Regret Guarantees for Linear Contextual Stochastic Shortest Path

Dor Polikar
School of Electrical Engineering,
Tel Aviv University

Alon Cohen
School of Electrical Engineering,
Tel Aviv University
Google Research, Tel Aviv

Abstract

We define the problem of linear Contextual Stochastic Shortest Path (CSSP), where at the beginning of each episode, the learner observes an adversarially chosen context that determines the MDP through a fixed but unknown linear function. The learner's objective is to reach a designated goal state with minimal expected cumulative loss, despite having no prior knowledge of the transition dynamics, loss functions, or the mapping from context to MDP. In this work, we propose LR-CSSP, an algorithm that achieves a regret bound of $\widetilde{O}(K^{2/3}d^{2/3}|S||A|^{1/3}B_\star^2 T_\star \log(1/\delta))$, where K is the number of episodes, d is the context dimension, S and A are the sets of states and actions respectively, $B_\star$ bounds the optimal cumulative loss and $T_\star$, unknown to the learner, bounds the expected time for the optimal policy to reach the goal. In the case where all costs exceed $\ell_{\min}$, LR-CSSP attains a regret of $\widetilde{O}(\sqrt{K \cdot d^2|S|^3|A|B_\star^3 \log(1/\delta)/\ell_{\min}})$. Unlike in contextual finite-horizon MDPs, where limited knowledge primarily leads to higher losses and regret, in the CSSP setting, insufficient knowledge can also prolong episodes and may even lead to non-terminating episodes. Our analysis reveals that LR-CSSP effectively handles continuous context spaces, while ensuring all episodes terminate within a reasonable number of time steps.

1 Introduction

We study the problem of learning Contextual Stochastic Shortest Path (CSSP), where the learner receives a context—additional side information—before each episode, that determines a specific SSP instance the learner faces during the episode. In each such instance, the learner attempts to reach a prespecified goal state while minimizing their expected cumulative loss.

Our problem is a special case of Reinforcement learning (RL)—a powerful framework for modeling sequential decision-making problems in unknown and stochastic environment. These environments are typically modeled as Markov decision processes (Puterman, 1994, MDP) that are assumed to remain fixed throughout the learning process. However, in many real-world applications—such as personalized medical treatment—the underlying decision-making dynamics may vary between individuals, with each instance effectively representing a different MDP.

Personalized medical treatment can be modeled as stochastic shortest path problem (SSP)—a special type of MDP, where a learner aims to reach a predefined goal state while minimizing their expected total loss. For example, consider a patient who has undergone a blood test revealing several abnormal biomarkers. The treatment process aims to bring these biomarkers into a healthy range, with each action (e.g., medication, lifestyle change) incurring a cost or risk. Importantly, both the transition dynamics and associated losses may differ between patients, reflecting variability in physiology and treatment response. The Contextual Stochastic Shortest Path (CSSP) model captures such personalized settings, allowing for individual-specific dynamics and objectives. More generally, SSP also generalizes both finite-horizon and discounted MDPs, and models a wide range of applications including game playing, autonomous driving, and finetuning of large language models.

Learning a single non-contextual instance of SSP, was originally suggested by Tarbouriech et al. (2020) who gave an algorithm with a $\widetilde{O}(K^{2/3})$ regret guarantee. Rosenberg et al. (2020); Cohen et al. (2021); Tarbouriech et al. (2021) later improved this to a tight bound of $\widetilde{\Theta}(\sqrt{(B_\star + B_\star^2)|S||A|K})$, where S and A de-

note the state and action space respectively, $B_\star$ is an upper bound on the expected cumulative loss of the optimal policy from any initial state and K is the number of episodes.

Contextual RL has also been extensively studied. Contextual Multi-Armed Bandits (CMABs) and Contextual finite horizon MDPs (CMDPs) extend classical decision-making models by incorporating context into the learning process. CMABs use context to guide action selection, with applications in online advertising and personalized recommendations (see, e.g., Bouneffouf et al., 2020). CMDPs, originally suggested in Hallak et al. (2015), generalize this approach by modeling state transitions, with context defining distinct MDPs. CMDPs were later studied in Levy et al. (2023); Modi et al. (2018); Qian et al. (2024); Levy et al. (2024), where the focus was on the context-to-dynamics mapping, and how to use it for computationally-efficient learning.

Our work borrows from Modi et al. (2018), and assumes that both the transition dynamics and the loss function decompose into linear context-dependent embeddings. The contexts are assumed to be generated arbitrarily, possibly adversarially. Our algorithm LR-CSSP is based on the well-established principle of *Optimism in the Face of Uncertainty* (see, e.g., Auer et al., 2008). It maintains confidence sets over the corresponding embeddings, which are guaranteed to contain the true parameters with high probability. LR-CSSP is guaranteed to reach the goal state while incurring a total regret of $\widetilde{O}(K^{2/3}d^{2/3}|S||A|^{1/3}B_\star^2 T_\star \log(1/\delta))$, where d is the context dimension, $B_\star$ bounds the optimal cumulative loss over all possible SSP instances and $T_\star$, unknown to the learner, bounds the expected time for the optimal policy to reach the goal. In the case where all losses exceed a minimal value $\ell_{\min}$, thus eliminating the risk of getting stuck in a loss-free loop, LR-CSSP attains regret of $\widetilde{O}(\sqrt{K \cdot d^2|S|^3|A|B_*^3 \log(1/\delta)/\ell_{\min}})$.

One of the main challenges in learning SSPs, which was also a challenge in previous work, is making sure that each episode terminates by reaching the goal state. Previous work mitigates this issue by dichotomizing state-action pairs into known, sufficiently sampled state-action pairs, and unknown. Once all state-action pairs are known, due to the optimistic approach, the algorithm is guaranteed to reach the goal state. Here, because of the contextual nature of the problem, there is an additional challenge of generalizing between different contexts in order to make sure episodes terminate quickly on new, possibly previously unseen, contexts. We still rely on known and unknown state-action pairs. However, as the contexts arrive arbitrarily, state-actions may fluctuate between known and unknown until they become permanently known.

An important future direction is extending LR - CSSP to settings with non-linear context dependencies. While linear embeddings offer tractability and theoretical guarantees, many real-world systems exhibit complex, non-linear dynamics. Incorporating kernel methods or deep learning-based representations could enhance expressiveness, allowing the model to capture richer structure while preserving sample efficiency through appropriate regularization or oracle-based strategies.

We see our work as a first step in studying general, other than finite horizon, CMDPs. Our hope is that future work tackles more general, non-linear, function approximation, possibly using oracle approaches such as in Levy et al. (2023).

1.1 Additional related work

There is a rich body of work on regret minimization in RL, much of which is based on the optimism principle. The majority of this literature focuses on the tabular setting (Auer et al., 2008; Azar et al., 2017; Jin et al., 2018; Fruit et al., 2018; Zanette and Brunskill, 2019; Efroni et al., 2019; Simchowitz and Jamieson, 2019). Recent advances have extended these ideas to function approximation under various structural assumptions, such as linearity, low-rank dynamics, or realizability (Yang and Wang, 2019; Jin et al., 2020; Zanette et al., 2020a,b; Vial et al., 2022; Chen et al., 2022; Min et al., 2022; Di et al., 2023). Min et al. (2022) study a stochastic shortest path problem with linear function approximation, providing a regret bound that characterizes learning performance under structural assumptions. Dann et al. (2023) proposed unified algorithmic frameworks that address a broad class of stochastic path problems, integrating various formulations such as SSPs and budgeted path planning into a single model with strong theoretical guarantees.

Sample complexity and regret guarantees for Contextual Decision Processes (CDPs) and Contextual MDPs (CMDPs) have been studied under various assumptions. OLIVE (Jiang et al., 2017) provides sample-efficient learning under the low Bellman rank assumption, while Witness Rank (Sun et al., 2019) enables PAC bounds for model-based learning. Generalization in smooth CMDPs and linear mixtures of MDPs is addressed by Modi et al. (2018), and Levy et al. (2023) provide a reduction from stochastic CMDPs to supervised learning. CMDPs extend Contextual Multi-Armed Bandits (CMAB), for which regret has been well characterized under linear assumptions (Xu and Zeevi, 2020; Chu et al., 2011) and function approximation tools (Foster and Rakhlin, 2020).

2 Preliminaries

2.1 Stochastic Shortest Path (SSP)

The Stochastic Shortest Path (SSP) is a Markov Decision Process (MDP) problem in which the learner interacts with an MDP $\mathcal{M} = (S, A, P, \ell, s_{\text{init}}, g)$, where S and A are finite sets of states and actions, respectively. The interaction with $\mathcal{M}$ starts in an initial state, at $s_0 = s_{\text{init}} \in S$. The learner transitions between states according to the dynamics: $P(\cdot \mid s, a)$ and the interaction is completed only when reaching $g \notin S$. During the course of its interaction, it suffers losses $\ell(s, a) \in [0, 1]$ every step according to $\ell(s, a) \sim \mathcal{D}_{s,a}$ with the objective of minimizing the expected cumulative loss on all steps and reaching g. The amount of steps required to reach g is, in general, unknown and depends both on $\mathcal{M}$ and the policy π. We denote this quantity I_k, which may be infinite.

A *stationary and deterministic policy* $\pi : S \mapsto A$ defines a mapping from states to actions. Given an MDP $\mathcal{M}$, any policy π induces a value function of a state s over the horizon H is defined as follows:

$$\mathcal{V}^{\pi}(s) = \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{t=0}^{I_k} \ell(s_t, a_t) \mid s_0 = s\right].$$

Proper policies. This work addresses the approximation of the optimal *proper* policy. Namely, a policy that guarantees reaching the goal state g from any initial state:

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

Note that the optimal proper policy may not necessarily be the optimal policy overall, in particular when there are loops with zero cumulative loss.

