Near-optimal Regret Bounds for Stochastic Shortest Path

Alon Cohen [1] **Haim Kaplan** [1 2] **Yishay Mansour** [1 2] **Aviv Rosenberg** [2]

Abstract

Stochastic shortest path (SSP) is a well-known problem in planning and control, in which an agent has to reach a goal state in minimum total expected cost. In the learning formulation of the problem, the agent is unaware of the environment dynamics (i.e., the transition function) and has to repeatedly play for a given number of episodes, while learning the problem's optimal solution. Unlike other well-studied models in reinforcement learning (RL), the length of an episode is not predetermined (or bounded) and is influenced by the agent's actions. Recently, Tarbouriech et al. (2020) studied this problem in the context of regret minimization, and provided an algorithm whose regret bound is inversely proportional to the square root of the minimum instantaneous cost. In this work we remove this dependence on the minimum cost—we give an algorithm that guarantees a regret bound of $\widetilde{O}(B_\star|S|\sqrt{|A|K})$, where $B_\star$ is an upper bound on the expected cost of the optimal policy, S is the set of states, A is the set of actions and K is the number of episodes. We additionally show that any learning algorithm must have at least $\Omega(B_\star\sqrt{|S||A|K})$ regret in the worst case.

1. Introduction

Stochastic shortest path (SSP) is one of the most basic models in reinforcement learning (RL). It includes the discounted return model and the finite-horizon model as special cases. In SSP the goal of the agent is to reach a predefined goal state in minimum expected cost. This setting captures a wide variety of realistic scenarios, such as car navigation, game playing and drone flying; i.e., tasks carried out in episodes that eventually terminate.

[1] Google Research, Tel Aviv [2]Tel Aviv University, Israel. Correspondence to: Alon Cohen <aloncohen@google.com>, Aviv Rosenberg <avivros007@gmail.com>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020.

The focus of this work is on regret minimization in SSP. It builds on extensive literature on theoretical aspects of online RL, and in particular on the copious works about regret minimization in either the average cost model or the finite-horizon model. A major contribution to this literature is the UCRL2 algorithm (Jaksch et al., 2010) that gives a general framework to achieve optimism in face of uncertainty for these settings. The main methodology is to define a confidence set that includes the true model parameters with high probability. The algorithm periodically computes an optimistic policy that minimizes the overall expected cost simultaneously over all policies and over all parameters within the confidence set, and proceeds to play this policy.

The only regret minimization algorithm specifically designed for SSP is that of Tarbouriech et al. (2020) that assumes that all costs are bounded away from zero (i.e., there is a $c_{\min} > 0$ such that all costs are in the range $[c_{\min}, 1]$). They show a regret bound that scales as $\widetilde{O}(D^{3/2}|S|\sqrt{|A|K/c_{\min}})$ where D is the minimum expected time of reaching the goal state from any state, S is the set of states, A is the set of actions and K is the number of episodes. In addition, they show that the algorithm's regret is $\widetilde{O}(K^{2/3})$ when the costs are arbitrary (namely, may be zero).

Here we improve upon the work of Tarbouriech et al. (2020) in several important aspects. First, we remove the dependency on $c_{\min}^{-1}$ and allow for zero costs while maintaining regret of $\widetilde{O}(\sqrt{K})$. Second, we give a much simpler algorithm in which the computation of the optimistic policy has a simple solution.[1] Our main regret term is $\widetilde{O}(B_\star|S|\sqrt{|A|K})$, where $B_\star$ is an upper bound on the expected cost of the optimal policy (note that $B_\star \leq D$). We show that this is almost optimal by giving a lower bound of $\Omega(B_\star\sqrt{|S||A|K})$.

In our work, we obtain a major improvement in the regret bound through the use of confidence sets that are based on Bernstein inequality (Azar et al., 2017), that is highly sensitive to variance, instead of Hoeffding inequality. In both our algorithm and the one of Tarbouriech et al. (2020), the regret scales with the square root of the total variance,

[1]There is a technical issue with the simpler optimistic policy computation. To get the regret bounds described in this paper, we need to compute the optimistic policy similarly to Tarbouriech et al. (2020) (see Section 3). We keep the simpler computation method in this paper, in hopes it can be proved to be sufficient.

i.e., $\widetilde{O}(\sqrt{\text{total variance}})$, where in variance we refer to the variance of the cost-to-go function (see Section 2) at a next-state given some state-action pair visited by the algorithm. When using Hoeffding-based confidence sets, similarly to UCRL2, this variance is trivially bounded by $B_\star^2$ at each step, which leads to a regret of $\widetilde{O}(\sqrt{B_\star^2 T})$, where T is the number of time-steps taken by the algorithm. However, the use of Bernstein inequalities enables us to bound the total expected variance in a time interval, of roughly $B_\star/c_{\min}$ timesteps, by an identical magnitude of $O(B_\star^2)$. Therefore, the regret bound for our algorithm improves upon the regret of Tarbouriech et al. (2020) by a factor of $\sqrt{B_\star/c_{\min}}$, that is, $\widetilde{O}(B_\star|S|\sqrt{|A|K})$ compared to $\widetilde{(D^{3/2}}|S|\sqrt{|A|K/c_{\min}})$.

Our technical contribution is as follows. To better explain our main Bernstein-based algorithm, we start by assuming that the costs are lower bounded by $c_{\min}$ and give an algorithm based on Hoeffding inequalities that is simple to analyze and achieves a regret bound of $\widetilde{O}(B_\star^{3/2}|S|\sqrt{|A|K/c_{\min}})$. Note that this bound is comparable to the one of Tarbouriech et al. (2020), yet our algorithm and its analysis are significantly simpler and more intuitive. In addition, its analysis contains many of the key ideas of the proof of the Bernstein-based algorithm, and is much easier to follow. We subsequently present the Bernstein-based algorithm. This algorithm is simpler than our first one mainly since picking the parameters of the optimistic model is particularly easy. The analysis, however, is somewhat more delicate (as mentioned earlier) – in the main body of the paper, we present only the main ideas and improvements over the Hoeffding-based algorithm, and differ the tedious technical details to the supplementary material. Eventually, we achieve our final regret bound by perturbing the instantaneous costs to be at least $\epsilon > 0$. The additional cost due to this perturbation has a small effect since the dependency of our regret on $c_{\min}^{-1}$ is additive and does not multiply any term depending on K.

1.1. Related work.

Early work by Bertsekas & Tsitsiklis (1991) studied the problem of planning in SSPs, that is, computing the optimal strategy efficiently in a known SSP instance. They established that, under certain assumptions, the optimal strategy is a deterministic stationary policy (a mapping from states to actions) and can be computed efficiently using standard planning algorithms, e.g., Value Iteration or Policy Iteration.

