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 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

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	$\tilde{O}(\sqrt{S^2ADT_\star K})$ (Theorem 1)	$\tilde{O}(\sqrt{DT_\star K} + DS\sqrt{AK})$	$\Omega(\sqrt{DT_\star K} + D\sqrt{SAK})$
Bandit feedback	$\tilde{O}(\sqrt{S^3A^2DT_\star K})$ (Theorem 2)	$\tilde{O}(\sqrt{SADT_\star K} + DS\sqrt{AK})$	$\Omega(\sqrt{SADT_\star K} + D\sqrt{SAK})$

$\sqrt{S}$ 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 $k = 1, \ldots, K$, 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 $i = 1, \ldots, I_k$.

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 $T_{\max} \leq \frac{D}{c_{\min}}$ where $c_{\min} = \min_k \min_{(s,a)} c_k(s,a)$ is the minimum possible cost. For ease of presentation, we assume $D \leq T_\star$ 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 $q : \Gamma \times (\mathcal{S}\cup\{g\}) \rightarrow [0,\infty)$, 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 $\{1,\ldots,n\}$.

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 $\tilde{\mathcal{O}}(1)$ 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 $H_1 \leq h < H$ 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 $h \leq H_1$ 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 $h \leq H_1$ 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:*

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

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 $H = \tilde{\mathcal{O}}(T_{\max})$. 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 $\delta$
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, $\phi_1 = \text{argmin}_{\phi\in\Omega_1}\psi(\phi)$ (where $\Omega_k$ 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 $\phi_{k+1} = \text{argmin}_{\phi\in\Omega_{k+1}}\langle\phi, c_k\rangle + D_\psi(\phi, \phi_k)$.
```

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:

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

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 $R_K = \tilde{O}(\sqrt{S^2ADTK} + H^3S^2A)$.*

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 $R_K = \tilde{O}\left(\sqrt{S^3A^2DTK} + H^3S^3A^2\right)$.*

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 $c_1, \ldots, c_K$ 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 $\tilde{\mathcal{O}}(\sqrt{DT_\star K})$. 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 $\delta$
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 $k = 1, \ldots, K$ do
    Compute the optimistic cost $\widehat{c}_k$ (Eq. (4)).
    Compute $\widehat{q}_k = \text{argmin}_{q \in \widetilde{\Delta}(T,\mathcal{P}_k)} \langle q, \widehat{c}_k \rangle$.
    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 $\widetilde{\pi}_k$: $J^{P_k, \widetilde{\pi}_k, \widehat{c}_k}(s,h) \leq D$, 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 $h \leq H_1$) 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 $J^{P_k, \widetilde{\pi}_k, \widehat{c}_k}(s,h) \leq D$ 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 $\tilde{\mathcal{O}}(\sqrt{S^2 A D^2 K})$ and improving over the $\tilde{\mathcal{O}}\left(\sqrt{S^2 A D T_\star K}\right)$ bound in Theorem 1. We summarize our results in the following theorem.

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

A. Loop-free reduction

In this section, we give full proofs of lemmas related to the proposed loop-free reduction.

A.1. Proof of Lemma 1

Proof. Denote by $N'_k(s,a)$ the number of visits to (s,a) during episode k before switching to Bernstein-SSP, by $N''_k(s,a)$ the number of visits to (s,a) after switching to Bernstein-SSP, and by N_f the number of episodes where Bernstein-SSP is invoked. We have: $N_k(s,a) = N'_k(s,a) + N''_k(s,a)$ and $\sum_{k=1}^K \left\langle \widetilde{N}_k, c_k \right\rangle = \sum_{k=1}^K \langle N'_k, c_k\rangle + H_2 N_f$. Recall that the regret of running Bernstein-SSP (Cohen et al., 2020) for K' episodes under uniform cost is of $\mathcal{O}\left(DS\sqrt{AK'}\ln\frac{K'DSA}{\delta} + \sqrt{D^3S^4A^2}\ln^2\frac{K'DSA}{\delta}\right)$ with probability at least $1-\delta$. Conditioned on the event above, we have:

$$
\begin{aligned}
\sum_{k=1}^K \langle N''_k, c_k\rangle - H_2N_f &\le \left(\sum_{k=1}^K \langle N''_k, c_k\rangle - DN_f\right) - DN_f \qquad (H_2 \ge 2D)\\
&\le \mathcal{O}\left(DS\sqrt{AN_f}\ln\frac{KDSA}{\delta} + \sqrt{D^3S^4A^2}\ln^2\frac{KDSA}{\delta}\right) - DN_f\\
&= \mathcal{O}\left(D^{3/2}S^2A\ln^2\frac{KDSA}{\delta}\right),
\end{aligned}
$$

where in the last inequality we solve for the maximum of a quadratic function with variable N_f. Therefore,

$$
\sum_{k=1}^K \langle N_k, c_k\rangle = \sum_{k=1}^K \langle N'_k, c_k\rangle + H_2N_f + \sum_{k=1}^K \langle N''_k, c_k\rangle - H_2N_f \le \sum_{k=1}^K \left\langle \widetilde{N}_k, c_k\right\rangle + \tilde{\mathcal{O}}\left(D^{3/2}S^2A\ln^2\frac{1}{\delta}\right).
$$

On the other hand, by Lemma 3, the probability that the goal state is not reached within H_1 steps when executing $\pi^\star$ is at most $2e^{-\frac{H_1}{4T_{\max}}} \le \frac{2}{K^2}$. Hence, the expected cost of $\pi^\star$ in M and the expected cost of $\widetilde{\pi}^\star$ in $\widetilde{M}$ is very similar:

$$
J_k^{\widetilde{\pi}^\star}(\widetilde{s}_0) \le J_k^{\pi^\star}(s_0) + \frac{2H_2}{K^2} = J_k^{\pi^\star}(s_0) + \tilde{\mathcal{O}}\left(\frac{1}{K}\right).
$$

Putting everything together, and by $K \ge D$, we get:

$$
R_K = \sum_{k=1}^K \langle N_k, c_k\rangle - J_k^{\pi^\star}(s_0) \le \sum_{k=1}^K \left\langle \widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k\right\rangle + \tilde{\mathcal{O}}\left(D^{3/2}S^2A\ln^2\frac{1}{\delta}\right).
$$

□

A.2. Proof of Lemma 2

Proof. With the inequality $(\sum_{i=1}^I a_i)^2 \le 2\sum_i a_i(\sum_{i'=i}^I a_{i'})$, we proceed as

$$
\begin{aligned}
&\mathbb{E}_k\left[\left(\sum_{(s,a)\in\widetilde{\Gamma},h} \widetilde{N}_k(s,a,h)c_k(s,a,h)\right)^2\right]\\
&\le 2\mathbb{E}_k\left[\sum_{h=1}^H\sum_{(s,a)\in\widetilde{\Gamma}} \widetilde{N}_k(s,a,h)c_k(s,a,h)\left(\sum_{h'=h}^H\sum_{(s',a')\in\widetilde{\Gamma}} \widetilde{N}_k(s',a',h')c_k(s',a',h')\right)\right]\\
&= 2\mathbb{E}_k\left[\sum_{h=1}^H\sum_{(s,a)\in\widetilde{\Gamma}} \widetilde{N}_k(s,a,h)c_k(s,a,h)\mathbb{E}\left[\sum_{h'=h}^H\sum_{(s',a')\in\widetilde{\Gamma}} \widetilde{N}_k(s',a',h')c_k(s',a',h')\middle|\widetilde{s}_k^h = (s,h), a_k^h = a\right]\right]\\
&= 2\mathbb{E}_k\left[\sum_{h=1}^H\sum_{(s,a)\in\widetilde{\Gamma}} \widetilde{N}_k(s,a,h)c_k(s,a,h)Q_k^{\widetilde{\pi}}(s,a,h)\right] = 2\left\langle q_{\widetilde{\pi}}, c_k \odot Q_k^{\widetilde{\pi}}\right\rangle
\end{aligned}
$$

$$\leq 2\mathbb{E}_k\left[\sum_{h=1}^{H}\sum_{(s,a)\in\widetilde{\Gamma}}\widetilde{N}_k(s,a,h)Q_k^{\widetilde{\pi}}(s,a,h)\right] = 2\sum_{h=1}^{H}\sum_{(s,a)\in\widetilde{\Gamma}} q_{\widetilde{\pi}}(s,a,h)Q_k^{\widetilde{\pi}}(s,a,h)$$
$$= 2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}\cup\{s_f\}} q_{\widetilde{\pi}}(s,h)\sum_a \widetilde{\pi}(a|s)Q_k^{\widetilde{\pi}}(s,a,h) = 2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}\cup\{s_f\}} q_{\widetilde{\pi}}(s,h)J_k^{\widetilde{\pi}}(s,h) = 2\left\langle q_\pi, J_k^\pi\right\rangle.$$

This proves the first two inequalities. Denote by $q_{\widetilde{\pi},(s,h)}$ the occupancy measure of policy $\widetilde{\pi}$ with initial state (s,h), so that

$$J_k^{\widetilde{\pi}}(s,h) = \sum_{(s',a')\in\widetilde{\Gamma}}\sum_{h'\geq h} q_{\widetilde{\pi},(s,h)}(s',a',h')c_k(s',a',h').$$

Then, we continue with the following equalities:

$$\sum_{h=1}^{H}\sum_{s\in\mathcal{S}\cup\{s_f\}} q_{\widetilde{\pi}}(s,h)J_k^{\widetilde{\pi}}(s,h)$$
$$= \sum_{h=1}^{H}\sum_{s\in\mathcal{S}\cup\{s_f\}} q_{\widetilde{\pi}}(s,h)\sum_{(s',a')\in\widetilde{\Gamma}}\sum_{h'\geq h} q_{\widetilde{\pi},(s,h)}(s',a',h')c_k(s',a',h')$$
$$= \sum_{h=1}^{H}\sum_{(s',a')\in\widetilde{\Gamma}}\sum_{h'\geq h}\left(\sum_{s\in\mathcal{S}\cup\{s_f\}} q_{\widetilde{\pi}}(s,h)q_{\widetilde{\pi},(s,h)}(s',a',h')\right)c_k(s',a',h')$$
$$= \sum_{h=1}^{H}\sum_{(s',a')\in\widetilde{\Gamma}}\sum_{h'\geq h} q_{\widetilde{\pi}}(s',a',h')c_k(s',a',h') = \sum_{h'=1}^{H}\sum_{(s',a')\in\widetilde{\Gamma}}\sum_{h\leq h'} q_{\widetilde{\pi}}(s',a',h')c_k(s',a',h')$$
$$= \sum_{h'=1}^{H}\sum_{(s',a')\in\widetilde{\Gamma}} h'\cdot q_{\widetilde{\pi}}(s',a',h')c_k(s',a',h') = \left\langle q_{\widetilde{\pi}}, \vec{h}\circ c_k\right\rangle,$$

where in the third line we use the equality $\sum_{s\in\mathcal{S}\cup\{s_f\}} q_{\widetilde{\pi}}(s,h)q_{\widetilde{\pi},(s,h)}(s',a',h') = q_{\widetilde{\pi}}(s',a',h')$ by definition (since both sides are the probability of visiting (s',a',h')). This proves the last equality and completes the proof. □

Lemma 3. *(Rosenberg and Mansour, 2020, Lemma E.1) Let π be a policy with expected hitting time at most τ starting from any state. Then, the probability that π takes more than m steps to reach the goal state is at most $2e^{-\frac{m}{4\tau}}$.*

B. Omitted details for Section 4

In this section, we provide all omitted algorithms and proofs for Section 4. We first introduce a lemma of a Bellman type law of total variance (Azar et al., 2017), which is the key in obtaining a regret bound without $T_{\max}$ dependency in the dominating term. Then in Section B.1, we provide details on transition estimation and prove the main lemma (Lemma 9) that gives a data dependent upper bound on the transition estimation error. Finally we prove Theorem 1 in Section B.2.

In the rest of this section, we use the shorthand Var_k for $\text{Var}[\cdot|\mathcal{F}_k]$.

Lemma 4. *Consider executing policy $\widetilde{\pi}_k$ induced by occupancy measure q_k in $\widetilde{M}$ in episode k with an arbitrary cost function $c_k : \Gamma \to [0,\infty)$, and define $\mathbb{V}_k(s,a,h) = \text{Var}_{S'\sim P(\cdot|s,a,h)}[J_k^{\widetilde{\pi}_k}(S',h+1)]$. Then, $\langle q_k, \mathbb{V}_k\rangle \leq \text{Var}_k\left[\left\langle \widetilde{N}_k, c_k\right\rangle\right]$.*

Proof. First, we have

$$\text{Var}_k\left[\left\langle \widetilde{N}_k, c_k\right\rangle\right] = \mathbb{E}_k\left[\left(\sum_{h=1}^{H} c_k(s^h,a^h,h) - J_k^{\widetilde{\pi}_k}(s^1,1)\right)^2\right]$$
$$= \mathbb{E}_k\left[\left(\sum_{h=2}^{H} c_k(s^h,a^h,h) - J_k^{\widetilde{\pi}_k}(s^2,2) + c_k(s^1,a^1,1) + J_k^{\widetilde{\pi}_k}(s^2,2) - J_k^{\widetilde{\pi}_k}(s^1,1)\right)^2\right].$$

Note that $c_k(s^1,a^1,1)+J_k^{\widetilde{\pi}_k}(s^2,2)-J_k^{\widetilde{\pi}_k}(s^1,1)\in\sigma(s^1,a^1,s^2)$ and $\mathbb{E}_k\left[\sum_{h=2}^H c_k(s^h,a^h,h)-J_k^{\widetilde{\pi}_k}(s^2,2)\middle|s^1,a^1,s^2\right]=0$. Thus,

$$\begin{aligned}&\mathbb{E}_k\left[\left(\sum_{h=2}^H c_k(s^h,a^h,h)-J_k^{\widetilde{\pi}_k}(s^2,2)+c_k(s^1,a^1,1)+J_k^{\widetilde{\pi}_k}(s^2,2)-J_k^{\widetilde{\pi}_k}(s^1,1)\right)^2\right]\\&=\mathbb{E}_k\left[\left(\sum_{h=2}^H c_k(s^h,a^h,h)-J_k^{\widetilde{\pi}_k}(s^2,2)\right)^2\right]+\mathbb{E}_k\left[\left(c_k(s^1,a^1,1)+J_k^{\widetilde{\pi}_k}(s^2,2)-J_k^{\widetilde{\pi}_k}(s^1,1)\right)^2\right].\end{aligned}$$

Moreover, $c_k(s^1,a^1,1)+\sum_{s'}P(s'|s^1,a^1,1)J_k^{\widetilde{\pi}_k}(s',2)-J_k^{\widetilde{\pi}_k}(s^1,1)\in\sigma(s^1,a^1)$ and

$$\mathbb{E}_k\left[J_k^{\widetilde{\pi}_k}(s^2,2)-\sum_{s'}P(s'|s^1,a^1,1)J_k^{\widetilde{\pi}_k}(s',2)\middle|s^1,a^1\right]=0.$$

Thus,

$$\begin{aligned}&\mathbb{E}_k\left[\left(c_k(s^1,a^1,1)+J_k^{\widetilde{\pi}_k}(s^2,2)-J_k^{\widetilde{\pi}_k}(s^1,1)\right)^2\right]\\&=\mathbb{E}_k\left[\left(c_k(s^1,a^1,1)+\sum_{s'}P(s'|s^1,a^1,1)J_k^{\widetilde{\pi}_k}(s',2)-J_k^{\widetilde{\pi}_k}(s^1,1)\right)^2\right]+\mathbb{E}_k\left[\left(J_k^{\widetilde{\pi}_k}(s^2,2)-\sum_{s'}P(s'|s^1,a^1,1)J_k^{\widetilde{\pi}_k}(s',2)\right)^2\right]\\&\geq\mathbb{E}_k\left[\left(J_k^{\widetilde{\pi}_k}(s^2,2)-\sum_{s'}P(s'|s^1,a^1,1)J_k^{\widetilde{\pi}_k}(s',2)\right)^2\right]=\mathbb{E}_k[\mathbb{V}_k(s^1,a^1,1)].\end{aligned}$$

Plugging these back, we obtain:

$$\begin{aligned}\mathrm{Var}_k\left[\left\langle\widetilde{N}_k,c_k\right\rangle\right]&=\mathbb{E}_k\left[\left(\sum_{h=1}^H c_k(s^h,a^h,h)-J_k^{\widetilde{\pi}_k}(s^1,1)\right)^2\right]\\&\geq\mathbb{E}_k\left[\left(\sum_{h=2}^H c_k(s^h,a^h,h)-J_k^{\widetilde{\pi}_k}(s^2,2)\right)^2\right]+\mathbb{E}_k[\mathbb{V}_k(s^1,a^1,1)].\end{aligned}$$

Doing this repeatedly, we get:

$$\mathrm{Var}_k\left[\left\langle\widetilde{N}_k,c_k\right\rangle\right]\geq\mathbb{E}_k\left[\sum_{h=1}^H\mathbb{V}_k(s^h,a^h,h)\right]=\sum_{(s,a),h}q_k(s,a,h)\mathbb{V}_k(s,a,h)=\langle q_k,\mathbb{V}_k\rangle,$$

finishing the proof. □

B.1. Omitted Details for Transition Estimation

In this subsection, we introduce a sub-procedure TransEst (Algorithm 4) used in all algorithms proposed in this paper for transition estimation. It takes the learner's trajectory as input and outputs a confidence set $\mathcal{P}_{k+1}$ (and additionally the counter $\widetilde{N}_k$ for the bandit setting) at the end of episode k. We first introduce some notations. Denote by Λ_P the set of valid transitions $\widetilde{P}$ for $\widetilde{M}$ based on its layer structure. Also define the entries of unknown transition in $\widetilde{M}$ as $U=\Gamma\times(\mathcal{S}\cup\{g\})\times[H_1-1]$, such that $\widetilde{P}(s'|s,a,h)=P(s'|s,a),\forall(s,a),s',h\in U$.

We implement the confidence sets of transition function by maintaining a separate (empirical) Bernstein confidence bound for each state, action, and next state in M around its empirical estimate:

$$\epsilon_k(s,a,s')=4\sqrt{\bar{P}_k(s'|s,a)A_k(s,a)}+28A_k(s,a),$$

Algorithm 4 TransEst$(\mathbf{N}, \mathbf{M}, \delta, H_1, H_2, \tau)$

```
**Input:** counters $\mathbf{N}, \mathbf{M}$, confidence parameter $\delta$, horizon parameters $H_1, H_2$, trajectory $\tau = \{s^1, a^1, s^2, a^2, \ldots, a^{h-1}, s^h\}$.
**Initialization:** $\widetilde{N}(s,a) = 0$, $\mathbf{N}_{k+1}(s,a) = \mathbf{N}_k(s,a)$, and $\mathbf{M}_{k+1}(s,a) = \mathbf{M}_k(s,a)$ for any $(s,a) \in \widetilde{\Gamma}$.
**for** $t = 1, \ldots, h-1$ **do**
    Update counters: $\mathbf{N}_{k+1}(s^t, a^t) \leftarrow \mathbf{N}_k(s^t, a^t) + 1$, $\mathbf{M}_{k+1}(s^t, a^t, s^{t+1}) \leftarrow \mathbf{M}_k(s^t, a^t, s^{t+1}) + 1$, $\widetilde{N}(s^t, a^t) \leftarrow \widetilde{N}(s^t, a^t) + 1$.
Compute confidence set $\mathcal{P}_{k+1}$ defined in Eq. (6).
**if** $s^h \neq g$ **then**
    $\widetilde{N}(s_f, a_f) \leftarrow H_2$.
