Finding the Stochastic Shortest Path with Low Regret: The Adversarial Cost and Unknown Transition Case

Liyu Chen [1] **Haipeng Luo** [1]

Abstract

We make significant progress toward the stochastic shortest path problem with adversarial costs and unknown transition. Specifically, we develop algorithms that achieve $\tilde{\mathcal{O}}(\sqrt{S^2ADT_\star K})$ regret for the full-information setting and $\tilde{\mathcal{O}}(\sqrt{S^3A^2DT_\star K})$ regret for the bandit feedback setting, where D is the diameter, $T_\star$ is the expected hitting time of the optimal policy, S is the number of states, A is the number of actions, and K is the number of episodes. Our work strictly improves (Rosenberg and Mansour, 2020) in the full information setting, extends (Chen et al., 2020) from known transition to unknown transition, and is also the first to consider the most challenging combination: bandit feedback with adversarial costs and unknown transition. To remedy the gap between our upper bounds and the current best lower bounds constructed via a stochastically oblivious adversary, we also propose algorithms with near-optimal regret for this special case.

1. Introduction

We study the stochastic shortest path (SSP) problem, where a learner aims to find the goal state with minimum total cost. The environment dynamics are modeled as a Markov Decision Process (MDP) with S states, A actions, and a fixed and unknown transition function. The learning proceeds in K episodes, where in each episode, starting from a fixed initial state, the learner sequentially selects an action, incurs a cost, and transits to the next state sampled from the transition function. The episode ends when the learner reaches a fixed goal state. We focus on regret minimization in SSP and measure the performance of the learner by the difference between her total cost over the K episodes and that of the best fixed policy in hindsight.

The special case of SSP where an episode is guaranteed to end within a fixed number of steps is extensively studied in recent years (often known as episodic finite-horizon reinforcement learning or loop-free SSP). The general (and also the more practical) case, on the other hand, has only been recently studied: Tarbouriech et al. (2020a) and Cohen et al. (2020) develop algorithms with sub-linear regret for the case with fixed or i.i.d. costs. Adversarial costs is later studied by Rosenberg and Mansour (2020) in the full-information setting (where the cost is revealed at the end of each episode). The minimax regret for adversarial costs and known transition is then fully characterized in a recent work by Chen et al. (2020), in both the full-information setting and the bandit feedback setting (where only the cost of visited state-action pairs is revealed).

In this work, we further extend our understanding of general SSP with adversarial costs and unknown transition, for both the full-information setting and the bandit setting. More specifically, our results are (see also Table 1):

- (Section 4) In the full-information setting, we develop an algorithm that achieves $\tilde{\mathcal{O}}(\sqrt{S^2ADT_\star K})$ regret with high probability, where D is the diameter of the MDP and $T_\star$ is the expected time for the optimal policy to reach the goal state. This improves over the best existing bound $\tilde{\mathcal{O}}(\frac{1}{c_{\min}}\sqrt{S^2AD^2K})$ or $\tilde{\mathcal{O}}(\sqrt{S^2AT_\star^2}K^{3/4} + D^2\sqrt{K})$ from (Rosenberg and Mansour, 2020), where $c_{\min} \in [0, 1]$ is a global lower bound of the cost for any state-action pair (it can be shown that $T_\star \leq D/c_{\min}$).
- (Section 5) In the bandit setting, we develop another algorithm that achieves $\tilde{\mathcal{O}}(\sqrt{S^3A^2DT_\star K})$ regret with high probability, which, as far as we know, is the first result for this most challenging setting (bandit feedback, adversarial costs, and unknown transition).
- (Section 6) By combining previous results, it can be shown that the lower bound for the full-information and the bandit setting are $\Omega(\sqrt{DT_\star K}+D\sqrt{SAK})$ and $\Omega(\sqrt{SADT_\star K}+D\sqrt{SAK})$ respectively, establishing a gap from our upper bounds. Noting that these lower bounds are constructed with a stochastically oblivious adversary, we propose another algorithm for this special case with near-optimal regret bounds that are only

[1]University of Southern California. Correspondence to: Liyu Chen <liyuc@usc.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Table 1. Summary of our results. Here, D, S, A are the diameter, number of states, and number of actions of the MDP, $T_\star$ is the expected hitting time of the optimal policy, and K is the number of episodes. All algorithms can be implemented efficiently. Our results strictly improve that of (Rosenberg and Mansour, 2020) in the full information setting, and are the first to consider the bandit setting with unknown transition. Lower bounds here are a direct combination of lower bounds for stochastic costs and known transition (Chen et al., 2020) and the lower bound for fixed costs and unknown transition (Cohen et al., 2020).

	Adversarial costs	Stochastic costs (Theorem 3)	Lower bounds
Full information	$\tilde{\mathcal{O}}(\sqrt{S^2ADT_\star K})$ (Theorem 1)	$\tilde{\mathcal{O}}(\sqrt{DT_\star K} + DS\sqrt{AK})$	$\Omega(\sqrt{DT_\star K} + D\sqrt{SAK})$
Bandit feedback	$\tilde{\mathcal{O}}(\sqrt{S^3A^2DT_\star K})$ (Theorem 2)	$\tilde{\mathcal{O}}(\sqrt{SADT_\star K} + DS\sqrt{AK})$	$\Omega(\sqrt{SADT_\star K} + D\sqrt{SAK})$

$\sqrt{S}$ factor larger than the lower bounds, a gap that is still open even for loop-free SSP (Rosenberg and Mansour, 2019; Jin et al., 2020). Note that this setting is slightly different from and harder than existing i.i.d. cost settings; see discussions in "Related work" below.

Technical contributions Our algorithms are largely based on those from the recent work of (Chen et al., 2020) for the known transition setting. However, learning with unknown transition and carefully controlling the transition estimation error requires several new ideas. First, we extend the loop-free reduction of (Chen et al., 2020) to the unknown transition setting (Section 3). Then, combining a Bellman type law of total variance (Azar et al., 2017) and a linear form of the variance of actual costs, we show that, importantly, the bias introduced by transition estimation is well controlled via the so-called skewed occupancy measure proposed by Chen et al. (2020). This leads to our algorithm for the full information setting. For the bandit setting, apart from the techniques above and those from (Chen et al., 2020), we further propose and utilize two optimistic cost estimators inspired by the idea of upper occupancy bounds from Jin et al. (2020) for loop-free SSP.

Finally, for the weaker stochastically oblivious adversaries, we further augment the loop-free reduction to allow the learner to switch to a fast policy at any time step if necessary, which is crucial to ensure the near-optimal regret for our simple optimism-based algorithm.

