for any number of episodes \( M \geq 1 \).

(ii) \( \omega_{ULCVI} = O(H^4 B_*^{-1}\|S\|) \).

Our analysis resembles the one in [Efroni et al. 2021], and is adapted to the stationary MDP setting (i.e., the transition function does not depend on the time step \( h \)), and to the setting where we have costs instead of rewards, and terminal costs (which do not appear in previous work). By the definition of the algorithm and the regret bound in Theorem 5.1, it is clear that properties (i)-(iii) in Definition [A.1] of admissible algorithms hold. For property (iv), we use standard concentration inequalities and the definition of the bonuses in Eq. (2) in order to show it holds for \( \omega_{ULCVI} = O(H^4 B_*^{-1}\|S\|) \).

To obtain a regret bound whose leading term depends on \( B_* \) and not \( H \), we start with a standard regret analysis for optimistic algorithms that establishes the regret scales with the square-root of the variance of the value functions of the agent’s policies, i.e.,

\[
\hat{R}_{ULCVI}(M) \lesssim \sqrt{\|S\|\|A\|} \sum_{m=1}^{M} \sum_{h=1}^{H} \text{Var}_{\tilde{P} \cdot (\cdot, a)}(J^m_{h+1}) + H^4 B_*^{-1}\|S\|^2\|A\|,
\]

up to logarithmic factors and lower order terms. This can be further bounded by the second moment of the cumulative cost in each episode as follows,

\[
\hat{R}_{ULCVI}(M) \lesssim \sqrt{\|S\|\|A\|} \sum_{m=1}^{M} \mathbb{E} \left[ \left( \sum_{h=1}^{H} C^m_h + \hat{c}_f(s^m_{H+1}) \right)^2 \left| \mathcal{U}^m \right) \right] + H^4 B_*^{-1}\|S\|^2\|A\|,
\]

up to logarithmic factors and lower order terms.
where $\bar{U}_m$ is the sequence of state-action pairs observed up to episode $m$. Leveraging our techniques for the SSP reduction (but independently), we show that the second moment of the cumulative cost until an unknown state-action pair is reached can be bounded by $O(B^2 \star + B \star)$. Therefore, we have at most $O(H^2B^2 |S||A|)$ episodes in which we bound the second moment trivially by $O(H^2)$, and in the rest of the episodes we can bound it by $O(B^2 \star + B \star)$. Together this yields the theorem as follows,

$$\hat{R}_{ULCVI} (M) \lesssim \sqrt{|S||A|} \sqrt{(B^2 \star + B \star)M + H^2 \cdot H^4B^2 |S||A|} \lesssim \sqrt{(B^2 \star + B \star)|S||A|M + H^4B^2 |S||A|}.$$

**Acknowledgements**

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University

**References**