**Return:** $\mathcal{P}_{k+1}$ (and additionally $\widetilde{N}_k$ for the bandit setting).
```

for any $(s,a) \in \Gamma, s' \in \mathcal{S} \cup \{g\}$, where $\bar{P}_k(s'|s,a) = \frac{\mathbf{M}_k(s,a,s')}{\mathbf{N}_k^+(s,a)}$ is the empirical transition estimation, $\mathbf{N}_k^+(s,a) = \max\{1, \mathbf{N}_k(s,a)\}$, $\mathbf{N}_k(s,a)$ is the number of visits to (s,a) before episode k, $\mathbf{M}_k(s,a.s')$ is the number of visits to $(s,a), s'$ before episode k, and $A_k(s,a) = \frac{\ln(\frac{HKSA}{\delta})}{\mathbf{N}_k^+(s,a)}$. We then define the confidence set in episode k as follows:

$$\mathcal{P}_k = \left\{ \widehat{P} \in \Lambda_P : |\widehat{P}(s'|s,a,h) - \bar{P}_k(s'|s,a)| \leq \epsilon_k(s,a,s'), \forall (s,a), s', h \in U \right\}. \tag{6}$$

Next, we introduce several lemmas useful for bounding the bias of transition estimation.

Lemma 5. *(Lemma 4.2 in (Cohen et al., 2020)) With probability at least $1-\delta$, $P \in \mathcal{P}_k$ for all k.*

Lemma 6. *Assume $\widehat{P} \in \mathcal{P}_k$. Then, under the event of Lemma 5 we have*

$$|\widehat{P}(s'|s,a,h) - P(s'|s,a,h)| \leq \mathbb{I}\{(s,a), s', h \in U\} \left(8\sqrt{P(s'|s,a,h)A_k(s,a)} + 136 A_k(s,a) \right) \overset{\text{def}}{=} \epsilon_k^\star(s,a,s').$$

Proof. First note that $\widehat{P}(s'|s,a,h) = P(s'|s,a,h)$ when $(s,a), s', h \notin U$. When $(s,a), s', h \in U$, the result follows from $P(s'|s,a,h) = P(s'|s,a)$, definition of $\epsilon_k(s,a,s')$ and (Cohen et al., 2020, Lemma B.13). □

Lemma 7. *For any occupancy measure q and $\widehat{q}$ induced by the same policy π but different transition functions P and $\widehat{P}$ respectively, we have:*

$$\widehat{q}(s,a,h) - q(s,a,h) = \sum_{(s',a'),s''} \sum_{m=1}^{h-1} q(s',a',m)(\widehat{P}(s''|s',a',m) - P(s''|s',a',m))\widehat{q}_{(s'',m+1)}(s,a,h),$$

where $\widehat{q}_{(s'',m+1)}$ is the occupancy measure with respect to π, $\widehat{P}$, and initial state $(s'', m+1)$. As a result, under the event of Lemma 5,

$$\begin{aligned} |\langle \widehat{q} - q, c \rangle| &= \left| \sum_{(s,a),s',h} q(s,a,h)(\widehat{P}(s'|s,a,h) - P(s'|s,a,h)) J^{\widehat{P},\pi}(s', h+1) \right| \\ &\leq \langle |\widehat{q} - q|, c \rangle \leq H \sum_{(s,a),s',h \in U} q(s,a,h)\epsilon_k^\star(s,a,s'). \end{aligned}$$

Proof. We prove the first statement by induction on h. Denote by $q_{(s,h)}$ the occupancy measure of π_q with initial state (s,h). When $h = 1$, the statement is true by $q(s,a,h) = \widehat{q}(s,a,h) = \pi(a|s,1)\mathbb{I}\{s = s_0\}$. For the induction step with $h > 1$:

$$\begin{aligned} &\widehat{q}(s,a,h) - q(s,a,h) = \pi(a|s,h)(\widehat{q}(s,h) - q(s,h)) \\ &= \pi(a|s,h) \sum_{(s',a')} \left(\widehat{P}(s|s',a',h-1)\widehat{q}(s',a',h-1) - P(s|s',a',h-1)q(s',a',h-1) \right) \\ &\qquad\qquad (q(s,h) = \textstyle\sum_{(s',a')} P(s|s',a',h-1)q(s',a',h-1)) \end{aligned}$$

$$= \underbrace{\pi(a|s,h) \sum_{(s',a')} \widehat{P}(s|s',a',h-1)\left(\widehat{q}(s',a',h-1) - q(s',a',h-1)\right)}_{\chi_1}$$

$$+ \underbrace{\pi(a|s,h) \sum_{(s',a')} q(s',a',h-1)\left(\widehat{P}(s|s',a',h-1) - P(s|s',a',h-1)\right)}_{\chi_2}$$

For χ_1, by the induction step, we have:

$$\widehat{q}(s',a',h-1) - q(s',a',h-1) = \sum_{(s'',a''),s'''} \sum_{m=1}^{h-2} q(s'',a'',m)(\widehat{P} - P)(s'''|s'',a'',m)\widehat{q}_{(s''',m+1)}(s',a',h-1).$$

Thus, by $\sum_{(s',a')} \widehat{q}_{(s''',m+1)}(s',a',h-1)\widehat{P}(s|s',a',h-1)\pi(a|s,h) = \widehat{q}_{(s''',m+1)}(s,a,h)$:

$$\begin{aligned}
\chi_1 &= \pi(a|s,h) \sum_{(s',a')} \widehat{P}(s|s',a',h-1) \sum_{(s'',a''),s'''} \sum_{m=1}^{h-2} q(s'',a'',m)(\widehat{P} - P)(s'''|s'',a'',m)\widehat{q}_{(s''',m+1)}(s',a',h-1) \\
&= \sum_{(s'',a''),s'''} \sum_{m=1}^{h-2} q(s'',a'',m)\left(\widehat{P}(s'''|s'',a'',m) - P(s'''|s'',a'',m)\right)\widehat{q}_{(s''',m+1)}(s,a,h).
\end{aligned}$$

For χ_2, note that $\pi(a|s'',h)\mathbb{I}\{s'' = s\} = q_{(s'',h)}(s,a,h)$. Thus,

$$\begin{aligned}
&\pi(a|s,h)\left(\widehat{P}(s|s',a',h-1) - P(s|s',a',h-1)\right) \\
&= \sum_{s''} \pi(a|s'',h)\mathbb{I}\{s'' = s\}\left(\widehat{P}(s''|s',a',h-1) - P(s''|s',a',h-1)\right) \\
&= \sum_{s''} q_{(s'',h)}(s,a,h)\left(\widehat{P}(s''|s',a',h-1) - P(s''|s',a',h-1)\right).
\end{aligned}$$

and,

$$\chi_2 = \sum_{(s',a'),s''} q(s',a',h-1)(\widehat{P} - P)(s''|s',a',h-1)q_{(s'',h)}(s,a,h).$$

Plugging these back and changing variables $(s'',a''),s'''$ in χ_1 to $(s',a'),s''$, we get:

$$\begin{aligned}
&\widehat{q}(s,a,h) - q(s,a,h) = \chi_1 + \chi_2 \\
&\qquad = \sum_{(s',a'),s''} \sum_{m=1}^{h-1} q(s',a',m)\left(\widehat{P}(s''|s',a',m) - P(s''|s',a',m)\right)\widehat{q}_{(s'',m+1)}(s,a,h).
\end{aligned}$$

This completes the proof of the first statement. For the second statement,

$$\begin{aligned}
\langle \widehat{q} - q, c\rangle &= \sum_{(s,a),h} \sum_{(s',a'),s''} \sum_{m=1}^{h-1} q(s',a',m)(\widehat{P}(s''|s',a',m) - P(s''|s',a',m))\widehat{q}_{(s'',m+1)}(s,a,h)c(s,a,h) \\
&= \sum_{m=1}^{H} \sum_{(s',a'),s''} q(s',a',m)(\widehat{P}(s''|s',a',m) - P(s''|s',a',m)) \sum_{(s,a),h>m} \widehat{q}_{(s'',m+1)}(s,a,h)c(s,a,h) \\
&= \sum_{m=1}^{H} \sum_{(s',a'),s''} q(s',a',m)(\widehat{P}(s''|s',a',m) - P(s''|s',a',m))J^{\widehat{P},\pi}(s'',m+1)
\end{aligned}$$

$$= \sum_{h=1}^{H} \sum_{(s,a),s'} q(s,a,h)(\widehat{P}(s'|s,a,h) - P(s'|s,a,h)) J^{\widehat{P},\pi}(s',h+1). \qquad \text{(change of variables)}$$

Similarly, $\langle|\widehat{q}-q|, c\rangle \leq \sum_{h=1}^{H} \sum_{(s,a),s'} q(s,a,h) \left|\widehat{P}(s'|s,a,h) - P(s'|s,a,h)\right| J^{\widehat{P},\pi}(s',h+1)$. Applying the fact $J^{\widehat{P},\pi}(s',h+1) \leq H$ and Lemma 6 completes the proof. □

Lemma 8. *With probability at least $1-\delta$ we have: $\sum_{(s,a)} \sum_{k=1}^{K} \frac{q_k(s,a)}{\mathbf{N}_k^+(s,a)} = \tilde{\mathcal{O}}(SAH)$.*

Proof. We first prove that for each (s,a), $\sum_{k=1}^{K} \frac{\widetilde{N}_k(s,a)}{\mathbf{N}_k^+(s,a)} = \mathcal{O}(\ln(HK)+H)$. Indeed, we have

$$\begin{aligned}
\sum_{k=1}^{K} \frac{\widetilde{N}_k(s,a)}{\mathbf{N}_k^+(s,a)} &= \sum_{k=1}^{K} \frac{\widetilde{N}_k(s,a)}{\mathbf{N}_{k+1}^+(s,a)} + \sum_{k=1}^{K} \widetilde{N}_k(s,a) \left(\frac{1}{\mathbf{N}_k^+(s,a)} - \frac{1}{\mathbf{N}_{k+1}^+(s,a)} \right) \\
&\leq \sum_{k=1}^{K} \frac{\widetilde{N}_k(s,a)}{\mathbf{N}_{k+1}^+(s,a)} + H \sum_{k=1}^{K} \left(\frac{1}{\mathbf{N}_k^+(s,a)} - \frac{1}{\mathbf{N}_{k+1}^+(s,a)} \right) \\
&\leq \sum_{k=1}^{K} \frac{\widetilde{N}_k(s,a)}{\mathbf{N}_{k+1}^+(s,a)} + H \\
&= \mathcal{O}(\ln(HK)+H). \qquad (\widetilde{N}_k(s,a) = \mathbf{N}_{k+1}(s,a) - \mathbf{N}_k(s,a))
\end{aligned}$$

Therefore, applying the fact $\mathbb{E}_k[\widetilde{N}_k(s,a)] = q_k(s,a)$ and Lemma 24, we have with probability at least $1-\delta$:

$$\sum_{(s,a)} \sum_{k=1}^{K} \frac{q_k(s,a)}{\mathbf{N}_k^+(s,a)} = \sum_{k=1}^{K} \mathbb{E}_k \left[\sum_{(s,a)} \frac{n_k(s,a)}{\mathbf{N}_k^+(s,a)} \right] \leq 2 \sum_{(s,a)} \sum_{k=1}^{K} \frac{n_k(s,a)}{\mathbf{N}_k^+(s,a)} + \mathcal{O}\left(H \ln \frac{2K}{\delta} \right) = \tilde{\mathcal{O}}(SAH),$$

competing the proof. □

Lemma 9. *Consider interacting with the environment for K episodes, where in episode k the executed policy is $\widetilde{\pi}_k$ and the cost function denoted by $c_k : \widehat{\Gamma} \to [0,1]$ is arbitrary. Also let P_k be any transition function within the confidence set $\mathcal{P}_k$ defined in Eq. (6), and define $q_k = q_{P,\widetilde{\pi}_k}$ and $\widehat{q}_k = q_{P_k,\widetilde{\pi}_k}$. Then with probability at least $1-4\delta$, we have*

$$\begin{aligned}
\sum_{k=1}^{K} |\langle q_k - \widehat{q}_k, c_k \rangle| &\leq 32 \sqrt{S^2 A \ln^2\left(\frac{HKSA}{\delta}\right) \left(\sum_{k=1}^{K} \mathrm{Var}_k \left[\left\langle \widetilde{N}_k, c_k \right\rangle \right] + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right) \right)} + \tilde{\mathcal{O}}\left(H^3 S^2 A\right) \\
&\leq 32 \sqrt{S^2 A \ln^2\left(\frac{HKSA}{\delta}\right) \left(\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right) \right)} + \tilde{\mathcal{O}}\left(H^3 S^2 A\right) \\
&\leq 16 \left(\lambda' S^2 A \left(\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right) \right) + \frac{\ln^2\left(\frac{HKSA}{\delta}\right)}{\lambda'} \right) + \tilde{\mathcal{O}}\left(H^3 S^2 A\right).
\end{aligned}$$

where $\lambda' > 0$ is arbitrary.

Proof. Define $\mu_k(s,a,h) = \sum_{s'} P(s'|s,a,h) J_k^{\widetilde{\pi}_k}(s',h+1)$. Then by Lemma 7:

$$\begin{aligned}
\sum_{k=1}^{K} |\langle q_k - \widehat{q}_k, c_k \rangle| &= \sum_{k=1}^{K} \left| \sum_{(s,a),s',h} q_k(s,a,h)(P(s'|s,a,h) - P_k(s'|s,a,h)) J_k^{P_k,\widetilde{\pi}_k}(s',h+1) \right| \\
&\leq \sum_{k=1}^{K} \left| \sum_{(s,a),s',h} q_k(s,a,h)(P(s'|s,a,h) - P_k(s'|s,a,h)) J_k^{\widetilde{\pi}_k}(s',h+1) \right| + \tilde{\mathcal{O}}\left(H^3 S^2 A\right) \qquad \text{(Lemma 10)}
\end{aligned}$$

$$= \sum_{k=1}^{K} \left| \sum_{(s,a),s',h} q_k(s,a,h)(P(s'|s,a,h) - P_k(s'|s,a,h)) \left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right) \right| + \tilde{\mathcal{O}}\left(H^3 S^2 A\right)$$

($\sum_{s'} P(s'|s,a,h) - P_k(s'|s,a,h) = 0$ and $\mu_k(s,a,h)$ is independent of s')

$$\leq \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} q_k(s,a,h)\epsilon_k^{\star}(s,a,s') \left| J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h) \right| + \tilde{\mathcal{O}}\left(H^3 S^2 A\right)$$

($P(s'|s,a,h) - P_k(s'|s,a,h) \leq \epsilon_k^{\star}(s,a,s')$ by Lemma 6)

$$\leq 8 \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} q_k(s,a,h) \sqrt{P(s'|s,a,h)A_k(s,a)\left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right)^2}$$
$$+ \tilde{\mathcal{O}}\left(HS \sum_{k=1}^{K} \sum_{(s,a)} \frac{q_k(s,a)}{\mathbf{N}_k^{+}(s,a)} + H^3 S^2 A \right)$$

(definition of $\epsilon_k^{\star}$ from Lemma 6)

$$\leq 8 \sum_{k=1}^{K} \mathbb{E}_k \left[\sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h) \sqrt{P(s'|s,a,h)A_k(s,a)\left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right)^2} \right] + \tilde{\mathcal{O}}\left(H^3 S^2 A\right)$$

(Lemma 8)

$$= 8 \sum_{k=1}^{K} \mathbb{E}_k X_k + \tilde{\mathcal{O}}\left(H^3 S^2 A\right).$$

($X_k = \sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h) \sqrt{P(s'|s,a,h)A_k(s,a)\left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right)^2}$)

Note that $0 \leq X_k = \tilde{\mathcal{O}}(\sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h)H) = \tilde{\mathcal{O}}(H^2 S)$. Hence, by Lemma 24, with probability at least $1-\delta$:

$$\sum_{k=1}^{K} \mathbb{E}_k X_k \leq 2 \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h) \sqrt{P(s'|s,a,h)A_k(s,a)\left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right)^2} + \tilde{\mathcal{O}}\left(H^2 S\right)$$
$$\leq 2 \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h) \sqrt{P(s'|s,a,h)A_{k+1}(s,a)\left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right)^2}$$
$$+ 2H^2 \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \left(\sqrt{A_k(s,a)} - \sqrt{A_{k+1}(s,a)}\right) + \tilde{\mathcal{O}}\left(H^2 S\right)$$
$$\leq 2 \sqrt{\sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h)P(s'|s,a,h)\left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right)^2} \sqrt{\sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h)A_{k+1}(s,a)}$$
$$+ \tilde{\mathcal{O}}\left(H^3 S^2 A\right).$$

(Cauchy-Schwarz inequality)

Note that

$$\sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h)A_{k+1}(s,a) = \ln\left(\frac{HKSA}{\delta}\right) \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \frac{\widetilde{N}_k(s,a,h)}{\mathbf{N}_{k+1}(s,a)}$$
$$= S \ln\left(\frac{HKSA}{\delta}\right) \sum_{k=1}^{K} \sum_{(s,a)} \frac{\mathbf{N}_{k+1}(s,a) - \mathbf{N}_k^{+}(s,a)}{\mathbf{N}_{k+1}(s,a)} \leq 2S^2 A \ln^2\left(\frac{HKSA}{\delta}\right).$$

Moreover, define $\mathbb{V}_k(s,a,h) = P(s'|s,a,h)\left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right)^2$, with probability at least $1-\delta$,

$$\sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h)P(s'|s,a,h)\left(J_k^{\widetilde{\pi}_k}(s',h+1) - \mu_k(s,a,h)\right)^2$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} \widetilde{N}_k(s,a,h)\mathbb{V}_k(s,a,h) \le \sum_{k=1}^{K} \sum_{(s,a),h} \widetilde{N}_k(s,a,h)\mathbb{V}_k(s,a,h) \\
&= \sum_{k=1}^{K} \sum_{(s,a),h} q_k(s,a,h)\mathbb{V}_k(s,a,h) + \sum_{k=1}^{K} \sum_{(s,a),h} (\widetilde{N}_k(s,a,h) - q_k(s,a,h))\mathbb{V}_k(s,a,h) \\
&\le 2\sum_{k=1}^{K} \text{Var}\left[\left\langle \widetilde{N}_k, c_k \right\rangle\right] + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right),
\end{aligned}
$$

where in the last inequality we apply Lemma 4 and Lemma 20 with

$$
\sum_{(s,a),h} |\widetilde{N}_k(s,a,h) - q_k(s,a,h)|\mathbb{V}_k(s,a,h) \le \sum_{(s,a),h} \widetilde{N}_k(s,a,h)H^2 + \sum_{(s,a),h} q_k(s,a,h)H^2 \le 2H^3.
$$

Therefore, combining everything we arrive at

$$
\left|\sum_{k=1}^{K} \langle q_k - \widehat{q}_k, c_k\rangle\right| \le 32\sqrt{S^2A\ln^2\left(\frac{HKSA}{\delta}\right)\left(\sum_{k=1}^{K} \text{Var}\left[\left\langle \widetilde{N}_k, c_k \right\rangle\right] + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^2A\right).
$$

When the value of c_k is at most 1, we have by Lemma 2 and $\text{Var}\left[\left\langle \widetilde{N}_k, c_k \right\rangle\right] \le \mathbb{E}_k\left[\left\langle \widetilde{N}_k, c_k \right\rangle^2\right]$:

$$
\begin{aligned}
\left|\sum_{k=1}^{K} \langle q_k - \widehat{q}_k, c_k\rangle\right| &\le 32\sqrt{S^2A\ln^2\left(\frac{HKSA}{\delta}\right)\left(\sum_{k=1}^{K} \left\langle q_k, \vec{h}\circ c_k \right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^2A\right) \\
&\le 16\left(\lambda' S^2A\left(\sum_{k=1}^{K} \left\langle q_k, \vec{h}\circ c_k \right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right)\right) + \frac{\ln^2\left(\frac{HKSA}{\delta}\right)}{\lambda'}\right) + \tilde{\mathcal{O}}\left(H^3S^2A\right),
\end{aligned}
$$

where the last step is by AM-GM inequality. □

Lemma 10. *Under the event of Lemma 5,*

$$
\sum_{k=1}^{K}\left|\sum_{(s,a),s',h} q_k(s,a,h)(P(s'|s,a,h) - P_k(s'|s,a,h))\left(J_k^{P_k,\widetilde{\pi}_k}(s',h+1) - J_k^{P,\widetilde{\pi}_k}(s',h+1)\right)\right| = \tilde{\mathcal{O}}\left(H^3S^2A\right).