Related work The SSP problem was studied earlier mostly from the control aspect where the goal is to find the optimal policy efficiently with all parameters known (Bertsekas and Tsitsiklis, 1991; Bertsekas and Yu, 2013). Regret minimization in SSP was first studied in (Tarbouriech et al., 2020a; Cohen et al., 2020), with fixed and known costs and unknown transition. Although their results can be generalized to i.i.d. costs as discussed in (Tarbouriech et al., 2020a, Appendix I.1), this is in fact different from our stochastic cost setting. Indeed, in their setting, the cost of each state-action pair is drawn (independently of other pairs and other episodes) every time it is visited, and is revealed to the learner immediately. On the other hand, in our stochastic setting, the costs of all state-action pairs in each episode are jointly drawn from a fixed distribution (independently of other episodes; but costs of different pairs could be correlated) and fixed throughout the episode, and any information about the costs is only revealed after the episode ends. As argued in (Chen et al., 2020, Section 3.1), our setting is information-theoretically harder as an extra dependence on $T_\star$ is unavoidable here, and thus our bounds for stochastic costs are incomparable to these two works. To distinguish these two different settings, we sometime refer to ours as a setting with a stochastically oblivious adversary.

(Rosenberg and Mansour, 2020) is the first work that studies SSP with adversarial costs with either known or unknown transition, but only in the full-information setting. Later, (Chen et al., 2020) develops efficient and minimax optimal algorithms for both the full-information setting and the bandit feedback setting, but only with known transition. As mentioned, our results significantly improve and extend these two works. One of the key technical contributions of (Chen et al., 2020) is the loop-free reduction, which, as discussed by the authors, is readily applied to the unknown transition case, but leads to suboptimal bounds with unnecessary dependence on other parameters if applied directly. Our algorithms are built on top of an extension of this loop-free reduction, and we overcome the technical difficulty they run into via a more careful analysis showing that the transition estimation error can in fact be well controlled using their idea of skewed occupancy measure.

As mentioned, the special case of loop-free SSP has been extensively studied in recent years, for both fixed or i.i.d. costs (see e.g., (Azar et al., 2017; Jin et al., 2018; Zanette and Brunskill, 2019; Shani et al., 2020a)) and adversarial costs (see e.g., (Neu et al., 2012; Zimin and Neu, 2013; Rosenberg and Mansour, 2019; Jin et al., 2020; Shani et al., 2020b; Cai et al., 2020)). In particular, the idea of upper occupancy bound from (Jin et al., 2020), used to construct an optimistic cost estimator with a confidence set of the transition, is also one key technique we adopt in the bandit setting.

2. Preliminaries

We largely follow the notations of (Chen et al., 2020). An SSP instance consists of an MDP $M = (\mathcal{S}, s_0, g, \mathcal{A}, P)$ and a sequence of K cost functions $\{c_k\}_{k=1}^K$. Here, $\mathcal{S}$ is a finite state space, $s_0 \in \mathcal{S}$ is the initial state, $g \notin \mathcal{S}$ is the goal state, $\mathcal{A} = \{\mathcal{A}_s\}_{s\in\mathcal{S}}$ is a finite action space where $\mathcal{A}_s$ is the available action set at state s. Let $\Gamma = \{(s,a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$ be the set of all valid state-action pairs. The transition function $P : \Gamma \times (\mathcal{S} \cup \{g\}) \to [0,1]$ is such that $P(s'|s,a)$ specifies the probability of transiting to the next state s' after taking action $a \in \mathcal{A}_s$ at state s, and we have $\sum_{s'\in\mathcal{S}\cup\{g\}} P(s'|s,a) = 1$ for each $(s,a) \in \Gamma$. Finally, $c_k : \Gamma \to [0,1]$ is the cost function that specifies the cost for each state-action pair during episode k. We denote by $S = |\mathcal{S}|$ and $A = (\sum_{s\in\mathcal{S}} |\mathcal{A}_s|)/S$ the total number of states and the average number of available actions respectively.

The learner interacts with the MDP through K episodes, not knowing the transition function P nor the cost functions $\{c_k\}_{k=1}^K$ ahead of time. In each episode $k = 1, \dots, K$, the adversary first decides the cost function c_k, which, for the majority of this work, can depend on the learner's algorithm and the randomness before episode k in an arbitrary way (known as an adaptive adversary). Only in Section 6, we switch to a weaker stochastically oblivious adversary who draws c_k independently from a fixed but unknown distribution. In any case, without knowing c_k, the learner decides which action to take in each step of the episode, starting from the initial state s_0 and ending at the goal state g. More precisely, in each step i of episode k, the learner observes its current state s_k^i (with $s_k^1 = s_0$). If $s_k^i \neq g$, the learner selects an action $a_k^i \in \mathcal{A}_{s_k^i}$ and transits to the next state s_k^{i+1} sampled from $P(\cdot|s_k^i, a_k^i)$; otherwise, the episode ends, and we let I_k be the number of steps in this episode such that $s_k^{I_k+1} = g$.

After each episode k ends, the learner receives some feedback on the cost function c_k. In the *full-information* setting, the learner observes the entire c_k, while in the more challenging *bandit feedback* setting, the learner only observes the costs of the visited state-action pairs, that is, $c_k(s_k^i, a_k^i)$ for $i = 1, \dots, I_k$.

Important concepts We introduce several necessary concepts before discussing the goal of the learner. A stationary policy is a mapping π such that $\pi(a|s)$ specifies the probability of taking action $a \in \mathcal{A}_s$ in state s. It is deterministic if for all s, $\pi(a|s) = 1$ holds for some action a (in which case we write $\pi(s) = a$). A policy is *proper* if executing it in the MDP starting from any state ensures that the goal state is reached within a finite number of steps with probability 1 (and improper otherwise). We denote by Π_{proper} the set of all deterministic and proper policies, and make the basic assumption $\Pi_{\text{proper}} \neq \emptyset$ following (Rosenberg and Mansour, 2020; Chen et al., 2020).

We denote by $T^\pi(s)$ the expected hitting time it takes for a stationary policy π to reach g starting from state s. The *fast policy* π^f is a deterministic policy that achieves the minimum expected hitting time starting from any state (among all stationary policies). The diameter of the MDP is defined as $D = \max_{s\in\mathcal{S}} \min_{\pi\in\Pi_{\text{proper}}} T^\pi(s) = \max_{s\in\mathcal{S}} T^{\pi^f}(s)$, which is the "largest shortest distance" between any state and the goal state.

Given a transition function P, a cost function c, and a proper policy π, we define the cost-to-go function $J^{P,\pi,c} : \mathcal{S} \to [0,\infty)$ such that $J^{P,\pi,c}(s) = \mathbb{E}\left[\sum_{i=1}^I c(s^i,a^i) \middle| P,\pi,s^1 = s\right]$, where the expectation is over the randomness of the action a^i drawn from $\pi(\cdot|s^i)$, the state s^{i+1} drawn from $P(\cdot|s^i,a^i)$, and the number of steps I before reaching g. Similarly, we also define the state-action value function $Q^{P,\pi,c} : \Gamma \to [0,\infty)$ such that $Q^{P,\pi,c}(s,a) = \mathbb{E}\left[\sum_{i=1}^I c(s^i,a^i) \middle| P,\pi,s^1 = s, a^1 = a\right]$. We use $J_k^{P,\pi}$ and $Q_k^{P,\pi}$ to denote the cost-to-go and state-action function with respect to the cost c_k. When there is no confusion, we also ignore the dependency on the transition function (especially when P is the true transition function of the MDP) and write $J^{P,\pi,c}$ as $J^{\pi,c}$, $J_k^{P,\pi}$ as J_k^π, $Q^{P,\pi,c}$ as $Q^{\pi,c}$, and $Q_k^{P,\pi}$ as Q_k^π.

