Stochastic Shortest Path with Adversarially Changing Costs

Aviv Rosenberg,[1] **Yishay Mansour**[1,2]
[1] Tel Aviv University, Israel
[2] Google Research, Israel
avivros007@gmail.com, mansour.yishay@gmail.com

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 this paper we consider adversarial SSPs that also account for adversarial changes in the costs over time, while the dynamics (i.e., transition function) remains unchanged. Formally, an agent interacts with an SSP environment for K episodes, the cost function changes arbitrarily between episodes, and the fixed dynamics are unknown to the agent. We give high probability regret bounds of $\widetilde{O}(\sqrt{K})$ assuming all costs are strictly positive, and $\widetilde{O}(K^{3/4})$ for the general case. To the best of our knowledge, we are the first to consider this natural setting of adversarial SSP and obtain sub-linear regret for it.

1 Introduction

Stochastic shortest path (SSP) is one of the most basic models in reinforcement learning (RL). In SSP the goal of the agent is to reach a predefined goal state in minimum expected cost, and it captures a wide variety of natural scenarios, such as car navigation and game playing.

An important aspect that the SSP model fails to capture is the changes in the environment over time (e.g., changes in traffic when navigating a car). Usually, this aspect of the environment is theoretically modeled by adversarial Markov decision processes (MDPs), in which the cost function may change arbitrarily over time, while still assuming a fixed transition function. In this work we present the adversarial SSP model that combines SSPs with adversarial MDPs.

In the adversarial SSP model, the agent interacts with an SSP environment for K episodes, but the cost function changes between episodes arbitrarily. The agent's objective is to reach the goal state in every episode while minimizing its total expected cost. Its performance is measured by the regret – the difference between the agent's total cost and the expected total cost of the best stationary policy in hindsight.

We propose the first algorithms for regret minimization in adversarial SSPs. Our algorithms take recent advances in learning SSP problems (Tarbouriech et al. 2019; Cohen et al. 2020) – that build upon the optimism in face of uncertainty principle, and combine them with the O-REPS framework (Zimin and Neu 2013; Rosenberg and Mansour 2019a,b; Jin and Luo 2019) for adversarial episodic MDPs – which implements the online mirror descent (OMD) algorithm for online convex optimization (OCO). We follow the strategy of Tarbouriech et al. (2019); Cohen et al. (2020) for SSPs – We start by assuming all costs are strictly positive and prove $\widetilde{O}(\sqrt{K})$ regret (which is optimal). Then, using a perturbation argument, we remove this assumption and show that our algorithms obtain $\widetilde{O}(K^{3/4})$ regret in the general case.

Our main technical contribution is the adaptation of OMD to the SSP environment. While a naive application of this method results in unbounded regret even in expectation, we show that small modifications obtain near-optimal regret bounds with high probability. Our second technical contribution is the combination of OMD with optimism, which is far more challenging than in the episodic setting (Rosenberg and Mansour 2019a), and requires clever modifications to the analysis. We hope that the framework created in this paper for handling adversarial costs in SSP environments will pave the way for future work to achieve tight regret bounds with more practical algorithms.

The rest of the paper is organized as follows. First, we consider a simplified case in which the transition function is known to the learner and the regret should be minimized in expectation. For this case, we establish an efficient O-REPS based algorithm and bound its expected regret. Then, we introduce an improvement that ensures the learner will not run too long before reaching the goal, and show that this yields a high probability regret bound. Next, we remove the known transition function assumption and combine our algorithm with the confidence set framework of UCRL2 (Jaksch, Ortner, and Auer 2010). This allows us to prove high probability regret bound without knowledge of the transitions. Finally, we remove the assumption of strictly positive costs and obtain high probability regret bounds for the general case.

1.1 Related work

Early work by Bertsekas and 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.

Recently the problem of learning SSPs was addressed by Tarbouriech et al. (2019) and then improved by Cohen et al. (2020). The latter show an efficient algorithm based on optimism and prove that it obtains a high probability regret bound of $\widetilde{O}(D|S|\sqrt{|A|K})$, where D is the diameter, S is the state space and A is the action space. They also prove a nearly matching lower bound of $\Omega(D\sqrt{|S||A|K})$.

Regret minimization in RL has been extensively studied, but the literature mainly focuses on the average-cost infinite-horizon model (Bartlett and Tewari 2009; Jaksch, Ortner, and Auer 2010; Fruit et al. 2018) and on the finite-horizon (episodic) model (Osband, Van Roy, and Wen 2016; Azar, Osband, and Munos 2017; Dann, Lattimore, and Brunskill 2017; Jin et al. 2018; Zanette and Brunskill 2019; Efroni et al. 2019). The extension of these ideas to SSPs is an important task due to the practical applications of this setting, and therefore we focus on extending the adversarial MDP literature to SSP.

Adversarial MDPs were first studied in the average-cost infinite-horizon model (Even-Dar, Kakade, and Mansour 2009; Yu, Mannor, and Shimkin 2009; Neu et al. 2014), before focusing on the episodic setting. Early work in the episodic setting by Neu, György, and Szepesvári (2010) used a reduction to multi-arm bandit (Auer et al. 2002), but then Zimin and Neu (2013) introduced the O-REPS framework. All these works assumed a known transition function, but more recent work (Neu, György, and Szepesvári 2012; Rosenberg and Mansour 2019a,b; Jin and Luo 2019) considered the more general case where the agent must learn the transition function from experience. Recently, Efroni et al. (2020); Cai et al. (2019) showed that policy optimization methods (that are widely used in practice) can also achieve near-optimal regret bounds in adversarial episodic MDPs.

It is important to point out that we are the first to consider SSP with adversarially changing costs, although some previous works on finite-horizon adversarial MDPs refer to it as "adversarial (loop-free) stochastic shortest path".

