Proceedings of Machine Learning Research vol 178:1–65, 2022 35th Annual Conference on Learning Theory

Policy Optimization for Stochastic Shortest Path

Liyu Chen LIYUC@USC.EDU
University of Southern California

Haipeng Luo HAIPENGL@USC.EDU
University of Southern California

Aviv Rosenberg AVIVROS007@GMAIL.COM
Tel-Aviv University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Policy optimization is among the most popular and successful reinforcement learning algorithms, and there is increasing interest in understanding its theoretical guarantees. In this work, we initiate the study of policy optimization for the stochastic shortest path (SSP) problem, a goal-oriented reinforcement learning model that strictly generalizes the finite-horizon model and better captures many applications. We consider a wide range of settings, including stochastic and adversarial environments under full information or bandit feedback, and propose a policy optimization algorithm for each setting that makes use of novel correction terms and/or variants of dilated bonuses (Luo et al., 2021). For most settings, our algorithm is shown to achieve a near-optimal regret bound.

One key technical contribution of this work is a new approximation scheme to tackle SSP problems that we call *stacked discounted approximation* and use in all our proposed algorithms. Unlike the finite-horizon approximation that is heavily used in recent SSP algorithms, our new approximation enables us to learn a near-stationary policy with only logarithmic changes during an episode and could lead to an exponential improvement in space complexity.

1. Introduction

Stochastic Shortest Path (SSP) is a goal-oriented reinforcement learning setting, where a learner tries to reach a goal state with minimum total cost. Compared to the heavily studied finite-horizon setting, SSP is often a better model for capturing many real-world applications such as games, car navigation, robotic manipulations, and others. We study the online learning problem in SSP, where the learner interacts with an environment with unknown cost and transition function for multiple episodes. In each episode, the learner starts from an initial state, sequentially takes an action, incurs a cost, and transits to the next state until the goal state is reached. The goal of the learner is to achieve low regret, defined as the difference between her total cost and the expected cost of the optimal policy. A unique challenge of learning SSP is to trade off between two objectives: reaching the goal state and minimizing the cost. Indeed, neither reaching the goal as fast as possible nor minimizing the cost alone solves the problem.

Policy Optimization (PO) is among the most popular methods in reinforcement learning due to its strong empirical performance and favorable theoretical properties. Unlike value-based approaches such as Q learning, PO-type methods directly optimize the policy in an incremental manner. Many widely used practical algorithms fall into this category, such as REINFORCE (Williams,

1992), NPG (Kakade, 2001), and TRPO (Schulman et al., 2015). They are also easy to implement and computationally efficient compared to other methods such as those operating over the occupancy measure space (e.g., (Zimin and Neu, 2013)). From a theoretical perspective, PO is a general framework that works for different types of environments, including stochastic costs or even adversarial costs (Shani et al., 2020), function approximation (Cai et al., 2020), and non-stationary environments (Fei et al., 2020). Despite its popularity in applications, most theoretical works on PO focus on simple models such as finite-horizon models (Cai et al., 2020; Shani et al., 2020; Luo et al., 2021) and discounted models (Liu et al., 2019; Wang et al., 2020; Agarwal et al., 2021), which are often oversimplifications of real-life applications. In particular, PO methods have not been applied to regret minimization in SSP as far as we know.

Motivated by this gap, in this work, we systematically study policy optimization in SSP. We consider a wide range of different settings and for each of them discuss how to design a policy optimization algorithm with a strong regret bound. Specifically, our main results are as follows:

- In Section 3, we first propose an important technique used in all our algorithms: *stacked discounted approximation*. It reduces any SSP instance to a special Markov Decision Process (MDP) with a stack of $\mathcal{O}(\ln K)$ layers (K is the total number of episodes), each of which contains a discounted MDP (hence the name) such that the learner stays in the same layer with a certain probability γ and proceeds to the next layer with probability $1-\gamma$. This approximation not only resolves the difficulty of having dynamic and potentially unbounded episode lengths in the PO analysis, but more importantly leads to a near-stationary policies with only $\mathcal{O}(\ln K)$ changes within an episode. Compared to the commonly used finite-horizon approximation (Chen et al., 2021d; Chen and Luo, 2021; Cohen et al., 2021) which changes the policy at every step of an episode, our approach could lead to an exponential improvement in space complexity and is also more natural since the optimal policy for SSP is indeed stationary.
- Building on the stacked discounted approximation, in Section 4, we design PO algorithms for two types of stochastic environments considered in the literature. In the first type (called stochastic costs), the cost for each visit of a state-action pair is an i.i.d. sample of an unknown distribution and is revealed to the learner immediately after the visit. Our algorithm achieves $\tilde{\mathcal{O}}(B_\star S\sqrt{AK})$ regret in this case, close to the minimax bound $\tilde{\mathcal{O}}(B_\star\sqrt{SAK})$ (Cohen et al., 2021), where S is the number of states, A is the number of actions, and $B_\star$ is the maximum expected cost of the optimal policy starting from any states. In the second type (called stochastic adversary following (Chen and Luo, 2021)), the cost function for each episode is fixed and an i.i.d. sample of an unknown distribution, and only at the end of the episode, the learner observes the entire cost function (full-information feedback) or the costs for all visited state-action pairs (bandit feedback). Our algorithm achieves $\tilde{\mathcal{O}}(\sqrt{DT_\star K}+DS\sqrt{AK})$ regret with full information and $\tilde{\mathcal{O}}(\sqrt{DT_\star SAK}+DS\sqrt{AK})$ regret with bandit feedback, where D is the diameter of the MDP and $T_\star$ is the expected hitting time of the optimal policy starting from the initial state. These bounds match the best existing results from (Chen and Luo, 2021) (and exhibit a $\sqrt{S}$ gap in the second term $DS\sqrt{AK}$ compared to their lower bounds).
- Finally, in Section 5, we further study SSP with adversarial costs and design PO algorithms that achieve $\tilde{\mathcal{O}}(T_\star\sqrt{DK}+\sqrt{DT_\star S^2AK})$ regret with full information and $\tilde{\mathcal{O}}(\sqrt{T_{\max}^5 S^2AK})$ regret with bandit feedback, where $T_{\max}$ is the maximum expected hitting time of the optimal policy over all states. The best existing bounds for these settings are $\tilde{\mathcal{O}}(\sqrt{DT_\star S^2AK})$ and $\tilde{\mathcal{O}}(\sqrt{DT_\star S^3A^2K})$ respectively (Chen and Luo, 2021).

Table 1: Comparison of regret bound, time complexity, and space complexity of different SSP algorithms. We consider five feedback types: SC (stochastic costs), SAF (Stochastic Adversary, Full information), SAB (Stochastic Adversary, Bandit feedback), AF (Adversarial, Full information), and AB (Adversarial, Bandit feedback). Operator $\tilde{\mathcal{O}}(\cdot)$ is hidden for simplicity. Time complexity of $\text{poly}(S, A, T_{\max})$ is due to optimization in the occupancy measure space.

	Regret	Time	Space	Feedback
Cohen et al. (2021)	$B_\star\sqrt{SAK}$	$S^3A^2T_{\max}$	$S^2AT_{\max}$	SC
Our work	$B_\star S\sqrt{AK}$	$S^2AT_{\max}K$	S^2A	
Chen and Luo (2021)	$\sqrt{DT_\star K} + DS\sqrt{AK}$	$\text{poly}(S, A, T_{\max}) \cdot K$	$S^2AT_{\max}$	SAF
Our work	$\sqrt{DT_\star K} + DS\sqrt{AK}$	$S^2AT_{\max}K$	S^2A	
Chen and Luo (2021)	$\sqrt{SADT_\star K} + DS\sqrt{AK}$	$\text{poly}(S, A, T_{\max}) \cdot K$	$S^2AT_{\max}$	SAB
Our work	$\sqrt{SADT_\star K} + DS\sqrt{AK}$	$S^2AT_{\max}K$	S^2A	
Chen and Luo (2021)	$\sqrt{S^2ADT_\star K}$	$\text{poly}(S, A, T_{\max}) \cdot K$	$S^2AT_{\max}$	AF
Our work	$\sqrt{(S^2A + T_\star)DT_\star K}$	$S^2AT_{\max}K$	S^2A	
Chen and Luo (2021)	$\sqrt{S^3A^2DT_\star K}$	$\text{poly}(S, A, T_{\max}) \cdot K$	$S^2AT_{\max}$	AB
Our work	$\sqrt{S^2AT_{\max}^5 K}$	$\text{poly}(S, A, T_{\max}) \cdot K$	S^2A	

We also include Table 1 with a comprehensive comparison between SSP algorithms for a better understanding of our contributions. While our regret bounds do not always match the state-of-the-art, we emphasize again that our algorithms are more space-efficient due to the stacked discounted approximation (S^2A versus $S^2AT_{\max}$ in Table 1). It is also more time-efficient in some cases (for feedback types SAF, SAB, and AF in Table 1). We also note that in the analysis of stacked discounted approximation, a regret bound starting from any state (not just the initial state) is important, and PO indeed provides such a guarantee while other methods based on occupancy measure do not. In other words, PO is especially compatible with our stacked discounted approximation. Moreover, our results also significantly improve our theoretical understanding on PO, and pave the way for future study on more challenging problems such as SSP with function approximation, where in some cases PO is the only method known to be computationally and statistically efficient (Luo et al., 2021).

Other Techniques To achieve our results for stochastic environments, we make two other technical contributions. First, in order to control the cost estimation error optimally, we derive a set of novel correction terms fed to the PO algorithm, which resolves some technical difficulties brought by PO due to its lack of optimism and also greatly simplifies the analysis. Second, due to the soft policy updates, the standard PO analysis leads to an undesirable dominating term related to $T_\star$ or even $T_{\max}$ in the regret, and we develop a refined analysis on the value difference between learner's policies and the optimal policy to reduce this to a lower order term.

To achieve our results for adversarial environments, we develop a tighter variance-aware bound for the stability term in the PO analysis, which plays a key role in removing the $T_{\max}$ dependency in the dominating term of the regret bound in the full information setting. We further extend the dilated bonuses of (Luo et al., 2021) (for the finite-horizon setting) to the stacked discounted MDPs, which is essential for both the full information setting and the bandit feedback setting.

Related Work Regret minimization in SSP has received much attention recently for both stochastic environment (Tarbouriech et al., 2020; Cohen et al., 2020, 2021; Tarbouriech et al., 2021; Chen et al., 2021a,b; Jafarnia-Jahromi et al., 2021) and adversarial environment (Rosenberg and Mansour, 2021; Chen et al., 2021d; Chen and Luo, 2021). All previous approaches are either value-based (e.g. Q learning) or occupancy-measure-based, while we take the first step in studying the more practical and versatile PO methods. Among numerous studies on PO, the closest to our work are the recent ones by Shani et al. (2020) and Luo et al. (2021) for the special case of finite-horizon MDPs.

The use of variance information (Lattimore and Hutter, 2012; Azar et al., 2017; Zhou et al., 2021; Zhang et al., 2021; Kim et al., 2021) and correction terms (Steinhardt and Liang, 2014; Wei and Luo, 2018; Chen et al., 2021c) is crucial for achieving optimal and adaptive regret bound in online learning. In this work we heavily make use of these ideas as mentioned.

2. Preliminaries

An SSP instance is defined by a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, s_{\text{init}}, g, \mathcal{A}, P)$. Here, $\mathcal{S}$ is the state space, $s_{\text{init}} \in \mathcal{S}$ is the initial state, $g \notin \mathcal{S}$ is the goal state, $\mathcal{A}$ is the action space, and $P = \{P_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ with $P_{s,a} \in \Delta_{\mathcal{S}_+}$ is the transition function, where $\mathcal{S}_+ = \mathcal{S} \cup \{g\}$ and $\Delta_{\mathcal{S}_+}$ is the simplex over $\mathcal{S}_+$.

The learning protocol is as follows: the learner interacts with the environment for K episodes. In episode k, the learner starts in initial state s_{init}, sequentially takes an action, incurs a cost (which might not be observed immediately), and transits to the next state until the goal state g is reached. Formally, at the i-th step of episode k, the learner observes state s_i^k (with $s_1^k = s_{\text{init}}$), takes action a_i^k, suffers cost c_i^k, and transits to the next state $s_{i+1}^k \sim P_{s_i^k, a_i^k}$. Denote by I_k the length of episode k, such that $s_{I_k+1}^k = g$ when I_k is finite. Note that the heavily studied finite-horizon setting is a special case of SSP where I_k is always guaranteed to be some fixed number.

Proper Policies and Related Concepts At a high level, the learner's goal is to reach the goal state with minimum cost. Thus, we focus on *proper policies*: a stationary policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is a mapping that assigns to each state a distribution over actions, and it is proper if following π from any initial state reaches the goal state with probability 1. Denote by Π the set of proper policies (assumed to be non-empty). Given a proper policy π, a transition function P, and a cost function $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$, we define its value function and action-value function as follows: $V^{\pi,P,c}(s) = \mathbb{E}\left[\sum_{i=1}^{I} c(s_i, a_i) \middle| \pi, P, s_1 = s\right]$ and $Q^{\pi,P,c}(s,a) = c(s,a) + \mathbb{E}_{s'\sim P_{s,a}}[V^{\pi,P,c}(s')]$, where the expectation in $V^{\pi,P,c}$ is over the randomness of action $a_i \sim \pi(\cdot|s_i)$, next state $s_{i+1} \sim P_{s_i,a_i}$, and the number of steps I before reaching g. Also define the *advantage function* $A^{\pi,P,c}(s,a) = Q^{\pi,P,c}(s,a) - V^{\pi,P,c}(s)$.

We consider two types of environments: stochastic environments and adversarial environments, which differ in the way costs are generated (and revealed), discussed in detail below.

Stochastic Environments We start with the simpler environment with a fixed "ground truth" cost: there exists an unknown mean cost function $c : \mathcal{S} \times \mathcal{A} \to [c_{\min}, 1]$, and the costs incurred by the learner are i.i.d samples from some distribution with support $[c_{\min}, 1]$ and mean c. Here, $c_{\min} \in [0, 1]$ is a global lower bound.[1] We consider the following three types of cost feedback.

1. Unlike many previous works for stochastic costs that require $c_{\min} > 0$ in their analysis, our methods allow $c_{\min} = 0$.

1. **Stochastic costs:** whenever the learner visits state-action pair (s, a), she immediately observes (and incurs) an i.i.d cost sampled from some unknown distribution with mean $c(s, a)$.

2. **Stochastic adversary, full information:** before learning starts, an adversary samples K i.i.d. cost functions $\{c_k\}_{k=1}^K$ from some unknown distribution with mean c. At the i-th step of episode k, the learner incurs cost $c_i^k = c_k(s_i^k, a_i^k)$. Only at the end of this episode (after the goal state is reached), the learner observes the entire cost function c_k.

3. **Stochastic adversary, bandit feedback:** this is the same as above, except that at the end of episode k, the learner only observes the costs of all visited state-action pairs: $\{c_k(s_i^k, a_i^k)\}_{i=1}^{I_k}$.

The learner's objective is to minimize her regret, defined as the difference between her total incurred cost and the total expected cost of the best proper policy: $R_K = \sum_{k=1}^K \sum_{i=1}^{I_k} c_i^k - K \cdot V^{\pi^\star, P, c}(s_{\text{init}})$, where $\pi^\star$ is the optimal proper policy satisfying $\pi^\star \in \operatorname{argmin}_{\pi \in \Pi} V^{\pi, P, c}(s)$ for all $s \in \mathcal{S}$.

Adversarial Environments We also consider the more challenging environment that adapts to learner's behavior in a possibly malicious manner. Specifically, in episode k, the environment decides an arbitrary cost function $c_k : \mathcal{S} \times \mathcal{A} \to [c_{\min}, 1]$ which could depend on the learner's algorithm as well as her randomness before episode k. The learner then suffers cost $c_i^k = c_k(s_i^k, a_i^k)$ at the i-th step of episode k. Similarly to the stochastic adversary case, the learner observes information on c_k only after she reaches the goal state in episode k, and she observes the entire c_k in the full-information setting or just the cost of visited state-action pairs $\{c_k(s_i^k, a_i^k)\}_{i=1}^{I_k}$ in the bandit setting. The objective is again to minimize her regret against the optimal proper policy in hindsight: $R_K = \sum_{k=1}^K \left(\sum_{i=1}^{I_k} c_i^k - V^{\pi^\star, P, c_k}(s_{\text{init}}) \right)$, where we overload the notation $\pi^\star$ to denote the overall optimal proper policy such that $\pi^\star \in \operatorname{argmin}_{\pi \in \Pi} \sum_{k=1}^K V^{\pi, P, c_k}(s)$ for all $s \in \mathcal{S}$.

Key Parameters and Notations Let $T^\pi(s)$ be one plus the expected number of steps to reach the goal if one follows policy π starting from state s. Four parameters play a key role in our analysis and regret bounds: $B_\star = \max_s V^{\pi^\star, P, c}(s)$, the maximum expected cost of the optimal policy starting from any state; $T_\star = T^{\pi^\star}(s_{\text{init}})$, the hitting time of the optimal policy starting from the initial state; $T_{\max} = \max_s T^{\pi^\star}(s)$, the maximum hitting time of the optimal policy starting from any state; and $D = \max_s \min_\pi T^\pi(s)$, the SSP-diameter. We also define the *fast policy* π_f such that $\pi_f \in \operatorname{argmin}_\pi T^\pi(s)$ for all state s. Similarly to previous works, in most discussions we assume the knowledge of all four parameters and the fast policy, and defer to Appendix E what we can achieve when some of these are unknown. We also assume $B_\star \geq 1$ for simplicity.

For $n \in \mathbb{N}_+$, we define $[n] = \{1, \ldots, n\}$. $\mathbb{E}_k[\cdot]$ denotes the conditional expectation given everything before episode k. The notation $\widetilde{\mathcal{O}}(\cdot)$ hides all logarithmic terms including $\ln K$ and $\ln \frac{1}{\delta}$ for some confidence level $\delta \in (0, 1)$. For a distribution $\widetilde{P} \in \Delta_{\mathcal{S}_+}$ and a function $V : \mathcal{S}_+ \to \mathbb{R}$, define $\widetilde{P}V = \mathbb{E}_{s \sim \widetilde{P}}[V(s)]$.

3. Stacked Discounted Approximation and Algorithm Template

Policy optimization algorithm have been naturally derived in many MDP models. In the finite-horizon setting, one can update the policy at the end of each episode using the cost for this episode that is always bounded. In the discounted setting or average reward setting with some ergodic assumption, one can also update the policy after a certain fixed number of steps since the short-term information is enough to predict the long-term behavior reasonably well. However, this is

not possible in SSP: the hitting time of an arbitrary policy can be arbitrarily large in SSP, and only looking at a fixed number of steps can not always provide accurate information.

A natural solution would be to approximate SSP by other MDP models, and then apply PO in the reduced model. Approximating SSP instances by finite-horizon MDPs (Chen et al., 2021a,d; Cohen et al., 2021) or discounted MDPs (Tarbouriech et al., 2021; Min et al., 2021) is a common practice in the literature, but both have their pros and cons. Finite-horizon approximation shrinks the estimation error exponentially fast and usually leads to optimal regret (Chen et al., 2021d; Cohen et al., 2021). However, it greatly increases the space complexity of the algorithm as it needs to store non-stationary policies with horizon of order $\tilde{\mathcal{O}}(T_{\max})$ or $\tilde{\mathcal{O}}(\frac{B_\star}{c_{\min}})$ (as shown in Table 1). Discounted approximation, on the other hand, produces stationary policies, but the estimation error decreases only linearly in the effective horizon $(1-\gamma)^{-1}$, where γ is the discounted factor. This often leads to sub-optimal regret bounds and large time complexity (Tarbouriech et al., 2021). We include a detailed discussion on limitations of existing approximation schemes in Appendix B.1. These issues greatly limit the practical potential of these methods, and PO methods built on top of them would be less interesting.

To address these issues and achieve optimal regret with small space complexity, we introduce a new approximation scheme called *Stacked Discounted Approximation*, which is a hybrid of finite-horizon and discounted approximations. The key idea is as follows: the finite-horizon approximation requires a horizon of order $\mathcal{O}(T_{\max} \ln K)$, but one can imagine that policies at nearby layers are close to each other and can be approximated by one stationary policy. Thus, we propose to achieve the best of both worlds by dividing the layers into $\mathcal{O}(\ln K)$ parts and performing discounted approximation within each part with an effective horizon $\mathcal{O}(T_{\max})$. Formally, we define the following.

Definition 1 *For an SSP instance $\mathcal{M} = (\mathcal{S}, s_{\text{init}}, g, \mathcal{A}, P)$, we define, for number of layers H, discounted factor γ, and terminal cost c_f, another SSP instance $\mathring{\mathcal{M}} = (\mathring{\mathcal{S}}, \mathring{s}_{init}, g, \mathcal{A}, \mathring{P})$ as follows:*

1. *$\mathring{\mathcal{S}} = \mathcal{S} \times [H+1]$, $\mathring{s}_{init} = (s_{\text{init}}, 1)$, and the goal state g remains the same.*
2. *Transition from (s,h) to (s',h') is only possible for $h' \in \{h, h+1\}$: for any $h \le H$ and $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have $\mathring{P}_{(s,h),a}(s',h) = \gamma P_{s,a}(s')$ (stay in the same layer with probability γ), $\mathring{P}_{(s,h),a}(s',h+1) = (1-\gamma)P_{s,a}(s')$ (proceed to the next layer with probability $1-\gamma$), and $\mathring{P}_{(s,h),a}(g) = P_{s,a}(g)$; for $h = H+1$, we have $\mathring{P}_{(s,H+1),a}(g) = 1$ for any (s,a) (immediately reach the goal if at layer $H+1$). For notational convenience, we also write $\mathring{P}_{(s,h),a}(s',h')$ as $P_{(s,h),a}(s',h')$ or $P_{s,a,h}(s',h')$, and $\mathring{P}_{(s,h),a}(g)$ as $P_{(s,h),a}(g)$ or $P_{s,a,h}(g)$.*
3. *For any cost function $c : \mathcal{S} \times \mathcal{A} \to [0,1]$ in $\mathcal{M}$, we define a cost function $\mathring{c}$ for $\mathring{\mathcal{M}}$ such that $\mathring{c}((s,h),a) = c(s,a)$ for $h \in [H]$ and $\mathring{c}((s,H+1),a) = c_f$ (terminal cost). For notational convenience, we also write $\mathring{c}((s,h),a)$ as $c((s,h),a)$ or $c(s,a,h)$.*

For any stationary policy π in $\mathring{\mathcal{M}}$, we write $\pi(a|(s,h))$ as $\pi(a|s,h)$, and we often abuse the notation $Q^{\pi,P,c}$ and $V^{\pi,P,c}$ to represent the value functions with respect to policy π, transition $\mathring{P}$, and cost function $\mathring{c}$. We also often use (s,a,h) in place of $((s,h),a)$ for function input, that is, we write $f((s,h),a)$ as $f(s,a,h)$.

Define $\mathring{\pi}^\star$ for $\mathring{\mathcal{M}}$ that mimics the behavior of $\pi^\star$, in the sense that $\mathring{\pi}^\star(\cdot|s,h) = \pi^\star(\cdot|s)$. If we set $\gamma = 1 - \frac{1}{2T_{\max}}$, by the definition of $T_{\max}$, it can be shown that the probability of $\mathring{\pi}^\star$ transiting to the next layer before reaching g is upper bounded by $1/2$. If we further set $H = \mathcal{O}(\ln K)$, then

the probability of transiting to the $(H+1)$-th layer before reaching g is at most $\frac{1}{2^H} = \tilde{\mathcal{O}}(1/K)$. As a result, the estimation error decreases exponentially in the number of layers while the policy only changes for $\mathcal{O}(\ln K)$ many times. More importantly, due to the discounted factor, the expected hitting time of any policy is of order $\mathcal{O}(\frac{H}{1-\gamma}) = \mathcal{O}(T_{\max} \ln K)$, which controls the cost of exploration and enables the learner to only update its policy at the end of an episode. We summarize the intuition above in the following lemma.

Lemma 2 *For any cost function $c : \mathcal{S} \times \mathcal{A} \to [0,1]$ and terminal cost c_f, we have $V^{\pi,P,c}(s,h) \le \frac{H-h+1}{1-\gamma} + c_f$ for any $h \in [H], s \in \mathcal{S}$, and policy π in $\mathring{\mathcal{M}}$. Moreover, if $\gamma = 1 - \frac{1}{2T_{\max}}$, we further have $Q^{\mathring{\pi}^\star,P,c}(s,a,h) \le Q^{\pi^\star,P,c}(s,a) + \frac{c_f}{2^{H-h+1}}$ for any $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

Proof The first statement is because in expectation it takes any policy $\frac{1}{1-\gamma}$ steps to transit from one layer to the next and each step incurs at most 1 cost (except for the terminal cost). For the second statement, note that $V^{\pi,P,c}(s,H+1) = Q^{\pi,P,c}(s,a,H+1) = c_f$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, and for any $h \in [H]$, $V^{\pi,P,c}(s,h) = \sum_{a \in \mathcal{A}} \pi(a|s,h) Q^{\pi,P,c}(s,a,h)$ and

$$Q^{\pi,P,c}(s,a,h) = c(s,a) + \gamma P_{s,a} V^{\pi,P,c}(\cdot,h) + (1-\gamma) P_{s,a} V^{\pi,P,c}(\cdot,h+1),$$

where we abuse the notation and define $V^{\pi,P,c}(g,h) = 0$ for all $h \in [H+1]$. Now we prove the second statement by induction for $h = H+1, \ldots, 1$. The base case $h = H+1$ is clearly true. For $h \le H$, we bound $Q^{\mathring{\pi}^\star,P,c}(s,a,h) - Q^{\pi^\star,P,c}(s,a)$ as follows:

$$\begin{aligned}
&\gamma P_{s,a} V^{\mathring{\pi}^\star,P,c}(\cdot,h) + (1-\gamma) P_{s,a} V^{\mathring{\pi}^\star,P,c}(\cdot,h+1) - P_{s,a} V^{\pi^\star,P,c} \\
&\le \gamma P_{s,a} (V^{\mathring{\pi}^\star,P,c}(\cdot,h) - V^{\pi^\star,P,c}) + (1-\gamma)\frac{c_f}{2^{H-h}} \\
&\qquad\qquad (V^{\mathring{\pi}^\star,P,c}(s,h+1) - V^{\pi^\star,P,c}(s) \le \tfrac{c_f}{2^{H-h}} \text{ by induction}) \\
&= \gamma \mathbb{E}_{s' \sim P_{s,a}, a' \sim \pi^\star(s')} \left[Q^{\mathring{\pi}^\star,P,c}(s',a',h) - Q^{\pi^\star,P,c}(s',a') \right] + (1-\gamma)\frac{c_f}{2^{H-h}}.
\end{aligned}$$

By repeating the arguments above, we arrive at

$$Q^{\mathring{\pi}^\star,P,c}(s,a,h) - Q^{\pi^\star,P,c}(s,a) \le \mathbb{E}\left[\sum_{t=1}^{I} \gamma^{t-1}(1-\gamma)\frac{c_f}{2^{H-h}} \,\middle|\, \pi^\star, P, s_1 = s, a_1 = a \right],$$

where I is the (random) number of steps it takes for $\pi^\star$ to reach the goal in $\mathcal{M}$ starting from (s,a). Bounding γ^{t-1} by 1 and $\mathbb{E}[I]$ by $T_{\max}$, we then obtain the upper bound $\frac{(1-\gamma)T_{\max} c_f}{2^{H-h}} = \frac{c_f}{2^{H-h+1}}$, which finishes the induction. ■

Remark 3 *Applying the first statement of Lemma 2 with $c(s,a) = 1$ and $c_f = 1$, we have the expected hitting time of any policy in $\mathring{\mathcal{M}}$ bounded by $\frac{H}{1-\gamma} + 1$ starting from any state in any layer.*

Now we complete the approximation by showing how to solve the original problem via solving its stacked discounted version. Given a policy π for $\mathring{\mathcal{M}}$, define a non-stationary randomized policy $\sigma(\pi)$ for $\mathcal{M}$ as follows: it maintains an internal counter h initialized as 1. In each time step before reaching the goal, it first follows $\pi(\cdot|s,h)$ for one step, where s is the current state. Then, it samples a Bernoulli random variable X with mean γ, and it increases h by 1 if $X = 0$. When $h = H+1$,

Algorithm 1 Template for Policy Optimization with Stacked Discounted Approximation

Initialize: $\mathcal{P}_1$, the set of all possible transition functions in $\mathring{\mathcal{M}}$ (Eq. (3)); $\eta > 0$, some learning rate.
for $k = 1, \dots, K$ **do**
  Compute $\pi_k(a|s,h) \propto \exp\left(-\eta \sum_{j=1}^{k-1} (\widetilde{Q}_j(s,a,h) - B_j(s,a,h))\right)$.
  Execute $\sigma(\pi_k)$ for one episode (see the paragraph before Lemma 4).
  Compute some optimistic action-value estimator $\widetilde{Q}_k$ and exploration bonus function B_k using $\mathcal{P}_k$ and observations from episode k.
  Compute transition confidence set $\mathcal{P}_{k+1}$, as defined in Eq. (4).

it executes the fast policy π_f until reaching the goal state. Clearly, the trajectory of $\sigma(\pi)$ indeed follows the same distribution of the trajectory of π in $\mathring{\mathcal{M}}$. We show that as long as H is large enough and c_f is of order $\tilde{\mathcal{O}}(D)$, this reduction makes sure that the regret between these two problems are similar. The proof is deferred to Appendix B.

Lemma 4 *Let* $\gamma = 1 - \frac{1}{2T_{\max}}$, $H = \lceil \log_2(c_f K) \rceil$, $c_f = \lceil 4D \ln \frac{2K}{\delta} \rceil$ *for some* $\delta \in (0,1)$, *and* $\pi_1, \dots, \pi_K$ *be policies for* $\mathring{\mathcal{M}}$. *Then the regret of executing* $\sigma(\pi_1), \dots, \sigma(\pi_K)$ *in* $\mathcal{M}$ *satisfies* $R_K \le \mathring{R}_K + \tilde{\mathcal{O}}(1)$ *with probability at least* $1-\delta$, *where* $\mathring{R}_K = \sum_{k=1}^{K} \left(\sum_{i=1}^{J_k} c_i^k + \mathring{c}_{J_k+1}^k - V^{\mathring{\pi}^\star, P, c}(s_1^k, 1) \right)$ *for stochastic environments, and* $\mathring{R}_K = \sum_{k=1}^{K} \left(\sum_{i=1}^{J_k} c_i^k + \mathring{c}_{J_k+1}^k - V^{\mathring{\pi}^\star, P, c_k}(s_1^k, 1) \right)$ *for adversarial environments. Here,* J_k *is the number of time steps in episode* k *before the learner reaching* g *or the counter of* $\sigma(\pi_k)$ *reaching* $H+1$, *and* $\mathring{c}_{J_k+1}^k = c_f \mathbb{I}\{s_{J_k+1}^k \neq g\}$.

Computing Fast Policy and Estimating Diameter For simplicity, we assume knowledge of the diameter and the fast policy above. When these are unknown, one can follow the ideas in (Chen and Luo, 2021) for estimating the fast policy with constant overhead and then adopt their template for learning without knowing the diameter; see (Chen and Luo, 2021, Lemma 1, Appendix E).

Policy Optimization in Stacked Discounted MDPs Now we describe a template of performing policy optimization with the stacked discounted approximation. The pseudocode is shown in Algorithm 1. To handle unknown transition, we maintain standard Bernstein-style transition confidence sets $\{\mathcal{P}_k\}_{k=1}^K$ whose definition is deferred to Appendix A.1. In episode k, the algorithm first computes policy π_k in $\mathring{\mathcal{M}}$ following the multiplicative weights update with some learning rate $\eta > 0$, such that $\pi_k(a|s,h) \propto e^{-\eta \sum_{j=1}^{k-1} (\widetilde{Q}_j(s,a,h) - B_j(s,a,h))}$ for some optimistic action-value estimator $\widetilde{Q}_j$ and exploration bonus function B_j (computed from past observations and confidence sets). Then, it executes $\sigma(\pi_k)$ for this episode. Finally, it computes confidence set $\mathcal{P}_{k+1}$. All algorithms introduced in this work follow this template and differ from each other in the definition of $\widetilde{Q}_k$ and B_k. Ideally, $\widetilde{Q}_k - B_k$ should be the action-value function with respect to the true transition, the true cost function, and policy π_k, but since the transition and cost functions are unknown, the key challenge lies in constructing accurate estimators that simultaneously encourage sufficient exploration.

Optimistic Transitions Our algorithms require using some optimistic transitions. Specifically, for a policy π, a confidence set $\mathcal{P}$, and a cost function c, let $\Gamma(\pi, \mathcal{P}, c)$ be the corresponding optimistic transition such that $\Gamma(\pi, \mathcal{P}, c) \in \operatorname{argmin}_{P \in \mathcal{P}} V^{\pi, P, c}(s,h)$ for all state (s,h). The existence of such an optimistic transition and how it can be efficiently approximated via Extended Value Iteration (in

at most $\widetilde{\mathcal{O}}(T_{\max})$ iterations) are deferred to Appendix A.2. We abuse the notation and denote by $V^{\pi,\mathcal{P},c}$ and $Q^{\pi,\mathcal{P},c}$ the value function $V^{\pi,\Gamma(\pi,\mathcal{P},c),c}$ and action-value function $Q^{\pi,\Gamma(\pi,\mathcal{P},c),c}$.