Learning objective The learner's goal is to minimize her *regret*, defined as the difference between her total cost and the total expected cost of the best deterministic proper policy in hindsight:

$$R_K = \sum_{k=1}^K \sum_{i=1}^{I_k} c_k(s_k^i, a_k^i) - \sum_{k=1}^K J_k^{\pi^\star}(s_0),$$

where $\pi^\star \in \operatorname{argmin}_{\pi\in\Pi_{\text{proper}}} \sum_{k=1}^K J_k^\pi(s_0)$ is the optimal stationary and proper policy, which is referred to as optimal policy in the rest of the paper. By the Markov property, $\pi^\star$ is in fact also the optimal policy starting from any other state, that is, $\pi^\star \in \operatorname{argmin}_{\pi\in\Pi_{\text{proper}}} \sum_{k=1}^K J_k^\pi(s)$ for any $s \in \mathcal{S}$. As in (Chen et al., 2020), the following two quantities related to $\pi^\star$ play an important role: its expected hitting time starting from the initial state $T_\star = T^{\pi^\star}(s_0)$ and its largest expected hitting time starting from any state $T_{\max} = \max_s T^{\pi^\star}(s)$. Chen et al. (2020) show that $T_{\max} \le \frac{D}{c_{\min}}$ where $c_{\min} = \min_k \min_{(s,a)} c_k(s,a)$ is the minimum possible cost. For ease of presentation, we assume $D \le T_\star$ to simplify our bounds.

Knowledge on key parameters Our algorithms require the knowledge of $T_\star$ and $T_{\max}$, similarly to most algorithms of (Chen et al., 2020). This requirement is seemingly restrictive, especially when against an adaptive adversary, in which

case $T_\star$ and $T_{\max}$ depend on the behavior of both the algorithm itself and the adversary. However, we argue that our results are still meaningful: First, for an oblivious adversary, $T_\star$ and $T_{\max}$ are fixed unknown quantity independent of the learner's behavior. Many works in online learning indeed start with assuming knowledge on such quantities to get a better understanding of the problem (and to tune these hyperparameters empirically), before one can eventually develop a fully parameter-free algorithm. Thus, as the first step, we believe that our work is still valuable. Second, in the lower bound construction (Chen et al., 2020), $T_\star$ is also known to the learner, meaning that knowing $T_\star$ does not make the problem any easier information-theoretically. Finally, to emphasize the difficulty of removing this requirement, we note that this is still open even with known transition when considering high-probability bounds. Chen et al. (2020) were able to resolve this for expected regret bounds, but extending their techniques to high-probability bounds is related to deriving a high-probability bound for the so-called multi-scale expert problem, which is also still open (Chen et al., 2021, Appendix A).

On the other hand, we also emphasize that our main improvement compared to (Rosenberg and Mansour, 2020) is not due to the knowledge of these parameters. Indeed, under the same setup where these parameters are unknown, we can still run our algorithms by replacing $T_\star$ with its upper bound $D/c_{\min}$ and $T_{\max}$ with some lower order term $o(K)$, and this still leads to better results compared to (Rosenberg and Mansour, 2020). Details are deferred to Appendix F.

Finally, for simplicity, we also assume that D is known, but our results can be extended even if D is unknown; see Appendix E.

Occupancy measure Occupancy measure plays a key role in solving SSP with adversarial costs, in both the loop-free case (Neu et al., 2012; Zimin and Neu, 2013; Rosenberg and Mansour, 2019; Jin et al., 2020) and the general case (Rosenberg and Mansour, 2020; Chen et al., 2020). A proper stationary policy π and a transition function P induce an occupancy measure $q_{P,\pi} \in \mathbb{R}_{\geq 0}^{\Gamma\times(\mathcal{S}\cup\{g\})}$ such that $q_{P,\pi}(s,a,s')$ is the expected number of visits to state-action-afterstate triplet (s,a,s') when executing π in an MDP with transition P, that is: $q_{P,\pi}(s,a,s') = \mathbb{E}\left[\sum_{i=1}^{I}\mathbb{I}\{s^i=s,a^i=a,s^{i+1}=s'\}\middle| P,\pi,s^1=s_0\right]$. When P is clear from the context (which is usually the case if it is the true transition), we omit the P dependence and only write q_π. We also let $q_\pi(s,a)=\sum_{s'}q_\pi(s,a,s')$ be the expected number of visits to state-action pair (s,a) and $q_\pi(s)=\sum_{a\in\mathcal{A}_s}q_\pi(s,a)$ be the expected number of visits to state s when executing π. Note that, given a function $q:\Gamma\times(\mathcal{S}\cup\{g\})\to[0,\infty)$, if it corresponds to an occupancy measure, then the corresponding policy π_q can be obtained via $\pi_q(a|s)\propto q(s,a)$, and the corresponding transition function can be obtained via $P_q(s'|s,a)\propto q(s,a,s')$. Also note that $T^\pi(s_0)=\sum_{(s,a)}q_\pi(s,a)=\sum_{s\in\mathcal{S}}q_\pi(s)$.

Occupancy measures allow one to turn the problem into a form of online linear optimization where Online Mirror Descent is a standard tool. Indeed, we have $J_k^\pi(s_0)=\sum_{(s,a)\in\Gamma}q_\pi(s,a)c_k(s,a)=\langle q_\pi,c_k\rangle$, and if the learner executes a stationary proper policy π_k in episode k, then the expected regret can be written as $\mathbb{E}[R_K]=\mathbb{E}\left[\sum_{k=1}^K J_k^{\pi_k}(s_0)-J_k^{\pi^\star}(s_0)\right]=\mathbb{E}\left[\sum_{k=1}^K\langle q_{\pi_k}-q_{\pi^\star},c_k\rangle\right]$, exactly in the form of online linear optimization.

Other notations We let $N_k(s,a)$ denote the (random) number of visits of the learner to (s,a) during episode k, so that the regret can be re-written as $R_K=\sum_{k=1}^K\langle N_k-q_{\pi^\star},c_k\rangle$. Denote by $\mathbb{I}_k(s,a)$ the indicator of whether $c_k(s,a)$ is revealed to the learner in episode k, so that in the full information setting $\mathbb{I}_k(s,a)=1$ always holds, and in the bandit feedback setting $\mathbb{I}_k(s,a)$ is also the indicator of whether (s,a) is ever visited by the learner. Throughout the paper, we use the notation $\langle f,g\rangle$ as a shorthand for $\sum_{s\in\mathcal{S}}f(s)g(s)$, $\sum_{(s,a)}f(s,a)g(s,a)$, $\sum_{h=1}^H\sum_{(s,a)}f(s,a,h)g(s,a,h)$, or $\sum_{(s,a)}\sum_{s'}\sum_{h=1}^H f(s,a,s',h)g(s,a,s',h)$ when f and g are functions in $\mathbb{R}^{\mathcal{S}}$, $\mathbb{R}^{\Gamma}$, $\mathbb{R}^{\Gamma\times[H]}$ or $\mathbb{R}^{\Gamma\times(\mathcal{S}\cup\{g\})\times[H]}$ (for some H) respectively. Denote $\odot$ as the Hadamard product of tensors, so that $(u\odot v)_i=u_i\cdot v_i$ (e.g. the feedback on cost for both settings is thus $c_k\odot\mathbb{I}_k$). Let $\mathcal{F}_k$ denote the σ-algebra of events up to the beginning of episode k, and $\mathbb{E}_k$ be a shorthand of $\mathbb{E}[\cdot|\mathcal{F}_k]$. To be specific, c_k and the learner's policy in episode k is already determined at the beginning of episode k, and the randomness in $\mathbb{E}[\cdot|\mathcal{F}_k]$ is w.r.t the learner's actual trajectory in episode k. For a convex function ψ, the Bregman divergence between u and v is defined as: $D_\psi(u,v)=\psi(u)-\psi(v)-\langle\nabla\psi(v),u-v\rangle$. For an integer n, $[n]$ denotes the set $\{1,\dots,n\}$.

