Improved Regret Bound for Safe Reinforcement Learning via Tighter Cost Pessimism and Reward Optimism

Kihyun Yu[1] khyu99@kaist.ac.kr
Duksang Lee[1] duksang@kaist.ac.kr
William Overman[2] wpo@stanford.edu
Dabeen Lee[1,†] dabeenl@kaist.ac.kr

[1] *Department of Industrial and Systems Engineering, KAIST, Daejeon 34141, South Korea*
[2] *Graduate School of Business, Stanford University, Stanford, CA 94305, United States*
[†] *Corresponding author*

Abstract

This paper studies the safe reinforcement learning problem formulated as an episodic finite-horizon tabular constrained Markov decision process with an unknown transition kernel and stochastic reward and cost functions. We propose a model-based algorithm based on novel cost and reward function estimators that provide tighter cost pessimism and reward optimism. While guaranteeing no constraint violation in every episode, our algorithm achieves a regret upper bound of $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^{2.5}S\sqrt{AK})$ where $\bar{C}$ is the cost budget for an episode, $\bar{C}_b$ is the expected cost under a safe baseline policy over an episode, H is the horizon, and S, A and K are the number of states, actions, and episodes, respectively. This improves upon the best-known regret upper bound, and when $\bar{C}-\bar{C}_b=\Omega(H)$, it nearly matches the regret lower bound of $\Omega(H^{1.5}\sqrt{SAK})$. We deduce our cost and reward function estimators via a Bellman-type law of total variance to obtain tight bounds on the expected sum of the variances of value function estimates. This leads to a tighter dependence on the horizon in the function estimators. We also present numerical results to demonstrate the computational effectiveness of our proposed framework.

1 Introduction

Safe reinforcement learning (RL) aims to learn a policy that maximizes the cumulative reward and, at the same time, ensures that some safety requirements are satisfied during the learning process. Safe RL provides modeling frameworks for many practical scenarios where violating a safety constraint results in a critical situation. For example, it is crucial to enforce collision avoidance for autonomous driving (Isele et al., 2018; Krasowski et al., 2020) and robotics (Fisac et al., 2018; García and Shafie, 2020). For financial planning, there exist legal and business regulations (Abe et al., 2010). For healthcare systems, service providers consider restrictions due to patients' conditions (Coronato et al., 2020).

The standard approach is to formulate a safe RL problem as a constrained Markov decision process (CMDP), where the objective is to maximize the expected reward over a time horizon while there is a constraint that the expected cost should be under budget (Altman, 1999). The presence of constraints, however, brings about challenges in developing solution methods for CMDPs. The Bellman optimality principle does not hold for CMDPs, and as a consequence, backward induction and the greedy operator cannot be directly applied to CMDPs (Altman, 1999). This makes online learning of CMDPs difficult, and we need significantly different frameworks and algorithms compared to the unconstrained setting (Efroni et al., 2020).

The first direction for online reinforcement learning of CMDPs is to consider *cumulative (or soft) constraint violation*, which sums up the constraint violations across episodes (Efroni et al., 2020). Here, the constraint violation in an episode is defined as the expected cost minus the budget. Then a policy can have a negative constraint violation, which means that a positive violation in one episode can be canceled out by a negative violation in another episode in the sum. This cancellation effect allows oscillating between such two cases, while still achieving zero cumulative constraint violation. This phenomenon can indeed be observed in practice (Stooke et al., 2020; Moskovitz et al., 2023).

The second direction attempts to remedy the issue of error cancellation with the notion of *hard constraint violation* (Efroni et al., 2020). It ignores episodes with a negative violation and takes the sum of only the positive constraint violations. Efroni et al. (2020) developed OptCMDP and its efficient variant, OptCMDP-bonus, that attain a regret upper bound and a hard constraint violation of $\widetilde{\mathcal{O}}(H^2\sqrt{S^2AK})$. Recently, Ghosh et al. (2024) proposed a model-free algorithm with the same asymptotic guarantees. However, as in the first setting, the algorithms cannot avoid episodes in which the constraint is violated. Thus, they are still not suitable for the aforementioned applications, where even a single incidence of violation can cause substantial problems.

The third approach seeks *zero (hard) constraint violation*, requiring that the constraint is satisfied in every episode (Simão et al., 2021). Satisfying constraints in the early stage is difficult when the model parameters, especially the transition kernel, are unknown. Simão et al. (2021) considered some abstraction of the transition model under which they showed an algorithm with no constraint violation, but no regret upper bound was presented. Then Liu et al. (2021) came up with the first algorithm, OptPess-LP, that achieves a sublinear regret with no constraint violation, assuming the knowledge of a *safe baseline policy*. Here, a safe baseline policy is a policy under which the expected cost is lower than the budget. OptPess-LP guarantees a regret upper bound of $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^3\sqrt{S^3AK})$ where $\bar{C}$ is the budget, $\bar{C}_b$ is the expected cost under the safe baseline policy, H is the length of the horizon, and S, A and K are the number of states, actions, and episodes, respectively. Bura et al. (2022) developed Doubly Optimistic Pessimistic Exploration (DOPE) with an improved regret upper bound of $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^3\sqrt{S^2AK})$. DOPE is based on designing tight optimistic reward function estimators (reward optimism) and conservative cost function estimators (cost pessimism).

While DOPE establishes a tight regret upper bound with no constraint violation, there is still room for improvement. The regret lower bound of $\Omega(H^{1.5}\sqrt{SAK})$ for the unconstrained case (Jin et al., 2018; Domingues et al., 2021) also works as a lower bound for the constrained setting because we may take trivial cost functions. However, even when $\bar{C}-\bar{C}_b=\Omega(H)$, the regret upper bound of DOPE is as low as $\widetilde{\mathcal{O}}(H^2\sqrt{S^2AK})$ which has a gap of $\widetilde{\mathcal{O}}(\sqrt{HS})$ from the lower bound. This naturally motivates the following question.

Is there an algorithm for learning CMDPs that guarantees no constraint violation during learning and achieves an improved regret upper bound?

Our Contributions We answer this question affirmatively with an algorithm that improves upon DOPE via tighter reward optimism and cost pessimism. Our results are summarized in Table 1 and as follows.

- Our algorithm, DOPE+, achieves a regret upper bound of $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^{2.5}\sqrt{S^2AK})$ and ensures no constraint violation in every episode, with the knowledge of a safe baseline policy. This improves upon the best-known regret upper bound $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^{3}\sqrt{S^2AK})$ attained by DOPE.

- When the gap $\bar{C}-\bar{C}_b$ between the budget and the expected cost under the safe baseline policy satisfies $\bar{C}-\bar{C}_b = \Omega(H)$, the regret upper bound becomes $\widetilde{\mathcal{O}}(H^{1.5}\sqrt{S^2AK})$. This nearly matches the regret lower bound $\Omega(H^{1.5}\sqrt{SAK})$, which shows that the regret upper bound achieves the optimal dependence on the horizon H.

- The improvement comes from our novel reward and cost function estimators with tighter reward optimism and cost pessimism. We deduce the function estimators by providing a tighter upper bound on the difference of value functions with respect to the true and estimated transition kernels. We analyze the difference based on the *value difference lemma* due to Dann et al. (2017). The key step is to apply a Bellman-type law of total variance to control the expected sum of the variance of value function estimates, inspired by Azar et al. (2017); Chen and Luo (2021).

Table 1: Comparison of Safe RL algorithms for the Hard Constraint Violation Setting: OptCMDP, OptCMDP-bonus (Efroni et al., 2020), AlwaysSafe (Simão et al., 2021), OptPess-LP (Liu et al., 2021), DOPE (Bura et al., 2022), and DOPE+ (Algorithm 1).

Algorithms	Regret	Hard Constraint Violation
OptCMDP, OptCMDP-bonus	$\widetilde{\mathcal{O}}(H^2\sqrt{S^2AK})$	$\widetilde{\mathcal{O}}(H^2\sqrt{S^2AK})$
AlwaysSafe	Unknown	0
OptPess-LP	$\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^3\sqrt{S^3AK})$	0
DOPE	$\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^3\sqrt{S^2AK})$	0
DOPE+	$\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^{2.5}\sqrt{S^2AK})$	0

DOPE by Bura et al. (2022) provides a framework for CMDPs to consider the error term in inferring the value function of a policy caused by the difference between the true transition kernel and an estimated one. They also applied the value difference lemma to analyze the difference between the value function under the true transition kernel and that under an estimated transition kernel. Then they take a naïve upper bound of H on any value function.

Instead, we take one step further to refine the analysis by considering the variance terms of value functions. We decompose the difference term and apply a Bellman-type law of total variance (Azar et al., 2017; Chen and Luo, 2021) to bound the expected sum of the variances of value function estimates. We explain this and how it leads to an improved regret bound in detail in Section 3.

A more comprehensive literature review on online reinforcement learning of CMDPs is given in the appendix.

2 Problem Setting

A finite-horizon tabular MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^{H-1}, p)$ where $\mathcal{S}$ is the finite state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the finite action space with $|\mathcal{A}| = A$, H is the finite-horizon, $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition kernel at step $h \in [H-1]$, and p is the known initial distribution of the states. Here, $P_h(s' \mid s, a)$ is the probability of transitioning to state s' from state s when the chosen action is a at step $h \in [H-1]$. Equivalently, we may define a single *non-stationary* transition kernel $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ with $P(s' \mid s, a, h) = P_h(s' \mid s, a)$ and $P(s' \mid s, a, H) = p(s')$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H-1]$. We assume that $\{P_h\}_{h=1}^{H-1}$ and thus P are *unknown*.

Before an episode begins, the agent prepares a *stochastic policy* $\pi : \mathcal{S} \times [H] \times \mathcal{A} \to [0,1]$ where $\pi(a \mid s, h)$ is the probability of taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at step h. Here, π can be viewed as a *non-stationary policy* as it may change over the horizon, and this is due to the non-stationarity of P over steps $h \in [H]$. Given a policy π_k for episode $k \in [K]$, the MDP proceeds with trajectory $\{s_h^{P,\pi_k}, a_h^{P,\pi_k}\}_{h \in [H]}$ generated by P.

The reward and cost functions are given by $f, g : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$, i.e., choosing action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$ and step $h \in [H]$ generates a reward $f(s,a,h)$ and cost $g(s,a,h)$. Here, functions f and g are non-stationary over $h \in [H]$. However, the agent observes the noisy reward and cost. We denote the observed noisy reward and cost for episode $k \in [K]$ by $f_k(s,a,h)$ and $g_k(s,a,h)$, respectively. As in Liu et al. (2021), we assume that $f_k(s,a,h)$ and $g_k(s,a,h)$ are determined by independent[1] noisy random variables $\zeta_k^f(s,a,h)$ and $\zeta_k^g(s,a,h)$ following a zero-mean 1/2-sub-Gaussian distribution, i.e., $f_k(s,a,h) = f(s,a,h) + \zeta_k^f(s,a,h)$ and $g_k(s,a,h) = g(s,a,h) + \zeta_k^g(s,a,h)$. We note that $1/2-$sub-Gaussian random variables ζ with zero mean satisfies $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}[\exp(\lambda\zeta)] \le \exp(\lambda^2/4)$. Then Hoeffding's inequality implies the following.