For a proper policy π, the value function $\mathcal{V}^{\pi}(s)$ is guaranteed to be finite for all $s \in S$ due to the finiteness of S. However, $\mathcal{V}^{\pi}(s)$ may be infinite if π is improper. Let $T^{\pi}(s)$ the expected time it takes for π to reach g starting at s. In particular, if π is proper then $T^{\pi}(s)$ is finite for all s; in contrast, if π is improper there exist at least one s such that $T^{\pi}(s) = \infty$.

The following Lemmas characterize the importance of proper policies:

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

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

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

Lemma 2 (Bertsekas and Tsitsiklis, 1991, Proposition 2)**.** Under the conditions of Lemma 1 the optimal policy $\pi^{\star}$ is stationary, deterministic, and proper. Moreover, a policy π is optimal if and only if it satisfies the Bellman optimality equations for all $s \in S$:

$$\mathcal{V}^{\pi}(s) = \min_{a \in A} \ell\big(s, a\big) + \sum_{s' \in S} P\big(s' \mid s, a\big)\mathcal{V}^{\pi}(s'), \qquad (1)$$

$$\pi(s) \quad \in \arg\min_{a \in A} \ell\big(s, a\big) + \sum_{s' \in S} P\big(s' \mid s, a\big)\mathcal{V}^{\pi}(s').$$

Lemma 1 gives us the conditions for a policy to be proper. Lemma 2 restricts the overall optimal policy to be *stationary, deterministic, and proper* and gives conditions for such policies to be optimal.

2.2 Concentration inequality for linear regression

Let $\theta^{\star} \in \mathbb{R}^d$ be an unknown but fixed parameter vector. At each time step $t = 1, 2, \ldots, T$, an input vector $X_t \in \mathbb{R}^d$ is possibly a random variable that is a function of $Y_1, Y_2, \ldots, Y_{t-1}$, and the learner observes:

$$Y_t = \langle \theta^{\star}, X_t \rangle + \eta_t,$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^d$, η_t is a bounded noise term conditioned on $X_1, X_2, \ldots, X_t$ with zero mean. Let $\hat{\theta}_t$ denote the ridge regression of $\theta^{\star}$ with regularization parameter $\lambda > 0$. Then, the estimate is given by:

$$\hat{\theta}_t = (X_{1:t}^T X_{1:t} + \lambda I)^{-1} X_{1:t}^T Y_{1:t}. \qquad (2)$$

where $X_{1:t} \in \mathbb{R}^{t \times d}$ is the matrix whose rows are the vectors $X_1, X_2, \ldots, X_t$, and $Y_{1:t} \in \mathbb{R}^t$ is the vector of corresponding observations $Y_1, Y_2, \ldots, Y_t$.

Theorem 2.1. *(Paraphrased from Abbasi-yadkori et al., 2011). Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a a zero mean variable bounded by R such that each η_t is F_t-measurable.*

Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1} measurable and let $\lambda > 0$ and for any $t \geq 0$ define: $\bar{V}_t = \lambda I + \sum_{s=1}^{t} X_s X_s^T, S_t = \sum_{s=1}^{t} \eta_s X_s$.

Define $Y_t = \langle X_t, \theta_* \rangle + \eta_t$, and assume that $\|\theta_*\|_2 \leq \beta$ and that for all $t \geq 1$, $\|X_t\|_2 \leq L$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$, θ_* lies in the set:

$$\Theta_t = \Big\{\theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{\bar{V}_t} \le R\sqrt{d \log \frac{1 + tL^2/\lambda}{\delta}} + \sqrt{\lambda}\,\beta\Big\}.$$

3 Problem Setup: Adversarial Linear-Contextual SSP

The Contextual Stochastic Shortest Path (CSSP) problem is a generalization of the SSP and the CMDP models, and is defined by the tuple $(\mathcal{C}, S, A, \mathcal{M})$, where $\mathcal{C}$, S, A are the context, state and action spaces, respectively. Unlike S and A, which are finite and discrete spaces, $\mathcal{C}$ may be uncountably infinite. The mapping $\mathcal{M}$ maps a context $c \in \mathcal{C}$ to an SSP $\mathcal{M}(c) = (S, A, P^c, \ell^c, s^c_{\text{init}}, g)$.

In every episode the learner receives a context $c \in C$, begins the interaction with the SSP $\mathcal{M}(c)$ in s^c_{init} and ends its interaction with $\mathcal{M}$ by arriving at the goal state g. Note the g is fixed across all $c \in C$ and $g \notin S$. Whenever the learner plays action a in state s, it incurs a loss $\ell^c(s,a) \in [0,1]$ and the next state $s' \in S \cup \{g\}$ is chosen with probability $P^c(s' \mid s, a)$. Crucially, both the loss function and the dynamics are context-dependent, making the optimal policy also context-dependent.

Similarly to the non-contextual SSP, we avoid explicitly addressing the goal state g. We assume that the probability of reaching the goal state by playing action a at state s is $P^c(g \mid s, a) = 1 - \sum_{s' \in S} P^c(s' \mid s, a)$.

Interaction protocol. The interaction between the learner and the environment is defined as follows. In each episode $k = 1, 2, ..., K$:

1. An adversary chooses a context $c_k \in C$.
2. The learner chooses an initial policy π^{c_k}.
3. The learner observes the trajectory generated by playing π^{c_k} in $\mathcal{M}$ until it reaches g during I_K steps. The learner may change its policy at any time-step.
4. The learner observes the trajectory from s^c_{init} to g denoted as:

$$\sigma^k = (c_k, s_{k,0}, a_{k,0}, \ell_{k,0}, \ldots, s_{k,I_{k-1}}, a_{k,I_{k-1}}, \ell_{I_{k-1}}, g).$$

We adopt the following assumptions to meet the assumptions of Lemmas 1 and 2:

Assumption 1. For every $c \in C$ there exists at least one proper policy.

Assumption 2. All losses are strictly positive: there exists some $\ell_{\min} > 0$ such that $\ell(s,a) \ge \ell_{\min}$ for all $(s,a) \in S \times A$.

Note that Assumption 2 can be relaxed (Section 4).

In addition to these assumptions, our main model-assumption is that both the loss function and the dynamics depend *linearly* on the context:

Assumption 3. A CSSP $(\mathcal{C}, S, A, \mathcal{M})$ is called linear, if there exists d such that $C = \{c \in \mathbb{R}^d \mid c_i \ge 0\ \forall i = 1, \ldots, d, \text{ and } \sum_{i=1}^d c_i = 1\}$ and for every context $c \in \mathcal{C}$ the associated SSP $\mathcal{M}(c) = (S, A, P^c, \ell^c, s^c_{\text{init}})$ can be decomposed as follows:

(i) $\mathbb{E}[\ell^c(s,a)] = \sum_{j=1}^d c_j \cdot \ell^\star_j(s,a) = \langle c,\ L^\star(s,a)\rangle$. for all $s, a \in S \times A$

(ii) For any $s, a, s' \in S \times A \times S$: $P^c(s' \mid s,a) = \sum_{j=1}^d c_j \cdot p^\star_j(s' \mid s,a) = \langle c,\ P^\star(s' \mid s,a)\rangle$, where each $p^\star_j(s' \mid s,a)$ defines a legal transition distribution.

In the above $L^\star(s,a) \in \mathbb{R}^d,\ \forall (s,a) \in S \times A$, and $P^\star(s' \mid s,a) \in \mathbb{R}^d,\ (s,a,s') \in S \times A \times S$ are fixed but unknown. We will sometimes write $P^\star(s,a) \in \mathbb{R}^{|S| \times d}$, referring to a matrix whose rows correspond to $P^\star(s' \mid s,a)$, for all $s' \in S$.

Learning formulation. The algorithm's performance is measured by the learner's *regret* over K episodes, that is, the sum of K differences between the total loss over the I_k steps during episode k and the expected loss of the optimal proper policy corresponding to the context c_k:

$$R_k = \sum_{k=1}^{K} \left[\sum_{t=1}^{I_k} \ell^{c_k}(s_{k,t}, a_{k,t}) - \min_{\pi \in \Pi_{\text{proper}}(c_k)} \mathcal{V}^\pi_{\mathcal{M}(c_k)}(s) \right],$$

where $\Pi_{\text{proper}}(c_k)$ is the set of all stationary, deterministic, and proper policies corresponding to $\mathcal{M}(c_k)$ (that is not empty by Assumption 1), and $(s_{k,t}, a_{k,t})$ is the t-th state-action pair at episode k. In the case where I_k is infinite for some k, we define $R_k = \infty$.

We denote the optimal proper policy for $\mathcal{M}(c)$ by $\pi^\star_c$, defined as:

$$\forall s \in S : \mathcal{V}^{\pi^\star}_c(s) = \min_{\pi \in \Pi_{\text{proper}}(c)} \mathcal{V}^\pi_{\mathcal{M}(c)}(s).$$

Let $B_\star > 0$ be an upper bound on the values of $\mathcal{V}^{\pi^\star}_c$, i.e., $B_\star \ge \max_{s \in S, c \in C} \mathcal{V}^{\pi^\star}_c(s)$ and assume that $B_\star \ge 1$.

We also define $T^\pi_{\mathcal{M}(c)}(s)$ the expected time it takes for π to reach g starting at s under context c. Let $T_\star > 0$ be an upper bound on the times $T^{\pi^\star_c}_{\mathcal{M}(c)}(s)$, i.e., $T_\star \ge \max_{s \in S, c \in C} T^{\pi^\star_c}_{\mathcal{M}(c)}(s)$.