Occupancy Measure Another important concept for subsequent discussions is *occupancy measure*. Given a policy $\pi : \mathring{\mathcal{S}} \to \Delta_{\mathcal{A}}$ and a transition function $P = \{P_{s,a,h}\}_{(s,h)\in\mathring{\mathcal{S}},a\in\mathcal{A}}$ with $P_{s,a,h} \in \Delta_{\mathring{\mathcal{S}}_+}$ and $\mathring{\mathcal{S}}_+ = \mathring{\mathcal{S}} \cup \{g\}$, define $q_{\pi,P} : \mathring{\mathcal{S}} \times \mathcal{A} \times \mathring{\mathcal{S}}_+ \to \mathbb{R}_+$ such that $q_{\pi,P}(\mathring{s}, a, \mathring{s}') = \mathbb{E}[\sum_{i=1}^{I} \mathbb{I}\{s_i = \mathring{s}, a_i = a, s_{i+1} = \mathring{s}'\}|\pi, P, s_1 = \mathring{s}_{\text{init}}]$ is the expected number of visits to $(\mathring{s}, a, \mathring{s}')$ following policy π in a stacked discounted MDP with transition P. We also let $q_{\pi,P}(s, a, h) = \sum_{\mathring{s}'} q_{\pi,P}((s, h), a, \mathring{s}')$ be the expected number of visits to $((s, h), a)$ and $q_{\pi,P}(s, h) = \sum_a q_{\pi,P}(s, a, h)$ be the number of visits to (s, h). Note that if a function $q : \mathring{\mathcal{S}} \times \mathcal{A} \times \mathring{\mathcal{S}}_+ \to \mathbb{R}_+$ is an occupancy measure, then the corresponding policy π_q satisfies $\pi_q(a|s, h) \propto q(s, a, h)$ and the corresponding transition function P_q satisfies $P_{q,s,a,h}(s', h') \propto q((s, h), a, (s', h'))$. Moreover, $V^{\pi,P,c}(\mathring{s}_{\text{init}}) = \langle q_{\pi,P}, c\rangle$ holds for any policy π, transition function P and cost function c.

Other Notations In the rest of the paper, following Lemma 4 we set $\gamma = 1 - \frac{1}{2T_{\max}}$, $H = \lceil \log_2(c_f K)\rceil$, and $c_f = \lceil 4D \ln \frac{2K}{\delta}\rceil$ for some failure probability $\delta \in (0, 1)$. Define $\chi = 2HT_{\max} + c_f$ as the value function upper bound in $\mathring{\mathcal{M}}$ (according to the first statement of Lemma 2). Also define $q_k = q_{\pi_k,P}$, $q^\star = q_{\mathring{\pi}^\star,P}$, and $L = \lceil \frac{8H}{1-\gamma} \ln(2T_{\max}K/\delta)\rceil$.

4. Algorithms and Results for Stochastic Environments

In this section, we consider policy optimization in stochastic environments with three types of feedback introduced in Section 2. We show that a simple policy optimization framework can be used to achieve near-optimal regret for all three settings. In contrast, previous works treat stochastic costs and stochastic adversaries as different problems and solve them via different approaches. Below, we start by describing the algorithm and its guarantees, followed by some explanation behind the algorithm design and then some key ideas and novelty in the analysis.

Algorithm As mentioned, the only elements left to be specified in Algorithm 1 are $\widetilde{Q}_k$ and B_k. For stochastic environments, we simply set $B_k(s, a, h) = 0$ for all (s, a, h) since exploration is relatively easier in this case. We now discuss how to construct $\widetilde{Q}_k$.

- **Action-value estimator** $\widetilde{Q}_k$ is defined as $Q^{\pi_k,\mathcal{P}_k,\widetilde{c}_k}$ for some corrected cost estimator $\widetilde{c}_k$:

$$\widetilde{c}_k(s, a, h) = (1 + \lambda\widehat{Q}_k(s, a, h))\widehat{c}_k(s, a, h) + e_k(s, a, h), \tag{1}$$

  where λ is some parameter, $\widehat{Q}_k = Q^{\pi_k,\mathcal{P}_k,\widehat{c}_k}$ is another action-value estimator with respect to some optimistic cost estimator $\widehat{c}_k$, and e_k is some correction term (all to be specified below).

- **Optimistic cost estimator** $\widehat{c}_k$ is defined as

$$\widehat{c}_k(s, a, h) = \widehat{c}_k(s, a)\mathbb{I}\{h \le H\} + c_f\mathbb{I}\{h = H + 1\},$$
$$\widehat{c}_k(s, a) = \max\left\{0, \bar{c}_k(s, a) - 2\sqrt{\bar{c}_k(s, a)\alpha_k(s, a)} - 7\alpha_k(s, a)\right\},$$

  where $\bar{c}_k(s, a)$ is the average of all costs that are observed for (s, a) in episode $j = 1, \ldots, k - 1$ before $\sigma(\pi_j)$ switches to the fast policy, and $\alpha_k(s, a)$ is $\iota = \ln(2SALK/\delta)$ divided by the number of samples used in computing $\bar{c}_k(s, a)$, such that $2\sqrt{\bar{c}_k(s, a)\alpha_k(s, a)} + 7\alpha_k(s, a)$ is a

standard Bernstein-style deviation term (thus making $\widehat{c}_k(s,a)$ an optimistic underestimator). We note that naturally, the way to compute $\bar{c}_k(s,a)$ is different for different types of feedback — for stochastic costs, we might have multiple samples for (s,a) in one episode, while for stochastic adversaries, we have exactly one sample in each episode in the full-information setting, and one or zero samples in the bandit setting.

- **Correction term** $e_k(s,a,h)$ is defined as 0 for stochastic costs; $(8\iota\sqrt{\widehat{c}_k(s,a,h)/k} + \beta'\widehat{Q}_k(s,a,h))\mathbb{I}\{h \le H\}$ with $\beta' = \min\{1/T_{\max}, 1/\sqrt{DT_\star K}\}$ for stochastic adversary with full information; and $\beta\widehat{Q}_k(s,a,h)\mathbb{I}\{h \le H\}$ with $\beta = \min\{1/T_{\max}, \sqrt{SA/DT_\star K}\}$ for stochastic adversary with bandit feedback.
- **Parameter tuning**: learning rate η (for the multiplicative weights update) is set to $\min\{1/3T_{\max}(8\iota+\chi/T_{\max})^2, 1/\sqrt{\lambda T_{\max}^4 K}\}$, and the parameter λ is set to $\min\{1/T_{\max}, \sqrt{S^2A/\square^2 K}\}$ where $\square$ is $B_\star$ for stochastic costs and D for stochastic adversaries.

We now state the regret guarantees of our algorithm for each of the three settings (proofs are deferred to Appendix C.2.1 to Appendix C.2.3).

Theorem 5 *For stochastic costs, Algorithm 1 with the instantiation above achieves* $R_K = \tilde{\mathcal{O}}(B_\star S\sqrt{AK} + T_{\max}^3(S^2AK)^{1/4} + S^4A^{2.5}T_{\max}^4)$ *with probability at least* $1 - 32\delta$.

Ignoring lower-order terms, our bound almost matches the minimax bound $\tilde{\mathcal{O}}(B_\star\sqrt{SAK})$ of (Cohen et al., 2021), with a $\sqrt{S}$ factor gap.

Theorem 6 *For stochastic adversary with full information, Algorithm 1 with the instantiation above achieves* $R_K = \tilde{\mathcal{O}}(\sqrt{DT_\star K} + DS\sqrt{AK} + T_{\max}^3(S^2A^3K)^{1/4} + S^4A^{2.5}T_{\max}^4)$ *with probability at least* $1 - 50\delta$.

Theorem 7 *For stochastic adversary with bandit feedback, Algorithm 1 with the instantiation above achieves* $R_K = \tilde{\mathcal{O}}(\sqrt{SADT_\star K} + DS\sqrt{AK} + T_{\max}^3 SA^{5/4}K^{1/4} + S^4A^{2.5}T_{\max}^4)$ *with probability at least* $1 - 50\delta$.

Ignoring lower-order terms again, these bounds for stochastic adversary match the best known results from (Chen and Luo, 2021), and they all exhibit a $\sqrt{S}$ gap in the term $DS\sqrt{AK}$ compared to the best existing lower bounds (Chen and Luo, 2021).

We emphasize again that besides the simplicity of PO, one algorithmic advantage of our method compared to those based on finite-horizon approximation is its low space complexity to store policies — the horizon H for our method is only $\mathcal{O}(\ln K)$, while the horizon for other works (Chen and Luo, 2021; Cohen et al., 2021) is $\tilde{\mathcal{O}}(T_{\max})$ when $T_{\max}$ is known or otherwise $\tilde{\mathcal{O}}(B_\star/c_{\min})$. Note that when $c_{\min} = 0$, a common technique is to perturb the cost and deal with a modified problem with $c_{\min} = 1/\text{poly}(K)$, in which case our space complexity is exponentially better. In fact, even for time complexity, although our method requires calculating optimistic transition and might need $\tilde{\mathcal{O}}(T_{\max})$ rounds of Extended Value Iteration, this procedure could terminate much earlier, while the finite-horizon approximation approaches always need at least $\Omega(T_{\max})$ time complexity since that is the horizon of the MDP they are dealing with.

Analysis highlights We start by explaining the design of the corrected cost estimator Eq. (1). Roughly speaking, standard analysis of PO (specifically, by (Chen and Luo, 2021, Lemma 9) and then Lemma 26) leads to a term of order $\lambda \sum_{k=1}^{K} \langle q_k, c \circ \widehat{Q}_k \rangle$ due to the transition estimation error, which can be prohibitively large (for functions f and g with the same domain, we define $(f \circ g)(x) = f(x)g(x)$). Introducing the correction bias $\lambda \widehat{Q}_k(s,a,h)\widehat{c}_k(s,a,h)$ in Eq. (1), on the other hand, has the effect of transforming this problematic term into its counterpart $\lambda \sum_{k=1}^{K} \langle q^\star, c \circ \widehat{Q}_k \rangle$ in terms of $q^\star$ instead of q_k. Bounding the latter term, however, requires a property that PO enjoys, that is, a regret bound for any initial state-action pair: $\sum_{k=1}^{K} (\widehat{Q}_k - Q^{\mathring{\pi}^\star, P, c})(s,a,h) = \tilde{\mathcal{O}}(\sqrt{K})$ for any (s,a,h). In contrast, approaches based on occupancy measure (Chen and Luo, 2021) only guarantee a regret bound starting from s_{init}. This makes PO especially compatible with our stacked discounted approximation. Based on this observation, we further have $\lambda \sum_{k=1}^{K} \langle q^\star, c \circ \widehat{Q}_k \rangle \approx \lambda \sum_{k=1}^{K} \langle q^\star, c \circ Q^{\hat{\pi}^\star, P, c} \rangle$, where the latter term is only about the behavior of the optimal policy and is thus nicely bounded (see e.g. Lemma 20). To sum up, the correction term $\lambda \widehat{Q}_k(s,a,h)\widehat{c}_k(s,a,h)$ in Eq. (1) together with a favorable property of PO helps us control the transition estimation error in a near-optimal way.

For stochastic adversaries, an extra complication arises due to the cost estimation error $\sum_{k=1}^{K} \langle q_k, c - \widehat{c}_k \rangle$, which results in the extra $\sqrt{DT_\star K}$ or $\sqrt{SADT_\star K}$ term in the minimax regret bound (depending on the feedback type). Obtaining this optimal cost estimation error requires us to add yet another correction term e_k in Eq. (1). Specifically, we show that $\sum_{k=1}^{K} \langle q_k, c - \widehat{c}_k \rangle \approx \sum_{k=1}^{K} \langle q_k, e_k \rangle$ for e_k defined as in our algorithm description. Then, the role of adding e_k in Eq. (1) is again to turn the term above to its counterpart $\sum_{k=1}^{K} \langle q^\star, e_k \rangle$ in terms of the optimal policy's behavior, which can then be nicely bounded. As a side product, we note that this also provides a much cleaner analysis on bounding the cost estimation error compared to (Chen and Luo, 2021), where they require explicitly forcing the expected hitting time of the learner's policy to be bounded.

Finally, we point out another novelty in our analysis. Compared to other approaches that act according to the exact optimal policy of an estimated MDP, PO incurs an additional cost due to only updating the policy incrementally in each episode. This cost is often of order $\tilde{\mathcal{O}}(\sqrt{K})$ and is one of the dominating terms in the regret bound; see e.g. (Shani et al., 2020; Wu et al., 2021) for the finite-horizon case. For SSP, this is undesirable because it also depends on $T_\star$ or even $T_{\max}$. Reducing this cost has been studied from the optimization perspective — for example, an improved $\tilde{\mathcal{O}}(1/K)$ convergence rate of PO has been established recently by (Agarwal et al., 2021). However, adopting their analysis to regret minimization requires additional efforts. Specifically, we need to carefully bound the bias from using an action-value estimator in the policy's update, which can be shown to be approximately bounded by $\sum_{k=1}^{K} (\widetilde{Q}_{k+1} - \widetilde{Q}_k)(s,a,h)$. In Lemma 25, we show that this term is of lower order by carefully analyzing the drift $(\widetilde{Q}_{k+1} - \widetilde{Q}_k)(s,a,h)$ in each episode.

Remark 8 *We remark that our algorithm can be applied to finite-horizon MDPs with inhomogeneous transition and gives a $\tilde{\mathcal{O}}(\sqrt{S^2AH^3K})$ regret bound, improving over that of (Shani et al., 2020) by a factor of $\sqrt{H}$ where H is the horizon. We omit the details but only mention that the improvement comes from two sources: first, the aforementioned improved PO analysis turns a $\tilde{\mathcal{O}}(H^2\sqrt{K})$ regret term into a lower order term; second, we use Bernstein-style transition confidence set to obtain an improved $\tilde{\mathcal{O}}(\sqrt{S^2AH^3K})$ transition estimation error.*

5. Algorithms and Results for Adversarial Environments

We move on to consider the more challenging environments with adversarial costs, where the extra exploration bonus function B_k in Algorithm 1 now plays an important role. Even in the finite-horizon setting, developing efficient PO methods in this case can be challenging, and Luo et al. (2021) proposed the so-called "dilated bonuses" to guide better exploration, which we also adopt and extend to SSP. Specifically, for a policy π, a transition confidence set $\mathcal{P}$, and some bonus function $b : \mathcal{S} \times \mathcal{A} \times [H+1] \to \mathbb{R}$, we define the corresponding dilated bonus function $B^{\pi,\mathcal{P},b} : \mathcal{S} \times \mathcal{A} \times [H+1] \to \mathbb{R}$ as: $B^{\pi,\mathcal{P},b}(s,a,H+1) = b(s,a,H+1)$ and for $h \in [H]$,

$$B^{\pi,\mathcal{P},b}(s,a,h) = b(s,a,h) + \left(1 + \frac{1}{H'}\right) \max_{\widehat{P} \in \mathcal{P}} \widehat{P}_{s,a,h} \left(\sum_{a'} \pi(a'|\cdot,\cdot) B^{\pi,\mathcal{P},b}(\cdot,a',\cdot) \right), \tag{2}$$

where $H' = \frac{8(H+1)\ln(2K)}{1-\gamma}$ is the dilated coefficient. Intuitively, $B^{\pi,\mathcal{P},b}$ is the dilated (by a factor of $1 + 1/H'$) and optimistic (by maximizing over $\mathcal{P}$) version of the action-value function with respect to π and b. In the finite-horizon setting (Luo et al., 2021), this can be computed directly via dynamic programming, but how to compute it in a stacked discounted MDP (or even why it exists) is less clear. Fortunately, we show that this can indeed be computed efficiently via a combination of dynamic programming and Extended Value Iteration; see Appendix D.4.

Algorithm (full information) We now describe our algorithm for the adversarial full-information case (where c_k is revealed at the end of episode k). It suffices to specify $\widetilde{Q}_k$ and B_k in Algorithm 1.

- **Action-value estimator** $\widetilde{Q}_k$ is defined as $\widetilde{Q}_k = Q^{\pi_k,\mathcal{P}_k,\widetilde{c}_k}$, where $\widetilde{c}_k(s,a,h) = (1 + \lambda \widehat{Q}_k(s,a,h)) c_k(s,a,h)$ for some parameter λ and $\widehat{Q}_k = Q^{\pi_k,\mathcal{P}_k,c_k}$.
- **Dilated bonus** B_k is defined as $B^{\pi_k,\mathcal{P}_k,b_k}$ with $b_k(s,a,h) = 2\eta \sum_{a \in \mathcal{A}} \pi_k(a|s,h) \widetilde{A}_k(s,a,h)^2$, where $\widetilde{A}_k(s,a,h) = \widetilde{Q}_k(s,a,h) - \widetilde{V}_k(s,h)$ (advantage function) and $\widetilde{V}_k = V^{\pi_k,\mathcal{P}_k,\widetilde{c}_k}$.
- **Parameter tuning**: $\eta = \min\{1/(64\chi^2\sqrt{HH'}), 1/\sqrt{DK}\}$ and $\lambda = \min\{1/\chi, 48\eta + \sqrt{S^2A/DT_\star K}\}$.

Our algorithm enjoys the following guarantee (whose proof can be found in Appendix D.1).

Theorem 9 *For adversarial costs with full information, Algorithm 1 with the instantiation above achieves* $R_K = \tilde{\mathcal{O}}(T_\star\sqrt{DK} + \sqrt{S^2ADT_\star K} + S^4A^2T_{\max}^5)$ *with probability at least* $1 - 20\delta$.

The best existing bound is $\tilde{\mathcal{O}}(\sqrt{S^2ADT_\star K})$ from (Chen and Luo, 2021). Ignoring the lower order term, our result matches theirs when $T_\star \le S^2A$ (and is worse by a $\sqrt{T_\star/S^2A}$ factor otherwise). Our algorithm enjoys better time and space complexity though, similar to earlier discussions.

Analysis highlights For simplicity we assume that the true transition is known, in which case our bound is only $\tilde{\mathcal{O}}(T_\star\sqrt{DK})$ (the other term $\sqrt{S^2ADT_\star K}$ is only due to transition estimation error). A naive way to implement PO would lead to a penalty term $T_\star/\eta$ plus a stability term $\eta \sum_{k=1}^K \sum_{s,h} q^\star(s,h) \sum_a \pi_k(a|s,h) Q^{\pi_k,P,c_k}(s,a,h)^2$, which eventually leads to a bound of order $\tilde{\mathcal{O}}(T_\star T_{\max}\sqrt{K})$ if one bounds $Q^{\pi_k,P,c_k}(s,a,h)$ by $\tilde{\mathcal{O}}(T_{\max})$. Our improvement comes from the following five steps: 1) first, through a careful shifting argument, we show that the stability term can be improved to $\eta \sum_{k=1}^K \sum_{s,h} q^\star(s,h) \sum_a \pi_k(a|s,h) A^{\pi_k,P,c_k}(s,a,h)^2$ (recall that A is the advantage function); 2) second, similarly to (Luo et al., 2021), the dilated bonus B_k helps transform

$q^\star$ to q_k in the term above, leading to $\eta \sum_{k=1}^K \langle q_k, (A^{\pi_k,P,c_k})^2 \rangle$; 3) third, in Lemma 26 we show that the previous term is bounded by the variance of the learner's cost, which in turn is at most $\eta \sum_{k=1}^K \langle q_k, c_k \circ Q^{\pi_k,P,c_k} \rangle$; 4) fourth, similarly to Section 4, the correction term $\lambda c_k \circ \widehat{Q}_k$ in the definition of $\widetilde{c}_k$ helps transform q_k back to $q^\star$, resulting in $\eta \sum_{k=1}^K \langle q^\star, c_k \circ Q^{\pi_k,P,c_k} \rangle$; 5) finally, since PO guarantees a regret bound for any initial state (as mentioned in Section 4), the previous term is close to $\eta \sum_{k=1}^K \langle q^\star, c_k \circ Q^{\mathring{\pi}_k^\star,P,c_k} \rangle$, which is now only related to the optimal policy and can be shown to be at most $\tilde{\mathcal{O}}(\eta D T_\star K)$. Combining this with the penalty term $T_\star/\eta$ and picking the best η then results in the claimed $\tilde{\mathcal{O}}(T_\star \sqrt{DK})$ regret bound.

Algorithm (Bandit Feedback) Finally, we describe our algorithm for the adversarial setting with bandit feedback, starting with the instantiation of B_k followed by that of $\widetilde{Q}_k$.

- **Dilated bonus** B_k is again defined as $B^{\pi_k,\mathcal{P}_k,b_k}$, but with a different b_k function similar to that of (Luo et al., 2021): $b_k(s,a,h) = L'\mathbb{I}\{h \le H\} \sum_{a'} \pi_k(a'|s,h) \frac{\overline{x}_k(s,a',h) - \underline{x}_k(s,a',h) + 4\theta}{\overline{x}_k(s,a',h) + \theta}$, for some parameters L' and θ. Here, $\overline{x}_k(s,a,h)$ and $\underline{x}_k(s,a,h)$ are respectively the largest and smallest possible probability that $((s,h),a)$ is ever visited in episode k following policy π_k if the transition lies in $\mathcal{P}_k$, and they can be computed efficiently as shown in Appendix D.5.
- **Action-value estimator** $\widetilde{Q}_k$ is defined as $\widetilde{Q}_k(s,a,h) = \frac{G_{k,s,a,h}}{\overline{x}_k(s,a,h)+\theta}\mathbb{I}\{h \le H\} + c_f \mathbb{I}\{h = H+1\}$, where $G_{k,s,a,h}$ is the learner's total cost in $\mathring{\mathcal{M}}$ starting from the first visit to $((s,h),a)$ during the first $L+1$ steps of episode k. Recall the definition of L stated at the end of Section 3, which is a high-probability upper bound on the number of steps any policy in $\mathring{\mathcal{M}}$ takes to reach the last layer (so counting only the first $L+1$ steps is simply to make sure that $G_{k,s,a,h}$ is always bounded).
- **Parameter tuning:** $\eta = \min\left\{1/(300HH'T_{\max}L'), \sqrt{1/T_{\max}^2 SAK}\right\}$, $\theta = 2\eta L'$, and $L' = L + c_f$.

We note that this algorithm is in spirit very similar to that of (Luo et al., 2021) for the finite-horizon case. Unfortunately, the correction terms we use throughout other algorithms in this work do not work here for technical reasons, resulting in the following sub-optimal guarantee which still has $T_{\max}$ dependency in the dominating term (see Appendix D.2 for the proof). We remark that the best existing bound is $\tilde{\mathcal{O}}(\sqrt{DT_\star S^3 A^2 K})$ from (Chen and Luo, 2021).

Theorem 10 *For adversarial costs with bandit feedback, Algorithm 1 with the instantiation above achieves* $R_K = \tilde{\mathcal{O}}(\sqrt{S^2 A T_{\max}^5 K} + S^{5.5} A^{3.5} T_{\max}^5)$ *with probability at least* $1 - 28\delta$.

6. Conclusion

Our work initiates the study of policy optimization for SSP and systematically develops a set of novel algorithms suitable for different settings. Many questions remain open, such as closing the gap between some of our results and the best known results achieved by other types of methods. Moreover, as mentioned, one of the reasons to study PO for SSP is that PO usually works well when combined with function approximation. Our stacked discounted approximation scheme also does not make use of any modeling assumption and should be applicable in more general settings. Although our work is only for the tabular setting, we believe that our results lay a solid foundation for future studies on SSP with function approximation.

Acknowledgments

LC thanks Chen-Yu Wei for many helpful discussions.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

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.

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 and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, 2021.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 2021a.

Liyu Chen, Rahul Jain, and Haipeng Luo. Improved no-regret algorithms for stochastic shortest path with linear MDP. *arXiv preprint arXiv:2112.09859*, 2021b.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, pages 1216–1259. PMLR, 2021c.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference on Learning Theory*, pages 1180–1215. PMLR, 2021d.

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*, volume 119, pages 8210–8219. PMLR, 2020.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34, 2021.

Yingjie Fei, Zhuoran Yang, Zhaoran Wang, and Qiaomin Xie. Dynamic regret of policy optimization in non-stationary environments. *Advances in Neural Information Processing Systems*, 33: 6743–6754, 2020.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Mehdi Jafarnia-Jahromi, Liyu Chen, Rahul Jain, and Haipeng Luo. Online learning for stochastic shortest path model via posterior sampling. *arXiv preprint arXiv:2106.05335*, 2021.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture MDPs. *arXiv preprint arXiv:2111.03289*, 2021.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in neural information processing systems*, 32, 2019.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial MDPs: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34, 2021.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. *arXiv preprint arXiv:2110.12727*, 2021.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2936–2942. ijcai.org, 2021. doi: 10.24963/ijcai.2021/404. URL https://doi.org/10.24963/ijcai.2021/404.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

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, 2020.

Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, pages 1593–1601. PMLR, 2014.

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

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34, 2021.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *International Conference on Learning Representations (ICLR)*, 2020.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10170–10180. PMLR, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Tianhao Wu, Yunchang Yang, Han Zhong, Liwei Wang, Simon S Du, and Jiantao Jiao. Nearly optimal policy optimization with stable at any time guarantee. *arXiv preprint arXiv:2112.10935*, 2021.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture MDP. *Advances in Neural Information Processing Systems*, 2021.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

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.

Appendix A. Preliminary for Appendix

Extra Notations Define $\mathring{s}_i^k = (s_i^k, h_i^k)$ as the i-th step in $\mathring{\mathcal{M}}$ in episode k. Define $n_k(s,a,h)$ as the number of visits to $((s,h),a)$ in $\mathring{\mathcal{M}}$ in episode k, and $n_k(s,a) = \sum_{h \le H} n_k(s,a,h)$ (excluding layer $H+1$). Define $\bar{J}_k = \min\{L, J_k\}$, $\bar{n}_k(s,a) = \min\{L, n_k(s,a)\}$, and $\bar{n}_k(s,a,h) = \min\{L, n_k(s,a,h)\}$. For any sequence of scalars or functions $\{z_k\}_k$, define $dz_k = z_{k+1} - z_k$. By default we assume $\sum_h = \sum_{h=1}^{H+1}$. For inner product $\langle u, v\rangle$, if $u(s,a)$, $u(s,a,h)$, $v(s,a)$, and $v(s,a,h)$ are all defined, we let $\langle u, v\rangle = \sum_{s,a,h} u(s,a,h)v(s,a,h)$. For functions f and g with the same domain, define function $(f \circ g)(x) = f(x)g(x)$. For any random variable X, define conditional variance $\mathrm{Var}_k[X] = \mathbb{E}_k[(X - \mathbb{E}_k[X])^2]$.

For an occupancy measure q w.r.t policy π and transition P, define $q_{(s,h)}$ as the occupancy measure w.r.t policy π, transition P, and initial state (s,h), and $q_{(s,a,h)}$ as the occupancy measure w.r.t policy π, transition P, initial state (s,h), and initial action a. Denote by $x_k(s,a,h)$ the probability that $((s,h),a)$ is ever visited in episode k, $x_k(s,a) = \sum_{h=1}^{H} x_k(s,a,h)$ the probability that (s,a) is ever visited before layer $H+1$ in episode k, and $y_k(s,a,h)$ the probability of visiting $((s,h),a)$ again if the agent starts from $((s,h),a)$. For any occupancy measure $q(s,a,h)$, we define $q(s,a) = \sum_{h \le H} q(s,a,h)$ (excluding layer $H+1$). Note that $q_k(s,a,h) = \frac{x_k(s,a,h)}{1-y_k(s,a,h)}$ and $y_k(s,a,h) \le \gamma$. Thus, we have $q_k(s,a,h) = \mathcal{O}(T_{\max} x_k(s,a,h))$.

Define $\Lambda_{\mathcal{M}}$ as the set of possible transition functions of $\mathring{\mathcal{M}}$:

$$\begin{aligned}\Lambda_{\mathcal{M}} = \Big\{ P = \{P_{s,a,h}\}_{(s,h)\in\mathring{\mathcal{S}}, a\in\mathcal{A}}, P_{s,a,h} \in \Delta_{\mathring{\mathcal{S}}_+} : P_{s,a,H+1}(g) = 1, \\ \sum_{s'\in\mathcal{S}} P_{s,a,h}(s',h) \le \gamma, \sum_{s'\in\mathcal{S}} P_{s,a,h}(s',h+1) \le 1-\gamma, \\ P_{s,a,h}(s',h') = 0, \forall (s,a) \in \mathcal{S}\times\mathcal{A}, h \in [H], h' \notin \{h, h+1\} \Big\}, \end{aligned} \tag{3}$$

where $\gamma \cdot \mathcal{X} = \{\gamma x : x \in \mathcal{X}\}$ for some set $\mathcal{X}$. By definition, the expected hitting time of any stationary policy in an MDP with transition $P \in \Lambda_{\mathcal{M}}$ is upper bounded by $(H+1)(1-\gamma)^{-1}$ starting from any state. Therefore, for any occupancy measure q with $P_q \in \Lambda_{\mathcal{M}}$ (for example, q_k and $q^\star$), we have $\sum_{s,a,h} q(s,a,h) \le (H+1)(1-\gamma)^{-1} = \tilde{\mathcal{O}}(T_{\max})$.

Finally define $\mathcal{C}_{\mathcal{M}}$ as the set of possible cost functions of $\mathring{\mathcal{M}}$:

$$\mathcal{C}_{\mathcal{M}} = \Big\{ c : \mathring{\mathcal{S}} \to \mathbb{R}_+ : c(s,a,h) = \tilde{\mathcal{O}}(1), \forall h \le H, \text{ and } \exists C_0 = \tilde{\mathcal{O}}(T_{\max}), c(s,a,H+1) = C_0, \forall a \Big\}.$$

A.1. Transition Estimation

In this section, we present important lemmas regarding the transition confidence sets $\{\mathcal{P}_k\}_{k=1}^K$. We first prove an auxiliary lemma saying that the number of steps taken by the learner before reaching g or switching to fast policy is well bounded with high probability.

Lemma 11 *With probability at least $1-\delta$, we have $J_k = \bar{J}_k$ for all $k \in [K]$.*

Proof We want to show that $J_k \le L = \lceil \frac{8H}{1-\gamma} \ln(2T_{\max}K/\delta) \rceil$ for all $k \in [K]$ with probability at least $1-\delta$. Let $k \in [K]$, it suffices to show that the expected hitting time of π_k is upper bounded by $\frac{H}{1-\gamma}$ starting from any (s,h), because then we can apply Lemma 31 and take a union bound over all K episodes.

Note that the expected hitting time (w.r.t J_k) is simply the value function with respect to a cost function that is 1 for all state-action pairs except for 0 cost in the goal state g and layer $H+1$ (i.e., $c_f = 0$). Thus, by Lemma 2, the expected hitting time starting from (s,h) is bounded by $\frac{H-h+1}{1-\gamma} \le \frac{H}{1-\gamma}$. ■

Definition of $\mathcal{P}_k$ We define $\mathcal{P}_k = \bigcap_{s,a,h\le H} \mathcal{P}_{k,s,a,h}$, where:

$$\begin{aligned}\mathcal{P}_{k,s,a,h} = \big\{P' \in \Lambda_{\mathcal{M}} : &\left|\bar{P}_{k,s,a}(s') - P'_{s,a,h}(s',h)/\gamma\right| \le \epsilon_k(s,a,s'),\\ &\left|\bar{P}_{k,s,a}(s') - P'_{s,a,h}(s',h+1)/(1-\gamma)\right| \le \epsilon_k(s,a,s'),\\ &\left|\bar{P}_{k,s,a}(g) - P'_{s,a,h}(g)\right| \le \epsilon_k(s,a,g), \forall s' \in \mathcal{S}\big\},\end{aligned} \tag{4}$$

where $\epsilon_k(s,a,s') = 4\sqrt{\bar{P}_{k,s,a}(s')\alpha'_k(s,a)} + 28\alpha'_k(s,a)$, $\alpha'_k(s,a) = \frac{\iota}{N_k^+(s,a)}$, $\bar{P}_{k,s,a}(s') = \frac{N_k(s,a,s')}{N_k^+(s,a)}$ is the empirical transition, $N_k^+(s,a) = \max\{1, N_k(s,a)\}$, $N_k(s,a)$ is the number of visits to (s,a) in episode $j = 1,\dots,k-1$ before $\sigma(\pi_j)$ switches to the fast policy, and $N_k(s,a,s')$ is the number of visits to (s,a,s') in episode $j = 1,\dots,k-1$ before $\sigma(\pi_j)$ switches to the fast policy.