2 Preliminaries

An adversarial SSP problem is defined by an MDP $M = (S, A, P, s_0, g)$ and a sequence $c_1, \ldots, c_K$ of cost functions, where $c_k : S \times A \to [0, 1]$. We do not make any statistical assumption on the cost functions, i.e., they can be chosen arbitrarily. S and A are finite state and action spaces, respectively, and P is a transition function such that $P(s' \mid s, a)$ is the probability to move to s' when taking action a in state s.

The learner interacts with M in episodes, where c_k is the cost function for episode k. However, it is revealed to the learner only in the end of the episode. In every episode, the learner begins at the initial state s_0[1], and ends the interaction with M by arriving at the goal state g (where $g \notin S$). The full interaction is described in Algorithm 1. To simplify the presentation we denote $S^+ = S \cup \{g\}$ and thus for every $(s, a) \in S \times A$ we have that $\sum_{s' \in S^+} P(s' \mid s, a) = 1$.

[1]For simplicity s_0 is fixed, but all the results generalize to any initial distribution.

Algorithm 1 Learner-Environment Interaction

Parameters: $M = (S, A, P, s_0, g), \{c_k\}_{k=1}^K$
for $k = 1$ **to** K **do**
  learner starts in state $s_1^k = s_0$, $i \leftarrow 1$
  while $s_i^k \neq g$ **do**
    learner chooses action $a_i^k \in A$
    learner observes state $s_{i+1}^k \sim P(\cdot \mid s_i^k, a_i^k)$, $i \leftarrow i+1$
  end while
  learner observes c_k and suffers cost $\sum_{j=1}^{i-1} c_k(s_j^k, a_j^k)$
end for

2.1 Proper policies

A stationary (stochastic) policy is a mapping $\pi : S \times A \to [0, 1]$, where $\pi(a \mid s)$ gives the probability that action a is selected in state s. Since reaching the goal is one of the main objectives of the learner along with minimizing its cost, we are interested in *proper* policies (otherwise we cannot guarantee finite cost, let alone finite regret).

Definition 1. A policy π is *proper* if playing according to π reaches the goal state with probability 1 when starting from any state. A policy is *improper* if it is not proper.

The set of all proper deterministic policies is denoted by Π_{proper}. In addition, we 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 $s \in S$, and if π is improper there must exist some $s' \in S$ such that $T^\pi(s') = \infty$. We make the basic assumption that the goal is reachable from every state (also implies that $\Pi_{\text{proper}} \neq \emptyset$), and formalize it as follows.

Assumption 1. There exists at least one proper policy.

When paired with a cost function $c : S \times A \to [0, 1]$, any policy π induces a *cost-to-go function* $J^\pi : S \to [0, \infty]$, where $J^\pi(s)$ is the expected cost when playing policy π and starting at state s, i.e., $J^\pi(s) = \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=1}^T c(s_t, a_t) \mid P, \pi, s_1 = s\right]$. For a proper policy π, it follows that $J^\pi(s)$ is finite for all $s \in S$. However, note that $J^\pi(s')$ may be finite for some state $s' \in S$ even if π is improper.

Under Assumption 1 and the assumption that every improper policy suffers infinite cost from some state, Bertsekas and Tsitsiklis (1991) show that the optimal policy is stationary, deterministic and proper; and that every proper policy π satisfies the following Bellman equations for every $s \in S$:

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

$$T^\pi(s) = 1 + \sum_{a \in A} \sum_{s' \in S} \pi(a \mid s) P(s' \mid s, a) T^\pi(s') \tag{2}$$

2.2 Learning formulation

The success of the learner is measured by the regret – the difference between the learner's total cost in K episodes and the total expected cost of the best *proper* policy in hindsight:

$$R_K = \sum_{k=1}^K \sum_{i=1}^{I^k} c_k(s_i^k, a_i^k) - \min_{\pi \in \Pi_{\text{proper}}} \sum_{k=1}^K J_k^\pi(s_0),$$

where I^k is the time it takes the learner to complete episode k (which may be infinite), (s_i^k, a_i^k) is the i-th state-action pair at episode k, and J_k^π is the cost-to-go of policy π with respect to (w.r.t) cost function c_k. In the case that I^k is infinite for some k, to ensure the goal must be reached, we define $R_K = \infty$.

We denote $\pi^\star = \arg\min_{\pi \in \Pi_{\text{proper}}} \sum_{k=1}^K J_k^\pi(s_0)$, and define the SSP-diameter (Tarbouriech et al. 2019), $D = \max_{s \in S} \min_{\pi \in \Pi_{\text{proper}}} T^\pi(s)$, which commonly appears in regret bounds but is unknown to the learning algorithms.

Our analysis makes use of the Bellman equations, that hold under the conditions described before Eq. (1). To make sure these are met, we make the assumption that the costs are strictly positive. In Section 6 we remove this assumption, by adding a small positive perturbation to the costs.

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

3 Occupancy measures

Every policy π induces an occupancy measure $q^\pi : S \times A \to [0, \infty]$ such that $q^\pi(s,a)$ is the expected number of times to visit state s and take action a when playing according to π,

$$q^\pi(s,a) = \lim_{T \to \infty} \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}\{s_t = s, a_t = a\} \mid P, \pi, s_1 = s_0\Big],$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Note that for a proper policy π, $q^\pi(s,a)$ is finite for all state-action pairs. Furthermore, this correspondence between proper policies and finite occupancy measures is 1-to-1, and its inverse for q is given by $\pi^q(a \mid s) = \frac{q(s,a)}{q(s)}$, where $q(s) \overset{\text{def}}{=} \sum_{a \in A} q(s,a)$. [2]