The extensive literature about regret minimization in RL focuses on the average-cost infinite-horizon model (Bartlett & Tewari, 2009; Jaksch et al., 2010; Zhang & Ji, 2019) and on the finite-horizon model (Osband et al., 2016; Azar et al., 2017; Dann et al., 2017; Zanette & Brunskill, 2019; Efroni et al., 2019). These recent works give algorithms with near-optimal regret bounds using Bernstein-type concentration bounds.

Another related model is that of loop-free SSP with adversarial costs (Neu et al., 2010; 2012; Zimin & Neu, 2013; Rosenberg & Mansour, 2019a;b; Jin & Luo, 2019). This model eliminates the challenge of avoiding policies that never terminate, but the adversarial costs pose a different, unrelated, challenge.

2. Preliminaries and Main Results

An instance of the SSP problem is a Markov decision process (MDP) $M = (S, A, P, c, s_{\text{init}})$ where S is the state space and A is the action space. The agent begins at the initial state s_{init}, 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 $c(s,a) \in [0,1]$ and the next state $s' \in S$ is chosen with probability $P(s' \mid s, a)$. Note that to simplify the presentation we avoid addressing the goal state g explicitly – we assume that the probability of reaching the goal state by playing action a at state s is $1 - \sum_{s' \in S} P(s' \mid s, a)$.

We now review planning in a known SSP instance. Under certain assumptions that we shall briefly discuss, the optimal behaviour of the agent, i.e., the policy that minimizes the expected total cost of reaching the goal state from *any* state, is a stationary, deterministic and proper policy. 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 proper policy is defined as follows.

Definition 1 (Proper and Improper Policies)**.** A policy π is *proper* if playing π reaches the goal state with probability 1 when starting from any state. A policy is *improper* if it is not proper.

Any policy π induces a *cost-to-go function* $J^\pi : S \mapsto [0, \infty]$ defined as $J^\pi(s) = \lim_{T\to\infty} \mathbb{E}_\pi\left[\sum_{t=1}^{T} c(s_t, a_t) \mid s_1 = s\right]$, where the expectation is taken w.r.t the random sequence of states generated by playing according to π when the initial state is s. For a proper policy π, since the number of states $|S|$ is finite, it follows that $J^\pi(s)$ is finite for all $s \in S$. However, note that $J^\pi(s)$ may be finite even if π is improper. We additionally denote by $T^\pi(s)$ the expected time it takes for π to reach g starting at s; in particular, if π is proper then $T^\pi(s)$ is finite for all s, and if π is improper there must exist some s such that $T^\pi(s) = \infty$. In this work we assume the following about the SSP model.

Assumption 1. There exists at least one proper policy.

With Assumption 1, we have the following important properties of proper policies. In particular, the first result shows that a policy is proper if and only if its cost-to-go function satisfies the Bellman equations. The second result proves that a policy is optimal if and only if it satisfies the Bellman optimality criterion. Note that they assume that every improper policy has high cost.

Lemma 2.1 (Bertsekas & Tsitsiklis, 1991, Lemma 1)**.** *Suppose that Assumption 1 holds and that for every improper policy π' there exists at least one state $s \in S$ such that $J^{\pi'}(s) = \infty$. Let π be any policy, then*

(i) If there exists some $J : S \mapsto \mathbb{R}$ such that $J(s) \geq c\big(s, \pi(s)\big) + \sum_{s' \in S} P\big(s' \mid s, \pi(s)\big) J(s')$ for all $s \in S$, then π is proper. Moreover, it holds that $J^{\pi}(s) \leq J(s)$, $\forall s \in S$.

(ii) If π is proper then J^{π} is the unique solution to the equations $J^{\pi}(s) = c\big(s, \pi(s)\big) + \sum_{s' \in S} P\big(s' \mid s, \pi(s)\big) J^{\pi}(s')$ for all $s \in S$.

Lemma 2.2 (Bertsekas & Tsitsiklis, 1991, Proposition 2)**.** *Under the conditions of Lemma 2.1 the optimal policy $\pi^\star$ is stationary, deterministic, and proper. Moreover, a policy π is optimal if and only if it satisfies the Bellman optimality equations for all $s \in S$:*

$$J^{\pi}(s) = \min_{a \in A} c\big(s, a\big) + \sum_{s' \in S} P\big(s' \mid s, a\big) J^{\pi}(s'), \tag{1}$$

$$\pi(s) \in \arg\min_{a \in A} c\big(s, a\big) + \sum_{s' \in S} P\big(s' \mid s, a\big) J^{\pi}(s').$$

In this work we are not interested in approximating the optimal policy overall, but rather the best *proper* policy. In this case the second requirement in the lemmas above, that for every improper policy π there exists some state $s \in S$ such that $J^{\pi}(s) = \infty$, can be circumvented in the following way (Bertsekas & Yu, 2013). First, note that this requirement is trivially satisfied when all instantaneous costs are strictly positive. Then, one can perturb the instantaneous costs by adding a small positive cost $\epsilon \in [0, 1]$, i.e., the new cost function is $c_\epsilon(s, a) = \max\{c(s, a), \epsilon\}$. After this perturbation, all proper policies remain proper, and every improper policy has infinite cost-to-go from some state (as all costs are positive). In the modified MDP, we apply Lemma 2.2 and obtain an optimal policy $\pi^\star_\epsilon$ that is stationary, deterministic and proper and has a cost-to-go function $J^\star_\epsilon$. Taking the limit as $\epsilon \to 0$, we have that $\pi^\star_\epsilon \to \pi^\star$ and $J^\star_\epsilon \to J^{\pi^\star}$, where $\pi^\star$ is the optimal *proper* policy in the original model that is also stationary and deterministic, and $J^{\pi^\star}$ denotes its cost-to-go function. We use this observation to obtain Corollaries 2.5 and 2.6 below that only require Assumption 1 to hold.

Learning formulation. We assume that the costs are deterministic and known to the learner, and the transition probabilities P are fixed but unknown to the learner. The learner interacts with the model in episodes: each episode starts at the initial state s_{init}, and ends when the learner reaches the goal state g (note that she might *never* reach the goal state). Success is measured by the learner'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:

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

where I^k is the time it takes the learner to complete episode k (which may be infinite), Π_{proper} is the set of all stationary, deterministic and proper policies (that is not empty by Assumption 1), and (s_i^k, a_i^k) is the i-th state-action pair at episode k. In the case that I^k is infinite for some k, we define $R_K = \infty$.