Lemma 2.1 *With probability at least $1 - 4\delta$, it holds that for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$,*

$$|f_k(s,a,h)|, |g_k(s,a,h)| \le 1 + \sqrt{\ln(HSAK/\delta)}.$$

We define the value function $V_h^\pi(s; \ell, P)$ at state $s \in \mathcal{S}$ and step $h \in [H]$ for a given policy π, function ℓ, and transition kernel P as

$$V_h^\pi(s; \ell, P) = \mathbb{E}\left[\sum_{j=h}^{H} \ell(s_j^{P,\pi}, a_j^{P,\pi}, j) \mid \ell, \pi, P, s_h^{P,\pi} = s\right].$$

Moreover, let $V_1^\pi(\ell, P) = \mathbb{E}_{s\sim p}\left[V_1^\pi(s; \ell, P) \mid \ell, \pi, P\right]$ where p is the known distribution of the initial state.

1. We may impose conditional independence.

The goal of the constrained Markov decision process is to learn an optimal policy π^* defined as

$$\pi^* \in \underset{\pi}{\operatorname{argmax}} \quad V_1^{\pi}(f, P) \quad \text{s.t.} \quad V_1^{\pi}(g, P) \leq \bar{C}$$

where $\bar{C}$ is the budget on the expected cost over the horizon. As the model parameters f, g, P are unknown, we develop a learning algorithm that computes policies over multiple episodes. For K episodes, we deduce policies $\pi_1, \ldots, \pi_K$ with the safety requirement that

$$V_1^{\pi_k}(g, P) \leq \bar{C} \quad \forall k \in [K]$$

holds with high probability. The safety requirement is equivalent to enforcing zero hard constraint violation where the hard constraint violation is defined as

$$\text{Violation}(\vec{\pi}) := \sum_{k=1}^{K} \max\left\{0, V_1^{\pi_k}(g, P) - \bar{C}\right\}$$

and $\vec{\pi} = (\pi_1, \ldots, \pi_K)$ is a shorthand notation for the K policies. As a performance metric for a learning algorithm, we use the following notion of regret.

$$\text{Regret}(\vec{\pi}) := \sum_{k=1}^{K} \left(V_1^{\pi^*}(f, P) - V_1^{\pi_k}(f, P)\right).$$

To satisfy the safety requirement, we assume that a *strictly safe baseline policy* π_b is given to the agent.

Assumption 1 *The agent knows a policy π_b and its expected cost $\bar{C}_b = V_1^{\pi_b}(g, P)$. We further assume that π_b is strictly feasible, i.e., $\bar{C}_b < \bar{C}$.*

This assumption is necessary because the learning agent has no information about the underlying MDP at the beginning. Without a safe baseline policy, it is difficult to satisfy the constraint in the initial phase of learning. It is a commonly assumed condition for learning CMDPs (Simão et al., 2021; Liu et al., 2021; Bura et al., 2022). We also remark that strict feasibility of π_b is related to Slater's condition in constrained optimization. Lastly, we assume that the budget $\bar{C}$ satisfies $\bar{C} \in (0, H)$. If $\bar{C} \geq H$, then as $V_1^{\pi}(g, P) \leq H$ for any policy π, the safety requirement is trivially satisfied. Moreover, we have $\bar{C}$ is strictly positive because Assumption 1 imposes that $\bar{C} > \bar{C}_b$ and $\bar{C}_b = V_1^{\pi_b}(g, P) \geq 0$.

3 Model Estimators

In this section, we present our model estimators. In Section 3.1, we define confidence sets for the transition kernel and confidence intervals for reward and cost functions. In Section 3.2, we deduce our tighter optimistic reward and pessimistic cost function estimators. Lastly, in Section 3.3, we sketch our technical proof for obtaining the tighter function estimators.

3.1 Confidence Sets and Intervals

We follow the standard Bernstein inequality-based confidence set construction for estimating the true transition kernel and use confidence intervals based on Hoeffding's inequality for estimating reward and cost functions (Jin et al., 2020; Cohen et al., 2020).

As in Efroni et al. (2020); Bura et al. (2022), we maintain counters to keep track of the number of visits to each tuple (s, a, h) and tuple (s, a, s', h). For each $k \in [K]$, we define $N_k(s, a, h)$ and $M_k(s, a, s', h)$ as the number of visits to tuple (s, a, h) and the number of visits to tuple (s, a, s', h) up to the first $k-1$ episodes, respectively, for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Given $N_k(s, a, h)$ and $M_k(s, a, s', h)$, we define the empirical transition kernel $\bar{P}_k$ for episode k as

$$\bar{P}_k(s' \mid s, a, h) = \frac{M_k(s, a, s', h)}{\max\{1, N_k(s, a, h)\}}.$$

Next, for some confidence parameter $\delta \in (0, 1)$, we define the confidence radius $\epsilon_k(s' \mid s, a, h)$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ and $k \in [K]$ as

$$\epsilon_k(s' \mid s, a, h) = 2\sqrt{\frac{\bar{P}_k(s' \mid s, a, h) \ln(HSAK/\delta)}{\max\{1, N_k(s, a, h) - 1\}}} + \frac{14 \ln(HSAK/\delta)}{3 \max\{1, N_k(s, a, h) - 1\}}.$$

Based on the empirical transition kernel and the radius, we define the confidence set $\mathcal{P}_k$ for episode k as

$$\mathcal{P}_k = \left\{ \widehat{P} : \ \left| \widehat{P}(s' \mid s, a, h) - \bar{P}_k(s' \mid s, a, h) \right| \leq \epsilon_k(s' \mid s, a, h) \ \ \forall (s, a, s', h) \ \right\}. \tag{1}$$

Then by the empirical Bernstein inequality due to Maurer and Pontil (2009), we can show the following.

Lemma 3.1 *With probability at least $1 - 4\delta$, the true transition kernel P is contained in the confidence set $\mathcal{P}_k$ for every episode $k \in [K]$.*

Next, for reward and cost functions, we define the confidence radius $R_k(s, a, h)$ for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $k \in [K]$ and $\delta \in (0, 1)$ as

$$R_k(s, a, h) = \sqrt{\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s, a, h)\}}}.$$

We define empirical estimators $\bar{f}_k$ and $\bar{g}_k$ as

$$\bar{f}_k(s, a, h) = \frac{\sum_{j=1}^{k-1} f_j(s, a, h) n_j(s, a, h)}{\max\{1, N_k(s, a, h)\}}, \quad \bar{g}_k(s, a, h) = \frac{\sum_{j=1}^{k-1} g_j(s, a, h) n_j(s, a, h)}{\max\{1, N_k(s, a, h)\}}$$

where $f_j(s, a, h)$, $g_j(s, a, h)$ are the instantaneous reward and cost for episode $j \in [k-1]$ and $n_j(s, a, h)$ is the indicator variable that returns 1 if the agent visited (s, a, h) in episode j and 0 otherwise. Then we may deduce the following from Hoeffding's inequality.

Lemma 3.2 *With probability at least $1 - 4\delta$, it holds that for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$,*

$$\left| \bar{f}_k(s, a, h) - f(s, a, h) \right| \leq R_k(s, a, h), \quad \left| \bar{g}_k(s, a, h) - g(s, a, h) \right| \leq R_k(s, a, h).$$

3.2 Tighter Function Estimators

Lemmas 3.1 and 3.2 motivate the following attempt to deduce feasible policies. For episode $k \in [K]$, we take a transition kernel P_k from the confidence set $\mathcal{P}_k$ and $\bar{g}_k + R_k$ as a pessimistic (or conservative) estimator of the cost function g. Then we may compute a policy π_k that satisfies $V_1^{\pi_k}(\bar{g}_k + R_k, P_k) \le \bar{C}$, which is an approximation of the constraint. However, even if $\bar{g}_k + R_k$ provides an upper bound on g, the issue is that $V_1^{\pi_k}(g, P) \not\le V_1^{\pi_k}(\bar{g}_k + R_k, P_k)$. This is because the difference between the true transition kernel P and P_k can make $V_1^{\pi_k}(g, P)$ greater than $V_1^{\pi_k}(\bar{g}_k + R_k, P_k)$. That said, π_k does not necessarily satisfy the constraint, although it satisfies the approximate constraint.

Inspired by the challenge, the next question is as to whether we can design an approximate constraint, satisfying which guarantees that the true constraint is also satisfied. Liu et al. (2021); Bura et al. (2022) considered this, and their idea was to add an extra pessimism to cost function estimators. Basically, we take functions of the form

$$\widehat{g}_k(s, a, h) = \bar{g}_k(s, a, h) + R_k(s, a, h) + U_k(s, a, h) \tag{2}$$

for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$ where U_k captures the error in estimating the true transition kernel P. In the above-discussed context, U_k considers the difference between P and P_k. Here, one needs to set U_k sufficiently large so that $V_1^{\pi_k}(g, P) \le V_1^{\pi_k}(\widehat{g}_k, P_k)$, in which case satisfying the corresponding approximate constraint $V_1^{\pi_k}(\widehat{g}_k, P_k) \le \bar{C}$ guarantees satisfaction of the true constraint.

On the other hand, choosing the right magnitude of U_k is important to control the regret function. When U_k is too large, $\widehat{g}_k$ is too conservative, and it prevents from getting a high reward. Indeed, Bura et al. (2022) improved upon Liu et al. (2021) by making U_k tighter. Our main contribution is to develop an even tighter U_k function than Bura et al. (2022).

Before we present our design of U_k, let us briefly discuss how to deduce the extra pessimism term U_k in general. As explained before, we want to guarantee $V_1^{\pi_k}(g, P) \le V_1^{\pi_k}(\widehat{g}_k, P_k)$ for any $P_k \in \mathcal{P}_k$. Then note that

$$V_1^{\pi_k}(g, P) \le V_1^{\pi_k}(g, P_k) + |V_1^{\pi_k}(g, P) - V_1^{\pi_k}(g, P_k)|.$$

If the statement of Lemma 3.2 holds, then $V_1^{\pi_k}(g, P_k)$ is bounded above by $V_1^{\pi_k}(\bar{g}_k + R_k, P_k)$. Therefore, once we come up with some U_k such that $|V_1^{\pi_k}(g, P) - V_1^{\pi_k}(g, P_k)| \le V_1^{\pi_k}(U_k, P_k)$, we get

$$V_1^{\pi_k}(g, P) \le V_1^{\pi_k}(\bar{g}_k + R_k + U_k, P_k).$$

In this case, $\widehat{g}_k = \bar{g}_k + R_k + U_k$ gives rise to a valid function estimator.

We devise our pessimism function U_k as follows.

Theorem 1 *Let π_k be any policy for episode k. Take*

$$U_k(s, a, h) = 8\sqrt{H}\varepsilon_k(s, a, h) + 4S\sqrt{HA/K} + \frac{2\sqrt{HK/A}\ln(HSAK/\delta) + \eta}{\max\{1, N_k(s, a, h) - 1\}} \tag{3}$$

for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$ where

$$\varepsilon_k(s, a, h) = 2\sqrt{\frac{S\ln(HSAK/\delta)}{\max\{1, N_k(s, a, h) - 1\}}} + \frac{14S\ln(HSAK/\delta)}{3\max\{1, N_k(s, a, h) - 1\}} \tag{4}$$

and $\eta = (19HS + 2H^{1.5}S + 10^4 H^2 S^2)(\ln(HSAK/\delta))^2$. *Then it holds with probability at least* $1 - 14\delta$ *that*

$$|V_1^{\pi_k}(g, P) - V_1^{\pi_k}(g, P_k)| \le V_1^{\pi_k}(U_k, P_k)$$

for any $P_k \in \mathcal{P}_k$ *and* $g : \mathcal{S} \times \mathcal{A} \times [H] \to [0, 1]$.

Next, we demonstrate that our U_k indeed improves upon Bura et al. (2022).