Lemma 12 *Under the event of Lemma 11, we have $\mathring{P} \in \mathcal{P}_k$ for any $k \in [K]$ with probability at least $1-\delta$.*

Proof Clearly $\mathring{P} \in \Lambda_{\mathcal{M}}$. Moreover, for any $(s,a) \in \mathcal{S}\times\mathcal{A}, s' \in \mathcal{S}_+$ by Lemma 51 and $N_{K+1}(s,a) \le LK$ under the event of Lemma 11, we have with probability at least $1 - \frac{\delta}{2S^2A}$,

$$\left|P_{s,a}(s') - \bar{P}_{k,s,a}(s')\right| \le \epsilon_k(s,a,s'). \tag{5}$$

By a union bound, we have Eq. (5) holds for any $(s,a) \in \mathcal{S}\times\mathcal{A}, s' \in \mathcal{S}_+$ with probability at least $1-\delta$. Then the statement is proved by $\mathring{P}_{s,a,h}(s',h) = \gamma P_{s,a}(s')$, $\mathring{P}_{s,a,h}(s',h+1) = (1-\gamma)P_{s,a}(s')$, and $\mathring{P}_{s,a,h}(g) = P_{s,a}(g)$. ■

Lemma 13 *Under the event of Lemma 12, for any $P' \in \mathcal{P}_k$, we have for any $\mathring{s}' \in \mathring{\mathcal{S}}_+$:*

$$\left|P'_{s,a,h}(\mathring{s}') - P_{s,a,h}(\mathring{s}')\right| \le 8\sqrt{P_{s,a,h}(\mathring{s}')\alpha'_k(s,a)} + 136\alpha'_k(s,a) \triangleq \epsilon^\star_k(s,a,h,\mathring{s}').$$

For simplicity, we also write $\epsilon^\star_k(s,a,h,(s',h'))$ as $\epsilon^\star_k(s,a,h,s',h')$ for $(s',h') \in \mathring{\mathcal{S}}$.

Proof Under the event of Lemma 12 and by Eq. (5), we have for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and $s' \in \mathcal{S}_+$:

$$\bar{P}_{k,s,a}(s') \le P_{s,a}(s') + 4\sqrt{\bar{P}_{k,s,a}(s')\alpha'_k(s,a)} + 28\alpha'_k(s,a).$$

Applying $x^2 \le ax + b \implies x \le a + \sqrt{b}$ with $a = 4\sqrt{\alpha'_k(s,a)}$ and $b = P_{s,a}(s') + 28\alpha'_k(s,a)$, we have

$$\sqrt{\bar{P}_{k,s,a}(s')} \le 4\sqrt{\alpha'_k(s,a)} + \sqrt{P_{s,a}(s') + 28\alpha'_k(s,a)} \le \sqrt{P_{s,a}(s')} + 10\sqrt{\alpha'_k(s,a)}.$$

Substituting this back to the definition of ϵ_k, we have

$$\epsilon_k(s,a,s') = 4\sqrt{\bar{P}_{k,s,a}(s')\alpha'_k(s,a)} + 28\alpha'_k(s,a) \le 4\sqrt{P_{s,a}(s')\alpha'_k(s,a)} + 68\alpha'_k(s,a).$$

Now we start to prove the statement. The statement is clearly true for $\mathring{s}' = (s',h')$ with $h' \notin \{h, h+1\}$ since the left-hand side equals to 0. Moreover, by the definition of $\mathcal{P}_k$, Lemma 12, and $x \le \sqrt{x}$ for $x \in (0,1)$,

$$\begin{aligned}\left|P'_{s,a,h}(s',h) - P_{s,a,h}(s',h)\right| &\le \left|P'_{s,a,h}(s',h) - \gamma\bar{P}_{k,s,a}(s')\right| + \left|\gamma\bar{P}_{k,s,a}(s') - P_{s,a,h}(s',h)\right| \\ &\le 2\gamma\epsilon_k(s,a,s') \le \epsilon^\star_k(s,a,h,s',h),\end{aligned}$$

$$\begin{aligned}&\left|P'_{s,a,h}(s',h+1) - P_{s,a,h}(s',h+1)\right| \\ &\le \left|P'_{s,a,h}(s',h+1) - (1-\gamma)\bar{P}_{k,s,a}(s')\right| + \left|(1-\gamma)\bar{P}_{k,s,a}(s') - P_{s,a,h}(s',h+1)\right| \\ &\le 2(1-\gamma)\epsilon_k(s,a,s') \le \epsilon^\star_k(s,a,h,s',h),\end{aligned}$$

$$\begin{aligned}\left|P'_{s,a,h}(g) - P_{s,a,h}(g)\right| &\le \left|P'_{s,a,h}(g) - \bar{P}_{k,s,a}(g)\right| + \left|\bar{P}_{k,s,a}(g) - P_{s,a,h}(g)\right| \\ &\le 2\epsilon_k(s,a,g) \le 2\epsilon^\star_k(s,a,h,g).\end{aligned}$$

This completes the proof. ■

A.2. Approximation of $Q^{\pi,\Gamma(\pi,\mathcal{P},c),c}$

We show that $Q^{\pi,\Gamma(\pi,\mathcal{P},c),c}$ can be approximated efficiently by Extended Value Iteration similar to (Jaksch et al., 2010). Note that finding $\Gamma(\pi,\mathcal{P},c)$ is equivalent to computing the optimal policy in an augmented MDP $\mathring{\mathcal{M}}$ with state space $\mathring{\mathcal{S}}$ and extended action space $\mathcal{P}$, such that for any extended action $P \in \mathcal{P}$, the cost at $((s,h),P)$ is $\sum_a \pi(a|s,h)c(s,a,h)$, and the transition probability to $\mathring{s}' \in \mathring{\mathcal{S}}_+$ is $\sum_a \pi(a|s,h)P_{s,a,h}(\mathring{s}')$. In this work, we have $\mathcal{P} \in \{\mathcal{P}_k\}_{k=1}^K$, and $\mathcal{P}_k = \bigcap_{s,a,h} \mathcal{P}_{k,s,a,h}$, where $\mathcal{P}_{k,s,a,h}$ is a convex set that specifies constraints on $((s,h),a)$. In other words, $\mathcal{P}_k$ is a product of constraints on each $((s,h),a)$ (note that $\Lambda_{\mathcal{M}}$ can also be decomposed into shared constraints on $P_{s,a,H+1}$ and independent constraints on each $s,a,h \le H$). Thus, any policy in $\mathring{\mathcal{M}}$ can be represented by an element $P \in \mathcal{P}$. We can now perform value iteration in $\mathring{\mathcal{M}}$ to approximate $Q^{\pi,\Gamma(\pi,\mathcal{P},c),c}$. The Bellman operator of $\mathring{\mathcal{M}}$ is $\mathcal{T}_0$ defined in Eq. (19) with $\min$ operator replaced by $\max$ operator. Also note that $\mathring{\mathcal{M}}$ is an SSP instance where all policies are proper. Thus, $V^{\pi,\Gamma(\pi,\mathcal{P},c),c}$ is the unique fixed point of $\mathcal{T}_0$ (Bertsekas and Yu, 2013). It is straightforward to show that Lemma 47 still holds with $\min$ operator replaced by $\max$ operator in Eq. (19) and let $V^0(s,H+1) = \max_a c(s,a,H+1)$. Thus, we can approximate $V^{\pi,\Gamma(\pi,\mathcal{P},c),c}$ efficiently.

Now suppose after n iterations of modified Eq. (19), we obtain V^n such that $\left\|V^n - V^{\pi,\Gamma(\pi,\mathcal{P},c),c}\right\|_\infty \le \epsilon$. Then we can simply use $Q(s,a,h) = c(s,a,h) + \min_{P\in\mathcal{P}} P_{s,a,h}V^n$ to approximate $Q^{\pi,\Gamma(\pi,\mathcal{P},c),c}$, since

$$\left|Q(s,a,h) - Q^{\pi,\Gamma(\pi,\mathcal{P},c),c}(s,a,h)\right| \overset{\text{(i)}}{=} \left|\min_{P\in\mathcal{P}} P_{s,a,h}V^n - \min_{P\in\mathcal{P}} P_{s,a,h}V^{\pi,\Gamma(\pi,\mathcal{P},c),c}\right|$$

$$\leq \max_{P\in\mathcal{P}} \left| P_{s,a,h}(V^n - V^{\pi,\Gamma(\pi,\mathcal{P},c),c}) \right| \leq \left\| V^n - V^{\pi,\Gamma(\pi,\mathcal{P},c),c} \right\|_\infty \leq \epsilon,$$

where (i) is by the definition of $\Gamma(\pi,\mathcal{P},c)$. In this work, setting $\epsilon = 1/K$ is enough for obtaining the desired regret bounds. Lemma 47 (modified) then implies that $\tilde{\mathcal{O}}(T_{\max})$ iterations of modified Eq. (19) suffices.

Appendix B. Omitted Details for Section 3

In this section, we provide omitted discussions and proofs for Section 3.

B.1. Limitation of Existing Approximation Schemes

Finite-Horizon Approximation Thanks to Lemma 31, the approximation error under finite-horizon approximation decreases exponentially. Specifically, we only need a horizon of order $\mathcal{O}(T_{\max}\ln K)$ to have approximation error of order $\mathcal{O}(\frac{1}{K})$. This gives optimal regret bound under both adversarial costs (Chen et al., 2021d) and stochastic costs (Chen et al., 2021a). However, it also clearly brings an extra $\tilde{\mathcal{O}}(T_{\max})$ dependency in the space complexity since we need to store non-stationary policies changing in different layers. Chen et al. (2021a) proposes an implicit finite-horizon approximation analysis that achieves optimal regret bound without storing non-stationary policies. Unfortunately, their approach does not work for adversarial costs.

Discounted Approximation Approximating an SSP by a discounted MDP clearly produces stationary policies. However, the approximation error scales with $1-\gamma$ (that is, inversely proportional to the effective horizon $(1-\gamma)^{-1}$) following similar arguments as in (Wei et al., 2020, Lemma 2), where γ is the discounted factor. This leads to a sub-optimal regret bound when the achieved regret bound in the discounted MDP has polynomial dependency on the horizon even in the lower order term (Wei et al., 2020). In Tarbouriech et al. (2021), they still achieve minimax optimal regret by deriving a horizon-free regret bound (no polynomial dependency on the horizon even in the lower order term), and approximately set $1-\gamma = \tilde{\mathcal{O}}(\frac{1}{K})$ to achieve small approximation error. The drawback, however, is that the time complexity of updating the learner's policy scales linearly w.r.t the effective horizon, which is of order $\tilde{\mathcal{O}}(K)$; see (Tarbouriech et al., 2021, Remark 1).

B.2. Proof of Lemma 4

Proof We only prove the statement for adversarial environment, and the statement for stochastic environment follows directly from setting $c_1 = \cdots c_K = c$. By Lemma 2, we have $V^{\mathring{\pi}^\star,P,c_k}(s,1) \leq V^{\pi^\star,P,c_k}(s) + \frac{1}{K}$ for any $k\in[K]$. Now by Lemma 31 and the fact that the expected hitting time of fast policy is upper bounded by D, we have with probability at least $1-\delta$, the learner reaches the goal within $J_k + c_f$ steps for each episode k. Thus by a union bound, we have with probability at least $1-\delta$, $\sum_{k=1}^K \sum_{i=1}^{I_k} c_i^k \leq \sum_{k=1}^K \left(\sum_{i=1}^{J_k} c_i^k + \mathring{c}_{J_k+1}^k\right)$. Putting everything together, we get:

$$\begin{aligned} R_K &= \sum_{k=1}^K \left(\sum_{i=1}^{I_k} c_i^k - V^{\pi^\star,P,c_k}(s_1^k) \right) \leq \sum_{k=1}^K \left(\sum_{i=1}^{J_k} c_i^k + \mathring{c}_{J_k+1}^k - V^{\mathring{\pi}^\star,P,c_k}(s_1^k,1) \right) + \tilde{\mathcal{O}}(1) \\ &= \mathring{R}_K + \tilde{\mathcal{O}}(1). \end{aligned}$$

This completes the proof. ■

Appendix C. Omitted Details for Section 4

In this section, we provide all proofs for Section 4. We first provide omitted details for cost estimation under various feedback types. Then, we establish the main results in Appendix C.2. Finally, we provide proofs of auxiliary lemmas in Appendix C.3.

Extra Notations Define optimistic transitions $\widetilde{P}_k = \Gamma(\pi_k, \mathcal{P}_k, \widetilde{c}_k)$ and $P_k = \Gamma(\pi_k, \mathcal{P}_k, \widehat{c}_k)$, such that $\widetilde{Q}_k = Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}$ and $\widehat{Q}_k = Q^{\pi_k, P_k, \widehat{c}_k}$. Also define $\widetilde{q}_k = q_{\pi_k, \widetilde{P}_k}$ and $Q_k = Q^{\pi_k, P, \widehat{c}_k}$.

C.1. Cost Estimation

We provide more details on the definition of $\widehat{c}_k$ for the subsequent analysis. Recall that $\widehat{c}_k(s,a) = \max\{0, \bar{c}_k(s,a) - 2\sqrt{\bar{c}_k(s,a)\alpha_k(s,a)} - 7\alpha_k(s,a)\}$. Here, $\bar{c}_k(s,a) = \frac{C_k(s,a)}{\mathfrak{N}_k^+(s,a)}$, where $C_k(s,a)$ is the accumulated costs that are observed at (s,a) in episode $j = 1, \ldots, k-1$ before $\sigma(\pi_j)$ switches to the fast policy, $\alpha_k(s,a) = \frac{\iota}{\mathfrak{N}_k^+(s,a)}$ (recall $\iota = \ln(2SALK/\delta)$), $\mathfrak{N}_k^+(s,a) = \max\{1, \mathfrak{N}_k(s,a)\}$, and $\mathfrak{N}_k$ is the number of times the learner observes cost at (s,a) in episode $j = 1, \ldots, k-1$ before $\sigma(\pi_j)$ switches to the fast policy. The definition of C_k and $\mathfrak{N}_k$ depends on the type of cost feedback. For stochastic costs, $C_k(s,a) = \sum_{j=1}^{k-1} \sum_{i=1}^{J_k} c_i^j \mathbb{I}\{s_i^j = s, a_i^j = a\}$ and $\mathfrak{N}_k = N_k(s,a)$. For stochastic adversary, $C_k(s,a) = \sum_{j=1}^{k-1} m_j(s,a) c_j(s,a)$, where $m_k(s,a)$ is the indicator of whether $c_k(s,a)$ is observed in episode k before $\sigma(\pi_k)$ switches to the fast policy, and $\mathfrak{N}_k(s,a) = M_k(s,a) \triangleq \sum_{j=1}^{k-1} m_j(s,a)$.

Below we show a lemma quantifying the cost estimation error.

Lemma 14 *Under the event of Lemma 11, we have with probability at least* $1-\delta$,

$$0 \le c(s,a) - \widehat{c}_k(s,a) \le 4\sqrt{\widehat{c}_k(s,a)\alpha_k(s,a)} + 34\alpha_k(s,a),$$

for all definitions of $\widehat{c}_k$.

Proof Only prove the stochastic cost case and the stochastic adversary case follows similarly. Note that under the event of Lemma 11, $N_{k+1}(s,a) \le LK$. Applying Lemma 51 with $X_k = c_k(s,a)$ for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ and then by a union bound over all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have with probability at least $1-\delta$, for all $k \in [K]$:

$$|\bar{c}_k(s,a) - c(s,a)| \le 2\sqrt{\alpha_k(s,a)\bar{c}_k(s,a)} + 7\alpha_k(s,a).$$

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

$$\sqrt{\bar{c}_k(s,a)} \le 2\sqrt{\alpha_k(s,a)} + \sqrt{c(s,a) + 7\alpha_k(s,a)} \le \sqrt{c(s,a)} + 5\sqrt{\alpha_k(s,a)},$$

Therefore, $2\sqrt{\alpha_k(s,a)\bar{c}_k(s,a)} + 7\alpha_k(s,a) \le 2\sqrt{\alpha_k(s,a)c(s,a)} + 17\alpha_k(s,a)$, and

$$\begin{aligned} c(s,a) - \widehat{c}_k(s,a) &= c(s,a) - \bar{c}_k(s,a) + \bar{c}_k(s,a) - \widehat{c}_k(s,a) \\ &\le 2 \cdot (2\sqrt{\alpha_k(s,a)\bar{c}_k(s,a)} + 7\alpha_k(s,a)) \le 4\sqrt{\alpha_k(s,a)c(s,a)} + 34\alpha_k(s,a). \end{aligned}$$

This completes the proof. ■

C.2. Main Results for Stochastic Costs and Stochastic Adversary

We first show a general regret bound agnostic to the feedback type (Theorem 15). Then, we present the proofs of Theorem 5 to Theorem 7 (Appendix C.2.1 to Appendix C.2.3) using the general regret bound.

Theorem 15 *Assuming that there exists a constant G such that for any s, h:*

$$\sum_{k=1}^{K-1} \left\langle \pi_{k+1}(\cdot|s,h), d\widetilde{Q}_k(s,\cdot,h) \right\rangle \leq G.$$

Then, Algorithm 1 in stochastic environments with $\lambda \leq \min\{1/T_{\max}, \sqrt{S^2A/K}\}$ ensures with probability at least $1-22\delta$,

$$\begin{aligned}
\mathring{R}_K = \tilde{\mathcal{O}}&\left(\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k - \widehat{c}_k(\mathring{s}_i^k, a_i^k)) - \sum_{k=1}^{K}\langle q_k, e_k\rangle + \sum_{k=1}^{K}\langle q^\star, e_k\rangle + \frac{S^2A}{\lambda}\right) \\
&+ \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^{K}\langle q_k, e_k \circ Q^{\pi_k,P,e_k}\rangle + S^4A^{2.5}T_{\max}^3 + \lambda\sum_{k=1}^{K}\left\langle q^\star, c\circ Q^{\mathring{\pi}^\star,P,\widehat{c}_k}\right\rangle}\right) \\
&+ \tilde{\mathcal{O}}\left(\frac{T_{\max}}{\eta} + T_{\max}G + \lambda\sum_{k=1}^{K}\left\langle q^\star, Q^{\mathring{\pi}^\star,P,e_k}\right\rangle\right).
\end{aligned}$$

Proof For notational convenience, define $\omega = S^4A^{2.5}T_{\max}^3$. By $\langle q^\star, \widehat{c}_k\rangle \leq \langle q^\star, c\rangle$ (Lemma 14) and Lemma 11 (under which $n_k = \bar{n}_k$), we have with probability at least $1-2\delta$,

$$\begin{aligned}
\mathring{R}_K &= \sum_{k=1}^{K}\left(\sum_{i=1}^{\bar{J}_k} c_i^k + \mathring{c}_{J_k+1}^k - V^{\mathring{\pi}^\star,P,c}(\mathring{s}_1^k)\right) \leq \sum_{k=1}^{K}\left(\sum_{i=1}^{\bar{J}_k} c_i^k + \mathring{c}_{J_k+1}^k - V^{\mathring{\pi}^\star,P,\widehat{c}_k}(\mathring{s}_1^k)\right) \\
&\leq \sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k - \widehat{c}_k(\mathring{s}_i^k, a_i^k)) + \sum_{k=1}^{K}\langle \bar{n}_k - q^\star, \widehat{c}_k\rangle.
\end{aligned}$$

For the second term, by the definition of $\widetilde{c}_k$,

$$\begin{aligned}
\sum_{k=1}^{K}\langle \bar{n}_k - q^\star, \widehat{c}_k\rangle &= \sum_{k=1}^{K}\langle \bar{n}_k - q_k, \widehat{c}_k\rangle + \sum_{k=1}^{K}\langle q_k - \widetilde{q}_k, \widetilde{c}_k\rangle + \sum_{k=1}^{K}\langle \widetilde{q}_k - q^\star, \widetilde{c}_k\rangle \\
&\quad - \sum_{k=1}^{K}\langle q_k, e_k\rangle + \sum_{k=1}^{K}\langle q^\star, e_k\rangle - \lambda\sum_{k=1}^{K}\left\langle q_k, \widehat{c}_k\circ\widehat{Q}_k\right\rangle + \lambda\sum_{k=1}^{K}\left\langle q^\star, \widehat{c}_k\circ\widehat{Q}_k\right\rangle \\
&\leq \underbrace{\sum_{k=1}^{K}\langle \bar{n}_k - q_k, \widehat{c}_k\rangle + \sum_{k=1}^{K}\langle q_k - \widetilde{q}_k, \widetilde{c}_k\rangle - \lambda\sum_{k=1}^{K}\langle q_k, \widehat{c}_k\circ Q_k\rangle}_{\xi_1} \\
&\quad - \sum_{k=1}^{K}\langle q_k, e_k\rangle + \sum_{k=1}^{K}\langle q^\star, e_k\rangle + \underbrace{\lambda\sum_{k=1}^{K}\left\langle q_k, \widehat{c}_k\circ(Q_k - \widehat{Q}_k)\right\rangle}_{\xi_2}
\end{aligned}$$

$$+\lambda\sum_{k=1}^{K}\left\langle q^\star, c\circ Q^{\mathring{\pi}^\star,P,\widehat{c}_k}\right\rangle+\underbrace{\sum_{k=1}^{K}\langle \widetilde{q}_k-q^\star,\widetilde{c}_k\rangle+\lambda\sum_{k=1}^{K}\left\langle q^\star, c\circ(\widehat{Q}_k-Q^{\mathring{\pi}^\star,P,\widehat{c}_k})\right\rangle}_{\xi_3}.$$

$$(\widehat{c}_k(s,a,h)\le c(s,a,h))$$

For ξ_1, with probability at least $1-17\delta$:

$$\sum_{k=1}^{K}\langle \bar{n}_k-q_k,\widehat{c}_k\rangle+\sum_{k=1}^{K}\langle q_k-\widetilde{q}_k,\widetilde{c}_k\rangle-\lambda\sum_{k=1}^{K}\langle q_k,\widehat{c}_k\circ Q_k\rangle\le\tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\langle q_k,\widehat{c}_k\circ Q_k\rangle}+SAT_{\max}\right)$$

$$+\sum_{k=1}^{K}\left\langle q_k-\widetilde{q}_k,(1+\lambda\widehat{Q}_k)\circ\widehat{c}_k\right\rangle+\sum_{k=1}^{K}\langle q_k-\widetilde{q}_k,e_k\rangle-\lambda\sum_{k=1}^{K}\langle q_k,\widehat{c}_k\circ Q_k\rangle$$

$$(\mathbb{E}_k[\bar{n}_k(s,a,h)]\le q_k(s,a,h)\text{, Lemma 50, Lemma 26, and }\bar{n}_k(s,a,h)\le L=\tilde{\mathcal{O}}(T_{\max}))$$

$$=\tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^{K}\langle q_k,\widehat{c}_k\circ Q_k\rangle}+\sqrt{S^2A\sum_{k=1}^{K}\langle q_k,e_k\circ Q^{\pi_k,P,e_k}\rangle}+\omega\right)-\lambda\sum_{k=1}^{K}\langle q_k,\widehat{c}_k\circ Q_k\rangle$$

$$(\text{Lemma 28 and }(1+\lambda\widehat{Q}_k(s,a,h))\widehat{c}_k(s,a,h)=\tilde{\mathcal{O}}(\widehat{c}_k(s,a,h)))$$

$$=\tilde{\mathcal{O}}\left(\frac{S^2A}{\lambda}+\sqrt{S^2A\sum_{k=1}^{K}\langle q_k,e_k\circ Q^{\pi_k,P,e_k}\rangle}+\omega\right).\qquad\text{(AM-GM inequality)}$$

For ξ_2, by Lemma 30 and Lemma 13, with probability at least $1-2\delta$,

$$Q_k(s,a,h)-\widehat{Q}_k(s,a,h)=\sum_{s',a',h'}q_{k,(s,a,h)}(s',a',h')(P_{s',a',h'}-P_{k,s',a',h'})V^{\pi_k,P_k,\widehat{c}_k}$$

$$=\tilde{\mathcal{O}}\left(\sum_{s',a'}q_{k,(s,a,h)}(s',a')\left(\frac{\sqrt{S}T_{\max}}{\sqrt{N_k^+(s',a')}}+\frac{ST_{\max}}{N_k^+(s',a')}\right)\right).\tag{6}$$

By $q_k(s,a,h)=\frac{x_k(s,a,h)}{1-y_k(s,a,h)}$ and $y_k(s,a,h)\le\gamma=1-\frac{1}{2T_{\max}}$, we have

$$\sum_{s,a,h\le H}q_k(s,a,h)q_{k,(s,a,h)}(s',a')\le 2T_{\max}\sum_{s,a,h\le H}x_k(s,a,h)q_{k,(s,a,h)}(s',a')$$

$$\le 2T_{\max}\sum_{s,a,h\le H}q_k(s',a')=2T_{\max}SAHq_k(s',a').\tag{7}$$

Therefore, with probability at least $1-\delta$,

$$\xi_2=\lambda\sum_{k=1}^{K}\left\langle q_k,\widehat{c}_k\circ(Q_k-\widehat{Q}_k)\right\rangle$$

$$=\tilde{\mathcal{O}}\left(\lambda\sum_{k=1}^{K}\sum_{s,a,h}q_k(s,a,h)\sum_{s',a'}q_{k,(s,a,h)}(s',a')\left(\frac{\sqrt{S}T_{\max}}{\sqrt{N_k^+(s',a')}}+\frac{ST_{\max}}{N_k^+(s',a')}\right)\right)$$

$$= \tilde{\mathcal{O}}\left(\lambda T_{\max}^2 S^{3/2} A \sum_{s',a'} \sum_{k=1}^{K} \frac{q_k(s',a')}{\sqrt{N_k^+(s',a')}} + \lambda T_{\max}^2 S^2 A \sum_{s',a'} \sum_{k=1}^{K} \frac{q_k(s',a')}{N_k^+(s',a')}\right) \qquad \text{(Eq. (7))}$$

$$= \tilde{\mathcal{O}}\left(\lambda T_{\max}^2 S^{3/2} A \sqrt{SAT_{\max}K} + \lambda S^3 A^2 T_{\max}^3\right) = \tilde{\mathcal{O}}\left(\omega\right).$$

(Lemma 32 and $\sum_{s,a} q_k(s,a) = \tilde{\mathcal{O}}(T_{\max})$)

For ξ_3, first note that $\left\|\widetilde{Q}_1\right\|_\infty = \tilde{\mathcal{O}}(T_{\max})$ under all definitions of $\widetilde{c}_k$, and by Lemma 30:

$$\sum_{k=1}^{K} \langle \widetilde{q}_k - q^\star, \widetilde{c}_k \rangle = \sum_{k=1}^{K} \sum_{s,h} q^\star(s,h) \sum_a \left(\pi_k(a|s,h) - \pi^\star(a|s,h)\right) \widetilde{Q}_k(s,a,h)$$
$$+ \sum_{k=1}^{K} \sum_{s,a,h} q^\star(s,a,h) \left(\widetilde{Q}_k(s,a,h) - \widetilde{c}_k(s,a,h) - P_{s,a,h} V^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}\right)$$
$$= \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta} + T_\star G + T_{\max}^2\right).$$

(Lemma 24, the definition of $\widetilde{P}_k$ and $\sum_{s,a,h} q^\star(s,a,h) = \tilde{\mathcal{O}}(T_\star)$ by Lemma 2)

Next, note that

$$\sum_{k=1}^{K} (\widehat{Q}_k(s,a,h) - Q^{\mathring{\pi}^\star, P, \widehat{c}_k}(s,a,h)) \le \sum_{k=1}^{K} (Q^{\pi_k, \widetilde{P}_k, \widehat{c}_k}(s,a,h) - Q^{\mathring{\pi}^\star, P, \widehat{c}_k}(s,a,h))$$

$(P_k, \widetilde{P}_k \in \mathcal{P}_k)$

$$\le \sum_{k=1}^{K} \left(Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,a,h) - Q^{\mathring{\pi}^\star, P, \widetilde{c}_k}(s,a,h)\right) + \sum_{k=1}^{K} Q^{\mathring{\pi}^\star, P, \lambda \widehat{Q}_k + e_k}(s,a,h) \quad \text{(definition of } \widetilde{c}_k)$$

Also note that $\lambda \sum_{k=1}^{K} \left\langle q^\star, Q^{\mathring{\pi}^\star, P, \lambda \widehat{Q}_k} \right\rangle = \tilde{\mathcal{O}}(\lambda^2 T_{\max}^3 K) = \tilde{\mathcal{O}}(S^2 A T_{\max}^3)$ by $\lambda \le \sqrt{S^2 A / K}$. Thus,

$$\lambda \sum_{k=1}^{K} \left\langle q^\star, c \circ (\widehat{Q}_k - Q^{\mathring{\pi}^\star, P, \widehat{c}_k}) \right\rangle$$
$$= \tilde{\mathcal{O}}\left(\lambda \sum_{k=1}^{K} \left\langle q^\star, c \circ (\widetilde{Q}_k - Q^{\mathring{\pi}^\star, P, \widetilde{c}_k}) \right\rangle + \lambda \sum_{k=1}^{K} \left\langle q^\star, Q^{\mathring{\pi}^\star, P, e_k} \right\rangle + S^2 A T_{\max}^3\right).$$

Now by Lemma 30 and the definition of $\widetilde{P}_k$:

$$\sum_{k=1}^{K} (\widetilde{Q}_k(s,a,h) - Q^{\mathring{\pi}^\star, P, \widetilde{c}_k}(s,a,h)) \tag{8}$$
$$\le \sum_{k=1}^{K} \sum_{s'',h''} P_{s,a,h}(s'',h'') \sum_{s',a',h'} q^\star_{(s'',h'')}(s',h') \left(\pi_k(a'|s',h') - \mathring{\pi}^\star(a'|s',h')\right) \widetilde{Q}_k(s',a',h')$$

$$= \tilde{\mathcal{O}}\left(\frac{T_{\max}}{\eta} + T_{\max}G + T_{\max}^2\right). \qquad \text{(Lemma 24)}$$

Thus, by $\lambda T_{\max} \le 1$, we have $\lambda \sum_{k=1}^{K} \left\langle q^\star, c \circ (\widetilde{Q}_k - Q^{\mathring{\pi}^\star, P, \widetilde{c}_k}) \right\rangle = \tilde{\mathcal{O}}(\frac{T_{\max}}{\eta} + T_{\max}G + T_{\max}^2)$. Putting everything together completes the proof. ■

C.2.1. PROOF OF THEOREM 5

Proof By Lemma 25 with $\mathfrak{n}_k = n_k$, $\mathfrak{N}_k = N_k$, and $e_k(s,a,h) = 0$, with probability at least $1 - 2\delta$:

$$\sum_{k=1}^{K-1} \left\langle \pi_{k+1}(\cdot|s,h), d\widetilde{Q}_k(s,\cdot,h) \right\rangle = \tilde{\mathcal{O}}\left(T_{\max}^2 \sum_{k=1}^{K} \sum_{s',a'} \frac{S n_k(s',a')}{N_k^+(s',a')} + \lambda\eta T_{\max}^4 K\right)$$
$$= \tilde{\mathcal{O}}\left(S^2 A T_{\max}^3 + T_{\max}^2 (S^2AK)^{1/4}\right).$$

(definition of λ and η, $n_k(s,a) = \bar{n}_k(s,a)$ under the event of Lemma 11, and Lemma 32)

Thus, by Theorem 15, Lemma 16, definition of λ, and replacing G by the bound above, we have with probability at least $1 - 28\delta$,

$$\mathring{R}_K = \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{k=1}^{K}\sum_{i=1}^{J_k} c_i^k} + \frac{S^2A}{\lambda} + T_{\max}^3 (S^2AK)^{1/4} + S^4A^{2.5}T_{\max}^4 + \lambda \sum_{k=1}^{K} \left\langle q^\star, c \circ Q^{\mathring{\pi}^\star, P, \widehat{c}_k} \right\rangle\right)$$
$$= \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{k=1}^{K}\sum_{i=1}^{J_k} c_i^k} + B_\star S\sqrt{AK} + T_{\max}^3 (S^2AK)^{1/4} + S^4A^{2.5}T_{\max}^4\right). \qquad \text{(Lemma 20)}$$

Now by $\mathring{R}_k = \sum_{k=1}^{K}\sum_{i=1}^{J_k} c_i^k - K \cdot V^{\mathring{\pi}^\star, P, c}(\mathring{s}_1^k)$ and Lemma 48, we have $\sum_{k=1}^{K}\sum_{i=1}^{J_k} c_i^k = \tilde{\mathcal{O}}(B_\star K)$. Plugging this back, we get $\mathring{R}_K = \tilde{\mathcal{O}}(B_\star S\sqrt{AK} + T_{\max}^3 (S^2AK)^{1/4} + S^4A^2T_{\max}^4)$. Applying Lemma 4 then completes the proof. ■

C.2.2. PROOF OF THEOREM 6

Proof First note that with probability at least $1 - 3\delta$,

$$\sum_{k=1}^{K} \|de_k\|_1 \le \sum_{k=1}^{K} \sum_{s,a,h\le H} \left| \sqrt{\frac{\widehat{c}_k(s,a,h)}{k}} - \sqrt{\frac{\widehat{c}_{k+1}(s,a,h)}{k+1}} \right| + \beta' \sum_{k=1}^{K} \sum_{s,a,h\le H} \left| d\widehat{Q}_k(s,a,h) \right|$$
$$= \tilde{\mathcal{O}}\left(S^3A^2T_{\max}^2 + S^{1/2}A^{3/4}T_{\max}K^{1/4}\right),$$

where in the last inequality we apply

$$\sum_{k=1}^{K} \sum_{s,a,h\le H} \left| \sqrt{\frac{\widehat{c}_k(s,a,h)}{k}} - \sqrt{\frac{\widehat{c}_{k+1}(s,a,h)}{k+1}} \right|$$

$$\leq \sum_{k=1}^{K} \sum_{s,a,h \leq H} \left(\frac{1}{\sqrt{k}} - \frac{1}{\sqrt{k+1}} \right) + \sum_{k=1}^{K} \sum_{s,a,h \leq H} \frac{\sqrt{|\widehat{c}_k(s,a,h) - \widehat{c}_{k+1}(s,a,h)|}}{\sqrt{k+1}}$$