$$

Proof. Define $q_{k,(s',h+1)} = q_{P,\widetilde{\pi}_k,(s',h+1)}$. Then,

$$
\begin{aligned}
&\sum_{k=1}^{K}\left|\sum_{(s,a),s',h} q_k(s,a,h)(P(s'|s,a,h) - P_k(s'|s,a,h))\left(J_k^{P_k,\widetilde{\pi}_k}(s',h+1) - J_k^{P,\widetilde{\pi}_k}(s',h+1)\right)\right| \\
&\le \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} q_k(s,a,h)\epsilon_k^\star(s,a,s')\left|\left\langle q_{P_k,\widetilde{\pi}_k,(s',h+1)} - q_{P,\widetilde{\pi}_k,(s',h+1)}, c_k\right\rangle\right|
\end{aligned}
$$

(Lemma 6 and $J_k^{P',\widetilde{\pi}_k}(s',h+1) = \left\langle q_{P',\widetilde{\pi}_k,(s',h+1)}, c_k\right\rangle$)

$$
\le \sum_{k=1}^{K} \sum_{(s,a),s',h\in U} q_k(s,a,h)\epsilon_k^\star(s,a,s') \sum_{(\widetilde{s},\widetilde{a}),\widetilde{s}',h'\in U} q_{k,(s',h+1)}(\widetilde{s},\widetilde{a},h')\epsilon_k^\star(\widetilde{s},\widetilde{a},\widetilde{s}')H
$$

(Lemma 7)

$$
= \tilde{\mathcal{O}}\left(H\sum_{k=1}^{K} \sum_{\substack{(s,a),s',h\in U\\(\widetilde{s},\widetilde{a}),\widetilde{s}',h'\in U}} q_k(s,a,h)\sqrt{\frac{P(s'|s,a,h)}{\mathbf{N}_k^+(s,a)}}\, q_{k,(s',h+1)}(\widetilde{s},\widetilde{a},h')\sqrt{\frac{P(\widetilde{s}'|\widetilde{s},\widetilde{a},h')}{\mathbf{N}_k^+(\widetilde{s},\widetilde{a})}}\right)
$$

$$= \tilde{\mathcal{O}}\left(H \sum_{k=1}^{K} \sum_{\substack{(s,a),s',h\in U\\ (\widetilde{s},\widetilde{a}),\widetilde{s}',h'\in U}} \sqrt{\frac{q_k(s,a,h)P(\widetilde{s}'|\widetilde{s},\widetilde{a},h')q_{k,(s',h+1)}(\widetilde{s},\widetilde{a},h')}{\mathbf{N}_k^+(s,a)}} \sqrt{\frac{q_k(s,a,h)P(s'|s,a,h)q_{k,(s',h+1)}(\widetilde{s},\widetilde{a},h')}{\mathbf{N}_k^+(\widetilde{s},\widetilde{a})}} \right)$$

$$= \tilde{\mathcal{O}}\left(H \sqrt{\sum_{\substack{k,(s,a),s',h\in U\\ (\widetilde{s},\widetilde{a}),\widetilde{s}',h'\in U}} \frac{q_k(s,a,h)P(\widetilde{s}'|\widetilde{s},\widetilde{a},h')q_{k,(s',h+1)}(\widetilde{s},\widetilde{a},h')}{\mathbf{N}_k^+(s,a)}} \sqrt{\sum_{\substack{k,(s,a),s',h\in U\\ (\widetilde{s},\widetilde{a}),\widetilde{s}',h'\in U}} \frac{q_k(s,a,h)P(s'|s,a,h)q_{k,(s',h+1)}(\widetilde{s},\widetilde{a},h')}{\mathbf{N}_k^+(\widetilde{s},\widetilde{a})}} \right)$$

(Cauchy-Schwarz inequality)

$$= \tilde{\mathcal{O}}\left(H \sqrt{HS \sum_{k,(s,a)} \frac{q_k(s,a)}{\mathbf{N}_k^+(s,a)}} \sqrt{HS \sum_{k,(\widetilde{s},\widetilde{a})} \frac{q_k(\widetilde{s},\widetilde{a})}{\mathbf{N}_k^+(\widetilde{s},\widetilde{a})}} \right) = \tilde{\mathcal{O}}\left(H^3S^2A\right). \qquad \text{(Lemma 8)}$$

where in the first square root of the last line we simply sum over $s', h, (\widetilde{s},\widetilde{a}), \widetilde{s}', h'$, and in the second square root of the last line we apply $\sum_{(s,a),s'} q_k(s,a,h)P(s'|s,a,h)q_{k,(s',h+1)}(\widetilde{s},\widetilde{a},h') = q_k(\widetilde{s},\widetilde{a},h')$ for any $\widetilde{s},\widetilde{a},h < h'$. □

B.2. Proof of Theorem 1

Proof. First apply Lemma 1: with probability at least $1-\delta$, we have

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

Define $P_k = P_{\widehat{q}_k}$, and $q_k = q_{P,\widetilde{\pi}_k}$. We decompose the regret in $\widetilde{M}$ into three terms:

$$\sum_{k=1}^{K} \left\langle \widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k \right\rangle = \sum_{k=1}^{K} \left\langle \widetilde{N}_k - q_k, c_k \right\rangle + \sum_{k=1}^{K} \langle q_k - \widehat{q}_k, c_k \rangle + \sum_{k=1}^{K} \langle \widehat{q}_k - q_{\widetilde{\pi}^\star}, c_k \rangle \tag{7}$$

For the first term, by Lemma 21 with $\frac{\lambda}{4} \leq \frac{1}{H}$ (because $K \geq 16S^2AH^2$) and Lemma 2, we have with probability at least $1-\delta$:

$$\sum_{k=1}^{K} \left\langle \widetilde{N}_k - q_k, c_k \right\rangle \leq \frac{\lambda}{2} \sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \frac{4\ln(1/\delta)}{\lambda}.$$

For the second term, by Lemma 9, with probability $1-4\delta$, we have for any $\lambda' > 0$:

$$\sum_{k=1}^{K} \langle q_k - \widehat{q}_k, c_k \rangle \leq 16\left(\lambda' S^2 A \left(\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K} \right) \right) + \frac{\ln^2\left(\frac{HKSA}{\delta}\right)}{\lambda'} \right) + \tilde{\mathcal{O}}\left(H^3S^2A\right).$$

For the third term, by standard OMD analysis (see for example Eq. (12) of (Rosenberg and Mansour, 2020)):

$$\sum_{k=1}^{K} \langle \phi_k - \phi_{\widetilde{\pi}^\star}, c_k \rangle \leq D_\psi(\phi_{\widetilde{\pi}^\star}, \phi_1) + \sum_{k=1}^{K} \left\langle \phi_k - \phi'_{k+1}, c_k \right\rangle, \tag{8}$$

where $\phi'_{k+1} = \operatorname{argmin}_{\phi\in\mathbb{R}^{\tilde{\Gamma}\times\mathcal{S}\times[H]}} \langle \phi, c_k \rangle + D_\psi(\phi, \phi_k)$. It can be shown that $\phi'_{k+1}(s,a,s',h) = \phi_k(s,a,s',h)e^{-\eta c_k(s,a)}$ with the choice of the entropy regularizer. Applying the inequality $1 - e^{-x} \leq x$, we get:

$$\sum_{k=1}^{K} \left\langle \phi_k - \phi'_{k+1}, c_k \right\rangle \leq \eta \sum_{k=1}^{K} \sum_{(s,a),s',h} \phi_k(s,a,s',h)c_k^2(s,a) \leq 2\eta \sum_{k=1}^{K} \sum_{(s,a)} \widehat{q}_k(s,a)c_k(s,a) = 2\eta \sum_{k=1}^{K} \langle \widehat{q}_k, c_k \rangle,$$

where in the last inequality we apply $\phi_k(s,a,s',h) = (1+\lambda h)\widehat{q}_k(s,a,s',h) \le 2\widehat{q}_k(s,a,s',h)$. To bound $D_\psi(\phi_{\widetilde{\pi}^\star}, \phi_1)$, note that $\langle \nabla\psi(\phi_1), \phi_{\widetilde{\pi}^\star} - \phi_1\rangle \ge 0$ by $\phi_1 = \operatorname{argmin}_{\phi\in\Omega_1} \psi(\phi)$. Thus,

$$\begin{aligned}
D_\psi(\phi_{\widetilde{\pi}^\star}, \phi_1) &\le \psi(\phi_{\widetilde{\pi}^\star}) - \psi(\phi_1) \\
&= \frac{1}{\eta}\sum_{(s,a),s',h} \phi_{\widetilde{\pi}^\star}(s,a,s',h)\ln\phi_{\widetilde{\pi}^\star}(s,a,s',h) - \frac{1}{\eta}\sum_{(s,a),s',h}\phi_1(s,a,s',h)\ln\phi_1(s,a,s',h) \\
&\le \frac{1}{\eta}\sum_{(s,a),s',h}\phi_{\widetilde{\pi}^\star}(s,a,s',h)\ln(2T) - \frac{2T}{\eta}\sum_{(s,a),s',h}\frac{\phi_1(s,a,s',h)}{2T}\ln\frac{\phi_1(s,a,s',h)}{2T} \\
&\le \frac{2T\ln(2T)}{\eta} + \frac{2T\ln(S^2AH)}{\eta} \le \frac{2T\ln(2S^2AHT)}{\eta}.
\end{aligned}$$

Substituting these back to Eq. (8) and rearranging terms, we get:

$$\begin{aligned}
\sum_{k=1}^K \langle \widehat{q}_k - q_{\widetilde{\pi}^\star}, c_k\rangle &\le \frac{1}{1-2\eta}\left(\frac{2T\ln(2S^2AHT)}{\eta} + 2\eta\sum_{k=1}^K\langle q_{\widetilde{\pi}^\star}, c_k\rangle + \sum_{k=1}^K \lambda\left\langle q_{\widetilde{\pi}^\star} - \widehat{q}_k, \vec{h}\circ c_k\right\rangle\right) \\
&\le \frac{4T\ln(2S^2AHT)}{\eta} + 4\eta DK + 2\lambda DTK - \lambda\sum_{k=1}^K\left\langle \widehat{q}_k, \vec{h}\circ c_k\right\rangle, \qquad (9)
\end{aligned}$$

where we apply $\frac{1}{1-2\eta} \le 2$, $\sum_{k=1}^K \langle q_{\widetilde{\pi}^\star}, c_k\rangle \le DK$ and $\sum_{k=1}^K \left\langle q_{\widetilde{\pi}^\star}, \vec{h}\circ c_k\right\rangle = \sum_{k=1}^K \langle q_{\widetilde{\pi}^\star}, J_k^{\widetilde{\pi}^\star}\rangle \le DTK$ in the last inequality. Substituting everything back to Eq. (7) and set $\lambda' = \frac{1}{8}\sqrt{\frac{1}{S^2ADTK}}$, we get:

$$\begin{aligned}
\sum_{k=1}^K \left\langle \widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k\right\rangle &\le \tilde{\mathcal{O}}\left(\frac{1}{\lambda} + \frac{1}{\lambda'} + \frac{T}{\eta}\right) + \left(16\lambda' S^2A - \frac{\lambda}{2}\right)\sum_{k=1}^K\left\langle q_k, \vec{h}\circ c_k\right\rangle \\
&\quad + \lambda\sum_{k=1}^K\left\langle q_k - \widehat{q}_k, \vec{h}\circ c_k\right\rangle + 4\eta DK + 2\lambda DTK + \tilde{\mathcal{O}}\left(\lambda' S^2AH^3\sqrt{K} + H^3S^2A\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{S^2ADTK} + \sqrt{DTK} + \lambda\sqrt{S^2AH^2K} + H^3S^2A\right) = \tilde{\mathcal{O}}\left(\sqrt{S^2ADTK} + H^3S^2A\right),
\end{aligned}$$

where in the last line we apply $\eta \le \sqrt{\frac{T}{DK}}$, $\lambda = 4\sqrt{\frac{S^2A}{DTK}}$, and

$$\begin{aligned}
\lambda\sum_{k=1}^K\left\langle q_k - \widehat{q}_k, \vec{h}\circ c_k\right\rangle &= \tilde{\mathcal{O}}\left(\sqrt{\frac{S^2A}{DTK}}\cdot\sqrt{S^2A\left(\sum_{k=1}^K\left\langle q_k, \vec{h}\circ c_k\right\rangle + H^3\sqrt{K}\right)}\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{\frac{S^2A}{K}\cdot S^2A\left(H^2K + H^3\sqrt{K}\right)}\right) = \tilde{\mathcal{O}}\left(H^3S^2A\right).
\end{aligned}$$

□

C. Omitted details for Section 5

In this section, we provide all omitted algorithms and proofs for Section 5. We first prove a bound of the bias induced by our optimistic cost estimator in the bandit setting. Then we gives the full proof of Theorem 2, which has a similar structure to (Chen et al., 2020, Theorem 11).

Lemma 11. *With probability at least $1-4\delta$, we have*

$$\sum_{k=1}^K\langle u_k - q_k, c_k\rangle \le 32\sqrt{S^3A^2\ln^2\left(\frac{HKS^2A^2}{\delta}\right)\left(\sum_{k=1}^K\left\langle q_k, \vec{h}\circ c_k\right\rangle + \tilde{\mathcal{O}}\left(SAH^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^3A^2\right).$$

Algorithm 5 Log-barrier Policy Search for SSP with unknown transition

```
Input: Upper bound on expected hitting time $T$, horizon parameter $H_1$, and confidence level $\delta$
Parameters: $H_2 = \lceil 2D \rceil$, $H = H_1 + H_2$, $C = \lceil \log_2(TK^4) \rceil \lceil \log_2(T^2K^9) \rceil$, $\beta = e^{\frac{1}{7\ln K}}$, $\eta = \sqrt{\frac{SA}{DTK}}$, $\gamma = 280(\ln K)\left(2C\sqrt{\ln\left(\frac{CSA}{\delta}\right)} + 1\right)^2 \eta$, $\lambda = 40\eta + 2\gamma + 33\sqrt{\frac{S^3A^2}{DTK}}$
Define: $\psi_k(\phi) = \sum_{(s,a)\in\widetilde{\Gamma}} \frac{1}{\eta_k(s,a)} \ln \frac{1}{\phi(s,a)}$, where $\phi(s,a) = \sum_{s'\in\mathcal{S}\cup\{s_f\}} \sum_{h=1}^{H} \phi(s,a,s',h)$
Define: $\Omega_k = \{\phi = q + \lambda\vec{h} \circ q : q \in \widetilde{\Delta}(T, \mathcal{P}_{i_k}),\ \ q(s,a) \geq \frac{1}{TK^4},\ \forall (s,a) \in \widetilde{\Gamma}\}$
Initialize: $\phi_1 = \text{argmin}_{\phi\in\Omega_1} \psi_1(\phi)$.
Initialize: for all $(s,a) \in \widetilde{\Gamma}$, $\eta_1(s,a) = \eta$, $\rho_1(s,a) = 2T$.
Initialize: $\mathbf{N}_1(s,a) = \mathbf{M}_1(s,a,s') = 0$ for all $(s,a,s') \in \Gamma \times (\mathcal{S} \cup \{g\})$. An instance of Bernstein-SSP $\mathcal{B}$.
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 $\widetilde{\pi}_k = \pi_{\widehat{q}_k}$.
    Execute policy: $\tau_k = \text{RUN}(\widetilde{\pi}_k, \mathcal{B})$, receive $c_k \odot \mathbb{I}_k$.
    $\mathcal{P}_{k+1}, \widetilde{N}_k = \text{TransEst}(\mathbf{N}, \mathbf{M}, \delta, H_1, H_2, \tau_k)$.
    Construct cost estimator $\widehat{c}_k \in \mathbb{R}^{\widetilde{\Gamma}}_{\geq 0}$ such that $\widehat{c}_k(s,a) = \frac{\widetilde{N}_k(s,a)c_k(s,a)}{u_k(s,a)}$, where $u_k(s,a) = \max_{\widehat{P}\in\mathcal{P}_k} q_{\widehat{P},\widetilde{\pi}_k}(s,a)$.
    Construct bias term $\widehat{b}_k \in \mathbb{R}^{\widetilde{\Gamma}}_{\geq 0}$ such that $\widehat{b}_k(s,a) = \frac{\sum_h hu'_k(s,a,h)\widehat{c}_k(s,a)}{u_k(s,a)}$ where $u'_k(s,a,h) = q_{P_k^{(s,a)},\widetilde{\pi}_k}(s,a,h)$ and $P_k^{(s,a)} = \text{argmax}_{\widehat{P}\in\mathcal{P}_k} \sum_h h \cdot q_{\widehat{P},\widetilde{\pi}_k}(s,a,h)$.
    Update
        $\phi_{k+1} = \underset{\phi\in\Omega_{k+1}}{\text{argmin}} \left\langle \phi, \widehat{c}_k - \gamma\widehat{b}_k \right\rangle + D_{\psi_k}(\phi, \phi_k).$
    for $\forall (s,a) \in \widetilde{\Gamma}$ do
        if $\frac{1}{u_{k+1}(s,a)} > \rho_k(s,a)$ then
            $\rho_{k+1}(s,a) = \frac{2}{u_{k+1}(s,a)}$, $\eta_{k+1}(s,a) = \beta \cdot \eta_k(s,a)$.
        else
            $\rho_{k+1}(s,a) = \rho_k(s,a)$, $\eta_{k+1}(s,a) = \eta_k(s,a)$.
```

Proof. Denote $c_k^{(s,a)}(s',a') = c_k(s',a')\mathbb{I}\{s' = s, a' = a\}$. Then by Lemma 9 (with δ there set to $\delta/|\Gamma|$) and a union bound over all (s,a), we have with probability $1 - 4\delta$, for any $(s,a) \in \Gamma$:

$$
\begin{aligned}
\sum_{k=1}^{K} (u_k(s,a) - q_k(s,a))c_k(s,a) &= \sum_{k=1}^{K} \left\langle u_k - q_k, c_k^{(s,a)} \right\rangle \\
&\leq 32\sqrt{S^2 A \ln^2\left(\frac{HKS^2A^2}{\delta}\right)\left(\sum_{k=1}^{K}\left\langle q_k, \vec{h} \circ c_k^{(s,a)}\right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^2A\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\sum_{k=1}^{K} \langle u_k - q_k, c_k \rangle &\leq 32 \sum_{(s,a)} \sqrt{S^2 A \ln^2\left(\frac{HKS^2A^2}{\delta}\right)\left(\sum_{k=1}^{K}\left\langle q_k, \vec{h} \circ c_k^{(s,a)}\right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^3A^2\right) \\
&\leq 32\sqrt{S^3 A^2 \ln^2\left(\frac{HKS^2A^2}{\delta}\right)\sum_{(s,a)}\left(\sum_{k=1}^{K}\left\langle q_k, \vec{h} \circ c_k^{(s,a)}\right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^3A^2\right) \\
&= 32\sqrt{S^3 A^2 \ln^2\left(\frac{HKS^2A^2}{\delta}\right)\left(\sum_{k=1}^{K}\left\langle q_k, \vec{h} \circ c_k\right\rangle + \tilde{\mathcal{O}}\left(SAH^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^3A^2\right),
\end{aligned}
$$

where in the second line we use Cauchy-Schwarz inequality. □

Below we present the proof of Theorem 2. It decomposes the regret into several terms, each of which is bounded by a lemma included after the proof.