4 Algorithm and Main Result

We present the *Linear-Regression Contextual Stochastic Shortest Path* (LR-CSSP, Algorithm 1). Our ap-

proach follows the well-known principle of optimism in the face of uncertainty: it maintains confidence sets that, with high probability, contain the true embedding of both the dynamics $P^\star$ and the loss $L^\star$. Each time the learner visits a state classified as "unknown" (described below), the algorithm selects optimistic embeddings from these sets by solving the following convex equations:

$$\hat{\ell}_m(s,a) \;= \arg\min_{L\in\mathbb{R}^d} \sum_{(c,\ell)\in D(s,a)} (c_m^\top L - \ell)^2 + \lambda\|L\|^2, \tag{3}$$

$$\hat{P}_m(s,a)' = \arg\min_{P\in\mathbb{R}^{|S|\times d}} \sum_{(s',c)\in D(s,a)} \|Pc - e_{s'}\|_2^2 + \lambda\|P\|_F^2. \tag{4}$$

Then, $\hat{P}_m(s,a)$ is derived by projecting $\hat{P}_m(s,a)'$ on the set of all stochastic matrices:

$$\hat{P}_m(s,a) = \arg\min_{P\in\Psi} \|P - \hat{P}'_m(s,a))\|_{\bar{V}_{s_h^m,a_h^m}^{\tau_{s_h^m,a_h^m}}}. \tag{5}$$

Here $\|P\|_V^2 = \mathrm{tr}(PVP^\top)$, for any $P \in \mathbb{R}^{|S|\times d}$ and $V \in \mathbb{R}^{d\times d}$. Finally, Algorithm 1 maps $\hat{\ell}_m(s,a)$ and $\hat{P}_m(s,a)$ to the context-specific dynamics and loss via the linear function approximation, and computes an optimistic policy in Algorithm 1. Let us denote $\bar{V}_m(s,a) \coloneqq \bar{V}_{s,a}^{\tau}$ where τ is the number of visits (s,a) had at the beginning of interval m. Formally, we define a state-action pair (s,a) as *known* in interval m, if:

$$\|c_m\|_{\bar{V}_m(s,a)^{-1}} < \frac{\ell_{\min}}{10\, B_\star \,\max\left\{\beta, \sqrt{\log(4m/\delta)}\right\}}, \tag{6}$$

where $\beta \coloneqq |S|(\sqrt{d\log\frac{8|S|^2|A|(1+\tau_{s,a}/\lambda)}{\delta}} + \sqrt{\lambda})$. Unlike in previous work on learning non-contextual SSP, in CSSPs a state-action pair is initially classified as unknown and may fluctuate between being known and unknown multiple times, until it is permanently known, as proved in Lemma 5. If all state-actions are known, the algorithm is guaranteed to reach the goal state in finite time as proved in Lemma 9 of Section A.1.

To ensure that the algorithm reliably reaches the goal state, each episode is divided into intervals. A new interval begins whenever the learner encounters an unknown state or reaches the goal state. At the start of each interval, new estimates of the environment's dynamics and associated loss are computed (Algorithm 1). The estimated dynamics are then projected onto the set of stochastic matrices to ensure it represents a valid probability distribution in Algorithm 1. Eqs. (3) to (5) can be solved efficiently using standard convex optimization techniques. Based on these projected dynamics and the estimated loss, a new optimistic policy is computed for the subsequent interval.

The following theorem bounds the regret of Algorithm 1.

Theorem 4.1. *Set $\lambda = 1$. With probability at least $1-\delta$ the regret of Algorithm 1 is bounded by:*

$$\widetilde{O}\left(\frac{d^2|A|B_\star^3|S|^3}{\ell_{\min}^2}\log\frac{1}{\delta} + \sqrt{K\cdot\frac{d^2|S|^3|A|B_\star^3\log\frac{1}{\delta}}{\ell_{\min}}}\right).$$

Prior Knowledge of $\ell_{\min}$ and $B_\star$. Assumption 2 can be circumvented by adding a small fixed loss to the loss of all state-action pairs. After this perturbation, the regret bounds from Theorem 4.1 continue to hold with $\ell_{\min} \leftarrow \epsilon$, $B_\star \leftarrow B_\star + \epsilon T_\star$ and an additional regret term $\epsilon K T_\star$. By selecting ϵ to balance these terms, we obtain the following corollary:

Corollary 4.1.1. Running Algorithm 1 using losses $\ell_\epsilon = \max(\ell(s,a),\epsilon)$ for $\epsilon = |S|\sqrt[3]{\frac{d^2|A|}{K}}$ gives the following regret bound with probability at least $1-\delta$:

$$R_k = \widetilde{O}\left(K^{2/3}d^{2/3}|S||A|^{1/3}B_\star^2 T_\star\log\frac{1}{\delta}\right).$$

The requirement of knowing $B_\star$ in advance can also be easily circumvented via the standard doubling trick. We initialize $B_\star$ to 1, and whenever the optimistic value function is observed to exceed this threshold, we reset the algorithm and update $B_\star$ to twice its previous value. This procedure increases the regret bound by only a factor of $\log B_\star$.

Notations. Let $D = \{(c_t, s_t, a_t, s_{t+1}, \ell_t) \mid t < T_{tot}(K)\}$ denote the set of all tuples collected over the K episodes, with $T_{tot}(K) = \sum_{k=1}^K I_k$ the total number of steps. We discard episode associations, treating all steps as a single time-ordered sequence. The context c_t remains fixed within each episode and changes only between episodes. For any state-action pair (s,a), define the subset: $D(s,a) = \{(c_t, s_t, a_t, s_{t+1}, \ell_t) \in D \mid (s_t, a_t) = (s,a)\}$.

Algorithm 1 LINEAR REGRESSION FOR CONTEXTUAL STOCHASTIC SHORTEST PATH (LR-CSSP)

input: state space S, action space A, context space $\mathcal{C}$, initial state s_{init}, confidence parameter δ, regularization parameter λ.
initialization: $\forall(s,a) : \tau_{s,a} \leftarrow 0$ count of visits for (s,a), $\bar{V}_{s,a}^{\tau_{s,a}} \leftarrow \lambda I$; $D = \emptyset$; an arbitrary policy $\tilde{\pi}$; $h \leftarrow 1$; $m \leftarrow 1$.
set $c_m \in C$ as the current context.
for $k = 1, 2, \ldots$ **do**
  $s_h^m = s_{\text{init}}$.
  while $s_h^m \neq g$ **do**
    choose $a_h^m = \hat{\pi}(s_h^m)$,
    observe loss $\ell_h \sim \ell^{c_m}(s_h^m, a_h^m)$
    and the next state: $s_{h+1}^m \sim P^{c_m}(\cdot \mid s_h^m, a_h^m)$.
    set $\tau_{s_h^m, a_h^m} \leftarrow \tau_{s_h^m, a_h^m} + 1$.
    set $\bar{V}_{s_h^m, a_h^m}^{\tau_{s_h^m, a_h^m}} \leftarrow \bar{V}_{s_h^m, a_h^m}^{\tau_{s_h^m, a_h^m}} + c_m c_m^\top$.
    Insert: $D \leftarrow (c_m, s_h^m, a_h^m, s_{h+1}^m, \ell_h)$
    if Eq. (6) is true or $s_{h+1}^m = g$ **then**
      # start a new interval
      if $s_{h+1}^m = g$ **then**
        set $c_{m+1} \leftarrow c_{k+1}$
      else
        $c_{m+1} \leftarrow c_m$
      end if
      $m \leftarrow m + 1$, $h \leftarrow 1$.
      for $(s,a) \in S \times A$ **do**
        Compute $\hat{\ell}_m(s,a)$ and $\hat{P}_m(s,a)'$ by solving Eqs. (3) and (4).
        Derive $\hat{P}_m(s,a)$ by projecting: $\hat{P}_m(s,a)'$ on the set of all stochastic matrices via Eq. (5)
      end for
      Compute optimistic policy as

$$(\pi_m, \tilde{\ell}_m, \widetilde{P}_m) \in \underset{\pi, \ell \in \Theta_m^\ell(c), P \in \Theta_m^P(c)}{\arg\min} \mathcal{V}_{\mathcal{M}(\ell,P)}^{\pi}(s),$$

      where $\mathcal{M}(\ell, P)$ is the instance of SSP defined with losses ℓ and dynamics P, and $\Theta_m^\ell(c), \Theta_m^P(c)$ are defined in Eq. (10) and Eq. (11) respectively.
    end if
    set $h \leftarrow h + 1$.
  end while
end for

Computational complexity. The optimistic optimal policy in Algorithm 1 of Algorithm 1 can be computed efficiently (see Tarbouriech et al., 2020). The core idea is to construct an augmented MDP in which the state space remains unchanged, but the action space is expanded to include tuples of the form $(a, \widetilde{P}, \widetilde{\ell})$ where $a \in A$, $\widetilde{P}$ and $\widetilde{\ell}$ are any transition and loss functions within the confidence sets $\Theta_m^P(c), \Theta_m^\ell(s,a)$ as defined in Eqs. (10) and (11). It can be shown that the optimistic policy along with the associated optimistic model—those that minimize the expected total cost over all policies and feasible transition functions—correspond to the optimal policy of the augmented MDP. This policy can be computed efficiently using Extended Value Iteration (EVI). Moreover, the updates in Algorithm 1 reduce to convex optimization problems, which can be solved in polynomial time using standard techniques. Consequently, the overall algorithm admits a polynomial-time implementation.

4.1 Regret analysis

We now sketch the proof of Theorem 4.1, and defer the full details to the supplementary material. The proof begins with the definition of the high probability event (HPE). Under the HPE, the remaining proof is reduced to a deterministic bound on the regret.

