Stochastic Shortest Path with Adversarially Changing Costs

Aviv Rosenberg[1] and **Yishay Mansour**[1,2]
[1]Tel Aviv University, Israel
[2]Google Research, Tel Aviv
{avivros007,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 present the adversarial SSP model that also accounts for adversarial changes in the costs over time, while the underlying transition function remains unchanged. Formally, an agent interacts with an SSP environment for K episodes, the cost function changes arbitrarily between episodes, and the transitions are unknown to the agent. We develop the first algorithms for adversarial SSPs and prove high probability regret bounds of square-root K assuming all costs are strictly positive, and sub-linear regret in the general case. 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). It features an agent that interacts with a Markov decision process (MDP) with the aim of reaching a predefined goal state in minimum total expected cost. Many important RL problems fall into the SSP framework, e.g., car navigation and Atari games, and yet it was only rarely studied from a theoretical point of view until very recently, mainly due to its challenging nature in comparison to finite-horizon, average-reward or discounted MDPs. For example, in SSP some polices might suffer infinite cost.

An important aspect that the standard SSP model fails to capture is changes in the environment over time (e.g., changes in traffic when navigating a car). In the finite-horizon setting, the adversarial MDP model was proposed to address changing environments, and has gained considerable popularity in recent years. It allows the cost function to change arbitrarily over time, while still assuming a fixed transition function.

In this work we present the adversarial SSP model that introduces adversarially changing costs to the classical SSP model. Formally, the agent interacts with an SSP instance for K episodes, and the cost function changes arbitrarily between episodes. The agent's objective is to reach the goal state in all episodes while minimizing its total expected cost. Its performance is measured by the *regret*, defined as the cumulative difference between the agent's total cost in K episodes and the expected total cost of the best policy in hindsight.

Finite-horizon MDPs are a special case of the general SSP problem where the agent is guaranteed to reach the goal state within a fixed number of steps H. This model is extensively studied in recent years for both stochastic and adversarial costs. In the adversarial MDP literature it is better known as the *loop-free* SSP model. While having a similar name, loop-free SSP follows the restrictive assumption that after H steps the goal will be reached and is thus far less challenging.

As pointed out by [Tarbouriech *et al.*, 2020], in the general SSP problem we face new challenges that do not arise in the loop-free version. Notably, it features two possibly conflicting objectives – reaching the goal vs minimizing cost; and it requires handling unbounded value functions and episode lengths. In the adversarial SSP model, these difficulties are further amplified as the adversary might encourage the learner to use "slow" policies and then punish her with large costs.

In this paper we propose the first algorithms for regret minimization in adversarial SSPs without any restrictive assumptions (namely, loop-free assumption). While we leverage algorithmic and technical tools from both SSP and finite-horizon adversarial MDP, tackling the general SSP problem in the presence of an adversary requires novel techniques and careful analysis. Our algorithms are based on the popular online mirror descent (OMD) framework for online convex optimization (OCO). However, naive application of OMD to SSP cannot overcome the challenges mentioned above as we later show, and we use carefully designed mechanisms to establish our theoretical guarantees.

The main contributions of this paper are as follows. First, we formalize the adversarial SSP model and define the notion of learning and regret. Second, we establish an efficient implementation of OMD in the SSP model with known transitions and study the conditions under which it guarantees near-optimal $\sqrt{K}$ expected regret, showing that some modifications are necessary. Then, we illustrate the challenge of obtaining regret bounds in high probability in adversarial SSPs, and present a novel method that allows OMD to obtain its regret with high probability. Finally, we tackle unknown transitions. We describe the crucial adaptations that allow OMD to be combined with optimistic estimates of the transition function and guarantee $\sqrt{K}$ regret when all costs are strictly positive, and $K^{3/4}$ regret in the general case. Hopefully, the

infrastructure created in this paper for handling adversarial costs in SSPs with unknown transition function paves the way for future work to achieve minimax optimal regret bounds.