(add and subtract $\sqrt{\widehat{c}_k(s,a,h)/(k+1)}$, and $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a-b|}$)

$$= \tilde{\mathcal{O}} \left(SA + \sqrt{\sum_{k=1}^{K} \sum_{s,a,h \leq H} \frac{1}{k+1}} \sqrt{\sum_{k=1}^{K} \sum_{s,a,h \leq H} \frac{m_k(s,a)}{M_k^+(s,a)}} \right) = \tilde{\mathcal{O}}\left(SA\right),$$

(Cauchy-Schwarz inequality, Lemma 25, and Lemma 32)

and by Lemma 11,

$$\beta' \sum_{k=1}^{K} \sum_{s,a,h \leq H} \left| d\widehat{Q}_k(s,a,h) \right|$$

$$= \tilde{\mathcal{O}} \left(\beta' \sum_{k=1}^{K} \sum_{s,a,h \leq H} \left(T_{\max}^2 \sum_{s',a'} \frac{S n_k(s',a')}{N_k^+(s',a')} + T_{\max} \sum_{s',a'} \frac{m_k(s',a')}{M_k^+(s',a')} + \eta T_{\max}^3 \right) \right)$$

(Lemma 25)

$$= \tilde{\mathcal{O}} \left(\beta' S^3 A^2 T_{\max}^3 + \beta' \eta S A T_{\max}^3 K \right) = \tilde{\mathcal{O}} \left(S^3 A^2 T_{\max}^2 + S^{1/2} A^{3/4} T_{\max} K^{1/4} \right). \quad \text{(Lemma 32)}$$

Moreover, by Lemma 25 with $\mathfrak{n}_k = m_k$, $\mathfrak{N}_k = M_k$, and $\lambda \leq \frac{1}{T_{\max}}$, we have with probability at least $1 - \delta$:

$$\sum_{k=1}^{K-1} \left\langle \pi_{k+1}(\cdot|s,h), d\widetilde{Q}_k(s,\cdot,h) \right\rangle$$

$$= \tilde{\mathcal{O}} \left(T_{\max}^2 \sum_{k=1}^{K} \sum_{s',a'} \frac{S n_k(s',a')}{N_k^+(s',a')} + \lambda T_{\max}^2 \sum_{k=1}^{K} \sum_{s',a'} \frac{m_k(s',a')}{M_k^+(s',a')} + \lambda \eta T_{\max}^4 K + T_{\max} \sum_{k=1}^{K} \|de_k\|_1 \right)$$

$$= \tilde{\mathcal{O}} \left(S^3 A^2 T_{\max}^3 + T_{\max}^2 (S^2 A^3 K)^{1/4} \right), \tag{9}$$

where the last step is by Lemma 32, the definition of η and λ, and the bound on $\sum_{k=1}^{K} \|de_k\|_1$. Moreover, by Lemma 17 and definition of e_k, we have with probability at least $1 - 16\delta$:

$$\sum_{k=1}^{K} \sum_{i=1}^{\bar{J}_k} \left(c_i^k - \widehat{c}_k(\mathring{s}_i^k, a_i^k) \right) - \sum_{k=1}^{K} \langle q_k, e_k \rangle = \tilde{\mathcal{O}} \left(\beta' \sum_{k=1}^{K} \left\langle q_k, Q_k - \widehat{Q}_k \right\rangle + \frac{1}{\beta'} + \sqrt{S^3 A^3 T_{\max}^3} \right)$$

$$= \tilde{\mathcal{O}} \left(S^3 A^2 T_{\max}^3 + \sqrt{D T_\star K} \right),$$

(Eq. (6), Eq. (7) similar to bounding ξ_2, and the definition of β')

$$\sum_{k=1}^{K} \langle q^\star, e_k \rangle = \tilde{\mathcal{O}} \left(\sum_{k=1}^{K} \sum_{s,a} q^\star(s,a) \sqrt{\frac{c(s,a)}{k}} + \beta' \sum_{k=1}^{K} \left\langle q^\star, \widehat{Q}_k \right\rangle \right) \quad \text{(Lemma 14)}$$

$$\overset{\text{(i)}}{=} \tilde{\mathcal{O}} \left(\sqrt{D T_\star K} + S^3 A^2 T_{\max}^4 \right),$$

$$\sqrt{S^2 A \sum_{k=1}^{K} \langle q_k, e_k \circ Q^{\pi_k, P, e_k} \rangle} = \tilde{\mathcal{O}} \left(\sqrt{S^2 A T_{\max}^4} \right),$$

$$\lambda \sum_{k=1}^{K} \left\langle q^\star, Q^{\mathring{\pi}^\star, P, e_k} \right\rangle = \tilde{\mathcal{O}}\left(\lambda T_{\max}^2 \sqrt{K} + \lambda \beta' T_{\max}^3 K\right) = \tilde{\mathcal{O}}\left(\sqrt{S^2 A} T_{\max}^3\right),$$

where (i) is by

$$\sum_{k=1}^{K} \sum_{s,a} q^\star(s,a) \sqrt{\frac{c(s,a)}{k}} = \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K} \sum_{s,a} q^\star(s,a) c(s,a)} \sqrt{\sum_{k=1}^{K} \sum_{s,a} \frac{q^\star(s,a)}{k}}\right) = \tilde{\mathcal{O}}\left(\sqrt{D T_\star K}\right),$$

(Cauchy-Schwarz inequality)

definition of β', and

$$\begin{aligned}
&\beta' \sum_{k=1}^{K} \left\langle q^\star, \widehat{Q}_k \right\rangle = \beta' \sum_{k=1}^{K} \left\langle q^\star, \widehat{Q}_k - Q^{\mathring{\pi}^\star, P, \widehat{c}_k} \right\rangle + \beta' \sum_{k=1}^{K} \left\langle q^\star, Q^{\mathring{\pi}^\star, P, \widehat{c}_k} \right\rangle \\
&= \tilde{\mathcal{O}}\left(\beta' \sum_{k=1}^{K} \sum_{s,a,h} q^\star(s,a,h) \left(Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,a,h) - Q^{\mathring{\pi}^\star, P, \widetilde{c}_k}(s,a,h)\right)\right) \\
&\quad + \tilde{\mathcal{O}}\left(\beta' \sum_{s,a,h} \sum_{k=1}^{K} q^\star(s,a,h) Q^{\mathring{\pi}^\star, P, \lambda \widehat{Q}_k + e_k}(s,a,h) + \sqrt{D T_\star K}\right), \\
&\qquad\qquad \left(\textstyle\sum_{s,a,h} q^\star(s,a,h) = \mathcal{O}(T_\star), \text{ and } \left\|Q^{\mathring{\pi}^\star, P, \widehat{c}_k}\right\|_\infty = \mathcal{O}(D)\right) \\
&= \tilde{\mathcal{O}}\left(\frac{\beta' T_{\max}^2}{\eta} + \beta' T_{\max}^2 G + \beta' T_{\max}^3 + (\lambda \beta' + \beta'^2) T_{\max}^3 K + \beta' T_{\max}^2 \sqrt{K} + \sqrt{D T_\star K}\right), \\
&\qquad\qquad \text{(Eq. (8))} \\
&= \tilde{\mathcal{O}}\left(S^3 A^2 T_{\max}^4 + \sqrt{D T_\star K}\right). \qquad \text{(replace } G \text{ by Eq. (9))}
\end{aligned}$$

Thus, by Theorem 15, Lemma 21, and definition of η, λ, we have with probability at least $1 - 22\delta$,

$$\mathring{R}_K = \tilde{\mathcal{O}}\left(\sqrt{D T_\star K} + DS\sqrt{AK} + T_{\max}^3 (S^2 A^3 K)^{1/4} + S^4 A^{2.5} T_{\max}^4\right).$$

Applying Lemma 4 completes the proof. ■

C.2.3. PROOF OF THEOREM 7

Proof By Lemma 25 with $\mathfrak{n}_k = m_k$, $\mathfrak{N}_k = M_k$, and $\lambda \le \frac{1}{T_{\max}}$, we have with probability at least $1 - 2\delta$:

$$\begin{aligned}
&\sum_{k=1}^{K-1} \left\langle \pi_{k+1}(\cdot|s,h), d\widetilde{Q}_k(s,\cdot,h) \right\rangle \qquad (10) \\
&= \tilde{\mathcal{O}}\left(T_{\max}^2 \sum_{k=1}^{K} \sum_{s',a'} \frac{S n_k(s',a')}{N_k^+(s',a')} + \lambda T_{\max}^2 \sum_{k=1}^{K} \sum_{s',a'} \frac{m_k(s',a')}{M_k^+(s',a')} + \lambda \eta T_{\max}^4 K + T_{\max} \sum_{k=1}^{K} \|de_k\|_1\right) \\
&= \tilde{\mathcal{O}}\left(S^3 A^2 T_{\max}^3 + T_{\max}^2 S A^{5/4} K^{1/4}\right), \qquad \text{(definition of } \eta \text{ and Lemma 32)}
\end{aligned}$$

where in the last step we apply

$$\sum_{k=1}^{K}\|de_k\|_1 = \beta\sum_{k=1}^{K}\sum_{s,a,h\le H}\left|d\widehat{Q}_k(s,a,h)\right|$$
$$= \tilde{\mathcal{O}}\left(\beta\sum_{k=1}^{K}\sum_{s,a,h\le H}\left(T_{\max}^2\sum_{s',a'}\frac{Sn_k(s',a')}{N_k^+(s',a')} + SAT_{\max}\frac{m_k(s,a)}{M_k^+(s,a)} + \eta T_{\max}^3\right)\right) \quad \text{(Lemma 25)}$$
$$= \tilde{\mathcal{O}}\left(\beta S^3A^2T_{\max}^3 + \beta\eta SAT_{\max}^3K\right) = \tilde{\mathcal{O}}\left(S^3A^2T_{\max}^2 + SA^{5/4}T_{\max}K^{1/4}\right). \quad \text{(Lemma 32)}$$

By Lemma 18 and the definition of e_k, we have with probability at least $1-11\delta$:

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}\left(c_i^k - \widehat{c}_k(\mathring{s}_i^k, a_i^k)\right) - \sum_{k=1}^{K}\langle q_k, e_k\rangle = \tilde{\mathcal{O}}\left(\beta\sum_{k=1}^{K}\left\langle q_k, Q_k - \widehat{Q}_k\right\rangle + \frac{SA}{\beta} + \sqrt{S^3A^3T_{\max}^3}\right)$$
$$= \tilde{\mathcal{O}}\left(\sqrt{SADT_\star K} + S^{2.5}A^2T_{\max}^3\right),$$
(Eq. (6), Eq. (7) similar to bounding ξ_2, and the definition of β)

$$\sum_{k=1}^{K}\langle q^\star, e_k\rangle \le \beta\sum_{k=1}^{K}\left\langle q^\star, \widehat{Q}_k - Q^{\mathring{\pi}^\star,P,\widehat{c}_k}\right\rangle + \beta\sum_{k=1}^{K}\left\langle q^\star, Q^{\mathring{\pi}^\star,P,\widehat{c}_k}\right\rangle$$
$$= \tilde{\mathcal{O}}\left(\beta\sum_{k=1}^{K}\sum_{s,a,h}q^\star(s,a,h)\left(Q^{\pi_k,\widetilde{P}_k,\widetilde{c}_k}(s,a,h) - Q^{\mathring{\pi}^\star,P,\widetilde{c}_k}(s,a,h)\right)\right)$$
$$+ \tilde{\mathcal{O}}\left(\beta\sum_{s,a,h}\sum_{k=1}^{K}q^\star(s,a,h)Q^{\mathring{\pi}^\star,P,\lambda\widehat{Q}_k+e_k}(s,a,h) + \sqrt{SADT_\star K}\right),$$
($\sum_{s,a,h}q^\star(s,a,h) = \mathcal{O}(T_\star)$, and $\left\|Q^{\mathring{\pi}^\star,P,\widehat{c}_k}\right\|_\infty = \mathcal{O}(D)$)

$$= \tilde{\mathcal{O}}\left(\frac{\beta T_{\max}^2}{\eta} + \beta T_{\max}^2G + \beta T_{\max}^3 + (\lambda\beta+\beta^2)T_{\max}^3K + \sqrt{SADT_\star K}\right), \quad \text{(Eq. (8))}$$
$$= \tilde{\mathcal{O}}\left(S^3A^2T_{\max}^4 + \sqrt{SADT_\star K}\right), \quad \text{(replace } G \text{ by Eq. (10))}$$

$$\sqrt{S^2A\sum_{k=1}^{K}\langle q_k, e_k\circ Q^{\pi_k,P,e_k}\rangle} = \tilde{\mathcal{O}}\left(\sqrt{\beta^2S^2AT_{\max}^4K}\right) = \tilde{\mathcal{O}}\left(\sqrt{S^3A^2T_{\max}^4}\right),$$
$$\lambda\sum_{k=1}^{K}\left\langle q^\star, Q^{\mathring{\pi}^\star,P,e_k}\right\rangle = \tilde{\mathcal{O}}\left(\lambda\beta T_{\max}^3K\right) = \tilde{\mathcal{O}}\left(S^{3/2}AT_{\max}^3\right).$$

Thus, by Theorem 15, definition of η, λ, and β, and Lemma 21, with probability at least $1-22\delta$,

$$\mathring{R}_K = \tilde{\mathcal{O}}\left(\sqrt{SADT_\star K} + DS\sqrt{AK} + T_{\max}^3SA^{5/4}K^{1/4} + S^4A^{2.5}T_{\max}^4\right).$$

Applying Lemma 4 completes the proof. ■

C.3. Extra Lemmas for Section 4

We give an outline of this section: Lemma 16 to Lemma 18 bound the term $\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k - \widehat{c}_k(\mathring{s}_i^k, a_i^k))$ the under various feedback types. Lemma 20 and Lemma 21 bound the term $\sum_{k=1}^{K}\left\langle q^\star, c \circ Q^{\mathring{\pi}^\star, P, c}\right\rangle$ under stochastic costs and stochastic adversaries respectively. Lemma 23 establishes stability of PO updates. Lemma 24 provide a refined analysis of PO. Lemma 25 bounds the drift of various quantities (such as $d\widehat{c}_k$ and $d\widetilde{Q}_k$) across episodes. Lemma 26 provide bounds on variance of learner's costs. Lemma 28 gives a bound on the estimation error of value functions due to transition estimation.

Lemma 16 *Under stochastic costs, we have with probability at least* $1-6\delta$*:*

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k - \widehat{c}_k(\mathring{s}_i^k, a_i^k)) = \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{k=1}^{K}\sum_{i=1}^{J_k} c_i^k} + SAT_{\max}\right).$$

Proof First note that:

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k - \widehat{c}_k(\mathring{s}_i^k, a_i^k)) = \sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k - c(s_i^k, a_i^k)) + \sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c(s_i^k, a_i^k) - \widehat{c}_k(s_i^k, a_i^k)).$$

For the first term, by Lemma 50 and Lemma 52, we have with probability at least $1-2\delta$,

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k - c(s_i^k, a_i^k)) &= \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}\mathbb{E}[(c_i^k)^2|s_i^k, a_i^k]}\right) = \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k} c(s_i^k, a_i^k)}\right)\\ &= \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{J_k} c_i^k}\right).\end{aligned}$$

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

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c(s_i^k, a_i^k) - \widehat{c}_k(\mathring{s}_i^k, a_i^k)) &= \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}\left(\sqrt{\frac{c(s_i^k, a_i^k)}{N_k^+(s_i^k, a_i^k)}} + \frac{1}{N_k^+(s_i^k, a_i^k)}\right)\right)\\ &\qquad\qquad \text{(Lemma 14 and } \widehat{c}_k(s,a) \le c(s,a)\text{)}\\ &= \tilde{\mathcal{O}}\left(\sum_{s,a}\sum_{k=1}^{K}\left(\bar{n}_k(s,a)\sqrt{\frac{c(s,a)}{N_k^+(s,a)}} + \frac{\bar{n}_k(s,a)}{N_k^+(s,a)}\right)\right)\\ &= \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k} c(s_i^k, a_i^k)} + SAT_{\max}\right) \qquad \text{(Lemma 32 and } J_k = \bar{J}_k\text{)}\\ &= \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k} c_i^k} + SAT_{\max}\right). \qquad \text{(Lemma 52)}\end{aligned}$$

This completes the proof. ■

Lemma 17 *Under stochastic adversary with full information, with probability at least* $1-8\delta$,

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k-\widehat{c}_k(\mathring{s}_i^k,a_i^k))=8\iota\cdot\sum_{k=1}^{K}\sum_{s,a}q_k(s,a)\sqrt{\widehat{c}_k(s,a)/k}$$
$$+\tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{s,a,h\le H}q_k(s,a,h)Q_k(s,a,h)}+\sqrt{S^3A^3T_{\max}^3}\right).$$

Proof First note that by $c_i^k=c_k(s_i^k,a_i^k)$:

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k-\widehat{c}_k(\mathring{s}_i^k,a_i^k))=\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_k(s_i^k,a_i^k)-c(s_i^k,a_i^k))+\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c(s_i^k,a_i^k)-\widehat{c}_k(s_i^k,a_i^k)).$$

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

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_k(s_i^k,a_i^k)-c(s_i^k,a_i^k))=\sum_{k=1}^{K}\sum_{s,a}\bar{n}_k(s,a)(c_k(s,a)-c(s,a))$$
$$\overset{\text{(i)}}{=}\tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\mathbb{E}_k\left[\left(\sum_{s,a}\bar{n}_k(s,a)c_k(s,a)\right)^2\right]}+T_{\max}\right)$$
$$=\tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\mathbb{E}_{c_k}\left[\sum_{s,a,h\le H}q_k(s,a,h)Q^{\pi_k,P,c_k}(s,a,h)\right]}+T_{\max}\right)$$
$$(\mathbb{E}_k[\cdot]=\mathbb{E}_{c_k,\bar{n}_k}[\cdot],\text{ Lemma 26 and }c_k(s,a)\le 1)$$
$$=\tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{s,a,h\le H}q_k(s,a,h)Q_k(s,a,h)+\sum_{k=1}^{K}\langle q_k,Q^{\pi_k,P,c}-Q_k\rangle}+T_{\max}\right),$$
$$(\widehat{c}_k(s,a,H+1)=c(s,a,H+1))$$

where in (i) we apply Lemma 50, $\mathbb{E}_k[\cdot]=\mathbb{E}_{c_k,\bar{n}_k}[\cdot]$, and

$$\mathbb{E}_{c_k}\left[\left(\sum_{s,a}\bar{n}_k(s,a)(c_k(s,a)-c(s,a))\right)^2\middle|\bar{n}_k\right]\le\mathbb{E}_{c_k}\left[\left(\sum_{s,a}\bar{n}_k(s,a)c_k(s,a)\right)^2\middle|\bar{n}_k\right].$$

Now note that for $h\le H$, by Lemma 30, Lemma 14, and $\widehat{c}_k(s,a,H+1)=c(s,a,H+1)$, we have with probability at least $1-2\delta$:

$$Q^{\pi_k,P,c}(s,a,h)-Q_k(s,a,h)=\sum_{s',a',h'\le H}q_{k,(s,a,h)}(s',a',h')(c(s',a',h')-\widehat{c}_k(s',a',h'))$$
$$=\tilde{\mathcal{O}}\left(\sum_{s',a'}q_{k,(s,a,h)}(s',a')\left(\sqrt{\frac{\widehat{c}_k(s',a')}{M_k^+(s',a')}}+\frac{1}{M_k^+(s',a')}\right)\right).$$

Note that $q_k(s,a,h)q_{k,(s,a,h)}(s',a') = \tilde{\mathcal{O}}(T_{\max}x_k(s,a,h)q_{k,(s,a,h)}(s',a')) = \tilde{\mathcal{O}}(T_{\max}q_k(s',a'))$. Therefore, we have with probability at least $1-\delta$:

$$\sum_{k=1}^{K}\left\langle q_k, Q^{\pi_k,P,c}-Q_k\right\rangle = \tilde{\mathcal{O}}\left(T_{\max}\sum_{k=1}^{K}\sum_{s,a,h\le H}\sum_{s',a'}q_k(s',a')\left(\sqrt{\frac{\widehat{c}_k(s',a')}{M_k^+(s',a')}}+\frac{1}{M_k^+(s',a')}\right)\right)$$

$$= \tilde{\mathcal{O}}\left(SAT_{\max}\sum_{k=1}^{K}\sum_{s',a'}q_k(s',a')\left(\sqrt{\frac{\widehat{c}_k(s',a')}{M_k^+(s',a')}}+\frac{1}{M_k^+(s',a')}\right)\right)$$

$(q_k(s',a',h') = \mathcal{O}(T_{\max}x_k(s',a',h')))$

$$= \tilde{\mathcal{O}}\left(SAT_{\max}\left(\sqrt{\sum_{k=1}^{K}\sum_{s',a'}\frac{q_k(s',a')}{M_k^+(s',a')}}\sqrt{\sum_{k=1}^{K}\sum_{s',a'}q_k(s',a')\widehat{c}_k(s',a')}+\sum_{k=1}^{K}\sum_{s',a'}\frac{q_k(s',a')}{M_k^+(s',a')}\right)\right)$$

(Cauchy-Schwarz inequality)

$$= \tilde{\mathcal{O}}\left(\sqrt{S^3A^3T_{\max}^3\sum_{k=1}^{K}\sum_{s',a'}q_k(s',a')\widehat{c}_k(s',a')+S^2A^2T_{\max}^2}\right).$$

$(q_k(s',a') \le T_{\max}x_k(s',a')$ and Lemma 32)

$$= \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{s',a'}q_k(s',a')\widehat{c}_k(s',a')+S^3A^3T_{\max}^3\right).$$ (AM-GM inequality)

Substituting these back, we have

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_k(s_i^k,a_i^k)-c(s_i^k,a_i^k))$$
$$= \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{s,a,h\le H}q_k(s,a,h)Q_k(s,a,h)}+\sqrt{S^3A^3T_{\max}^3}\right). \tag{11}$$

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

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c(s_i^k,a_i^k)-\widehat{c}_k(\mathring{s}_i^k,a_i^k)) \le \sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}\left(4\sqrt{\frac{\widehat{c}_k(s_i^k,a_i^k)\iota}{M_k^+(s_i^k,a_i^k)}}+\frac{34\iota}{M_k^+(s_i^k,a_i^k)}\right) \quad \text{(Lemma 14)}$$

$$\le \sum_{k=1}^{K}\sum_{s,a}8\cdot q_k(s,a)\iota\sqrt{\frac{\widehat{c}_k(s,a)}{k}}+\sum_{k=1}^{K}\sum_{s,a}\frac{68q_k(s,a)\iota}{k}+\tilde{\mathcal{O}}\left(T_{\max}\right) \quad \text{(Lemma 52)}$$

$$\le 8\iota\cdot\sum_{k=1}^{K}\sum_{s,a}q_k(s,a)\sqrt{\widehat{c}_k(s,a)/k}+\tilde{\mathcal{O}}\left(T_{\max}\right).$$

Putting everything together completes the proof. ■

Lemma 18 *Under stochastic adversary with bandit feedback, with probability at least* $1-8\delta$,

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k-\widehat{c}_k(\mathring{s}_i^k,a_i^k))=\tilde{\mathcal{O}}\left(\sqrt{SA\sum_{k=1}^{K}\sum_{s,a}\sum_{h=1}^{H}q_k(s,a,h)Q_k(s,a,h)}+\sqrt{S^3A^3T_{\max}^3}\right).$$

Proof First note that by $c_i^k=c_k(s_i^k,a_i^k)$:

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_i^k-\widehat{c}_k(\mathring{s}_i^k,a_i^k))=\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c_k(s_i^k,a_i^k)-c(s_i^k,a_i^k))+\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c(s_i^k,a_i^k)-\widehat{c}_k(s_i^k,a_i^k)).$$

For the first term, Eq. (11) holds by the same arguments as in Lemma 17 with probability at least $1-4\delta$. For the second term, we have with probability at least $1-4\delta$,

$$\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}(c(s_i^k,a_i^k)-\widehat{c}_k(\mathring{s}_i^k,a_i^k))=\tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{i=1}^{\bar{J}_k}\left(\sqrt{\frac{\widehat{c}_k(s_i^k,a_i^k)}{M_k^+(s_i^k,a_i^k)}}+\frac{1}{M_k^+(s_i^k,a_i^k)}\right)\right)$$

(Lemma 14)

$$=\tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{s,a}\sum_{h=1}^{H}q_k(s,a,h)\sqrt{\frac{\widehat{c}_k(s,a)}{M_k^+(s,a)}}+\sum_{k=1}^{K}\sum_{s,a}\frac{q_k(s,a)}{M_k^+(s,a)}+T_{\max}\right) \quad \text{(Lemma 52)}$$

$$=\tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{s,a}\sum_{h=1}^{H}\frac{q_k^2(s,a,h)}{x_k(s,a,h)}\widehat{c}_k(s,a)}\sqrt{\sum_{k=1}^{K}\sum_{s,a}\sum_{h=1}^{H}\frac{x_k(s,a,h)}{M_k^+(s,a)}}+SAT_{\max}\right)$$

(Cauchy-Schwarz inequality, Lemma 32, and $q_k(s,a)=\tilde{\mathcal{O}}\left(T_{\max}x_k(s,a)\right)$)

$$=\tilde{\mathcal{O}}\left(\sqrt{SA\sum_{k=1}^{K}\sum_{s,a}\sum_{h=1}^{H}q_k(s,a,h)Q_k(s,a,h)}+SAT_{\max}\right).$$

(Lemma 32 and $\frac{q_k(s,a,h)}{x_k(s,a,h)}\widehat{c}_k(s,a)\le Q_k(s,a,h)$)

■

Lemma 19 *For* $h\in[H+1]$, *we have* $\sum_{s,a}q^\star(s,a,h)\le(\frac{1}{2})^{h-1}T_{\max}$.

Proof Denote by $p(s)$ the probability that the learner starts at state s in layer h and eventually reaches layer $h+1$ following $\mathring{\pi}^\star$. Clearly, $p(g)=0$, and

$$p(s)\le 1-\gamma+\gamma P_{s,\pi^\star(s)}p\overset{\text{(i)}}{\le}\mathbb{E}\left[\sum_{t=1}^{I}(1-\gamma)\gamma^{t-1}\middle|\pi^\star,P,s_1=s\right]\le\frac{1}{2},$$

where (i) is by repeatedly applying the first inequality. By a recursive argument, we have the probability of reaching layer h is upper bounded by $(\frac{1}{2})^{h-1}$. Then by $\sum_{s,a}q^\star_{(s',h)}(s,a,h)\le T_{\max}$ for any s', we have $\sum_{s,a}q^\star(s,a,h)\le(\frac{1}{2})^{h-1}T_{\max}$. ■

Lemma 20 *Under stochastic costs,* $\left\langle q^\star, c \circ Q^{\mathring{\pi}^\star, P, c} \right\rangle \le 2B_\star^2 + \frac{(H+1)T_{\max}}{K}$.

Proof By Lemma 2 and Lemma 19, we have:

$$\begin{aligned}
\left\langle q^\star, c \circ Q^{\mathring{\pi}^\star, P, c} \right\rangle &= \sum_{h=1}^{H} \sum_{s,a} q^\star(s,a,h) c(s,a) Q^{\mathring{\pi}^\star, P, c}(s,a,h) + \sum_s q^\star(s, H+1) c_f \\
&\le \sum_{h=1}^{H} \sum_{s,a} q^\star(s,a,h) c(s,a) \left(Q^{\pi^\star, P, c}(s,a) + \frac{c_f}{2^{H-h+1}} \right) + \frac{c_f T_{\max}}{2^H} \\
&\le 2B_\star^2 + \sum_{h=1}^{H} \frac{T_{\max}}{2^{h-1}} \frac{c_f}{2^{H-h+1}} + \frac{c_f T_{\max}}{2^H} \\
&\qquad \left(\textstyle\sum_{h=1}^{H} q^\star(s,a,h) c(s,a) \le B_\star \text{ and } Q^{\pi^\star, P, c}(s,a) \le 1 + B_\star\right) \\
&\le 2B_\star^2 + (H+1) \frac{c_f T_{\max}}{2^H} \le 2B_\star^2 + \frac{(H+1)T_{\max}}{K}.
\end{aligned}$$

■

Lemma 21 *For stochastic adversary, we have* $\sum_{k=1}^{K} \left\langle q^\star, c \circ Q^{\mathring{\pi}^\star, P, c} \right\rangle = \tilde{\mathcal{O}}\left(D^2 K\right)$.

Proof $\sum_{k=1}^{K} \left\langle q^\star, c \circ Q^{\mathring{\pi}^\star, P, c} \right\rangle = \tilde{\mathcal{O}}\left(DK \left\langle q^\star, c \right\rangle\right) = \tilde{\mathcal{O}}\left(D^2 K\right)$. ■

Lemma 22 $\eta \left\| \widetilde{Q}_k \right\|_\infty \le 1$ *under all definitions of* $\widetilde{c}_k$.

Proof It suffices to bound $\left\| \widetilde{Q}_k \right\|_\infty$. By Lemma 2, $\widehat{Q}_k(s,a,h) \le \frac{H}{1-\gamma} + c_f = \chi$. Therefore, $e_k(s,a,h) \le 8\iota + \chi/T_{\max}$ under all feedback types. This gives $\widetilde{c}_k(s,a,h) \le (1 + \lambda \widehat{Q}_k(s,a,h)) + e_k(s,a,h) \le 3(8\iota + \chi/T_{\max})$ for $h \le H$ and $\widetilde{c}_k(s,a,H+1) \le (1 + \lambda \widehat{Q}_k(s,a,H+1)) c_f \le 3c_f \chi / T_{\max}$. Lemma 2 then gives $\widetilde{Q}_k(s,a,h) \le \frac{H}{1-\gamma} \cdot 3(8\iota + \chi/T_{\max}) + 3c_f\chi/T_{\max} \le 3T_{\max}(8\iota + \chi/T_{\max})^2$, and the statement is proved by the definition of η. ■

Lemma 23 *Under all definitions of* $\widetilde{c}_k$, *we have* $|d\pi_k(a|s,h)| = \tilde{\mathcal{O}}(\eta T_{\max} \pi_k(a|s,h))$ *and* $\left\| dQ^{\pi_k, P', c'} \right\|_\infty = \tilde{\mathcal{O}}(\eta T_{\max}^3)$ *for* $P' \in \Lambda_{\mathcal{M}}$ *and* $c' \in \mathcal{C}_{\mathcal{M}}$.