Proof of Theorem 2. By Lemma 1, with probability at least $1-\delta$,

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

We define a slightly perturbed benchmark $q^\star = (1-\frac{1}{TK})q_{\widetilde{\pi}^\star} + \frac{1}{TK}q_0 \in \widetilde{\Delta}(T, \{P\})$ (note that $\widetilde{\Delta}(T,\{P\}) \subseteq \widetilde{\Delta}(T, \mathcal{P}_{i_k}), \forall k$ under the event of Lemma 5) for some $q_0 \in \widetilde{\Delta}(T,\{P\})$ with $q_0(s,a) \ge \frac{1}{K^3}$ for all $(s,a) \in \widetilde{\Gamma}$, so that $\phi^\star = q^\star + \lambda \vec{h} \circ q^\star \in \Omega_k, \forall k$. Also define $b_k \in \mathbb{R}^{\widetilde{\Gamma}}$ such that $b_k(s,a) = \frac{\sum_h h u'_k(s,a,h) c_k(s,a)}{u_k(s,a)}$, which clearly satisfies $\mathbb{E}_k[\widehat{b}_k] = b_k$. We then decompose $\sum_{k=1}^K \left\langle \widetilde{N}_k - q^\star, c_k \right\rangle$ as

$$\begin{aligned}
&\sum_{k=1}^{K} \left\langle \widetilde{N}_k - q^\star, c_k \right\rangle \\
&= \sum_{k=1}^{K} \langle u_k, \widehat{c}_k \rangle - \sum_{k=1}^{K} \langle q^\star, c_k \rangle && (\langle \widetilde{N}_k, c_k \rangle = \langle u_k, \widehat{c}_k \rangle) \\
&= \sum_{k=1}^{K} \langle u_k - \widehat{q}_k, \widehat{c}_k \rangle + \sum_{k=1}^{K} \langle \widehat{q}_k, \widehat{c}_k \rangle - \langle q^\star, c_k \rangle \\
&= \textsc{Err}_1 + \sum_{k=1}^{K} \langle \phi_k - \phi^\star, \widehat{c}_k \rangle + \sum_{k=1}^{K} \langle \phi^\star, \widehat{c}_k - c_k \rangle + \lambda \sum_{k=1}^{K} \left\langle \vec{h} \circ q^\star, c_k \right\rangle - \lambda \sum_{k=1}^{K} \left\langle \vec{h} \circ \widehat{q}_k, \widehat{c}_k \right\rangle \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left(\text{define } \textsc{Err}_1 = \textstyle\sum_{k=1}^K \langle u_k - \widehat{q}_k, \widehat{c}_k \rangle\right) \\
&= \textsc{Err}_1 + \sum_{k=1}^{K} \langle \phi_k - \phi^\star, \widehat{c}_k \rangle + \sum_{k=1}^{K} \langle \phi^\star, \widehat{c}_k - c_k \rangle + \tilde{\mathcal{O}}\left(\lambda DTK\right) - \lambda \sum_{k=1}^{K} \left\langle \vec{h} \circ \widehat{q}_k, \widehat{c}_k \right\rangle \\
&= \sum_{k=1}^{K} \langle \phi_k - \phi^\star, \widehat{c}_k \rangle + \tilde{\mathcal{O}}\left(\lambda DTK\right) + \textsc{Err}_1 + \textsc{Bias}_1 + \textsc{Bias}_2 - \lambda \sum_{k=1}^{K} \left\langle \vec{h} \circ \widehat{q}_k, c_k \right\rangle \\
&\qquad\qquad \left(\text{define } \textsc{Bias}_1 = \textstyle\sum_{k=1}^K \langle \phi^\star, \widehat{c}_k - c_k \rangle \text{ and } \textsc{Bias}_2 = \lambda \sum_{k=1}^K \left\langle \vec{h} \circ \widehat{q}_k, c_k - \widehat{c}_k \right\rangle\right) \\
&= \textsc{Reg}_\phi + \tilde{\mathcal{O}}\left(\lambda DTK\right) + \textsc{Err}_1 + \textsc{Bias}_1 + \textsc{Bias}_2 + \gamma \sum_{k=1}^{K} \left\langle \phi_k - \phi^\star, \widehat{b}_k \right\rangle - \lambda \sum_{k=1}^{K} \left\langle \vec{h} \circ \widehat{q}_k, c_k \right\rangle \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left(\text{define } \textsc{Reg}_\phi = \textstyle\sum_{k=1}^K \langle \phi_k - \phi^\star, \widehat{c}_k - \gamma \widehat{b}_k \rangle\right) \\
&= \textsc{Reg}_\phi + \tilde{\mathcal{O}}\left(\lambda DTK\right) + \textsc{Err}_1 + \textsc{Bias}_1 + \textsc{Bias}_2 + \textsc{Bias}_3 + \textsc{Bias}_4 + \gamma \sum_{k=1}^{K} \langle \phi_k - \phi^\star, b_k \rangle - \lambda \sum_{k=1}^{K} \left\langle \vec{h} \circ \widehat{q}_k, c_k \right\rangle \\
&\qquad\qquad \left(\text{define } \textsc{Bias}_3 = \gamma \textstyle\sum_{k=1}^K \langle \phi_k, \widehat{b}_k - b_k \rangle \text{ and } \textsc{Bias}_4 = \gamma \sum_{k=1}^K \langle \phi^\star, b_k - \widehat{b}_k \rangle\right) \\
&\le \textsc{Reg}_\phi + \tilde{\mathcal{O}}\left(\lambda DTK\right) + \textsc{Err}_1 + \textsc{Bias}_1 + \textsc{Bias}_2 + \textsc{Bias}_3 + \textsc{Bias}_4 \\
&\qquad\qquad + 2\gamma \sum_{k=1}^{K} \langle \widehat{q}_k, b_k \rangle - \gamma \sum_{k=1}^{K} \langle \phi^\star, b_k \rangle - \lambda \sum_{k=1}^{K} \left\langle \vec{h} \circ \widehat{q}_k, c_k \right\rangle \\
&\le \textsc{Reg}_\phi + \tilde{\mathcal{O}}\left(\lambda DTK\right) + \textsc{Err}_1 + \textsc{Bias}_1 + \textsc{Bias}_2 + \textsc{Bias}_3 + \textsc{Bias}_4 \\
&\qquad\qquad + (2\gamma - \lambda) \sum_{k=1}^{K} \left\langle \widehat{q}_k, \vec{h} \circ c_k \right\rangle + 2\gamma \sum_{k=1}^{K} \left\langle u'_k - \widehat{q}_k, \vec{h} \circ c_k \right\rangle - \gamma \sum_{k=1}^{K} \langle \phi^\star, b_k \rangle \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left(\langle \widehat{q}_k, b_k \rangle \le \langle u_k, b_k \rangle = \left\langle u'_k, \vec{h} \circ c_k \right\rangle\right) \\
&= \textsc{Reg}_\phi + \tilde{\mathcal{O}}\left(\lambda DTK\right) + \textsc{Err}_1 + \textsc{Err}_2 + \textsc{Bias}_1 + \textsc{Bias}_2 + \textsc{Bias}_3 + \textsc{Bias}_4
\end{aligned}$$

$$+ (2\gamma - \lambda)\sum_{k=1}^{K}\left\langle \widehat{q}_k, \vec{h}\circ c_k \right\rangle - \gamma\sum_{k=1}^{K}\left\langle \phi^\star, b_k \right\rangle .$$

(define $\textsc{Err}_2 = 2\gamma\sum_{k=1}^{K}\left\langle u'_k - \widehat{q}_k, \vec{h}\circ c_k \right\rangle$)

The $\textsc{Reg}_\phi$ term can be upper bounded by the OMD analysis (see Lemma 12), the four bias terms $\textsc{Bias}_1, \textsc{Bias}_2, \textsc{Bias}_3$, and $\textsc{Bias}_4$ can be bounded using Azuma's or Freedman's inequality (see Lemma 13 and Lemma 14), and $\textsc{Err}_1 + \textsc{Err}_2$ can be bounded by Lemma 11 (see Lemma 15). Combining everything, we obtain

$$\begin{aligned}
R_K &\le \tilde{\mathcal{O}}\left(\frac{SA}{\eta}\right) - \frac{\langle \phi^\star, \rho_K\rangle}{140\eta\ln K} + 40\eta\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k \right\rangle + \tilde{\mathcal{O}}\left(\lambda DTK\right) \\
&\quad + 33\lambda'\left(\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k \right\rangle + \tilde{\mathcal{O}}\left(SAH^3\sqrt{K}\right)\right) + \tilde{\mathcal{O}}\left(\frac{S^3A^2}{\lambda'}\right) \\
&\quad + 2C\sqrt{\ln\left(\frac{CSA}{\delta}\right)}\left(\frac{\langle \phi^\star, \rho_K\rangle}{\eta'} + \eta'\left\langle \phi^\star, \sum_{k=1}^{K} b_k \right\rangle\right) + 2CH\ln\left(\frac{CSA}{\delta}\right)\langle \phi^\star, \rho_K\rangle \\
&\quad + \left(\frac{1}{\eta'} + 1\right)\langle \phi^\star, \rho_K\rangle + \eta'\left\langle \phi^\star, \sum_{k=1}^{K} b_k \right\rangle + (2\gamma - \lambda)\sum_{k=1}^{K}\left\langle \widehat{q}_k, \vec{h}\circ c_k \right\rangle - \gamma\sum_{k=1}^{K}\langle \phi^\star, b_k\rangle + \tilde{\mathcal{O}}\left(H^3S^3A^2\right) \\
&= \tilde{\mathcal{O}}\left(\frac{SA}{\eta} + \lambda DTK + \lambda' SAH^3\sqrt{K} + \frac{S^3A^2}{\lambda'} + H^3S^3A^2\right) + (40\eta + 33\lambda' + 2\gamma - \lambda)\sum_{k=1}^{K}\left\langle \widehat{q}_k, \vec{h}\circ c_k \right\rangle \\
&\quad + \left(\frac{2C\sqrt{\ln\left(\frac{CSA}{\delta}\right)} + 1}{\eta'} + 2CH\ln\left(\frac{CSA}{\delta}\right) + 1 - \frac{1}{140\eta\ln K}\right)\langle \phi^\star, \rho_K\rangle \\
&\quad + \left(2C\sqrt{\ln\left(\frac{CSA}{\delta}\right)}\eta' + \eta' - \gamma\right)\left\langle \phi^\star, \sum_{k=1}^{K} b_k \right\rangle + (40\eta + 33\lambda')\sum_{k=1}^{K}\left\langle q_k - \widehat{q}_k, \vec{h}\circ c_k \right\rangle .
\end{aligned}$$

Finally, taking $\eta' = \gamma / \left(2C\sqrt{\ln\left(\frac{CSA}{\delta}\right)} + 1\right), \lambda' = \sqrt{\frac{S^3A^2}{DTK}}$, and noticing $\frac{1}{\eta'} \ge 2CH\ln\left(\frac{CSA}{\delta}\right) + 1$, we have the coefficients multiplying $\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k \right\rangle, \langle \phi^\star, \rho_K\rangle$, and $\left\langle \phi^\star, \sum_{k=1}^{K} b_k \right\rangle$ are all non-positive. Moreover, by Lemma 9, $\frac{1}{H}(\vec{h}\circ c_k)(s,a) \le c_k(s,a)$, and $\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k \right\rangle \le H^2K$:

$$\begin{aligned}
(40\eta + 33\lambda')\sum_{k=1}^{K}\left\langle q_k - \widehat{q}_k, \vec{h}\circ c_k \right\rangle &= (40\eta + 33\lambda')H\sum_{k=1}^{K}\left\langle q_k - \widehat{q}_k, \frac{\vec{h}\circ c_k}{H} \right\rangle \\
&= \tilde{\mathcal{O}}\left((40\eta + 33\lambda')H\sqrt{S^2A(H^2K + H^3\sqrt{K})}\right) = \tilde{\mathcal{O}}\left(H^3S^3A^2\right).
\end{aligned}$$

Thus, we arrive at $R_K = \tilde{\mathcal{O}}\left(\sqrt{S^3A^2DTK} + H^3S^3A^2\right)$. □

Lemma 12. *Algorithm 5 ensures with probability at least* $1 - \delta$*:*

$$\textsc{Reg}_\phi \le \tilde{\mathcal{O}}\left(\frac{SA}{\eta}\right) - \frac{\langle \phi^\star, \rho_K\rangle}{140\eta\ln K} + 40\eta\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k \right\rangle + \tilde{\mathcal{O}}\left(H^2\sqrt{SA}\right).$$

Proof. Denote by $n(s,a)$ the number of times the learning rate for (s,a) increases, such that $\eta_K(s,a) = \eta\beta^{n(s,a)}$, and by $k_1, \ldots, k_{n(s,a)}$ the episodes where $\eta_k(s,a)$ is increased, such that $\eta_{k_t+1}(s,a) = \beta\cdot\eta_{k_t}(s,a)$. Since $\rho_1(s,a) = 2T$ and

$$\rho_1(s,a)2^{n(s,a)-1} \le \cdots \le \rho_{k_{n(s,a)}}(s,a) \le \frac{1}{u_{k_{n(s,a)}+1}(s,a)} \le \frac{1}{\widehat{q}_{k_{n(s,a)}+1}(s,a)} \le TK^4,$$

we have $n(s,a) \leq 1 + \log_2 \frac{K^4}{2} \leq 7\log_2 K$. Therefore, $\eta_K(s,a) \leq \eta e^{\frac{7\log_2 K}{7\ln K}} \leq 5\eta$.

Now, notice that $\sum_h u'_k(s,a,h) \leq u_k(s,a)$ by definition and $\gamma \leq \frac{1}{H}$ for large enough K ($K \gtrsim S^3A^2H^2$). Thus,

$$\gamma\widehat{b}_k(s,a) \leq \frac{\gamma H \sum_h u'_k(s,a,h)\widehat{c}_k(s,a)}{u_k(s,a)} \leq \gamma H\widehat{c}_k(s,a) \leq \widehat{c}_k(s,a).$$

This means that the cost $\widehat{c}_k - \gamma\widehat{b}_k$ we feed to OMD is always non-negative, and thus by the same argument of (Agarwal et al., 2017, Lemma 12), we have

$$\begin{aligned}
&\textsc{Reg}_\phi = \sum_{k=1}^K \left\langle \phi_k - \phi^\star, \widehat{c}_k - \gamma\widehat{b}_k \right\rangle \\
&\leq \sum_{k=1}^K D_{\psi_k}(\phi^\star,\phi_k) - D_{\psi_k}(\phi^\star,\phi_{k+1}) + \sum_{k=1}^K\sum_{(s,a)} \eta_k(s,a)\phi_k^2(s,a)(\widehat{c}_k(s,a) - \gamma\widehat{b}_k(s,a))^2 \\
&\leq D_{\psi_1}(\phi^\star,\phi_1) + \sum_{k=1}^{K-1}\left(D_{\psi_{k+1}}(\phi^\star,\phi_{k+1}) - D_{\psi_k}(\phi^\star,\phi_{k+1})\right) + 5\eta\sum_{k=1}^K\sum_{(s,a)}\phi_k^2(s,a)\widehat{c}_k^2(s,a) \\
&\leq D_{\psi_1}(\phi^\star,\phi_1) + \sum_{k=1}^{K-1}\left(D_{\psi_{k+1}}(\phi^\star,\phi_{k+1}) - D_{\psi_k}(\phi^\star,\phi_{k+1})\right) + 20\eta\sum_{k=1}^K\sum_{(s,a)}\widehat{q}_k^2(s,a)\widehat{c}_k^2(s,a) \\
&\leq D_{\psi_1}(\phi^\star,\phi_1) + \sum_{k=1}^{K-1}\left(D_{\psi_{k+1}}(\phi^\star,\phi_{k+1}) - D_{\psi_k}(\phi^\star,\phi_{k+1})\right) + 20\eta\sum_{k=1}^K\sum_{(s,a)}\widetilde{N}_k^2(s,a)c_k^2(s,a). \quad (u_k(s,a) \geq \widehat{q}_k(s,a))
\end{aligned}$$

For the first term, since ϕ_1 minimizes ψ_1 and thus $\langle\nabla\psi_1(\phi_1), \phi^\star - \phi_1\rangle \geq 0$, we have

$$D_{\psi_1}(\phi^\star,\phi_1) \leq \psi_1(\phi^\star) - \psi_1(\phi_1) = \frac{1}{\eta}\sum_{(s,a)}\ln\frac{\phi_1(s,a)}{\phi^\star(s,a)} \leq \frac{1}{\eta}\sum_{(s,a)}\ln\frac{2H}{q^\star(s,a)} = \tilde{\mathcal{O}}\left(\frac{SA}{\eta}\right).$$

For the second term, we define $\chi(y) = y - 1 - \ln y$ and proceed similarly to (Agarwal et al., 2017):

$$\begin{aligned}
&\sum_{k=1}^{K-1} D_{\psi_{k+1}}(\phi^\star,\phi_{k+1}) - D_{\psi_k}(\phi^\star,\phi_{k+1}) \\
&= \sum_{k=1}^{K-1}\sum_{(s,a)}\left(\frac{1}{\eta_{k+1}(s,a)} - \frac{1}{\eta_k(s,a)}\right)\chi\left(\frac{\phi^\star(s,a)}{\phi_{k+1}(s,a)}\right) \\
&\leq \sum_{(s,a)}\frac{1-\beta}{\eta\beta^{n(s,a)}}\chi\left(\frac{\phi^\star(s,a)}{\phi_{k_{n(s,a)}+1}(s,a)}\right) \\
&= \sum_{(s,a)}\frac{1-\beta}{\eta\beta^{n(s,a)}}\left(\frac{\phi^\star(s,a)}{\phi_{k_{n(s,a)}+1}(s,a)} - 1 - \ln\frac{\phi^\star(s,a)}{\phi_{k_{n(s,a)}+1}(s,a)}\right) \\
&\leq -\frac{1}{35\eta\ln K}\sum_{(s,a)}\left(\phi^\star(s,a)\frac{\rho_K(s,a)}{4} - 1 - \ln\frac{\phi^\star(s,a)}{\phi_{k_{n(s,a)}+1}(s,a)}\right) \\
&\leq \frac{SA(1+6\ln K)}{35\eta\ln K} - \frac{\langle\phi^\star,\rho_K\rangle}{140\eta\ln K} = \tilde{\mathcal{O}}\left(\frac{SA}{\eta}\right) - \frac{\langle\phi^\star,\rho_K\rangle}{140\eta\ln K},
\end{aligned}$$

where in the last two lines we use the facts $1-\beta \leq -\frac{1}{7\ln K}$, $\beta^{n(s,a)} \leq 5$, $\rho_K(s,a) = \frac{2}{u_{k_{n(s,a)}+1}(s,a)}$, and $\ln\frac{\phi^\star(s,a)}{\phi_{k_{n(s,a)}+1}(s,a)} \leq \ln(HTK^4) \leq 6\ln K$.