We denote the optimal proper policy by $\pi^\star$, i.e., $J^{\pi^\star}(s) = \arg\min_{\pi \in \Pi_{\text{proper}}} J^{\pi}(s)$ for all $s \in S$. Moreover, let $B_\star > 0$ be an upper bound on the values of $J^{\pi^\star}$ and let $T_\star > 0$ be an upper bound on the times $T^{\pi^\star}$, i.e., $B_\star \geq \max_{s \in S} J^{\pi^\star}(s)$ and $T_\star \geq \max_{s \in S} T^{\pi^\star}(s)$.

2.1. Summary of our results

In Section 3 we present our Hoeffding-based algorithms (Algorithms 1 and 3) and their analysis. While they achieve similar regret bounds to Tarbouriech et al. (2020), their presentation is important in order to lay the foundations for our Bernstein-based algorithm (Algorithm 2) and its improved regret bound shown in Section 4. Finally, in Section 5 we give a lower bound on the learner's regret showing that Algorithm 2 is near-optimal.

The learner must reach the goal state otherwise she has infinite regret. Therefore, she has to trade-off two objectives, one is to reach the goal state and the other is to minimize the cost. Under the following assumption, the two objectives essentially coincide.

Assumption 2. All costs are positive, i.e., there exists $c_{\min} > 0$ such that $c(s, a) \geq c_{\min}$ for every $(s, a) \in S \times A$.

This assumption allows us to upper bound the running time of the algorithm by its total cost up to a factor of $c_{\min}^{-1}$. In particular, it guarantees that any policy that does not reach the goal state has infinite cost, so any bounded regret algorithm has to reach the goal state. We eventually relax Assumption 2 by a technique similar to that of Bertsekas & Yu (2013). We add a small positive perturbation to the instantaneous costs and run our algorithms on the model with the perturbed costs. This provides a regret bound that scales with the expected running time of the optimal policy.

We now summarize our results. For ease of comparison, we first present our regret bounds for both the Hoeffding and Bernstein-based algorithms when Assumption 2 holds, and subsequently show the regret bounds of both algorithms for the general case. In order to simplify the presentation of our results, we assume that $|S| \geq 2$, $|A| \geq 2$ and $K \geq |S|^2|A|$ throughout. In addition, we denote

$L = \log(KB_\star|S||A|/\delta c_{\min})$. The complete proofs of all statements is found in the supplementary material.

Positive costs. The following results hold when Assumption 2 holds (recall that we always assume Assumption 1). In particular, when this assumption holds the optimal policy overall is proper (Lemma 2.2) hence the regret bounds below are with respect to the best overall policy.

Theorem 2.3. *Suppose that Assumption 2 holds. With probability at least $1-\delta$ the regret of Algorithm 3 is bounded as follows:*

$$R_K = O\left(\sqrt{\frac{B_\star^3|S|^2|A|K}{c_{min}}}L + \frac{B_\star^3|S|^2|A|}{c_{min}^2}L^2\right).$$

The main issue with the regret bound in Theorem 2.3 is that it scales with $\sqrt{K/c_{\min}}$ which cannot be avoided regardless of how large K is with respect to $c_{\min}^{-1}$. This problem is alleviated in Algorithm 2 that uses the tighter Bernstein-based confidence bounds.

Theorem 2.4. *Assume that Assumption 2 holds. With probability at least $1-\delta$ the regret of Algorithm 2 is bounded as follows:*

$$R_K = O\left(B_\star|S|\sqrt{|A|K}L + \sqrt{\frac{B_\star^3|S|^4|A|^2}{c_{min}}}L^2\right).$$

Note that when $K \gg B_\star|S|^2|A|/c_{\min}$, the regret bound above scales as $\widetilde{O}(B_\star|S|\sqrt{|A|K})$ thus obtaining a near-optimal rate.

Arbitrary costs. Recall that in this case we can no longer assume that the optimal policy is proper. Therefore, the regret bounds below are with comparison to the best *proper* policy. Assumption 2 can be easily alleviated by adding a small fixed cost to the cost of all state-action pairs. Following the perturbation of the costs, we obtain regret bounds from Theorems 2.3 and 2.4 with $c_{\min} \leftarrow \epsilon$ and $B_\star \leftarrow B_\star + \epsilon T_\star$, and the learner also suffers an additional cost of $\epsilon T_\star K$ due to the misspecification of the model caused by the perturbation. By picking ϵ to balance these terms we get the following corollaries (letting $\widetilde{L} = \log(KB_\star T_\star|S||A|/\delta)$).

Corollary 2.5. *Running Algorithm 3 using costs $c_\epsilon(s,a) = \max\{c(s,a),\epsilon\}$ for $\epsilon = (|S|^2|A|/K)^{1/3}$ gives the following regret bound with probability at least $1-\delta$:*

$$R_K = O\left(T_\star^3|S|^{2/3}|A|^{1/3}K^{2/3}\widetilde{L} + T_\star^3|S|^2|A|\widetilde{L}^2\right).$$

Corollary 2.6. *Running Algorithm 2 using costs $c_\epsilon(s,a) = \max\{c(s,a),\epsilon\}$ for $\epsilon = |S|^2|A|/K$ gives the following regret bound with probability at least $1-\delta$:*

$$R_K = O\left(B_\star^{3/2}|S|\sqrt{|A|K\widetilde{L}} + T_\star^{3/2}|S|^2|A|\widetilde{L}^2\right).$$

Moreover, when the algorithm knows $B_\star$ and $K \gg |S|^2|A|T_\star^2$, then choosing $\epsilon = B_\star|S|^2|A|/K$ gives a near-optimal regret bound of $\widetilde{O}(B_\star|S|\sqrt{|A|K})$.

Lower bound. In Section 5 we show that Corollary 2.6 is nearly-tight using the following theorem.

Theorem 2.7. *There exists an SSP problem instance $M = (S, A, P, c, s_{init})$ in which $J^{\pi^\star}(s) \le B_\star$ for all $s \in S$, $|S| \ge 2$, $|A| \ge 16$, $B_\star \ge 2$, $K \ge |S||A|$, and $c(s,a) = 1$ for all $s \in S, a \in A$, such the expected regret of any learner after K episodes satisfies*

$$\mathbb{E}[R_K] \ge \frac{1}{1024}B_\star\sqrt{|S||A|K}.$$

3. Hoeffding-type Confidence Bounds

We start with a simpler case in which $B_\star$ is known to the learner. In Section 3.2 we alleviate this assumption with a penalty of an additional log-factor in the regret bound. For now, we prove the following bound on the learner's regret.