Proof Note that:

$$\begin{aligned}
&\pi_{k+1}(a|s,h) - \pi_k(a|s,h) = \frac{\pi_k(a|s,h) \exp(-\eta \widetilde{Q}_k(s,a,h))}{\sum_{a'} \pi_k(a'|s,h) \exp(-\eta \widetilde{Q}_k(s,a',h))} - \pi_k(a|s,h) \\
&\le \frac{\pi_k(a|s,h)}{\sum_{a'} \pi_k(a'|s,h)} \exp(\max_{a'} |\eta \widetilde{Q}_k(s,a',h)|) - \pi_k(a|s,h) = \tilde{\mathcal{O}}\left(\eta T_{\max} \pi_k(a|s,h)\right). \\
&\qquad \text{(Lemma 22 and } |e^x - 1| \le 2|x| \text{ for } x \in [-1,1])
\end{aligned}$$

The other direction can be proved similarly. Then by Lemma 30,

$$\begin{aligned}
&\left|Q^{\pi_{k+1},P',c'}(s,a,h) - Q^{\pi_k,P',c'}(s,a,h)\right| \\
&= \left|\sum_{s'',h''} P_{s,a,h}(s'',h'') \sum_{s',a',h'} q_{\pi_k,P',(s'',h'')}(s',h') \left(d\pi_k(a'|s',h')\right) Q^{\pi_{k+1},P',c'}(s',a',h')\right| \\
&= \tilde{\mathcal{O}}\left(\eta T_{\max}^3\right).
\end{aligned}$$

This completes the proof. ■

Lemma 24 *Suppose* $\pi_k(a|s,h) \propto \exp(\sum_{j<k} \widetilde{Q}_j(s,a,h))$. *Then,*

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{a\in\mathcal{A}} (\pi_k(a|s,h) - \pi^\star(a|s,h))\widetilde{Q}_k(s,a,h) \\
&\leq \frac{\ln A}{\eta} + \left\langle \pi_1(\cdot|s,h), \widetilde{Q}_1(s,\cdot,h)\right\rangle + \sum_{k=1}^{K-1} \left\langle \pi_{k+1}(\cdot|s,h), \widetilde{Q}_{k+1}(s,\cdot,h) - \widetilde{Q}_k(s,\cdot,h)\right\rangle.
\end{aligned}$$

Proof First note that:

$$\pi_{k+1}(\cdot|s,h) = \operatorname*{argmin}_{\pi(\cdot|s,h)\in\Delta(A)} \eta \left\langle \pi(\cdot|s,h), \widetilde{Q}_k(s,\cdot,h)\right\rangle + \mathrm{KL}(\pi(\cdot|s,h), \pi_k(\cdot|s,h)), \tag{12}$$

where $\mathrm{KL}(p,q) = \sum_a (p(a)\ln\frac{p(a)}{q(a)} - p(a) + q(a))$, and

$$\pi_{k+1}(\cdot|s,h) \propto \pi'_{k+1}(a|s,h) \triangleq \pi_k(a|s,h)\exp(-\eta\widetilde{Q}_k(s,a,h)),$$

where π'_{k+1} is the solution of the unconstrained variant of Eq. (12) (that is, replacing $\operatorname{argmin}_{\pi(\cdot|s,h)\in\Delta(A)}$ by $\operatorname{argmin}_{\pi(\cdot|s,h)\in\mathbb{R}^A}$). It is easy to verify that:

$$\begin{aligned}
&\mathrm{KL}(\pi_k(\cdot|s,h), \pi_{k+1}(\cdot|s,h)) + \mathrm{KL}(\pi_{k+1}(\cdot|s,h), \pi_k(\cdot|s,h)) \\
&= \left\langle \pi_k(\cdot|s,h), \ln\frac{\pi_k(\cdot|s,h)}{\pi_{k+1}(\cdot|s,h)}\right\rangle + \left\langle \pi_{k+1}(\cdot|s,h), \ln\frac{\pi_{k+1}(\cdot|s,h)}{\pi_k(\cdot|s,h)}\right\rangle \\
&= \left\langle \pi_k(\cdot|s,h) - \pi_{k+1}(\cdot|s,h), \ln\frac{\pi_k(\cdot|s,h)}{\pi'_{k+1}(\cdot|s,h)}\right\rangle \qquad (\pi_{k+1}(\cdot|s,h) \propto \pi'_{k+1}(\cdot|s,h)) \\
&= \left\langle \pi_k(\cdot|s,h) - \pi_{k+1}(\cdot|s,h), \eta\widetilde{Q}_k(s,\cdot,h)\right\rangle \geq 0.
\end{aligned} \tag{13}$$

By the standard OMD analysis (Hazan et al., 2016) (note that KL is the Bregman divergence w.r.t the negative entropy regularizer),

$$\sum_{k=1}^{K} \left\langle \pi_k(\cdot|s,h) - \pi^\star(\cdot|s,h), \widetilde{Q}_k(s,\cdot,h)\right\rangle$$

$$
\begin{aligned}
&= \frac{1}{\eta}\sum_{k=1}^{K}\left(\mathrm{KL}(\pi^\star(\cdot|s,h),\pi_k(\cdot|s,h)) - \mathrm{KL}(\pi^\star(\cdot|s,h),\pi'_{k+1}(\cdot|s,h)) + \mathrm{KL}(\pi_k(\cdot|s,h),\pi'_{k+1}(\cdot|s,h))\right)\\
&= \frac{1}{\eta}\sum_{k=1}^{K}\left(\mathrm{KL}(\pi^\star(\cdot|s,h),\pi_k(\cdot|s,h)) - \mathrm{KL}(\pi^\star(\cdot|s,h),\pi_{k+1}(\cdot|s,h)) + \mathrm{KL}(\pi_k(\cdot|s,h),\pi_{k+1}(\cdot|s,h))\right)\\
&\le \frac{\mathrm{KL}(\pi^\star(\cdot|s,h),\pi_1(\cdot|s,h))}{\eta} + \sum_{k=1}^{K}\left\langle \pi_k(\cdot|s,h) - \pi_{k+1}(\cdot|s,h), \widetilde{Q}_k(s,\cdot,h)\right\rangle \qquad \text{(Eq. (13))}\\
&\le \frac{\ln A}{\eta} + \sum_{k=1}^{K-1}\left\langle \pi_{k+1}(\cdot|s,h), \widetilde{Q}_{k+1}(s,\cdot,h) - \widetilde{Q}_k(s,\cdot,h)\right\rangle\\
&\quad + \left\langle \pi_1(\cdot|s,h), \widetilde{Q}_1(s,\cdot,h)\right\rangle - \left\langle \pi_{K+1}(\cdot|s,h), \widetilde{Q}_K(s,\cdot,h)\right\rangle.
\end{aligned}
$$

This completes the proof. ■

Lemma 25 *Define* $\mathfrak{n}_k(s,a) = \mathfrak{N}_{k+1}(s,a) - \mathfrak{N}_k(s,a)$. *We have:*

$$
\begin{aligned}
&|d\widehat{c}_k(s,a)| = \mathcal{O}\left(\frac{\mathfrak{n}_k(s,a)\iota}{\mathfrak{N}_k^+(s,a)}\right),\\
&\left|d\widehat{Q}_k(s,a,h)\right| = \mathcal{O}\left(T_{\max}^2\sum_{s',a'}\frac{Sn_k(s',a')\iota}{N_k^+(s',a')} + T_{\max}\sum_{s',a'}\frac{\mathfrak{n}_k(s',a')\iota}{\mathfrak{N}_k^+(s',a')} + \eta T_{\max}^3\right),\\
&|d\widetilde{c}_k(s,a)|\\
&= \mathcal{O}\left(\frac{\mathfrak{n}_k(s,a)\iota}{\mathfrak{N}_k^+(s,a)} + \lambda T_{\max}^2\sum_{s',a'}\frac{Sn_k(s',a')\iota}{N_k^+(s',a')} + \lambda T_{\max}\sum_{s',a'}\frac{\mathfrak{n}_k(s',a')}{\mathfrak{N}_k^+(s',a')} + \lambda\eta T_{\max}^3 + |de_k(s,a,h)|\right),\\
&d\widetilde{Q}_k(s,a,h) = \mathcal{O}\left(T_{\max}^2\sum_{s',a'}\frac{Sn_k(s',a')\iota}{N_k^+(s',a')} + \lambda T_{\max}^2\sum_{s',a'}\frac{\mathfrak{n}_k(s',a')}{\mathfrak{N}_k^+(s',a')} + \lambda\eta T_{\max}^4 + T_{\max}\left\|de_k\right\|_1\right).
\end{aligned}
$$

Proof **First statement:** Note that for all definitions of $\widehat{c}_k$ used in this paper, we have $\|\widehat{c}_k\|_\infty \le 1$. Then by the definition of $\widehat{c}_k$ and $|\max\{0,a\} - \max\{0,b\}| \le |a-b|$:

$$
\begin{aligned}
&|\widehat{c}_{k+1}(s,a) - \widehat{c}_k(s,a)|\\
&= \mathcal{O}\left(|\bar{c}_{k+1}(s,a) - \bar{c}_k(s,a)| + \left|\sqrt{\frac{\bar{c}_k(s,a)\iota}{\mathfrak{N}_k^+(s,a)}} - \sqrt{\frac{\bar{c}_{k+1}(s,a)\iota}{\mathfrak{N}_{k+1}^+(s,a)}}\right| + \frac{\iota}{\mathfrak{N}_k^+(s,a)} - \frac{\iota}{\mathfrak{N}_{k+1}^+(s,a)}\right).
\end{aligned}
$$

Note that:

$$
\begin{aligned}
|\bar{c}_{k+1}(s,a) - \bar{c}_k(s,a)| &= \left|\frac{C_{k+1}(s,a)}{\mathfrak{N}_{k+1}^+(s,a)} - \frac{C_k(s,a)}{\mathfrak{N}_k^+(s,a)}\right|\\
&\le \left|\frac{C_{k+1}(s,a) - C_k(s,a)}{\mathfrak{N}_{k+1}^+(s,a)}\right| + \mathfrak{N}_k(s,a)\left|\frac{1}{\mathfrak{N}_k^+(s,a)} - \frac{1}{\mathfrak{N}_{k+1}^+(s,a)}\right| \qquad (C_k(s,a) \le \mathfrak{N}_k(s,a))
\end{aligned}
$$

$$\leq \frac{\mathfrak{n}_k(s,a)}{\mathfrak{N}_{k+1}^+(s,a)} + \frac{\mathfrak{N}_k(s,a)\mathfrak{n}_k(s,a)}{\mathfrak{N}_k^+(s,a)\mathfrak{N}_{k+1}^+(s,a)} \leq \frac{2\mathfrak{n}_k(s,a)}{\mathfrak{N}_k^+(s,a)},$$

and by $|\sqrt{a}-\sqrt{b}| \leq \sqrt{|a-b|}$, $\mathfrak{n}_k(s,a) \in \mathbb{N}$:

$$\begin{aligned}
&\left|\sqrt{\frac{\bar{c}_k(s,a)\iota}{\mathfrak{N}_k^+(s,a)}} - \sqrt{\frac{\bar{c}_{k+1}(s,a)\iota}{\mathfrak{N}_{k+1}^+(s,a)}}\right| \\
&\leq \sqrt{\frac{|\bar{c}_k(s,a)-\bar{c}_{k+1}(s,a)|\iota}{\mathfrak{N}_k^+(s,a)}} + \sqrt{\bar{c}_{k+1}(s,a)\iota}\left(\frac{1}{\sqrt{\mathfrak{N}_k^+(s,a)}} - \frac{1}{\sqrt{\mathfrak{N}_{k+1}^+(s,a)}}\right) \\
&\leq \frac{2\mathfrak{n}_k(s,a)\iota}{\mathfrak{N}_k^+(s,a)} + \left(\sqrt{\frac{\iota}{\mathfrak{N}_k^+(s,a)}} - \sqrt{\frac{\iota}{\mathfrak{N}_{k+1}^+(s,a)}}\right) = \mathcal{O}\left(\frac{\mathfrak{n}_k(s,a)\iota}{\mathfrak{N}_k^+(s,a)}\right),
\end{aligned}$$

where in the last inequality we apply

$$\begin{aligned}
\frac{1}{\sqrt{\mathfrak{N}_k^+(s,a)}} - \frac{1}{\sqrt{\mathfrak{N}_{k+1}^+(s,a)}} &= \left(\frac{1}{\mathfrak{N}_k^+(s,a)} - \frac{1}{\mathfrak{N}_{k+1}^+(s,a)}\right) \Big/ \left(\frac{1}{\sqrt{\mathfrak{N}_k^+(s,a)}} + \frac{1}{\sqrt{\mathfrak{N}_{k+1}^+(s,a)}}\right) \\
&\leq \sqrt{\mathfrak{N}_{k+1}^+(s,a)} \cdot \frac{\mathfrak{n}_k(s,a)}{\mathfrak{N}_k^+(s,a)\mathfrak{N}_{k+1}^+(s,a)} \leq \frac{\mathfrak{n}_k(s,a)}{\mathfrak{N}_k^+(s,a)}. \qquad (14)
\end{aligned}$$

Thus, $|d\widehat{c}_k(s,a)| = \mathcal{O}\left(\frac{\mathfrak{n}_k(s,a)\iota}{\mathfrak{N}_k^+(s,a)}\right)$.

Second statement: Define $\Pi_k(P') = \operatorname{argmin}_{P'' \in \mathcal{P}_{k+1}} \sum_{s,a,h} \left\| P''_{s,a,h} - P'_{s,a,h} \right\|_1$ for any $P' \in \mathcal{P}_k$. By the definition of $\mathcal{P}_k$, we have (note that $P'_{s,a,h}(s',h') = 0$ for $h' \notin \{h, h+1\}$):

$$\left\|\Pi_k(P')_{s,a,h} - P'_{s,a,h}\right\|_1 \leq 2\sum_{s'} \left|\bar{P}_{k,s,a}(s') - \bar{P}_{k+1,s,a}(s')\right| + 2\sum_{s'} \left|\epsilon_{k+1}(s,a,s') - \epsilon_k(s,a,s')\right|.$$

Denote by $n_k(s,a,s')$ the number of visits to (s,a,s') (before policy switch or goal state is reached) in episode k. Note that:

$$\begin{aligned}
\left|\bar{P}_{k,s,a}(s') - \bar{P}_{k+1,s,a}(s')\right| &= \left|\frac{N_k(s,a,s') + n_k(s,a,s')}{N_{k+1}^+(s,a)} - \frac{N_k(s,a,s')}{N_k^+(s,a)}\right| \\
&\leq N_k(s,a,s')\left(\frac{1}{N_k^+(s,a)} - \frac{1}{N_{k+1}^+(s,a)}\right) + \frac{n_k(s,a,s')}{N_{k+1}^+(s,a)} \leq \frac{2n_k(s,a)}{N_k^+(s,a)}.
\end{aligned}$$

and by $|\sqrt{a}-\sqrt{b}| \leq \sqrt{|a-b|}$,

$$\begin{aligned}
&\left|\epsilon_k(s,a,s') - \epsilon_{k+1}(s,a,s')\right| = \mathcal{O}\left(\left|\sqrt{\frac{\bar{P}_{k,s,a}(s')\iota}{N_k^+(s,a)}} - \sqrt{\frac{\bar{P}_{k+1,s,a}(s')\iota}{N_{k+1}^+(s,a)}}\right| + d\left(\frac{-\iota}{N_k^+(s,a)}\right)\right) \\
&= \mathcal{O}\left(\sqrt{\frac{\left|\bar{P}_{k,s,a}(s') - \bar{P}_{k+1,s,a}(s')\right|\iota}{N_k^+(s,a)}} + \sqrt{\bar{P}_{k+1,s,a}(s')\iota}d\left(\frac{-1}{\sqrt{N_k^+(s,a)}}\right) + d\left(\frac{-\iota}{N_k^+(s,a)}\right)\right)
\end{aligned}$$

$$= \mathcal{O}\left(\frac{n_k(s,a)\iota}{N_k^+(s,a)} + \sqrt{\bar{P}_{k+1,s,a}(s')}d\left(\frac{-\sqrt{\iota}}{\sqrt{N_k^+(s,a)}}\right)\right).$$

Plugging these back, and by Cauchy-Schwarz inequality and Eq. (14) with $\mathfrak{N}_k = N_k$, we have

$$\left\|\Pi_k(P')_{s,a,h} - P'_{s,a,h}\right\|_1 = \mathcal{O}\left(\frac{Sn_k(s,a)\iota}{N_k^+(s,a)} + d\left(\frac{-\sqrt{S\iota}}{\sqrt{N_k^+(s,a)}}\right)\right) = \mathcal{O}\left(\frac{Sn_k(s,a)\iota}{N_k^+(s,a)}\right). \quad (15)$$

Thus, for any policy π' and cost function $c' \in \mathcal{C}_{\mathcal{M}}$ with $c'(s,a,h) \in [0,1]$ for $h \le H$, by Lemma 30 and Eq. (15),

$$\begin{aligned}
&\left|Q^{\pi',\Pi_k(P'),c'}(s,a,h) - Q^{\pi',P',c'}(s,a,h)\right| \\
&= \left|\sum_{s',a',h'} q_{\pi',P',(s,a,h)}(s',a',h')(\Pi_k(P')_{s',a',h'} - P'_{s',a',h'})V^{\pi',\Pi_k(P'),c'}\right| \\
&= \mathcal{O}\left(T_{\max}^2 \sum_{s',a'} \frac{Sn_k(s',a')\iota}{N_k^+(s',a')}\right). \qquad (16)
\end{aligned}$$

Now define $P'_k = \Pi_k(P_k)$. We have

$$\begin{aligned}
&\widehat{Q}_{k+1}(s,a,h) - \widehat{Q}_k(s,a,h) = Q^{\pi_{k+1},P_{k+1},\widehat{c}_{k+1}}(s,a,h) - Q^{\pi_k,P_k,\widehat{c}_k}(s,a,h) \\
&\le Q^{\pi_{k+1},P'_k,\widehat{c}_{k+1}}(s,a,h) - Q^{\pi_{k+1},P_k,\widehat{c}_{k+1}}(s,a,h) + Q^{\pi_{k+1},P_k,\widehat{c}_{k+1}}(s,a,h) - Q^{\pi_k,P_k,\widehat{c}_k}(s,a,h) \\
&= \mathcal{O}\left(T_{\max}^2 \sum_{s',a'} \frac{Sn_k(s',a')\iota}{N_k^+(s',a')}\right) + (Q^{\pi_{k+1},P_k,\widehat{c}_{k+1}}(s,a,h) - Q^{\pi_{k+1},P_k,\widehat{c}_k}(s,a,h)) \qquad \text{(Eq. (16))} \\
&\quad + (Q^{\pi_{k+1},P_k,\widehat{c}_k}(s,a,h) - Q^{\pi_k,P_k,\widehat{c}_k}(s,a,h)) \\
&= \mathcal{O}\left(T_{\max}^2 \sum_{s',a'} \frac{Sn_k(s',a')\iota}{N_k^+(s',a')} + T_{\max} \sum_{s',a'} \left|\widehat{c}_{k+1}(s',a') - \widehat{c}_k(s',a')\right| + \eta T_{\max}^3\right) \\
&\qquad \text{(Lemma 30 and Lemma 23)} \\
&= \mathcal{O}\left(T_{\max}^2 \sum_{s',a'} \frac{Sn_k(s',a')\iota}{N_k^+(s',a')} + T_{\max} \sum_{s',a'} \frac{\mathfrak{n}_k(s',a')\iota}{\mathfrak{N}_k^+(s',a')} + \eta T_{\max}^3\right).
\end{aligned}$$

The other direction can be proved similarly.

Third statement: Note that $|d\widetilde{c}_k(s,a,H+1)| = 0$, and for $h \le H$,

$$\begin{aligned}
&|\widetilde{c}_{k+1}(s,a,h) - \widetilde{c}_k(s,a,h)| \\
&\le |d\widehat{c}_k(s,a)| + \lambda\left|\widehat{c}_{k+1}(s,a)\widehat{Q}_{k+1}(s,a,h) - \widehat{Q}_k(s,a,h)\widehat{c}_k(s,a)\right| + |de_k(s,a,h)| \\
&\le |d\widehat{c}_k(s,a)| + \lambda\widehat{Q}_{k+1}(s,a,h)\,|d\widehat{c}_k(s,a)| + \lambda\widehat{c}_k(s,a)\left|d\widehat{Q}_k(s,a,h)\right| + |de_k(s,a,h)|
\end{aligned}$$

$$= \mathcal{O}\left(\frac{\mathfrak{n}_k(s,a)\iota}{\mathfrak{N}_k^+(s,a)} + \lambda T_{\max}^2 \sum_{s',a'} \frac{Sn_k(s',a')\iota}{N_k^+(s',a')} + \lambda T_{\max} \sum_{s',a'} \frac{\mathfrak{n}_k(s',a')}{\mathfrak{N}_k^+(s',a')} + \lambda \eta T_{\max}^3 + |de_k(s,a,h)|\right).$$

Fourth statement: Define $\widetilde{P}'_k = \Pi_k(\widetilde{P}_k)$. By Eq. (15), $\left\|\widetilde{P}'_{k,s,a,h} - \widetilde{P}_{k,s,a,h}\right\|_1 = \mathcal{O}\left(\frac{Sn_k(s,a)\iota}{N_k^+(s,a)}\right)$, and $\lambda \le 1/T_{\max}$, we have

$$\begin{aligned}
&\widetilde{Q}_{k+1}(s,a,h) - \widetilde{Q}_k(s,a,h) \le Q^{\pi_{k+1},\widetilde{P}'_k,\widetilde{c}_{k+1}}(s,a,h) - Q^{\pi_k,\widetilde{P}_k,\widetilde{c}_k}(s,a,h) \\
&= \left(Q^{\pi_{k+1},\widetilde{P}'_k,\widetilde{c}_{k+1}}(s,a,h) - Q^{\pi_{k+1},\widetilde{P}_k,\widetilde{c}_{k+1}}(s,a,h)\right) \\
&\quad + \left(Q^{\pi_{k+1},\widetilde{P}_k,\widetilde{c}_{k+1}}(s,a,h) - Q^{\pi_{k+1},\widetilde{P}_k,\widetilde{c}_k}(s,a,h)\right) + \left(Q^{\pi_{k+1},\widetilde{P}_k,\widetilde{c}_k}(s,a,h) - Q^{\pi_k,\widetilde{P}_k,\widetilde{c}_k}(s,a,h)\right) \\
&\overset{\text{(i)}}{\le} \mathcal{O}\left(T_{\max}^2 \sum_{s',a'} \frac{Sn_k(s',a')\iota}{N_k^+(s',a')}\right) + \sum_{s',a',h'} q_{\pi_{k+1},\widetilde{P}_k,(s,a,h)}(s',a',h') \left|\widetilde{c}_{k+1}(s',a',h') - \widetilde{c}_k(s',a',h')\right| \\
&= \mathcal{O}\left(T_{\max}^2 \sum_{s',a'} \frac{Sn_k(s',a')\iota}{N_k^+(s',a')} + \lambda T_{\max}^2 \sum_{s',a'} \frac{\mathfrak{n}_k(s',a')}{\mathfrak{N}_k^+(s',a')} + \lambda \eta T_{\max}^4 + T_{\max} \|de_k\|_1\right),
\end{aligned}$$

where in (i) we apply Eq. (16), Lemma 30, and

$$\begin{aligned}
&Q^{\pi_{k+1},\widetilde{P}_k,\widetilde{c}_k}(s,a,h) - Q^{\pi_k,\widetilde{P}_k,\widetilde{c}_k}(s,a,h) \\
&= \sum_{s'',h''} \widetilde{P}_{k,s,a,h}(s'',h'') \sum_{s',a',h'} q_{\pi_{k+1},\widetilde{P}_k,(s'',h'')}(s',h') \left(d\pi_k(a'|s',h')\right) Q^{\pi_k,\widetilde{P}_k,\widetilde{c}_k}(s',a',h') \\
&\qquad\qquad \text{(Lemma 30)} \\
&\le 0. \qquad\qquad \text{(Eq. (13))}
\end{aligned}$$

This completes the proof. ■

Lemma 26 *For any cost function c in $\mathring{\mathcal{M}}$ such that $c((s,h),a) \ge 0$, we have:*

$$\begin{aligned}
\mathrm{Var}_k[\langle n_k, c\rangle] &= \sum_{s,a,h} q_k(s,a,h)(A^{\pi_k,P,c}(s,a,h)^2 + \mathbb{V}(P_{s,a,h}, V^{\pi_k,P,c})) \\
&\le \mathbb{E}_k[\langle n_k, c\rangle^2] \le 2\left\langle q_k, c \circ Q^{\pi_k,P,c}\right\rangle.
\end{aligned}$$

Proof Let $Q = Q^{\pi_k,P,c}$, $V = V^{\pi_k,P,c}$, $A = A^{\pi_k,P,c}$ and define $c(g,a) = 0$. Then,

$$\begin{aligned}
\mathrm{Var}_k[\langle n_k, c\rangle] &= \mathbb{E}_k\left[\left(\sum_{i=1}^{J_k+1} c(\mathring{s}_i^k, a_i^k) - V(\mathring{s}_1^k)\right)^2\right] \\
&= \mathbb{E}_k\left[\left(\sum_{i=2}^{J_k+1} c(\mathring{s}_i^k, a_i^k) + Q(\mathring{s}_1^k, a_1^k) - P_{\mathring{s}_1^k, a_1^k} V - V(\mathring{s}_1^k)\right)^2\right] \qquad (Q(\mathring{s},a) = c(\mathring{s},a) + P_{\mathring{s},a}V)
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{=} \mathbb{E}_k\left[\left(Q(\mathring{s}_1^k, a_1^k) - V(\mathring{s}_1^k)\right)^2\right] + \mathbb{E}_k\left[\left(\sum_{i=2}^{J_k+1} c(\mathring{s}_i^k, a_i^k) - P_{\mathring{s}_1^k, a_1^k} V\right)^2\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_k\left[\left(Q(\mathring{s}_1^k, a_1^k) - V(\mathring{s}_1^k)\right)^2\right] + \mathbb{E}_k\left[\left(\sum_{i=2}^{J_k+1} c(\mathring{s}_i^k, a_i^k) - V(\mathring{s}_2^k)\right)^2\right] + \mathbb{E}_k\left[\left(V(\mathring{s}_2^k) - P_{\mathring{s}_1^k, a_1^k} V\right)^2\right] \\
&= \mathbb{E}_k\left[\sum_{i=1}^{J_k+1}\left[\left(Q(\mathring{s}_i^k, a_i^k) - V(\mathring{s}_i^k)\right)^2 + \left(V(\mathring{s}_{i+1}^k) - P_{\mathring{s}_i^k, a_i^k} V\right)^2\right]\right] \qquad \text{(recursive argument)} \\
&= \sum_{s,a,h} q_k(s,a,h)\left(A^2(s,a,h) + \mathbb{V}(P_{s,a,h}, V)\right),
\end{aligned}
$$

where (i) is by $Q(\mathring{s}_1^k, a_1) - V(\mathring{s}_1^k) \in \sigma(\mathring{s}_1^k, a_1^k)$ (the σ-algebra of events defined on $(\mathring{s}_1^k, a_1^k)$) and

$$
\mathbb{E}_k\left[\sum_{i=2}^{J_k+1} c(\mathring{s}_i^k, a_i^k) - P_{\mathring{s}_1^k, a_1^k} V \,\middle|\, \mathring{s}_1^k, a_1^k\right] = 0;
$$

(ii) is by $V(\mathring{s}_2^k) - P_{\mathring{s}_1^k, a_1^k} V \in \sigma(\mathring{s}_1^k, a_1^k, \mathring{s}_2^k)$ and

$$
\mathbb{E}_k\left[\sum_{i=2}^{J_k+1} c(\mathring{s}_i^k, a_i^k) - V(\mathring{s}_2^k) \,\middle|\, \mathring{s}_1^k, a_1^k, \mathring{s}_2^k\right] = 0.
$$

Moreover, by $(\sum_{i=1}^n a_i)^2 \le 2a_i(\sum_{i'=i}^n a_{i'})$ for any $n \ge 1$ and $P(J_k = \infty) = 0$,

$$
\begin{aligned}
\mathrm{Var}_k[\langle n_k, c\rangle] \le \mathbb{E}_k[\langle n_k, c\rangle^2] &= \mathbb{E}_k\left[\left(\sum_{i=1}^{J_k+1} c(\mathring{s}_i^k, a_i^k)\right)^2\right] \le 2\mathbb{E}_k\left[\sum_{i=1}^{J_k+1} c(\mathring{s}_i^k, a_i^k) \sum_{i'=i}^{J_k+1} c(\mathring{s}_{i'}^k, a_{i'}^k)\right] \\
&= 2\mathbb{E}_k\left[\sum_{i=1}^{\infty} \mathbb{I}\{J_k + 1 \ge i\} c(\mathring{s}_i^k, a_i^k) \sum_{i'=i}^{J_k+1} c(\mathring{s}_{i'}^k, a_{i'}^k)\right] \\
&\overset{\text{(i)}}{=} 2\mathbb{E}_k\left[\sum_{i=1}^{J_k+1} c(\mathring{s}_i^k, a_i^k) Q(\mathring{s}_i^k, a_i^k)\right] = 2\langle q_k, c \circ Q\rangle,
\end{aligned}
$$

where in (i) we apply $Q(\mathring{s}_i^k, a_i^k) = \mathbb{E}[\sum_{i'=i}^{J_k+1} c(\mathring{s}_{i'}^k, a_{i'}^k) | \mathring{s}_1^k, a_1^k, \ldots, \mathring{s}_i^k, a_i^k]$ and $\{J_k + 1 \ge i\} \in \sigma(\mathring{s}_1^k, a_1^k, \ldots, \mathring{s}_i^k, a_i^k)$. ∎

Lemma 27 *For every $k \in [K]$ it holds that $q_k(s,a,h) \le \mathbb{E}_k[\bar{n}_k(s,a,h)] + \tilde{\mathcal{O}}(1/K)$.*

Proof By definition of $n_k(s,a,h)$, $x_k(s,a,h)$, and $y_k(s,a,h)$ we have:

$$
\begin{aligned}
\Pr(n_k(s,a,h) > n) &= \Pr(n_k(s,a,h) > n \mid n_k(s,a,h) > n-1) \Pr(n_k(s,a,h) > n-1) \\
&= \Pr(\text{return to } (s,a,h)) \Pr(n_k(s,a,h) > n-1) \\
&= y_k(s,a,h) \Pr(n_k(s,a,h) > n-1)
\end{aligned}
$$

$$= \cdots = y_k^n(s,a,h)\Pr\left(n_k(s,a,h) > 0\right) = y_k^n(s,a,h)x_k(s,a,h).$$

Now, since $q_k(s,a,h)$ is the expected number of visits to (s,a,h),

$$q_k(s,a,h) = \mathbb{E}_k[n_k(s,a,h)] = \sum_{n=0}^{\infty}\Pr\left(n_k(s,a,h) > n\right) = x_k(s,a,h)\sum_{n=0}^{\infty} y_k^n(s,a,h)$$
$$= x_k(s,a,h)\sum_{n=0}^{L-1} y_k^n(s,a,h) + x_k(s,a,h)\sum_{n=L}^{\infty} y_k^n(s,a,h).$$

To finish we bound each of the sums separately. By definition of $\bar{n}_k(s,a,h)$:

$$x_k(s,a,h)\sum_{n=0}^{L-1} y_k^n(s,a,h) = \sum_{n=0}^{L-1}\Pr\left(n_k(s,a,h) > n\right) = \sum_{n=0}^{L-1}\Pr\left(\min\{L, n_k(s,a,h)\} > n\right)$$
$$\leq \sum_{n=0}^{\infty}\Pr\left(\min\{L, n_k(s,a,h)\} > n\right)$$
$$= \sum_{n=0}^{\infty}\Pr\left(\bar{n}_k(s,a,h) > n\right) = \mathbb{E}_k[\bar{n}_k(s,a,h)].$$

In each step there's a probability of at most γ to stay in layer h. So $y_k(s,a,h) \leq \gamma$, which implies:

$$x_k(s,a,h)\sum_{n=L}^{\infty} y_k^n(s,a,h) \leq \sum_{n=L}^{\infty}\gamma^n = \frac{\gamma^L}{1-\gamma} \leq \frac{\gamma^{\frac{8H}{1-\gamma}\ln(2T_{\max}K/\delta)}}{1-\gamma} \leq \frac{e^{-8H\ln(2T_{\max}K/\delta)}}{1-\gamma}$$
$$\leq 2T_{\max}\left(\frac{\delta}{2T_{\max}K}\right)^{8\log_2(c_f K)} = \tilde{\mathcal{O}}\left(1/K\right),$$

where the second inequality uses $\gamma^{\frac{1}{1-\gamma}} \leq e^{-1}$. ■

Lemma 28 *Consider a sequence of cost functions $\{c_k\}_{k=1}^K$ and transition functions $\{P_k\}_{k=1}^K$ such that $c_k \in \mathcal{C}_{\mathcal{M}}$ and $P_k \in \mathcal{P}_k$. Also define $\widehat{q}_k = q_{\pi_k,P_k}$. Then with probability at least $1-8\delta$,*

$$\sum_{k=1}^{K}|\langle q_k - \widehat{q}_k, c_k\rangle| = \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^{K}\langle q_k, c_k \circ Q^{\pi_k,P,c_k}\rangle} + S^{2.5}A^{1.5}T_{\max}^3\right).$$

Proof Define $v_{k,s,a,h}(\mathring{s}') = V^{\pi_k,P,c_k}(\mathring{s}') - P_{s,a,h}V^{\pi_k,P,c_k}$ for $\mathring{s}' \in \mathring{\mathcal{S}}_+$. Note that with probability at least $1-4\delta$:

$$\sum_{k=1}^{K}|\langle q_k - \widehat{q}_k, c_k\rangle| = \sum_{k=1}^{K}\left|\sum_{s,a,h} q_k(s,a,h)(P_{s,a,h} - P_{k,s,a,h})V^{\pi_k,P_k,c_k}\right| \qquad \text{(Lemma 30)}$$
$$= \sum_{k=1}^{K}\left|\sum_{s,a,h} q_k(s,a,h)(P_{s,a,h} - P_{k,s,a,h})V^{\pi_k,P,c_k}\right| + \tilde{\mathcal{O}}\left(S^{2.5}A^{1.5}T_{\max}^3\right).$$

(Lemma 13 and Lemma 29)

Below we bound the first term. We continue with:

$$= \sum_{k=1}^{K} \left| \sum_{s,a} \sum_{h=1}^{H} q_k(s,a,h)(P_{s,a,h} - P_{k,s,a,h}) v_{k,s,a,h} \right| \qquad (P_{s,a,H+1} = P_{k,s,a,H+1})$$

$$= \tilde{\mathcal{O}} \left(\sum_{k=1}^{K} \sum_{s,a,h \le H, \mathring{s}'} q_k(s,a,h) \sqrt{\frac{P_{s,a,h}(\mathring{s}') v_{k,s,a,h}^2(\mathring{s}')}{N_k^+(s,a)}} + S T_{\max} \sum_{k=1}^{K} \sum_{s,a} \frac{q_k(s,a)}{N_k^+(s,a)} \right). \qquad \text{(Lemma 13)}$$

By Lemma 27, we have $q_k(s,a,h) \le \mathbb{E}_k[\bar{n}_k(s,a,h)] + \tilde{\mathcal{O}}(1/K)$. Therefore, we continue with

$$= \tilde{\mathcal{O}} \left(\sum_{k=1}^{K} \sum_{s,a,h \le H, \mathring{s}'} \mathbb{E}_k[\bar{n}_k(s,a,h)] \sqrt{\frac{P_{s,a,h}(\mathring{s}') v_{k,s,a,h}^2(\mathring{s}')}{N_k^+(s,a)}} + S^2 A T_{\max} \right) \qquad \text{(Lemma 32)}$$

$$= \tilde{\mathcal{O}} \left(\sum_{k=1}^{K} \sum_{s,a,h \le H, \mathring{s}'} \bar{n}_k(s,a,h) \sqrt{\frac{P_{s,a,h}(\mathring{s}') v_{k,s,a,h}^2(\mathring{s}')}{N_k^+(s,a)}} + S^2 A T_{\max}^2 \right) \qquad \text{(Lemma 52)}$$

$$= \tilde{\mathcal{O}} \left(\sum_{k=1}^{K} \sum_{s,a,h \le H, \mathring{s}'} \bar{n}_k(s,a,h) \sqrt{\frac{P_{s,a,h}(\mathring{s}') v_{k,s,a,h}^2(\mathring{s}')}{N_{k+1}^+(s,a)}} \right)$$

$$+ \tilde{\mathcal{O}} \left(S T_{\max}^2 \sum_{s,a} \sum_{k=1}^{K} \left(\frac{1}{\sqrt{N_k^+(s,a)}} - \frac{1}{\sqrt{N_{k+1}^+(s,a)}} \right) + S^2 A T_{\max}^2 \right)$$

$$= \tilde{\mathcal{O}} \left(\sqrt{\sum_{k=1}^{K} \sum_{s,a,\mathring{s}'} \frac{\bar{n}_k(s,a)}{N_{k+1}^+(s,a)}} \sqrt{\sum_{k=1}^{K} \sum_{s,a,h \le H, \mathring{s}'} \bar{n}_k(s,a,h) P_{s,a,h}(\mathring{s}') v_{k,s,a,h}^2(\mathring{s}')} + S^2 A T_{\max}^2 \right)$$