3. Loop-free Reduction with Unknown Transition

When the transition is known, (Chen et al., 2020) show that it is possible to approximate a general SSP by a loop-free SSP in a way such that any policy in the loop-free instance can be transformed to a policy in the original instance with only $\tilde{O}(1)$ additional overhead in the final regret. More importantly, this loop-free reduction provides simpler forms for some variance-related quantities, which is the key in achieving high probability bounds and dealing with bandit feedback. As the first step, we extend this loop-free reduction to the unknown transition setting, and show that the

additional regret is also very small.

Loop-free instance The construction of the converted loop-free SSP instance is essentially the same as that in (Chen et al., 2020): for the first H_1 steps, we duplicate each state by attaching it with a time step h, then we connect all states to some virtual *fast* state that lasts for another H_2 steps. We show the definition below for completeness (with slight modifications for our purposes), and then discuss what the necessary changes are to complete the reduction using this loop-free SSP when the transition is unknown.

Definition 1. *(Chen et al., 2020, Definition 5) For an SSP instance $M = (\mathcal{S}, s_0, g, \mathcal{A}, P)$ with cost functions $c_{1:K}$, we define, for horizon parameters $H_1, H_2 \in \mathbb{N}$, another loop-free SSP instance $\widetilde{M} = (\widetilde{\mathcal{S}}, \widetilde{s}_0, g, \widetilde{\mathcal{A}}, \widetilde{P})$ with cost function $\widetilde{c}_{1:K}$ as follows:*

- $\widetilde{\mathcal{S}} = \mathcal{X} \times [H]$ *where* $\mathcal{X} = \mathcal{S} \cup \{s_f\}$, s_f *is an artificially added "fast" state, and* $H = H_1 + H_2$.
- $\widetilde{s}_0 = (s_0, 1)$, *and the goal state g remains the same.*
- $\widetilde{\mathcal{A}} = \mathcal{A} \cup \{a_f\}$, *where a_f is an artificially added action. The available action set at (s,h) is $\mathcal{A}_s$ for all $s \neq s_f$ and $h \in [H]$, and the only available action at (s_f, h) for $h \in [H]$ is a_f.*
- *Transition from (s,h) to (s',h') is only possible when $h' = h+1$: for the first H_1 layers, the transition follows the original MDP in the sense that $\widetilde{P}((s', h+1)|(s,h),a) = P(s'|s,a)$ and $\widetilde{P}(g|(s,h),a) = P(g|s,a)$ for all $h < H_1$ and $(s,a) \in \Gamma$; from layer H_1 to layer H, all states transit to the fast state: $\widetilde{P}((s_f, h+1)|(s,h),a) = 1$ for all $H_1 \le h < H$ and $(s,a) \in \widetilde{\Gamma} \triangleq \Gamma \cup \{(s_f, a_f)\}$; finally, the last layer transits to the goal state always: $\widetilde{P}(g|(s,H),a) = 1$ for all $(s,a) \in \widetilde{\Gamma}$. For notational convenience, we also write $\widetilde{P}((s', h+1)|(s,h),a)$ as $P(s'|s,a,h)$, and $\widetilde{P}(g|(s,h),a)$ as $P(g|s,a,h)$.*
- *Cost function is such that $\widetilde{c}_k((s,h),a) = c_k(s,a)$ and $\widetilde{c}_k((s_f,h),a_f) = 1$ for all $(s,a) \in \Gamma$ and $h \in [H]$. For notational convenience, we also write $\widetilde{c}_k((s,h),a)$ as $c_k(s,a,h)$.*

For notations related to the loop-free version, we often use a tilde symbol to distinguish them from the original counterparts (such as $\widetilde{M}$ and $\widetilde{\mathcal{S}}$), and for a function $\widetilde{f}((s,h),a)$ or $\widetilde{f}((s,h),a,(s',h+1))$ that takes a state-action pair or a state-action-afterstate triplet in $\widetilde{M}$ as input, we often simplify it as $f(s,a,h)$ (such as c_k) or $f(s,a,s',h)$ (such as q and P). For such a function, we will also use the notation $\vec{h} \circ f \in \mathbb{R}^{\widetilde{\Gamma} \times [H]}$ (or $\vec{h} \circ f \in \mathbb{R}^{\widetilde{\Gamma} \times \mathcal{X} \times [H]}$) such that $(\vec{h} \circ f)(s,a,h) = h \cdot f(s,a,h)$ (or $(\vec{h} \circ f)(s,a,s',h) = h \cdot f(s,a,s',h)$). Similarly, for a function $f \in \mathbb{R}^{\widetilde{\Gamma}}$, we use the same notation $\vec{h} \circ f \in \mathbb{R}^{\widetilde{\Gamma} \times [H]}$ such that $(\vec{h} \circ f)(s,a,h) = h \cdot f(s,a)$. Finally, for a occupancy measure $q \in [0,1]^{\widetilde{\Gamma} \times \mathcal{X} \times [H]}$ of $\widetilde{M}$, we write $q(s,a,h) = \sum_{s' \in \mathcal{X}} q(s,a,s',h)$ and $q(s,a) = \sum_{h=1}^{H} q(s,a,h)$.

Algorithm 1 RUN$(\widetilde{\pi}, \mathcal{B})$

Input: a policy $\widetilde{\pi}$ for $\widetilde{M}$ and a Bernstein-SSP instance $\mathcal{B}$.
Initialize: $s^1 = s_0$ and $h = 1$.
while $s^h \neq g$ *and* $h \le H_1$ **do**
  Draw action $a^h \sim \widetilde{\pi}(\cdot|(s^h,h))$. If $a^h = a_f$, break.[2]
  Play a^h, observe s^{h+1}, increment $h \leftarrow h+1$.
if $s^h \neq g$ **then**
  Invoke $\mathcal{B}$ with a new episode starting with state s^h, follow its decision until reaching g, and always feed it cost 1 for all state-action pairs.
Return: trajectory $\{s^1, a^1, s^2, a^2, \ldots, a^{h-1}, s^h\}$.