High probability event. We let Ω denote the event that the following holds:

1. $\forall(s,a) \in S \times A$, for all $m \geq 0$, $\vec{L}^\star(s,a)$ lies in the set:

$$\Theta_m^\ell(s,a) = \left\{ L \in \mathbb{R}^d \;\middle|\; \|L(s,a) - \hat{L}_m(s,a)\|_{\bar{V}_m(s,a)} \leq \sqrt{d \log \frac{8|S||A|(1+\tau_{s,a})/\lambda}{\delta}} + \sqrt{\lambda} \right\}. \tag{7}$$

   where $\tau_{s,a}$ here is the number of visits to s, a at the start of interval m.
2. $\forall(s,a) \in S \times A$, for all $m \geq 0$, $\vec{P}^\star(\cdot \mid s,a)$ lies in the set:

$$\Theta_m^P(s,a) = \left\{ P \in \mathbb{R}^{|S| \times d} \;\middle|\; \|P - \hat{P}_m(s,a)\|_{\bar{V}_m(s,a)} \leq \sqrt{d \log \frac{8|S|^2|A|(1+\tau_{s,a}/\lambda)}{\delta}} + \sqrt{\lambda} \right\}. \tag{8}$$

3. Let H^m be the length of interval m.

$$\sum_{h=1}^{H_m} \ell(s_h^m, a_h^m) \leq 48 B_\star \log \frac{4m}{\delta}.$$

   for all intervals m simultaneously.
4. For every interval m, define $\widetilde{V}_m$ as the value function of policy π_m over the optimistic model $\mathcal{M}(\tilde{\ell}_m, \widetilde{P}_m)$ associated with the context c_m. Then:

$$\sum_{m=1}^{M}\left(\sum_{h=1}^{H^m}\widetilde{V}_m(s_{h+1}) - \sum_{s'\in S} P^{c_k}(s' \mid s_h^m, a_h^m)\,\widetilde{V}_m(s')\right) \le B_\star\sqrt{T\log\tfrac{4T}{\delta}}. \tag{9}$$

Eqs. (7) and (8) guarantee that the confidence sets contain the true embeddings. Eq. (9) allows to bound the difference between the realized and expected cumulative cost of the algorithm.

We also define the derived confidence sets for every context c as:

$$\Theta_m^\ell(c) = \big\{\ell : S\times A\to\mathbb{R} \;\big|\; \ell(s,a) = L\cdot c,\ L\in\Theta_m^\ell(s,a),\ \forall(s,a)\in S\times A\big\}, \tag{10}$$

$$\Theta_m^P(c) = \big\{P' : S\times A\to\Delta(S) \;\big|\; P'(\cdot\mid s,a) = Pc,\ P\in\Theta_m^P(s,a),\ \forall(s,a)\in S\times A\big\}. \tag{11}$$

The following lemma ensures that the HPE holds with high probability (proved in Section A.1):

Lemma 3. The HPE holds with probability at least $1-\delta$.

The remainder of the proof bounds the regret deterministically, conditioned on the event Ω. We divide the proof into two parts: the first focuses on bounding the overall number of steps T, and the second focuses on bounding the regret.

First, we use the principle of optimism to show that the value function induced by the optimistic policy in the optimistic model (Algorithm 1) is upper bounded by $B_\star$ (Lemma 8). Then, we prove that the policies generated by the algorithm are guaranteed to be proper if all state-action pairs are known (Lemmas 9, 11 and 12). Indeed, when all state-action pairs are known, the true and the optimistic model are "close" enough, so Lemma 1 implies that any proper policy in the optimistic model is also proper in the true model. However, it is important to note that not all state-action pairs may be visited. In such cases, the learner may still reach the goal state g, but properness can no longer be guaranteed. Nevertheless, we bound the number of such occurrences in Lemma 5, ensuring that almost all episodes terminate at the goal state.

Algorithm 1 projects the dynamics embedding on the set Ψ of $|S|\times d$ matrices with stochastic columns. As C is defined as the probability simplex in $\mathbb{R}^d$, the projection ensures that $\widehat{P}_m(s,a)\cdot c$ is indeed a probability distribution over S for any $c\in C$. Moreover, since Ψ is convex, the projection preserves the statistical guarantee attained by the least squares estimate, as in the following Lemma.

Lemma 4.

$$\|P^\star(s,a) - \hat{P}_m(s,a)\|^2_{\bar{V}_m(s,a)} \le |S|\left(\sqrt{d\log\frac{8|S|^2|A|(1+\tau_{s,a}/\lambda)}{\delta}} + \sqrt{\lambda}\right)^2.$$

Proof. From Lemma 10 in Section A.1 we have with probability at least $1-\delta/8$:

$$\|P^\star(s,a) - \hat{P}_m(s,a)'\|^2_{\bar{V}_m(s,a)} \le |S|\left(\sqrt{d\log\frac{8|S|^2|A|(1+\tau_{s,a}/\lambda)}{\delta}} + \sqrt{\lambda}\right)^2.$$

The next step in Algorithm 1, Algorithm 1, is to project $\hat{P}_m(s,a)'$ on the set Ψ of matrices with stochastic columns, which contains $P^\star$. Therefore: $\|P^\star(s,a) - \hat{P}_m(s,a)'\|^2_{\bar{V}_m(s,a)} \le \|P^\star(s,a) - \hat{P}_m(s,a))\|^2_{\bar{V}_m(s,a)}$. □

Next, Lemma 5 bounds the number of time steps required for a state to become known. We include the full proof below:

Lemma 5. Let $N^+_{s,a}$ denote the number of times the pair (s,a) is unknown. Then:

$$N^+_{s,a} \le \frac{8dB_\star^2|S|^2}{\ell_{\min}^2}\log(1+\frac{\tau_{s,a}}{\lambda d})\cdot \left(\sqrt{d\log(8|S|^2|A|\left(\frac{1+\tau_{s,a}/\lambda}{\delta}\right)} + \sqrt{\lambda}\right)^2.$$

Proof. Denote:

$$a \coloneqq \frac{\ell_{\min}}{4\,B_\star\,|S|\left(\sqrt{d\log\frac{8|S|^2|A|(1+\tau_{s,a}/\lambda)}{\delta}} + \sqrt{\lambda}\right)}.$$

It follows that: $a^2N^+_{s,a} < \sum_{m=1}^{M} n_m(s,a)\|c_m\|^2_{\bar{V}_m^{-1}} \le 2d\log(1+\frac{\tau_{s,a}}{\lambda d})$, where the right-hand inequality is from Lemma 13 in Section A.1. Finally, isolating $N^+_{s,a}$ gives: $N^+_{s,a} < \frac{2d}{a^2}\log\left(1+\frac{\tau_{s,a}}{\lambda d}\right)$. □

By examining the algorithm, we obtain the following bound on the number of intervals:

Observation 1. The number of intervals, M, is bounded as: $M \le K + |S||A|N^+$ where N^+ is a bound on the number of visits it takes for any (s,a) to be known.

Now we are ready to state a key lemma of our analysis, a bound on the time it takes to complete K episodes. The proof is a combination of Lemma 5, Observation 1 and Item 3 of the HPE (full proof in Section A.1).

Lemma 6. Suppose the HPE holds. Then the total number of steps of the algorithm is:

$$T = \widetilde{O}\left(\frac{KB_\star}{\ell_{\min}} + \frac{|S|^3|A|d^2B_*^3}{\ell_{\min}^3}\log^2\frac{1}{\delta}\right). \tag{12}$$

We move now to bound the regret. Our regret analysis starts with the following regret decomposition:

$$R_K = \sum_{m=1}^{M}\sum_{h=1}^{H^m}\left(\widetilde{V}_m(s_{h^m}) - \widetilde{V}_m(s_{h+1}^m)\right) \tag{13}$$
$$- \sum_{k=1}^{K}\min_{\pi\in\Pi_{\text{proper}}(c_k)}\widetilde{V}_m(s_{\text{init}})$$
$$+ \sum_{m=1}^{M}\sum_{h=1}^{H^m}\sum_{s'\in S}\widetilde{V}_m(s'_m)\cdot \tag{14}$$
$$\left(P^{c_k}(s' \mid s_h^m, a_h^m) - \widetilde{P}_m^{c_k}(s' \mid s_h^m, a_h^m)\right)$$

$$+ \sum_{m=1}^{M}\left(\sum_{h=1}^{H^m}\widetilde{V}_m(s_{h+1}^m) \tag{15}\right.$$
$$\left. - \sum_{s'\in S}P^{c_k}(s' \mid s_h^m, a_h^m)\widetilde{V}_m(s'_m)\right).$$
$$+ \sum_{m=1}^{M}\sum_{h=1}^{H^m}\ell^{c_k}(s,a) - \widetilde{\ell}^{c_k}(s,a). \tag{16}$$

Eq. (13) captures the loss incurred from switching policies whenever an unknown state-action pair or the goal state are encountered.

Eq. (14) bounds the transition model misspecification weighted by the true value function. The proof leverages concentration inequalities applied to the empirical transition estimates, along with the definition of the confidence sets. It closely parallels the proof for the error in estimating the cost function, which we provide below for completeness.