Finally, for the third term, since $\sum_{(s,a)} \widetilde{N}_k^2(s,a)c_k^2(s,a) \le \left(\sum_{(s,a)} \widetilde{N}_k(s,a)\right)^2 \le H^2$, we apply Azuma's inequality (Lemma 20) and obtain, with probability at least $1-\delta$:

$$\begin{aligned}
\eta\sum_{k=1}^{K}\sum_{(s,a)} \widetilde{N}_k^2(s,a)c_k^2(s,a) &\le \eta\sum_{k=1}^{K}\mathbb{E}_k\left[\sum_{(s,a)} \widetilde{N}_k^2(s,a)c_k^2(s,a)\right] + \tilde{\mathcal{O}}\left(\eta H^2\sqrt{K}\right) \\
&\le \eta\sum_{k=1}^{K}\mathbb{E}_k\left[\left\langle \widetilde{N}_k, c_k\right\rangle^2\right] + \tilde{\mathcal{O}}\left(H^2\sqrt{SA}\right) \\
&\qquad\qquad\qquad (\textstyle\sum_i a_i^2 \le (\sum_i a_i)^2 \text{ for } a_i > 0 \text{ and } \eta = \sqrt{\frac{SA}{DTK}}) \\
&\le 2\eta\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k\right\rangle + \tilde{\mathcal{O}}\left(H^2\sqrt{SA}\right). \qquad \text{(Lemma 2)}
\end{aligned}$$

Combining everything shows

$$\textsc{Reg}_\phi \le \tilde{\mathcal{O}}\left(\frac{SA}{\eta}\right) - \frac{\langle\phi^\star, \rho_K\rangle}{140\eta\ln K} + 40\eta\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k\right\rangle + \tilde{\mathcal{O}}\left(H^2\sqrt{SA}\right).$$

finishing the proof. □

Lemma 13. *For any $\eta' > 0$, with probability at least $1-\delta$,*

$$\textsc{Bias}_1 \le 2C\sqrt{\ln\left(\frac{CSA}{\delta}\right)}\left(\frac{\langle\phi^\star, \rho_K\rangle}{\eta'} + \eta'\left\langle \phi^\star, \sum_{k=1}^{K} b_k\right\rangle\right) + 2CH\ln\left(\frac{CSA}{\delta}\right)\langle\phi^\star, \rho_K\rangle.$$

Also, with probability at least $1-5\delta$, $\textsc{Bias}_2 = \tilde{\mathcal{O}}\left(S^3A^2H^2\right)$.

Proof. Define $X_k(s,a) = \widehat{c}_k(s,a) - \mathbb{E}_k[\widehat{c}_k(s,a)]$. Note that $X_k(s,a) \le \frac{H}{u_k(s,a)} \le HTK^4$, and

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}_k\left[X_k^2(s,a)\right] &\le \sum_{k=1}^{K}\frac{\mathbb{E}_k\left[\widetilde{N}_k^2(s,a)c_k^2(s,a)\right]}{u_k^2(s,a)} \\
&\le \rho_k(s,a)\sum_{k=1}^{K}\frac{\mathbb{E}_k\left[\widetilde{N}_k^2(s,a)c_k^2(s,a)\right]}{u_k(s,a)} \\
&\le 2\rho_k(s,a)\sum_{k=1}^{K} b_k(s,a). \qquad \text{(Lemma 16)}
\end{aligned}$$

Therefore, by applying a strengthened Freedman's inequality (Lemma 23) with $b = HTK^4$, $B_k = H\rho_k(s,a), \max_k B_k = H\rho_K(s,a)$, and $V = 2\rho_k(s,a)\sum_{k=1}^{K} b_k(s,a)$, we have with probability at least $1-\delta/(SA)$,

$$\begin{aligned}
&\sum_{k=1}^{K}\widehat{c}_k(s,a) - \mathbb{E}_k[\widehat{c}_k(s,a)] \\
&\le C\left(4\sqrt{\rho_K(s,a)\sum_{k=1}^{K} b_k(s,a)\ln\left(\frac{CSA}{\delta}\right)} + 2H\rho_K(s,a)\ln\left(\frac{CSA}{\delta}\right)\right) \\
&\le 2C\sqrt{\ln\left(\frac{CSA}{\delta}\right)}\left(\frac{\rho_K(s,a)}{\eta'} + \eta'\sum_{k=1}^{K} b_k(s,a)\right) + 2CH\rho_K(s,a)\ln\left(\frac{CSA}{\delta}\right),
\end{aligned}$$

where the last step is by AM-GM inequality. Further using a union bound shows that the above holds for all $(s,a) \in \widetilde{\Gamma}$ with probability at least $1-\delta$. Thus, by $\mathbb{E}_k[\widehat{c}_k(s,a)] \leq c_k(s,a)$,

$$\begin{aligned}\textsc{Bias}_1 &= \sum_{k=1}^K \langle \phi^\star, \widehat{c}_k - c_k \rangle \leq \sum_{k=1}^K \langle \phi^\star, \widehat{c}_k - \mathbb{E}_k[\widehat{c}_k] \rangle \\ &\leq 2C\sqrt{\ln\left(\frac{CSA}{\delta}\right)} \left(\frac{\langle \phi^\star, \rho_K \rangle}{\eta'} + \eta' \left\langle \phi^\star, \sum_{k=1}^K b_k \right\rangle \right) + 2CH \ln\left(\frac{CSA}{\delta}\right) \langle \phi^\star, \rho_K \rangle .\end{aligned}$$

To bound $\textsc{Bias}_2$, simply note that $\left|\left\langle \vec{h} \circ \widehat{q}_k, \mathbb{E}_k[\widehat{c}_k] - \widehat{c}_k \right\rangle\right| \leq 2H^2$ and apply Azuma's inequality (Lemma 20): with probability $1-5\delta$,

$$\begin{aligned}\textsc{Bias}_2 &= \lambda \sum_{k=1}^K \left\langle \vec{h} \circ \widehat{q}_k, c_k - \mathbb{E}_k[\widehat{c}_k] \right\rangle + \lambda \sum_{k=1}^K \left\langle \vec{h} \circ \widehat{q}_k, \mathbb{E}_k[\widehat{c}_k] - \widehat{c}_k \right\rangle \\ &\leq \lambda H \sum_{k=1}^K \langle u_k - q_k, c_k \rangle + \tilde{\mathcal{O}}\left(\lambda H^2 \sqrt{K}\right) \\ &= \tilde{\mathcal{O}}\left(\lambda H \sqrt{S^3A^2 \sum_{k=1}^K \left\langle q_k, \vec{h} \circ c_k \right\rangle} + H^2\sqrt{S^3A^2} \right) = \tilde{\mathcal{O}}\left(S^3A^2H^2\right), \\ &\qquad\qquad \left(\lambda = \tilde{\mathcal{O}}\left(\sqrt{\tfrac{S^3A^2}{DTK}}\right), \textstyle\sum_{k=1}^K \left\langle q_k, \vec{h} \circ c_k \right\rangle = \mathcal{O}\left(H^2K\right) \text{ and Lemma 11}\right)\end{aligned}$$

where in the second line we use:

$$\begin{aligned}\sum_{k=1}^K \left\langle \vec{h} \circ \widehat{q}_k, c_k - \mathbb{E}_k[\widehat{c}_k] \right\rangle &= \sum_{k=1}^K \sum_{(s,a),h} h \cdot \widehat{q}_k(s,a,h) c_k(s,a) \left(1 - \frac{q_k(s,a)}{u_k(s,a)}\right) \\ &\leq H \sum_{k=1}^K \sum_{(s,a),h} \widehat{q}_k(s,a,h) c_k(s,a) \frac{u_k(s,a) - q_k(s,a)}{u_k(s,a)} \\ &\leq H \sum_{k=1}^K \langle u_k - q_k, c_k \rangle . \qquad (u_k(s,a) \geq \max\{q_k(s,a), \widehat{q}_k(s,a)\})\end{aligned}$$

□

Lemma 14. *With probability at least* $1-\delta$*, we have* $\textsc{Bias}_3 = \tilde{\mathcal{O}}\left(H^2(SA)^{3/2}\right)$*. Also, for any* $\eta' > 0$*, with probability at least* $1-\delta$*, we have*

$$\textsc{Bias}_4 \leq \left(\frac{1}{\eta'} + 1\right) \langle \phi^\star, \rho_K \rangle + \eta' \left\langle \phi^\star, \sum_{k=1}^K b_k \right\rangle + \tilde{\mathcal{O}}(1) .$$

Proof. To bound $\textsc{Bias}_3$, simply note that $\mathbb{E}_k[\widehat{b}_k(s,a)] \leq b_k(s,a)$,

$$\begin{aligned}&\left|\left\langle \phi_k, \widehat{b}_k - \mathbb{E}_k[\widehat{b}_k] \right\rangle\right| \leq \left|\left\langle \phi_k, \widehat{b}_k \right\rangle\right| + \left|\left\langle \phi_k, \mathbb{E}_k\widehat{b}_k \right\rangle\right| \\ &\leq 2\left|\sum_{(s,a)}\sum_h h \cdot u'_k(s,a,h)\widehat{c}_k(s,a)\right| + 2\left|\sum_{(s,a)}\sum_h h \cdot u'_k(s,a,h)\mathbb{E}_k[\widehat{c}_k(s,a)]\right| \qquad (\phi_k(s,a) \leq 2\widehat{q}_k(s,a) \leq 2u_k(s,a)) \\ &\leq 2H\left|\sum_{(s,a)} \widetilde{N}_k(s,a)c_k(s,a)\right| + 2H\left|\sum_{(s,a)}\sum_h u'_k(s,a,h)c_k(s,a)\right| \leq 4SAH^2, \\ &\qquad \left(\textstyle\sum_h u'_k(s,a,h) \leq u_k(s,a), q_k(s,a) \leq u_k(s,a), \sum_{(s,a)} \widetilde{N}_k(s,a) \leq H, \text{ and } \sum_h u'_k(s,a,h) \leq H\right)\end{aligned}$$

$\gamma = \tilde{\mathcal{O}}\left(\sqrt{\frac{SA}{DTK}}\right)$, and apply Azuma's inequality (Lemma 20): with probability at least $1-\delta$,

$$\textsc{Bias}_3 = \gamma\sum_{k=1}^K \langle \phi_k, \widehat{b}_k - b_k\rangle \le \gamma\sum_{k=1}^K \left\langle \phi_k, \widehat{b}_k - \mathbb{E}_k[\widehat{b}_k]\right\rangle = \tilde{\mathcal{O}}\left(\gamma SAH^2\sqrt{K}\right) = \tilde{\mathcal{O}}\left(H^2(SA)^{3/2}\right).$$

To bound $\textsc{Bias}_4 = \gamma\sum_{k=1}^K \left\langle \phi^\star, b_k - \mathbb{E}_k[\widehat{b}_k]\right\rangle + \gamma\sum_{k=1}^K \left\langle \phi^\star, \mathbb{E}_k[\widehat{b}_k] - \widehat{b}_k\right\rangle$, first note that

$$\begin{aligned}
\gamma\sum_{k=1}^K \left\langle \phi^\star, b_k - \mathbb{E}_k[\widehat{b}_k]\right\rangle &\le \gamma\sum_{k=1}^K\sum_{(s,a)} \langle \phi^\star, b_k\rangle = \gamma\sum_{k=1}^K\sum_{(s,a)} \phi^\star(s,a)\frac{\sum_{h=1}^H h\cdot u_k'(s,a,h)}{u_k(s,a)}c_k(s,a)\\
&\le \gamma H\sum_{k=1}^K\sum_{(s,a)}\phi^\star(s,a) \le \sum_{k=1}^K \langle\phi^\star, \rho_K\rangle. \qquad (\gamma H\le 1 \text{ and } \rho_K(s,a)\ge\rho_1(s,a)\ge 1)
\end{aligned}$$

Then, we note that $\mathbb{E}_k[\widehat{b}_k(s,a)] - \widehat{b}_k(s,a) \le \mathbb{E}_k[\widehat{b}_k(s,a)] \le H$, and

$$\begin{aligned}
\mathbb{E}_k\left[\left(\mathbb{E}_k[\widehat{b}_k(s,a)] - \widehat{b}_k(s,a)\right)^2\right] &\le \mathbb{E}_k[\widehat{b}_k^2(s,a)] \le \frac{\mathbb{E}_k\left[(\sum_h h\cdot u_k'(s,a,h)\widehat{c}_k(s,a))^2\right]}{u_k^2(s,a)}\\
&\le \frac{H^2\mathbb{E}_k\left[\widetilde{N}_k^2(s,a)c_k^2(s,a)\right]}{u_k^2(s,a)} \qquad (h\le H \text{ and } \textstyle\sum_h u_k'(s,a,h)\le u_k(s,a))\\
&\le H^2\rho_K(s,a)\frac{\mathbb{E}_k\left[\widetilde{N}_k^2(s,a)c_k^2(s,a)\right]}{u_k(s,a)}\\
&\le 2H^2\rho_K(s,a)b_k(s,a). \qquad \text{(Lemma 16)}
\end{aligned}$$

Hence, applying a strengthened Freedman's inequality (Lemma 23) with $b = B_k = H$, $V = 2H^2\rho_K(s,a)\sum_{k=1}^K b_k(s,a)$, and $C' = \lceil\log_2 H\rceil\lceil\log_2(H^2K)\rceil$, we have with probability at least $1-\delta/(SA)$,

$$\begin{aligned}
&\sum_{k=1}^K b_k(s,a) - \widehat{b}_k(s,a)\\
&\le 4C'H\sqrt{\ln\left(\frac{C'SA}{\delta}\right)}\sqrt{\rho_K(s,a)\sum_{k=1}^K b_k(s,a)} + 2C'H\ln\left(\frac{C'SA}{\delta}\right)\\
&= 2C'H\sqrt{\ln\left(\frac{C'SA}{\delta}\right)}\left(\frac{\rho_K(s,a)}{\eta'} + \eta'\sum_{k=1}^K b_k(s,a)\right) + 2C'H\ln\left(\frac{C'SA}{\delta}\right),
\end{aligned}$$

where the last step is by AM-GM inequality. Finally, applying a union bound shows that the above holds for all $(s,a)\in\widetilde{\Gamma}$ with probability at least $1-\delta$ and thus

$$\textsc{Bias}_4 \le \gamma\sum_{k=1}^K\left\langle\phi^\star, b_k - \widehat{b}_k\right\rangle \le \left(\frac{1}{\eta'}+1\right)\langle\phi^\star,\rho_K\rangle + \eta'\left\langle\phi^\star, \sum_{k=1}^K b_k\right\rangle + \tilde{\mathcal{O}}(1),$$

where we bound $2\gamma C'H\sqrt{\ln\left(\frac{C'SA}{\delta}\right)}$ by $\tilde{\mathcal{O}}(1)$ since $\gamma H\le 1$ when K is large enough ($K \gtrsim S^3A^2H^2$). □

Lemma 15. *For any $\lambda'\in(0,\frac{2}{H}]$, with probability at least $1-9\delta$ we have*

$$\textsc{Err}_1 \le 33\lambda'\left(\sum_{k=1}^K\left\langle q_k, \vec{h}\circ c_k\right\rangle + \tilde{\mathcal{O}}\left(SAH^3\sqrt{K}\right)\right) + \frac{34S^3A^2\ln^2\left(\frac{HKS^2A^2}{\delta}\right)}{\lambda'} + \tilde{\mathcal{O}}\left(H^3S^3A^2\right).$$

Also, with probability at least $1-8\delta$, we have $\textsc{Err}_2 = \tilde{\mathcal{O}}\left(H^3S^3A^2\right)$, *Proof.* We write $\textsc{Err}_1$ as

$$\textsc{Err}_1 = \sum_{k=1}^{K} \langle u_k - \widehat{q}_k, \widehat{c}_k - \mathbb{E}_k[\widehat{c}_k] \rangle + \sum_{k=1}^{K} \langle u_k - \widehat{q}_k, \mathbb{E}_k[\widehat{c}_k] \rangle .$$

For the first term, note that $\langle u_k - \widehat{q}_k, \widehat{c}_k \rangle \le \sum_{(s,a)} \widetilde{N}_k(s,a) \le H$, and

$$\mathbb{E}_k\left[\left(\sum_{(s,a)} (u_k(s,a) - \widehat{q}_k(s,a)) \frac{\widetilde{N}_k(s,a)}{u_k(s,a)} c_k(s,a)\right)^2\right] \le \mathbb{E}_k\left[\left(\sum_{(s,a)} \widetilde{N}_k(s,a) c_k(s,a)\right)^2\right] \le 2\left\langle q_k, \vec{h} \circ c_k \right\rangle .$$

Hence, by Freedman's inequality (Lemma 21), for any $0 < \lambda' \le \frac{2}{H}$, with probability at least $1 - \delta$ we have:

$$\sum_{k=1}^{K} \langle u_k - \widehat{q}_k, \widehat{c}_k - \mathbb{E}_k[\widehat{c}_k] \rangle = \lambda' \sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \frac{2\ln(1/\delta)}{\lambda'} .$$

For the second term, with probability at least $1 - 8\delta$:

$$\begin{aligned}
\sum_{k=1}^{K} \langle u_k - \widehat{q}_k, \mathbb{E}_k[\widehat{c}_k] \rangle &\le \sum_{k=1}^{K} \langle u_k - \widehat{q}_k, c_k \rangle = \sum_{k=1}^{K} \langle u_k - q_k, c_k \rangle + \langle q_k - \widehat{q}_k, c_k \rangle \\
&\le 64 \sqrt{S^3 A^2 \ln^2\left(\frac{HKS^2A^2}{\delta}\right) \left(\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \tilde{\mathcal{O}}\left(SAH^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3 S^3 A^2\right) . \\
&\qquad \text{(Lemma 11 and Lemma 9)} \\
&\le 32\lambda' \left(\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \tilde{\mathcal{O}}\left(SAH^3\sqrt{K}\right)\right) + \frac{32 S^3 A^2 \ln^2\left(\frac{HKS^2A^2}{\delta}\right)}{\lambda'} + \tilde{\mathcal{O}}\left(H^3 S^3 A^2\right) . \\
&\qquad \text{(AM-GM inequality)}
\end{aligned}$$

Putting everything together, we have:

$$\textsc{Err}_1 \le 33\lambda' \left(\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \tilde{\mathcal{O}}\left(SAH^3\sqrt{K}\right)\right) + \frac{34 S^3 A^2 \ln^2\left(\frac{HKS^2A^2}{\delta}\right)}{\lambda'} + \tilde{\mathcal{O}}\left(H^3 S^3 A^2\right) .$$

For $\textsc{Err}_2$, use the bound for $\sum_{k=1}^{K} \langle u_k - \widehat{q}_k, c_k \rangle$ from above, we have with probability at least $1 - 8\delta$:

$$\begin{aligned}
\textsc{Err}_2 &= 2\gamma \sum_{k=1}^{K} \left\langle u'_k - \widehat{q}_k, \vec{h} \circ c_k \right\rangle \le 2\gamma H \sum_{k=1}^{K} \langle u_k - \widehat{q}_k, c_k \rangle \\
&= \tilde{\mathcal{O}}\left(\gamma H \sqrt{S^3 A^2 \left(\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle + \tilde{\mathcal{O}}\left(SAH^3\sqrt{K}\right)\right)} + \gamma H^4 S^3 A^2\right) && \text{(Lemma 11 and Lemma 9)} \\
&= \tilde{\mathcal{O}}\left(\gamma H \sqrt{S^3 A^2 (H^2 K + SAH^3\sqrt{K})}\right) = \tilde{\mathcal{O}}\left(H^3 S^3 A^2\right) . && \left(\gamma = \tilde{\mathcal{O}}\left(\sqrt{\tfrac{SA}{DTK}}\right) \text{ and } \gamma H \le 1\right)
\end{aligned}$$

□

Lemma 16. *For any episode k and $(s,a) \in \widetilde{\Gamma}$, we have $\mathbb{E}_k\left[\widetilde{N}_k(s,a)^2 c_k(s,a)^2\right] \le 2u_k(s,a)b_k(s,a)$.*

Proof. We use the inequality $(\sum_{i=1}^{I} a_i)^2 \le 2\sum_i a_i(\sum_{i'=i}^{I} a_{i'})$:

$$\mathbb{E}_k\left[\widetilde{N}_k(s,a)^2 c_k(s,a)^2\right] \le \mathbb{E}_k\left[\left(\sum_{h=1}^{H} \widetilde{N}_k(s,a,h)\right)^2 c_k(s,a)\right]$$

$$\leq 2\mathbb{E}_k\left[\left(\sum_{h=1}^{H}\widetilde{N}_k(s,a,h)\right)\left(\sum_{h'\geq h}^{H}\widetilde{N}_k(s,a,h')c_k(s,a)\right)\right]$$

$$\leq 2\mathbb{E}_k\left[\sum_{h=1}^{H}\sum_{h'\geq h}^{H}\widetilde{N}_k(s,a,h')c_k(s,a)\right] \qquad (\widetilde{N}_k(s,a,h)\leq 1)$$

$$= 2\sum_{h=1}^{H}\sum_{h'\geq h}^{H} q_k(s,a,h')c_k(s,a) = 2\sum_{h=1}^{H} h\cdot q_k(s,a,h)c_k(s,a)$$

$$\leq 2\sum_{h=1}^{H} h\cdot u'_k(s,a,h)c_k(s,a) = 2u_k(s,a)b_k(s,a),$$

where the last step is by the definition of b_k. □

D. Omitted details for Section 6

In this section, we provide all omitted proofs for Section 6. We first introduce modifications to the loop-free SSP which allows the learner to switch to Bernstein-SSP at any state and any time. We add a new state s'_f in $\widetilde{M}$ (that is, add s'_f into $\mathcal{X}$) and add a_f to $\mathcal{A}_{(s,h)}$ for all $(s,h)\in(\mathcal{S}\cup\{s'_f\})\times[H_1]$, such that $\widetilde{P}((s'_f,h+1)|(s,h),a_f)=1,\forall s\in\mathcal{S}\cup\{s'_f\},h\in[H_1]$ and $\widetilde{c}_k(s,a_f,h)=0,\forall s\in\mathcal{S}\cup\{s'_f\},h\in[H]$. Also add (s,a_f) to $\widetilde{\Gamma}$ for all $s\in\mathcal{S}$. Executing a_f in state (s,h) leads the agent into state $(s'_f,h+1)$ and stay at s'_f until it reaches (s_f,H_1+1). This is equivalent to directly transit to (s_f,H_1+1) when taking action a'_f at (s,h), but we pad states $(s'_f,h+1),\ldots,(s'_f,H_1)$ to preserve the layer structure (which simplifies notation). Therefore, executing a_f at (s,h) with $h\leq H_1$ corresponds to directly switching to Bernstein-SSP at current state in M (see Algorithm 1 and Footnote 2). Eventually, this modification only amounts to the following slightly different definition of $\widetilde{\Delta}(T,\mathcal{P})$ for the algorithm (so that taking action a_f does not count towards actual time steps in M):

$$\widetilde{\Delta}(T,\mathcal{P})=\Bigg\{q\in[0,1]^{\widetilde{\Gamma}\times\mathcal{X}\times[H]}:\sum_{h=1}^{H}\sum_{(s,a)\in\widetilde{\Gamma}\setminus((\mathcal{S}\cup\{s'_f\})\times\{a_f\})}q(s,a,h)\leq T,$$
$$\sum_{a\in\widetilde{\mathcal{A}}_{(s,h)}}q(s,a,h)-\sum_{(s',a')\in\widetilde{\Gamma}}q(s',a',s,h-1)=\mathbb{I}\{(s,h)=\widetilde{s}_0\},\forall h>1;$$
$$\sum_{a\in\widetilde{\mathcal{A}}_{(s,1)}}q(s,a,1)=\mathbb{I}\{s=s_0\},\ \forall s\in\mathcal{X};\ P_q\in\mathcal{P}\Bigg\}.$$

With this new definition, it is not hard to obtain something similar to Lemma 1 for the pseudo-regret.