Theorem 3.1. *Suppose that Assumption 2 holds. With probability at least $1-\delta$ the regret of Algorithm 1 is bounded as follows:*

$$R_K = O\left(\sqrt{\frac{B_\star^3|S|^2|A|K}{c_{min}}}L + \frac{B_\star^3|S|^2|A|}{c_{min}^2}L^{3/2}\right).$$

Our algorithm follows the known concept of optimism in face of uncertainty. That is, it maintains confidence sets that contain the true transition function with high probability and picks an optimistic optimal policy—a policy that minimizes the expected cost over all policies and all transition functions in the current confidence set. The computation of the optimistic optimal policy can be done efficiently as shown by Tarbouriech et al. (2020). Construct an augmented MDP whose states are S and its action set consists of tuples $(a, \widetilde{P})$ where $a \in A$ and $\widetilde{P}$ is any transition function such that

$$\|\widetilde{P}(\cdot|s,a) - \bar{P}(\cdot|s,a)\|_1 \le 5\sqrt{\frac{|S|\log(|S||A|N_+(s,a)/\delta)}{N_+(s,a)}} \quad (2)$$

where $\bar{P}$ is the empirical estimate of P. It can be shown that the optimistic policy and the optimistic model, i.e., those that minimize the expected total cost over all policies and feasible transition functions, correspond to the optimal policy of the augmented MDP.

To ensure that the algorithm reaches the goal state in every episode, we define a state-action pair (s,a) as *known* if the number of visits to this pair is at least $\frac{5000B_\star^2|S|}{c_{\min}^2}\log\frac{B_\star|S||A|}{\delta c_{\min}}$ and as *unknown* otherwise. We show with high probability

Algorithm 1 HOEFFDING-TYPE CONFIDENCE BOUNDS AND KNOWN $B_\star$

input: state space S, action space A, bound on cost-to-go of optimal policy $B_\star$, confidence parameter δ.
initialization: $\forall (s,a,s') \in S \times A \times S : N(s,a,s') \leftarrow 0, N(s,a) \leftarrow 0$, an arbitrary policy $\tilde{\pi}$, $t \leftarrow 1$.
for $k = 1, 2, \ldots$ **do**
  set $s_t \leftarrow s_{\text{init}}$.
  while $s_t \neq g$ **do**
    follow optimistic optimal policy: $a_t \leftarrow \tilde{\pi}(s_t)$.
    observe next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.
    update: $N(s_t, a_t, s_{t+1}) \leftarrow N(s_t, a_t, s_{t+1}) + 1$, $N(s_t, a_t) \leftarrow N(s_t, a_t) + 1$.
    if $N(s_{t+1}, \tilde{\pi}(s_{t+1})) \leq \frac{5000 B_\star^2 |S|}{c_{\min}^2} \log \frac{B_\star |S||A|}{\delta c_{\min}}$ or $s_{t+1} = g$ **then**
      # start new interval
      compute empirical transition function $\bar{P}$ as $\bar{P}(s'|s,a) = N(s,a,s')/N_+(s,a)$ where $N_+(s,a) = \max\{N(s,a), 1\}$.
      compute optimistic policy $\tilde{\pi}$ by minimizing expected cost over transition functions $\widetilde{P}$ that satisfy Eq. (2).
    end if
    set $t \leftarrow t + 1$.
  end while
end for

the optimistic policy chosen by the algorithm will be proper once all state-action pairs are known. However, when some pairs are still unknown, our chosen policies may be improper. This implies that the strategy of keeping the policy fixed throughout an episode, as done usually in episodic RL, will fail. Consequently, our algorithm changes policies at the start of every episode and also every time we reach an unknown state-action pair.

Formally, we split the time into *intervals*. The first interval begins at the first time step, and every interval ends by reaching the goal state or a state s such that $(s, \tilde{\pi}(s))$ is unknown (where $\tilde{\pi}$ is the current policy followed by the learner). Recall that once all state-action pairs are known, the optimistic policy will eventually reach the goal state. Therefore, recomputing the optimistic policy at the end of every interval ensures that the algorithm will eventually reach the goal state with high probability. Note that the total number of intervals is at most the number of visits to an unknown state-action pair plus the number of episodes.

Observation 3.2. The total number of intervals, M, is

$$O\left(K + \frac{B_\star^2 |S|^2 |A|}{c_{\min}^2} \log \frac{B_\star |S||A|}{\delta c_{\min}}\right).$$

3.1. Analysis

The proof of Theorem 3.1 begins by defining the "good event" in which our confidence sets contain the true transition function and the total cost in every interval is bounded. This in turn implies that all episodes end in finite time. We prove that the good event holds with high probability.

Then, independently, we give a high-probability bound on the regret of the algorithm when the good event holds. To do so, recall that at the beginning of every interval m, the learner computes an optimistic policy by minimizing over all policies and over all transition functions within the current confidence set. We denote the chosen policy by $\tilde{\pi}^m$ and let $\widetilde{P}_m$ be the minimizing transition function (i.e., the optimistic model). A key observation is that by the definition of our confidence sets, $\widetilde{P}_m$ is such that there is always some positive probability to transition to the goal state directly from any state-action. This implies that all policies are proper in the optimistic model and that the cost-to-go function of $\tilde{\pi}^m$ defined with respect to $\widetilde{P}_m$, and denoted by $\widetilde{J}^m$, is finite. By Lemma 2.1, the following Bellman optimality equations hold for all $s \in S$,

$$\widetilde{J}^m(s) = \min_{a \in A} c(s,a) + \sum_{s' \in S} \widetilde{P}_m(s' \mid s,a) \widetilde{J}^m(s'). \tag{3}$$

High probability events. For every interval m, we let Ω^m denote the event that the confidence set for interval m contains the true transition function P. Formally, let $\bar{P}_m$ denote the empirical estimate of the transition function at the beginning of interval m, let $N_m(s,a)$ denote the number of visits to state-action pair (s,a) up to interval m (not including), and let $n_m(s,a)$ be the number of visits to (s,a) during interval m. Then we say that Ω^m holds if for all $(s,a) \in S \times A$, we have $(N_+^m(s,a) = \max\{1, N_m(s,a)\})$

$$\|P(\cdot|s,a) - \bar{P}_m(\cdot|s,a)\|_1 \leq 5 \sqrt{\frac{|S| \log\left(|S||A| N_+^m(s,a)/\delta\right)}{N_+^m(s,a)}}. \tag{4}$$

In the following lemma we show that, with high probability, the events Ω^m hold and that the total cost in each interval is bounded. Combining this with Observation 3.2 we get that all episodes terminate within a finite number of steps, with high probability.