Related work. Early work by [Bertsekas and Tsitsiklis, 1991] studied the planning problem in SSPs, i.e., computing the optimal strategy efficiently when parameters are known. Under certain assumptions, they established that the optimal strategy is a deterministic stationary policy (a mapping from states to actions) and can be computed efficiently using standard planning algorithms, e.g., Value Iteration and LP.

Recently [Tarbouriech *et al.*, 2020] presented the problem of learning SSPs (with stochastic costs) and provided the first algorithms with sub-linear regret but with dependence on the minimal cost $c_{\min}$. Their results were further improved by [Rosenberg *et al.*, 2020] that eliminate the $c_{\min}$ dependence and prove high probability regret bound of $\widetilde{O}(D|S|\sqrt{|A|K})$ complemented by a nearly matching lower bound of $\Omega(D\sqrt{|S||A|K})$, where D is the diameter, S is the state space and A is the action space.

As mentioned before, regret minimization in RL is extensively studied in recent years, but the literature mainly focuses on the average-reward infinite-horizon model [Bartlett and Tewari, 2009; Jaksch *et al.*, 2010] and on the finite-horizon model [Osband *et al.*, 2016; Azar *et al.*, 2017; Dann *et al.*, 2017; Jin *et al.*, 2018; Zanette and Brunskill, 2019; Efroni *et al.*, 2019]. Adversarial MDPs were also first studied in the average-reward model [Even-Dar *et al.*, 2009; Neu *et al.*, 2014], before focusing on the finite-horizon setting which is typically referred to as loop-free SSP. Early work in this setting by [Neu *et al.*, 2010] used a reduction to multi-arm bandit [Auer *et al.*, 2002], but then [Zimin and Neu, 2013] introduced the O-REPS framework, which is the implementation of OMD in finite-horizon MDPs. All these works assume known transition function, but more recently [Neu *et al.*, 2012; Rosenberg and Mansour, 2019a; Rosenberg and Mansour, 2019b; Jin *et al.*, 2020; Shani *et al.*, 2020; Cai *et al.*, 2020] consider unknown transitions.

We stress that all previous work in the adversarial setting made the restrictive loop-free assumption, avoiding the main challenges tackled in this paper. Building on our methodologies, [Chen *et al.*, 2020] recently extended our work and obtained minimax optimal $\sqrt{K}$ regret with known transitions. However, they do not consider the more challenging unknown transitions case, and also assume that the learner knows in advance the running time of the best policy in hindsight.

2 Preliminaries

An adversarial SSP problem is defined by an MDP $M = (S, A, P, s_0, g)$ and a sequence $\{c_k : S \times A \to [0,1]\}_{k=1}^{K}$ of cost functions. S and A are finite state and action spaces, respectively, $s_0 \in S$ is an initial state and $g \notin S$ is the goal state. P is a transition function such that $P(s' \mid s, a)$ gives the probability to move to s' when taking action a in state s, and thus $\sum_{s' \in S \cup \{g\}} P(s' \mid s, a) = 1$ for every $(s, a) \in S \times A$.

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. Formally, the learner starts each episode k at the initial state[1] $s_1^k = s_0$. In each step i of the episode, the learner observes its current state s_i^k, picks an action a_i^k and moves to the next state s_{i+1}^k sampled from $P(\cdot \mid s_i^k, a_i^k)$. The episode ends when the goal state g is reached, and then the learner observes c_k and suffers cost $\sum_{i=1}^{I^k} c_k(s_i^k, a_i^k)$ where I^k is the length of the episode. Importantly, I^k is a random variable that might be infinite. This is the unique challenge of SSP compared to finite-horizon.

Proper Policies. A stationary policy $\pi : A \times S \to [0,1]$ is a mapping such that $\pi(a \mid s)$ gives the probability that action a is selected in state s. A policy π is called *proper* if playing according to π ensures that the goal state is reached with probability 1 when starting from any state (otherwise it is *improper*). Since reaching the goal is one of the learner's main objectives, we make the basic assumption that there exists at least one proper policy. This is equivalent to the assumption that the goal state is reachable from every state, which is clearly a necessary assumption.