(Cauchy-Schwarz inequality)

$$= \tilde{\mathcal{O}} \left(\sqrt{S^2 A} \sqrt{\sum_{k=1}^{K} \sum_{s,a,h \le H, \mathring{s}'} q_k(s,a,h) P_{s,a,h}(\mathring{s}') v_{k,s,a,h}^2(\mathring{s}') + S A T_{\max}^3} + S^2 A T_{\max}^2 \right) \qquad \text{(Lemma 52)}$$

$$= \tilde{\mathcal{O}} \left(\sqrt{S^2 A \sum_{k=1}^{K} \mathrm{Var}_k[\langle n_k, c_k \rangle]} + S^2 A T_{\max}^2 \right)$$

$$\left(\textstyle\sum_{\mathring{s}'} P_{s,a,h}(\mathring{s}') v_{k,s,a,h}^2(\mathring{s}') = \mathbb{V}(P_{s,a,h}, V^{\pi_k, P, c_k}) \text{ and Lemma 26}\right)$$

$$= \tilde{\mathcal{O}} \left(\sqrt{S^2 A \sum_{k=1}^{K} \langle q_k, c_k \circ Q^{\pi_k, P, c_k} \rangle} + S^2 A T_{\max}^2 \right). \qquad \text{(Lemma 26)}$$

Substituting these back completes the proof. ■

Lemma 29 *Consider a sequence of cost functions $\{c_k\}_{k=1}^K$ and transition functions $\{P_k\}_{k=1}^K$ such that $c_k \in \mathcal{C}_\mathcal{M}$ and $P_k \in \mathcal{P}_k$. Then, we have with probability at least $1-4\delta$:*

$$\sum_{k=1}^{K} \sum_{s,a,h,s',h'} q_k(s,a,h)\epsilon_k^\star(s,a,h,s',h') \left|V^{\pi_k,P,c_k}(s',h') - V^{\pi_k,P_k,c_k}(s',h')\right| = \tilde{\mathcal{O}}\left(S^{2.5}A^{1.5}T_{\max}^3\right).$$

Proof Below $\lesssim$ is equivalent to $\tilde{\mathcal{O}}(\cdot)$. Also denote $z = (s,a,h,s',h')$ and $\tilde{z} = (\widetilde{s},\widetilde{a},\widetilde{h},\widetilde{s}',\widetilde{h}')$. By Lemma 30 we have with probability at least $1-2\delta$:

$$\begin{aligned}
\left|V^{\pi_k,P,c_k}(s',h') - V^{\pi_k,P_k,c_k}(s',h')\right| &\lesssim \sum_{\widetilde{s},\widetilde{a},\widetilde{h}} q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h}) \left|P_{\widetilde{s},\widetilde{a},\widetilde{h}}V^{\pi_k,P,c_k} - P_{k,\widetilde{s},\widetilde{a},\widetilde{h}}V^{\pi_k,P,c_k}\right| \\
&\lesssim T_{\max} \sum_{\widetilde{s},\widetilde{a},\widetilde{h}} q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h}) \left\|P_{\widetilde{s},\widetilde{a},\widetilde{h}} - P_{k,\widetilde{s},\widetilde{a},\widetilde{h}}\right\|_1 \\
&\lesssim T_{\max} \sum_{\widetilde{s},\widetilde{a},\widetilde{h},\widetilde{s}',\widetilde{h}'} q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})\epsilon_k^\star(\widetilde{s},\widetilde{a},\widetilde{h},\widetilde{s}',\widetilde{h}'),
\end{aligned}$$

where the second inequality is by Lemma 2, and the third is by Lemma 13. Thus, using Lemma 13 and the Cauchy-Schwarz inequality, we get:

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{s,a,h,s',h'} q_k(s,a,h)\epsilon_k^\star(s,a,h,s',h') \left|V^{\pi_k,P,c_k}(s',h') - V^{\pi_k,P_k,c_k}(s',h')\right| \\
&\lesssim T_{\max} \sum_{k=1}^{K} \sum_{z} q_k(s,a,h)\epsilon_k^\star(s,a,h,s',h') \sum_{\tilde{z}} q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})\epsilon_k^\star(\widetilde{s},\widetilde{a},\widetilde{h},\widetilde{s}',\widetilde{h}') \\
&\lesssim T_{\max} \sum_{k=1}^{K} \sum_{z} q_k(s,a,h)\sqrt{\frac{P_{s,a,h}(s',h')}{N_k^+(s,a)}} \sum_{\tilde{z}} q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})\sqrt{\frac{P_{\widetilde{s},\widetilde{a},\widetilde{h}}(\widetilde{s}',\widetilde{h}')}{N_k^+(\widetilde{s},\widetilde{a})}} \\
&\lesssim T_{\max} \sqrt{\sum_{k,z,\tilde{z}} \frac{q_k(s,a,h)P_{\widetilde{s},\widetilde{a},\widetilde{h}}(\widetilde{s}',\widetilde{h}')q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})}{N_k^+(s,a)}} \sqrt{\sum_{k,z,\tilde{z}} \frac{q_k(s,a,h)P_{s,a,h}(s',h')q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})}{N_k^+(\widetilde{s},\widetilde{a})}}.
\end{aligned}$$

Note that we ignore some lower order terms in the calculation above. To finish the proof we bound each of the terms separately. For the first term we have with probability at least $1-\delta$:

$$\begin{aligned}
&\sum_{k,z,\tilde{z}} \frac{q_k(s,a,h)P_{\widetilde{s},\widetilde{a},\widetilde{h}}(\widetilde{s}',\widetilde{h}')q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})}{N_k^+(s,a)} \\
&= \sum_{k,s,a} \frac{\left(\sum_h q_k(s,a,h)\right)\sum_{s',h',\widetilde{s},\widetilde{a},\widetilde{h}} q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h}) \sum_{\widetilde{s}',\widetilde{h}'} P_{\widetilde{s},\widetilde{a},\widetilde{h}}(\widetilde{s}',\widetilde{h}')}{N_k^+(s,a)} \\
&\lesssim T_{\max} S \sum_{k,s,a} \frac{q_k(s,a)}{N_k^+(s,a)} \lesssim T_{\max}^2 S^2 A,
\end{aligned}$$

where the last inequality is by Lemma 32. For the second term we have with probability at least $1-\delta$:

$$\begin{aligned}
&\sum_{k,z,\widetilde{z}} \frac{q_k(s,a,h)P_{s,a,h}(s',h')q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})}{N_k^+(\widetilde{s},\widetilde{a})} \\
&\lesssim S \sum_{k,s,a,h,\widetilde{s},\widetilde{a},\widetilde{h}} \frac{q_k(s,a,h)\sum_{s',h'}P_{s,a,h}(s',h')q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})}{N_k^+(\widetilde{s},\widetilde{a})} \\
&\lesssim T_{\max} S \sum_{k,s,a,h,\widetilde{s},\widetilde{a},\widetilde{h}} \frac{x_k(s,a,h)\sum_{s',h'}P_{s,a,h}(s',h')q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h})}{N_k^+(\widetilde{s},\widetilde{a})} \\
&\lesssim T_{\max} S \sum_{k,s,a,h,\widetilde{s},\widetilde{a},\widetilde{h}} \frac{q_k(\widetilde{s},\widetilde{a},\widetilde{h})}{N_k^+(\widetilde{s},\widetilde{a})} \lesssim T_{\max} S^2 A \sum_{k,\widetilde{s},\widetilde{a}} \frac{q_k(\widetilde{s},\widetilde{a})}{N_k^+(\widetilde{s},\widetilde{a})} \lesssim S^3A^2T_{\max}^2,
\end{aligned}$$

where the second inequality follows by $q_k(s,a,h) \lesssim T_{\max}x_k(s,a,h)$, the third by $x_k(s,a,h)\sum_{s',h'}P_{s,a,h}(s',h')q_{k,(s',h')}(\widetilde{s},\widetilde{a},\widetilde{h}) \le q_k(\widetilde{s},\widetilde{a},\widetilde{h})$, and the last one by Lemma 32. ■

Lemma 30 (Extended Value Difference) *For any policies π, π', transitions P, P', and cost functions c, c' in $\mathring{\mathcal{M}}$, we have:*

$$\begin{aligned}
&Q^{\pi,P,c}(s,a,h) - Q^{\pi',P',c'}(s,a,h) \\
&= \sum_{s'',h''} P'_{s,a,h}(s'',h'') \sum_{s',h'} q_{\pi',P',(s'',h'')}(s',h') \sum_{a'} \left(\pi(a'|s',h') - \pi'(a'|s',h')\right) Q^{\pi,P,c}(s',a',h') \\
&\quad + \sum_{s',a',h'} q_{\pi',P',(s,a,h)}(s',a',h') \left(Q^{\pi,P,c}(s',a',h') - c'(s',a',h') - P'_{s',a',h'}V^{\pi,P,c}\right).
\end{aligned}$$

and

$$\begin{aligned}
&V^{\pi,P,c}(s,h) - V^{\pi',P',c'}(s,h) \\
&= \sum_{s',h'} q_{\pi',P',(s,h)}(s',h') \sum_{a'} \left(\pi(a'|s',h') - \pi'(a'|s',h')\right) Q^{\pi,P,c}(s',a',h') \\
&\quad + \sum_{s',a',h'} q_{\pi',P',(s,h)}(s',a',h') \left(Q^{\pi,P,c}(s',a',h') - c'(s',a',h') - P'_{s',a',h'}V^{\pi,P,c}\right).
\end{aligned}$$

Proof We first prove the second statement, note that:

$$\begin{aligned}
V^{\pi,P,c}(s,h) - V^{\pi',P',c'}(s,h) &= \sum_{a'} \left(\pi(a'|s,h) - \pi'(a'|s,h)\right) Q^{\pi,P,c}(s,a',h) \\
&\quad + \sum_{a'} \pi'(a'|s,h)(Q^{\pi,P,c}(s,a',h) - Q^{\pi',P',c'}(s,a',h)) \\
&= \sum_{a'} \left(\pi(a'|s,h) - \pi'(a'|s,h)\right) Q^{\pi,P,c}(s,a',h)
\end{aligned}$$

$$+\sum_{a'}\pi'(a'|s,h)\left(Q^{\pi,P,c}(s,a',h)-c'(s,a',h)-P'_{s,a',h}V^{\pi,P,c}\right)$$
$$+\sum_{a'}\pi'(a'|s,h)P'_{s,a',h}(V^{\pi,P,c}-V^{\pi',P',c'}).$$

Applying the equality above recursively and by the definition of $q_{\pi',P',(s,h)}$, we prove the second statement. For the first statement, note that:

$$Q^{\pi,P,c}(s,a,h)-Q^{\pi',P',c'}(s,a,h)$$
$$=\left(Q^{\pi,P,c}(s,a,h)-c'(s,a,h)-P'_{s,a,h}V^{\pi,P,c}\right)+P'_{s,a,h}(V^{\pi,P,c}-V^{\pi',P',c'}).$$

Applying the second statement and the definition of $q_{\pi',P',(s,a,h)}$ completes the proof. ■

Lemma 31 *(Rosenberg and Mansour, 2021, Lemma 6) Let π be a policy with expected hitting time at most τ starting from any state. Then for any $\delta\in(0,1)$, with probability at least $1-\delta$, π takes no more than $4\tau\ln\frac{2}{\delta}$ steps to reach the goal state.*

Lemma 32 *For any $z_k:\mathcal{S}\times\mathcal{A}\to[0,1]$, with probability at least $1-\delta$,*

$$\sum_{k=1}^{K}\sum_{s,a}\frac{\bar{n}_k(s,a)\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}}=\tilde{\mathcal{O}}\left(SAT_{\max}+\sqrt{SA\sum_k\sum_{s,a}\bar{n}_k(s,a)z_k(s,a)}\right)$$
$$=\tilde{\mathcal{O}}\left(SAT_{\max}+\sqrt{SA\sum_k\sum_{s,a}q_k(s,a)z_k(s,a)}\right),$$
$$\sum_{k=1}^{K}\sum_{s,a}\frac{q_k(s,a)\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}}=\tilde{\mathcal{O}}\left(SAT_{\max}+\sqrt{SA\sum_k\sum_{s,a}\bar{n}_k(s,a)z_k(s,a)}\right)$$
$$=\tilde{\mathcal{O}}\left(SAT_{\max}+\sqrt{SA\sum_k\sum_{s,a}q_k(s,a)z_k(s,a)}\right),$$
$$\sum_{k=1}^{K}\sum_{s,a}\frac{\bar{n}_k(s,a)}{N_k^+(s,a)}=\tilde{\mathcal{O}}\left(SAT_{\max}\right),\qquad\sum_{k=1}^{K}\sum_{s,a}\frac{q_k(s,a)}{N_k^+(s,a)}=\tilde{\mathcal{O}}\left(SAT_{\max}\right),$$
$$\sum_{k=1}^{K}\sum_{s,a}\frac{m_k(s,a)}{M_k^+(s,a)}=\tilde{\mathcal{O}}\left(SA\right),\qquad\sum_{k=1}^{K}\sum_{s,a}\frac{x_k(s,a)}{M_k^+(s,a)}=\tilde{\mathcal{O}}\left(SA\right).$$

Proof **First statement:** Since $z_k(s,a)\le 1$ and $\bar{n}_k(s,a)\le L=\tilde{\mathcal{O}}\left(T_{\max}\right)$ we have:

$$\sum_{k=1}^{K}\frac{\bar{n}_k(s,a)\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}}\le\sum_{k=1}^{K}\frac{\bar{n}_k(s,a)\sqrt{z_k(s,a)}}{\sqrt{N_{k+1}^+(s,a)}}+\sum_{k=1}^{K}L\left(\frac{1}{\sqrt{N_k^+(s,a)}}-\frac{1}{\sqrt{N_{k+1}^+(s,a)}}\right)$$
$$\le\sum_{k=1}^{K}\frac{\bar{n}_k(s,a)\sqrt{z_k(s,a)}}{\sqrt{N_{k+1}^+(s,a)}}+\tilde{\mathcal{O}}\left(T_{\max}\right).$$

By Cauchy-Schwarz inequality this implies:

$$\sum_{s,a}\sum_{k=1}^{K}\frac{\bar{n}_k(s,a)\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}} = \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^{K}\sum_{s,a}\frac{\bar{n}_k(s,a)}{N_{k+1}^+(s,a)}}\sqrt{\sum_{k=1}^{K}\sum_{s,a}\bar{n}_k(s,a)z_k(s,a)} + SAT_{\max}\right)$$
$$= \tilde{\mathcal{O}}\left(\sqrt{SA\sum_{k}\sum_{s,a}\bar{n}_k(s,a)z_k(s,a)} + SAT_{\max}\right).$$

Finally, $\sum_{s,a}\sum_k \bar{n}_k(s,a)z_k(s,a) = \tilde{\mathcal{O}}\left(\sum_{s,a}\sum_k q_k(s,a)z_k(s,a) + SAT_{\max}\right)$ with high probability by Lemma 52.

Second statement: By Lemma 27 we have:

$$\sum_{k=1}^{K}\sum_{s,a}\frac{q_k(s,a)\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}} \le \sum_{k=1}^{K}\sum_{s,a}\frac{\mathbb{E}_k[\bar{n}_k(s,a)]\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}} + \tilde{\mathcal{O}}\left(1/K\right)\sum_{k=1}^{K}\sum_{s,a}\frac{\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}}$$
$$\le \sum_{k=1}^{K}\sum_{s,a}\frac{\mathbb{E}_k[\bar{n}_k(s,a)]\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}} + \tilde{\mathcal{O}}\left(SA\right)$$
$$= \tilde{\mathcal{O}}\left(\sum_{k=1}^{K}\sum_{s,a}\frac{\bar{n}_k(s,a)\sqrt{z_k(s,a)}}{\sqrt{N_k^+(s,a)}} + T_{\max}SA\right),$$

where the last relation holds with high probability by Lemma 52. Now the statement follows by the first statement.

Third and forth statements: Similarly to the first statement,

$$\sum_{k=1}^{K}\frac{\bar{n}_k(s,a)}{N_k^+(s,a)} \le \sum_{k=1}^{K}\frac{\bar{n}_k(s,a)}{N_{k+1}^+(s,a)} + \sum_{k=1}^{K}L\left(\frac{1}{N_k^+(s,a)} - \frac{1}{N_{k+1}^+(s,a)}\right)$$
$$\le \sum_{k=1}^{K}\frac{\bar{n}_k(s,a)}{\max\{1,\sum_{i\le k}\bar{n}_i(s,a)\}} + \tilde{\mathcal{O}}\left(T_{\max}\right) = \tilde{\mathcal{O}}\left(T_{\max}\right).$$

Summing over (s,a) proves the third statement. The forth statement is then proved similarly to the second statement.

Fifth and sixth statements: Similarly to the third statement,

$$\sum_{k=1}^{K}\frac{m_k(s,a)}{M_k^+(s,a)} \le \sum_{k=1}^{K}\frac{m_k(s,a)}{M_{k+1}^+(s,a)} + \sum_{k=1}^{K}\left(\frac{1}{M_k^+(s,a)} - \frac{1}{M_{k+1}^+(s,a)}\right)$$
$$\le \sum_{k=1}^{K}\frac{m_k(s,a)}{\max\{1,\sum_{i\le k}m_i(s,a)\}} + 1 = \tilde{\mathcal{O}}\left(1\right).$$

Summing over (s,a) proves the fifth statement. The sixth statement is again obtained with high probability by Lemma 52. ■

Appendix D. Omitted Details for Section 5

Extra Notations Define $\widetilde{q}_k = q_{\pi_k, \widetilde{P}_k}$, $Q_k = Q^{\pi_k, P, c_k}$, $V_k = V^{\pi_k, P, c_k}$, and $A_k = A^{\pi_k, P, c_k}$.

D.1. Proof of Theorem 9

In this part, define $\widetilde{P}_k = \Gamma(\pi_k, \mathcal{P}_k, \widetilde{c}_k)$ and $P_k = \Gamma(\pi_k, \mathcal{P}_k, c_k)$, such that $\widetilde{Q}_k = Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}$, $\widetilde{V}_k = V^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}$, and $\widehat{Q}_k = Q^{\pi_k, P_k, c_k}$. We first provide bounds on some important quantities.

Lemma 33 *$\widetilde{c}_k \in \mathcal{C}_{\mathcal{M}}$, $\eta \left\| \widetilde{A}_k - B_k \right\|_\infty \le 1$, and $\eta \left\| B_k \right\|_\infty \le \frac{1}{2H'}$.*

Proof For the first statement, by $P_k \in \Lambda_{\mathcal{M}}$, we have $\widehat{Q}_k(s,a,h) \le \frac{H}{1-\gamma} + c_f = \chi$. Therefore, $\lambda \widehat{Q}_k(s,a,h) \le 1$ and $\widetilde{c}_k \in \mathcal{C}_{\mathcal{M}}$. For the second statement, by $\widetilde{c}_k(s,a,h) \le 2$ for $h \le H$, we have $|\widetilde{A}_k(s,a,h)| \le |\widetilde{Q}_k(s,a,h)| + |\widetilde{V}_k(s,h)| \le 4(\frac{H}{1-\gamma} + c_f) = 4\chi$ for $h \le H$. Therefore, $\|b_k\|_\infty \le 32\eta\chi^2$, and by Lemma 45, we have $\|B_k\|_\infty \le \frac{15H\|b_k\|_\infty}{1-\gamma} \le 960\eta H T_{\max}\chi^2$. Thus by the definition of η, we have $\eta \left\| B_k \right\|_\infty \le \frac{1}{2H'}$ and $\eta \left\| \widetilde{A}_k - B_k \right\|_\infty \le \eta(\left\| \widetilde{A}_k \right\|_\infty + \|B_k\|_\infty) \le 1$. ■

We are now ready to prove Theorem 9. The proof decomposes the regret into several terms, each of which is bounded by a lemma included after the proof.

Proof [Proof of Theorem 9] With probability at least $1 - 10\delta$, we decompose the regret as follows:

$$\begin{aligned}
\mathring{R}_K &= \sum_{k=1}^K \langle n_k - q_k, c_k \rangle + \langle q_k - q^\star, c_k \rangle \overset{\text{(i)}}{\le} \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^K \langle q_k, c_k \circ Q_k \rangle} + SAT_{\max} \right) \\
&\quad + \sum_{k=1}^K \langle q_k - \widetilde{q}_k, \widetilde{c}_k \rangle + \sum_{k=1}^K \langle \widetilde{q}_k - q^\star, \widetilde{c}_k \rangle - \lambda \sum_{k=1}^K \left\langle q_k, c_k \circ \widehat{Q}_k \right\rangle + \lambda \sum_{k=1}^K \left\langle q^\star, c_k \circ \widehat{Q}_k \right\rangle \\
&\overset{\text{(ii)}}{=} \tilde{\mathcal{O}}\left(\sqrt{S^2 A \sum_{k=1}^K \langle q_k, c_k \circ Q_k \rangle} + S^{2.5} A^{1.5} T_{\max}^3 \right) + \sum_{k=1}^K \langle \widetilde{q}_k - q^\star, \widetilde{c}_k \rangle - \lambda \sum_{k=1}^K \langle q_k, c_k \circ Q_k \rangle \\
&\quad + \lambda \sum_{k=1}^K \left\langle q_k, c_k \circ (Q_k - \widehat{Q}_k) \right\rangle + \lambda \sum_{k=1}^K \left\langle q^\star, c_k \circ Q^{\mathring{\pi}^\star, P, c_k} \right\rangle + \lambda \sum_{k=1}^K \left\langle q^\star, c_k \circ (\widehat{Q}_k - Q^{\mathring{\pi}^\star, P, c_k}) \right\rangle,
\end{aligned}$$

where in (i) we apply Lemma 11 and Lemma 50 to have

$$\begin{aligned}
\sum_{k=1}^K \langle n_k - q_k, c_k \rangle = \sum_{k=1}^K \langle \bar{n}_k - q_k, c_k \rangle &= \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^K \mathbb{E}_k[\langle \bar{n}_k, c_k \rangle^2]} + SAT_{\max} \right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{\sum_{k=1}^K \langle q_k, c_k \circ Q_k \rangle} + SAT_{\max} \right), \qquad \text{(Lemma 26)}
\end{aligned}$$

and in (ii) we apply Lemma 28 and $\widetilde{c}_k \in \mathcal{C}_{\mathcal{M}}$ for $h \le H$ to have

$$\sum_{k=1}^K \langle q_k - \widetilde{q}_k, \widetilde{c}_k \rangle = \tilde{\mathcal{O}}\left(\sqrt{S^2 A \sum_{k=1}^K \left\langle q_k, \widetilde{c}_k \circ Q^{\pi_k, P, \widetilde{c}_k} \right\rangle} + S^{2.5} A^{1.5} T_{\max}^3 \right)$$

$$= \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^{K}\langle q_k, c_k \circ Q^{\pi_k,P,c_k}\rangle} + S^{2.5}A^{1.5}T_{\max}^3\right). \qquad (\widetilde{c}_k(s,a,h) \le 2c_k(s,a,h))$$

Define $\lambda' = \sqrt{\frac{S^2A}{DT_\star K}}$. By Lemma 34, Lemma 35, Lemma 36, and definition of λ, η, with probability at least $1-9\delta$:

$$\begin{aligned}
\mathring{R}_K &\le \tilde{\mathcal{O}}\left(\sqrt{S^2A\sum_{k=1}^{K}\langle q_k, c_k \circ Q_k\rangle} + \frac{T_\star}{\eta} + S^4A^2T_{\max}^5\right) \\
&\quad + 24\eta\sum_{k=1}^{K}\left\langle q_k, A_k^2\right\rangle - \lambda\sum_{k=1}^{K}\langle q_k, c_k \circ Q_k\rangle + \lambda\sum_{k=1}^{K}\left\langle q^\star, c_k \circ Q^{\mathring{\pi}^\star,P,c_k}\right\rangle \\
&= \tilde{\mathcal{O}}\left(\frac{S^2A}{\lambda'}\right) + \lambda'\sum_{k=1}^{K}\langle q_k, c_k \circ Q_k\rangle + \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta} + S^4A^2T_{\max}^5\right) \\
&\quad + 48\eta\sum_{k=1}^{K}\langle q_k, c_k \circ Q_k\rangle - \lambda\sum_{k=1}^{K}\langle q_k, c_k \circ Q_k\rangle + \mathcal{O}\left(\lambda DT_\star K\right) \\
&\qquad\qquad \text{(AM-GM inequality, Lemma 26, and Lemma 37)} \\
&= \tilde{\mathcal{O}}\left(T_\star\sqrt{DK} + \sqrt{S^2ADT_\star K} + S^4A^2T_{\max}^5\right). \quad (K = \tilde{\mathcal{O}}(S^2AT_{\max}^2) \text{ when } \lambda < 48\eta + \lambda')
\end{aligned}$$

Applying Lemma 4 completes the proof. ■

Lemma 34 *With probability at least* $1-6\delta$,

$$\sum_{k=1}^{K}\langle \widetilde{q}_k - q^\star, \widetilde{c}_k\rangle = 24\eta\left\langle q_k, A_k^2\right\rangle + \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta} + S^4A^2T_{\max}^{3.5}\right).$$

Proof Note that by Lemma 49 and Lemma 33:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{s,h} q^\star(s,h)\sum_{a\in\mathcal{A}}(\pi_k(a|s,h) - \mathring{\pi}^\star(a|s,h))\left(\widetilde{A}_k(s,a,h) - B_k(s,a,h)\right) \\
&\le \sum_{s,h} q^\star(s,h)\left(\frac{\ln A}{\eta} + \eta\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\pi_k(a|s,h)\left(\widetilde{A}_k(s,a,h) - B_k(s,a,h)\right)^2\right) \\
&\le \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta}\right) + 2\eta\sum_{s,h} q^\star(s,h)\left(\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\pi_k(a|s,h)\widetilde{A}_k(s,a,h)^2 + \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\pi_k(a|s,h)B_k(s,a,h)^2\right) \\
&= \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta}\right) + \sum_{k=1}^{K}\langle q^\star, b_k\rangle + \frac{1}{H'}\sum_{s,h} q^\star(s,h)\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\pi_k(a|s,h)B_k(s,a,h). \qquad \text{(Lemma 33)}
\end{aligned}$$

Define $\widehat{q}'_k = q_{\pi_k, P'_k}$, where P'_k is the optimistic transition defined in B_k. We have

$$\begin{aligned}
\sum_{k=1}^{K} \langle \widetilde{q}_k - q^\star, \widetilde{c}_k \rangle &= \sum_{k=1}^{K} \sum_{s,h} q^\star(s,h) \sum_{a \in \mathcal{A}} (\pi_k(a|s,h) - \mathring{\pi}^\star(a|s,h)) \left(\widetilde{A}_k(s,a,h) - B_k(s,a,h) \right) \\
&\quad + \sum_{k=1}^{K} \sum_{s,a,h} q^\star(s,a,h) \left(\widetilde{Q}_k(s,a,h) - \widetilde{c}_k(s,a,h) - P_{s,a,h} \widetilde{V}_k \right) \\
&\quad + \sum_{k=1}^{K} \sum_{s,h} q^\star(s,h) \sum_{a \in \mathcal{A}} (\pi_k(a|s,h) - \mathring{\pi}^\star(a|s,h)) B_k(s,a,h) \\
&\qquad\qquad \text{(shifting argument and Lemma 30)} \\
&\overset{\text{(i)}}{\leq} \tilde{\mathcal{O}} \left(\frac{T_\star}{\eta} \right) + 3 \sum_{k=1}^{K} \langle \widehat{q}'_k, b_k \rangle + \tilde{\mathcal{O}} \left(T_{\max} \right) \\
&= \tilde{\mathcal{O}} \left(\frac{T_\star}{\eta} \right) + 6\eta \sum_{k=1}^{K} \left\langle q_k, \widetilde{A}_k^2 \right\rangle + 3 \sum_{k=1}^{K} \langle \widehat{q}'_k - q_k, b_k \rangle ,
\end{aligned}$$

where in (i) we apply Lemma 46, $b_k(s,a,h) = \tilde{\mathcal{O}}(1)$, and the definition of $\widetilde{P}_k$ so that

$$\sum_{k=1}^{K} \sum_{s,a,h} q^\star(s,a,h) \left(\widetilde{Q}_k(s,a,h) - \widetilde{c}_k(s,a,h) - P_{s,a,h} \widetilde{V}_k \right) \leq 0.$$

For the second term, by $(a+b+c)^2 \leq 2a^2 + 2(b+c)^2 \leq 2a^2 + 4b^2 + 4c^2$,

$$\begin{aligned}
\eta \sum_{k=1}^{K} \sum_{s,a,h} q_k(s,a,h) \widetilde{A}_k(s,a,h)^2 &\leq 2\eta \sum_{k=1}^{K} \sum_{s,a,h} q_k(s,a,h) A^{\pi_k, P, \widetilde{c}_k}(s,a,h)^2 \\
&\quad + 4\eta \sum_{k=1}^{K} \sum_{s,a,h} q_k(s,a,h) \left(Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,a,h) - Q^{\pi_k, P, \widetilde{c}_k}(s,a,h) \right)^2 \\
&\quad + 4\eta \sum_{k=1}^{K} \sum_{s,h} q_k(s,h) \left(V^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,h) - V^{\pi_k, P, \widetilde{c}_k}(s,h) \right)^2 \\
&\leq 2\eta \sum_{k=1}^{K} \sum_{s,a,h} q_k(s,a,h) A^{\pi_k, P, \widetilde{c}_k}(s,a,h)^2 \\
&\quad + 8\eta \sum_{k=1}^{K} \sum_{s,a,h} q_k(s,a,h) \left(Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,a,h) - Q^{\pi_k, P, \widetilde{c}_k}(s,a,h) \right)^2 ,
\end{aligned}$$

where in the last step we apply Cauchy-Schwarz inequality to obtain

$$\left(V^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,h) - V^{\pi_k, P, \widetilde{c}_k}(s,h) \right)^2 = \left(\sum_a \pi_k(a|s,h) (Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,a,h) - Q^{\pi_k, P, \widetilde{c}_k}(s,a,h)) \right)^2$$

$$\leq \sum_a \pi_k(a|s,h)\left(Q^{\pi_k,\widetilde{P}_k,\widetilde{c}_k}(s,a,h) - Q^{\pi_k,P,\widetilde{c}_k}(s,a,h)\right)^2.$$

Note that with probability at least $1-2\delta$,

$$\begin{aligned}
&\left|Q^{\pi_k,\widetilde{P}_k,\widetilde{c}_k}(s,a,h) - Q^{\pi_k,P,\widetilde{c}_k}(s,a,h)\right| && (17)\\
&= \tilde{\mathcal{O}}\left(T_{\max}\sum_{s',a',h'\leq H} q_{\pi_k,P,(s,a,h)}(s',a',h')\left\|P_{s',a',h'} - \widetilde{P}_{k,s',a',h'}\right\|_1\right)\\
&\qquad\qquad \text{(Lemma 30, Hölder's inequality, and } V^{\pi_k,\widetilde{P}_k,\widetilde{c}_k} = \tilde{\mathcal{O}}(T_{\max}))\\
&= \tilde{\mathcal{O}}\left(T_{\max}S\sum_{s',a'}\frac{q_{\pi_k,P,(s,a,h)}(s',a')}{\sqrt{N_k^+(s',a')}}\right). && \text{(Lemma 13)}
\end{aligned}$$

Therefore, with probability at least $1-\delta$,

$$\begin{aligned}
&\eta\sum_{k=1}^K\sum_{s,a,h} q_k(s,a,h)\left(Q^{\pi_k,\widetilde{P}_k,\widetilde{c}_k}(s,a,h) - Q^{\pi_k,P,\widetilde{c}_k}(s,a,h)\right)^2\\
&\leq \eta\sum_{k=1}^K\sum_{s,a,h\leq H} q_k(s,a,h)T_{\max}^3S^2\sum_{s',a'}\frac{q_{\pi_k,P,(s,a,h)}(s',a')}{N_k^+(s',a')} && \text{(Cauchy-Schwarz inequality)}\\
&\overset{\text{(i)}}{=} \tilde{\mathcal{O}}\left(\eta T_{\max}^4S^3A\sum_{k=1}^K\sum_{s',a'}\frac{q_{\pi_k,P}(s',a')}{N_k^+(s',a')}\right) \overset{\text{(ii)}}{=} \tilde{\mathcal{O}}\left(\eta T_{\max}^5S^4A^2\right),
\end{aligned}$$

where in (i) we apply $q_k(s,a,h) \leq 2T_{\max}x_k(s,a,h)$ and $x_k(s,a,h)q_{\pi_k,P,(s,a,h)}(s',a') \leq q_{\pi_k,P}(s',a')$, and in (ii) we apply Lemma 32. Plugging these back, we get:

$$\begin{aligned}
&\eta\sum_{k=1}^K\sum_{s,a,h} q_k(s,a,h)\widetilde{A}_k(s,a,h)^2 \leq 2\eta\sum_{k=1}^K\left\langle q_k,(A^{\pi_k,P,\widetilde{c}_k})^2\right\rangle + \tilde{\mathcal{O}}\left(\eta T_{\max}^5S^4A^2\right)\\
&\leq 4\eta\sum_{k=1}^K\left\langle q_k,A_k^2\right\rangle + 4\eta\lambda^2\sum_{k=1}^K\left\langle q_k,(A^{\pi_k,P,\widehat{Q}_k})^2\right\rangle + \tilde{\mathcal{O}}\left(\eta T_{\max}^5S^4A^2\right) \quad ((a+b)^2\leq 2a^2+2b^2)\\
&\leq 4\eta\left\langle q_k,A_k^2\right\rangle + \tilde{\mathcal{O}}\left(\eta T_{\max}^5S^4A^2 + \eta\lambda^2T_{\max}^5K\right) = 4\eta\left\langle q_k,A_k^2\right\rangle + \tilde{\mathcal{O}}\left(S^4A^2T_{\max}^3\right).
\end{aligned}$$

For the third term, with probability at least $1-3\delta$,

$$\begin{aligned}
&\sum_{k=1}^K\left\langle \widehat{q}_k' - q_k, b_k\right\rangle \leq \sum_{k=1}^K\sum_{s,a,h\leq H} q_k(s,a,h)\left\|\widehat{P}'_{k,s,a,h} - P_{s,a,h}\right\|_1\left\|V^{\pi_k,\widehat{P}'_k,b_k}\right\|_\infty\\
&\qquad\qquad \text{(Lemma 30 and Hölder's inequality)}\\
&= \tilde{\mathcal{O}}\left(\eta ST_{\max}^3\sum_{k=1}^K\sum_{s,a,h\leq H}\frac{q_k(s,a,h)}{\sqrt{N_k^+(s,a)}}\right) \qquad (b_k(s,a,h) = \tilde{\mathcal{O}}(\eta T_{\max}^2) \text{ and Lemma 13})
\end{aligned}$$

$$= \tilde{\mathcal{O}}\left(\eta S T_{\max}^3 \sqrt{SAT_{\max}K} + \eta S^2 A T_{\max}^4\right) = \tilde{\mathcal{O}}\left(S^2 A T_{\max}^{3.5}\right). \qquad \text{(Lemma 32)}$$

Putting everything together completes the proof. ■

Lemma 35 $\lambda \sum_{k=1}^K \left\langle q^\star, c_k \circ (\widehat{Q}_k - Q^{\mathring{\pi}^\star, P, c_k}) \right\rangle = \tilde{\mathcal{O}}\left(S^2 A T_{\max}^5\right)$.