The reduction Now, we are ready to describe the reduction, that is, how one can convert an algorithm for $\widetilde{M}$ to an algorithm for M. Specifically, given policies $\widetilde{\pi}_1, \ldots, \widetilde{\pi}_K$ for $\widetilde{M}$, we define a sequence of *non-stationary* policies $\sigma(\widetilde{\pi}_1), \ldots, \sigma(\widetilde{\pi}_K)$ for M as follows. For each episode k, during the first $h \le H_1$ steps, we follow $\widetilde{\pi}(\cdot|(s,h))$ when at state s. After the first H_1 steps (if not reaching g yet), Chen et al. (2020) simply execute the fast policy π^f, available since the transition is known, to reach the goal state as soon as possible. In our case with unknown transition, we propose to approximate the fast policy's behavior by running the Bernstein-base algorithm of (Cohen et al., 2020) designed for the fixed cost setting and pretending that all costs are 1. More precisely, we initialize a copy of their algorithm (that we call Bernstein-SSP) for M (not $\widetilde{M}$) ahead of time, and whenever the learner does not reach the goal within H_1 steps in some episode, we invoke Bernstein-SSP as if this is a new episode for this algorithm, follow its decisions until reaching g, and always feed it a cost of 1 for all state-action pairs.[1] We describe this converted policy in the procedure RUN (Algorithm 1).

The rationale of using Bernstein-SSP in this way is simply because when the costs are all 1, the fast policy is exactly the optimal policy, and since Bernstein-SSP guarantees low regret against the optimal policy in the fixed cost setting, it behaves similarly to the fast policy in the long run in our reduction.

[1] This means that Bernstein-SSP is dealing with different initial states for different episodes, which is not exactly the same setting as the original work of (Cohen et al., 2020) but makes no real difference in their regret guarantee as pointed out in (Tarbouriech et al., 2020b, Appendix C).

[2] This if statement is only necessary for Section 6.

This allows us to mostly preserve the properties of the reduction of (Chen et al., 2020). To state these properties, we need the following notations. When executing $\sigma(\widetilde{\pi}_k)$ in M for episode k, we adopt the notation $\widetilde{N}_k$ and let $\widetilde{N}_k(s,a,h)$ be 1 if (s,a) is visited at time step $h \le H_1$, or 0 otherwise; and $\widetilde{N}_k(s_f,a_f,h)$ be 1 if $H_1 < h \le H$ and the goal state g is not reached within H_1 steps, or 0 otherwise. Clearly, $\widetilde{N}_k$ for $\widetilde{M}$ is the analogue of N_k for M, and $\widetilde{N}_k(s,a,h)$ follows the same distribution as the number of visits to state-action pair $((s,h),a)$ when executing $\widetilde{\pi}$ in $\widetilde{M}$. In addition, define a deterministic policy $\widetilde{\pi}^\star$ for $\widetilde{M}$ that mimics the behavior of $\pi^\star$ in the sense that $\widetilde{\pi}^\star(s,h) = \pi^\star(s)$ for $s \in \mathcal{S}$ and $h \le H_1$ (for larger h, s has to be s_f and the only available action is a_f). With these notations, the next lemma shows that the reduction introduces little regret overhead when the horizon parameters H_1 and H_2 are set appropriately.

Lemma 1. *Suppose $H_1 \ge 8T_{\max}\ln K$, $H_2 = \lceil 2D \rceil$, $K \ge D$, and $\widetilde{\pi}_1,\dots,\widetilde{\pi}_K$ are policies for $\widetilde{M}$. Then with probability at least $1-\delta$, the regret of executing $\sigma(\widetilde{\pi}_1),\dots,\sigma(\widetilde{\pi}_K)$ in M satisfies:*

$$R_K \le \sum_{k=1}^{K}\langle \widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k\rangle + \widetilde{\mathcal{O}}\left(D^{3/2}S^2A\left(\ln\tfrac{1}{\delta}\right)^2\right).$$

Reduction alone is not enough While all of our algorithms make use of this reduction, it is worth emphasizing that the reduction alone is not enough. Put differently, applying existing loop-free algorithms to $\widetilde{M}$ directly only leads to sub-optimal bounds with dependence on $H = \widetilde{\mathcal{O}}(T_{\max})$. This is true already in the known transition case (Chen et al., 2020), and is even more so in our unknown transition case where one needs to estimate the transition. On the other hand, what the reduction accomplishes is to make sure that some important variance-related quantities take a simple form that is linear in both the occupancy measure and the cost function. For example, we will make use of the following important lemma, which is essentially taken from (Chen et al., 2020) but includes an extra intermediate result (the first inequality) important for Section 6. In Section 4, we will see another important property of the reduction.

Lemma 2. *Consider executing a policy $\sigma(\widetilde{\pi})$ in episode k. Then $\mathbb{E}_k[\langle \widetilde{N}_k, c_k\rangle^2] \le 2\langle q_{\widetilde{\pi}}, c_k \odot Q_k^{\widetilde{\pi}}\rangle \le 2\langle q_{\widetilde{\pi}}, J_k^{\widetilde{\pi}}\rangle = 2\langle q_{\widetilde{\pi}}, \vec{h}\circ c_k\rangle$.*

4. Adversarial Costs with Full Information

In the full-information setting, the algorithm of (Chen et al., 2020) maintains a sequence of occupancy measures $q_1,\dots,q_K$ for $\widetilde{M}$, obtained via Online Mirror Descent (OMD) over a sophisticated *skewed occupancy measure* space. In their analysis, the regret for $\widetilde{M}$ from Lemma 1 is decomposed as $\sum_{k=1}^K\langle \widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k\rangle = \sum_{k=1}^K\langle \widetilde{N}_k - q_k, c_k\rangle + \sum_{k=1}^K\langle q_k - q_{\widetilde{\pi}^\star}, c_k\rangle$, where the first term is the sum of a martingale difference sequence whose variance can be bounded using Lemma 2, and the second term is controlled by the standard OMD analysis. Importantly, due to the skewed occupancy measure, the bound for the second term contains a negative bias in terms of $-\langle q_k, \vec{h}\circ c_k\rangle$, which can then cancel the variance from the first term in light of Lemma 2.

When the transition is unknown, we follow the ideas of the SSP-O-REPS algorithm (Rosenberg and Mansour, 2020) and maintain a confidence set of plausible transition functions, which contains the true transition P with high probability. This step is conducted via the procedure TransEst (Algorithm 4), which takes a trajectory returned by RUN (along with other statistics) and outputs an updated confidence set based on standard concentration inequalities. We defer the details to Section B.1.

With a confidence set $\mathcal{P}$ at hand, we define the set of plausible occupancy measures $\widetilde{\Delta}(T,\mathcal{P})$ as follows, which is parameterized by $\mathcal{P}$ and a size parameter T (recall the shorthand $q(s,a,h) = \sum_{s'} q(s,a,s',h)$):

$$\Bigg\{ q \in [0,1]^{\widetilde{\Gamma}\times\mathcal{X}\times[H]} : \sum_{h=1}^{H}\sum_{(s,a)\in\widetilde{\Gamma}} q(s,a,h) \le T;$$
$$\sum_{a\in\widetilde{\mathcal{A}}_{(s,h)}} q(s,a,h) = \sum_{(s',a')\in\widetilde{\Gamma}} q(s',a',s,h-1),\ \forall h>1;$$
$$\sum_{a\in\widetilde{\mathcal{A}}_{(s,1)}} q(s,a,1) = \mathbb{I}\{s=s_0\},\ \forall s\in\mathcal{X};\ P_q \in \mathcal{P}\Bigg\}. \quad (1)$$

When $\mathcal{P} = \{P\}$, this is equivalent to the set used by (Chen et al., 2020), where the first inequality constraint makes sure that the induced policy reaches the goal within T steps in expectation, the equality constraints make sure that q is a valid occupancy measure, and the last constraint $P_q = P$ makes sure that the induced transition is consistent with the true one. We naturally generalize the set to the unknown transition case by enforcing the induced transition P_q to be within a given confidence set.