The aforementioned equivalence between policies and occupancy measures is well-known for MDPs (see, e.g., Zimin and Neu (2013)), but also holds for SSPs by linear programming formulation (Manne 1960; d'Epenoux 1963). Notice that the expected cost of policy π is linear w.r.t q^π,

$$\begin{aligned} J_k^{\pi_k}(s_0) &= \mathbb{E}\Big[\sum_{i=1}^{I^k} c_k(s_i^k, a_i^k) \mid P, \pi_k, s_1 = s_0\Big] \\ &= \sum_{s \in S} \sum_{a \in A} q^{\pi_k}(s,a) c_k(s,a) \overset{\text{def}}{=} \langle q^{\pi_k}, c_k \rangle. \end{aligned} \tag{3}$$

Thus, minimizing the expected regret can be written as an instance of online linear optimization (using tower property),

$$\begin{aligned} \mathbb{E}[R_K] &= \mathbb{E}\left[\sum_{k=1}^K J_k^{\pi_k}(s_0) - \sum_{k=1}^K J_k^{\pi^\star}(s_0)\right] \\ &= \mathbb{E}\left[\sum_{k=1}^K \langle q^{\pi_k} - q^{\pi^\star}, c_k \rangle\right]. \end{aligned}$$

[2] If $q(s) = 0$ for some state s then the inverse mapping is not well-defined. However, since s will not be reached, we can pick the action there arbitrarily. More precisely, the correspondence holds when restricting to reachable states.

4 Known transition function

Before tackling our main challenge of adversarial SSP with unknown transition function in Section 5, we use the simpler case of known dynamics to develop our main techniques. In Section 4.1 we establish the implementation of the OMD method in SSP, and then in Section 4.2 we show how to convert its expected regret bound into a high probability bound.

4.1 Online mirror descent for SSP

Online mirror descent is a popular framework for OCO and its application to occupancy measures yields the O-REPS algorithms (Zimin and Neu 2013; Rosenberg and Mansour 2019a,b; Jin and Luo 2019). Usually these algorithms operate w.r.t to the set of all occupancy measures (which corresponds to the set of all policies), but a naive application of this kind may result in the learner playing improper policies.

Thus, we propose to use the set $\Delta(D/c_{\min})$ – occupancy measures of policies π that reach the goal in expected time $T^\pi(s_0) \leq D/c_{\min}$, which has a compact representation as we show next. As mentioned before, while this is not a concern in finite-horizon RL, limiting the running time of our policies will be crucial in the regret analysis. The parameter $D/c_{\min}$ is chosen because it is the smallest such that $q^{\pi^\star} \in \Delta(D/c_{\min})$, i.e., $T^{\pi^\star}(s_0) \leq D/c_{\min}$ (see Appendix C[3]). Another approach may be to estimate $T^{\pi^\star}(s_0)$ on the fly, but this seems unlikely due to the adversarial change in the cost (that changes the best policy in hindsight).

Our algorithm, called SSP-O-REPS, follows the O-REPS framework. In each episode we pick an occupancy measure (and thus a policy) from $\Delta(D/c_{\min})$ which minimizes a trade-off between the current cost function and the distance to the previously chosen occupancy measure. Formally,

$$q_{k+1} = q^{\pi_{k+1}} = \operatorname*{arg\,min}_{q \in \Delta(D/c_{\min})} \eta \langle q, c_k \rangle + \mathrm{KL}(q \parallel q_k), \tag{4}$$

where $\mathrm{KL}(\cdot||\cdot)$ is the un-normalized KL-divergence defined by $\mathrm{KL}(q \parallel q') = \sum_{s,a} q(s,a) \log \frac{q(s,a)}{q'(s,a)} + q'(s,a) - q(s,a)$, and $\eta > 0$ is a learning rate. To compute q_{k+1} we first find the unconstrained minimizer and then project it into $\Delta(D/c_{\min})$ (see Zimin and Neu (2013)), i.e.,

$$q'_{k+1} = \operatorname*{arg\,min}_{q} \eta \langle q, c_k \rangle + \mathrm{KL}(q \parallel q_k) \tag{5}$$

$$q_{k+1} = \operatorname*{arg\,min}_{q \in \Delta(D/c_{\min})} \mathrm{KL}(q \parallel q'_{k+1}). \tag{6}$$

Eq. (5) has a closed form $q'_{k+1}(s,a) = q_k(s,a) e^{-\eta c_k(s,a)}$, and Eq. (6) can be formalized as a constrained convex optimization problem using the following constraints:

$$\forall s \in S \quad \sum_{a \in A} q(s,a) - \mathbb{I}\{s = s_0\} = \tag{7}$$

$$= \sum_{s' \in S} \sum_{a' \in A} q(s',a') P(s \mid s', a')$$

$$\sum_{s \in S} \sum_{a \in A} q(s,a) \leq \frac{D}{c_{\min}}, \tag{8}$$

[3] The appendix can be found in the full version of the paper on Arxiv: https://arxiv.org/abs/2006.11561.

where we omitted non-negativity constraints, constraint (7) is standard, and constraint (8) ensures $T^{\pi^q}(s_0) \leq D/c_{\min}$. In Appendix A, we show how to solve this problem efficiently and describe the implementation details of the algorithm. In addition, we describe the efficient computation of D by finding the optimal policy w.r.t the constant cost function $c(s,a) = 1$. High-level pseudocode is found in Algorithm 2, and fully detailed pseudocode in Appendix B.

We follow the analysis of OMD to obtain the algorithm's expected regret bound. Moreover, we show that all the policies chosen by the algorithm must be proper and therefore the goal state will be reached with probability 1 in all episodes. Proofs are in Appendix C.