We denote by $T^\pi(s)$ the expected hitting time of g when playing according to π and 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 $s' \in S$ such that $T^\pi(s') = \infty$. 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., $J^\pi(s) = \lim_{T \to \infty} \mathbb{E}\big[\sum_{t=1}^{T} c(s_t, a_t) \mid P, \pi, s_1 = s\big]$. For a proper policy π, it follows that $J^\pi(s)$ is finite for all s.

Under the additional assumption that every improper policy suffers infinite expected 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 $s \in S$:

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

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

Learning Formulation. The learner's goal is to minimize its total cost. Its performance 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 J_k^π is the cost-to-go of policy π with respect to (w.r.t) cost function c_k, and Π_{proper} is the set of proper policies. If I^k is infinite for some k, we define $R_K = \infty$ forcing the learner to reach the goal in every episode. We also denote by $\pi^\star = \arg\min_{\pi \in \Pi_{\text{proper}}} \sum_{k=1}^{K} J_k^\pi(s_0)$ the best policy in hindsight.

Our analysis makes use of the Bellman equations, that hold under the conditions described before Eq. (1). To make sure these are met, we assume that the costs are strictly positive.

Assumption 1. *All costs are positive, i.e., there exists $c_{min} > 0$ such that $c_k(s,a) \geq c_{min}$ for every k and $(s,a) \in S \times A$.*

[1] Our algorithms readily extend to a fixed initial distribution.

We can easily eliminate Assumption 1 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 $\epsilon > 0$. This ensures that the effective minimal cost is $c_{\min} = \epsilon$, at the price of introducing additional bias. Choosing $\epsilon = \Theta(K^{-1/4})$ ensures that all our algorithms obtain regret bounds of $\widetilde{O}(K^{3/4})$ in the general case. See details in Appendix K and discussion about $c_{\min}$ in Section 5.

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, i.e.,

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

where $\mathbb{I}\{\cdot\}$ is the indicator function. Note that for a proper policy π, $q^\pi(s,a)$ is finite for every (s,a). In fact, the correspondence between proper policies and finite occupancy measures is 1-to-1, and its inverse[2] for q is given by $\pi^q(a \mid s) = \frac{q(s,a)}{q(s)}$ where $q(s) = \sum_{a\in A} q(s,a)$ is the expected number of visits to s. The 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]. Notice that the expected cost of policy π is linear w.r.t q^π, i.e.,

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

Thus, minimizing the expected regret can be written as an instance of online linear optimization in the following manner,

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

3 Known Transition Function

We start with the simpler (yet surprisingly challenging) case where P is known to the learner. Recall that while the transition function is known, the costs change arbitrarily between episodes. In Section 3.1 we establish the implementation of the OMD method in SSP, and in Section 3.2 we use it to obtain a high probability regret bound.

3.1 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; Rosenberg and Mansour, 2019b; Jin *et al.*, 2020]. 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 fails in SSP because it does not guarantee that the learner plays proper policies. For example, in the first episode these algorithms play the uniform policy which may suffer exponential cost (see Appendix A).

Thus, we propose to apply OMD to the set $\Delta(\tau)$ – occupancy measures of policies π that reach the goal in expected time $T^\pi(s_0) \le \tau$. This set is convex and has a compact representation as we show shortly. Our algorithm SSP-O-REPS operates as follows. In the beginning of episode k, it picks an occupancy measure q_k from $\Delta(\tau)$ which minimizes a trade-off between the current cost function and the distance to the previously chosen occupancy measure. Then, it extracts the policy $\pi_k = \pi^{q_k}$ and plays it through the episode. Formally,

$$q_k = q^{\pi_k} = \arg\min_{q\in\Delta(\tau)} \eta\langle q, c_{k-1}\rangle + \mathrm{KL}(q \parallel q_{k-1}), \qquad (2)$$

where $\mathrm{KL}(\cdot||\cdot)$ is the KL-divergence, and $\eta > 0$ is a learning rate. Computing q_k is implemented in two steps: first find the unconstrained minimizer and then project it into $\Delta(\tau)$, i.e.,