Then, in each episode k, with $\mathcal{P}_k$ being the current confidence set, we define the skew occupancy measure space for some parameter λ as

$$\Omega_k = \left\{\phi = q + \lambda\vec{h}\circ q : q \in \widetilde{\Delta}(T,\mathcal{P}_k)\right\}. \quad (2)$$

which is again a direct generalization of (Chen et al., 2020) from $\{P\}$ to $\mathcal{P}_k$. Our algorithm then maintains a sequence of skewed occupancy measures $\phi_1,\dots,\phi_K$ based on the standard OMD framework:

$$\phi_{k+1} = \operatorname*{argmin}_{\phi\in\Omega_{k+1}} \langle \phi, c_k\rangle + D_\psi(\phi,\phi_k)$$

Algorithm 2 SSP-O-REPS with Loop-free Reduction and Skewed Occupancy Measure

Input: Upper bound on expected hitting time T, horizon parameter H_1, confidence level δ

Parameters: $\eta = \min\left\{\frac{1}{8}, \sqrt{\frac{T}{DK}}\right\}, \lambda = 4\sqrt{\frac{S^2A}{DTK}}, H_2 = \lceil 2D \rceil, H = H_1 + H_2$

Define: regularizer

$$\psi(\phi) = \frac{1}{\eta}\sum_{h=1}^{H}\sum_{(s,a)\in\widetilde{\Gamma}}\sum_{s'\in\mathcal{X}\cup\{g\}}\phi(s,a,s',h)\ln\phi(s,a,s',h)$$

Initialize: $\mathbf{N}_1(s,a) = \mathbf{M}_1(s,a,s') = 0$ for all $(s,a,s') \in \Gamma \times (\mathcal{S}\cup\{g\})$, a Bernstein-SSP instance $\mathcal{B}$, $\mathcal{P}_1$ is the set of all possible transition functions, $\phi_1 = \operatorname{argmin}_{\phi\in\Omega_1}\psi(\phi)$ (where Ω_k is defined in Eq. (2)).

for $k = 1, \ldots, K$ **do**

Extract $\widehat{q}_k$ from $\phi_k = \widehat{q}_k + \lambda\vec{h}\circ\widehat{q}_k$ and let $\widetilde{\pi}_k = \widetilde{\pi}_{\widehat{q}_k}$.

Execute policy $\widetilde{\pi}_k$: $\tau_k = \text{RUN}(\widetilde{\pi}_k, \mathcal{B})$, receive c_k.

Update $\mathcal{P}_{k+1} = \text{TransEst}(\mathbf{N}, \mathbf{M}, \delta, H_1, H_2, \tau_k)$.

Update $\phi_{k+1} = \operatorname{argmin}_{\phi\in\Omega_{k+1}}\langle\phi, c_k\rangle + D_\psi(\phi, \phi_k)$.

where ψ is the negative entropy regularizer. In each episode, extracting $\widehat{q}_k$ from $\phi_k = \widehat{q}_k + \lambda\vec{h}\circ\widehat{q}_k$, we obtain a policy $\widetilde{\pi}_{\widehat{q}_k}$ for $\widetilde{M}$, and then execute it via the RUN procedure (Algorithm 1). The complete pseudocode of our algorithm is presented in Algorithm 2, which can be efficiently implemented (see related discussion in (Rosenberg and Mansour, 2020)).

Analysis Let q_k be the occupancy measure with respect to the policy $\widetilde{\pi}_k$ and the true transition P. We can then decompose the regret from Lemma 1 as $\sum_{k=1}^K\langle\widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k\rangle = \sum_{k=1}^K\langle\widetilde{N}_k - q_k, c_k\rangle + \sum_{k=1}^K\langle\widehat{q}_k - q_{\widetilde{\pi}^\star}, c_k\rangle + \sum_{k=1}^K\langle q_k - \widehat{q}_k, c_k\rangle$, where the last term measures the difference between q_k and $\widehat{q}_k$ due to the transition estimation error and is the only extra term compared to the known transition case discussed at the beginning of this section. One of our key technical contributions is to prove that, thanks to the structure of the loop-free instance $\widetilde{M}$, this term is in fact also bounded by the variance term seen earlier in Lemma 2:

$$\sum_{k=1}^K\langle q_k - \widehat{q}_k, c_k\rangle = \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^K\mathbb{E}_k[\langle\widetilde{N}_k, c_k\rangle^2]}\right). \tag{3}$$

See Lemma 9 for the complete statement, whose proof makes use of a Bellman type law of total variance for Bernstein-based confidence sets (Lemma 4).

With this result and Lemma 2, one can see that just like the first term $\sum_{k=1}^K\langle\widetilde{N}_k - q_k, c_k\rangle$, the extra transition error term can also be handled by the negative bias introduced by the skewed occupancy measure space as discussed earlier. This leads to our final regret guarantee of Algorithm 2.

Theorem 1. *If $T \geq T_\star + 1, H_1 \geq 8T_{\max}\ln K$, and $K \geq 16S^2AH^2$, then with probability at least $1-6\delta$, Algorithm 2 ensures $R_K = \tilde{\mathcal{O}}(\sqrt{S^2ADTK} + H^3S^2A)$.*

We emphasize that our way to handle the transition estimation error $\sum_{k=1}^K\langle q_k - \widehat{q}_k, c_k\rangle$ is novel. Specifically, all previous works directly upper bound this error using the definition of confidence interval, which in our case introduces an undesirable $T_{\max}$ dependency. Instead, we derive a specific upper bound (Eq. (3)) of the transition estimation error that can be cancelled out by the negative term introduced by the skewed occupancy measure. This technique is especially useful in obtaining data-dependent bound in the unknown transition case, since it replaces the error by a term related to the optimal policy, which is hard to achieve if we directly upper bound the error.

Besides this new way to handle the transition estimation error, another source of improvement compared to the analysis of (Rosenberg and Mansour, 2020) is to make use of the fact $\sum_{k=1}^K\langle q_{\pi^\star}, c_k\rangle \leq DK$ in the OMD analysis. Again, we emphasize that even without the knowledge of $T_\star$ or $T_{\max}$, our analysis leads to better bounds compared to theirs; see Appendix F.