Lemma 3.3. *With probability at least $1 - \delta/2$, for all intervals m simultaneously, we have that Ω^m holds and that $\sum_{h=1}^{H^m} c(s_h^m, a_h^m) \leq 24 B_\star \log \frac{4m}{\delta}$, where H^m denotes the length of interval m, s_h^m is the observed state at time h of interval m and $a_h^m = \tilde{\pi}^m(s_h^m)$ is the chosen action. This implies that the total number of steps of the algorithm is*

$$T = O\left(\frac{K B_\star}{c_{min}} L + \frac{B_\star^3 |S|^2 |A|}{c_{min}^3} L^2\right).$$

Proof sketch. The events Ω^m hold with high probability due to standard concentration inequalities, and thus it remains to address the high probability bound on the total cost within each interval.

This proof consists of three parts. In the first, we show that when Ω^m occurs we have that $\widetilde{J}^m(s) \le J^{\pi^\star}(s) \le B_\star$ for all $s \in S$ due to the optimistic nature of the computation of $\tilde{\pi}^m$. In the second part, we postulate that had all state-action pairs been known, then having Ω^m hold implies that $J^m(s) \le 2B_\star$ for all $s \in S$. That is, when all state-action pairs are known, not only $\tilde{\pi}^m$ is proper in the true model, but its expected cumulative cost is at most $2B_\star$.

The third part of the proof deals with the general case when not all state-action pairs are known. Fix some interval m. Since the interval ends when we reach an unknown state-action, it must be that all but the first state-action pair visited during the interval are known. For this unknown first state-action pair, it follows from the Bellman equations (Eq. (3)) and from $\widetilde{J}^m(s) \le B_\star$ for all $s \in S$ that $\tilde{\pi}^m$ never picks an action whose instantaneous cost is larger than $B_\star$. Therefore, the cost of this first unknown state-action pair is at most $B_\star$, and we focus on bounding the total cost in the remaining time steps with high probability.

To that end, we define the following modified MDP $M^{\text{know}} = (S^{\text{know}}, A, P^{\text{know}}, c, s_{\text{init}})$ in which every state $s \in S$ such that $(s, \tilde{\pi}^m(s))$ is unknown is contracted to the goal state. Let P^{know} be the transition function induced in M^{know} by P, and let J^m_{know} be the cost-to-go of $\tilde{\pi}^m$ in M^{know} w.r.t P^{know}. Similarly, define $\widetilde{P}^{\text{know}}_m$ as the transition function induced in M^{know} by $\widetilde{P}_m$, and $\widetilde{J}^m_{\text{know}}$ as the cost-to-go of $\tilde{\pi}^m$ in M^{know} w.r.t $\widetilde{P}^{\text{know}}_m$. It is clear that $\widetilde{J}^m_{\text{know}}(s) \le \widetilde{J}^m(s)$ for every $s \in S$ from whence $\widetilde{J}^m_{\text{know}}(s) \le B_\star$. Moreover, since all states $s \in S$ for which $(s, \tilde{\pi}^m(s))$ is unknown were contracted to the goal state, in M^{know} all remaining states-action pairs are known. Therefore, by the second part of the proof, $J^m_{\text{know}}(s) \le 2B_\star$ for all $s \in S$. Note that reaching the goal state in M^{know} is equivalent to reaching either the goal state or an unknown state-action pair in the true model hence the latter argument shows that the total expected cost in doing so is at most $2B_\star$. We further obtain the high probability bound by a probabilistic amplification argument using the Markov property of the MDP. □

Regret analysis. In what follows, instead of bounding R_K, we bound $\widetilde{R}_K = \sum_{m=1}^{M} \sum_{h=1}^{H^m} c(s_h^m, a_h^m) \mathbb{I}\{\Omega^m\} - K \cdot J^{\pi^\star}(s_{\text{init}})$, where $\mathbb{I}$ is the indicator function. Note that according to Lemma 3.3, we have that $\widetilde{R}_K = R_K$ with high probability.

The definition of $\widetilde{R}_K$ allows the analysis to disentangle two dependent probabilistic events. The first is the intersection of the events Ω^m which is dealt with in Lemma 3.3. The second holds when, for a fixed policy, the costs suffered by the learner do not deviate significantly from their expectation. In the following lemma we bound $\widetilde{R}_K$.

Lemma 3.4. *With probability at least* $1 - \delta/2$*, we have*

$$\widetilde{R}_K \le O\Bigg(\underbrace{\frac{B_\star^3 |S|^2 |A|}{c_{min}^2} \log \frac{B_\star |S||A|}{c_{min}\delta}}_{(1)} + B_\star \sqrt{T \log \frac{T}{\delta}} + B_\star \sqrt{|S| \log \frac{|S||A|T}{\delta}} \underbrace{\sum_{s,a} \sum_{m=1}^{M} \frac{n_m(s,a)}{\sqrt{N_+^m(s,a)}}}_{(2)} \Bigg).$$

Here we only explain how to interpret the resulting bound. The term (1) bounds the total cost spent in intervals that ended in unknown state-action pairs (it does not depend on K). The term (2) is at most $O(\sqrt{|S||A|T})$ when Lemma 3.3 holds, and then the dominant term in Lemma 3.4 becomes $\widetilde{O}(B|S|\sqrt{|A|T})$. Theorem 3.1 is finally obtained by applying a union bound on Lemmas 3.3 and 3.4 and using Lemma 3.3 to bound T.

3.2. Unknown Cost Bound

In this section we relax the assumption that $B_\star$ is known to the learner. Instead, we keep an estimate $\widetilde{B}$ that is initialized to $c_{\min}$ and doubles every time the cost in interval m (denoted as C_m) reaches $24\widetilde{B} \log \frac{4m}{\delta}$. By Lemma 3.3, with high probability, $\widetilde{B} \le 2B_\star$. We end an interval as before (once the goal state is reached or an unknown state-action pair is reached), but also when $\widetilde{B}$ is doubled. The algorithm for this case is presented in the supplementary material (Algorithm 3). Since $\widetilde{B}$ changes, every state-action pair can become known once for every different value of $\widetilde{B}$.