Remark 1 Bura et al. (2022) set U_k as

$$U_k(s, a, h) = 2H \sum_{s' \in \mathcal{S}} \epsilon_k(s' \mid s, a, h). \tag{5}$$

However, there is a minor issue with this choice. We need the property that U_k is nonincreasing in k to show Lemma 4.1 and (Proposition 4, Bura et al., 2022), but U_k given in (5) can increase as $M_k(s, a, s', h)/N_k(s, a, h)^2$ can increase. As a fix, we may take $U_k(s, a, h) = 2H\varepsilon_k(s, a, h)$ where ε_k is given in (4). By the Cauchy-Schwarz inequality, we have $\sum_{s' \in \mathcal{S}} \epsilon_k(s' \mid s, a, h) \le \varepsilon_k(s, a, h)$. Moreover, $\varepsilon_k(s, a, h)$ is nonincreasing in k, as desired. Comparing $2H\varepsilon_k(s, a, h)$ with our construction from Theorem 1, we have coefficient $8\sqrt{H}$ instead of $2H$ for the sum. Although we have additional terms for U_k, the reduction of $\mathcal{O}(\sqrt{H})$ in the coefficient translates to the improvement of $\mathcal{O}(\sqrt{H})$ factor in the regret upper bound.□

Now we present our optimistic reward function estimator $\widehat{f}_k$. On top of $\bar{f}_k + R_k$, we take additional optimistic terms for the reward function to compensate for the extra pessimism in $\widehat{g}_k$, which reduces the search space of policies and hinders exploration. We define the optimistic reward function estimator $\widehat{f}_k$ as

$$\widehat{f}_k(s, a, h) = \min\left\{B,\ \bar{f}_k(s, a, h) + \frac{3H}{\bar{C} - \bar{C}_b} R_k(s, a, h) + \frac{H}{\bar{C} - \bar{C}_b} U_k(s, a, h)\right\} \tag{6}$$

where $B = 1 + \sqrt{\ln(HSAK/\delta)}$. Here, the extra optimism in $\widehat{f}_k$ is designed to promote exploration.

3.3 Proof Outline of Theorem 1

The value difference lemma (Dann et al., 2017) implies

$$V_1^{\pi_k}(g, P) - V_1^{\pi_k}(g, P_k) = \mathbb{E}\left[\sum_{h=1}^{H} \ell(s_h^{P_k, \pi_k}, a_h^{P_k, \pi_k}, h) \mid \pi_k, P_k\right]$$

where $\ell(s, a, h)$ is given by

$$\sum\nolimits_{s' \in \mathcal{S}} (P - P_k)(s' \mid s, a, h) V_{h+1}^{\pi_k}(s'; g, P) \tag{7}$$

with $V_{H+1}^{\pi_k} = 0$ and $(P - P_k)(s' \mid s, a, h) = P(s' \mid s, a, h) - P_k(s' \mid s, a, h)$. Here, Bura et al. (2022) used that $V_{h+1}^{\pi_k} \le H$ and $|P - P_k| \le |P - \bar{P}_k| + |\bar{P}_k - P_k| \le 2\epsilon_k$ by Lemma 3.1. Then it follows that

$$|V_1^{\pi_k}(g, P) - V_1^{\pi_k}(g, P_k)| \le \mathbb{E}\left[\sum_{h=1}^{H} 2H \sum_{s' \in \mathcal{S}} \epsilon_k(s' \mid s_h^{P_k, \pi_k}, a_h^{P_k, \pi_k}, h) \mid \epsilon_k, \pi_k, P_k\right]$$

whose right-hand side equals $V_1^{\pi_k}(U_k, P_k)$ where U_k is given as in (5). This explains how Bura et al. (2022) deduced their pessimistic cost estimators.

To prove Theorem 1 that establishes the validity of our choice of tighter U_k in (3), we need a more refined analysis of the difference term $|V_1^{\pi_k}(g, P) - V_1^{\pi_k}(g, P_k)|$. Note that $\ell(s, a, h)$ in (7) satisfies

$$|\ell(s,a,h)| \leq \left|\sum_{s' \in \mathcal{S}} (P - P_k)(s' \mid s,a,h) V_{h+1}^{\pi_k}(s'; g, P_k)\right| + \left|\sum_{s' \in \mathcal{S}} (P - P_k)(s' \mid s,a,h) W_{h+1}^{\pi_k}(s'; g)\right|$$

where $W_{h+1}^{\pi_k}(s'; g) = V_{h+1}^{\pi_k}(s'; g, P) - V_{h+1}^{\pi_k}(s'; g, P_k)$. We may argue that the second term on the right-hand side is a small value. That said, let us focus on the first term which is the dominant one. Since P and P_k both define transition functions, the first term equals

$$\left|\sum_{s' \in \mathcal{S}} (P - P_k)(s' \mid s,a,h)(V_{h+1}^{\pi_k}(s'; g, P_k) - \widehat{\mu}_k(s,a,h))\right|$$

where $\widehat{\mu}_k(s,a,h) = \mathbb{E}_{s' \sim P_k(\cdot \mid s,a,h)}[V_{h+1}^{\pi_k}(s'; g, P_k)]$. Next, we observe that

$$\left|(P - P_k)(s' \mid s,a,h)\right| \leq 2\epsilon_k(s' \mid s,a,h)$$

due to Lemma 3.1. Recall that $\epsilon_k(s' \mid s,a,h)$ contains the term $\sqrt{\bar{P}_k(s' \mid s,a,h)}$. As $P_k \in \mathcal{P}_k$ we deduce that $\sqrt{\bar{P}_k(s' \mid s,a,h)} \leq \sqrt{P_k(s' \mid s,a,h) + \epsilon_k(s' \mid s,a,h)}$. As a result, by the Cauchy-Schwarz inequality, the analysis boils down to providing an upper bound on the term

$$\sum_{s' \in \mathcal{S}} P_k(s' \mid s,a,h)(V_{h+1}^{\pi_k}(s'; g, P_k) - \widehat{\mu}_k(s,a,h))^2,$$

which equals

$$\widehat{\mathbb{V}}_k(s,a,h) := \operatorname*{Var}_{s' \sim P_k(\cdot \mid s,a,h)}[V_{h+1}^{\pi_k}(s'; g, P_k)].$$

Furthermore, our proof reveals that $V_1^{\pi_k}(\widehat{\mathbb{V}}_k, P_k)$ is the important quantity to control. Applying a naïve upper bound on value functions as in Bura et al. (2022) gives $\widehat{\mathbb{V}}_k \leq H^2$ and thus $V_1^{\pi_k}(\widehat{\mathbb{V}}_k, P_k) \leq H^3$. However, this bound is not tight enough. Instead, we prove the following lemma based on a Bellman-type law of total variance (Azar et al., 2017; Chen and Luo, 2021).

Lemma 3.3 *Let π_k be a policy for episode k. Then*

$$V_1^{\pi_k}(\widehat{\mathbb{V}}_k, P_k) \leq 2H^2$$

for any $P_k \in \mathcal{P}_k$ and $g : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$.

This improvement in the variance term leads to our tighter estimators.

Algorithm 1 Doubly Optimistic Pessimistic Exploration with Tighter Function Estimators (DOPE+)

Initialize: episode counter $k = 1$, counters $N(s, a, h) = 0$ and $M(s, a, s', h) = 0$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, safe baseline policy π_b and its expected cost for a single episode $\bar{C}_b$, and the number K_0 of episodes for the initial phase.
for $k = 1, \ldots, K$ **do**
  Set counters $N_k \leftarrow N$ and $M_k \leftarrow M$.
  Compute $\bar{P}_k$, ϵ_k, and $\mathcal{P}_k$ (Section 3.1).
  if $k \leq K_0$ **then**
    $\pi_k = \pi_b$
  end if
  if $k > K_0$ **then**
    Compute estimators $\widehat{f}_k$ and $\widehat{g}_k$ (Section 3.2).
    Deduce π_k, P_k from (8).
  end if
  Sample state s_1 from distribution p.
  for $h = 1, \ldots, H$ **do**
    Sample a_h from $\pi_k(\cdot \mid s_h, h)$
    Observe $f_k(s_h, a_h, h)$ and $g_k(s_h, a_h, h)$.
    Observe s_{h+1} determined by $P(\cdot \mid s_h, a_h, h)$.
    Update counters $N(s_h, a_h, h) \leftarrow N(s_h, a_h, h) + 1$ and $M(s_h, a_h, s_{h+1}, h) \leftarrow M(s_h, a_h, s_{h+1}, h) + 1$.
  end for
end for

4 Algorithm

DOPE+, given by Algorithm 1, is a variant of DOPE by Bura et al. (2022) with our novel reward and cost function estimators from Section 3. Recall that our pessimistic cost estimator $\widehat{g}_k$ is given by (2) with the extra pessimism term U_k given in (3) and our optimistic reward estimator $\widehat{f}_k$ is given in (6).
As in Efroni et al. (2020); Bura et al. (2022), we compute our policy π_k for episode $k \in [K]$ by solving the following optimization problem.

$$(\pi_k, P_k) \in \underset{(\pi,Q)\in\Pi\times\mathcal{P}_k}{\operatorname{argmax}} \left\{ V_1^\pi(\widehat{f}_k, Q) : V_1^\pi(\widehat{g}_k, Q) \leq \bar{C} \right\} \tag{8}$$

where $\mathcal{P}_k$ is the confidence set given by (1) and

$$\Pi = \left\{ \pi : \sum\nolimits_{a\in\mathcal{A}} \pi(a \mid s, h) = 1,\ \pi(a \mid s, h) \geq 0 \qquad \forall (s, a, h) \right\}$$

is the set of valid policies. Note that (8) takes the optimistic objective with $\widehat{f}_k$ and the pessimistic (or conservative) constraint with $\widehat{g}_k$. The optimistic objective induces exploration while the pessimistic constraint guarantees constraint satisfaction. Moreover, we optimize over the space of the confidence set $\mathcal{P}_k$, which also encourages exploration.

Now that we have the optimization formulation (8), the next question is as to how we solve it. We take the standard approach of using *occupancy measures* (Altman, 1999). An occupancy measure is essentially a joint probability distribution for the event that we observe the state-action pair (s, a) at step h and state s' at step $h + 1$. An occupancy measure defines a policy and a transition kernel. The converse is true in that we can define the occupancy measure associated with a given pair of a policy and a transition kernel. Introducing occupancy measure, we can reformulate (8) as an optimization problem in terms of occupancy measures.

In the reformulation, the objective and the constraint become linear in an occupancy measure. Hence, the reformulation is a linear program, and we refer to it as the *extended linear program* (Altman, 1999). Again, by solving it, we deduce an optimal occupancy measure, which corresponds to an optimal solution to (8). We defer the formal description of the extended linear program and occupancy measures to the appendix.

One issue, however, is that (8) can be infeasible at the beginning of the algorithm as $\widehat{g}_k$ can be too large to guarantee feasibility of (8). Hence, the algorithm executes the safe baseline policy π_b for the first few episodes until sufficient information is gathered so that (8) becomes feasible. The following lemma characterizes a sufficient number of episodes running the safe baseline policy to guarantee feasibility of (8).

Lemma 4.1 *With probability at least $1 - 14\delta$, (π_b, P) is a feasible solution of* (8) *for any $k > K_0$ where*

$$K_0 = \widetilde{\mathcal{O}}\left(\frac{H^3 S^2 A}{(\bar{C} - \bar{C}_b)^2}\right) \tag{9}$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides factors polynomial in $\ln(HSAK/\delta)$.