Proof Define $q'_{s,a,h}(s',h') = \sum_{s'',h''} P_{s,a,h}(s'',h'') q^\star_{(s'',h'')}(s',h')$. We have for $h \le H$:

$$\begin{aligned}
&\lambda \sum_{k=1}^K (\widehat{Q}_k(s,a,h) - Q^{\mathring{\pi}^\star, P, c_k}(s,a,h)) \le \lambda \sum_{k=1}^K (Q^{\pi_k, \widetilde{P}_k, c_k}(s,a,h) - Q^{\mathring{\pi}^\star, P, c_k}(s,a,h)) \\
&\le \lambda \sum_{k=1}^K \left(Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,a,h) - Q^{\mathring{\pi}^\star, P, \widetilde{c}_k}(s,a,h)\right) + \lambda^2 \sum_{k=1}^K Q^{\mathring{\pi}^\star, P, \widehat{Q}_k}(s,a,h) \\
&\le \lambda \sum_{k=1}^K \left(Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s,a,h) - Q^{\mathring{\pi}^\star, P, \widetilde{c}_k}(s,a,h)\right) + \tilde{\mathcal{O}}\left(\lambda^2 T_{\max}^2 K\right) \\
&\le \lambda \sum_{s',h'} q'_{s,a,h}(s',h') \sum_{k=1}^K \sum_{a'} \left(\pi_k(a'|s',h') - \mathring{\pi}^\star(a'|s',h')\right) \left(\widetilde{A}_k(s',a',h') - B_k(s',a',h')\right) \\
&\quad + \lambda \sum_{s',a',h'} q^\star_{(s,a,h)}(s',a',h') \sum_{k=1}^K \left(Q^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}(s',a',h') - \widetilde{c}_k(s',a',h') - P_{s',a',h'} V^{\pi_k, \widetilde{P}_k, \widetilde{c}_k}\right) \\
&\quad + \lambda \sum_{s',h'} q'_{s,a,h}(s',h') \sum_{k=1}^K \sum_{a'} \left(\pi_k(a'|s',h') - \mathring{\pi}^\star(a'|s',h')\right) B_k(s',a',h') + \tilde{\mathcal{O}}\left(\lambda^2 T_{\max}^2 K\right) \\
&\qquad\qquad \text{(Lemma 30)} \\
&= \tilde{\mathcal{O}}\left(\lambda \sum_{s',h'} q'_{s,a,h}(s',h') \left(\frac{T_\star}{\eta} + \eta \sum_{k=1}^K \sum_{a \in \mathcal{A}} \pi_k(a|s,h) T_{\max}^2\right) + \lambda \eta T_{\max}^4 K + \lambda^2 T_{\max}^2 K\right) \\
&\qquad\qquad \left(\left\|\widetilde{A}_k\right\|_\infty = \tilde{\mathcal{O}}(T_{\max}), \text{ definition of } \widetilde{P}_k, \text{ and } B_k(s,a,h) = \tilde{\mathcal{O}}(\eta T_{\max}^3)\right) \\
&= \tilde{\mathcal{O}}\left(\frac{\lambda T_{\max}^2}{\eta} + \lambda \eta T_{\max}^4 K + \lambda^2 T_{\max}^2 K\right).
\end{aligned}$$

Plugging this back and by the definition of λ, η:

$$\lambda \sum_{k=1}^K \left\langle q^\star, \widehat{Q}_k - Q^{\mathring{\pi}^\star, P, c_k} \right\rangle = \tilde{\mathcal{O}}\left(\frac{\lambda T_{\max}^3}{\eta} + \lambda \eta T_{\max}^5 K + \lambda^2 T_{\max}^3 K\right) = \tilde{\mathcal{O}}\left(S^2 A T_{\max}^5\right).$$

This completes the proof. ■

Lemma 36 *With probability at least* $1 - 3\delta$, $\lambda \sum_{k=1}^K \left\langle q_k, c_k \circ (Q_k - \widehat{Q}_k) \right\rangle = \tilde{\mathcal{O}}\left(S^{3.5} A^2 T_{\max}^3\right)$.

Proof By similar arguments as in Eq. (17) with $\widetilde{P}_k$ replaced by P_k and $\widetilde{c}_k$ replaced by c_k, with probability at least $1-3\delta$:

$$\lambda \sum_{k=1}^{K} \left\langle q_k, Q_k - \widehat{Q}_k \right\rangle = \lambda T_{\max} S \sum_{k=1}^{K} \sum_{s,a,h\le H} q_k(s,a,h) \sum_{s',a'} \frac{q_{\pi_k,P,(s,a,h)}(s',a')}{\sqrt{N_k^+(s',a')}}$$

$$= \tilde{\mathcal{O}}\left(\lambda S^2 A T_{\max}^2 \sum_{k=1}^{K} \sum_{s',a'} \frac{q_k(s',a')}{\sqrt{N_k^+(s',a')}}\right)$$

$$(q_k(s,a,h) = \mathcal{O}(T_{\max} x_k(s,a,h)) \text{ and } x_k(s,a,h) q_{\pi_k,P,(s,a,h)}(s',a') \le q_k(s',a'))$$

$$= \tilde{\mathcal{O}}\left(\lambda S^2 A T_{\max}^2 \sqrt{SAT_{\max}K} + \lambda S^3 A^2 T_{\max}^3\right) = \tilde{\mathcal{O}}\left(S^{3.5} A^2 T_{\max}^3\right). \qquad \text{(Lemma 32)}$$

■

Lemma 37 $\sum_{k=1}^{K} \left\langle q^\star, c_k \circ Q^{\mathring{\pi}^\star,P,c_k} \right\rangle = \tilde{\mathcal{O}}\left(DT_\star K\right) + T_{\max}$.

Proof By Lemma 2, for $h \le H$, $\sum_{k=1}^{K} c_k(s,a,h) Q^{\mathring{\pi}^\star,P,c_k}(s,a,h) \le \sum_{k=1}^{K} (Q^{\pi^\star,P,c_k}(s,a) + c_f) = \tilde{\mathcal{O}}(DK)$. Therefore,

$$\sum_{k=1}^{K} \left\langle q^\star, c_k \circ Q^{\mathring{\pi}^\star,P,c_k} \right\rangle$$

$$= \sum_{k=1}^{K} \sum_{s,a,h\le H} q^\star(s,a,h) c_k(s,a,h) Q^{\mathring{\pi}^\star,P,c_k}(s,a,h) + \sum_{k=1}^{K} \sum_{s,a} q^\star(s,a,H+1) c_f$$

$$\le \tilde{\mathcal{O}}\left(DT_\star K\right) + T_{\max},$$

where the last step is by $\sum_{s,a,h\le H} q^\star(s,a,h) = \mathcal{O}(T_\star)$, $\sum_{k=1}^{K} Q^{\mathring{\pi}^\star,P,c_k}(s,a,h) = \mathcal{O}(DK)$, and Lemma 19. ■

D.2. Proof of Theorem 10

Here we denote by $\widehat{q}'_k$ the occupancy measure w.r.t policy π_k and the optimistic transition defined in B_k. Also define $\bar{Q}_k(s,a,h) = \mathbb{E}_k[\sum_{i=1}^{\min\{J_k,L\}+1} c(s_i^k, a_i^k, h_i^k) | \pi_k, P, s_1^k = s, a_1^k = a, h_1^k = h]$, such that $\mathbb{E}_k[G_{k,s,a,h}] = x_k(s,a,h)\bar{Q}_k(s,a,h)$. We again decompose the regret into several terms, each of which is bounded by a lemma included after the proof.

Proof With probability at least $1-2\delta$, we decompose the regret as follows:

$$\mathring{R}_K = \sum_{k=1}^{K} \langle n_k - q^\star, c_k \rangle = \sum_{k=1}^{K} \langle \bar{n}_k - q_k, c_k \rangle + \sum_{k=1}^{K} \langle q_k - q^\star, c_k \rangle \qquad \text{(Lemma 11)}$$

$$= \tilde{\mathcal{O}}\left(T_{\max}\sqrt{K} + SAT_{\max}\right) + \sum_{k=1}^{K} \sum_{s,h} q^\star(s,h) \sum_{a\in\mathcal{A}} (\pi_k(a|s,h) - \mathring{\pi}^\star(a|s,h)) Q_k(s,a,h)$$

(Lemma 50 and Lemma 30)

$$= \tilde{\mathcal{O}}\left(T_{\max}\sqrt{K} + SAT_{\max}\right) + \underbrace{\sum_{k=1}^{K}\sum_{s,h} q^\star(s,h)\sum_{a\in\mathcal{A}}(\pi_k(a|s,h) - \mathring{\pi}^\star(a|s,h))\widetilde{Q}_k(s,a,h)}_{\text{REG}}$$
$$+ \underbrace{\sum_{k=1}^{K}\sum_{s,h} q^\star(s,h)\sum_{a\in\mathcal{A}}\pi_k(a|s,h)(Q_k(s,a,h) - \widetilde{Q}_k(s,a,h))}_{\text{BIAS}_1}$$
$$+ \underbrace{\sum_{k=1}^{K}\sum_{s,h} q^\star(s,h)\sum_{a\in\mathcal{A}}\mathring{\pi}^\star(a|s,h)(\widetilde{Q}_k(s,a,h) - Q_k(s,a,h))}_{\text{BIAS}_2}.$$

Therefore, by Lemma 40, we have $\mathring{R}_K = \tilde{\mathcal{O}}(\sqrt{S^2AT_{\max}^5K} + S^{5.5}A^{3.5}T_{\max}^5)$ with probability at least $1 - 25\delta$. Applying Lemma 4 completes the proof. ■

Lemma 38 $\left\|\widetilde{Q}_k\right\|_\infty \le L'/\theta$, $\|b_k\|_\infty \le 5L'$, $\|B_k\|_\infty \le 150HT_{\max}L'$, *and* $\eta\left\|\widetilde{Q}_k - B_k\right\|_\infty \le 1$.

Proof The first statement is by the definition of $\widetilde{Q}_k$ and $G_{k,s,a,h} \le L'$. For the second statement, $b_k \le 5L'$ by definition. For the third statement, by Lemma 45, we have $\|B_k\|_\infty \le \frac{15H\|b_k\|_\infty}{1-\gamma} \le 150HT_{\max}L'$. For the fourth statement, we have $\eta\left\|\widetilde{Q}_k - B_k\right\|_\infty \le \eta(\left\|\widetilde{Q}_k\right\|_\infty + \|B_k\|_\infty) \le 1/2 + \eta 150HT_{\max}L' \le 1$. ■

Lemma 39 $Q_k(s,a,h) - \bar{Q}_k(s,a,h) = \tilde{\mathcal{O}}(1/K)$.

Proof Note that:

$$Q_k(s,a,h) - \bar{Q}_k(s,a,h) = \mathbb{E}_k\left[\sum_{i=\bar{J}_k+2}^{J_k+1} c(s_i^k, a_i^k, h_i^k)\middle|\pi_k, P, s_1^k = s, a_1^k = a, h_1^k = h\right]$$
$$= \tilde{\mathcal{O}}\left(\frac{T_{\max}}{T_{\max}K}\right) = \tilde{\mathcal{O}}(1/K). \qquad \text{(Lemma 31)}$$

■

Lemma 40 *With probability at least* $1 - 25\delta$,

$$\text{REG} + \text{BIAS}_1 + \text{BIAS}_2 = \tilde{\mathcal{O}}\left(\sqrt{S^2AT_{\max}^5K} + S^{5.5}A^{3.5}T_{\max}^5\right).$$

Proof Define $\xi_B = \sum_{k=1}^{K}\sum_{s,h\le H} q^\star(s,h)\sum_{a\in\mathcal{A}}(\pi_k(a|s,h) - \mathring{\pi}^\star(a|s,h))B_k(s,a,h)$. By Lemma 38 and Lemma 49, with probability at least $1-\delta$,

$$\text{REG} = \sum_{k=1}^{K}\sum_{s,h\le H} q^\star(s,h)\sum_{a\in\mathcal{A}}(\pi_k(a|s,h) - \mathring{\pi}^\star(a|s,h))\left(\widetilde{Q}_k(s,a,h) - B_k(s,a,h)\right) + \xi_B$$

$$
\begin{aligned}
&\leq \frac{T_\star \ln A}{\eta} + \eta \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \pi_k(a|s,h) \left(\widetilde{Q}_k(s,a,h) - B_k(s,a,h)\right)^2 + \xi_B \\
&\leq \frac{T_\star \ln A}{\eta} + 2\eta \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \pi_k(a|s,h) \widetilde{Q}_k^2(s,a,h) \\
&\quad + \frac{1}{H'} \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \pi_k(a|s,h) B_k(s,a,h) + \xi_B \\
&\qquad\qquad ((a+b)^2 \leq 2a^2 + 2b^2 \text{ and } \eta \left\|B_k\right\|_\infty \leq \tfrac{1}{H'}) \\
&\leq \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta}\right) + \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \pi_k(a|s,h) \frac{2\theta L'}{\overline{x}_k(s,a,h)+\theta} \\
&\quad + \frac{1}{H'} \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \pi_k(a|s,h) B_k(s,a,h) + \xi_B,
\end{aligned}
$$

where in the last inequality we apply:

$$
\begin{aligned}
&2\eta \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \pi_k(a|s,h) \widetilde{Q}_k^2(s,a,h) \\
&= 2\eta \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \pi_k(a|s,h) \frac{G_{k,s,a,h}^2}{(\overline{x}_k(s,a,h)+\theta)^2} \\
&\leq L' \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \frac{\theta \pi_k(a|s,h)}{\overline{x}_k(s,a,h)+\theta} \frac{m_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta} \\
&\qquad\qquad (2\eta = \theta/L' \text{ and } G_{k,s,a,h} \leq L' m_k(s,a,h)) \\
&\leq L' \sum_{s,h\leq H} q^\star(s,h) \left(2 \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \frac{\theta \pi_k(a|s,h)}{\overline{x}_k(s,a,h)+\theta} + \tilde{\mathcal{O}}\left(\frac{1}{\theta}\right)\right) \quad \text{(Lemma 52 and } \tfrac{x_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta} \leq 1) \\
&\leq \sum_{s,h\leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a\in\mathcal{A}} \frac{2\theta L' \pi_k(a|s,h)}{\overline{x}_k(s,a,h)+\theta} + \tilde{\mathcal{O}}\left(\frac{T_\star L'}{\theta}\right).
\end{aligned}
$$

Therefore, by Lemma 42, Lemma 43, and Lemma 46, with probability at least $1 - 24\delta$,

$$
\begin{aligned}
&\textsc{Reg} + \textsc{Bias}_1 + \textsc{Bias}_2 \leq \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta}\right) + 3\sum_{k=1}^{K} \left\langle \widehat{q}'_k, b_k \right\rangle + \tilde{\mathcal{O}}\left(T_{\max}^2\right) \\
&\qquad\qquad (\theta = 2\eta L' \text{ and } \left\|b_k\right\|_\infty = \tilde{\mathcal{O}}(T_{\max}) \text{ by Lemma 38}) \\
&\leq \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta} + \sum_{k=1}^{K} \sum_{s,a,h\leq H} \widehat{q}'_k(s,a,h) \frac{L'(\overline{x}_k(s,a,h) - \underline{x}_k(s,a,h)) + \theta L'}{\overline{x}_k(s,a,h)+\theta} + T_{\max}^2\right)
\end{aligned}
$$

$$\leq \tilde{\mathcal{O}}\left(\frac{T_\star}{\eta} + L'\sum_{k=1}^{K}\sum_{s,a,h\leq H}(\overline{x}_k(s,a,h) - \underline{x}_k(s,a,h))\widehat{q}'_{k,(s,a,h)}(s,a,h) + \theta T_{\max}L'SAK + T_{\max}^2\right).$$

$$(\widehat{q}'_k(s,a,h) \leq \overline{x}_k(s,a,h)\widehat{q}'_{k,(s,a,h)}(s,a,h) \text{ and } \widehat{q}'_k(s,a,h) = \tilde{\mathcal{O}}(T_{\max}\overline{x}_k(s,a,h)))$$

$$= \tilde{\mathcal{O}}\left(\sqrt{S^2AT_{\max}^5K} + S^{5.5}A^{3.5}T_{\max}^5\right). \qquad \text{(Lemma 41 and definition of } \eta, \theta)$$

This completes the proof. ■

Lemma 41 *With probability at least* $1 - 22\delta$,

$$\sum_{k=1}^{K}\sum_{s,a,h\leq H}(\overline{x}_k(s,a,h) - \underline{x}_k(s,a,h))\widehat{q}'_{k,(s,a,h)}(s,a,h) = \tilde{\mathcal{O}}(\sqrt{S^2AT_{\max}^3K} + S^{5.5}A^{3.5}T_{\max}^4).$$

Proof For any $z \in \mathcal{S}\times\mathcal{A}\times[H]$, denote by $\overline{q}_k^z$ / $\underline{q}_k^z$ the occupancy measure w.r.t the policy and transition defined in $\overline{x}_k(z)$ / $\underline{x}_k(z)$ (transition at (s,a,h) can be randomly pick as long as $P_{\overline{q}_k^z}, P_{\underline{q}_k^z} \in \mathcal{P}_k$). For a fixed tuple $z = (s,a,h)$,

$$\overline{x}_k(s,a,h)\widehat{q}'_{k,(s,a,h)}(s,a,h) = \overline{q}_k^z(s,a,h) + \overline{x}_k(s,a,h)(\widehat{q}'_{k,(s,a,h)}(s,a,h) - \overline{q}^z_{k,(s,a,h)}(s,a,h))$$

$$\leq \overline{q}_k^z(s,a,h) + 2\overline{x}_k(s,a,h)\sum_{s',a',h'}\overline{q}^z_{k,(s,a,h)}(s',a',h')\sum_{s'',h''}\epsilon_k^\star(s',a',h',s'',h'')\widehat{q}'_{k,(s'',h'')}(s,a,h)$$

(Lemma 30 and Lemma 13)

$$\leq \overline{q}_k^z(s,a,h) + 2\sum_{s',a',h'}\overline{q}_k^z(s',a',h')\sum_{s'',h''}\epsilon_k^\star(s',a',h',s'',h'')\widehat{q}'_{k,(s'',h'')}(s,a,h)$$

$$(\overline{x}_k(s,a,h)\overline{q}^z_{k,(s,a,h)}(s',a',h') \leq \overline{q}_k^z(s',a',h'))$$

$$= \overline{q}_k^z(s,a,h) + 2\sum_{s',a',h'}q_k(s',a',h')\sum_{s'',h''}\epsilon_k^\star(s',a',h',s'',h'')\widehat{q}'_{k,(s'',h'')}(s,a,h)$$

$$+ 2\sum_{s',a',h'}(\overline{q}_k^z(s',a',h') - q_k(s',a',h'))\sum_{s'',h''}\epsilon_k^\star(s',a',h',s'',h'')\widehat{q}'_{k,(s'',h'')}(s,a,h).$$

Therefore, with probability at least $1 - 7\delta$,

$$\sum_{k=1}^{K}\sum_{s,a,h\leq H}\overline{x}_k(s,a,h)\widehat{q}'_{k,(s,a,h)}(s,a,h) \overset{\text{(i)}}{\leq} \sum_{k=1}^{K}\sum_{s,a,h\leq H}\overline{q}_k^z(s,a,h)$$

$$+ \tilde{\mathcal{O}}\left(T_{\max}\sum_{k=1}^{K}\sum_{s',a',h'}q_k(s',a',h')\sum_{s'',h''}\epsilon_k^\star(s',a',h',s'',h'') + S^{5.5}A^{3.5}T_{\max}^4\right)$$

$$\overset{\text{(ii)}}{\leq} \sum_{k=1}^{K}\sum_{s,a,h\leq H}\overline{q}_k^z(s,a,h) + \tilde{\mathcal{O}}\left(\sqrt{S^2AT_{\max}^3K} + S^{5.5}A^{3.5}T_{\max}^4\right),$$

where in (i) we apply $\sum_{s,a,h\leq H}\widehat{q}'_{k,(s'',h'')}(s,a,h) = \tilde{\mathcal{O}}(T_{\max})$ and ($z = (s,a,h)$ iterates over $\mathcal{S}\times\mathcal{A}\times[H]$):

$$\sum_{k,z}\sum_{s',a',h'}(\overline{q}_k^z(s',a',h') - q_k(s',a',h'))\sum_{s'',h''}\epsilon_k^\star(s',a',h',s'',h'')\widehat{q}'_{k,(s'',h'')}(z)$$

$$\leq \sum_{k,z} \sum_{\substack{\widetilde{s},\widetilde{a},\widetilde{h} \\ \widetilde{s}',\widetilde{h}'}} \sum_{\substack{s',a',h' \\ s'',h''}} q_k(\widetilde{s},\widetilde{a},\widetilde{h}) \epsilon_k^\star(\widetilde{s},\widetilde{a},\widetilde{h},\widetilde{s}',\widetilde{h}') \overline{q}^z_{k,(\widetilde{s}',\widetilde{h}')}(s',a',h') \epsilon_k^\star(s',a',h',s'',h'') \widehat{q}'_{k,(s'',h'')}(z)$$

(Lemma 30 and Lemma 13)

$$\leq \sum_{k,z} \sum_{\substack{s',a',h' \\ s'',h''}} \sum_{\substack{\widetilde{s},\widetilde{a},\widetilde{h} \\ \widetilde{s}',\widetilde{h}'}} q_k(\widetilde{s},\widetilde{a},\widetilde{h}) \epsilon_k^\star(\widetilde{s},\widetilde{a},\widetilde{h},\widetilde{s}',\widetilde{h}') q_{k,(\widetilde{s}',\widetilde{h}')}(s',a',h') \epsilon_k^\star(s',a',h',s'',h'') \widehat{q}'_{k,(s'',h'')}(z)$$

$$+ \tilde{\mathcal{O}}\left(S^{2.5} A^{1.5} T_{\max}^3 \sum_z \sum_{\substack{s',a',h' \\ s'',h''}} T_{\max} \right)$$

($\epsilon_k^\star(s',a',h',s'',h'') \widehat{q}'_{k,(s'',h'')}(z) = \tilde{\mathcal{O}}(T_{\max})$ and Lemma 29)

$$= \tilde{\mathcal{O}}\left(S^{5.5} A^{3.5} T_{\max}^4 \right).$$

and in (ii) we apply:

$$T_{\max} \sum_{k=1}^K \sum_{s',a',h'} q_k(s',a',h') \sum_{s'',h''} \epsilon_k^\star(s',a',h',s'',h'')$$

$$= \tilde{\mathcal{O}}\left(T_{\max} \sum_{k=1}^K \sum_{s',a',h' \leq H} q_k(s',a',h') \sum_{s'',h''} \left(\sqrt{\frac{P_{s',a',h'}(s'',h'')}{N_k^+(s',a')}} + \frac{1}{N_k^+(s',a')} \right) \right)$$

(definition of $\epsilon_k^\star$)

$$= \tilde{\mathcal{O}}\left(T_{\max} \sqrt{S} \sum_{k=1}^K \sum_{s',a'} \frac{q_k(s',a')}{\sqrt{N_k^+(s',a')}} + T_{\max} S \sum_{k=1}^K \sum_{s',a'} \frac{q_k(s',a')}{N_k^+(s',a')} \right)$$

$$= \tilde{\mathcal{O}}\left(\sqrt{S^2 A T_{\max}^3 K} + S^2 A T_{\max}^2 \right). \qquad \text{(Lemma 32)}$$

By similar arguments, we also have with probability at least $1 - 7\delta$,

$$-\sum_{k=1}^K \sum_{s,a,h \leq H} \underline{x}_k(s,a,h) \widehat{q}'_{k,(s,a,h)}(s,a,h)$$

$$\leq -\sum_{k=1}^K \sum_{s,a,h \leq H} \underline{q}^z_k(s,a,h) + \tilde{\mathcal{O}}\left(\sqrt{S^2 A T_{\max}^3 K} + S^{5.5} A^{3.5} T_{\max}^4 \right).$$

Therefore, with probability at least $1 - 7\delta$,

$$\sum_{k=1}^K \sum_{s,a,h \leq H} (\overline{x}_k(s,a,h) - \underline{x}_k(s,a,h)) \widehat{q}'_{k,(s,a,h)}(s,a,h)$$

$$= \tilde{\mathcal{O}}\left(\sum_{k=1}^K \sum_{s,a,h \leq H} (\overline{q}_k^{(s,a,h)}(s,a,h) - \underline{q}_k^{(s,a,h)}(s,a,h)) + \sqrt{S^2 A T_{\max}^3 K} + S^{5.5} A^{3.5} T_{\max}^4 \right)$$

$$= \tilde{\mathcal{O}}\left(\sqrt{S^2AT_{\max}^3K} + S^{5.5}A^{3.5}T_{\max}^4\right),$$

where in the last inequality we apply (similarly for $\sum_{k=1}^K (q_k(s,a,h) - \underline{q}_k^{(s,a,h)}(s,a,h))$):

$$\sum_{k=1}^K \sum_{s,a,h\le H} (\overline{q}_k^{(s,a,h)}(s,a,h) - q_k(s,a,h))$$

$$\le \sum_{k,s,a,h\le H} \sum_{s',a',h'} q_k(s',a',h') \sum_{s'',h''} \epsilon_k^\star(s',a',h',s'',h'')\overline{q}_{k,(s'',h'')}^{(s,a,h)}(s,a,h)$$

(Lemma 30 and Lemma 13)

$$\le \sum_{k,s,a,h\le H} \sum_{s',a',h'} q_k(s',a',h') \sum_{s'',h''} \epsilon_k^\star(s',a',h',s'',h'')q_{k,(s'',h'')}(s,a,h) + \tilde{\mathcal{O}}\left(\sum_{s,a,h} S^{2.5}A^{1.5}T_{\max}^3\right)$$

(Lemma 29)

$$= \tilde{\mathcal{O}}\left(T_{\max} \sum_{k,s',a',h'\le H} q_k(s',a',h') \sum_{s'',h''} \left(\sqrt{\frac{P_{s',a',h'}(s'',h'')}{N_k^+(s',a')}} + \frac{1}{N_k^+(s',a')}\right) + S^{3.5}A^{2.5}T_{\max}^3\right)$$

(definition of $\epsilon_k^\star$)

$$= \tilde{\mathcal{O}}\left(\sqrt{S^2AT_{\max}^3K} + S^{3.5}A^{2.5}T_{\max}^3\right).$$ (Cauchy-Schwarz inequality and Lemma 32)

This completes the proof. ■

Lemma 42 *With probability at least* $1-\delta$,

$$\textsc{Bias}_1 \le \sum_{k=1}^K \sum_{s,h\le H} q^\star(s,h) \sum_{a\in\mathcal{A}} \pi_k(a|s,h) \frac{L'(\overline{x}_k(s,a,h) - \underline{x}_k(s,a,h)) + 2\theta L'}{\overline{x}_k(s,a,h) + \theta} + \tilde{\mathcal{O}}\left(\frac{T_\star L'}{\theta}\right).$$

Proof Note that:

$$\textsc{Bias}_1 = \sum_{k=1}^K \sum_{s,h\le H} q^\star(s,h) \sum_{a\in\mathcal{A}} \pi_k(a|s,h)(Q_k(s,a,h) - \mathbb{E}_k[\widetilde{Q}_k(s,a,h)])$$
$$+ \sum_{k=1}^K \sum_{s,h\le H} q^\star(s,h) \sum_{a\in\mathcal{A}} \pi_k(a|s,h)(\mathbb{E}_k[\widetilde{Q}_k(s,a,h)] - \widetilde{Q}_k(s,a,h)).$$

For the first term,

$$\sum_{k=1}^K \sum_{s,h\le H} q^\star(s,h) \sum_{a\in\mathcal{A}} \pi_k(a|s,h)(Q_k(s,a,h) - \mathbb{E}_k[\widetilde{Q}_k(s,a,h)])$$

$$\le \sum_{k=1}^K \sum_{s,h\le H} q^\star(s,h) \sum_{a\in\mathcal{A}} \pi_k(a|s,h)\bar{Q}_k(s,a,h)\left(1 - \frac{x_k(s,a,h)}{\overline{x}_k(s,a,h) + \theta}\right) + \tilde{\mathcal{O}}(T_\star)$$

(Lemma 39 and $\mathbb{E}_k[G_{k,s,a,h}] = x_k(s,a,h)\bar{Q}_k(s,a,h)$)

$$\leq \sum_{k=1}^{K} \sum_{s,h \leq H} q^\star(s,h) \sum_{a \in \mathcal{A}} \pi_k(a|s,h) \frac{L'(\overline{x}_k(s,a,h) - \underline{x}_k(s,a,h) + \theta)}{\overline{x}_k(s,a,h) + \theta} + \tilde{\mathcal{O}}\left(T_\star\right).$$

For the second term, first note that $G_{k,s,a,h} \leq L' m_k(s,a,h)$ and

$$\begin{aligned}
&\mathrm{Var}_k\left[\left\langle \pi_k(\cdot|s,h), \widetilde{Q}_k(s,\cdot,h)\right\rangle\right] \leq \mathbb{E}_k\left[\left\langle \pi_k(\cdot|s,h), \widetilde{Q}_k(s,\cdot,h)\right\rangle^2\right] \\
&\leq \sum_{a \in \mathcal{A}} \pi_k(a|s,h) \frac{\mathbb{E}_k[G_{k,s,a,h}^2]}{(\overline{x}_k(s,a,h)+\theta)^2} \leq \sum_{a \in \mathcal{A}} \pi_k(a|s,h) \frac{L'^2}{\overline{x}_k(s,a,h)+\theta}. \\
&\qquad\qquad \text{(Cauchy-Schwarz inequality and } \mathbb{E}_k[G_{k,s,a,h}] \leq L' x_k(s,a,h))
\end{aligned}$$

Therefore, by Lemma 50, with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{s,h \leq H} q^\star(s,h) \sum_{a \in \mathcal{A}} \pi_k(a|s,h)(\mathbb{E}_k[\widetilde{Q}_k(s,a,h)] - \widetilde{Q}_k(s,a,h)) \\
&= \tilde{\mathcal{O}}\left(\sum_{s,h \leq H} q^\star(s,h) \left(\sqrt{\sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \pi_k(a|s,h) \frac{L'^2}{\overline{x}_k(s,a,h)+\theta}} + \frac{L'}{\theta}\right)\right) \\
&\leq \sum_{s,h \leq H} q^\star(s,h) \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \pi_k(a|s,h) \frac{\theta L'}{\overline{x}_k(s,a,h)+\theta} + \tilde{\mathcal{O}}\left(\frac{T_\star L'}{\theta}\right). \qquad \text{(AM-GM inequality)}
\end{aligned}$$

Summing these two terms, we have:

$$\textsc{Bias}_1 \leq \sum_{k=1}^{K} \sum_{s,h \leq H} q^\star(s,h) \sum_{a \in \mathcal{A}} \pi_k(a|s,h) \frac{L'(\overline{x}_k(s,a,h) - \underline{x}_k(s,a,h)) + 2\theta L'}{\overline{x}_k(s,a,h) + \theta} + \tilde{\mathcal{O}}\left(\frac{T_\star L'}{\theta}\right).$$

This completes the proof. ■

Lemma 43 *With probability at least* $1-\delta$, $\textsc{Bias}_2 = \tilde{\mathcal{O}}\left(T_\star L'/\theta\right)$.