Lemma 17. *Suppose $H_1\geq 8T_{\max}\ln K, H_2=\lceil 2D\rceil$ and $K\geq D$, and $\widetilde{\pi}_1,\ldots,\widetilde{\pi}_k$ are policies for $\widetilde{M}$. Then with probability at least $1-\delta$, the regret of executing $\sigma(\widetilde{\pi}_1),\ldots,\sigma(\widetilde{\pi}_K)$ in M satisfies,*

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

where $\pi^\star\in\operatorname{argmin}_{\pi\in\Pi_{\text{proper}}}J^{\pi,c}(s_0)$ and $\widetilde{\pi}^\star(s,h)=\pi^\star(s)$.

Proof. By similar arguments in the proof of Lemma 1, we have with probability at least $1-\delta$:

$$\sum_{k=1}^{K}\langle N_k,c_k\rangle\leq\sum_{k=1}^{K}\left\langle\widetilde{N}_k,c_k\right\rangle+\tilde{\mathcal{O}}\left(D^{3/2}S^2A(\ln\tfrac{1}{\delta})^2\right).$$

and (the same cost for all K episodes)

$$J^{\widetilde{\pi}^\star,c}(\widetilde{s}_0)\leq J^{\pi^\star,c}(s_0)+\frac{2H_2}{K^2}=J^{\pi^\star,c}(s_0)+\tilde{\mathcal{O}}\left(\frac{1}{K}\right).$$

Putting everything together, we get:

$$\widetilde{R}_K = \sum_{k=1}^{K} \langle N_k, c_k \rangle - J^{\pi^\star, c}(s_0) \leq \sum_{k=1}^{K} \left\langle \widetilde{N}_k, c_k \right\rangle - \langle q_{\widetilde{\pi}^\star}, c \rangle + \tilde{\mathcal{O}}\left(D^{3/2} S^2 A (\ln \tfrac{1}{\delta})^2\right).$$

□

With these modifications, the state value w.r.t optimistic transition/cost is of $\mathcal{O}\left(D\right)$.

Lemma 18. *Assume $T \geq T_\star$. For $\widehat{q}_k$ obtained from Eq. (5), we have $\langle \widehat{q}_k, \widehat{c}_k \rangle \leq \langle q_{\widetilde{\pi}^\star}, \widehat{c}_k \rangle$ and $J^{P_{\widehat{q}_k}, \widetilde{\pi}_k, \widehat{c}_k}(s, h) \leq H_2$ for all (s, h) with $\widehat{q}_k(s, h) > 0$.*

Proof. The first inequality is by the definition of $\widehat{q}_k = \operatorname{argmin}_{q \in \widetilde{\Delta}(T, \mathcal{P}_k)} \langle q, \widehat{c}_k \rangle$ and $q_{\widetilde{\pi}^\star} \in \widetilde{\Delta}(T, \mathcal{P}_k)$. The second inequality can be proven by contradiction: suppose $J^{P_{\widehat{q}_k}, \widetilde{\pi}_k, \widehat{c}_k}(s, h) > H_2$. Define $\pi'(a'_f | s, h) = 1$ and $\pi' = \widetilde{\pi}_k$ for all other entries. Note that $q_{P_{\widehat{q}_k}, \pi'} \in \widetilde{\Delta}(T, \mathcal{P}_{i_k})$, since the expected hitting time of π' should be no more than that of $\widetilde{\pi}_k$. Moreover, by the structure of $\widetilde{M}$, we have $J^{P_{\widehat{q}_k}, \pi', \widehat{c}_k}(s, h) = H_2 < J^{P_{\widehat{q}_k}, \widetilde{\pi}_k, \widehat{c}_k}(s, h)$. Since, π' differs from $\widetilde{\pi}_k$ in only one entry, we have $J^{P_{\widehat{q}_k}, \pi', \widehat{c}_k}(s_0, 1) < J^{P_{\widehat{q}_k}, \widetilde{\pi}_k, \widehat{c}_k}(s_0, 1)$, a contradiction to the definition of $\widehat{q}_k$. □

We next present a lemma similar to Lemma 6 but for cost estimation.

Lemma 19. *With probability at least $1 - \delta$, for any $(s, a) \in \Gamma, k \in [K]$, $0 \leq c(s, a) - \widehat{c}_k(s, a) \leq 8\sqrt{A^c_k(s, a) c(s, a)} + 34 A^c_k(s, a)$.*

Proof. Applying Lemma 22 with $X_k = c_k(s, a)$ for each $(s, a) \in \Gamma$ and then by a union bound over all $(s, a) \in \Gamma$, we have with probability at least $1 - \delta$, for all $k \in [K]$:

$$|\bar{c}_k(s, a) - c(s, a)| \leq 2\sqrt{A^c_k(s, a) \bar{c}_k(s, a)} + 7 A^c_k(s, a).$$

Hence, $c(s, a) \geq \widehat{c}_k(s, a)$ by the definition of $\widehat{c}_k$. Applying $x \leq a\sqrt{x} + b \implies x \leq (a + \sqrt{b})^2$ with $x = \bar{c}_k(s, a)$ to the inequality above (ignoring the absolute value operator), we obtain

$$\bar{c}_k(s, a) \leq c(s, a) + 4\sqrt{A^c_k(s, a) c(s, a)} + 23 A^c_k(s, a) \leq 3c(s, a) + 25 A^c_k(s, a),$$

where we apply $\sqrt{ab} \leq \frac{a+b}{2}, \forall a, b > 0$ for the last inequality. Therefore, $2\sqrt{A^c_k(s, a) \bar{c}_k(s, a)} + 7 A^c_k(s, a) \leq 4\sqrt{A^c_k(s, a) c(s, a)} + 17 A^c_k(s, a)$, and

$$c(s, a) - \widehat{c}_k(s, a) = c(s, a) - \bar{c}_k(s, a) + \bar{c}_k(s, a) - \widehat{c}_k(s, a) \leq 8\sqrt{A^c_k(s, a) c(s, a)} + 34 A^c_k(s, a).$$

□

We are now ready to prove Theorem 3. We first introduce some notations used in the proof below. Define $P_k = P_{\widehat{q}_k}$, and $x_k(s, a) = \mathbb{E}_k[\mathbb{I}_k(s, a)]$ as the probability that (s, a) is ever visited in episode k. Also denote by $q_{k,(s,a,h)}$ (or $\widehat{q}_{k,(s,a,h)}$) the occupancy measure of $\widetilde{\pi}_k$ with transition $\widetilde{P}$ (or P_k) and initial state-action pair $((s, h), a)$. Without loss of generality, we assume c_k is sampled right before the beginning of episode k instead of before learning starts.

Proof of Theorem 3. We condition on the event of Lemma 5 and Lemma 19, which happens with probability at least $1 - 2\delta$. We first decompose the regret in $\widetilde{M}$ using the fact $\langle q_{\widetilde{\pi}^\star}, c \rangle \geq \langle q_{\widetilde{\pi}^\star}, \widehat{c}_k \rangle \geq \langle \widehat{q}_k, \widehat{c}_k \rangle$ derived from Lemma 19 and Lemma 18:

$$\begin{aligned} \sum_{k=1}^{K} \left\langle \widetilde{N}_k, c_k \right\rangle - \langle q_{\widetilde{\pi}^\star}, c \rangle &\leq \sum_{k=1}^{K} \left\langle \widetilde{N}_k, c_k \right\rangle - \langle \widehat{q}_k, \widehat{c}_k \rangle \\ &\leq \sum_{k=1}^{K} \left\langle \widetilde{N}_k, c_k \right\rangle - \langle q_k, c \rangle + \sum_{k=1}^{K} \langle q_k, c - \widehat{c}_k \rangle + \sum_{k=1}^{K} \langle q_k - \widehat{q}_k, \widehat{c}_k \rangle. \end{aligned} \quad (10)$$

The first term in Eq. (10) is a martingale difference sequence where the randomness in episode k is w.r.t the learner's trajectory in episode k and sampling of c_k. Thus, it suffices to bound its second moment according to Freedman's inequality. Note that

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}_{c_k}\left[\mathbb{E}_k\left[\left\langle\widetilde{N}_k, c_k\right\rangle^2\right]\right] &\le 2\sum_{k=1}^{K}\mathbb{E}_{c_k}\left[\left\langle q_k, \vec{h}\circ c_k\right\rangle\right] = 2\left\langle q_k, \vec{h}\circ c\right\rangle && \text{(Lemma 2)}\\ &= 2\underbrace{\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ(c-\widehat{c}_k)\right\rangle}_{\zeta_1} + 2\underbrace{\sum_{k=1}^{K}\left\langle q_k-\widehat{q}_k, \vec{h}\circ\widehat{c}_k\right\rangle}_{\zeta_2} + 2\underbrace{\sum_{k=1}^{K}\left\langle \widehat{q}_k, \vec{h}\circ\widehat{c}_k\right\rangle}_{\zeta_3}.\end{aligned}$$

We can bound ζ_1 as follows with probability at least $1-\delta$:

$$\begin{aligned}\zeta_1 &\le H\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)(c(s,a)-\widehat{c}_k(s,a)) && (\widehat{c}_k(s,a)\le c(s,a)\text{ by Lemma 19})\\ &= \tilde{\mathcal{O}}\left(H^2\sum_{k=1}^{K}\sum_{(s,a)} x_k(s,a)\left(\sqrt{\frac{c(s,a)}{\mathbf{N}_k^c(s,a)}}+\frac{1}{\mathbf{N}_k^c(s,a)}\right)\right) && (q_k(s,a)\le Hx_k(s,a)\text{ and Lemma 19})\\ &= \tilde{\mathcal{O}}\left(H^2\sum_{k=1}^{K}\sum_{(s,a)} \mathbb{I}_k(s,a)\left(\sqrt{\frac{c(s,a)}{\mathbf{N}_k^c(s,a)}}+\frac{1}{\mathbf{N}_k^c(s,a)}\right)+H^2SA\right) && \text{(Lemma 24)}\\ &= \tilde{\mathcal{O}}\left(H^2SA\sqrt{K}\right) = \tilde{\mathcal{O}}\left(H^4S^2A^2+K\right), && \text{(AM-GM inequality)}\end{aligned}$$

where the last inequality is by $\sum_{k=1}^{K}\sum_{(s,a)}\frac{\mathbb{I}_k(s,a)}{\sqrt{\mathbf{N}_k^c(s,a)}} = \tilde{\mathcal{O}}\left(\sum_{(s,a)}\sqrt{\mathbf{N}_K^c(s,a)}\right) = \tilde{\mathcal{O}}\left(SA\sqrt{K}\right)$. To bound ζ_2, we apply Lemma 9 with costs $\{\frac{\vec{h}\circ\widehat{c}_k}{H}\}_k$ to obtain with probability at least $1-4\delta$:

$$\begin{aligned}\zeta_2 &= \tilde{\mathcal{O}}\left(H\sqrt{S^2A\left(\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k\right\rangle + H^3\sqrt{K}\right)}+H^4S^2A\right) && (\tfrac{1}{H}(\vec{h}\circ\widehat{c}_k)(s,a)\le c_k(s,a))\\ &= \tilde{\mathcal{O}}\left(\sqrt{S^2AH^5K}+H^4S^2A\right) = \tilde{\mathcal{O}}\left(K+H^5S^2A\right). && \text{(AM-GM inequality)}\end{aligned}$$

Finally, by Lemma 2, Lemma 18 and $\sum_{(s,a)}\widehat{q}_k(s,a)\le T$, we have $\zeta_3 = \sum_{k=1}^{K}\langle\widehat{q}_k, J^{P_k,\widetilde{\pi}_k,\widehat{c}_k}\rangle = \tilde{\mathcal{O}}(DTK)$. Putting everything together, we have: $\sum_{k=1}^{K}\mathbb{E}_{c_k}\left[\mathbb{E}_k\left[\left\langle\widetilde{N}_k, c_k\right\rangle^2\right]\right] = \tilde{\mathcal{O}}\left(H^5S^2A^2+DTK\right)$. Hence, by Freedman's inequality with $\lambda = \frac{1}{\sqrt{DTK+H^5S^2A^2}}\le\frac{1}{H}$, we have with probability at least $1-\delta$:

$$\sum_{k=1}^{K}\left\langle\widetilde{N}_k, c_k\right\rangle - \langle q_k, c\rangle = \tilde{\mathcal{O}}\left(\sqrt{DTK}+H^3SA\right).$$

For the second term in Eq. (10), by Lemma 19 and the definition of A_k^c:

$$\begin{aligned}\sum_{k=1}^{K}\langle q_k, c-\widehat{c}_k\rangle &= \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{c(s,a)}{\mathbf{N}_k^c(s,a)}}+\sum_{k=1}^{K}\sum_{(s,a)}\frac{q_k(s,a)}{\mathbf{N}_k^c(s,a)}\right)\\ &= \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}}+\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{c(s,a)-\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}}+H\sum_{k=1}^{K}\sum_{(s,a)}\frac{x_k(s,a)}{\mathbf{N}_k^c(s,a)}\right)\\ &\qquad (\sqrt{x+y}\le\sqrt{x}+\sqrt{y},\ q_k(s,a)\le Hx_k(s,a))\\ &= \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}}+\sqrt{K}+H^2S^2A^2\right),\end{aligned}$$

where in the last line we use with probability at least $1-\delta$:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{c(s,a)-\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}} + H\sum_{k=1}^{K}\sum_{(s,a)}\frac{x_k(s,a)}{\mathbf{N}_k^c(s,a)}\\
&= \tilde{\mathcal{O}}\left(H\sum_{k=1}^{K}\sum_{(s,a)}\frac{x_k(s,a)}{\mathbf{N}_k^c(s,a)^{3/4}}\right) && \left(q_k(s,a)\le Hx_k(s,a)\text{, Lemma 19, and } \tfrac{1}{\mathbf{N}_k^c(s,a)}\le\tfrac{1}{\mathbf{N}_k^c(s,a)^{3/4}}\right)\\
&= \tilde{\mathcal{O}}\left(H\sum_{k=1}^{K}\sum_{(s,a)}\frac{\mathbb{I}_k(s,a)}{\mathbf{N}_k^c(s,a)} + HSA\right) && \text{(Lemma 24)}\\
&= \mathcal{O}\left(SAHK^{1/4}\right) = \mathcal{O}\left(H^2S^2A^2+\sqrt{K}\right). && \text{(AM-GM inequality)}
\end{aligned}$$

It is left to bound $\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}}$, for which we need to discuss the type of the feedback model. In the full information setting, we have $\mathbf{N}_k^c(s,a)=\max\{1,k-1\}$. We decompose it into two terms:

$$\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}} = \sum_{k=1}^{K}\sum_{(s,a)} \widehat{q}_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}} + \sum_{k=1}^{K}\sum_{(s,a)}(q_k(s,a)-\widehat{q}_k(s,a))\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}}.$$

For the first term, by Cauchy-Schwarz inequality, $\sum_{(s,a)}\widehat{q}_k(s,a)\le T$, and $\langle \widehat{q}_k,\widehat{c}_k\rangle \le H_2 = \mathcal{O}(D)$ (Lemma 18):

$$\sum_{k=1}^{K}\sum_{(s,a)} \widehat{q}_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}} = \sqrt{\sum_{k=1}^{K}\sum_{(s,a)}\widehat{q}_k(s,a)}\sqrt{\sum_{k=1}^{K}\frac{\langle\widehat{q}_k,\widehat{c}_k\rangle}{\max\{1,k-1\}}} = \tilde{\mathcal{O}}\left(\sqrt{DTK}\right).$$

For the second term, we apply Lemma 9 with costs $\{\sqrt{\widehat{c}_k/\mathbf{N}_k^c}\}_k$: with probability at least $1-4\delta$,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{(s,a)}(q_k(s,a)-\widehat{q}_k(s,a))\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}}\\
&= \tilde{\mathcal{O}}\left(\sqrt{S^2A\left(\sum_{k=1}^{K}\sum_{(s,a)} h\cdot q_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}} + H^3\sqrt{K}\right)} + H^3S^2A\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{S^2A\left(\sum_{k=1}^{K}\frac{H^2}{\sqrt{k}} + H^3\sqrt{K}\right)} + H^3S^2A\right) && \left(h\le H \text{ and } \textstyle\sum_{(s,a)} q_k(s,a)\le H\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{H^3S^2A\sqrt{K}} + H^3S^2A\right) = \tilde{\mathcal{O}}\left(\sqrt{K}+H^3S^2A\right). && \left(\textstyle\sum_{k=1}^{K}\frac{1}{\sqrt{k}}=\mathcal{O}\left(\sqrt{K}\right)\text{ and AM-GM inequality}\right)
\end{aligned}$$

Therefore, $\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}} = \tilde{\mathcal{O}}\left(\sqrt{DTK}+H^3S^2A\right)$ in the full information setting.