5 Regret Analysis of DOPE+

Let us state our theoretical guarantees for DOPE+.

Theorem 2 *Let $\vec{\pi} = (\pi_1, \ldots, \pi_K)$ denote policies computed by DOPE+ with K_0 given in* (9). *Then*

$$\text{Violation}(\vec{\pi}) = 0$$

with probability at least $1 - 14\delta$.

Hence, DOPE+ achieves no constraint violation. The next theorem shows a regret upper bound for DOPE+.

Theorem 3 *Let $\vec{\pi} = (\pi_1, \ldots, \pi_K)$ denote policies computed by DOPE+ with K_0 given in* (9). *Then, with probability at least $1 - 16\delta$, we have*

$$\text{Regret}\,(\vec{\pi}) = \widetilde{\mathcal{O}}\left(\frac{H}{\bar{C} - \bar{C}_b}\left(H^{1.5} S\sqrt{AK} + \frac{H^4 S^3 A}{\bar{C} - \bar{C}_b}\right)\right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides factors polynomial in $\ln(HSAK/\delta)$.

Remark 2 Note that there is a gap of $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H\sqrt{S})$ factor between our regret upper bound and the lower bound $\Omega(H^{3/2}\sqrt{SAK})$ due to Jin et al. (2020); Domingues et al. (2021). In fact, the instance from Domingues et al. (2021) is an unconstrained MDP. We observe that the $\mathcal{O}(H/(\bar{C}-\bar{C}_b))$ factor in our regret upper bound is due to the constraint, which becomes a constant if $\bar{C}-\bar{C}_b = \Omega(H)$. Hence, our regret upper bound nearly matches in terms of H when $\bar{C}-\bar{C}_b = \Omega(H)$.$\square$

5.1 Constraint Violation Analysis

We prove Theorem 2 as follows. For episode k with $k \leq K_0$, Algorithm 1 takes the safe baseline policy π_b, so no constraint violation is guaranteed. Then let us consider episode k with $k > K_0$. As explained in Section 3.2, we argue that

$$\begin{aligned}
V_1^{\pi_k}(g,P) &\leq V_1^{\pi_k}(g,P_k) + |V_1^{\pi_k}(g,P) - V_1^{\pi_k}(g,P_k)| \\
&\leq V_1^{\pi_k}(\bar{g}_k + R_k, P_k) + V_1^{\pi_k}(U_k, P_k) \\
&= V_1^{\pi_k}(\widehat{g}_k, P_k)
\end{aligned}$$

where the second inequality is due to Lemma 3.2 and Theorem 1. Since (π_k, P_k) is a solution to (8), it holds that $V_1^{\pi_k}(\widehat{g}_k, P_k) \leq \bar{C}$. Therefore, it follows that $V_1^{\pi_k}(g,P) \leq \bar{C}$ and thus the constraint is satisfied.

5.2 Regret Decomposition

We provide an overview of the proof of Theorem 3. Since we execute the safe baseline policy π_b for the first K_0 episodes, we decompose the regret function as follows.

$$\begin{aligned}
\text{Regret}\,(\vec{\pi}) = &\underbrace{\sum_{k=1}^{K_0}\Big(V_1^{\pi^*}(f,P) - V_1^{\pi_b}(f,P)\Big)}_{\text{(I)}} + \underbrace{\sum_{k=K_0+1}^{K}\Big(V_1^{\pi^*}(f,P) - V_1^{\pi_k}(\widehat{f}_k,P_k)\Big)}_{\text{(II)}} \\
&+ \underbrace{\sum_{k=K_0+1}^{K}\Big(V_1^{\pi_k}(\widehat{f}_k,P_k) - V_1^{\pi_k}(\widehat{f}_k,P)\Big)}_{\text{(III)}} + \underbrace{\sum_{k=K_0+1}^{K}\Big(V_1^{\pi_k}(\widehat{f}_k,P) - V_1^{\pi_k}(f,P)\Big)}_{\text{(IV)}}.
\end{aligned}$$

Term (I) is due to executing π_b for K_0 episodes for feasibility. By Lemma 4.1, term (I) can be bounded by $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-2}(H^4S^2A))$ as $V_1^{\pi} \leq H$ for any policy π.
For term (II), we provide the following upper bound.

Lemma 5.1 *With probability at least* $1-14\delta$*,*

$$\sum_{k=K_0+1}^{K}\Big(V_1^{\pi^*}(f,P) - V_1^{\pi_k}(\widehat{f}_k,P_k)\Big) = \widetilde{\mathcal{O}}\left(\frac{H}{\bar{C}-\bar{C}_b}\left(H^{1.5}S\sqrt{AK} + H^3S^3A\right)\right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ *hides factor* $(\ln(HSAK/\delta))^3$.

To prove the lemma, we define a new policy $\pi_k^{\alpha_k}$ for $k \in [K]$, which is a convex combination of the occupancy measures associated with (π^*, P) and (π_b, P) with coefficients $\alpha_k, 1-\alpha_k \in (0,1)$. We choose the value of α_k so that $(\pi_k^{\alpha_k}, P)$ is feasible to (8). Then the optimality of (π_k, P_k) implies $V_1^{\pi_k^{\alpha_k}}(\widehat{f}_k, P) \leq V_1^{\pi_k}(\widehat{f}_k, P_k)$, which lets us to analyze $V_1^{\pi^*}(f,P) - V_1^{\pi_k^{\alpha_k}}(\widehat{f}_k, P)$ with the same transition kernel P.
Term (III) comes from learning the unknown transition kernel. We apply a Bellman-type law of total variance to provide an upper bound on term (III).

Lemma 5.2 *With probability at least* $1-16\delta$,

$$\sum_{k=K_0+1}^{K} \left(V_1^{\pi_k}(\widehat{f}_k, P_k) - V_1^{\pi_k}(\widehat{f}_k, P) \right) = \widetilde{\mathcal{O}}\left(H^{1.5} S\sqrt{AK} + H^3 S^3 A \right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ *hides factor* $(\ln(HSAK/\delta))^4$.

Term (IV) is due to the difference between f and our estimator $\widehat{f}_k$.

Lemma 5.3 *With probability at least* $1-14\delta$,

$$\sum_{k=K_0+1}^{K} \left(V_1^{\pi_k}(\widehat{f}_k, P) - V_1^{\pi_k}(f, P) \right) = \widetilde{\mathcal{O}}\left(\frac{H}{\bar{C} - \bar{C}_b} \left(H^{1.5} S\sqrt{AK} + H^3 S^3 A \right) \right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ *hides factor* $(\ln(HSAK/\delta))^3$.

6 Numerical Experiment

We evaluate DOPE+ on the three-state CMDP instance of Zheng and Ratliff (2020); Simão et al. (2021); Bura et al. (2022) with a few modifications. In Figure 1, we compare regret and constraint violation under DOPE+ and DOPE for $200,000$ episodes when $H = 30$. We consider DOPE as a benchmark algorithm because it provides the best regret bound among the existing algorithms while ensuring zero constraint violation. Our results are averaged across 5 runs with different random seeds, and we display the 95% confidence interval with shaded regions. More details of the experiment setup can be found in the appendix including the MDP instance and algorithm parameters.
In Figure 1, DOPE+ outperforms DOPE in terms of regret. This result demonstrates that DOPE+ improves upon DOPE computationally, in addition to our theoretical improvement. Figure 1 shows that both algorithms achieve zero constraint violation.

7 Conclusion

In this paper, we investigate safe RL formulated as an episodic finite-horizon tabular CMDP. We propose novel reward and cost function estimators with tighter reward optimism and cost pessimism. Based on them, we develop DOPE+, which is a variant of the LP-based DOPE due to (Bura et al., 2022). We prove that DOPE+ achieves regret upper bound $\widetilde{\mathcal{O}}((\bar{C} - \bar{C}_b)^{-1} H^{2.5} S\sqrt{AK})$ and zero hard constraint violation. The regret upper bound improves upon the best-known bound by a multiplicative factor of $\widetilde{\mathcal{O}}(\sqrt{H})$ factor. When

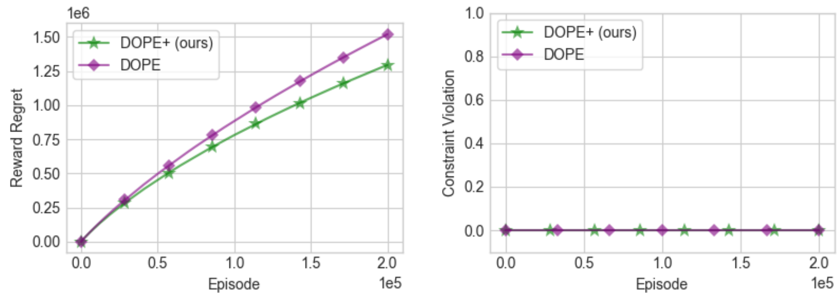

Figure 1: Comparison of DOPE+ and DOPE: Regret (Left) and Hard Constraint Violation (Right)

$\bar{C} - \bar{C}_b = \Omega(H)$, DOPE+ nearly matches the lower bound $\Omega(H^{1.5}\sqrt{SAK})$ (Jin et al., 2020; Domingues et al., 2021). We also present numerical results that demonstrate the computational effectiveness of DOPE+ compared to DOPE.

Acknowledgments and Disclosure of Funding

This research is supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00350703).

References

N. Abe, P. Melville, C. Pendus, C. K. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk, M. Domick, and T. Gardinier. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 75–84, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300551. doi: 10.1145/1835804.1835817. URL https://doi.org/10.1145/1835804.1835817.

E. Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/azar17a.html.

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 19–26, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/beygelzimer11a.html.

K. Brantley, M. Dudík, T. Lykouris, S. Miryoosefi, M. Simchowitz, A. Slivkins, and W. Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

A. Bura, A. Hasanzadezonuzy, D. Kalathil, S. Shakkottai, and J.-F. Chamberland. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=U4BUMoVTrB2.

L. Chen and H. Luo. Finding the stochastic shortest path with low regret: the adversarial cost and unknown transition case. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1651–1660. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/chen21l.html.

Y. Chen, J. Dong, and Z. Wang. A primal-dual approach to constrained markov decision processes, 2021. URL https://arxiv.org/abs/2101.10895.

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

A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109: 101964, 2020. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2020.101964. URL https://www.sciencedirect.com/science/article/pii/S093336572031229X.

C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/17d8da815fa21c57af9829fb0a869602-Paper.pdf.

D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International conference on artificial intelligence and statistics*, pages 3304–3312. PMLR, 2021.

O. D. Domingues, P. Ménard, E. Kaufmann, and M. Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/domingues21a.html.

Y. Efroni, S. Mannor, and M. Pirotta. Exploration-exploitation in constrained mdps, 2020.

J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

J. García and D. Shafie. Teaching a humanoid robot to walk faster through safe reinforcement learning. *Engineering Applications of Artificial Intelligence*, 88:103360, 2020.

A. Ghosh, X. Zhou, and N. Shroff. Provably efficient model-free constrained rl with linear function approximation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13303–13315. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/56b8f22d895c45f60eaac9580152afd9-Paper-Conference.pdf.

A. Ghosh, X. Zhou, and N. Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free rl. In *International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR, 2024.

D. Isele, A. Nakhaei, and K. Fujimura. Safe reinforcement learning on autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6. IEEE, 2018.

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

C. Jin, T. Jin, H. Luo, S. Sra, and T. Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/jin20c.html.

K. C. Kalagarla, R. Jain, and P. Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8030–8037, May 2021. doi: 10.1609/aaai.v35i9.16979. URL https://ojs.aaai.org/index.php/AAAI/article/view/16979.