Since Chen et al. (2020) show a lower bound of $\Omega(\sqrt{DT_\star K})$ for stochastic costs and known transition, and Cohen et al. (2020) show a lower bound of $\Omega(D\sqrt{SAK})$ for fixed costs and unknown transition, we know that in our setting, $\Omega(\sqrt{DT_\star K} + D\sqrt{SAK})$ is a lower bound, which shows a gap of $\sqrt{ST_\star/D}$ from our upper bound. Closing the $\sqrt{S}$ gap is still open even for the loop-free case (Rosenberg and Mansour, 2019; Jin et al., 2020). On the other hand, closing the $\sqrt{T_\star/D}$ gap also seems rather challenging for adversarial costs, but is indeed possible for stochastic costs as we show in Section 6 (note that the lower bound is indeed constructed with stochastic costs).

5. Adversarial Costs with Bandit Feedback

We now consider the bandit feedback setting which, even when the transition is known, is quite challenging already and requires several new techniques as shown by Chen et al. (2020). Our algorithm is built on top of their Log-barrier Policy Search algorithm with the transition estimation component integrated in a similar way as in Section 4. We defer most details to Appendix C but only highlight two important new ingredients below.

A standard technique to deal with adversarial costs and bandit feedback in online learning is to feed the OMD algorithm with importance-weighted cost estimators (since c_k is now only partially observed). Specifically, the Log-barrier Policy Search algorithm of Chen et al. (2020) feeds OMD

with cost $\widehat{c}_k - \gamma \widehat{b}_k$ (for some parameter γ), where $\widehat{c}_k(s,a) = \frac{\widetilde{N}_k(s,a)}{q_k(s,a)} c_k(s,a)$ and $\widehat{b}_k(s,a) = \frac{\sum_h h \cdot q_k(s,a,h)\widehat{c}_k(s,a)}{q_k(s,a)}$ are two importance-weighted estimators. Here, $q_k(s,a)$ is defined as $\sum_{h=1}^H q_k(s,a,h)$ and $\widetilde{N}_k$ is defined above Lemma 1 with mean $q_k(s,a)$, so that $\widehat{c}_k$ is an unbiased estimator of c_k. The reason of having $\widehat{b}_k$, on the other hand, is relatively technical, but it eventually serves as a way of reducing variance by introducing a negative bias. The immediate challenge to generalize these estimators to the unknown transition setting is that q_k, the occupancy measure with respect to the policy $\widetilde{\pi}_k$ for episode k and the true transition P, is now unknown.

To address this issue for $\widehat{c}_k$, we follow the idea of (Jin et al., 2020) and construct the following *optimistic* biased estimator: $\widehat{c}_k(s,a) = \frac{\widetilde{N}_k(s,a)}{u_k(s,a)} c_k(s,a)$ where $u_k(s,a) = \max_{\widehat{P} \in \mathcal{P}_k} q_{\widehat{P},\widetilde{\pi}_k}(s,a)$, called the *upper occupancy bound*, is the largest possible expected number of visits to (s,a) of policy $\widetilde{\pi}_k$ under a plausible transition from the confidence set $\mathcal{P}_k$. Clearly, $q_k(s,a) \le u_k(s,a)$ holds (with high probability), making $\widehat{c}_k(s,a)$ an optimistic underestimator which is important in reducing variance as shown in Jin et al. (2020). Note that u_k can be efficiently computed since it boils down to solving a linear program.[3]

On the other hand, $\widehat{b}_k$ does not appear before in the loop-free setting of Jin et al. (2020) and requires some more careful thinking. Other than replacing q_k in the denominator with u_k, we also need to deal with $q_k(s,a,h)$ in the numerator. It turns out that the right generalization is to let

$$\widehat{b}_k(s,a) = \frac{\max_{\widehat{P} \in \mathcal{P}_k} \sum_h h \cdot q_{\widehat{P},\widetilde{\pi}_k}(s,a,h)\widehat{c}_k(s,a)}{u_k(s,a)},$$

so that $\sum_h h \cdot q_k(s,a,h)\widehat{c}_k(s,a) \le u_k(s,a)\widehat{b}_k(s,a)$ holds (with high probability), which in turn makes sure that the bias introduced by $\widehat{b}_k$ is large enough to cancel some important variance term, as shown in Lemma 16. Similarly, $\widehat{b}_k$ can also be computed efficiently (*c.f.* Footnote 3).

Our final algorithm is summarized in Algorithm 5 of Appendix C. Noting that the bias introduced by the upper occupancy bounds is eventually also related to the transition estimation error that has been analyzed in Lemma 9, we are able to prove the following regret guarantee.

Theorem 2. *If $T \ge T_\star + 1, H_1 \ge 8T_{\max} \ln K$, and K is large enough ($K \gtrsim S^3A^2H^2$), then with probability at least $1 - 30\delta$, Algorithm 5 ensures $R_K = \widetilde{\mathcal{O}}\left(\sqrt{S^3A^2DTK} + H^3S^3A^2\right)$.*

Compared to the full-information setting, here we pay an extra $\sqrt{SA}$ factor in the regret bound, a price that does not exist in the loop-free setting (Rosenberg and Mansour, 2019; Jin et al., 2020). This comes from a technical lemma on bounding $\sum_{k=1}^K \langle u_k - q_k, c_k \rangle$ in terms of $\sum_{k=1}^K \langle q_k, \vec{h} \circ c_k \rangle$ so that it can be canceled by the skew occupancy measure; see Lemma 11. Removing this extra factor is an important future direction. On the other hand, by combing the lower bounds of (Chen et al., 2020) and (Cohen et al., 2020) again, we have that $\Omega(\sqrt{SADT_\star K} + D\sqrt{SAK})$ is the best existing lower bound for this setting.

6. Stochastically Oblivious Adversary

Given the gap between our upper and lower bounds, in this section, we consider a weaker stochastically oblivious adversary and develop a simple algorithm with regret bounds only $\sqrt{S}$ times larger than the aforementioned lower bounds. Specifically, in this setting the adversary generates ahead of the time the cost functions $c_1, \dots, c_K$ as i.i.d. samples from a fixed and unknown distribution with mean $c : \Gamma \to [0,1]$. The regret measure is also changed to the more standard pseudo-regret $\widetilde{R}_K = \sum_{k=1}^K \langle N_k, c_k \rangle - \langle q_{\pi^\star}, c \rangle$ where $\pi^\star \in \operatorname{argmin}_\pi J^{\pi,c}(s_0)$.[4] We remind the readers that the lower bound is indeed for the pseudo-regret and is constructed via this weaker adversary, and also that this is slightly different from the setting studied in (Tarbouriech et al., 2020a; Cohen et al., 2020) as mentioned in Section 1.

Our algorithm is based on the well-known *optimism in face of uncertainty* principle, which finds the best policy among all plausible MDPs subject to some additional constraints. First, we compute an optimistic cost function $\widehat{c}_k$ defined via $\widehat{c}_k(s,a)$ being[5]

$$\max\left\{\bar{c}_k(s,a) - 2\sqrt{A_k^c(s,a)\bar{c}_k(s,a)} - 7A_k^c(s,a), 0\right\}, \tag{4}$$

where $\bar{c}_k(s,a) = \frac{\sum_{j=1}^{k-1} c_j(s,a)\mathbb{I}_j(s,a)}{\mathbf{N}_k^c(s,a)}$ is the empirical cost mean, $\mathbf{N}_k^c(s,a) = \max\left\{\sum_{j=1}^{k-1} \mathbb{I}_j(s,a), 1\right\}$ is the number of times the cost at (s,a) was revealed (covering both the full-information and the bandit settings), and $A_k^c(s,a) = \frac{\ln(2SAK/\delta)}{\mathbf{N}_k^c(s,a)}$. Then, we find the best occupancy measure with respect to this optimistic cost, with the same constraint $\widetilde{\Delta}(T, \mathcal{P}_k)$ as in previous sections:

$$\widehat{q}_k = \operatorname*{argmin}_{q \in \widetilde{\Delta}(T,\mathcal{P}_k)} \langle q, \widehat{c}_k \rangle, \tag{5}$$

and finally execute the induced policy $\widetilde{\pi}_k = \widetilde{\pi}_{\widehat{q}_k}$ as before.