In the bandit feedback setting, denote by $x_k(s,a,h)$ the probability that the first visit to (s,a) is at time step h under transition $\widetilde{P}$ and policy $\widetilde{\pi}_k$. Then, we have $q_k(s,a)=\sum_{h=1}^{H}x_k(s,a,h)q_{k,(s,a,h)}(s,a)$, and

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}} = \sum_{k=1}^{K}\sum_{(s,a)}\sqrt{\frac{q_k(s,a)}{\mathbf{N}_k^c(s,a)}}\sqrt{\sum_{h=1}^{H}x_k(s,a,h)q_{k,(s,a,h)}(s,a)\widehat{c}_k(s,a)}\\
&= \sum_{k=1}^{K}\sum_{(s,a)}\sqrt{\frac{q_k(s,a)}{\mathbf{N}_k^c(s,a)}}\sqrt{\sum_{h=1}^{H}x_k(s,a,h)\left(\widehat{q}_{k,(s,a,h)}(s,a)\widehat{c}_k(s,a)+(q_{k,(s,a,h)}(s,a)-\widehat{q}_{k,(s,a,h)}(s,a))\widehat{c}_k(s,a)\right)}
\end{aligned}$$

$$\le \underbrace{\sum_{k=1}^{K}\sum_{(s,a)}\sqrt{q_k(s,a)}\sqrt{\frac{\sum_{h=1}^{H} x_k(s,a,h)\widehat{q}_{k,(s,a,h)}(s,a)\widehat{c}_k(s,a)}{\mathbf{N}^c_k(s,a)}}}_{\zeta_4}$$

$$+\underbrace{\sum_{k=1}^{K}\sum_{(s,a)}\sqrt{\frac{q_k(s,a)}{\mathbf{N}^c_k(s,a)}}\sqrt{\sum_{h=1}^{H} x_k(s,a,h)\left|q_{k,(s,a,h)}(s,a)-\widehat{q}_{k,(s,a,h)}(s,a)\right|\widehat{c}_k(s,a)}}_{\zeta_5}.$$

For ζ_4, first notice that

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a) &= \sum_{k=1}^{K}\sum_{(s,a)} \widehat{q}_k(s,a) + \sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)-\widehat{q}_k(s,a)\\
&= \tilde{\mathcal{O}}\left(TK\right) + H\sum_{(s,a),s',h\in U} q_k(s,a,h)\epsilon^\star_k(s,a,s') && \textstyle\left(\sum_{(s,a)}\widehat{q}_k(s,a)\le T \text{ and Lemma 7}\right)\\
&= \tilde{\mathcal{O}}\left(TK + HS\sum_{(s,a)}\frac{q_k(s,a)}{\sqrt{\mathbf{N}^+_k(s,a)}}\right) && \text{(by the definition of } \epsilon^\star_k \text{ in Lemma 6)}\\
&\overset{\text{(i)}}{=} \tilde{\mathcal{O}}\left(TK + HS\sqrt{SAHK} + H^2SA\right) \overset{\text{(ii)}}{=} \tilde{\mathcal{O}}\left(TK + H^3S^3A\right), && (11)
\end{aligned}$$

where in (ii) we apply AM-GM inequality, and in (i) we apply with probability at least $1-\delta$:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{(s,a)}\frac{q_k(s,a)}{\sqrt{\mathbf{N}^+_k(s,a)}} = \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{(s,a)}\frac{\widetilde{N}_k(s,a)}{\sqrt{\mathbf{N}^+_k(s,a)}} + H\right) && \text{(Lemma 24)}\\
&= \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{(s,a)}\frac{\widetilde{N}_k(s,a)}{\sqrt{\mathbf{N}^+_{k+1}(s,a)}} + H\sum_{k=1}^{K}\sum_{(s,a)}\left(\frac{1}{\sqrt{\mathbf{N}^+_k(s,a)}}-\frac{1}{\sqrt{\mathbf{N}^+_{k+1}(s,a)}}\right)+H\right) && (\widetilde{N}_k(s,a)\le H)\\
&= \tilde{\mathcal{O}}\left(\sum_{(s,a)}\sqrt{\mathbf{N}^+_{K+1}(s,a)} + HSA\right) = \tilde{\mathcal{O}}\left(\sqrt{SAHK} + HSA\right). && (12)
\end{aligned}$$

Moreover, by Lemma 24, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\sum_{(s,a)}\frac{x_k(s,a)}{\mathbf{N}^c_k(s,a)} = \mathcal{O}\left(\sum_{k=1}^{K}\sum_{(s,a)}\frac{\mathbb{I}_k(s,a)}{\mathbf{N}^c_k(s,a)}+1\right) = \tilde{\mathcal{O}}\left(SA\right). \tag{13}$$

Therefore,

$$\begin{aligned}
\zeta_4 &= \sum_{k=1}^{K}\sum_{(s,a)}\sqrt{q_k(s,a)}\sqrt{\frac{\sum_{h=1}^{H} x_k(s,a,h)\widehat{q}_{k,(s,a,h)}(s,a)\widehat{c}_k(s,a)}{\mathbf{N}^c_k(s,a)}}\\
&= \mathcal{O}\left(\sqrt{\sum_{k=1}^{K}\sum_{(s,a)} q_k(s,a)}\sqrt{D\sum_{k=1}^{K}\sum_{(s,a)}\frac{x_k(s,a)}{\mathbf{N}^c_k(s,a)}}\right)\\
&\qquad \textstyle\left(\text{Cauchy-Schwarz inequality, } \widehat{q}_{k,(s,a,h)}(s,a)\widehat{c}_k(s,a)\le Q^{\widetilde{\pi}_k,P_k,\widehat{c}_k}(s,a)=\tilde{\mathcal{O}}(D)\text{, and } \sum_{h=1}^{H}x_k(s,a,h)=x_k(s,a)\right)\\
&= \tilde{\mathcal{O}}\left(\sqrt{TK+H^3S^3A}\sqrt{DSA}\right) = \tilde{\mathcal{O}}\left(\sqrt{DTSAK}+H^2S^2A\right). && \text{(Eq. (11), Eq. (13) and } D\le H)
\end{aligned}$$

For ζ_5, we have

$$\sum_{k=1}^{K}\sum_{(s,a)}\sqrt{\frac{q_k(s,a)}{\mathbf{N}^c_k(s,a)}}\sqrt{\sum_{h=1}^{H} x_k(s,a,h)\left|q_{k,(s,a,h)}(s,a)-\widehat{q}_{k,(s,a,h)}(s,a)\right|\widehat{c}_k(s,a)}$$

$$\le \sum_{k=1}^{K}\sum_{(s,a)}\sqrt{\frac{q_k(s,a)}{\mathbf{N}_k^c(s,a)}}\sqrt{\sum_{h=1}^{H}x_k(s,a,h)\left\langle|q_{k,(s,a,h)}-\widehat{q}_{k,(s,a,h)}|,\widehat{c}_k\right\rangle}$$

$$\le \sum_{k=1}^{K}\sum_{(s,a)}\sqrt{\frac{q_k(s,a)}{\mathbf{N}_k^c(s,a)}}\sqrt{H\sum_{h=1}^{H}x_k(s,a,h)\sum_{s',a',s'',h'\in U}q_{k,(s,a,h)}(s',a',h')\epsilon_k^\star(s',a',s'')}$$

(Lemma 7)

$$=\tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{(s,a)}\sqrt{\frac{q_k(s,a)}{\mathbf{N}_k^c(s,a)}}\sqrt{H\sum_{h=1}^{H}x_k(s,a,h)\sum_{s',a',s'',h'\in U}\frac{q_{k,(s,a,h)}(s',a',h')}{\sqrt{\mathbf{N}_k^+(s',a')}}}\right)$$

(by definition of $\epsilon_k^\star$ in Lemma 6)

$$\le\sqrt{H\sum_{k=1}^{K}\sum_{(s,a)}\frac{x_k(s,a)}{\mathbf{N}_k^c(s,a)}}\sqrt{H\sum_{k=1}^{K}\sum_{(s,a)}\sum_{s',a',s'',h'\in U}\frac{q_k(s',a',h')}{\sqrt{\mathbf{N}_k^+(s',a')}}}$$

(Cauchy-Schwarz inequality, $q_k(s,a)\le Hx_k(s,a)$, and $\sum_{h=1}^{H}x_k(s,a,h)q_{k,(s,a,h)}(s',a',h')\le q_k(s',a',h')$)

$$=\tilde{\mathcal{O}}\left(\sqrt{HSA}\sqrt{HS^2A\sum_{k=1}^{K}\sum_{(s',a')}\frac{q_k(s',a')}{\sqrt{\mathbf{N}_k^+(s',a')}}}\right)=\tilde{\mathcal{O}}\left(\sqrt{H^2S^3A^2\left(\sqrt{SAHK}+HSA\right)}\right).$$

(Eq. (13) and Eq. (12))

$$=\tilde{\mathcal{O}}\left(\sqrt{SAK+H^5S^6A^4}\right)=\tilde{\mathcal{O}}\left(\sqrt{SAK}+H^3S^3A^2\right).$$

(AM-GM inequality and $\sqrt{x+y}\le\sqrt{x}+\sqrt{y}$)

Putting everything together, we have the second term in Eq. (10) bounded by:

$$\sum_{k=1}^{K}\langle q_k,c-\widehat{c}_k\rangle=\tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{(s,a)}q_k(s,a)\sqrt{\frac{\widehat{c}_k(s,a)}{\mathbf{N}_k^c(s,a)}}+\sqrt{K}+H^2S^2A^2\right)$$
$$=\begin{cases}\tilde{\mathcal{O}}\left(\sqrt{DTK}+H^3S^2A^2\right), & \text{full information setting,}\\ \tilde{\mathcal{O}}\left(\sqrt{DTSAK}+H^3S^3A^2\right), & \text{bandit feedback setting.}\end{cases}$$

For the third term in Eq. (10), by Lemma 9 with costs $\{\widehat{c}_k\}_k$ and Lemma 2, we have with probability at least $1-4\delta$:

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

($\sqrt{x+y}\le\sqrt{x}+\sqrt{y}$ and AM-GM inequality)

Note that:

$$\sum_{k=1}^{K}\left\langle q_k,\widehat{c}_k\odot Q^{\widetilde{\pi}_k,\widehat{c}_k}\right\rangle=\sum_{k=1}^{K}\left\langle q_k,\widehat{c}_k\odot(Q^{\widetilde{\pi}_k,\widehat{c}_k}-Q^{P_k,\widetilde{\pi}_k,\widehat{c}_k})\right\rangle+\sum_{k=1}^{K}\left\langle q_k,\widehat{c}_k\odot Q^{P_k,\widetilde{\pi}_k,\widehat{c}_k}\right\rangle.$$

For the first term, note that by Lemma 7:

$$Q^{\widetilde{\pi}_k,\widehat{c}_k}(s,a,h)-Q^{P_k,\widetilde{\pi}_k,\widehat{c}_k}(s,a,h)=\langle q_{k,(s,a,h)}-\widehat{q}_{k,(s,a,h)},\widehat{c}_k\rangle\le\sum_{(s',a'),s'',h'\in U}q_{k,(s,a,h)}(s',a',h')\epsilon_k^\star(s',a',s'')H.$$

Therefore,

$$\sum_{k=1}^{K}\left\langle q_k,\widehat{c}_k\odot(Q^{\widetilde{\pi}_k,\widehat{c}_k}-Q^{P_k,\widetilde{\pi}_k,\widehat{c}_k})\right\rangle\le\sum_{k=1}^{K}\sum_{(s,a),h}q_k(s,a,h)\sum_{(s',a'),s'',h'\in U}q_{k,(s,a,h)}(s',a',h')\epsilon_k^\star(s',a',s'')H$$

$$\leq H^2 \sum_{k=1}^{K} \sum_{(s',a'),s'',h' \in U} q_k(s',a',h')\epsilon_k^\star(s',a',s'') \quad \left(\textstyle\sum_{(s,a)} q_k(s,a,h) q_{k,(s,a,h)}(s',a',h') = q_k(s',a',h') \text{ for } h \leq h'\right)$$

$$= \tilde{\mathcal{O}}\left(H^2 S \sum_{k=1}^{K} \sum_{(s',a')} \frac{q_k(s',a')}{\sqrt{\mathbf{N}_k^+(s',a')}}\right) = \tilde{\mathcal{O}}\left(H^2 S(\sqrt{SAHK} + HSA)\right) = \tilde{\mathcal{O}}\left(H^5 S^3 A + K\right).$$

(Lemma 6, Eq. (12), and AM-GM inequality)

For the second term, with probability at least $1 - 4\delta$,

$$\begin{aligned}
\sum_{k=1}^{K} \left\langle q_k, \widehat{c}_k \odot Q^{P_k,\widetilde{\pi}_k,\widehat{c}_k} \right\rangle &= \mathcal{O}\left(D \sum_{k=1}^{K} \langle q_k, \widehat{c}_k \rangle\right) \qquad (Q^{P_k,\widetilde{\pi}_k,\widehat{c}_k}(s,a,h) \leq H_2 + 1 = \mathcal{O}(D) \text{ by Lemma 18}) \\
&= \mathcal{O}\left(D \sum_{k=1}^{K} \langle \widehat{q}_k, \widehat{c}_k \rangle + D \sum_{k=1}^{K} \langle q_k - \widehat{q}_k, \widehat{c}_k \rangle\right) \\
&= \tilde{\mathcal{O}}\left(D^2 K + D\sqrt{S^2 A \left(\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ \widehat{c}_k \right\rangle + H^3\sqrt{K}\right)} + DH^3 S^2 A\right) \\
&\qquad (\langle \widehat{q}_k, \widehat{c}_k \rangle \leq D \text{ (Lemma 18) and Lemma 9 with costs } \{\widehat{c}_k\}_k) \\
&= \tilde{\mathcal{O}}\left(D^2 K + D\sqrt{H^3 S^2 A K} + DH^3 S^2 A\right) = \tilde{\mathcal{O}}\left(D^2 K + DH^3 S^2 A\right). \\
&\qquad \left(\textstyle\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle \leq H^2 K \text{ and AM-GM inequality}\right)
\end{aligned}$$

Putting everything together, we obtain:

$$\sum_{k=1}^{K} \langle q_k - \widehat{q}_k, \widehat{c}_k \rangle = \tilde{\mathcal{O}}\left(\sqrt{S^2 A(H^5 S^3 A^2 + K + D^2 K + DH^3 S^2 A)} + \sqrt{K} + H^3 S^2 A\right) = \tilde{\mathcal{O}}\left(DS\sqrt{AK} + H^3 S^3 A^2\right).$$

Substituting everything back to Eq. (10) and applying Lemma 17, we have with probability at least $1 - \delta$,

$$\widetilde{R}_K = \begin{cases} \tilde{\mathcal{O}}\left(\sqrt{DTK} + DS\sqrt{AK} + H^3 S^3 A^2\right), & \text{full information setting,} \\ \tilde{\mathcal{O}}\left(\sqrt{DTSAK} + DS\sqrt{AK} + H^3 S^3 A^2\right), & \text{bandit feedback setting.} \end{cases}$$

This completes the proof. □

E. Learning without knowing SSP-diameter

In this section, we present a general idea on how to learn without knowing the SSP-diameter D. To give a concrete example, we apply this idea to Algorithm 2 and obtain an algorithm (Algorithm 6) that achieves the same regret without the knowledge of D. The same idea can also be applied to other algorithms proposed in this paper (details omitted). The main ideas of our proposed algorithm are as follows:

1. We partition K episodes into M_p phases. In phase m, we learn on a virtual MDP $\Pi(M, \mathcal{S}_m)$. We call states in $\mathcal{S}_m \subseteq \mathcal{S}$ known states, and states in $\mathcal{S} \setminus \mathcal{S}_m$ unknown states which are all treated as goal states. That is, $\Pi(M, \mathcal{S}_m)$ is obtained by modifying the transition and cost function of M in unknown states so that $P(s|s,a) = 1$ and $c(s,a) = 0$ for any $s \in \mathcal{S} \setminus \mathcal{S}_m$. The correspondence between M and $\Pi(M, \mathcal{S}_m)$ is as follows: every time we reach a state $s \in \mathcal{S} \setminus \mathcal{S}_m$ in $\Pi(M, \mathcal{S}_m)$, we run Bernstein-SSP until the goal state is reached. In phase m, we run Algorithm 2 on $\widetilde{\Pi}(M, \mathcal{S}_m)$ with $H_2 = \lceil 2 \max_{s \in \mathcal{S}_m} \widetilde{D}_s \rceil$, where $\widetilde{\Pi}(M, \mathcal{S}_m)$ is the loop-free reduction of $\Pi(M, \mathcal{S}_m)$ and $\widetilde{D}_s$ is defined below.

2. In (Rosenberg and Mansour, 2020, Appendix I.3), they show that we can get an estimate $\widetilde{D}_s$ of $T^{\pi^f}(s)$ such that $T^{\pi^f}(s) \leq \widetilde{D}_s = \mathcal{O}(D)$ for any $s \in \mathcal{S}$ as long as we run Bernstein-SSP starting from state s for some $L \geq \max\left\{\frac{2400 D^2}{T^{\pi^f}(s_0)^2} \ln^3 \frac{4K}{\delta}, S^2 A\sqrt{D} \ln^2 \frac{KDSA}{\delta}\right\}$ episodes. We thus concretely define known states as follows: define

Algorithm 6 Learning SSP with unknown diameter

```
Input: Upper bound on expected hitting time $T$, horizon parameter $H_1$, confidence level $\delta$.
Define: $\eta = \min\left\{\frac{1}{8}, \sqrt{\frac{ST}{\widetilde{D}_{s_0}K}}\right\}, \lambda_j = 4\sqrt{\frac{S^2A}{D_jTK}}, L = 2400\sqrt{AK}\ln^3\frac{4KH_1SA}{\delta}$.
Initialization: $\mathbf{N}_1(s,a) = \mathbf{M}_1(s,a,s') = 0$ for all $(s,a,s') \in \Gamma \times (\mathcal{S} \cup \{g\})$.
Initialization: for each state $s$, a Bernstein-SSP instance $\mathcal{B}_s$ that uses $s$ as the initial state and treats all costs as 1.
Initialization: Compute $\widetilde{D}_{s_0}$ by executing $\mathcal{B}_{s_0}$ for the first $L$ episodes.
Initialization: $\mathcal{S}_1 = \{s_0\}, m = 1, j = 1, D_1 = 2\widetilde{D}_{s_0}$.
Initialization: $A = \mathcal{A}(T, H_1, \delta; \eta, \lambda_1, \mathbf{N}, \mathbf{M}, \mathcal{S}_m)$, where $\mathcal{A}$ is a variant of Algorithm 2 which also takes $\eta, \lambda, \mathbf{N}, \mathbf{M}$ and $\mathcal{S}_m$ as inputs.
Initialization: $N_f(s) = L\mathbb{I}\{s = s_0\}, \forall s \in \mathcal{S}$.
for $k = L+1, \ldots, K$ do
    Execute $A$ on $\Pi(M, \mathcal{S}_m)$ for episode $k$.
    if $A$ stops at an unknown state $e$ then
        Invoke $\mathcal{B}_e$ (as a new episode for it) and follow its decision until reaching $g$.
        $N_f(e) \leftarrow N_f(e) + 1$.
        if $N_f(e) = L$ then
            Compute $\widetilde{D}_e$ using previous data (Rosenberg and Mansour, 2020, Appendix I.3).
            $\mathcal{S}_{m+1} \leftarrow \mathcal{S}_m \cup \{e\}$.