Theorem 4.1. *Under Assumptions 1 and 2, the expected regret of SSP-O-REPS with known transition function and* $\eta = \sqrt{\frac{3\log(D|S||A|/c_{min})}{K}}$ *is*

$$\mathbb{E}[R_K] \leq O\left(\frac{D}{c_{min}}\sqrt{K\log\frac{D|S||A|}{c_{min}}}\right) = \widetilde{O}\left(\frac{D}{c_{min}}\sqrt{K}\right).$$

4.2 High probability regret bound

In contrast to the finite-horizon setting where the cost is always bounded by the horizon H, a regret bound in expectation in the SSP setting does not guarantee any concrete bound on the actual cost. Thus, it is of great importance to bound the regret with high probability, which requires us to bound the deviation of the suffered cost from its expectation.

The following lemma shows that this deviation is closely related to the expected time of reaching the goal from any state. Its proof is based on an adaptation of Azuma inequality to martingales that are bounded only with high probability (Theorem J.5), and might be of independent interest.

Lemma 4.2. *Denote by σ_k the learner's strategy in episode k, and assume that the expected time of reaching the goal from any state when playing σ_k is at most D/c_{min}. Then, with probability at least $1-\delta$,*

$$\sum_{k=1}^{K}\sum_{i=1}^{I^k} c_k(s_i^k,a_i^k) \leq \sum_{k=1}^{K}\mathbb{E}\Big[\sum_{i=1}^{I^k} c_k(s_i^k,a_i^k)|P,\sigma_k,s_1^k=s_0\Big] + O\left(\frac{D}{c_{min}}\sqrt{K\log^3\frac{K}{\delta}}\right).$$

Thus, we would like to bound $T^{\pi_k}(s) \leq D/c_{\min}$ for all $s \in S$ and not just s_0. However, representing this constraint with occupancy measures gives a non-convex set. We bypass this issue with the algorithm SSP-O-REPS2 that operates as follows. We start every episode k by playing the policy π_k chosen by SSP-O-REPS (i.e., Eq. (4)), but once we reach a state s whose expected time to the goal is too long (i.e., $T^{\pi_k}(s) \geq D/c_{\min}$), we switch to the fast policy π^f.

The fast policy minimizes the time to the goal from any state and can be computed efficiently similarly to the SSP-diameter D (see Appendix A). Computing T^{π_k} is also efficient. Notice that if $q^{\pi_k}(s) > 0$ then $T^{\pi_k}(s)$ must be finite, otherwise $T^{\pi_k}(s_0) = \infty$. Thus, we can compute T^{π_k} as follows: Ignore states that are not reachable from s_0 using π_k, and solve the (linear) Bellman equations (Eq. (2)). Due to lack of space, we defer to the pseudocode in Appendix D.

Algorithm 2 SSP-O-REPS

Input: $S, A, P, c_{\min}, \eta$.
Initialization:
  Compute D (see Appendix A.2).
  Set $q_0(s,a) = 1$ and $c_0(s,a) = 0$ for every (s,a).
for $k = 1, 2, \ldots$ **do**
  Perform OMD step (see Appendix A):

$$q_k = \operatorname*{arg\,min}_{q \in \Delta(D/c_{\min})} \eta\langle q, c_{k-1}\rangle + \mathrm{KL}(q \,\|\, q_{k-1}).$$

  Compute $\pi_k(a \mid s) = \frac{q_k(s,a)}{\sum_{a' \in A} q_k(s,a')}$ for every (s,a).
  Set $s_1^k \leftarrow s_0$, $i \leftarrow 1$.
  while $s_i^k \neq g$ **do**
    Play action according to π_k, i.e., $a_i^k \sim \pi_k(\cdot \mid s_i^k)$.
    Observe next state $s_{i+1}^k \sim P(\cdot \mid s_i^k, a_i^k)$, $i \leftarrow i+1$.
  end while
  Observe c_k and suffer cost $\sum_{j=1}^{i-1} c_k(s_j^k, a_j^k)$.
end for

We denote by σ_k the strategy of playing π_k until reaching a "bad" state and then switching to the fast policy. Now Lemma 4.2 bounds the deviation of our suffered cost from its expectation. Next, we again turn to bounding the expected regret. We cannot apply OMD analysis immediately since we did not play π_k all through the episode. However, Lemma 4.3 shows that our mid-episode policy switch only decreases the expected cost, and this leads to the high probability regret bound in Theorem 4.4 (proofs in Appendix E).

The technique of switching to the fast policy was already used by Tarbouriech et al. (2019) for non-adversarial SSP. Our key novelty is to make the switch without suffering extra cost, by switching only in "bad" states. While Tarbouriech et al. (2019) leverage the non-adversarial environment to guarantee that the number of switches is finite with high probability, here it is crucial to avoid extra cost since the switch may occur in every episode.

Lemma 4.3. *For every $k = 1, \ldots, K$ it holds that*

$$\mathbb{E}\Big[\sum_{i=1}^{I^k} c_k(s_i^k,a_i^k) \mid P,\sigma_k,s_1^k=s_0\Big] \leq \leq \mathbb{E}\Big[\sum_{i=1}^{I^k} c_k(s_i^k,a_i^k) \mid P,\pi_k,s_1^k=s_0\Big] = J_k^{\pi_k}(s_0).$$

Theorem 4.4. *Under Assumptions 1 and 2, with probability at least $1-\delta$, the regret of SSP-O-REPS2 with known transition function and* $\eta = \sqrt{\frac{3\log(D|S||A|/c_{min})}{K}}$ *is*

$$R_K \leq O\left(\frac{D}{c_{min}}\sqrt{K\log^3\frac{KD|S||A|}{\delta c_{min}}}\right) = \widetilde{O}\left(\frac{D}{c_{min}}\sqrt{K}\right).$$

5 Unknown transition function

Our algorithms used the transition function in the definition of $\Delta(D/c_{\min})$, and in the computation of T^{π_k} and the fast policy π^f. When P is unknown, all of these must be performed w.r.t some estimation. Thus, our algorithm SSP-O-REPS3 keeps confidence sets that contain P with high probability, similarly to UCRL2 (Jaksch, Ortner, and Auer 2010).