Lemma 7. Suppose that Ω holds. It holds that:

$$\sum_{m=1}^{M}\sum_{h=1}^{H^m}\sum_{s'\in S}\widetilde{V}_m(s'_m)\Big(P^{c_k}(s' \mid s_h^m, a_h^m)$$
$$- \widetilde{P}_m^{c_k}(s' \mid s_h^m, a_h^m)\Big)$$
$$\leq B_\star|S|\left(\sqrt{d\log\frac{8|S|^2|A|(1+T/\lambda)}{\delta}} + \sqrt{\lambda}\right)\cdot$$
$$\sqrt{d|S||A||T|\log(1+\frac{T}{d\lambda})}.$$

Proof.

$$\sum_{m=1}^{T}\sum_{h=1}^{H^m}\sum_{s'\in S}\widetilde{\mathcal{V}}^{\pi}_{\mathcal{M}(c_k)}(s')\cdot$$
$$\left(P^{c_m}(s' \mid s_h^m, a_h^m) - \widetilde{P}^{c_m}(s' \mid s_h^m, a_h^m)\right)$$
$$\leq B_\star\sum_{s\in S}\sum_{a\in A}\sum_{m=1}^{M}n_m(s_h^m, a_h^m)\cdot$$
$$\|P^{c_m}(\cdot \mid s_h^m, a_h^m) - \widetilde{P}^{c_m}(\cdot \mid s_h^m, a_h^m)\|_1$$
(Hölder's inequality)
$$\leq B_\star\sum_{s\in S}\sum_{a\in A}\sum_{m=1}^{M}n_m(s_h^m, a_h^m)\cdot$$
$$\|c_m\|_{\bar{V}_m^{-1}(s,a)}\|P^\star(s_h^m, a_h^m) - \widetilde{P}_m(s_h^m, a_h^m)\|_{\bar{V}_m(s,a)}$$
(Hölder's inequality)
$$\leq B_\star\sum_{s\in S}\sum_{a\in A}\sum_{m=1}^{M}n_m(s_h^m, a_h^m)\cdot$$
$$\|c_m\|_{\bar{V}_m^{-1}(s,a)}\left(\sqrt{d\log\frac{8|S|^2|A|(1+\tau_{s,a}/\lambda)}{\delta}}\right.$$
$$\left. + \sqrt{\lambda}\right) \qquad \text{(Theorem 2.1)}$$
$$\leq B_\star\left(\sqrt{d\log\frac{8|S|^2|A|(1+T/\lambda)}{\delta}} + \sqrt{\lambda}\right)\cdot$$
$$\sum_{s\in S}\sum_{a\in A}\left[\sum_{m=1}^{M}n_m(s,a)\|c_m\|_{\bar{V}_m^{-1}(s,a)}\right]$$
$$\leq B_\star\left(\sqrt{d\log\frac{8|S|^2|A|(1+T/\lambda)}{\delta}} + \sqrt{\lambda}\right)\cdot$$
$$\sqrt{d|S||A|T\log(1+\frac{T}{d\lambda})},$$

where the last equality follows by Lemma 19 in Section A.2. □

Eq. (15) measures the difference between observed value at the next state and the expected value under the transition model. It is given directly from the HPE, and its proof is straightforward from Azuma's inequality and supplied in Lemma 20.

Eq. (16) represents the the error in estimating the loss function, and is proved in a similar way to Lemma 7.

Finally, Theorem 4.1 is the combination of Lemmas 7, 18, 20 and 21 when placing the bound on T and M from Lemma 6. The complete proof is found in Section A.2.

5 Acknowledgements

This project is supported by the Israel Science Foundation (ISF, grant number 2250/22).

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/e4a6222cdb5b34375400904f03d8e6a5-Paper.pdf.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/azar17a.html.

Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991. doi: 10.1287/moor.16.3.580.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2020. doi: 10.1109/CEC48606.2020.9185782.

Liyu Chen, Rahul Jain, and Haipeng Luo. Improved no-regret algorithms for stochastic shortest path with linear MDP. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3204–3245. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/chen22h.html.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *Advances in neural information processing systems*, 34:28350–28361, 2021.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. A unified algorithm for stochastic path problems. In Shipra Agrawal and Francesco Orabona, editors, *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 510–567. PMLR, 20–23 Feb 2023. URL https://proceedings.mlr.press/v201/dann23b.html.

Qiwei Di, Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning linear mixture stochastic shortest path. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7837–7864. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/di23a.html.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/f

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3199–3210. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning, 2018. URL https://arxiv.org/abs/1802.04020.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes, 2015. URL https://arxiv.org/abs/1502.02259.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In Doina Precup and Yee Whye

Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/jiang17c.html.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Orin Levy, Alon Cohen, Asaf Cassel, and Yishay Mansour. Efficient rate optimal regret for adversarial contextual MDPs using online function approximation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19287–19314. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/levy23a.html.

Orin Levy, Asaf Cassel, Alon Cohen, and Yishay Mansour. Eluder-based regret for stochastic contextual mdps. In *International Conference on Machine Learning*, pages 27326–27350. PMLR, 2024.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15584–15629. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/min22a.html.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 597–618. PMLR, 07–09 Apr 2018. URL https://proceedings.mlr.press/v83/modi18a.html.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Jian Qian, Haichen Hu, and David Simchi-Levi. Offline oracle-efficient learning for contextual mdps via layerwise exploration-exploitation tradeoff. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8210–8219. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/rosenberg20a.html.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/f

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/sun19a.html.

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in neural information processing systems*, 34:6843–6855, 2021.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Regret bounds for stochastic shortest path problems with linear function approximation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22203–22233. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/vial22a.html.

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *CoRR*, abs/2007.07876, 2020. URL `https://arxiv.org/abs/2007.07876`.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6995–7004. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/yang19b.html`.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/zanette19a.html`.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1954–1964. PMLR, 26–28 Aug 2020a. URL `https://proceedings.mlr.press/v108/zanette20a.html`.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR, 13–18 Jul 2020b. URL `https://proceedings.mlr.press/v119/zanette20a.html`.

A Proofs

A.1 Proof of Lemma 6

Lemma (Restatement of Lemma 6) Assume $\lambda \geq 1$. Then, Ω holds With probability at least $1-\delta$. This implies that the total number of steps of the algorithm is:

$$T = \widetilde{O}\left(\frac{KB_\star}{\ell_{\min}} + \frac{|S|^3|A|d^2B_*^3}{\ell_{\min}^3}\log^2\frac{1}{\delta}\right).$$

Lemma 8. Let m be an interval. If Ω holds, then $\widetilde{\mathcal{V}}^m(s) \leq \mathcal{V}^{\pi^\star}(s) \leq B_\star$ for every $s \in S$.

Proof. Since Ω holds, Equations 7 and 8 hold, which implies that the true dynamics $P^\star$ and the true loss $L^\star$ where considered in the minimization. Therefore, the chosen value function in Algorithm 1 in Algorithm 1 must be lower then that of $P^\star$ and $L^\star$, $\mathcal{V}^{\pi^\star}(s) \leq B_\star$ for every $s \in S$. □

Lemma 9. Let $\widetilde{\pi}$ be a policy, $\widetilde{P}$ be a transition function and $\widetilde{\ell}$ a loss function. Denote the loss-to-go of $\widetilde{\pi}$ with respect to $\widetilde{P}$ by $\widetilde{\mathcal{V}}$. Assume that for every $s \in S$, $\widetilde{\mathcal{V}}(s) \leq B_\star$ and

$$\left\|\widetilde{P}(\cdot \mid s, \tilde{\pi}(s)) - P(\cdot \mid s, \tilde{\pi}(s))\right\|_1 \leq \frac{\ell(s,\tilde{\pi}(s))}{2B_\star}. \tag{17}$$

$$|\widetilde{\ell}(s,\tilde{\pi}(s)) - \ell(s,\tilde{\pi}(s))| \leq \frac{\ell(s,\tilde{\pi}(s))}{4}. \tag{18}$$

Then, $\tilde{\pi}$ is proper (with respect to P), and it holds that $\mathcal{V}^{\tilde{\pi}}(s) \leq 4B_\star$ for every $s \in S$.

Proof. Consider the Bellman equations of $\tilde{\pi}$ with respect to transition function $\widetilde{P}$ and the loss $\widetilde{\ell}(s,a)$ at some state $s \in S$ (see Lemma 1) defined as:

$$\begin{aligned}\widetilde{\mathcal{V}}(s) &= \widetilde{\ell}(s,\tilde{\pi}(s)) + \sum_{s' \in S}\widetilde{P}(s' \mid s, \tilde{\pi}(s))\widetilde{\mathcal{V}}(s') \\ &= \widetilde{\ell}(s,\tilde{\pi}(s)) + \sum_{s' \in S}P(s' \mid s, \tilde{\pi}(s))\widetilde{\mathcal{V}}(s') + \sum_{s' \in S}\widetilde{\mathcal{V}}(s')\left(\widetilde{P}(s' \mid s, \tilde{\pi}(s)) - P(s' \mid s, \tilde{\pi}(s))\right). \end{aligned} \tag{19}$$

Notice that by our assumptions and using Hölder inequality,

$$\begin{aligned}\left|\sum_{s' \in S}\widetilde{\mathcal{V}}(s')\left(\widetilde{P}(s' \mid s, \tilde{\pi}(s)) - P(s' \mid s, \tilde{\pi}(s))\right)\right| &\leq \|\widetilde{P}(\cdot \mid s, \tilde{\pi}(s)) - P(\cdot \mid s, \tilde{\pi}(s))\|_1 \cdot \|\widetilde{\mathcal{V}}\|_\infty \\ &\leq \frac{\ell(s,\tilde{\pi}(s))}{2B_\star} \cdot B_\star = \frac{\ell(s,\tilde{\pi}(s))}{2}\,.\end{aligned}$$