$$q_k' = \arg\min_{q} \eta\langle q, c_{k-1}\rangle + \mathrm{KL}(q \parallel q_{k-1}) \qquad (3)$$

$$q_k = \arg\min_{q\in\Delta(\tau)} \mathrm{KL}(q \parallel q_k'). \qquad (4)$$

Eq. (3) has a closed form $q_k'(s,a) = q_{k-1}(s,a)e^{-\eta c_{k-1}(s,a)}$, and Eq. (4) can be formalized as a constrained convex optimization problem with the following linear constraints:

$$\forall s. \sum_{a\in A} q(s,a) - \sum_{s'\in S}\sum_{a'\in A} q(s',a')P(s|s',a') = \mathbb{I}\{s = s_0\}$$

$$\sum_{s\in S}\sum_{a\in A} q(s,a) \le \tau, \qquad (5)$$

where we omitted non-negativity constraints. The first set of constraints are standard flow constraints, while the novel constraint (5) ensures that $T^{\pi^q}(s_0) \le \tau$. In Appendix B we show how to solve this problem efficiently and describe implementation details for the algorithm. Pseudocode in Appendix C.

Finally, we need to pick the parameter τ. While it needs to upper bound $T^{\pi^\star}(s_0)$ in order to have $q^{\pi^\star} \in \Delta(\tau)$, we want it to be as small as possible to get tighter regret guarantees. To that end, define the SSP-diameter [Tarbouriech *et al.*, 2020] $D = \max_{s\in S} \min_{\pi\in\Pi_{\text{proper}}} T^\pi(s)$ and pick $\tau = D/c_{\min}$. The diameter can be computed efficiently by finding the optimal policy w.r.t the constant cost function $c(s,a) = 1$ (see Appendix B). We refer to this policy as the fast policy π^f, and it holds that $D = \max_{s\in S} T^{\pi^f}(s)$.

Indeed $q^{\pi^\star} \in \Delta(D/c_{\min})$ because the total cost of the best policy in hindsight in K episodes is upper bounded by the total cost of any other policy, e.g., the fast policy (which is at most DK), and is lower bounded by the expected time of $\pi^\star$ times the minimal cost, i.e., $J_k^{\pi^\star}(s_0) \ge c_{\min} T^{\pi^\star}(s_0)$ (see Appendix D). In Appendix A we also show that this choice of τ cannot be smaller in general.

In Appendix D we provide the full analysis of the algorithm yielding the following regret bound in expectation. Moreover,

[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.

we show that all the chosen policies must be proper and therefore the goal is reached with probability 1 in all episodes.

Theorem 1. *Under Assumption 1, the expected regret of SSP-O-REPS with known transition function and $\eta = \widetilde{\Theta}(\frac{1}{\sqrt{K}})$ is*

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

3.2 High Probability Regret Bound

To obtain high probability regret bounds, we must control the deviation between the learner's suffered cost and its expected value. While this is easily achievable in the finite-horizon setting through an application of Azuma inequality, it appears a major challenge in SSP since there is no finite upper bound on the learner's cost. In fact, Appendix A illustrates a simple example with 0 expected regret, but constant probability to suffer large regret (linear in K). The idea here is that even though a policy has small cost in expectation, there might be a tiny probability that it suffers huge cost (this cannot happen in finite-horizon since the cost is always bounded by H). Finally, even an event with tiny probability will happen at least once if there is a large number of episodes K.

Our strategy to control the deviation between the learner's actual suffered cost and its expected value is based on the observation that this quantity is closely related to the expected time to reach the goal from any state. This is illustrated by the following lemma whose proof is based on an adaptation of Azuma inequality to unbounded martingales (Theorem 11) which may be of independent interest.