K. C. Kalagarla, R. Jain, and P. Nuzzo. Safe posterior sampling for constrained mdps with bounded constraint violation, 2023. URL https://arxiv.org/abs/2301.11547.

H. Krasowski, X. Wang, and M. Althoff. Safe reinforcement learning for autonomous lane changing using set-based prediction. In *2020 IEEE 23rd international conference on Intelligent Transportation Systems (ITSC)*, pages 1–7. IEEE, 2020.

T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. In A. Beygelzimer, Y. Dauphin, P. Liang, and

J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Nl7VO_Y7K4Q.

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

S. Miryoosefi and C. Jin. A simple reward-free approach to constrained reinforcement learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15666–15698. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/miryoosefi22a.html.

T. Moskovitz, B. O'Donoghue, V. Veeriah, S. Flennerhag, S. Singh, and T. Zahavy. ReLOAD: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained MDPs. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25303–25336. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/moskovitz23a.html.

A. Müller, P. Alatur, V. Cevher, G. Ramponi, and N. He. Truly no-regret learning in constrained MDPs. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36605–36653. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/muller24b.html.

S. Qiu, X. Wei, Z. Yang, J. Ye, and Z. Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15277–15287. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ae95296e27d7f695f891cd26b4f37078-Paper.pdf.

A. Rosenberg and Y. Mansour. Online convex optimization in adversarial Markov decision processes. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5478–5486. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rosenberg19a.html.

T. D. Simão, N. Jansen, and M. T. J. Spaan. Alwayssafe: Reinforcement learning without safety constraint violations during training. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 1226–1235, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

R. Singh, A. Gupta, and N. B. Shroff. Learning in constrained markov decision processes. *IEEE Transactions on Control of Network Systems*, 10(1):441–453, 2023. doi: 10.1109/TCNS.2022.3203361.

A. Stooke, J. Achiam, and P. Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9133–9143. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

H. Wei, X. Liu, and L. Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):3868–3876, Jun. 2022a. doi: 10.1609/aaai.v36i4.20302. URL https://ojs.aaai.org/index.php/AAAI/article/view/20302.

H. Wei, X. Liu, and L. Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pages 3274–3307. PMLR, 2022b.

H. Wei, A. Ghosh, N. Shroff, L. Ying, and X. Zhou. Provably efficient model-free algorithms for non-stationary cmdps. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6527–6570. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/wei23b.html.

T. Yu, Y. Tian, J. Zhang, and S. Sra. Provably efficient algorithms for multi-objective competitive rl. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12167–12176. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yu21b.html.

L. Zheng and L. Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pages 620–629. PMLR, 2020.

Appendix A. Related Work

In this section, we provide a more detailed discussion of related work to online learning of constrained Markov decision processes (CMDPs). As explained in the introduction, we review previous works for the three frameworks, cumulative constraint violation, hard constraint violation, and zero constraint violation.

Cumulative Constraint Violation Starting with the work of Efroni et al. (2020), online learning of CMDPs has been an active area of research in reinforcement learning, especially with the framework of cumulative (or soft) constraint violation (Brantley et al., 2020; Qiu et al., 2020; Zheng and Ratliff, 2020; Kalagarla et al., 2021; Ding et al., 2021; Chen et al., 2021; Yu et al., 2021; Liu et al., 2021; Wei et al., 2022a,b; Singh et al., 2023; Miryoosefi and Jin, 2022; Ghosh et al., 2022; Wei et al., 2023; Kalagarla et al., 2023). Among these works, Brantley et al. (2020) studied a knapsack constrained formulation, and Qiu et al. (2020) studied the setting where the reward functions are adversarially given and the cost functions are sampled from a fixed but unknown distribution. Moreover, Zheng and Ratliff (2020) considered the case where the transition kernel is known to the agent, and Kalagarla et al. (2021) studied a PAC bound for learning CMDPs. Ding et al. (2021); Chen et al. (2021) developed model-free algorithms for CMDPs, although these approaches require access to simulators, while Yu et al. (2021) studied vector-valued Markov games for a variant of constrained MDPs. Liu et al. (2021) introduced the first algorithm that achieves zero cumulative constraint violation. Wei et al. (2022a) and Singh et al. (2023) considered the infinite-horizon average-reward setting. Moreover, Wei et al. (2022b) came up with a model-free algorithm for finite-horizon episodic tabular CMDPs. Miryoosefi and Jin (2022) studied the reward-free setting, and Ghosh et al. (2022) proposed an algorithm for the linear MDP setting, which leads to a model-free algorithm for tabular CMDPs. Lastly, Wei et al. (2023) considered non-stationary CMDPs, while Kalagarla et al. (2023) developed a posterior sampling-based algorithm that guarantees a Bayesian regret upper bound.

Wei et al. (2022b) introduced model-free and simulator-free algorithms to solve tabular CMDPs. These algorithms were analyzed under soft constraint violations, thus they do not guarantee safety in all episodes. In contrast, Müller et al. (2024); Ghosh et al. (2024) presented PD-based algorithms with hard constraint violations, though these suffer from high regret and constraint violations. On the other hand, Liu et al. (2021) proposed the LP-based algorithm OptPess-LP, which achieves zero hard constraint violations with sublinear regret by employing *optimistic pessimism in the face of uncertainty (OPFU)*. The pessimism in the cost function estimator ensures safety but hampers exploration. To address this, Bura et al. (2022) recently proposed DOPE, incorporating optimism for the transition kernel to improve the regret bound.

Hard Constraint Violation The notion of hard constraint violation was introduced by Efroni et al. (2020). Efroni et al. (2020) developed an LP-based algorithm for controlling hard constraint violation and raised an open question of whether there exists a primal-dual algorithm for the setting. Recently, Ghosh et al. (2024) established an algorithm that guarantees a sublinear regret upper bound and a sublinear upper bound on hard constraint violation. Their algorithm is for the linear MDP setting, and it provides a model-free algorithm for the tabular setting. In fact, their analysis shows that for the tabular case, one may get a tighter performance guarantees. Müller et al. (2024) developed a simpler primal-dual algorithm that guarantees a sublinear regret upper bound and a sublinear upper bound on hard constraint violation, answering the question of Efroni et al. (2020).

Zero Constraint Violation Simão et al. (2021) considered the importance of achieving no constraint violation, which is equivalent to zero hard constraint violation. They showed an algorithm that guarantees no constraint violation, but their result relies on the assumption of some abstraction of the transition model, and moreover, there is no regret upper bound given for the algorithm. Liu et al. (2021) established the first algorithm that achieves a sublinear regret while guaranteeing zero hard constraint violation. After Liu et al. (2021), (Bura et al., 2022) proposed their algorithm, DOPE, which improves upon Liu et al. (2021) to show a smaller regret upper bound.

Appendix B. Auxiliary Measures and Notations

In this section, we first summarize notations in Table 2. Next, we define some auxiliary measures and notations that are useful for the analysis of DOPE+.
We define the *state-action value function* for $(s,a) \in \mathcal{S} \times \mathcal{A}$ at step h with a function $\ell : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ and transition kernel P as follows.

$$Q_h^\pi(s,a;\ell,P) = \mathbb{E}\left[\sum_{j=h}^{H} \ell\left(s_j^{P,\pi}, a_j^{P,\pi}, j\right) \mid \ell, \pi, P, s_h^{P,\pi} = s, a_h^{P,\pi} = a\right].$$

Let $\boldsymbol{Q}^{\boldsymbol{P},\boldsymbol{\pi},\boldsymbol{\ell}}$ denote the $(S \times A \times H)$-dimensional vector whose coordinates are for $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$(\boldsymbol{Q}^{\boldsymbol{P},\boldsymbol{\pi},\boldsymbol{\ell}})_{(s,a,h)} = Q_h^\pi(s,a;\ell,P).$$

Given a policy π and transition kernel P, we define $q^{P,\pi}(s,a,h \mid s', m)$ as for $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $1 \le m \le h \le H$,

$$q^{P,\pi}\left(s,a,h \mid s', m\right) = \mathbb{P}\left[s_h^{P,\pi} = s,\ a_h^{P,\pi} = a \mid \pi, P, s_m^{P,\pi} = s'\right].$$

Given two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{S\times A\times H}$, let $\boldsymbol{u} \odot \boldsymbol{v}$, $\boldsymbol{u} \wedge \boldsymbol{v}$ be defined as the vector obtained from coordinate-wise products and coordinate-wise minimization of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively, i.e., for $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$(\boldsymbol{u} \odot \boldsymbol{v})_{(s,a,h)} = \boldsymbol{u}_{(s,a,h)} \times \boldsymbol{v}_{(s,a,h)}, \quad (\boldsymbol{u} \wedge \boldsymbol{v})_{(s,a,h)} = \min\{\boldsymbol{u}_{(s,a,h)}, \boldsymbol{v}_{(s,a,h)}\}.$$

Let $\vec{\boldsymbol{h}}$ and $\vec{\boldsymbol{B}}$ be $(S \times A \times H)$-dimensional vectors all of whose coordinates are h and $1 + \sqrt{\ln(HSAK/\delta)}$, respectively, i.e., for $(s,a,j) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$\vec{\boldsymbol{h}}_{(s,a,j)} = j,\ \vec{\boldsymbol{B}}_{(s,a,j)} = 1 + \sqrt{\ln(HSAK/\delta)}.$$

Appendix C. Extended Linear Program

In this section, we provide a formal definition of occupancy measures for a finite-horizon MDP. Then we provide a reformulation of (8) using occupancy measures, which is called the extended linear program (Efroni et al., 2020; Bura et al., 2022).

Table 2: Summary of Notations

Notation	Definition
K	The number of episodes
H	The finite horizon
$[H]$	The set $\{1, 2, \ldots, H\}$
$\mathcal{S},\ S$	The finite state space $\mathcal{S}$ and the number of states $S = \|\mathcal{S}\|$
$\mathcal{A},\ A$	The finite action space $\mathcal{A}$ and the number of actions $A = \|\mathcal{A}\|$
P	The true transition kernel $P(s, a, s', h) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0, 1]$
p	The initial distribution of the states
$\mathcal{P}_k$	The confidence set of the transition kernel for episode $k \in [K]$
P_k	The transition kernel obtained from DOPE+ for episode $k \in [K]$, $P_k \in \mathcal{P}_k$
$f,\ g$	The reward and cost function
$f_k,\ g_k$	The instantaneous reward and cost for episode $k \in [K]$
$\bar{f}_k,\ \bar{g}_k$	The empirical estimators of f, g for episode $k \in [K]$
$\widehat{f}_k,\ \widehat{g}_k$	The optimistic/pessimistic estimators of f, g for episode $k \in [K]$
$V_h^\pi(s; f, P)$	The value function at state s and step h under f and P
$Q_h^\pi(s, a; f, P)$	The state-action value function at state s and step h for action a under f and P
$N_k(s, a, h)$	The number of visits (s, a, h) up to the first $k - 1$ episodes
$M_k(s, a, s', h)$	The number of visits (s, a, s', h) up to the first $k - 1$ episodes
$n_k(s, a, h)$	The indicator variable for visits (s, a, h) for episode $k \in [K]$
π^*	The benchmark policy
π_k	The policy obtained from DOPE+ for episode $k \in [K]$
π_b	The safe baseline policy
$\bar{C}_b$	The expected cost of π_b for a single episode
$\bar{C}$	The budget on the expected cost
$q^{P,\pi}$	The occupancy measure with respect to policy π and transition kernel P
q^*	The occupancy measure q^{P,π^*}
q_b	The occupancy measure q^{P,π_b}
q_k	The occupancy measure q^{P,π_k}
$\widehat{q}_k$	The occupancy measure q^{P_k,π_k}
$\Delta(P)$	The set of occupancy measures inducing any policy π and the true transition kernel P
$\Delta(P, k)$	The set of occupancy measures inducing any policy π and transition kernel $P_k \in \mathcal{P}_k$