            $m \leftarrow m + 1$.
            if $\widetilde{D}_e > D_j$ then
                $D_{j+1} \leftarrow 2\widetilde{D}_e$
                $j \leftarrow j + 1$.
            $A = \mathcal{A}(T, H_1, \delta;\ \eta, \lambda_j, \mathbf{N}, \mathbf{M}, \mathcal{S}_m)$.
```

$N_f(s)$ as the number of episodes where state s is the first visited unknown state in $\Pi(M, \mathcal{S}_m)$ for any m. A state s is known if $N_f(s) \geq L$. By the definition of $\Pi(M, \mathcal{S}_m)$ and known states, if state s is known, then we have run Bernstein-SSP for at least L episodes starting from s and thus $\widetilde{D}_s$ satisfying $T^{\pi^f}(s) \leq \widetilde{D}_s$ can be computed. When a new known state is found, the algorithm enters into the next phase (that is, increment m). This construction also implies that $\mathcal{S}_m \subset \mathcal{S}_{m+1}$, $M_p \leq S$, and the diameter of $\Pi(M, \mathcal{S}_m)$ is upper bounded by $\max_{s\in\mathcal{S}_m} \widetilde{D}_s$.

3. When running Algorithm 2, we set the learning rate η using $\widetilde{D}_{s_0}$ and the parameter λ (for the skewed occupancy measure) using a doubling trick. Specifically, we replace the diameter D in λ by an estimate D_j to obtain λ_j, and run Algorithm 2 with λ_j in place of λ (starting with $D_1 = 2\widetilde{D}_{s_0}$). We double the estimate every time we realize that D_j is not an upper bound of D: suppose at the end of an episode, state s becomes a known state, and $\widetilde{D}_s > D_j$, then we set $D_{j+1} = 2\widetilde{D}_s$ and increase j by 1. Denote by I_p the value of j at the end of episode K. Clearly, $I_p = \mathcal{O}(\log_2 D)$.

4. The last important change is that instead of reinitializing the confidence sets and counters in each phase, we directly inherit them from the previous phase. Denote by $\widehat{q}_k^m$ the occupancy measure computed by Algorithm 2 in $\widetilde{\Pi}(M, \mathcal{S}_m)$ in episode k, and by q_k^m the occupancy measure of executing $\pi_{\widehat{q}_k^m}$ in $\widetilde{\Pi}(M, \mathcal{S}_m)$. Following the proof of Lemma 9, it is straightforward to verify that for any cost functions $\{c_k\}_{k=1}^K$ in $[0, 1]$, the following holds,

$$\sum_{k=1}^{K} |\langle q_k^m - \widehat{q}_k^m, c_k\rangle| \leq 32\sqrt{S^2A\ln(HK)\left(\sum_{k=1}^{K}\left\langle q_k^m, \vec{h} \circ c_k\right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^2A\right). \tag{14}$$

We summarize the ideas above in Algorithm 6. Now we proceed to show that Algorithm 6 ensures the same regret guarantee as Algorithm 2 without knowing D.

Theorem 4. *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 6 ensures $R_K = \tilde{\mathcal{O}}(\sqrt{S^2ADTK} + H^3S^3A + D^4S^4AH)$.*

Proof. When $K < \max\left\{DS^4A, \frac{D^4}{T^{\pi^f}(s_0)^4A}\right\}$, by the regret guarantee of Bernstein-SSP (Cohen et al., 2020), we have:

$$R_K \leq KH_1 + KD + \tilde{\mathcal{O}}\left(DS\sqrt{AK} + D^{\frac{3}{2}}S^2A\right) = \tilde{\mathcal{O}}\left(D^4S^4AH\right).$$

Otherwise, we have $L \geq \max\left\{\frac{2400D^2}{T^{\pi^f}(s_0)^2}\ln^3\frac{4K}{\delta}, S^2A\sqrt{D}\ln^2\frac{KDSA}{\delta}\right\}$ as desired. Denote by $N^m_k, \widetilde{N}^m_k$ the number of steps taken in $\Pi(M,\mathcal{S}_m), \widetilde{\Pi}(M,\mathcal{S}_m)$ in episode k respectively. We have $N^m_k < N_k$ only if the learner reaches an unknown state. Hence, There are at most at most SL episodes such that $N^m_k < N_k$, and $N_k - N^m_k$ is the number of steps of executing Bernstein-SSP after reaching an unknown state in episode k. Denote by q^m_π the occupancy measure of executing policy π in $\Pi(M,\mathcal{S}_m)$ or $\widetilde{\Pi}(M,\mathcal{S}_m)$ based on the policy π. We decompose the regret as follows:

$$\begin{aligned}
\sum_{k=1}^K \langle N_k - q_{\pi^\star}, c_k\rangle &= \sum_{k=1}^K \langle N^m_k - q_{\pi^\star}, c_k\rangle + \sum_{k=1}^K \langle N_k - N^m_k, c_k\rangle \\
&\leq \sum_{k=1}^K \langle N^m_k - q^m_{\pi^\star}, c_k\rangle + \tilde{\mathcal{O}}\left(S(DL + DS\sqrt{AL} + D^{3/2}S^2A)\right) \\
&\qquad \text{(by } q^m_{\pi^\star}(s,a) \leq q_{\pi^\star}(s,a) \text{ and the regret guarantee of Bernstein-SSP)} \\
&\leq \sum_{k=1}^K \left\langle \widetilde{N}^m_k - q^m_{\widetilde{\pi}^\star}, c_k\right\rangle + \tilde{\mathcal{O}}\left(DSL + D^2S^3A\right), \\
&\qquad \text{(by loop-free reduction (Lemma 1) on } \Pi(M,\mathcal{S}_m))
\end{aligned}$$

Note that using the last inequality above, we can already get a parameter-free algorithm with an extra S dependency by completely restarting Algorithm 2 in every new phase. To obtain a tighter bound, we need a more careful analysis on the sum of regret of all phases. Denote by $\mathcal{I}_m$ the set of episodes in phase m, and by $\mathcal{I}'_j$ the set of episodes using parameter λ_j. Note that we can treat j as a function of phase m, and m as a function of episode k. Following the arguments in the proof of Theorem 1, with probability at least $1-6\delta$,

$$\begin{aligned}
&\sum_{m=1}^{M_p}\sum_{k\in\mathcal{I}_m}\left\langle \widetilde{N}^m_k - q^m_{\widetilde{\pi}^\star}, c_k\right\rangle \\
&= \sum_{k=1}^K \left\langle \widetilde{N}^m_k - q^m_k, c_k\right\rangle + \sum_{k=1}^K \langle q^m_k - \widehat{q}^m_k, c_k\rangle + \sum_{m=1}^{M_p}\sum_{k\in\mathcal{I}_m}\langle \widehat{q}^m_k - q^m_{\widetilde{\pi}^\star}, c_k\rangle \\
&\leq \lambda_{I_p}\sum_{k=1}^K \left\langle q^m_k, \vec{h}\circ c_k\right\rangle + \tilde{\mathcal{O}}\left(\frac{1}{\lambda_{I_p}}\right) + 32\sqrt{S^2A\ln(HK)\left(\sum_{k=1}^K\left\langle q^m_k, \vec{h}\circ c_k\right\rangle + \tilde{\mathcal{O}}\left(H^3\sqrt{K}\right)\right)} + \tilde{\mathcal{O}}\left(H^3S^2A\right) \\
&\qquad \text{(Freedman's inequality and Eq. (14))} \\
&\quad + \sum_{m=1}^{M_p}\left(\tilde{\mathcal{O}}\left(\frac{T}{\eta} + \eta\sum_{k\in\mathcal{I}_m}\langle q^m_{\widetilde{\pi}^\star}, c_k\rangle\right) + 2\lambda_j\sum_{k\in\mathcal{I}_m}\left\langle q^m_{\widetilde{\pi}^\star}, \vec{h}\circ c_k\right\rangle - \lambda_j\sum_{k\in\mathcal{I}_m}\left\langle \widehat{q}^m_k, \vec{h}\circ c_k\right\rangle\right) \qquad \text{(by Eq. (9))} \\
&\leq \tilde{\mathcal{O}}\left(\sqrt{SDTK}\right) + \sum_{j=1}^{I_p}\left(2\lambda_j\sum_{k\in\mathcal{I}'_j}\left\langle q^m_{\widetilde{\pi}^\star}, \vec{h}\circ c_k\right\rangle + \tilde{\mathcal{O}}\left(\frac{S^2A}{\lambda_j}\right) + \lambda_j\sum_{k\in\mathcal{I}'_j}\left\langle q^m_k - \widehat{q}^m_k, \vec{h}\circ c_k\right\rangle\right) + \tilde{\mathcal{O}}\left(H^3S^2A\right) \\
&\qquad \text{(} |M_p| \leq S, \textstyle\sum_{k=1}^K\langle q^m_{\widetilde{\pi}^\star}, c_k\rangle \leq \sum_{k=1}^K\langle q_{\widetilde{\pi}^\star}, c_k\rangle \leq DK\text{, and AM-GM inequality)} \\
&\overset{\text{(i)}}{=} \tilde{\mathcal{O}}\left(\sqrt{SDTK} + \sum_{j=1}^{I_p}\sqrt{S^2AD_jTK} + H^3S^2A\right) = \tilde{\mathcal{O}}\left(\sqrt{S^2ADTK} + H^3S^2A\right).
\end{aligned}$$

In (i), we denote by $J^{m,\pi}_k(s,h)$ the expected cost of policy π starting from state (s,h) w.r.t cost c_k and $\widetilde{\Pi}(M,\mathcal{S}_m)$. Then, $\sum_{k=1}^K J^{m,\widetilde{\pi}^\star}_k(s,h) \leq \sum_{k=1}^K J^{\widetilde{\pi}^\star}_k(s,h)\mathbb{I}\{s\in\mathcal{S}_m\} \leq \sum_{k=1}^K(J^{\pi^\star}_k(s) + 3D_j)\mathbb{I}\{s\in\mathcal{S}_m\} \leq 4D_jK$ for all (s,h), and by

Lemma 2,

$$\sum_{k\in\mathcal{I}'_j}\left\langle q^m_{\widetilde{\pi}^\star}, \vec{h}\circ c_k\right\rangle \leq \sum_{k=1}^{K}\left\langle q^m_{\widetilde{\pi}^\star}, \vec{h}\circ c_k\right\rangle = \sum_{k=1}^{K}\left\langle q^m_{\widetilde{\pi}^\star}, J_k^{m,\widetilde{\pi}^\star}\right\rangle = \mathcal{O}\left(D_jTK\right).$$

Moreover, by Eq. (14),

$$\begin{aligned}\lambda_j\sum_{k\in\mathcal{I}'_j}\left\langle q^m_k-\widehat{q}^m_k, \vec{h}\circ c_k\right\rangle &= \lambda_j H\sum_{k\in\mathcal{I}'_j}\left\langle q^m_k-\widehat{q}^m_k, \frac{\vec{h}\circ c_k}{H}\right\rangle\\ &= \tilde{\mathcal{O}}\left(\lambda_j H\sqrt{S^2A\left(\sum_{k=1}^{K}\left\langle q^m_k, \vec{h}\circ c_k\right\rangle + H^3\sqrt{K}\right)} + \lambda_j H^4S^2A\right)\\ &= \tilde{\mathcal{O}}\left(\lambda_j H\sqrt{S^2AH^3K} + \lambda_j H^4S^2A\right) = \tilde{\mathcal{O}}\left(H^3S^2A\right).\\ &\qquad\qquad (\textstyle\sum_{k=1}^{K}\left\langle q^m_k, \vec{h}\circ c_k\right\rangle \leq H^2K,\ \lambda_j = 4\sqrt{\frac{S^2A}{D_jTK}}, \text{ and } K\geq 16S^2AH^2)\end{aligned}$$

Combining both cases, we have:

$$\begin{aligned}R_K &= \tilde{\mathcal{O}}\left(\sqrt{S^2ADTK} + H^3S^2A + DSL + D^2S^3A + D^4S^4AH\right)\\ &= \tilde{\mathcal{O}}\left(\sqrt{S^2ADTK} + H^3S^3A + DS\sqrt{AK} + D^4S^4AH\right)\\ &= \tilde{\mathcal{O}}\left(\sqrt{S^2ADTK} + H^3S^3A + D^4S^4AH\right). \qquad (D\leq T)\end{aligned}$$

This completes the proof. □

F. Learning without knowing $T_\star$ or $T_{\max}$

In this section, we discuss how to instantiate our proposed algorithms without knowing $T_\star$ or $T_{\max}$ (or both). We apply our ideas to Algorithm 2 to give concrete examples, and they are applicable to other proposed algorithms similarly. The modifications described below can be applied separately or jointly depending on the knowledge we have. Moreover, they are compatible with ideas in Section E for learning without knowing the SSP-diameter D.

Learning without knowing $T_\star$ We assume knowledge of $c_{\min} = \min_{s,a,k} c_k(s,a)$ and $c_{\min} > 0$, which is the assumption made in (Rosenberg and Mansour, 2020). Then by the inequality $T_\star \leq \frac{T^{\pi^f}(s_0)}{c_{\min}}$, it suffices to obtain an upper bound of $T^{\pi^f}(s_0)$ to obtain an upper bound of $T_\star$. To this end, we first run a Bernstein-SSP instance for $L = \tilde{\mathcal{O}}(\sqrt{AK})$ episodes with uniform costs equal to 1 and obtain $\widetilde{D}_{s_0}$, where both L and $\widetilde{D}_{s_0}$ are defined in Section E. Then, we simply run Algorithm 2 with $T = \widetilde{D}_{s_0}/c_{\min}$. Following the arguments in Section E, we know that the extra costs of estimating $\widetilde{D}_{s_0}$ in the regret is $\tilde{\mathcal{O}}(DL + D^{3/2}S^2A + D^4S^4AH) = \tilde{\mathcal{O}}(D\sqrt{AK} + D^4S^4AH)$. Thus, we obtain the following result:

Theorem 5. *If $H_1 \geq 8T_{\max}\ln K$, and $K \geq 16S^2AH^2$, then with probability at least $1-6\delta$, the algorithm described above ensures $R_K = \tilde{\mathcal{O}}(D\sqrt{\frac{S^2AK}{c_{\min}}} + H^3S^2A + D^4S^4AH)$.*

Compared to (Rosenberg and Mansour, 2020), the bound above is $\sqrt{\frac{1}{c_{\min}}}$ better in the dominating term. When $c_{\min} = 0$, similarly to (Rosenberg and Mansour, 2020), we can solve a modified MDP with perturbed cost functions $c'_k(s,a) = \max\{c_k(s,a), \epsilon\}$, where $\epsilon = K^{-1/3}$. The bias brought by the perturbation is of order $\tilde{\mathcal{O}}\left(\epsilon T_\star K\right)$, and thus the overall regret is $\tilde{\mathcal{O}}(K^{2/3})$ ignoring other parameters and constant terms. This is asymptotically better than the $\tilde{\mathcal{O}}\left(K^{3/4}\right)$ regret in (Rosenberg and Mansour, 2020) for $c_{\min} = 0$. We conclude that Algorithm 2 improves over previous work even without knowledge of $T_\star$.

Learning without knowing $T_{\max}$ Similarly to (Chen et al., 2020), we run Algorithm 2 with $H_1 = 8(K/S^2A)^{1/6}\ln K$. Note that when $K \leq T^6_{\max}S^2A$, by the regret guarantee of Bernstein-SSP, we have:

$$R_K \leq KH_1 + KD + \tilde{\mathcal{O}}\left(DS\sqrt{AK} + D^{\frac{3}{2}}S^2A\right) = \tilde{\mathcal{O}}\left(T^7_{\max}S^2A\right).$$

Otherwise, $H_1 \geq T_{\max}$, and by the regret guarantee of Algorithm 2, we have:

$$R_K = \tilde{\mathcal{O}}\left(\sqrt{S^2ADTK} + H^3S^2A\right) = \tilde{\mathcal{O}}\left(\sqrt{S^2ADTK}\right).$$

Combining these two cases, we have the following result:

Theorem 6. *If $T \geq T_\star + 1$, and $K \geq 16S^2AH^2$, then with probability at least $1 - 6\delta$, running Algorithm 2 with $H_1 = 8(K/S^2A)^{1/6} \ln K$ ensures $R_K = \tilde{\mathcal{O}}(\sqrt{S^2ADTK} + T_{\max}^7 S^2A)$.*

Applying both modifications to Algorithm 2, we obtain: $R_K = \tilde{\mathcal{O}}\left(D\sqrt{\frac{S^2AK}{c_{\min}}} + D^4S^4AK^{1/6} + T_{\max}^7 S^2A\right)$, which is still asymptomatically better compared to (Rosenberg and Mansour, 2020).

G. Concentration Inequalities

In this section, for a sequence of random variables $\{X_i\}_{i=1}^\infty$ adapted to a filtration $\{\mathcal{F}_i\}_{i=1}^\infty$, we define $\mathbb{E}_i[X_i] = \mathbb{E}[X_i|\mathcal{F}_i]$.

Lemma 20. *(Azuma's inequality) Let $X_{1:n}$ be a martingale difference sequence and $|X_i| \leq B$ holds for $i = 1, \ldots, n$ and some fixed $B > 0$. Then, with probability at least $1 - \delta$:*

$$\left|\sum_{i=1}^n X_i\right| \leq B\sqrt{2n \ln \frac{2}{\delta}}.$$

Lemma 21. *(A version of Freedman's inequality from (Beygelzimer et al., 2011)) Let $X_{1:n}$ be a martingale difference sequence and $X_i \leq B$ holds for $i = 1, \ldots, n$ and some fixed $B > 0$. Denote $V = \sum_{i=1}^n \mathbb{E}_i[X_i^2]$. Then, for any $\lambda \in [0, 1/B]$, with probability at least $1 - \delta$:*

$$\sum_{i=1}^n X_i \leq \lambda V + \frac{\ln(1/\delta)}{\lambda}.$$

Lemma 22. *(A version of anytime Bernstein's inequality from (Cohen et al., 2020, Theorem D.3)) Let $X_{1:n}$ be a sequence of i.i.d. random variables with expectation μ. Assume $X_n \in [0, B]$ almost surely. Then with probability $1 - \delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\left|\sum_{i=1}^n (X_i - \mu)\right| \leq 2\sqrt{B\mu n \ln \frac{2n}{\delta}} + B \ln \frac{2n}{\delta}.$$

$$\left|\sum_{i=1}^n (X_i - \mu)\right| \leq 2\sqrt{B\sum_{i=1}^n X_i \ln \frac{2n}{\delta}} + 7B \ln \frac{2n}{\delta}.$$

Lemma 23. *(Strengthened Freedman's inequality from (Lee et al., 2020, Theorem 2.2)) Let $X_{1:n}$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ such that $\mathbb{E}[X_i|\mathcal{F}_i] = 0$. Suppose $B_i \in [1, b]$ for a fixed constant b is $\mathcal{F}_i$-measurable and such that $X_i \leq B_i$ holds almost surely. Then with probability at least $1 - \delta$ we have*

$$\sum_{i=1}^n X_i \leq C\left(\sqrt{8V \ln(C/\delta)} + 2B^\star \ln(C/\delta)\right),$$

where $V = \max\left\{1, \sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_i]\right\}$, $B^\star = \max_{i \in [n]} B_i$, and $C = \lceil \log_2(b) \rceil \lceil \log_2(nb^2) \rceil$.

Lemma 24. *Let $\{X_i\}_{i=1}^\infty$ be a sequence of random variables adapted to the filtration $\{\mathcal{F}_i\}_{i=1}^\infty$, and $0 \leq X_i \leq B$ almost surely. Then with probability at least $1 - \delta$, for all $n \geq 1$ simultaneously:*

$$\sum_{i=1}^n \mathbb{E}_i[X_i] \leq 2\sum_{i=1}^n X_i + 4B \ln \frac{2n}{\delta}.$$