Observation 3.5. When $B_\star$ is unknown to the learner, the number of times a state-action pair can become known is at most $\log_2(B_\star/c_{\min})$. The number of intervals M is

$$O\left(K + \frac{B_\star^2 |S|^2 |A|}{c_{\min}^2} \log^2 \frac{B_\star |S||A|}{\delta c_{\min}} \right).$$

Lemma 3.6. *When $B_\star$ is unknown, with probability at least $1 - \delta/2$, for all intervals m simultaneously, we have that Ω^m holds and that $\sum_{h=1}^{H^m} c(s_h^m, a_h^m) \le 24 B_\star \log \frac{4m}{\delta}$. This implies that the total number of steps of the algorithm is*

$$T = O\left(\frac{KB_\star}{c_{min}} L + \frac{B_\star^3 |S|^2 |A|}{c_{min}^3} L^3 \right).$$

The analysis follows that of Algorithm 1. In particular, Lemma 3.4 still holds (with $2B_\star$ instead of $B_\star$), and jointly with Lemma 3.6 imply Theorem 2.3.

4. Bernstein-type Confidence Bounds

Algorithm 1 has two drawbacks. The first one is the use of Hoeffding-style confidence bounds which we improve with Bernstein-style confidence bounds. The second is the number of times the optimistic optimal policy is computed. In this section we propose to compute it in a way similar to UCRL2, i.e., once the number of visits to some state-action pair is doubled. Note that this change also eliminates the need to know or to estimate $B_\star$.

The algorithm is presented in Algorithm 2. It consists of *epochs*. The first epoch starts at the first time step, and each epoch ends once the number of visits to some state-action pair is doubled. An optimistic policy is computed at the end of every epoch using (empirical) Bernstein confidence bounds. In contrast to Algorithm 1, Algorithm 2 defines a confidence range for each state, action, and next state, separately, around its empirical estimate (i.e., we use an L_∞ "ball" rather than an L_1 "ball" around the empirical estimates). This allows us to disentangle the computation of the optimistic policy from the computation of the optimistic model. Indeed, the computation of the optimistic model becomes very easy: one simply has to maximize the probability of transition directly to the goal state at every state-action pair which means minimizing the probability of transition to all other states and setting them at the lowest possible value of their confidence range. This results in the following formula for $\widetilde{P}(s' \mid s,a)$ for every $(s,a,s') \in S \times A \times S$:

$$\max\{\bar{P}(s'|s,a) - 28A(s,a) - 4\sqrt{\bar{P}(s'|s,a)A(s,a)}, 0\}, \quad (5)$$

where $A(s,a) = \log(|S||A|N_+(s,a)/\delta)/N_+(s,a)$ and the remaining probability mass goes to $\widetilde{P}(g \mid s,a)$.[2] The optimistic policy is then the optimal policy in the SSP model defined by the transition function $\widetilde{P}$.

4.1. Analysis

In this section we prove Theorem 2.4. We start by showing that our new confidence sets contain P with high probability which implies that each episode ends in finite time with high probability. Consequently, we are able to bound the regret through summation of our confidence bounds.

We once again distinguish between *known* and *unknown* state-action pairs similarly to Algorithm 1. A state-action pair (s,a) becomes *known* at the end of an epoch if the total number of visits to (s,a) has passed $\alpha \cdot \frac{B_\star|S|}{c_{\min}} \log \frac{B_\star|S||A|}{\delta c_{\min}}$ at some time step during the epoch (for some constant $\alpha > 0$). Note that at the end of the epoch, the visit count of (s,a) may be strictly larger than $\alpha \cdot \frac{B_\star|S|}{c_{\min}} \log \frac{B_\star|S||A|}{\delta c_{\min}}$ but at most twice as much by the definition of our algorithm. Furthermore,

[2] The technical issue with this method is that the difference between $\widetilde{P}(g \mid s,a)$ and $\bar{P}(g \mid s,a)$ might be too large.

Algorithm 2 BERNSTEIN-TYPE CONFIDENCE BOUNDS

input: state space S, action space A and confidence parameter δ.
initialization: $i \leftarrow 1$, $t \leftarrow 1$, arbitrary policy $\tilde{\pi}_1$, $\forall (s,a,s') \in S \times A \times S: \; N_1(s,a,s') \leftarrow 0, N_1(s,a) \leftarrow 0$, $n_1(s,a,s') \leftarrow 0$, $n_1(s,a) \leftarrow 0$.
for $k = 1, 2, \ldots$ **do**
  set $s_t \leftarrow s_{\text{init}}$.
  while $s_t \neq g$ **do**
    follow optimistic optimal policy: $a_t \leftarrow \tilde{\pi}_i(s_t)$.
    observe next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.
    set: $n_i(s_t,a_t) \leftarrow n_i(s_t,a_t) + 1$, $n_i(s_t,a_t,s_{t+1}) \leftarrow n_i(s_t,a_t,s_{t+1}) + 1$.
    if $n_i(s_{t+1}, \tilde{\pi}_i(s_{t+1})) < N_i(s_{t+1}, \tilde{\pi}_i(s_{t+1}))$ **then**
      set $t \leftarrow t+1$ and **continue**.
    end if
    # start new epoch
    set: $N_{i+1}(s,a,s') \leftarrow N_i(s,a,s') + n_i(s,a,s')$, $N_{i+1}(s,a) \leftarrow N_i(s,a) + n_i(s,a)$, $n_{i+1}(s,a) \leftarrow 0$, $n_{i+1}(s,a,s') \leftarrow 0$ for all $(s,a,s') \in S \times A \times S$.
    compute empirical transition function $\bar{P}$ as $\bar{P}(s' \mid s,a) = N(s,a,s')/N_+(s,a)$ for every $(s,a,s') \in S \times A \times S$ where $N_+(s,a) = \max\{N(s,a),1\}$.
    compute optimistic transition function $\widetilde{P}$ using Eq. (5).
    compute optimal policy $\tilde{\pi}$ w.r.t $\widetilde{P}$.
    $i \leftarrow i+1$, $t \leftarrow t+1$.
  end while
end for

we split each epoch into *intervals* similar to what did in Section 3. The first interval starts at the first time step and each interval ends once (1) the total cost in the interval accumulates to at least $B_\star$; (2) an unknown state-action pair is reached; (3) the current episode ends; or (4) the current epoch ends. We have the following observation.

Observation 4.1. Let C_M denote the cost of the learner after M intervals. Observe that the total cost in each interval is at least $B_\star$ unless the interval ends in the goal state, in an unknown state-action pair or the epoch ends. Thus the total number of intervals satisfies

$$M \leq \frac{C_M}{B_\star} + 2|S||A| \log T + K + O\left(\frac{B_\star|S|^2|A|}{c_{\min}} \log \frac{B_\star|S||A|}{\delta c_{\min}}\right),$$

and the total time satisfies $T \leq C_M/c_{\min}$.