Given a policy π and a transition kernel P, let $\bar{q}^{P,\pi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0, 1]$ be defined as $\bar{q}^{P,\pi}(s, a, s', h) = \mathbb{P}[(s_h^{P,\pi}, a_h^{P,\pi},\ s_{h+1}^{P,\pi}) = (s, a, s') \mid \pi, P]$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Note that any $\bar{q}$ defined as the above equation has the following properties. (C1) $\sum_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \bar{q}(s, a, s', h) = 1$, (C2) $\sum_{(s',a) \in \mathcal{S} \times \mathcal{A}} \bar{q}(s, a, s', h) = \sum_{(s',a) \in \mathcal{S} \times \mathcal{A}} \bar{q}(s', a, s, h-1)$, $\quad s \in \mathcal{S},\ h = 2, \ldots, H$. The *occupancy measure* $q^{P,\pi} : \mathcal{S} \times \mathcal{A} \times [H] \to [0, 1]$ associated with policy π and transition kernel P is defined as (C3) $q^{P,\pi}(s, a, h) = \sum_{s' \in \mathcal{S}} \bar{q}^{P,\pi}(s, a, s', h)$. Then it follows that $q^{P,\pi}(s, a, h) = \mathbb{P}[(s_h^{P,\pi}, a_h^{P,\pi}) = (s, a) \mid \pi, P]$. Hence, if a policy π is chosen, then the occupancy measure for a finite-horizon MDP with transition kernel P is determined. Conversely, any $q \in \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ with $\bar{q} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ satisfying (C1), (C2), and (C3) induces a transition kernel P^q and a policy π^q given as follows.

$$P^q(s' \mid s,a,h) = \frac{\bar{q}(s,a,s',h)}{\sum_{s'' \in \mathcal{S}} \bar{q}(s,a,s'',h)}, \quad \pi^q(a \mid s,h) = \frac{q(s,a,h)}{\sum_{b \in \mathcal{A}} q(s,b,h)}. \tag{10}$$

Next, we provide a lemma that characterizes valid occupancy measures for a finite-horizon MDP.

Lemma C.1 *Let $q : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$. Then q is a valid occupancy measure that induces transition kernel P if and only if there exists $\bar{q} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ that satisfies (C1), (C2), (C3), and $P^q = P$.*

Proof. Given the finite-horizon MDP associated with transition kernel P, we may define a loop-free MDP as follows. We define its state space as $\mathcal{S}' := \mathcal{S} \times [H+1]$, which can be viewed as $H+1$ layers $\mathcal{S} \times \{h\}$ for $h \in [H+1]$. Its transition kernel P' is given by $P'((s', h+1) \mid (s,h), a) = P(s' \mid s,a,h)$ for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Next, given $\bar{q}$, we may define an occupancy measure q' for the loop-free MDP as $q'((s,h), a, (s', h+1)) = \bar{q}(s,a,s',h)$ for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Then it follows from (Rosenberg and Mansour, 2019, Lemma 3.1) that q' is a valid occupancy measure for the loop-free MDP with transition kernel P' if and only if q' satisfies

$$\sum_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} q'((s,h), a, (s', h+1)) = 1 \quad \text{for } h = 1, \ldots, H, \tag{C1'}$$

$$\sum_{(s',a) \in \mathcal{S} \times \mathcal{A}} q'((s,h), a, (s', h+1)) = \sum_{(s',a) \in \mathcal{S} \times \mathcal{A}} q'((s', h-1), a, (s,h)) \quad \text{for } s \in \mathcal{S},\ h \in \{2, \ldots, H\}, \tag{C2'}$$

and $P^{q'} = P'$ where $P^{q'}$ is given by

$$P^{q'}((s', h+1) \mid (s,h), a) = \frac{q'((s,h), a, (s', h+1))}{\sum_{s'' \in \mathcal{S}} q'((s,h), a, (s'', h+1))} = \frac{\bar{q}(s,a,s',h)}{\sum_{s'' \in \mathcal{S}} \bar{q}(s,a,s'',h)}.$$

Here, the conditions are equivalent to (C1), (C2), and $P^{\bar{q}} = P$. Moreover, q' is a valid occupancy measure with P' if and only if q is a valid occupancy measure with P, as required. ■

Therefore, there is a one-to-one correspondence between the set of policies and the set of occupancy measures that give rise to transition kernel P. We define $\Delta(P)$ as the set of occupancy measures inducing the true transition kernel P.

$$\Delta(P) = \{\boldsymbol{q} : \ \exists \bar{\boldsymbol{q}} \text{ satisfying (C1),(C2),(C3)}, P^q = P\}.$$

Moreover, the value function for reward function f, policy π_k, and transition kernel P can be written in terms of occupancy measure q^{P,π_k} as $V_1^{\pi_k}(f,P) = \sum_{(s,a,h)} q^{P,\pi_k}(s,a,h) f(s,a,h)$. Let $\boldsymbol{q}^{\boldsymbol{P},\boldsymbol{\pi}}, \boldsymbol{f}$ denote $(S \times A \times H)$-dimensional vector representations for $q^{P,\pi}, f$, respectively.

Then it follows that $V_1^{\pi_k}(f,P) = \langle \boldsymbol{f}, \boldsymbol{q^{P,\pi_k}} \rangle$ where $\langle \cdot, \cdot \rangle$ is the inner product. Consequently, (8) is equivalent to

$$\max_{q \in \Delta(P,k)} \left\{ \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}}, \boldsymbol{q} \rangle : \langle \widehat{\boldsymbol{g}}_{\boldsymbol{k}}, \boldsymbol{q} \rangle \leq \bar{C} \right\} \tag{11}$$

where $\widehat{\boldsymbol{f}}_{\boldsymbol{k}}, \widehat{\boldsymbol{g}}_{\boldsymbol{k}}$ are the vector representations of $\widehat{f}_k, \widehat{g}_k$, respectively, and

$$\Delta(P,k) = \{\boldsymbol{q} : \ \exists \bar{\boldsymbol{q}} \text{ satisfying (C1),(C2),(C3)}, P^q \in \mathcal{P}_k\}.$$

Next, we reformulate (8) as an extended linear program. Due to the definition of $\Delta(P,k)$, (11) is equivalent to the following linear program. Given $\widehat{f}_k(s,a,h)$, $\widehat{g}_k(s,a,h)$, $\bar{P}_k(s' \mid s,a,h)$, $\epsilon_k(s' \mid s,a,h)$, $p(s)$ for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$,

$$\begin{aligned}
\max \quad & \sum_{(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} \widehat{f}_k(s,a,h)\bar{q}(s,a,s',h) \\
\text{s.t.} \quad & \sum_{(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} \widehat{g}_k(s,a,h)\bar{q}(s,a,s',h) \leq \bar{C}, \\
& \sum_{(a,s') \in \mathcal{A}\times\mathcal{S}} \bar{q}(s,a,s',h) = \sum_{(a,s') \in \mathcal{A}\times\mathcal{S}} \bar{q}(s',a,s,h-1) \quad \forall s \in \mathcal{S},\ h = 2,\ldots,H, \\
& \sum_{(a,s') \in \mathcal{A}\times\mathcal{S}} \bar{q}(s,a,s',1) = p(s) \quad \forall s \in \mathcal{S}, \\
& \bar{q}(s,a,s',h) \leq \left(\bar{P}_k(s' \mid s,a,h) + \epsilon_k(s' \mid s,a,h)\right) \sum_{s' \in \mathcal{S}} \bar{q}(s,a,s',h) \quad \forall (s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H], \\
& \bar{q}(s,a,s',h) \geq \left(\bar{P}_k(s' \mid s,a,h) - \epsilon_k(s' \mid s,a,h)\right) \sum_{s' \in \mathcal{S}} \bar{q}(s,a,s',h) \quad \forall (s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H], \\
& 0 \leq \bar{q}(s,a,s',h) \quad \forall (s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H].
\end{aligned}$$

In fact, the constraint $\sum_{(s,a,s')} \bar{q}(s,a,s',h) = 1$ for $h \in [H]$ corresponding to (C1) is not necessary, because we can derive it from other constraints. To be more specific, the third constraint implies that $\sum_{(s,a,s')} \bar{q}(s,a,s',1) = 1$ as $\sum_s p(s) = 1$. Then we can deduce from the second constraint that $\sum_{(s,a,s')} \bar{q}(s,a,s',h) = 1$ for $h \in [H]$. Additionally, we call the above linear program as an extended linear program due to the fifth and sixth constraints.

Appendix D. Good Event

In this section, we first prove Lemma 2.1 which ensures that all instantaneous reward and cost values are bounded. Then we prove Lemma 3.1 that describes important properties of the confidence sets estimating the true transition kernel. Next, we show Lemma 3.2 which delineates the accuracy of our estimators of the reward function f and the cost function g. Furthermore, we prove Lemma D.1 that is useful to bound value functions with respect to estimated reward and cost functions. Then we define the notion of the *good event* $\mathcal{E}$ that the statements of Lemmas 2.1, 3.1, 3.2 and D.1 hold. Taking the union bound, we deduce that the good event $\mathcal{E}$ holds with probability at least $1 - 14\delta$ (Lemma D.2).
Lastly, we prove Lemma D.3 which considers the difference between the true transition kernel and any $\widehat{P}$ contained in the confidence set $\mathcal{P}_k$.