Lemma 1. *Assume that in each episode k the learner plays a strategy σ_k such that the expected time to reach the goal from any state is at most τ. Then, with probability at least $1-\delta$,*

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

Thus, bounding the regret in high probability boils down to guaranteeing that $T^{\pi_k}(s) \leq D/c_{\min}$ for all $s \in S$ and not just s_0. Unfortunately, these constraints admit a non-convex set of occupancy measures. To bypass this issue we propose the SSP-O-REPS2 algorithm that operates as follows: start every episode k by playing the policy π_k chosen by SSP-O-REPS (i.e., Eq. (2)), but once we reach a state s whose expected time to the goal is too long (i.e., $T^{\pi_k}(s) \geq D/c_{\min}$), switch to the fast policy π^f. We defer to the pseudocode in Appendix E.

Now the conditions of Lemma 1 are clearly met, so it remains to relate the expected cost of our new strategy σ_k to this of π_k. The key novelty of our mid-episode policy switch is the timing. The naive approach would be to perform the switch when the policy takes too long, but then there is no way to bound the excess cost when compared to that of π_k. Performing the switch only once a "bad" state is reached ensures that the expected cost of σ_k can only be better than π_k. The analysis in Appendix F makes these claims formal and proves the following high probability regret bound.

Theorem 2. *Under Assumption 1, with probability $1-\delta$, the regret of SSP-O-REPS2 with known transition function is*

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

4 Unknown Transition Function

A standard technique to deal with unknown transition function in adversarial MDPs is to use optimistic estimates of P. We follow this approach but, as in the known transitions case, crucial modifications are necessary to apply optimism and obtain regret guarantees. In this section we describe our SSP-O-REPS3 algorithm for unknown transitions.

We start by describing the confidence sets and transition estimates used by the algorithm. SSP-O-REPS3 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 by $N^e(s,a,s')$ the number of times this was followed by a transition to s'. Let $N_+^e(s,a) = \max\{N^e(s,a),1\}$ and define the empirical transition function for epoch e by $\bar{P}_e(s'|s,a) = N^e(s,a,s')/N_+^e(s,a)$. Finally, define the confidence set for epoch e as the set of all transition functions P' such that for every $(s,a,s') \in S \times A \times (S \cup \{g\})$,

$$|P'(s' \mid s,a) - \bar{P}_e(s' \mid s,a)| \leq \epsilon_e(s' \mid s,a),$$

where $\epsilon_e(s'|s,a) = 4\sqrt{\bar{P}_e(s'|s,a)A^e(s,a)} + 28A^e(s,a)$ is the confidence set radius for $A^e(s,a) = \frac{\log\left(|S||A|N_+^e(s,a)/\delta\right)}{N_+^e(s,a)}$. By Bernstein inequality (see, e.g., [Azar *et al.*, 2017]), these confidence sets contain P with probability $1-\delta$ for all epochs.

Next, we extend our OMD implementation to the unknown transitions case. We follow the elegant approach of [Rosenberg and Mansour, 2019a] that use occupancy measures that are extended to include a transition function as well, that is,

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

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 transition function-policy pair with the inverse mapping given by

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

where $q(s,a) = \sum_{s'\in S\cup\{g\}} q(s,a,s')$ is the expected number of visits to (s,a) w.r.t P^q when playing π^q. We extend the set $\Delta(\tau)$ (which we cannot compute without knowing P), and perform OMD on the set $\widetilde{\Delta}_e(\tau)$ that changes through epochs. $\widetilde{\Delta}_e(\tau)$ is defined as the set of 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 τ. This set is again convex with a compact representation, and it admits the following OMD update step,

$$q_k = q^{P_k,\pi_k} = \operatorname*{arg\,min}_{q\in\widetilde{\Delta}_{e(k)}(\tau)} \eta\langle q, c_{k-1}\rangle + \mathrm{KL}(q \parallel q_{k-1}), \quad (6)$$

where $e(k)$ denotes the first epoch in episode k. Similarly to the known transitions case, this update can be performed efficiently. See Appendix G for details of the implementation.

In contrast to the known transitions case, this version of OMD cannot even guarantee bounded regret in expectation, because without knowledge of the transition function there is no guarantee that the chosen policies are even proper. Note that in the easier loop-free SSP setting, this OMD version is enough to guarantee a high probability regret bound even with unknown transitions. We now describe the mechanisms that need to be combined with OMD to obtain our regret bound.