Recall that in the analysis of Algorithm 1 we show that once all state-action pairs are known, the optimistic policies generated by the algorithm are proper in the true MDP. The same holds true for Algorithm 2, yet we never prove this directly. Instead, our proof goes as follows.[3] We prove

[3] We neglect low order terms here.

that C_M, the cost accumulated by the learner during the first M intervals, is at most $K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{M}$ with high probability as long as no more than K episodes have been completed during these M intervals. We notice that once all state-action pairs are known, the total cost in each interval is at least $B_\star$ (ignoring intervals that end with the end of an epoch or an episode), which implies that the total number of intervals M is bounded by $C_M/B_\star$. This allows us to get a bound on C_M that is independent of the number of intervals by solving the inequality $C_M \lesssim K \cdot J^{\pi^\star}(s_{\text{init}}) + B_\star\sqrt{M} \lesssim K \cdot J^{\pi^\star}(s_{\text{init}}) + \sqrt{B_\star \cdot C_M}$. From this, and since the instantaneous costs are strictly positive (by Assumption 2), it must be that the learner eventually completes all K episodes; i.e., there must be a time from which Algorithm 2 generates only proper policies.

Notation. The epoch that interval m belongs to is denoted by $i(m)$, other notations are as in Section 3.1. Note that since the optimistic policy is computed at the end of an epoch and not at the end of an interval, it follows that $\tilde{\pi}^m = \tilde{\pi}^{i(m)}$ and $\widetilde{J}^m = \widetilde{J}^{i(m)}$. The trajectory visited in interval m is denoted by $U^m = (s_1^m, a_1^m, \ldots, s_{H^m}^m, a_{H^m}^m, s_{H^m+1}^m)$, where a_h^m is the action taken in s_h^m, and H^m is the length of the interval. In addition, the concatenation of the trajectories of the intervals up to and including interval m is denoted by $\bar{U}^m$, that is $\bar{U}^m = \cup_{m'=1}^m U^{m'}$.

High probability events. Throughout the analysis we denote $S^+ = S \cup \{g\}$. For every interval m we let Ω^m denote the event that the confidence set for epoch $i = i(m)$ contains the actual transition function P. Formally, if Ω^m holds then for all $(s,a,s') \in S \times A \times S^+$, we have (denote $N_+^m(s,a) = \max\{1, N^m(s,a)\}$, $A_h^m = A(s_h^m, a_h^m)$)

$$|P(s'|s,a) - \bar{P}_m(s'|s,a)| \le 28A_h^m + 4\sqrt{\bar{P}_m(s'|s,a)A_h^m}. \quad (6)$$

In the following lemma we show that the events Ω^m hold with high probability.

Lemma 4.2. *With probability at least $1 - \delta/2$, Ω^m holds for all intervals m simultaneously.*

Regret analysis. In the following section, instead of bounding R_K, we bound $\widetilde{R}_M = \sum_{m=1}^M \sum_{h=1}^{H^m} c(s_h^m, a_h^m)\mathbb{I}\{\Omega^m\} - KJ^{\pi^\star}(s_{\text{init}})$ for any number of intervals M. This implies Theorem 2.4 by the following argument. Lemma 4.2 implies that $\widetilde{R}_M = R_M$ with high probability for any number of intervals M (R_M is the true regret within the first M intervals). In particular, when M is the number of intervals in which the first K episodes elapse, this implies Theorem 2.4 (we show that the learner indeed completes these K episodes).

To bound $\widetilde{R}_M$, we use the next lemma to decompose $\widetilde{R}_M$ into two terms which we bound independently.

Lemma 4.3. *It holds that $\widetilde{R}_M = \sum_{m=1}^M \widetilde{R}_m^1 + \sum_{m=1}^M \widetilde{R}_m^2 - K \cdot J^{\pi^\star}(s_{init})$, where*

$$\widetilde{R}_m^1 = \left(\widetilde{J}^m(s_1^m) - \widetilde{J}^m(s_{H^m+1}^m)\right)\mathbb{I}\{\Omega^m\}, \quad \text{and}$$

$$\widetilde{R}_m^2 = \left(\sum_{h=1}^{H^m} \widetilde{J}^m(s_{h+1}^m) - \sum_{s'\in S} \widetilde{P}_m(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')\right)\mathbb{I}\{\Omega^m\}.$$

The lemma breaks down $\widetilde{R}_M$ into two terms. The first term accounts for the number of times in which the learner changes her policy in the middle of an episode which is at most the number of epochs. The second term sums the errors between the cost-to-go of the observed next state and its estimated expectation.

Indeed, $\sum_{m=1}^M \widetilde{R}_m^1$ is related to the total number of epochs which is at most $|S||A|\log_2 T$ due to the following lemma.

Lemma 4.4. *It holds that $\sum_{m=1}^M \widetilde{R}_m^1 \le 2B_\star|S||A|\log T + KJ^{\pi^\star}(s_{init})$.*

The next lemma shows that $\sum_{m=1}^M \widetilde{R}_m^2$ does not deviate from $\sum_{m=1}^M \mathbb{E}[\widetilde{R}_m^2 \mid \bar{U}^{m-1}]$ significantly.

Lemma 4.5. *With probability at least $1 - \delta/4$,*

$$\sum_{m=1}^M \widetilde{R}_m^2 \le \sum_{m=1}^M \mathbb{E}\left[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\right] + 3B_\star\sqrt{M\log\frac{8M}{\delta}}.$$

The key property of the lemma is that the deviations between $\sum_{m=1}^M \widetilde{R}_m^2$ and its corresponding expectation is of order $\sqrt{M}$ and do not scale with T.

To prove the lemma, we recall that an interval ends at most at the first time step in which the accumulated cost in the interval surpasses $B_\star$. We show in our analysis that $\widetilde{J}^m(s) \le J^{\pi^\star}(s) \le B_\star$ for all $s \in S$ due to the optimistic computation of $\tilde{\pi}^m$. Therefore, $\tilde{\pi}^m$ never picks an action whose instantaneous cost is more than $B_\star$. This implies that the total cost within each interval is at most $2B_\star$. Then, we use the Bellman equations to bound $\widetilde{R}_m^2$ by order of the total cost in the interval, and the lemma follows by an application of Azuma's concentration inequality.

Lemma 4.6 below bounds $\mathbb{E}\left[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\right]$ for every interval m by a sum of the confidence bounds used in Algorithm 2.