Proof. [**Proof of Lemma 2.1**] It follows from Hoeffding's inequality (Lemma I.1) and the union bound that for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and $k\in[K]$,

$$\mathbb{P}\left(|f_k(s,a,h)-f(s,a,h)| \geq \sqrt{\ln(HSAK/\delta)}\right) \leq 2\cdot\exp\left(-\ln(HSAK/\delta)\right) = \frac{2\delta}{HSAK}.$$

Likewise, for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and $k\in[K]$,

$$\mathbb{P}\left(|g_k(s,a,h)-g(s,a,h)| \geq \sqrt{\ln(HSAK/\delta)}\right) \leq 2\cdot\exp\left(-\ln(HSAK/\delta)\right) = \frac{2\delta}{HSAK}.$$

Taking the union bound, it follows that with probability at least $1-4\delta$,

$$|f_k(s,a,h)-f(s,a,h)|\,, |g_k(s,a,h)-g(s,a,h)| \leq \sqrt{\ln(HSAK/\delta)}$$

holds for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ and $k\in[K]$. Since $f(s,a,h), g(s,a,h)\in[0,1]$ for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, it holds with probability at least $1-4\delta$ that

$$|f_k(s,a,h)|\,, |g_k(s,a,h)| \leq 1+\sqrt{\ln(HSAK/\delta)},$$

as required. ∎

The following lemma is a modification of (Jin et al., 2020, Lemma 8) to our finite-horizon MDP setting.
Proof. [**Proof of Lemma 3.1**] We will show that with probability at least $1-4\delta$,

$$\left|P(s'\mid s,a,h)-\bar{P}_k(s'\mid s,a,h)\right| \leq \epsilon_k(s'\mid s,a,h) \tag{12}$$

where

$$\epsilon_k(s'\mid s,a,h) = 2\sqrt{\frac{\bar{P}_k(s'\mid s,a,h)\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)-1\}}} + \frac{14\ln(HSAK/\delta)}{3\max\{1,N_k(s,a,h)-1\}}$$

holds for every $(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$ and every episode $k\in[K]$.
Let us first consider the case $N_k(s,a,h)\leq 1$. As we may assume that $HSAK\geq 2$, it follows that

$$\epsilon_k(s'\mid s,a,h) = \frac{14\ln(HSAK/\delta)}{3\max\{1,N_k(s,a,h)-1\}} \geq \frac{14}{3}\ln 2 > 1.$$

Then (12) holds because $0\leq P(s'\mid s,a,h), \bar{P}_k(s'\mid s,a,h)\leq 1$.
Assume that $n=N_k(s,a,h)\geq 2$. Then we define $Z_1,\dots,Z_n$ as follows.

$$Z_j = \begin{cases} 1, & \text{if the transition after the } j\text{th visit to } (s,a,h) \text{ is } s', \\ 0, & \text{otherwise.}\end{cases}$$

Then $Z_1,\dots,Z_n$ are i.i.d. with mean $P(s'\mid s,a,h)$, and we have

$$\sum_{j=1}^n Z_j = M_k(s,a,s',h).$$

Moreover, the sample variance V_n of $Z_1, \ldots, Z_n$ is given by

$$\begin{aligned} V_n &= \frac{1}{N_k(s,a,h)(N_k(s,a,h)-1)} M_k(s,a,s',h)\left(N_k(s,a,h) - M_k(s,a,s',h)\right) \\ &= \frac{N_k(s,a,h)}{(N_k(s,a,h)-1)} \bar{P}_k(s' \mid s,a,h)\left(1 - \bar{P}_k(s' \mid s,a,h)\right). \end{aligned} \tag{13}$$

Then it follows from Lemma I.2 that with probability at least $1 - 2\delta/(HS^2AK)$,

$$P(s' \mid s,a,h) - \bar{P}_k(s' \mid s,a,h) \leq \sqrt{\frac{2\bar{P}_k(s' \mid s,a,h)\left(1 - \bar{P}_k(s' \mid s,a,h)\right)\ln\left(HS^2AK/\delta\right)}{N_k(s,a,h)-1}} + \frac{7\ln\left(HS^2AK/\delta\right)}{3(N_k(s,a,h)-1)}. \tag{14}$$

Here, as we assumed that $N_k(s,a,h) \geq 2$, we have $N_k(s,a,h) - 1 = \max\{1, N_k(s,a,h)-1\}$. In addition, we know that $1 - \bar{P}_k(s' \mid s,a,h) \leq 1$ and that $\ln\left(HS^2AK/\delta\right) \leq 2\ln\left(HSAK/\delta\right)$. Then (14) implies that with probability at least $1 - 2\delta/(HS^2AK)$,

$$P(s' \mid s,a,h) - \bar{P}_k(s' \mid s,a,h) \leq \epsilon_k(s' \mid s,a,h). \tag{15}$$

Next, we apply Lemma I.2 to variables $1 - Z_1, \ldots, 1 - Z_n$ that are i.i.d. and have mean $1 - \bar{P}_k(s' \mid s,a,h)$. Moreover, the sample variance of $1 - Z_1, \ldots, 1 - Z_n$ is also equal to V_n defined as in (13). Therefore, based on the same argument, we deduce that with probability at least $1 - 2\delta/(HS^2AK)$,

$$-P(s' \mid s,a,h) + \bar{P}_k(s' \mid s,a,h) \leq \epsilon_k(s' \mid s,a,h). \tag{16}$$

By applying union bound to (15) and (16), with probability at least $1 - 4\delta/(HS^2AK)$, (12) holds for (s,a,s',h). Furthermore, by applying union bound over all $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, it follows that with probability at least $1 - 4\delta$, (12) holds for every $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, as required. ■

Next, we state the proof of Lemma 3.2 based on Hoeffding's inequality.

Proof. [**Proof of Lemma 3.2**] If $N_k(s,a,h) = \sum_{j=1}^{k-1} n_j(s,a,h) = 0$, then $\bar{f}_k(s,a,h) = \bar{g}_k(s,a,h) = 0$ while $R_k(s,a,h) \geq 1$ when we may assume that $HSAK \geq 4$. In this case, the statements trivially hold. Now we consider when $\sum_{j=1}^{k-1} n_j(s,a,h) \geq 1$. Note that $f_k(s,a,h) = f(s,a,h) + \zeta_k^f(s,a,h)$ and $g_k(s,a,h) = g(s,a,h) + \zeta_k^g(s,a,h)$ where $\zeta_k^f(s,a,h)$ and $\zeta_k^g(s,a,h)$ are i.i.d. 1/2-sub-Gaussian random variables with zero mean for each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$. Then it follows from the Hoeffding's inequality provided in Lemma I.1 that for a given $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$,

$$\left|\bar{f}_k(s,a,h) - f(s,a,h)\right| \leq R_k(s,a,h) \tag{17}$$

with probability at least $1 - 2\delta/(HSAK)$. By applying union bound, (17) holds with probability at least $1 - 2\delta$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$. Likewise, we deduce for g that with probability at least $1 - 2\delta$,

$$\left|\bar{g}_k(s,a,h) - g(s,a,h)\right| \leq R_k(s,a,h)$$

for $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$ as desired. ■

Next, using Lemma I.3 that states the Bernstein-type concentration inequality for a martingale difference sequence, we prove the following lemma that is useful for our analysis. Lemma D.1 is a modification of (Jin et al., 2020, Lemma 10) and (Chen and Luo, 2021, Lemma 8) to our finite-horizon MDP setting.

Lemma D.1 *With probability at least $1-2\delta$, we have*

$$\sum_{k=1}^{K} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{q_k(s,a,h)}{\max\{1, N_k(s,a,h)\}} \le 2HSA\ln K + 2HSA + 4H\ln(H/\delta) \tag{18}$$

$$\sum_{k=1}^{K} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{q_k(s,a,h)}{\sqrt{\max\{1, N_k(s,a,h)\}}} \le 2H\sqrt{SAK} + 2HSA\ln K + 3HSA + 5H\ln(H/\delta) \tag{19}$$

Proof. We define ξ_1 as $\xi_1 = \emptyset$ and for $k \ge 2$, we define ξ_k as

$$\left\{ s_h^{P,\pi_{k-1}}, a_h^{P,\pi_{k-1}}, f_{k-1}(s_h^{P,\pi_{k-1}}, a_h^{P,\pi_{k-1}}, h), g_{k-1}(s_h^{P,\pi_{k-1}}, a_h^{P,\pi_{k-1}}, h) \right\}_{h=1}^{H}$$

where π_{k-1} denotes the policy for episode $k-1$ and

$$\left(s_1^{P,\pi_{k-1}}, a_1^{P,\pi_{k-1}}, \ldots, s_h^{P,\pi_{k-1}}, a_h^{P,\pi_{k-1}} \right)$$

is the trajectory generated under policy π_{k-1} and transition kernel P. Then for $k \in [K]$, let $\mathcal{F}_k$ be defined as the σ-algebra generated by the random variables in $\xi_1 \cup \cdots \cup \xi_k$. Then it follows that $\mathcal{F}_1, \ldots, \mathcal{F}_k$ give rise to a filtration.
Note that

$$\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_k(s,a,h)}{\max\{1, N_k(s,a,h)\}} = \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_k(s,a,h)}{\max\{1, N_k(s,a,h)\}} + \sum_{k=1}^{K} Y_k \tag{20}$$

where

$$Y_k = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{-n_k(s,a,h) + q_k(s,a,h)}{\max\{1, N_k(s,a,h)\}}.$$

As $\mathbb{E}[n_k(s,a,h) \mid \pi_k, P] = q_k(s,a,h)$ holds for every $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, we know that $Y_1, \ldots, Y_K$ is a martingale difference sequence. We know that $Y_k \le 1$ for each $k \in [K]$. Let $\mathbb{E}_k[\cdot]$ denote $\mathbb{E}[\cdot \mid \mathcal{F}_k, P]$. Since π_k is $\mathcal{F}_k$-measurable, we have $\mathbb{E}_k[n_k(s,a,h)] = q_k(s,a,h)$.

Then we deduce

$$
\begin{aligned}
\mathbb{E}_k\left[Y_k^2\right] &= \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_k\left[(n_k(s,a,h)-q_k(s,a,h))(n_k(s',a',h)-q_k(s',a',h))\right]}{\max\{1,N_k(s,a,h)\}\cdot\max\{1,N_k(s',a',h)\}} \\
&= \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_k\left[n_k(s,a,h)n_k(s',a',h)-q_k(s,a,h)q_k(s',a',h)\right]}{\max\{1,N_k(s,a,h)\}\cdot\max\{1,N_k(s',a',h)\}} \\
&\le \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_k\left[n_k(s,a,h)n_k(s',a',h)\right]}{\max\{1,N_k(s,a,h)\}\cdot\max\{1,N_k(s',a',h)\}} \\
&\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_k\left[n_k(s,a,h)\right]}{\max\{1,N_k(s,a,h)\}} \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}
\end{aligned}
$$

where the second equality holds because it follows from $\mathbb{E}_k[n_k(s,a,h)] = q_k(s,a,h)$ for $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ that

$$\mathbb{E}_k\left[q_k(s,a,h)n_k(s',a',h)\right] = \mathbb{E}_k\left[q_k(s',a',h)n_k(s,a,h)\right] = q_k(s,a,h)q_k(s',a',h),$$

the second inequality holds because $n_k(s,a,h)n_k(s',a',h) = 0$ if $(s,a)\neq(s',a')$, and the last equality holds true because $\mathbb{E}_k[n_k(s,a,h)] = q_k(s,a,h)$ for any $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$. Then we may apply Lemma I.3 with $\lambda = 1/2$, and we deduce that with probability at least $1-\delta/H$,

$$\sum_{k=1}^K Y_k \le \frac{1}{2}\sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}} + 2\ln(H/\delta).$$

Plugging this inequality to (20), it follows that

$$\sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}} = 2\sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_k(s,a,h)}{\max\{1,N_k(s,a,h)\}} + 4\ln(H/\delta).$$

Here, the first term on the right-hand side can be bounded as follows. We have

$$
\begin{aligned}
\sum_{k=1}^K \frac{n_k(s,a,h)}{\max\{1,N_k(s,a,h)\}} &= \sum_{k=1}^K \frac{n_k(s,a,h)}{\max\{1,N_{k+1}(s,a,h)\}} + \sum_{k=1}^K \left(\frac{n_k(s,a,h)}{\max\{1,N_k(s,a,h)\}} - \frac{n_k(s,a,h)}{\max\{1,N_{k+1}(s,a,h)\}}\right) \\
&\le \sum_{k=1}^K \frac{n_k(s,a,h)}{\max\{1,N_{k+1}(s,a,h)\}} + \sum_{k=1}^K \left(\frac{1}{\max\{1,N_k(s,a,h)\}} - \frac{1}{\max\{1,N_{k+1}(s,a,h)\}}\right) \\
&\le \sum_{k=1}^K \frac{n_k(s,a,h)}{\max\{1,N_{k+1}(s,a,h)\}} + 1 \\
&\le \ln K + 1.
\end{aligned}
$$

where the first inequality is due to $n_k(s,a,h)\le 1$ and the last inequality holds because

$$n_k(s,a,h) = N_{k+1}(s,a,h) - N_k(s,a,h) \quad \text{and} \quad N_K(s,a,h) + n_K(s,a,h) \le K.$$

Therefore, it follows that

$$\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_k(s,a,h)}{\max\{1, N_k(s,a,h)\}} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{k=1}^{K} \frac{n_k(s,a,h)}{\max\{1, N_k(s,a,h)\}} = SA\ln K + SA.$$

As a result, for any fixed $h \in [H]$,

$$\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_k(s,a,h)}{\max\{1, N_k(s,a,h)\}} \le 2SA\ln K + 2SA + 4\ln(H/\delta)$$

holds with probability at least $1 - \delta/H$. By union bound, (18) holds with probability at least $1 - \delta$.