Plugging this into Eq. 19, we obtain

$$\begin{aligned}
\widetilde{\mathcal{V}}(s) &\geq \widetilde{\ell}(s,\tilde{\pi}(s)) + \sum_{s'\in S} P(s' \mid s,\tilde{\pi}(s))\widetilde{\mathcal{V}}(s') - \frac{\ell(s,\tilde{\pi}(s))}{2} \\
&= \widetilde{\ell}(s,\tilde{\pi}(s)) - \frac{\ell(s,\tilde{\pi}(s))}{2} + \sum_{s'\in S} P(s' \mid s,\tilde{\pi}(s))\widetilde{\mathcal{V}}(s') \\
&\geq -|\widetilde{\ell}(s,\tilde{\pi}(s)) - \ell(s,\tilde{\pi}(s))| + \frac{\ell(s,\tilde{\pi}(s))}{2} + \sum_{s'\in S} P(s' \mid s,\tilde{\pi}(s))\widetilde{\mathcal{V}}(s') \\
&\geq -\frac{\ell(s,\tilde{\pi}(s))}{4} + \frac{\ell(s,\tilde{\pi}(s))}{2} + \sum_{s'\in S} P(s' \mid s,\tilde{\pi}(s))\widetilde{\mathcal{V}}(s') && \text{(Eq. (18))} \\
&= \frac{\ell(s,\tilde{\pi}(s))}{4} + \sum_{s'\in S} P(s' \mid s,\tilde{\pi}(s))\widetilde{\mathcal{V}}(s').
\end{aligned}$$

Therefore, defining $\mathcal{V}' = 4\widetilde{\mathcal{V}}$, then $\mathcal{V}'(s) \geq \ell(s,\tilde{\pi}(s)) + \sum_{s'\in S} P(s' \mid s,\tilde{\pi}(s))\mathcal{V}'(s')$ for all $s \in S$. The statement now follows by Lemma 1. □

Lemma 10. Suppose that state-action pair s, a was visited $\tau_{s,a}$ times during interval m. Then:

$$\begin{aligned}
&\operatorname{tr}\left((P^\star(s,a) - \hat{P}_m(s,a)')\bar{V}_{\tau_{s,a}}(P^\star(s,a) - \hat{P}_m(s,a)')^\top \right) \\
&\qquad \leq |S|\left(\sqrt{d\log\frac{8|S|^2|A|(1+\tau_{s,a})/\lambda)}{\delta}} + \sqrt{\lambda} \right)^2,
\end{aligned}$$

and the confidence sets of the dynamics and loss, defined in Eqs. (7) and (8) contain $P^\star(s,a)$, $L^\star(s,a)$ for all intervals m simultaneously with probability at least $1 - \frac{\delta}{4}$.

Proof. by Line 22 in Algorithm 1 and Eq. (4):

$$\begin{aligned}
\hat{P}_m(s,a)' &= \operatorname*{arg\,min}_{P\in\mathbb{R}^{|S|\times d}} \sum_{(s',c)\in D(s,a)} \|Pc - e_{s'}\|_2^2 + \lambda\|P\|_F^2 \\
&= \operatorname*{arg\,min}_{P\in\mathbb{R}^{|S|\times d}} \sum_{(s',c)\in D(s,a)} \sum_{s''\in S} \|P_{s''}c - e_{s'}(s'')\|_2^2 + \lambda \sum_{s''\in S} \|P_{s''}\|^2 \\
&= \operatorname*{arg\,min}_{P\in\mathbb{R}^{|S|\times d}} \sum_{s''\in S} \left(\sum_{(s',c)\in D(s,a)} \|P_{s''}c - e_{s'}(s'')\|_2^2 + \lambda\|P_{s''}\|^2 \right),
\end{aligned}$$

where $P_{s''}$ is the row of P corresponding to state s''. Therefore, we can solve for each $P_{s''}$ separately:

$$\hat{P}_m(s'' \mid s,a)' = \operatorname*{arg\,min}_{P_{s''}\in\mathbb{R}^d} \sum_{(s',c)\in D(s,a)} \|P_{s''}c - e_{s'}(s'')\|_2^2 + \lambda\|P_{s''}\|^2.$$

by Theorem 2.1 we have with probability at least $1-\delta'$ where $\delta' = \frac{\delta}{8|S|^2|A|}$ for each (s', a, s) of $P^\star(s' \mid s, a)$:

$$\|P^\star(s' \mid s,a) - \hat{P}_m(s' \mid s,a)'\|_{\bar{V}_m(s,a)} \leq \sqrt{d\log\frac{1+t/\lambda}{\delta'}} + \sqrt{\lambda},$$

Now aggregate all rows to the matrix $\hat{P}(s,a)' \in \mathbb{R}^{|S|\times d}$ by the union bound. By Lemma 4 we have that the distance is not increased:

$$\|P^\star(s,a) - \hat{P}_m(s,a)\|_{\bar{V}_m(s,a)}^2 \leq |S|\left(\sqrt{d\log\frac{8|S|^2|A|(1+t/\lambda)}{\delta}} + \sqrt{\lambda} \right)^2.$$

Finally, use the union bound for all s, a together with the loss for all s, a ($\delta'' = \frac{\delta}{8|S||A|}$). □

Lemma 11. Let m be an interval and suppose that Ω holds. Denote the context at interval m as c. If (s,a) are a known state-action pair, then:

$$\|\hat{P}_m(\cdot \mid s,a)\cdot c - P^\star(\cdot \mid s,a)\cdot c\|_1 \leq \frac{\ell(s,a)}{2B_\star}\ .$$

Proof. From Lemma 4 we have with probability at least $1-\delta$:

$$\begin{aligned}
&\mathrm{tr}((P^\star(s,a)-\hat{P}_m(s,a))\bar{V}_m(s,a)(P^\star(s,a)-\hat{P}(s,a))^\top) \\
&\quad\leq \mathrm{tr}((P^\star(s,a)-\hat{P}_m(s,a)')\bar{V}_m(s,a)(P^\star(s,a)-\hat{P}_m(s,a)')^\top) \\
&\quad\leq |S|\left(\sqrt{d\log\frac{8|S|^2|A|(1+\tau_{s,a}/\lambda)}{\delta}}+\sqrt{\lambda}\right)^2.
\end{aligned}$$

Therefore:

$$\begin{aligned}
\|P^\star(s,a)\cdot c_m - \hat{P}_m(s,a)\cdot c_m\|_1 &= \sum_{s'\in S}|P^\star(s'\mid s,a)\cdot c_m - \hat{P}_m(s'\mid s,a)\cdot c_m| \\
&\leq \sum_{s'\in S}\|P^\star(s'\mid s,a)-\hat{P}_m(s'\mid s,a)\|_{\bar{V}_m(s,a)}\|c_m\|_{\bar{V}_m(s,a)^{-1}} \quad \text{(Hölder's inequality)} \\
&\leq \|c_m\|_{\bar{V}_m(s,a)^{-1}}\cdot|S|\left(\sqrt{d\log\frac{8|S|^2|A|(1+\frac{\tau_{s,a}}{\lambda})}{\delta}}+\sqrt{\lambda}\right). \qquad (20)
\end{aligned}$$

Since the pair (s,a) is known, we have:

$$\|c_m\|_{\bar{V}_m(s,a)^{-1}} < \frac{\ell_{\min}}{10B_\star\max\left\{|S|(\sqrt{d\log\frac{8|S|^2|A|(1+\tau_{s,a}/\lambda)}{\delta}}+\sqrt{\lambda}),\sqrt{\log(4m/\delta)}\right\}}. \qquad (21)$$

Injecting Eq. (21) to Eq. (20):

$$\|P^\star_m(s,a)\cdot c_m - \hat{P}_m(s,a)\cdot c_m\|_1 \leq \frac{\ell_{\min}}{10B_\star}\leq\frac{\ell(s,a)}{2B_\star}. \quad \square$$

Lemma 12. Let m be an interval in episode with context c, and suppose that Ω holds. If (s,a) is a known state-action pair, then:

$$|\ell^\star(s,a)\cdot c_m - \hat{\ell}_m(s,a)\cdot c_m| \leq \frac{\ell(s,a)}{4}.$$

Proof. Following Line 22 in Algorithm 1 and Eq. (3). we apply Theorem 2.1) to get for probability at least $1-\delta$:

$$\|\ell^\star(s,a)-\hat{\ell}_m(s,a)\|_{\bar{V}_m(s,a)} \leq \sqrt{d\log\frac{|S|^2|A|(1+\frac{\tau_{s,a}}{\lambda})}{\delta}}+\sqrt{\lambda}.$$

Thus, we have:

$$\begin{aligned}
|\ell^\star(s,a)\cdot c - \hat{\ell}(s,a)\cdot c| &\underset{\{\mathrm{i}\}}{\leq} \|\ell^\star(s,a)-\hat{\ell}_i(s,a)\|_{\bar{V}_m(s,a)}\|c\|_{\bar{V}_m(s,a)^{-1}} \\
&= \|c\|_{\bar{V}_m^{-1}(s,a)}\cdot\left(\sqrt{d\log\frac{8|S|^2|A|(1+\frac{\tau_{s,a}}{\lambda})}{\delta}}+\sqrt{\lambda}\right) \qquad (22) \\
&\underset{\{\mathrm{ii}\}}{\leq} \|c\|_{\bar{V}_m^{-1}(s,a)}\cdot\left(\sqrt{d\log\frac{8|S|^2|A|(1+\frac{T}{\lambda})}{\delta}}+\sqrt{\lambda}\right),
\end{aligned}$$

where {i} follows from Hölder's inequality and {ii} because $t \leq T$. Now inject Eq. (6), since the pair (s,a) is known, to Eq. (22).

$$|\ell^\star(s,a) \cdot c - \hat{\ell}_m(s,a) \cdot c| \leq \frac{\ell_{\min}}{10} \leq \frac{\ell(s,a)}{4}. \quad \square$$