Similarly to Section 3.2, we must make sure that the learner does not take too much time to reach the goal. The problem now is that we cannot compute its expected time T^{π_k} since P is unknown. Instead, we use the expected time of π_k w.r.t P_k (denoted by $\widetilde{T}_k^{\pi_k}$) which is an estimate of T^{π_k}, but not necessarily an optimistic one. Once a state s is reached such that $\widetilde{T}_k^{\pi_k}(s) \geq D/c_{\min}$ we want to switch to the fast policy π^f which again cannot be computed without knowing P. This policy is replaced with its optimistic estimate $\widetilde{\pi}_e^f$, which we refer to as the optimistic fast policy. Together with the optimistic fast transition function $\widetilde{P}_e^f$, this policy minimizes the expected time to the goal out of all pairs of policies and transition functions from the confidence set of epoch e. The details of computing the optimistic fast policy are in Appendix G.

If we were in the known transitions case, this would have been enough. So it seems that it should also suffice with unknown transitions, if we recompute the optimistic fast policy in the end of every epoch similarly to [Rosenberg *et al.*, 2020]. However, in the adversarial setting this approach fails for two main reasons. First, we cannot guarantee that $\widetilde{T}_k^{\pi_k}$ is a good enough estimate of T^{π_k} in all states. Second, the learner's policy is stochastic which means that we cannot guarantee all actions are being explored enough (as opposed to [Rosenberg *et al.*, 2020] that only play deterministic policies since they do not tackle adversarial costs). To overcome these challenges, we propose to force exploration in the following manner. Define a state to be *unknown* until every action was played at least $\Phi = \alpha \frac{D|S|}{c_{\min}^2} \log \frac{D|S||A|}{\delta c_{\min}}$ times in this state (for some constant $\alpha > 0$), and *known* afterwards. When reaching an unknown state, we play the least played action so far (forcing exploration), and only then switch to the optimistic fast policy. The idea behind this forced exploration is inspired by [Rosenberg *et al.*, 2020] that show that once all states are known, the optimistic fast policy is proper with high probability.

To summarize, SSP-O-REPS3 operates as follows. We start each episode k by playing the policy π_k computed in Eq. (6), 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. We defer to the full pseudocode in Appendix H and to the full analysis in Appendix I that yields the following regret bound.

Theorem 3. *Under Assumption 1, with probability $1-\delta$, the regret of SSP-O-REPS3 with known SSP-diameter D is*

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

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

Our analysis builds on ideas from [Rosenberg *et al.*, 2020] that analyze optimistic algorithms in SSP with stochastic costs. However, for the many reasons described in this paper and because our algorithm is not optimistic, many novel technical adaptions are needed in order to tackle the new challenges that arise when both the costs are adversarial and the transition function is unknown. Due to lack of space these are mostly presented in Appendix I, but here we give a short overview of the analysis.

Recall that the learner has two objectives in SSP: minimizing cost and reaching the goal. When transitions were known, we used Lemma 1 to say that (with high probability) the goal is reached in every episode, and then we could simply focus on bounding the regret. With unknown transitions, the argument for bounding the total time becomes more involved. The idea is that (with high probability) the number of steps between policy switches cannot be too long, as a consequence of our added mechanisms. To that end, 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 policy switch is made due to reaching a "bad" state. Intuitively, we bound the length of every interval by $\widetilde{O}(D/c_{\min})$ with high probability, and then use fact that the number of intervals is bounded by $\widetilde{O}(K + D|S|^2|A|/c_{\min}^2)$ to bound the total time. Then, we show that the regret of the learner can be bounded by the regret of OMD (analyzed in Section 3) plus the square root of the total variance (times $|S|^2|A|$). Finally, we obtain our regret bound by noticing that the total variance is equal to the variance in each interval times the number of intervals, and bounding the variance in an interval by $O(D^2/c_{\min}^2)$.