The algorithm proceeds in *epochs* and updates the confidence set at the beginning of every epoch. The first epoch begins at the first time step, and an epoch ends once an episode ends or the number of visits to some state-action pair is doubled. Denote by $N^e(s,a)$ the number of visits to (s,a) up to (and not including) epoch e, and similarly $N^e(s,a,s')$. The empirical transition function for epoch e is defined by $\bar{P}_e(s' \mid s,a) = \frac{N^e(s,a,s')}{N^e_+(s,a)}$, where $N^e_+(s,a) = \max\{N^e(s,a),1\}$. The confidence set for epoch e contains all transition functions P' such that for every $(s,a,s') \in S \times A \times S^+$,

$$|P'(s' \mid s,a) - \bar{P}_e(s' \mid s,a)| \le \epsilon_e(s' \mid s,a) \stackrel{\text{def}}{=} \stackrel{\text{def}}{=} 4\sqrt{\bar{P}_e(s' \mid s,a)A^e(s,a)} + 28A^e(s,a),$$

where $A^e(s,a) = \frac{\log\left(|S||A|N^e_+(s,a)/\delta\right)}{N^e_+(s,a)}$, and the size of the confidence set is controlled by $\epsilon_e(s' \mid s,a)$.

In order to use our confidence sets together with OMD, we must extend occupancy measures to state-action-state triplets (Rosenberg and Mansour 2019a) as follows,

$$q^{P,\pi}(s,a,s') = \lim_{T\to\infty} \mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}\{s_t = s, a_t = a, s_{t+1} = s'\}\Big],$$

where $\mathbb{E}[\cdot]$ is shorthand for $\mathbb{E}[\cdot \mid P, \pi, s_1 = s_0]$ here. Now, an occupancy measure q corresponds to a policy-transition function pair with the inverse mapping given by

$$\pi^q(a \mid s) = \frac{q(s,a)}{q(s)} \;;\; P^q(s' \mid s,a) = \frac{q(s,a,s')}{q(s,a)},$$

where $q(s,a) \stackrel{\text{def}}{=} \sum_{s' \in S^+} q(s,a,s')$. Thus, we can incorporate the confidence sets into the OMD update of Eq. (4) by replacing the set $\Delta(D/c_{\min})$ with the set $\widetilde{\Delta}_e(D/c_{\min})$ – occupancy measures q whose induced transition function P^q is in the confidence set of epoch e and the expected time of π^q (w.r.t P^q) from s_0 to the goal is at most $D/c_{\min}$, i.e.,

$$q_{k+1} = q^{P_{k+1},\pi_{k+1}} = \underset{q \in \widetilde{\Delta}_{e(k+1)}(D/c_{\min})}{\arg\min} \eta\langle q, c_k\rangle + \mathrm{KL}(q \parallel q_k), \tag{9}$$

where $e(k)$ denotes the first epoch in episode k.

As in Section 4.1, this update can be performed in two steps, where the unconstrained minimization step is identical to Eq. (5) and the projection step is implemented similarly to Eq. (6) but with different constraints. Specifically, we accommodate the constraints (7) and (8) for the extended occupancy measures (see Rosenberg and Mansour (2019a)), and show that a set of linear constraints can express the condition that P^q is in the confidence set (see details in Appendix F). Note that D cannot be computed without knowing the transition function. Here we assume D is known, and in Section 6 we remove this assumption.

Similarly to SSP-O-REPS2, once we reach a state whose expected time to the goal is too long, we want to switch to the fast policy. However, since P is unknown we cannot compute T^{π_k} or the fast policy. Instead, we use the expected time of π_k w.r.t P_k which we denote by $\widetilde{T}_k^{\pi_k}$, and the optimistic fast policy $\widetilde{\pi}_e^f$. This policy, together with the optimistic fast transition function $\widetilde{P}_e^f$, minimizes the expected time to the goal out of all pairs of policies and transition functions from the confidence set. It is computed similarly to Cohen et al. (2020) (described in details in Appendix F).

Algorithm 3 SSP-O-REPS3

Input: $S, A, c_{\min}, \eta, \delta$.
Initialization:
  Get from user or estimate D (see Section 6).
  Initialize epoch counter $e \leftarrow 0$ and policy $\widetilde{\pi}_1^f$.
  Set $q_0(s,a,s') = 1$ and $c_0(s,a,s') = 0 \;\forall (s,a,s')$.
for $k = 1, 2, \ldots$ **do**
  Start new epoch $e \leftarrow e+1$, update confidence sets.
  Perform OMD step (see Appendix F):

$$q_k = \underset{q \in \widetilde{\Delta}_e(D/c_{\min})}{\arg\min} \eta\langle q, c_{k-1}\rangle + \mathrm{KL}(q \parallel q_{k-1}).$$

  Compute $\pi_k(a \mid s)$ and $P_k(s' \mid s,a)$ for every (s,a,s').
  Compute $\widetilde{T}_k^{\pi_k}(s)$ for every $s \in S$ using Eq. (2).
  set $s_1^k \leftarrow s_0$, $i \leftarrow 1$.
  while $s_i^k \neq g$ and $\widetilde{T}_k^{\pi_k}(s_i^k) < \frac{D}{c_{\min}}$ and s_i^k is known **do**
    Play action according to π_k, i.e., $a_i^k \sim \pi_k(\cdot \mid s_i^k)$.
    Observe next state $s_{i+1}^k \sim P(\cdot \mid s_i^k, a_i^k)$, $i \leftarrow i+1$.
    if number of visits to some (s,a) is doubled **then**
      Start new epoch $e \leftarrow e+1$, update confidence sets, and compute optimistic fast policy $\widetilde{\pi}_e^f$.
      Break
    end if
  end while
  while $s_i^k \neq g$ **do**
    if s_i^k is unknown **then**
      Play action a_i^k that was played the least in state s_i^k.
    else
      Play action according to $\widetilde{\pi}_e^f$, i.e., $a_i^k \sim \widetilde{\pi}_e^f(\cdot \mid s_i^k)$.
    end if
    Observe next state $s_{i+1}^k \sim P(\cdot \mid s_i^k, a_i^k)$, $i \leftarrow i+1$.
    if number of visits to some (s,a) is doubled **then**
      Start new epoch $e \leftarrow e+1$, update confidence sets, and compute optimistic fast policy $\widetilde{\pi}_e^f$.
    end if
  end while
  Observe c_k and suffer cost $\sum_{j=1}^{i-1} c_k(s_j^k, a_j^k)$.
end for