Lemma 13. Assume $\lambda \geq 1$. Let $n_m(s,a)$ be the number of visits to s,a during interval m. Then:

$$\sum_{m=1}^{M} n_m(s,a) \|c_m\|^2_{\bar{V}_m^{-1}(s,a)} \leq d \log(1 + \frac{\tau_{s,a}}{\lambda d}).$$

Proof. Let c_m be the context associated with the interval m and let $\bar{V}_m$ be the matrix $V_{\tau_{s,a}}$ that is set at the beginning of the interval m.

$$\begin{aligned}
\sum_{m=1}^{M} n_m(s,a) \|c_m\|^2_{\bar{V}_m(s,a)^{-1}} &= \sum_{m=1}^{M} n_m(s,a) c_m^\top \bar{V}_m^{-1}(s,a) c_m \\
&= \sum_{m=1}^{M} \operatorname{tr}(\bar{V}_m^{-1}(s,a) n_m(s,a) c_m c_m^\top) \\
&= \sum_{m=1}^{M} \operatorname{tr}(\bar{V}_m(s,a)^{-1}(\bar{V}_{m+1}(s,a) - \bar{V}_m(s,a)) \\
&\qquad\qquad (\bar{V}_{m+1}(s,a) = \bar{V}_m(s,a) + n_m(s,a) c_m c_m^\top) \\
&\leq 2 \sum_{m=1}^{M} (\log\det(\bar{V}_{m+1}(s,a)) - \log\det(\bar{V}_m(s,a))) \qquad (*) \\
&= 2(\log\det(\bar{V}_{M+1}(s,a)) - \log\det(\bar{V}_1(s,a))) \qquad \text{(Telescopic sum)} \\
&= 2(\log\det(\lambda^{-1}\bar{V}_{M+1}(s,a)),
\end{aligned}$$

where (*) follows due to concavity of $\log\det$, and the fact that $n_m(s,a) c_m^\top \bar{V}_m^{-1}(s,a) c_m \leq 1$, that holds because of the following. If (s,a) is unknown states, $n_m(s,a) \leq 1$, and $c_m^\top \bar{V}_m^{-1} c_m \leq \|c_m\|^2/\lambda \leq 1$, since $\lambda \geq 1$ and $\|c_m\| \leq 1$ by assumption. For known (s,a), by Item 3 of the HPE together with Eq. (6) yields

$$n_m(s,a) c_m^\top \bar{V}_m^{-1} c_m \leq \frac{48 B_\star \log(4m/\delta)}{\ell_{\min}} \cdot \frac{\ell_{\min}^2}{100 B_\star^2 \log(4m/\delta)} \leq 1,$$

since $\ell_{\min} \leq 1$ and $B_\star \geq 1$ by assumption. This yields inequality (i) of the Lemma.

Next,

$$\begin{aligned}
\log\det(\lambda^{-1}\bar{V}_{M+1}(s,a)) &\leq \log\operatorname{tr}\left(\frac{1}{\lambda d}\bar{V}_{M+1}(s,a)\right)^d \qquad \text{(AM-GM)} \\
&= d \log\operatorname{tr}\left(\frac{1}{\lambda d}(\lambda I + \sum_{t=1}^{\tau_{s,a}} c_t c_t^T)\right) \\
&= d \log(1 + \frac{1}{\lambda d}\sum_{t=1}^{\tau_{s,a}} \|c_t\|^2) \\
&\leq d \log(1 + \frac{\tau_{s,a}}{\lambda d}). \qquad \square
\end{aligned}$$

Lemma 14 (Rosenberg et al., 2020)**.** Let π be a proper policy such that for some $v > 0$, $\mathcal{V}^\pi(s) \leq v$ for every $s \in S$. Then, the probability that the loss of π to reach the goal state from any state s is more than m, is at most $2e^{-m/4v}$ for all $m \geq 0$. Note that a total loss of at most m implies that the number of steps is bounded above by $\frac{m}{\ell_{\min}}$.

Let us denote the trajectory visited in interval m by $U^m = (s_1^m, a_1^m, \ldots, s_{H^m}^m, a_{H^m}^m, s_{H^m+1}^m)$ where a_h^m is the action taken in s_h^m, and H^m is the length of the interval. In addition, the concatenation of the trajectories in the intervals up to and including interval m is denoted by $\bar{U}^m = \cup_{m'=1}^m U^{m'}$.

Lemma 15. Let m be an interval during an episode associated with c_k. For all $r \geq 0$, we have

$$\Pr\left[\sum_{h=1}^{H^m} \ell\big(s_h^m, a_h^m\big) > r \mid \bar{U}^{m-1}\right] \leq 3e^{-r/16B_\star}.$$

Proof. For a given SSP $\mathcal{M}(c_k)$ define the contracted SSP $\mathcal{M}^{\text{know}}(c_k) = (S^{\text{know}}, A, P_{c_k}^{\text{know}}, \ell_{c_k}, s_{\text{init}}^{c_k})$ in which every state $s \in S$ such that $(s, \tilde{\pi}^m(s))$ is unknown is contracted into the goal state. Let $\mathcal{M}(c_k), \widetilde{\mathcal{M}}^{\text{know}}(c_k)$ be the contracted SSP's corresponding to $P(\cdot)$ and $\widetilde{P}(\cdot)$ respectively; Let $\mathcal{V}_{\text{know}}^m$ and $\widetilde{\mathcal{V}}_{\text{know}}^m$ be the cost-to-go of $\tilde{\pi}^m$ in $\mathcal{M}^{\text{know}}(c_k)$ and $\widetilde{\mathcal{M}}^{\text{know}}(c_k)$, respectively. Now apply Lemma 9 in $\mathcal{M}^{\text{know}}(c_k)$ and obtain that $\mathcal{V}_{\text{know}}^m(s) \leq 4B_\star$ for every $s \in S^{\text{know}}$.

Notice that reaching the goal state in $\mathcal{M}^{\text{know}}(c_k)$ is equivalent to reaching the goal state or an unknown state-action pair in M, and also recall that all state-action pairs in the interval are known except for the first one. Thus, from Lemma 14,

$$\mathbb{P}\left[\sum_{h=2}^{H^m} \ell\big(s_h^m, a_h^m\big)\mathbb{I}\{\Omega^m\} > r-1 \mid \bar{U}^{m-1}\right] \leq 2e^{-(r-1)/16B_\star} < 3e^{-r/16B_\star}.$$

Where the last inequality is because $B_\star \geq 1$. Since the first state-action in the interval might not be known, its loss is at most 1, and therefore

$$\begin{aligned}\mathbb{P}\left[\sum_{h=1}^{H^m} \ell\big(s_h^m, a_h^m\big)\mathbb{I}\{\Omega^m\} > r \mid \bar{U}^{m-1}\right] &\leq \mathbb{P}\left[\sum_{h=2}^{H^m} \ell\big(s_h^m, a_h^m\big)\mathbb{I}\{\Omega^m\} > r-1 \mid \bar{U}^{m-1}\right] \\ &\leq 3e^{-r/16B_\star}.\end{aligned}$$

□

Lemma 16. $\sum_{h=1}^{H_m} \ell(s_h^m, a_h^m) \leq 48B_\star \log\frac{4m}{\delta}$ for all intervals m simultaneously with probability at least $1 - \frac{\delta}{4}$.

Proof. From Lemma 15, we have:

$$\mathbb{P}\left[\sum_{h=1}^{H^m} \ell\big(s_h^m, a_h^m\big)\mathbb{I}\{\Omega^m\} \leq 48B_\star \log\frac{4m}{\delta} \mid \bar{U}^{m-1}\right] \geq 1 - \frac{\delta}{16m^2}. \tag{23}$$

By the union bound Eq. (23) holds for all intervals m simultaneously with cumulative probability of at least $1 - \frac{\delta}{4}$. □

Lemma (Restatement of Lemma 3) The HPE holds with probability at least $1 - \delta$.

Proof. Apply Lemmas 10, 16 and 20 and the union bound. □

Now, we are ready to conclude the proof of Lemma 6.

Proof of Lemma 6. Section A.1 guarantees that Ω holds for all m with probability at least $1 - \delta$. Lemma 5 bounds the number of times the pair (s, a) is unknown. Since Ω holds, the combination of Lemmas 9, 11 and 12 ensures that the policy calculated in Algorithm 1 in Algorithm 1 is proper and therefore reaches the goal state or an unknown state-action pair with probability 1. If the cumulative loss of all intervals is bounded by the HPE (and therefore so is the length of the interval), we can use Observation 1 to conclude that:

$$T = \widetilde{\mathcal{O}}\left(\frac{KB_\star}{\ell_{\min}} + \frac{|S|^3|A|d^2B_\star^3}{\ell_{\min}^3}(\log^2\frac{1}{\delta})\right)$$

□

A.2 Proof of Theorem 4.1

Theorem (Restatement of Theorem 4.1). With probability at least $1-\delta$ the regret of algorithm 1 is bounded as follows:

$$R_k = \widetilde{O}\left(\frac{B_\star^3 d^2 |S|^3 |A|}{\ell_{\min}^2} \log \frac{1}{\delta} + \sqrt{K \cdot \frac{d|S|^3|A|B_\star^2 \log(1/\delta)}{\ell_{\min}}} \right). \tag{24}$$

Lemma 17. Suppose that the HPE holds, then:

$$R_K \le B_\star |S||A| N_{s,a} + 2B_\star |S| \left(\sqrt{d \log \frac{8|S|^2|A|(1+T/\lambda)}{\delta}} + \sqrt{\lambda} \right) \sqrt{d|S||A||T| \log(1 + \frac{T}{d\lambda})}$$
$$+ B_\star \sqrt{T \log \frac{4T}{\delta}}.$$