[3] To see this, note that $u_k(s,a)$ is equivalent to $\max_q q(s,a)$ where the maximization is over the set $\{q \in \widetilde{\Delta}(\infty, \mathcal{P}_k) : \pi_q = \widetilde{\pi}_k\}$, which consists of polynomially many linear constraints.

[4] We can get a bound for the standard regret with an extra cost of order $\widetilde{\mathcal{O}}(\sqrt{DT_\star K})$. Therefore, the standard regret and the pseudo regret are of the same order. We use the latter only for simplicity and convention.

[5] This is not to be confused with the estimator used in Section 5 with the same notation overloaded.

Algorithm 3 A near-optimal algorithm for stochastically oblivious adversary

Input: Upper bound on expected hitting time T, horizon parameter H_1 and confidence level δ
Parameters: $H_2 = \lceil 2D \rceil, H = H_1 + H_2$.
Initialization: $\mathbf{N}_1(s,a) = \mathbf{M}_1(s,a,s') = 0$ for all $(s,a,s') \in \Gamma \times (\mathcal{S} \cup \{g\})$, a Bernstein-SSP instance $\mathcal{B}$, $\mathcal{P}_1$ is the set of all possible transition functions.
for $k = 1, \dots, K$ **do**
  Compute the optimistic cost $\widehat{c}_k$ (Eq. (4)).
  Compute $\widehat{q}_k = \operatorname{argmin}_{q \in \widetilde{\Delta}(T, \mathcal{P}_k)} \langle q, \widehat{c}_k \rangle$.
  Execute $\widetilde{\pi}_k = \widetilde{\pi}_{\widehat{q}_k}$: $\tau_k = \text{RUN}(\widetilde{\pi}_k, \mathcal{B})$, receive $c_k \odot \mathbb{I}_k$.
  Update $\mathcal{P}_{k+1} = \text{TransEst}(\mathbf{N}, \mathbf{M}, \delta, H_1, H_2, \tau_k)$.

There is, however, one caveat in the approach above. Our analysis relies on one crucial property of $\widetilde{\pi}_k$: $J^{P_k, \widetilde{\pi}_k, \widehat{c}_k}(s,h) \le D$, that is, its state value with respect to the optimistic transition/cost is always no more than the diameter D. This holds automatically if we did not impose the hitting time constraint in Eq. (5), due to the existence of the fast policy π^f whose state value is never worse than D. With the hitting time constraint, however, this might not hold anymore. To address this, we slightly modify the loop-free instance $\widetilde{M}$ and give every state (s,h) (for $h \le H_1$) a *shortcut* to directly transit to (s_f, H_1+1) by taking action a_f, which is equivalent to allowing the learner to switch to Bernstein-SSP (whose role is similar to the fast policy) at any state and any time (*c.f.* Footnote 2). This ensures $J^{P_k, \widetilde{\pi}_k, \widehat{c}_k}(s,h) \lesssim D$ as desired; see Lemma 18. This modification can be implemented by a small change to the definition of $\widetilde{\Delta}$, and we defer the details to Appendix D. With this in mind, our final algorithm is presented in Algorithm 3.

Analysis The key reason that we can improve our regret bounds in this stochastic setting is as follows. First, since the estimated cost converges to the true cost fast enough, the previous dominating term $\sum_{k=1}^K \langle q_k - \widehat{q}_k, c_k \rangle$ can now be replaced by $\sum_{k=1}^K \langle q_k - \widehat{q}_k, \widehat{c}_k \rangle$. Then, similar to Eq. (3), the latter is in the order of $\sqrt{S^2 A \sum_{k=1}^K \mathbb{E}_k[\langle \widetilde{N}_k, \widehat{c}_k \rangle^2]}$, which is further bounded by $\sqrt{S^2 A \sum_{k=1}^K \langle q_k, \widehat{c}_k \odot Q^{\widetilde{\pi}_k, \widehat{c}_k} \rangle}$ according to the first inequality of Lemma 2. Finally, we make use of the aforementioned property $J^{P_k, \widetilde{\pi}_k, \widehat{c}_k}(s,h) \le D$ to show that $\langle q_k, \widehat{c}_k \odot Q^{\widetilde{\pi}_k, \widehat{c}_k} \rangle$ is roughly D^2, leading to a final bound of $\widetilde{\mathcal{O}}(\sqrt{S^2 A D^2 K})$ and improving over the $\widetilde{\mathcal{O}}\left(\sqrt{S^2 A D T_\star K}\right)$ bound in Theorem 1. We summarize our results in the following theorem.

Theorem 3. *If $T \ge T_\star + 1, H_1 \ge 8 T_{\max} \ln K$, and $K \ge H^2$, then Algorithm 3 ensures with probability at least $1 - 30\delta$, $\widetilde{R}_K = \widetilde{\mathcal{O}}(\sqrt{DTK} + DS\sqrt{AK} + H^3 S^3 A^2)$ in the full information setting and $\widetilde{R}_K = \widetilde{\mathcal{O}}(\sqrt{DTSAK} + DS\sqrt{AK} + H^3 S^3 A^2)$ in the bandit feedback setting.*

Comparing with the lower bounds, one sees that our bounds are only $\sqrt{S}$ factor larger, a gap that also appears in other settings such as (Cohen et al., 2020). Unfortunately, we are not able to obtain the same improvement in the general adversarial setting, and we in fact conjecture that the lower bound there can be improved to at least $\Omega\left(\sqrt{SADT_\star K}\right)$, which, if true, would require a lower bound construction that is actually adversarial, instead of being stochastic as in most existing lower bound proofs.

Acknowledgements

We thank Aviv Rosenberg, Chen-Yu Wei, and the anonymous reviewers for many helpful discussions and feedback. This work is supported by NSF Award IIS-1943607 and a Google Faculty Research Award.

References

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, 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*, pages 263–272, 2017.

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

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

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2011.

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.

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.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. *arXiv preprint arXiv:2102.01046*, 2021.

Alon Cohen, Haim Kaplan, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret bounds for stochastic shortest path. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8210–8219, 2020.

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

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4860–4869, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 33, 2020.

Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813, 2012.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5478–5486, 2019.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. *arXiv preprint arXiv:2006.11561*, 2020.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8604–8613, 2020a.

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, 2020b.

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, 2020a.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. A provably efficient sample collection strategy for reinforcement learning. *arXiv preprint arXiv:2007.06437*, 2020b.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7304–7312, 2019.

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.