To make sure the optimistic fast policy reaches the goal

with high probability we distinguish between *known* and *unknown* states. Cohen et al. (2020) show that once all state-action pairs were visited at least $\Phi = \alpha \frac{D|S|}{c_{\min}^2} \log \frac{D|S||A|}{\delta c_{\min}}$ times (for some constant $\alpha > 0$), the optimistic fast policy is proper with high probability. Therefore, we say that a state s is *known* if (s, a) was visited Φ times for all actions $a \in A$. When reaching an unknown state, we play the least played action so far to make sure it will become known as fast as possible. In Cohen et al. (2020) known states are only considered implicitly in the analysis, but in the adversarial setting we must address them explicitly because our policies must be stochastic and therefore we are not guaranteed that actions that were not played enough will reach Φ visits.

To summarize, we start each episode k by playing π_k computed in Eq. (9), and maintain confidence sets that are updated at the beginning of every epoch. When we reach a state s such that $\widetilde{T}_k^{\pi_k}(s) \geq D/c_{\min}$, we switch to the optimistic fast policy. In addition, when an unknown state is reached we play the least played action up to this point and then switch to the optimistic fast policy. Finally, we also make the switch to the optimistic fast policy once the number of visits to some state-action pair is doubled, at which point we also recompute it. High-level pseudocode in Algorithm 3, and full pseudocode in Appendix G. Next, we give an overview of the regret analysis for SSP-O-REPS3, which yields the following regret bound (full proof in Appendix H).

Theorem 5.1. *Under Assumptions 1 and 2, with probability at least $1 - \delta$, the regret of SSP-O-REPS3 with known SSP-diameter D and $\eta = \sqrt{\frac{6 \log(D|S||A|/c_{min})}{K}}$ is*

$$R_K \leq \widetilde{O}\left(\frac{D|S|}{c_{min}} \sqrt{|A|K} + \frac{D^2|S|^2|A|}{c_{min}^2} \right) = \widetilde{O}\left(\frac{D|S|}{c_{min}} \sqrt{|A|K} \right),$$

where the last equality holds for $K \geq D^2|S|^2|A|/c_{min}^2$.

Our analysis follows the framework of Cohen et al. (2020) for analyzing optimism in SSPs, but makes the crucial adaptations needed to handle the adversarial environment.

We have two objectives: bounding the number of steps T taken by the algorithm (to show that we reach the goal in every episode) and bounding the regret. To bound the total time we split the time steps into *intervals*. The first interval begins at the first time step, and an interval ends once (1) an episode ends, (2) an epoch ends, (3) an unknown state is reached, or (4) a state s such that $\widetilde{T}_k^{\pi_k}(s) \geq D/c_{\min}$ is reached when playing π_k in episode k, i.e., there is a switch.

Intuitively, we will bound the length of every interval by $\widetilde{O}(D/c_{\min})$ with high probability, and then use the number of intervals $\widetilde{O}(K + D|S|^2|A|/c_{\min}^2)$ to bound the total time T. Finally, we will show that the regret scales with the square root of the total variance (which is the number of intervals times the variance in each interval) to finish the proof. While intuitive, this approach is technically difficult and therefore we apply these principles in a different way.

We start by showing that the confidence sets contain P with high probability, which is a common result (see, e.g., Zanette and Brunskill (2019); Efroni et al. (2019)). Define Ω^m the event that P is in the confidence set of the epoch that interval m belongs to.

Lemma 5.2 (Cohen et al. (2020), Lemma 4.2)**.** *With probability at least $1 - \delta/2$, the event Ω^m holds for all intervals m simultaneously.*

There are two dependant probabilistic events that are important for the analysis. The first are the events Ω^m, and the second is that the deviation in the cost of a given policy from its expected value is not large. To disentangle these events we define an alternative regret for every $M = 1, 2, \ldots,$

$$\widetilde{R}_M = \sum_{m=1}^{M} \sum_{h=1}^{H^m} \sum_{a \in A} \tilde{\pi}_m(a \mid s_h^m) c_m(s_h^m, a) \mathbb{I}\{\Omega^m\} - \sum_{k=1}^{K} J_k^{\pi^\star}(s_0),$$

where $c_m = c_k$ for the episode k that interval m belongs to, $\tilde{\pi}_m$ is the policy followed by the learner in interval m, H^m is the length of interval m, and the trajectory visited in interval m is $U^m = (s_1^m, a_1^m, \ldots, s_{H^m}^m, a_{H^m}^m, s_{H^m+1}^m)$.

We focus on bounding $\widetilde{R}_M$ because we can use it to obtain a bound on R_K. This is done using Lemma 5.2 and an application of Azuma inequality, when M is the number of intervals in which the first K episodes elapse (we show that the learner indeed completes these K episodes).

As mentioned, bounding the length of each interval complicates the analysis, and therefore we introduce artificial intervals. That is, an interval m also ends at the first time step H such that $\sum_{h=1}^{H} \sum_{a \in A} \tilde{\pi}_m(a \mid s_h^m) c_m(s_h^m, a) \geq D/c_{\min}$. The artificial intervals are only introduced for the analysis and do not affect the algorithm. Now, the length of each interval is bounded by $2D/c_{\min}^2$ and we can bound the number of intervals as follows.