Estimating the SSP-diameter. When the transition function is unknown, we cannot compute the diameter D. However, a careful look at our algorithms shows that we use it only twice. First, we pick $\tau = D/c_{\min}$ as an upper bound on the expected time of the best policy in hindsight. For this purpose it is enough to use $T^{\pi^f}(s_0)/c_{\min}$, and therefore we shall dedicate the first L episodes to computing an estimate $\widetilde{D}(s_0)$ of $T^{\pi^f}(s_0)$ before running SSP-O-REPS3. Second, D is used to make a switch when a "bad" or unknown state s is reached, but again it is enough to use $T^{\pi^f}(s)$ instead. Similarly, we use the first L visits to s to estimate $T^{\pi^f}(s)$ and then continue executing the algorithm with $\widetilde{D}(s)$ instead of D.

To compute $\widetilde{D}(s)$ we run the algorithm of [Rosenberg *et al.*, 2020] for regret minimization in SSP with constant cost of 1 (since it measures time). By their regret bound, we can set $L \approx \sqrt{K}$ and suffer negligible additional regret. This is

also enough to yield the two properties we need in order to keep the same regret bound (with high probability): $\widetilde{D}(s)$ is an upper bound on $T^{\pi^f}(s)$ for any $s \in S$, and $\widetilde{D}(s) \leq O(D)$ (i.e., it is not too large). Details and full proofs in Appendix J.

5 Discussion

Lower bound and future work. In this paper we presented the first algorithms to achieve sub-linear regret in SSP with adversarially changing costs. Building on some of our ideas, [Chen *et al.*, 2020] recently proposed sophisticated algorithms with minimax optimal regret of $\widetilde{O}(\sqrt{DT_\star K})$ in the known transitions case, where $T_\star$ is the expected time of the best policy in hindsight. Interestingly, their lower bound reveals a gap from the stochastic setting (and from finite-horizon adversarial MDPs), showing that the adversarial SSP model is indeed significantly more challenging than previous models. Moreover, it shows that our regret bounds are near-optimal (up to $1/\sqrt{c_{\min}}$) in the hard case where the expected time of $\pi^\star$ is as large as $D/c_{\min}$ (see example in Appendix A).

There are still many interesting open problems in adversarial SSPs. Achieving minimax optimal regret with unknown dynamics is an important open problem that can hopefully be solved using some of the techniques presented here. The known transitions case is still far from solved as well. The algorithm of [Chen *et al.*, 2020] requires knowing $T_\star$ in advance which is a very restrictive assumption. Estimating $T_\star$ on the fly is another important open problem which seems very challenging due to the adversarially changing costs.

SSP vs finite-horizon. As this paper and the works of [Tarbouriech *et al.*, 2020; Rosenberg *et al.*, 2020] attempt to show, the SSP problem presents very different challenges than finite-horizon MDPs (or equivalently loop-free SSPs) although they are seemingly similar in structure. These differences stem from the double objective that the agent has to face in SSP, i.e., minimizing cost vs reaching the goal, while the only focus of the finite-horizon model is minimizing cost (the time of each episode is bounded by H by definition). Apart from the conceptual difference, this leads to numerous technical challenges, where the biggest one is unbounded value functions and episode lengths. Note that almost every online learning problem has some boundness assumptions and therefore novel technical tools must be used here (or at least non-trivial adaptations of existing tools, e.g., Theorem 11).

Dealing with adversarial costs in SSP is challenging even when the transition function is known to the learner. As described in this paper, using occupancy measures, this becomes an online linear optimization problem. However, unlike the finite-horizon case, in the SSP setting the decision set (i.e., the set of occupancy measures) does not have a bounded diameter (in finite-horizon it has diameter H), and this is the source of the unique challenges. To address these issues, we proposed to limit the decision set so it has a finite diameter (but still contains the best occupancy measure in hindsight). Surprisingly this is not enough to obtain high probability regret bounds (see example in Appendix A), because we cannot constrain the expected time from all states, and thus we used a novel notion of switching policy when reaching "bad" states.