Proof By Lemma 44 with $Z_k(s,a,h) = G_{k,s,a,h}/L'$ and $\mathbb{E}_k[G_{k,s,a,h}] = x_k(s,a,h)\bar{Q}_k(s,a,h)$, we have with probability at least $1-\delta$:

$$\sum_{k=1}^{K} \widetilde{Q}_k(s,a,h) - \bar{Q}_k(s,a,h) = \tilde{\mathcal{O}}\left(L'/\theta\right),$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}, h \leq H$. Therefore,

$$\textsc{Bias}_2 = \sum_{k=1}^{K} \sum_{s,h \leq H} q^\star(s,h) \sum_{a \in \mathcal{A}} \pi^\star(a|s,h)(\widetilde{Q}_k(s,a,h) - Q_k(s,a,h)) = \tilde{\mathcal{O}}\left(T_\star L'/\theta\right).$$

■

Lemma 44 *For any random variable $Z_k(s,a,h)$ depending on interaction before episode k such that $Z_k(s,a,h) \in [0,1]$, $\mathbb{E}_k[Z_k(s,a,h)] = z_k(s,a,h) \le x_k(s,a,h)$, we have with probability at least $1-\delta$:*

$$\sum_{k=1}^{K}\left(\frac{Z_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta} - \frac{z_k(s,a,h)}{\overline{x}_k(s,a,h)}\right) \le \frac{\ln\frac{1}{\delta}}{2\theta}.$$

Proof The statement is clearly true when $x_k(s,a,h) = 0$. When $x_k(s,a,h) > 0$, we also have $\overline{x}_k(s,a,h) > 0$. By $\frac{z}{1+z/2} \le \ln(1+z)$ for $z \ge 0$, we have:

$$\begin{aligned}\frac{2\theta Z_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta} &\le \frac{2\theta Z_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta Z_k(s,a,h)} = \frac{2\theta Z_k(s,a,h)/\overline{x}_k(s,a,h)}{1+\theta Z_k(s,a,h)/\overline{x}_k(s,a,h)} \\ &\le \ln\left(1+2\theta Z_k(s,a,h)/\overline{x}_k(s,a,h)\right).\end{aligned}$$

This gives

$$\begin{aligned}\mathbb{E}_k\left[\exp\left(\frac{2\theta Z_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta}\right)\right] &\le \mathbb{E}_k\left[1+\frac{2\theta Z_k(s,a,h)}{\overline{x}_k(s,a,h)}\right] = 1+\frac{2\theta z_k(s,a,h)}{\overline{x}_k(s,a,h)} \\ &\le \exp(2\theta z_k(s,a,h)/\overline{x}_k(s,a,h)). \qquad (1+z\le e^z)\end{aligned}$$

Therefore, by Markov inequality,

$$\begin{aligned}&P\left(\sum_{k=1}^{K}\frac{2\theta Z_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta} - \frac{2\theta z_k(s,a,h)}{\overline{x}_k(s,a,h)} > \ln\frac{1}{\delta}\right) \\ &\le \delta\cdot\mathbb{E}\left[\exp\left(\sum_{k=1}^{K}\frac{2\theta Z_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta} - \frac{2\theta z_k(s,a,h)}{\overline{x}_k(s,a,h)}\right)\right] \le \delta.\end{aligned}$$

Thus, with probability at least $1-\delta$, $\sum_{k=1}^{K}\frac{Z_k(s,a,h)}{\overline{x}_k(s,a,h)+\theta} - \frac{z_k(s,a,h)}{\overline{x}_k(s,a,h)} \le \frac{\ln\frac{1}{\delta}}{2\theta}$. ■

D.3. Dilated Bonus in SDA

Below we present lemmas related to dilated bonus in $\mathring{\mathcal{M}}$. We first show that a form of dilated value function is well-defined.

Lemma 45 *For some policy π in $\mathring{\mathcal{M}}$, transition $P \in \Lambda_{\mathcal{M}}$, and bonus function $b : \mathcal{S}\times\mathcal{A}\times[H] \to [0,\rho]$ for some $\rho > 0$, define $B(s,a,h) = b(s,a,h) + \left(1+\frac{1}{H'}\right)P_{s,a,h}B$, $B(s,h) = \sum_a \pi(a|s,h)B(s,a,h)$ and $B(g) = B(s,a,H+1) = 0$. Then, $\max_{s,a} B(s,a,h) \le \frac{15\rho(H-h+1)}{1-\gamma}$.*

Proof Define $\gamma' = (1+\frac{1}{H'})\gamma$ and recall that $H' = \frac{8(H+1)\ln(2K)}{1-\gamma}$. Now note that $\frac{1}{1-\gamma'} \le \frac{1+\frac{1}{H}}{1-\gamma}$ by simple algebra. Finally, define $\bar{b}(s,a,h) = \left(1+\frac{1}{H'}\right)\langle P_{s,a,h}(\cdot,h+1), B(\cdot,h+1)\rangle$ for $h \le H$, and $P'_{s,a,h}(s') = (1+\frac{1}{H'})P_{s,a,h}(s',h)$.

We prove that B is well defined and the statement holds by induction on $h = H+1, \dots, 1$. The base case is true by definition $B(s,a,H+1) = 0$. For $h \le H$ we have:

$$\begin{aligned} B(s,a,h) &= b(s,a,h) + \left(1 + \frac{1}{H'}\right)\left(\langle P_{s,a,h}(\cdot,h), B(\cdot,h)\rangle + \langle P_{s,a,h}(\cdot,h+1), B(\cdot,h+1)\rangle\right) \\ &= b(s,a,h) + \bar{b}(s,a,h) + P'_{s,a,h} B(\cdot,h). \end{aligned}$$

Therefore, $B(\cdot,\cdot,h)$ can be treated as the action-value function in an SSP with cost $(b+\bar{b})(\cdot,\cdot,h)$ and transition function P' (thus well defined). By $\sum_{s'} P'_{s,a,h}(s',h) \le \gamma'$, we have the expected hitting time of any policy starting from any state in an SSP with transition P' is upper bounded by $\frac{1}{1-\gamma'} \le \frac{1+\frac{1}{H}}{1-\gamma}$. Let $R(h) = \max_{s,a} B(s,a,h)$ and note that $R(H+1) = 0$. Since $b(s,a,h) \le \rho$ and $\bar{b}(s,a,h) \le \left(1+\frac{1}{H'}\right)(1-\gamma)R(h+1)$ by $\sum_{s'} P_{s,a,h}(s',h+1) \le 1-\gamma$, we have:

$$\begin{aligned} R(h) &\le \frac{\rho + \left(1+\frac{1}{H'}\right)(1-\gamma)R(h+1)}{1-\gamma'} \le \frac{\rho}{1-\gamma'} + \left(1+\frac{1}{H'}\right)\left(1+\frac{1}{H}\right)R(h+1) \\ &\le \frac{\rho\left(1+\frac{1}{H}\right)}{1-\gamma} + \left(1+\frac{1}{H'}\right)\left(1+\frac{1}{H}\right)R(h+1), \end{aligned}$$

where the two last inequalities follow because $\frac{1}{1-\gamma'} \le \frac{1+\frac{1}{H}}{1-\gamma}$. The proof is now finished by solving the recursion and obtaining:

$$R(h) \le \frac{\rho}{1-\gamma} \sum_{i=0}^{H-h} \left(1+\frac{1}{H'}\right)^i \left(1+\frac{1}{H}\right)^{i+1},$$

which implies that $R(h) \le \frac{15\rho(H-h+1)}{1-\gamma}$ since $(1+\frac{1}{H})^{H+1}(1+\frac{1}{H'})^H \le 2e^2 \le 15$. ■

Lemma 46 *Let π be a policy in $\mathring{\mathcal{M}}$ and b be a non-negative cost function in $\mathring{\mathcal{M}}$ such that $b(s,a,H+1) = 0$ and $b(s,a,h) \le \rho$. Moreover, let $\widehat{P} \in \Lambda_{\mathcal{M}}$ be an optimistic transition so that*

$$B(s,a,h) = b(s,a,h) + \left(1+\frac{1}{H'}\right)\widehat{P}_{s,a,h} B \ge b(s,a,h) + \left(1+\frac{1}{H'}\right) P_{s,a,h} B,$$

where $B(s,h) = \sum_{a\in\mathcal{A}} \pi(a|s,h) B(s,a,h)$ and $B(g) = B(s,H+1) = 0$. Then,

$$\begin{aligned} &\sum_{s,h} q^\star(s,h) \sum_{a\in\mathcal{A}} (\pi(a|s,h) - \mathring{\pi}^\star(a|s,h)) B(s,a,h) + \frac{1}{H'}\sum_{s,h} q^\star(s,h) B(s,h) \\ &\qquad + \sum_{s,a,h} q^\star(s,a,h) b(s,a,h) \le 3V^{\pi,\widehat{P},b}(s_{\text{init}},1) + \tilde{\mathcal{O}}\left(\frac{H\rho}{K(1-\gamma)}\right). \end{aligned}$$

Proof By the optimism property of $\widehat{P}$, we have:

$$\sum_{s,h} q^\star(s,h) \sum_{a\in\mathcal{A}} \left(\pi(a|s,h) - \mathring{\pi}^\star(a|s,h)\right) B(s,a,h)$$

$$
\begin{aligned}
&+ \frac{1}{H'} \sum_{s,h} q^\star(s,h) \sum_{a \in \mathcal{A}} \pi(a|s,h) B(s,a,h) + \sum_{s,a,h} q^\star(s,a,h) b(s,a,h) \\
&\leq \left(1 + \frac{1}{H'}\right) \sum_{s,h} q^\star(s,h) \sum_{a \in \mathcal{A}} \pi(a|s,h) B(s,a,h) + \sum_{s,a,h} q^\star(s,a,h) b(s,a,h) \\
&\quad - \sum_{s,a,h} q^\star(s,a,h) \left(b(s,a,h) + \left(1 + \frac{1}{H'}\right) \sum_{s',h'} P_{s,a,h}(s',h') B(s',h') \right) \\
&= \left(1 + \frac{1}{H'}\right) \sum_{s',h'} \left(q^\star(s',h') - \sum_{s,a,h} q^\star(s,a,h) P_{s,a,h}(s',h') \right) B(s',h') \\
&= \left(1 + \frac{1}{H'}\right) B(s_{\text{init}}, 1). \qquad (18)
\end{aligned}
$$

The last relation is by $q^\star(s,h) - \sum_{s',a',h'} q^\star(s',a',h') P_{s',a',h'}(s,h) = \mathbb{I}\{(s,h) = (s_{\text{init}},1)\}$ (see (Rosenberg and Mansour, 2021, Appendix B.1)).

Let J be the number of steps until the goal state g is reached in $\mathring{\mathcal{M}}$, and $n = \frac{8H}{1-\gamma} \ln(2K)$. Now note that for any policy, the expected hitting time in an SSP with transition $\widehat{P}$ is upper bounded by $\frac{H}{1-\gamma} + 1$ by $\widehat{P} \in \Lambda_{\mathcal{M}}$. Therefore, by Lemma 31, $P(J \geq n) \leq \frac{1}{K}$, and

$$
\begin{aligned}
B(s,h) &= \mathbb{E}\left[\sum_{t=1}^{J} \left(1 + \frac{1}{H'}\right)^{t-1} b(s_t,a_t,h_t) \middle| \pi, \widehat{P}, (s_1,h_1) = (s,h) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \left(1 + \frac{1}{H'}\right)^{t-1} b(s_t,a_t,h_t) + \left(1 + \frac{1}{H'}\right)^{n} B(s_{t+1},h_{t+1}) \middle| \pi, \widehat{P}, (s_1,h_1) = (s,h) \right] \\
&\leq \left(1 + \frac{1}{H'}\right)^{n-1} V^{\pi,\widehat{P},b}(s,h) + \tilde{\mathcal{O}}\left(\frac{H\rho}{K(1-\gamma)}\right). \qquad \text{(Lemma 45)}
\end{aligned}
$$

Plugging this back into Eq. (18) and by $(1 + 1/H')^n \leq e < 3$, we get the desired result. ■

D.4. Computation of B_k

We study an operator on value function, from which B_k can be computed as a fixed point. For any policy π, cost function c, transition confidence set $\mathcal{P} \subseteq \Lambda_{\mathcal{M}}$, and interest factor $\rho \geq 0$, we define the dilated Bellman operator $\mathcal{T}_\rho$ that maps any value function $V : \mathring{\mathcal{S}}_+ \to \mathbb{R}_+$ to another value function $\mathcal{T}_\rho V : \mathring{\mathcal{S}}_+ \to \mathbb{R}_+$, such that:

$$
\begin{aligned}
&(\mathcal{T}_\rho V)(s,h) = \sum_a \pi(a|s,h) \left(c(s,a,h) + (1+\rho) \max_{P \in \mathcal{P}} P_{s,a,h} V \right), \\
&(\mathcal{T}_\rho V)(g) = 0, \ (\mathcal{T}_\rho V)(s,H+1) = \max_a c(s,a,H+1). \qquad (19)
\end{aligned}
$$

In this work, we have $\mathcal{P} \in \{\mathcal{P}_k\}_{k=1}^K$, and $\mathcal{P}_k = \bigcap_{s,a,h} \mathcal{P}_{k,s,a,h}$, where $\mathcal{P}_{k,s,a,h}$ is a convex set that specifies constraints on $((s,h),a)$. In other words, $\mathcal{P}_k$ is a product of constraints on each $((s,h),a)$

(note that $\Lambda_{\mathcal{M}}$ can also be decomposed into shared constraints on $P_{s,a,H+1}$ and independent constraints on each $s, a, h \le H$). Thus, there exists $P' \in \mathcal{P}$ that satisfies $P' = \operatorname{argmax}_{P\in\mathcal{P}} P_{s,a,h}V$ in Eq. (19) for all $((s,h),a)$ simultaneously. Moreover, finding such P' can be done by linear programming for each $((s,h),a)$ independently. Now we show that iteratively applying $\mathcal{T}_\rho$ to some initial value function converges to a fixed point sufficiently fast.

Lemma 47 *Define value function $V^0 : \mathring{\mathcal{S}}_+ \to \mathbb{R}_+$ such that $V^0(s,h) = V^0(g) = 0$ for any $(s,h) \in \mathcal{S}\times[H]$ and $V^0(s,H+1) = \max_a c(s,a,H+1)$. Then for any $\rho \ge 0$ such that $\gamma' = (1+\rho)\gamma < 1$, the limit $V_\rho = \lim_{n\to\infty} \mathcal{T}_\rho^n V^0$ exists. Moreover, when $n \ge Hl$ with $l = \lceil \frac{\ln\frac{1}{\epsilon}}{1-\gamma'} \rceil$ for some $\epsilon > 0$, we have $\left\|\mathcal{T}_\rho^n V^0 - V_\rho\right\|_\infty \le H\left(\frac{(1+\rho)(1-\gamma)}{1-\gamma'}\right)^{H-1}\kappa\epsilon$, where $\kappa = \sum_{j=0}^{H-1}(\frac{(1+\rho)(1-\gamma)}{1-\gamma'})^j \frac{\|c\|_\infty}{1-\gamma'}$.*

Proof Define a sequence of value functions $\{V^i\}_{i=0}^\infty$ such that $V^{i+1} = \mathcal{T}_\rho V^i$. We first show that $\left\|V^i(\cdot,h)\right\|_\infty \le \sum_{j=0}^{H-h}(\frac{(1+\rho)(1-\gamma)}{1-\gamma'})^j \frac{\|c\|_\infty}{1-\gamma'}$ for $i \ge 0$ and $h \le H$. We prove this by induction on i. Note that this is clearly true when $i = 0$. For $i > 0$, by $\mathcal{P} \subseteq \Lambda_{\mathcal{M}}$ and Eq. (19), we have:

$$\begin{aligned}
V^i(s,h) = (\mathcal{T}_\rho V^{i-1})(s,h) &\le \|c\|_\infty + \gamma' \left\|V^{i-1}(\cdot,h)\right\|_\infty + (1+\rho)(1-\gamma)\left\|V^{i-1}(\cdot,h+1)\right\|_\infty \\
&\le \|c\|_\infty + \gamma' \sum_{j=0}^{H-h}\left(\frac{(1+\rho)(1-\gamma)}{1-\gamma'}\right)^j \frac{\|c\|_\infty}{1-\gamma'} + \sum_{j=1}^{H-h}\left(\frac{(1+\rho)(1-\gamma)}{1-\gamma'}\right)^j \|c\|_\infty \\
&\le \sum_{j=0}^{H-h}\left(\frac{(1+\rho)(1-\gamma)}{1-\gamma'}\right)^j \frac{\|c\|_\infty}{1-\gamma'}.
\end{aligned}$$

Therefore, $\left\|V^i\right\|_\infty \le \kappa$. We now show that $\{V^i\}_i$ converges to a fixed point. Specifically, we show that for some $\epsilon > 0$ and any $i, j \in \mathbb{N}$, when $n \ge (H-h+1)l$, we have $\left\|(\mathcal{T}_\rho^n V^i)(\cdot,h) - (\mathcal{T}_\rho^n V^j)(\cdot,h)\right\|_\infty \le (H-h+1)(\frac{(1+\rho)(1-\gamma)}{1-\gamma'})^{H-h}\kappa\epsilon$ (note that $\frac{(1+\rho)(1-\gamma)}{1-\gamma'} > 1$). Therefore, when $n \ge Hl$, we have $\left\|\mathcal{T}_\rho^n V^i - \mathcal{T}_\rho^n V^j\right\|_\infty \le H(\frac{(1+\rho)(1-\gamma)}{1-\gamma'})^{H-1}\kappa\epsilon$. Setting $\epsilon \to 0$, the statement above implies that for any $\mathring{s} \in \mathring{\mathcal{S}}$, $\{V^i(\mathring{s})\}_{i=1}^\infty$ is a Cauchy sequence and thus converges. Moreover, letting $j \to \infty$ implies that $\{V^i\}_i$ converges to V_ρ with the rate shown above. We prove the statement above by induction on $h = H, \ldots, 1$. First note that for any $s \in \mathcal{S}, h \in [H]$:

$$\begin{aligned}
&\left|(\mathcal{T}_\rho V^i)(s,h) - (\mathcal{T}_\rho V^j)(s,h)\right| = (1+\rho)\left|\sum_a \pi(a|s,h)\left(\max_{P\in\mathcal{P}} P_{s,a,h}V^i - \max_{P\in\mathcal{P}} P_{s,a,h}V^j\right)\right| \\
&\le (1+\rho)\sum_a \pi(a|s,h)\max_{P\in\mathcal{P}}\left|P_{s,a,h}(V^i - V^j)\right| \\
&\le \gamma'\left\|V^i(\cdot,h) - V^j(\cdot,h)\right\|_\infty + (1+\rho)(1-\gamma)\left\|V^i(\cdot,h+1) - V^j(\cdot,h+1)\right\|_\infty, \qquad (20)
\end{aligned}$$

where the last inequality is by $\sum_{s'} P_{s,a,h}(s',h) \le \gamma$, $\sum_{s'} P_{s,a,h}(s',h+1) \le 1-\gamma$, and $P_{s,a,h}(s',h') = 0$ for $h' \notin \{h, h+1\}$, for any $P \in \Lambda_{\mathcal{M}}$. Now for the base case $h = H$, Eq. (20) implies $\left\|(\mathcal{T}_\rho V^i)(\cdot,H) - (\mathcal{T}_\rho V^j)(\cdot,H)\right\|_\infty \le \gamma'\left\|V^i(\cdot,H) - V^j(\cdot,H)\right\|_\infty$. Thus for $n \ge l$, $\left\|(\mathcal{T}_\rho^n V^i)(\cdot,H) - (\mathcal{T}_\rho^n V^j)(\cdot,H)\right\|_\infty \le \gamma'^n \cdot \kappa \le \kappa\epsilon$. For the induction step $h < H$, if $n \ge (H-h+1)l$, then Eq. (20) implies:

$$\left|(\mathcal{T}_\rho^n V^i)(s,h) - (\mathcal{T}_\rho^n V^j)(s,h)\right|$$

$$\leq \gamma'^{l} \left\| (\mathcal{T}_\rho^{n-l} V^i)(s,h) - (\mathcal{T}_\rho^{n-l} V^j)(s,h) \right\|_\infty + (1-\gamma') \left(\frac{(1+\rho)(1-\gamma)}{1-\gamma'} \right)^{H-h} \sum_{i=0}^{l-1} \gamma'^{i} (H-h) \kappa\epsilon$$

(by the induction assumption)

$$\leq (H-h+1) \left(\frac{(1+\rho)(1-\gamma)}{1-\gamma'} \right)^{H-h} \kappa\epsilon.$$

This completes the proof of the statement above. ■

Now note that B_k is a fixed point of $\mathcal{T}_\rho$ with $\pi = \pi_k$, $\mathcal{P} = \mathcal{P}_k$, $c = b_k$, and $\rho = 1/H'$. Thus, B_k can be approximated efficiently.

D.5. Computation of $\overline{x}_k$ and $\underline{x}_k$

Note that $\overline{x}_k(s,a,h)$ can be computed by solving the following linear program (it is straightforward to verify that the constraints on π_q and P_q are linear):

$$\begin{aligned}
&\max_{q \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}\times[H]\times\mathring{\mathcal{S}}_+}} \sum_{\mathring{s}' \in \mathring{\mathcal{S}}_+} q(s,a,h,\mathring{s}') \\
\text{s.t.} \quad &\sum_{a' \in \mathcal{A}, \mathring{s}' \in \mathcal{S}_+} q(s',a',h',\mathring{s}') \\
&\quad - \sum_{(s'',h'') \in \mathring{\mathcal{S}}, a'' \in \mathcal{A}} q(s'',a'',h'',(s',h')) = \mathbb{I}\{(s',h') = (s_{\text{init}},1)\}, \forall (s',h') \\
&\pi_q = \pi_k, \quad P_q \in \bigcap_{(s',a',h') \in (\mathcal{S}\times\mathcal{A}\times[H]) \setminus \{(s,a,h)\}} \mathcal{P}_{k,s',a',h'}, \quad P_{q,s,a,h}(g) = 1
\end{aligned}$$

That is, we try to compute the occupancy measure that maximizes the number of visits to (s,a,h) in an augmented MDP, where the transition lies in $\mathcal{P}_k$ except that taking action a at state (s,h) directly transits to the goal state (so that the number of visits to (s,a,h) is at most 1 and the occupancy measure at (s,a,h) is the probability of visiting (s,a,h)). The computation of $\underline{x}_k(s,a,h)$ is similar. Thus, both $\overline{x}_k$ and $\underline{x}_k$ can be computed efficiently (in a weakly polynomial time).

Appendix E. Learning without Some Parameters

In this section, we discuss the achievable regret guarantee without knowing some of the parameters assumed to be known. For simplicity, we only describe the high level ideas. We first describe the general ideas of dealing with each parameter being unknown, which are applicable under all types of feedback.

- **Unknown D and unknown fast policy**: we can simply follow the ideas in (Chen and Luo, 2021) to estimate D and fast policy. For unknown fast policy, we maintain an instance of Bernstein-SSP (Cohen et al., 2020) $\mathcal{B}_f$. When we need to switch to the fast policy, we simply involve $\mathcal{B}_f$ as if this is a new episode for this algorithm, follow its decision until reaching g, and always feed cost 1 for all state-action pairs. Following the arguments in (Chen and Luo, 2021, Lemma 1), the scheme above only incurs constant extra regret. For unknown D, we maintain an estimate of it and update the algorithm's parameters whenever the estimate

is updated. Specifically, we separate the state space into known states and unknown states. A state is known if the number of visits to it is more than some threshold, and it is unknown otherwise. Whenever the learner visits an unknown state, it involves a Bernstein-SSP instance to approximate the behavior of fast policy until reaching g. When an unknown state s becomes known, we update the diameter estimate by incorporating an estimate of $T^{\pi^f}(s)$, and then updates the algorithm's parameters with respect to the new estimate. In terms of regret, this approach does not affect the transition estimation error, but brings an extra $\sqrt{S}$ factor in the regret from policy optimization due to at most S updates to the algorithm's parameters.

- **Unknown** $B_\star$: We can estimate $B_\star$ following the procedure in (Cohen et al., 2021, Appendix C). The main idea is pretty similar to the unknown D case: we again maintain an estimate of $B_\star$ and separate states into known states and unknown states based on how many times a state has been visited. The learner updates algorithm's parameters whenever the estimate of $B_\star$ is updated. Similarly, this approach brings an extra $\sqrt{S}$ factor in the regret from policy optimization.

- **Unknown** $T_\star$: We can replace $T_\star$ in parameters by $B_\star/c_{\min}$ in stochastic costs setting and $D/c_{\min}$ in other settings since $T_\star \le B_\star/c_{\min}$ (or $T_\star \le D/c_{\min}$). How to estimate D or $B_\star$ is discussed above.

- **Unknown** $T_{\max}$: Similar to (Chen and Luo, 2021), we simply replace $T_{\max}$ in parameters by K^p for some $p \in (0, \frac{1}{2})$.

Next, we describe under each setting, what regret guarantee we can achieve with each parameter being unknown by applying the corresponding method above.

Stochastic Costs In this setting, we need the knowledge of D, $B_\star$ and $T_{\max}$.

- **Unknown** D: Since the regret from policy optimization is a lower order term, the dominating term of the final regret remains to be $\tilde{\mathcal{O}}(B_\star S\sqrt{AK})$.

- **Unknown** $B_\star$: Since the regret from policy optimization is a lower order term, the dominating term of the final regret remains to be $\tilde{\mathcal{O}}(B_\star S\sqrt{AK})$.

- **Unknown** $T_{\max}$: We replace $T_{\max}$ in parameters by $K^{1/12}$. If $K^{1/12} \le T_{\max}$, then clearly the regret is of order $\tilde{\mathcal{O}}(LK) = \tilde{\mathcal{O}}(T_{\max}^{13})$. Otherwise, by Theorem 5 we have $R_K = \tilde{\mathcal{O}}(B_\star S\sqrt{AK} + S^4A^{2.5}K^{1/3})$.

Stochastic Adversary In this setting, we need the knowledge of D, $T_\star$, and $T_{\max}$. We consider the following cases:

- **Unknown** D: Since the regret from policy optimization is a lower order term, the dominating term of the final regret remains to be $\tilde{\mathcal{O}}(\sqrt{DT_\star K} + DS\sqrt{AK})$ in the full information setting, and $\tilde{\mathcal{O}}(\sqrt{DT_\star SAK} + DS\sqrt{AK})$ in the bandit feedback setting.

- **Unknown** $T_\star$: Ignoring the lower order terms, we have $R_K = \tilde{\mathcal{O}}(D\sqrt{K/c_{\min}} + DS\sqrt{AK})$ in the full information setting by Theorem 6, and $R_K = \tilde{\mathcal{O}}(D\sqrt{SAK/c_{\min}} + DS\sqrt{AK})$ in the bandit feedback setting by Theorem 7.

- **Unknown** $T_{\max}$: We replace $T_{\max}$ in parameters by $K^{1/13}$. If $K^{1/13} \le T_{\max}$, then R_K is of order $\tilde{\mathcal{O}}(LK) = \tilde{\mathcal{O}}(T_{\max}^{14})$. Otherwise, we have $R_K = \tilde{\mathcal{O}}(\sqrt{DT_\star K} + DS\sqrt{AK} + (S^2A^3)^{1/4}K^{25/52} + S^4A^{2.5}K^{4/13})$ in the full information setting by Theorem 6, and $R_K = \tilde{\mathcal{O}}(\sqrt{SADT_\star K} + DS\sqrt{AK} + SA^{5/4}K^{25/52} + S^4A^{2.5}K^{4/13})$ in the bandit feedback setting by Theorem 7.

Adversarial Costs, Full Information In this setting, we need the knowledge of D, $T_\star$, and $T_{\max}$. We consider the following cases:

- **Unknown** D: With an extra $\sqrt{S}$ factor in the policy optimization term, we have $R_K = \tilde{\mathcal{O}}(T_\star\sqrt{SDK} + \sqrt{S^2ADT_\star K})$ ignoring the lower order terms.
- **Unknown** $T_\star$: Ignoring the lower order terms, we have $R_K = \tilde{\mathcal{O}}(\frac{D^{1.5}}{c_{\min}}\sqrt{K} + D\sqrt{S^2AK/c_{\min}})$.
- **Unknown** $T_{\max}$: We replace $T_{\max}$ in parameters by $K^{1/11}$. If $K^{1/11} \le T_{\max}$, then clearly the regret is of order $\tilde{\mathcal{O}}(LK) = \tilde{\mathcal{O}}(T_{\max}^{12})$. Otherwise, by Theorem 9 we have $R_K = \tilde{\mathcal{O}}(T_\star\sqrt{DK} + \sqrt{S^2ADT_\star K} + S^4A^2K^{5/11})$.

Adversarial Costs, Bandit Feedback In this setting, we need the knowledge of D and $T_{\max}$. We consider the following cases:

- **Unknown** D: Tracing the proof of Theorem 10, the regret from policy optimization is of order $\tilde{\mathcal{O}}(\sqrt{SAT_{\max}^4 K})$. With an extra $\sqrt{S}$ factor in the policy optimization term, we still have $R_K = \tilde{\mathcal{O}}(\sqrt{S^2AT_{\max}^5 K})$ ignoring the lower order terms.
- **Unknown** $T_{\max}$: We replace $T_{\max}$ in parameters by K^p for any $p \in (0, \frac{1}{5})$. If $K^p \le T_{\max}$, then clearly the regret is of order $\tilde{\mathcal{O}}(LK) = \tilde{\mathcal{O}}(T_{\max}^{1+1/p})$. Otherwise, by Theorem 10 we have $R_K = \tilde{\mathcal{O}}(\sqrt{S^2AK^{1+5p}} + S^{5.5}A^{3.5}K^{5p})$.

Appendix F. Auxiliary Lemma

Lemma 48 *If $x \le (a\sqrt{x} + b)\ln^p(cx)$ for some $a, b, c > 0$ and absolute constant $p \ge 0$, then $x = \tilde{\mathcal{O}}(a^2 + b)$. Specifically, $x \le a\sqrt{x} + b$ implies $x \le (a + \sqrt{b})^2 \le 2a^2 + 2b$.*

Lemma 49 *(Luo et al., 2021, Lemma A.4) Let $\eta > 0$, $\pi_k \in \Delta(A)$, and $\ell_k \in \mathbb{R}^A$ satisfy the following for all $k \in [K]$ and $a \in \mathcal{A}$:*

$$\pi_1(a) = \frac{1}{A}, \quad \pi_{k+1}(a) \propto \pi_k(a)\exp(-\eta\ell_k(a)), \quad |\eta\ell_k(a)| \le 1.$$

Then for any $\pi^\star \in \Delta(A)$, $\sum_{k=1}^K \langle \pi_k - \pi^\star, \ell_k\rangle \le \frac{\ln A}{\eta} + \eta\sum_{k=1}^K \sum_{a\in\mathcal{A}} \pi_k(a)\ell_k^2(a)$.

Lemma 50 *(Chen et al., 2021b, Lemma 38) Let $\{X_i\}_{i=1}^\infty$ be a martingale difference sequence adapted to the filtration $\{\mathcal{F}_i\}_{i=0}^\infty$ and $|X_i| \le B$ for some $B > 0$. Then with probability at least $1-\delta$, for all $n \ge 1$ simultaneously,*

$$\left|\sum_{i=1}^n X_i\right| \le 3\sqrt{\sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_{i-1}]\ln\frac{4B^2n^3}{\delta}} + 2B\ln\frac{4B^2n^3}{\delta}.$$

Lemma 51 *(Cohen et al., 2020, Theorem D.3) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d random variables with expectation μ and $X_n \in [0, B]$ almost surely. Then with probability at least $1-\delta$, for any $n \geq 1$:*

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

Lemma 52 *(Cohen et al., 2020, Lemma D.4) and (Cohen et al., 2021, Lemma C.2) Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables w.r.t to the filtration $\{\mathcal{F}_i\}_{i=0}^{\infty}$ and $X_i \in [0, B]$ almost surely. Then with probability at least $1-\delta$, for all $n \geq 1$ simultaneously:*

$$\sum_{i=1}^{n}\mathbb{E}[X_i|\mathcal{F}_{i-1}] \leq 2\sum_{i=1}^{n}X_i + 4B\ln\frac{4n}{\delta},$$

$$\sum_{i=1}^{n}X_i \leq 2\sum_{i=1}^{n}\mathbb{E}[X_i|\mathcal{F}_{i-1}] + 8B\ln\frac{4n}{\delta}.$$