Observation 5.3. Let $\widetilde{C}_M = \sum_{m=1}^{M} \sum_{h=1}^{H^m} \sum_{a \in A} \tilde{\pi}_m(a \mid s_h^m) c_m(s_h^m, a)$. The total time satisfies $T \leq \widetilde{C}_M / c_{\min}$ and the total number of intervals satisfies

$$M \leq \frac{c_{\min} \widetilde{C}_M}{D} + 2|S||A| \log T + 2K + 2\alpha \frac{D|S|^2|A|}{c_{\min}^2} \log \frac{D|S||A|}{\delta c_{\min}}.$$

Note that a confidence set update occurs only in the end of an epoch and thus $\Omega^m = \Omega^{m-1}$ for most intervals. Also, for artificial intervals the policy does not change. Next we bound $\widetilde{C}_M$ as a function of the number of intervals M. Through summation of our confidence bounds, and by showing that the variance in each interval is bounded by $D^2/c_{\min}^2$ we are able to obtain the following, when Lemma 5.2 holds,

$$\widetilde{C}_M \leq \sum_{k=1}^{K} \langle q_k, c_k \rangle + \widetilde{O}\left(\frac{D|S|}{c_{\min}} \sqrt{M|A|} + \frac{D^2|S|^2|A|}{c_{\min}^2} \right).$$

Substituting in Observation 5.3 and solving for $\widetilde{C}_M$ we get

$$\widetilde{R}_M = \widetilde{C}_M - \sum_{k=1}^{K} J_k^{\pi^\star}(s_0) \leq \sum_{k=1}^{K} \langle q_k - q^{P, \pi^\star}, c_k \rangle + \widetilde{O}\left(\frac{D|S|}{c_{\min}} \sqrt{|A|K} + \frac{D^2|S|^2|A|}{c_{\min}^2} \right),$$

Notice that the first term on the RHS of the inequality is exactly the regret of OMD, and therefore analyzing it similarly to Theorem 4.1 gives the final bound (see Appendix H.6).

5.1 Adversarial vs. stochastic costs in SSP

The main challenge in SSP with stochastic costs is estimating the transition function, since estimating the cost is faster. Therefore, Tarbouriech et al. (2019); Cohen et al. (2020) follow the optimism principle w.r.t the estimated cost function. This approach has two benefits: the cost in each interval is at most $\widetilde{O}(D)$ (since it is bounded by the cost of the optimal policy), and the optimism bypasses the need to know D.

In adversarial SSP our main mechanism must be OMD (or similar methods from online learning) to handle the arbitrarily changing cost functions. The optimism is used as a secondary mechanism, as we still need to estimate the fixed transition function and make sure that the learner reaches the goal state in every episode. The main challenge is accommodating the optimistic framework and known states tracking of Cohen et al. (2020), to the main method in which we pick policies – online mirror descent.

Thus, in the adversarial setting we cannot enjoy the same benefits as in the stochastic setting. The cost in each interval is bounded by $\widetilde{O}(D/c_{\min})$ (since this is a bound on the time of the best policy in hindsight, and the cost cannot be estimated), and we need to know (or estimate) D in order to force our policies to reach the goal fast enough. This leads to the extra $1/c_{\min}$ factor in our regret compared to Cohen et al. (2020). However, it is worth mentioning that their bound depends on $c_{\min}$ in the additive term, and that Tarbouriech et al. (2019) have a $1/\sqrt{c_{\min}}$ factor in the main term of the regret.

On a technical level, we have to make subtle changes in order to use OMD in the optimistic framework of Cohen et al. (2020). While known states are only implicit in the analysis of Cohen et al. (2020), using stochastic policies (which is necessary in adversarial environments) forces us to make the known states tracking explicit, i.e., play the least played action in unknown states to make sure the visits count advances for all actions. Furthermore, when the dynamics are known it is clear that switching to the fast policy in "bad" states does not suffer more cost, but here the switch is w.r.t the optimistic transition function, and showing that this does not hurt the regret is a subtle argument (see details in Appendix H). Moreover, showing that the variance in each interval is of order $D^2/c_{\min}^2$ even when OMD is involved also requires some sophisticated adaptations, e.g., an alternative definition of artificial intervals.

In terms of computational complexity, we compute the optimistic fast policy in the end of each epoch which is similar to Cohen et al. (2020). However, in the beginning of an episode they compute an optimistic policy while we perform an OMD step. While this is more costly, it is unavoidable, similarly to adversarial episodic MDPs. In fact, an OMD step here is more efficient than in the episodic setting since our policies are not time-dependent, and therefore our optimization problem has $O(|S|^2|A|)$ variables compared to $O(H|S|^2|A|)$ (where H is the episode length).

6 Relaxation of assumptions

Estimating the SSP-diameter. We use the SSP-diameter D only in the definition of the sets $\widetilde{\Delta}_e(D/c_{\min})$. A key point in the analysis is that the occupancy measure of the best policy in hindsight $q^{P,\pi^\star}$ is contained in the sets on which we perform OMD $\widetilde{\Delta}_e(D/c_{\min})$ (with high probability). To that end, we chose $D/c_{\min}$ as upper bound on $T^{\pi^\star}(s_0)$ (see Appendix E). Once D is unknown, we want to compute an alternative upper bound $\widetilde{D}$ on the expected time of the fast policy $T^{\pi^f}(s_0)$, and then $\widetilde{D}/c_{\min}$ will upper bound $T^{\pi^\star}(s_0)$.

We dedicate the first L episodes to estimating this upper bound $\widetilde{D}$, before running SSP-O-REPS3. Notice that π^f is the optimal policy w.r.t the constant cost function $c(s,a) = 1$, and its expected cost is $T^{\pi^f}(s_0)$. Thus, to compute $\widetilde{D}$ we run an algorithm for regret minimization in regular SSPs for L episodes with this cost function, and set $\widetilde{D}$ to be the average cost per episode times a constant, 10 for example.