Lemma 4.6. *For every interval m,*

$$\mathbb{E}\left[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\right] \le 16\mathbb{E}\left[\sum_{h=1}^{H^m}\sqrt{|S|\mathbb{V}_h^m A_h^m}\mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1}\right] + 272\mathbb{E}\left[\sum_{h=1}^{H^m} B_\star|S|A_h^m\mathbb{I}\{\Omega^m\} \;\middle|\; \bar{U}^{m-1}\right], \quad (7)$$

where $\mathbb{V}_h^m$ is the empirical variance defined as $\mathbb{V}_h^m = \sum_{s'\in S^+} P(s' \mid s_h^m, a_h^m)\left(\widetilde{J}^m(s') - \mu_h^m\right)^2$, and $\mu_h^m = \sum_{s'\in S^+} P(s' \mid s_h^m, a_h^m)\widetilde{J}^m(s')$.

The next step is the part of our proof in which our analysis departs from that of Algorithm 1. Note that when Ω^m holds, $\mathbb{V}_h^m \leq B_\star^2$. Using this bound for each time step separately will result in a bound similar to that of Theorem 2.3. However, this bound is loose due to the following intuitive argument. Suppose that we replace $\widetilde{J}^m$ with the true cost-to-go function of $\tilde{\pi}^m$, J^m, in the definition of $\mathbb{V}_h^m$. Note that from the Bellman equations (Eq. (1)) we have $J^m(s_h^m) > J^m(s_{h+1}^m)$ in expectation on consecutive time steps h and $h+1$ hence we surmise that in expectation $\mathbb{V}_h^m$ would also decrease on consecutive time steps. A similar argument holds when in reality we use $\widetilde{J}^m$ because all-but-one of the state-action pairs in the interval are known, and $\widetilde{J}^m$ is a "close enough" approximation of J^m on known state-action pairs since they have been sampled sufficiently many times. Indeed, in Lemma 4.7 we use the technique of Azar et al. (2017) to show that (up to a constant) $B_\star^2$ bounds the expected sum of the variances over the time steps of an interval.

Lemma 4.7. $\mathbb{E}\left[\sum_{h=1}^{H^m} \mathbb{V}_h^m \mathbb{I}\{\Omega^m\} \mid \bar{U}^{m-1}\right] \leq 44 B_\star^2$.

Armed with Lemma 4.7, we upper bound $\sum_{m=1}^{M} \mathbb{E}\left[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\right]$ by applying some algebraic manipulation on Eq. (7), and summing over all intervals which gives the next lemma.

Lemma 4.8. *With probability at least* $1-\delta/4$,

$$\sum_{m=1}^{M} \mathbb{E}\left[\widetilde{R}_m^2 \mid \bar{U}^{m-1}\right] \leq 614 B_\star \sqrt{M|S|^2|A| \log^2 \frac{T|S||A|}{\delta}} + 8160 B_\star |S|^2 |A| \log^2 \frac{T|S||A|}{\delta}.$$

Theorem 2.4 is obtained by first applying a union bound on Lemmas 4.2, 4.5 and 4.8, plugging in the bounds of Lemmas 4.4, 4.5 and 4.8 into Lemma 4.3, and bounding T and M using Observation 4.1. This results in a quadratic inequality in $\sqrt{C_M}$ and solving it yields the theorem.

5. Lower Bound

In this section we give an overview of the proof for Theorem 2.7. For clarity we restate the theorem.

Theorem (restatement of Theorem 2.7)**.** *There exists an SSP problem instance* $M = (S, A, P, c, s_{init})$ *in which* $J^{\pi^\star}(s) \leq B_\star$ *for all* $s \in S$, $|S| \geq 2$, $|A| \geq 16$, $B_\star \geq 2$, $K \geq |S||A|$, *and* $c(s,a) = 1$ *for all* $s \in S, a \in A$, *such the expected regret of any learner after* K *episodes satisfies*

$$\mathbb{E}[R_K] \geq \frac{1}{1024} B_\star \sqrt{|S||A|K}.$$

The proof of our lower bound takes similar steps to the one in Jaksch et al. (2010). Note that one cannot simply use a reduction to the average-cost setting in our case because the number of steps taken by the algorithm is potentially unbounded, and not the same as the number of steps taken by the optimal policy.

Still, our lower bound matches the one for finite-horizon MDPs of $\Omega(\sqrt{H|S||A|T})$, where H is the horizon and T is the total number of time steps. Since the length of each episode is H, we have that $T = HK$ and the lower bound takes the form of $\Omega(H\sqrt{|S||A|K})$. In our case, $B_\star$ replaces the horizon H as an upper bound on the expected cost of the optimal policy, and we get the same linear dependence in this parameter.

Before constructing an instance for which we can prove the general lower bound, we consider a simpler instance that consists of only the initial state s_{init} and the goal state g. The actions are all the same, except for one optimal action $a^\star$ which is chosen uniformly at random. While all actions (including $a^\star$) suffer a cost of 1, $a^\star$ has a better probability of transitioning to the goal state, that is, $P(g \mid s_{\text{init}}, a^\star) = 1/B_\star$ compared to $P(g \mid s_{\text{init}}, a) = (1-\epsilon)/B_\star$ for all other actions $a \neq a^\star$.

Notice that the optimal policy $\pi^\star$ chooses $a^\star$ and has an expected cost of exactly $B_\star$. Therefore, the job of the learner is simply to identify $a^\star$. In the supplementary material we show that the regret of the learner in this case must be $\Omega(B_\star\sqrt{|A|K})$.

Subsequently, we build our general hard instance by taking $|S|$ copies of the aforementioned simple MDP and picking the initial state in every episode uniformly at random. Since the copies are not connected in any way, the lower bound applies to each of them separately. Notice that every state will be visited $K/|S|$ times in expectation, so the expected regret will be

$$\Omega\left(\sum_{s\in S} B_\star \sqrt{|A|\frac{K}{|S|}}\right) = \Omega(B_\star\sqrt{|S||A|K}).$$

Interestingly enough, although playing proper policies was a major concern in the construction of our algorithms, the hard instance we have built does not have any improper polices at all.

Acknowledgements

Haim Kaplan is supported in part by the Israeli Science Foundation (ISF) grant 1595/19. Yishay Mansour is supported in part by an Israeli Science Foundation (ISF) grant.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

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.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 35–42. AUAI Press, 2009.

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

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

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 12203–12213, 2019.

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, T. and Luo, H. Learning adversarial mdps with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pp. 231–243, 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.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386, 2016.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pp. 2209–2218, 2019a.

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

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

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.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pp. 2823–2832, 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 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 1583–1591, 2013.