When the transitions are unknown, all these challenges become harder because in order to estimate the expected cost of a policy to reasonable error (even just to determine whether it is proper), one needs very good estimation of the transition function. While in the finite-horizon setting OMD is easily generalized to unknown transitions through optimistic estimates, in adversarial SSP further adaptations are necessary.

Adversarial vs stochastic costs in SSP. In this paper we studied the effects of adversarially changing costs on the general SSP model without any restrictive assumptions, previously studied only under stochastic costs [Tarbouriech *et al.*, 2020; Rosenberg *et al.*, 2020]. The recent lower bound [Chen *et al.*, 2020] shows that adversarial costs in SSP pose significant new challenges, as opposed to finite-horizon where the lower bound for adversarial or stochastic costs is the same.

Both [Tarbouriech *et al.*, 2020; Rosenberg *et al.*, 2020] use optimism w.r.t the costs, which ensures them that the time is also bounded since they use the positive costs assumption, i.e., Assumption 1. While $c_{\min}$ appears in their regret bounds, the latter is able to push it to an additive term (independent of K) and thus keep a regret of $\widetilde{O}(\sqrt{K})$ in the general case (after applying perturbation). Since we are dealing with adversarial costs, we cannot use optimism. Instead we use the OMD method to handle the adversary, and must make sure that we do so while reaching the goal with high probability. For this reason we incorporate explicit constraints on the time, and these cause us to suffer regret that depends on $D/c_{\min}$ instead of D since $T^{\pi^\star}$ is not bounded by D even though $J^{\pi^\star}$ is. This dependence is unavoidable in the adversarial case, and it also requires the additional challenge of estimating D, while optimistic estimates are bounded by D (with high probability).

Technically, our analysis follows the framework of [Rosenberg *et al.*, 2020] since we need to show the goal is reached with high probability. Yet, the mechanisms we introduced are necessary to make this framework useful in the adversarial case, and even then careful analysis is needed. Hopefully, the framework we introduced here will help obtain minimax optimal regret with unknown transitions. In this context, two notable mechanisms are forced exploration and policy switch in "bad" states. Forced exploration is key to handle large variance stochastic policies might have in SSP (without adversarial costs deterministic policies suffice). It ensures that we can determine whether our policies are proper as soon as possible and finish intervals early. While the motivation for switching in "bad" states is clear from known transitions, when dynamics are unknown this switch becomes problematic as we cannot guarantee it actually occurs in "bad" states (our estimate for the time is not even optimistic). More ideas are required in order to bound the excess cost that comes from switching policies in falsely estimated "bad" states (see Appendix I).

Acknowledgments

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

References

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[Azar *et al.*, 2017] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

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

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

[Cai *et al.*, 2020] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[Chen *et al.*, 2020] Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. *arXiv preprint arXiv:2012.04053*, 2020.

[Dann *et al.*, 2017] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

[Efroni *et al.*, 2019] Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pages 12203–12213, 2019.

[Even-Dar *et al.*, 2009] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

[Jaksch *et al.*, 2010] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[Jin *et al.*, 2018] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[Jin *et al.*, 2020] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.

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

[Neu *et al.*, 2010] Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In *Conference on Learning Theory (COLT)*, pages 231–243, 2010.

[Neu *et al.*, 2012] Gergely Neu, András György, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, (AISTATS)*, pages 805–813, 2012.

[Neu *et al.*, 2014] Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online Markov Decision Processes under bandit feedback. *IEEE Trans. Automat. Contr.*, 59(3):676–691, 2014.

[Osband *et al.*, 2016] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386, 2016.

[Rosenberg and Mansour, 2019a] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486, 2019.

[Rosenberg and Mansour, 2019b] Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pages 2209–2218, 2019.

[Rosenberg *et al.*, 2020] Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pages 8210–8219. PMLR, 2020.

[Shani *et al.*, 2020] Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

[Tarbouriech *et al.*, 2020] Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pages 9428–9437. PMLR, 2020.

[Zanette and Brunskill, 2019] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

[Zimin and Neu, 2013] Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2013.