To analyze R_K, we begin by plugging in the Bellman optimality equation of $\tilde{\pi}_m$ with respect to $\widetilde{P}_m$ into R_K. This allows us to decompose $R_K$into four terms as follows.

$$R_K = \sum_{k=1}^{K} \sum_{t=1}^{I^k} \left(\widetilde{V}_m(s_h^m) - \widetilde{V}_m(s_{h+1}^m) \right) - \sum_{k=1}^{K} \min_{\pi \in \Pi_{\text{proper}}(c_k)} \widetilde{V}_m(s_{\text{init}}) \tag{25}$$

$$= \sum_{m=1}^{M} \sum_{h=1}^{H^m} \left(\widetilde{V}_m(s_h^m) - \widetilde{V}_m(s_{h+1}^m) \right) - \sum_{k=1}^{K} \min_{\pi \in \Pi_{\text{proper}}(c_k)} \widetilde{V}_m(s_{\text{init}}) \tag{26}$$

$$+ \sum_{m=1}^{M} \sum_{h=1}^{H^m} \sum_{s' \in S} \widetilde{V}_m(s'_m) \left(P^{c_k}(s' \mid s_h^m, a_h^m) - \widetilde{P}_m^{c_k}(s' \mid s_h^m, a_h^m) \right) \tag{27}$$

$$+ \sum_{m=1}^{M} \left(\sum_{h=1}^{H^m} \widetilde{V}_m(s_{h+1}^m) - \sum_{s' \in S} P^{c_k}(s' \mid s_h^m, a_h^m) \widetilde{V}_m(s'_m) \right). \tag{28}$$

$$+ \sum_{m=1}^{M} \sum_{h=1}^{H^m} \ell^{c_k}(s,a) - \widetilde{\ell}^{c_k}(s,a). \tag{29}$$

Lemma 18.

$$\sum_{m=1}^{M} \sum_{h=1}^{H^m} \left(\widetilde{V}_m(s_h^m) - \widetilde{V}_m(s_{h+1}^m)) \right) \le B_\star |S||A| \cdot N_{s,a} + \sum_{k=1}^{K} \min_{\pi \in \Pi_{\text{proper}}(c_k)} \mathcal{V}^\pi_{\mathcal{M}(c_k)}(s_{\text{init}})$$

Proof. Note that per interval $\sum_{m=1}^{M} \sum_{h=1}^{H^m} \left(\widetilde{V}_m(s_h^m) - \widetilde{V}_m(s_{h+1}^m)) \right)$ is a telescopic sum which equals $\widetilde{V}_m(s_1^m) - \widetilde{V}_m(s_{H^{m+1}}^m)$. Furthermore, for every two consecutive intervals $m, m+1$ one of the following occurs:

(i) If interval m ended in the goal state then $\widetilde{V}_m(s_{h+1}^m) = \widetilde{V}_m(g) = 0$ and $\widetilde{V}_m(s_1^{m+1}) = \widetilde{V}_m(s_{\text{init}})$ Thus,

$$\begin{aligned} \widetilde{V}_m(s_1^{m+1}) - \widetilde{V}_m(s_{H^m+1}^m) &= \widetilde{V}_m(s_1^{m+1}) \\ &= \widetilde{V}_m(s_{\text{init}}) \\ &\le \min_{\pi \in \Pi_{\text{proper}}(c_k)} V^\pi_{M(c_k)}(s_{\text{init}}), \end{aligned} \qquad \text{(Lemma 8)}$$

which adds up to $\sum_{k=1}^{K} \min_{\pi \in \Pi_{\text{proper}}(c_k)} \widetilde{V}_m(s_{\text{init}})$.

(ii) If interval m ended in an unknown state then

$$\widetilde{V}_m(s_1^{m+1}) - \widetilde{V}_m(s_{H^m+1}^m) \le \widetilde{V}_m(s_1^{m+1}) \le B_\star.$$

This happens at most $|S||A|N_{s,a}$ times. □

Lemma 19.

$$\sum_{s,a}\sum_{m=1} n_m(s,a)\|c_m\|_{\bar{V}_m^{-1}} \le \sqrt{d|S||A||T|\log(1+\frac{T}{d\lambda})}.$$

Proof.

$$\begin{aligned}
\sum_{m=1} n_m(s,a)\|c_m\|_{\bar{V}_m^{-1}(s,a)} &= \sum_{m=1}\sqrt{n_m(s,a)}\sqrt{n_m(s,a)}\|c_m\|_{\bar{V}_m^{-1}(s,a)} \\
&\le \sqrt{\sum_{m=1} n_m(s,a)}\sqrt{\sum_{m=1} n_m(s,a)\cdot\|c_m\|^2_{\bar{V}_m^{-1}(s,a)}} && \text{(Cauchy-Schwartz)} \\
&\le \sqrt{\tau_{s,a}}\sqrt{d\log\Big(1+\frac{\tau_{s,a}}{d\lambda}\Big)}. && \text{(Lemma 5)}
\end{aligned}$$

Therefore,

$$\sum_{s,a}\sum_{m=1} n_m(s,a)\|c_m\|_{\bar{V}_m^{-1}(s,a)} \le \sum_{s,a}\sqrt{d\log(1+\frac{T}{d\lambda})}\sqrt{\tau_{s,a}} \le \sqrt{d|S||A|T\log(1+\frac{T}{d\lambda})}.$$

Where the last inequality holds because we can apply Cauchy-Schwartz inequality on $\sum_{s,a}\sqrt{d\log(1+\frac{T}{d\lambda})}$ is fixed for all s,t and because $\sum_{s,a}\tau_{s,a}=T$. □

Lemma 20. With probability at least $1-\delta/2$,

$$\sum_{m=1}^{M}\left(\sum_{h=1}^{H^m}\widetilde{V}_m(s_{h+1}) - \sum_{s'\in S}P^{c_k}(s'\mid s_h^m,a_h^m)\widetilde{V}_m(s')\right) \le B_\star\sqrt{T\log\frac{4T}{\delta}}. \tag{30}$$

Proof. Consider the infinite sequence of random variables

$$X_t = \left(\sum_{h=1}^{H^m}\widetilde{V}_m(s_{h+1}) - \sum_{s'\in S}P^{c_k}(s'\mid s_h^m,a_h^m)\widetilde{V}_m(s')\right),$$

where m is the interval containing time t, and h is the index of time step t within interval m. Notice that this is a martingale difference sequence, and $|X_t|\le B_\star$ by Lemma 8. Now, we apply anytime Azuma's inequality (Theorem B.1) to any prefix of the sequence $\{X_t\}_{t=1}^{\infty}$. Thus, with probability at least $1-\delta/2$, for every T:

$$\sum_{t=1}^{T} X_t \le B_\star\sqrt{T\log\frac{4T}{\delta}}.$$ □

Lastly, the following Lemma bounds the difference between the estimated loss to the true loss, reflected in Equation 29:

Lemma 21. Suppose that Ω holds. It holds than that:

$$\sum_{m=1}^{M}\sum_{h=1}^{H^m}\ell^{c_k}(s,a)-\widetilde{\ell}^{c_k}(s,a) \le \left(\sqrt{d\log\frac{8|S||A|(1+T/\lambda)}{\delta}}+\sqrt{\lambda}\right)\sqrt{d|S||A||T|\log(1+\frac{T}{d\lambda})}$$

Proof.

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H^m} \ell^{c_k}(s,a) - \widetilde{\ell}^{c_k}(s,a) \\
&\quad\le \sum_{s\in S}\sum_{a\in A}\sum_{m=1}^{M} n_m(s,a)\|c_t\|_{\bar{V}^{-1}_{\tau_{s,a}}} \|\ell^{c_k}(s,a) - \widetilde{\ell}^{c_k}(s,a)\|_{\bar{V}_{\tau_{s,a}}} \\
&\quad\le \sum_{s\in S}\sum_{a\in A}\sum_{m=1}^{M} n_m(s,a)\|c_t\|_{\bar{V}^{-1}_{\tau_{s,a}}} \left(\sqrt{d\log\frac{8|S||A|(1+\tau_{s,a}\lambda)}{\delta}} + \sqrt{\lambda}\right) \\
&\quad\le \left(\sqrt{d\log\frac{8|S||A|(1+T/\lambda)}{\delta}} + \sqrt{\lambda}\right)\left(\sum_{s,a}\sum_{m=1}^{M} n_m(s,a)\|c_t\|_{\bar{V}^{-1}_{\tau_{s,a}}}\right)
\end{aligned}$$

Using Lemma 19, we conclude:

$$\sum_{m=1}^{M}\sum_{h=1}^{H^m} \ell^{c_k}(s,a) - \widetilde{\ell}^{c_k}(s,a) \le \left(\sqrt{d\log\frac{8|S||A|(1+T/\lambda)}{\delta}} + \sqrt{\lambda}\right)\sqrt{d|S||A||T|\log(1+\frac{T}{d\lambda})}.$$

□

Proof of Theorem 4.1. Lemma 18 Lemma 7, Lemma 20, Lemma 21 and the fact that $B_\star \ge 1$ prove Lemma 17. A.1 guarantees that the HPE occurs with probability at least $1-\delta$. We conclude the proof by injecting Lemma 6 to Lemma 17. □

B Concentration inequalities

Theorem B.1 (Anytime Azuma)**.** *(Rosenberg et al. (2020)) Let $(X_n)_{n=1}^{\infty}$ be a martingale difference sequence with respect to the filtration $(\mathcal{F})_{n=0}^{\infty}$ such that $|X_n| \le B$ almost surely. Then with probability at least $1-\delta$,*

$$\left|\sum_{i=1}^{n} X_i\right| \le B\sqrt{n\log\frac{2n}{\delta}}, \qquad \forall n \ge 1.$$