By the regret bound of Cohen et al. (2020), we can set $L = \widetilde{\Theta}\big(\max\{|S|\sqrt{|A|K}/c_{\min}, |S|^2|A|/c_{\min}^2\}\big)$ without suffering additional regret. In Appendix I, we show that this yields the two properties we desire, with high probability, for large enough K. First, $q^{P,\pi^\star} \in \widetilde{\Delta}_e(\widetilde{D}/c_{\min})$, i.e., $\widetilde{D}$ is an upper bound on $T^{\pi^f}(s_0)$. Second, $\widetilde{D}$ is not too large, i.e., $\widetilde{D} \leq O(D)$. Therefore, we get the same regret bound as in Theorem 5.1. When K is not large enough, we show that the regret is bounded by the constant factor $\widetilde{O}(D^4|S|^2|A|/c_{\min}^2)$.

Zero costs. Similarly to Tarbouriech et al. (2019); Cohen et al. (2020), we can eliminate Assumption 2 by applying a perturbation to the instantaneous costs. That is, instead of c_k we use the cost function $\tilde{c}_k(s,a) = \max\{c_k(s,a), \epsilon\}$ for some $\epsilon > 0$. Thus, Assumption 2 holds with $c_{\min} = \epsilon$, but we introduced additional bias into the model. Choosing $\epsilon = \Theta(K^{-1/4})$ ensures that all our algorithms obtain regret bounds of $\widetilde{O}(K^{3/4})$ for the general case (see in Appendix I).

7 Conclusions and future work

In this paper we present the first algorithms to achieve sub-linear regret for stochastic shortest path with arbitrarily changing costs. We show efficient algorithms with high probability regret bounds of $\widetilde{O}(\sqrt{K})$ when costs are strictly positive and $\widetilde{O}(K^{3/4})$ in the general case. We hope this paper paves the way to achieve tight regret bounds with practical algorithms in this important setting of adversarial SSP.

Closing the gap from the lower bound is the natural open problem that arises from this paper. The second direction that should be studied is bandit feedback. In this work we assumed that the entire cost function is revealed to the learner in the end of the episode, i.e., full information feedback. However, in many natural applications, the learner only observes the costs associated with the actions it took – this is called bandit feedback. Extending our results to bandit feedback is not trivial, even when the transition function is known, and is left for future work. Finally, it is of great importance to see if policy optimization methods can also obtain regret bounds in adversarial SSPs as done in adversarial MDPs recently (Cai et al. 2019; Efroni et al. 2020), since they are widely used in practice.

References

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

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

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

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

Boyd, S.; and Vandenberghe, L. 2004. *Convex Optimization*. New York, NY, USA: Cambridge University Press. ISBN 0521833787.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2019. Provably Efficient Exploration in Policy Optimization. *CoRR* abs/1912.05830.

Cohen, A.; Kaplan, H.; Mansour, Y.; and Rosenberg, A. 2020. Near-optimal Regret Bounds for Stochastic Shortest Path. *CoRR* abs/2002.09869.

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

d'Epenoux, F. 1963. A probabilistic production and inventory problem. *Management Science* 10(1): 98–108.

Efroni, Y.; Merlis, N.; Ghavamzadeh, M.; and Mannor, S. 2019. 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*, 12203–12213.

Efroni, Y.; Shani, L.; Rosenberg, A.; and Mannor, S. 2020. Optimistic Policy Optimization with Bandit Feedback. *arXiv preprint arXiv:2002.08243* .

Even-Dar, E.; Kakade, S. M.; and Mansour, Y. 2009. Online Markov decision processes. *Mathematics of Operations Research* 34(3): 726–736.

Fruit, R.; Pirotta, M.; Lazaric, A.; and Ortner, R. 2018. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. *arXiv preprint arXiv:1802.04020* .

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

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 4863–4873.

Jin, T.; and Luo, H. 2019. Learning Adversarial MDPs with Bandit Feedback and Unknown Transition. *CoRR* abs/1912.01192.

Manne, A. S. 1960. Linear programming and sequential decisions. *Management Science* 6(3): 259–267.

Neu, G.; György, A.; and Szepesvári, C. 2010. The Online Loop-free Stochastic Shortest-Path Problem. In *Conference on Learning Theory (COLT)*, 231–243.

Neu, G.; György, A.; and Szepesvári, C. 2012. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 805–813.

Neu, G.; György, A.; Szepesvári, C.; and Antos, A. 2014. Online Markov Decision Processes Under Bandit Feedback. *IEEE Trans. Automat. Contr.* 59(3): 676–691.

Osband, I.; Van Roy, B.; and Wen, Z. 2016. Generalization and Exploration via Randomized Value Functions. In *International Conference on Machine Learning*, 2377–2386.

Rosenberg, A.; and Mansour, Y. 2019a. Online Convex Optimization in Adversarial Markov Decision Processes. In *International Conference on Machine Learning*, 5478–5486.

Rosenberg, A.; and Mansour, Y. 2019b. Online Stochastic Shortest Path with Bandit Feedback and Unknown Transition Function. In *Advances in Neural Information Processing Systems*, 2209–2218.

Tarbouriech, J.; Garcelon, E.; Valko, M.; Pirotta, M.; and Lazaric, A. 2019. No-Regret Exploration in Goal-Oriented Reinforcement Learning.

Yu, J. Y.; Mannor, S.; and Shimkin, N. 2009. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research* 34(3): 737–757.

Zanette, A.; and Brunskill, E. 2019. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. In *International Conference on Machine Learning*, 7304–7312.

Zimin, A.; and Neu, G. 2013. 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*, 1583–1591.