Next, we will show that (19) holds.

$$\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_k(s,a,h)}{\sqrt{\max\{1, N_k(s,a,h)\}}} = \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{n_k(s,a,h)}{\sqrt{\max\{1, N_k(s,a,h)\}}} + \sum_{k=1}^{K} Z_k \quad (21)$$

where

$$Z_k = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{-n_k(s,a,h) + q_k(s,a,h)}{\sqrt{\max\{1, N_k(s,a,h)\}}}.$$

As $\mathbb{E}_k[n_k(s,a,h)] = q_k(s,a,h)$ holds for every $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, we know that $Z_1, \ldots, Z_K$ is a martingale difference sequence. We know that $Z_k \le 1$ for each $k \in [K]$. Then we deduce

$$\begin{aligned}
\mathbb{E}_k\left[Z_k^2\right] &\le \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_k[n_k(s,a,h)n_k(s',a',h)]}{\sqrt{\max\{1, N_k(s,a,h)\}} \cdot \sqrt{\max\{1, N_k(s',a',h)\}}} \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}_k[n_k(s,a,h)]}{\max\{1, N_k(s,a,h)\}} \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_k(s,a,h)}{\max\{1, N_k(s,a,h)\}}
\end{aligned}$$

where the first inequality is derived by the same argument when bounding $\mathbb{E}_k[Y_k^2]$, the first equality holds because $n_k(s,a,h)n_k(s',a',h) = 0$ if $(s,a) \ne (s',a')$, and the last equality holds true because $\mathbb{E}_k[n_k(s,a,h)] = q_k(s,a,h)$ for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. Then we may apply Lemma I.3 with $\lambda = 1$, and we deduce that with probability at least $1 - \delta/H$,

$$\sum_{k=1}^{K} Z_k \le \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{q_k(s,a,h)}{\max\{1, N_k(s,a,h)\}} + \ln(H/\delta).$$

Then with probability at least $1 - \delta$, (18) holds and

$$\begin{aligned}
\sum_{h\in[H]} \sum_{k=1}^{K} Z_k &\le \sum_{k=1}^{K} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{q_k(s,a,h)}{\max\{1, N_k(s,a,h)\}} + H\ln(H/\delta) \\
&= 2HSA\ln K + 2HSA + 5H\ln(H/\delta).
\end{aligned} \quad (22)$$

holds. Moreover, we have

$$
\begin{aligned}
&\sum_{k=1}^{K} \frac{n_k(s,a,h)}{\sqrt{\max\{1, N_k(s,a,h)\}}} \\
&= \sum_{k=1}^{K} \frac{n_k(s,a,h)}{\sqrt{\max\{1, N_{k+1}(s,a,h)\}}} + \sum_{k=1}^{K} \left(\frac{n_k(s,a,h)}{\sqrt{\max\{1, N_k(s,a,h)\}}} - \frac{n_k(s,a,h)}{\sqrt{\max\{1, N_{k+1}(s,a,h)\}}} \right) \\
&\leq \sum_{k=1}^{K} \frac{n_k(s,a,h)}{\sqrt{\max\{1, N_{k+1}(s,a,h)\}}} + \sum_{k=1}^{K} \left(\frac{1}{\sqrt{\max\{1, N_k(s,a,h)\}}} - \frac{1}{\sqrt{\max\{1, N_{k+1}(s,a,h)\}}} \right) \\
&\leq \sum_{k=1}^{K} \frac{n_k(s,a,h)}{\sqrt{\max\{1, N_{k+1}(s,a,h)\}}} + 1 \\
&\leq 2\sqrt{N_{K+1}(s,a,h)} + 1.
\end{aligned}
$$

where the last equality holds because $n_k(s,a,h) = N_{k+1}(s,a,h) - N_k(s,a,h)$. Then

$$
\begin{aligned}
\sum_{k=1}^{K} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{n_k(s,a,h)}{\sqrt{\max\{1, N_k(s,a,h)\}}} &\leq \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} 2\sqrt{N_{K+1}(s,a,h)} + HSA \\
&\leq 2\sqrt{HSA \sum_{(s,a,h)} N_{K+1}(s,a,h)} + HSA \\
&\leq 2H\sqrt{SAK} + HSA
\end{aligned}
$$

where the second equality is due to the Cauchy-Schwarz inequality. Then it follows from (21) and (22) that (19) holds. ■

Recall that the good event $\mathcal{E}$ is the event that the statements of Lemmas 2.1, 3.1, 3.2 and D.1 hold.

Lemma D.2 *The good event $\mathcal{E}$ holds with probability at least $1 - 14\delta$, i.e., $\mathbb{P}[\mathcal{E}] \geq 1 - 14\delta$.*

Proof. The proof follows from the union bound. ■

Lemma 3.1 bounds the difference between the true transition kernel P and the empirical transition kernel $\bar{P}_k$. Based on Lemma 3.1, the next lemma bounds the difference between the true transition kernel and any $\widehat{P}$ contained in the confidence set $\mathcal{P}_k$. Lemma D.3 is a modification of (Jin et al., 2020, Lemma 8) to our finite-horizon MDP setting.

Lemma D.3 *Under the good event $\mathcal{E}$, we have*

$$
\left| \widehat{P}(s' \mid s,a,h) - P(s' \mid s,a,h) \right| \leq \epsilon_k^\star(s' \mid s,a,h) \tag{23}
$$

where

$$
\epsilon_k^\star(s' \mid s,a,h) = 6\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + 94\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}
$$

for every $\widehat{P} \in \mathcal{P}_k$ *and every* $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.

Proof. We follow the proof of (Cohen et al., 2020, Lemma B.13). Note that

$$\max\{1, N_k(s,a,h) - 1\} \geq \frac{1}{2} \cdot \max\{1, N_k(s,a,h)\}$$

holds for any value of $N_k(s,a,h)$. As we assumed that $P \in \mathcal{P}_k$, we have that

$$\bar{P}_k(s' \mid s,a,h) \leq P(s' \mid s,a,h) + \sqrt{\frac{8\bar{P}_k(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + \frac{28\ln(HSAK/\delta)}{3\max\{1, N_k(s,a,h)\}}.$$

We may view this as a quadratic inequality in terms of $x = \sqrt{\bar{P}_k(s' \mid s,a,h)}$. Note that $x^2 \leq ax + b + c$ for any $a, b, c \geq 0$ implies that $x \leq a + \sqrt{b} + \sqrt{c}$. Therefore, we deduce that

$$\begin{aligned}
\sqrt{\bar{P}_k(s' \mid s,a,h)} &\leq \sqrt{P(s' \mid s,a,h)} + \left(2\sqrt{2} + \sqrt{\frac{28}{3}}\right)\sqrt{\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} \\
&\leq \sqrt{P(s' \mid s,a,h)} + 13\sqrt{\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}}.
\end{aligned}$$

Using this bound on $\sqrt{\bar{P}_k(s' \mid s,a,h)}$, we obtain the following.

$$\begin{aligned}
\epsilon_k(s' \mid s,a,h) &\leq \sqrt{\frac{8\bar{P}_k(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + \frac{28\ln(HSAK/\delta)}{3\max\{1, N_k(s,a,h)\}} \\
&\leq \sqrt{\frac{8P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + \left(13\sqrt{8} + \frac{28}{3}\right)\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}} \qquad (24) \\
&\leq 3\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + 47\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}} \\
&= \frac{1}{2} \cdot \epsilon_k^{\star}(s' \mid s,a,h)
\end{aligned}$$

Since we assumed that $P \in \mathcal{P}_k$,

$$\left|P(s' \mid s,a,h) - \bar{P}_k(s' \mid s,a,h)\right| \leq \frac{1}{2} \cdot \epsilon_k^{\star}(s' \mid s,a,h).$$

Moreover, for any $\widehat{P} \in \mathcal{P}_k$, we have

$$\left|\widehat{P}(s' \mid s,a,h) - \bar{P}_k(s' \mid s,a,h)\right| \leq \epsilon_k(s' \mid s,a,h) \leq \frac{1}{2} \cdot \epsilon_k^{\star}(s' \mid s,a,h).$$

By the triangle inequality, it follows that

$$\left|\widehat{P}(s' \mid s,a,h) - P(s' \mid s,a,h)\right| \leq \epsilon_k^{\star}(s' \mid s,a,h),$$

as required. ∎

We note that the above lemma holds when we replace $P(s' \mid s,a,h)$ of $\epsilon_k^\star(s' \mid s,a,h)$ into $\widehat{P}(s' \mid s,a,h)$ for any $\widehat{P} \in \mathcal{P}_k$. Specifically, under the good event $\mathcal{E}$, we have for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$,

$$\left|\widehat{P}(s' \mid s,a,h) - P(s' \mid s,a,h)\right| \leq 6\sqrt{\frac{\widehat{P}(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + 94\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}. \tag{25}$$

It can be obtained by applying

$$\bar{P}_k(s' \mid s,a,h) \leq \widehat{P}(s' \mid s,a,h) + \sqrt{\frac{8\bar{P}_k(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + \frac{28\ln(HSAK/\delta)}{3\max\{1, N_k(s,a,h)\}}$$

with the same argument for the remaining part of the proof.

Appendix E. Missing Proofs for Section 3: Tighter Function Estimators

In this section, we first show Lemma 3.3 that bounds the expected sum of the variance values of value function estimates.

Proof. [**Proof of Lemma 3.3**] Let π_k be a policy for episode k. Moreover, let $P_k \in \mathcal{P}_k$, and let $g : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ be an arbitrary cost function. Then we may define the occupancy measure $\widehat{q}_k = q^{P_k,\pi_k}$ associated with policy π_k and transitional kernel P_k. Then we know that $V_1^{\pi_k}(\widehat{\mathbb{V}}_k, P_k) = \langle \widehat{\boldsymbol{q}}_{\boldsymbol{k}}, \widehat{\mathbb{V}}_{\boldsymbol{k}}\rangle$. Moreover, it follows from Lemma H.5 that

$$\langle \widehat{\boldsymbol{q}}_{\boldsymbol{k}}, \widehat{\mathbb{V}}_{\boldsymbol{k}}\rangle \leq \mathrm{Var}\left[\langle \widehat{\boldsymbol{n}}_{\boldsymbol{k}}, \boldsymbol{g}\rangle \mid g, \pi_k, P_k\right]$$

where $\widehat{\boldsymbol{n}}_{\boldsymbol{k}}$ is a vector representation of $\widehat{n}_k = n^{P_k,\pi_k}$. Furthermore, by Lemma H.1 with $B = 1$, we have

$$\begin{aligned}
\mathrm{Var}\left[\langle \widehat{\boldsymbol{n}}_{\boldsymbol{k}}, \boldsymbol{g}\rangle \mid g, \pi_k, P_k\right] &\leq \mathbb{E}[\langle \widehat{\boldsymbol{n}}_{\boldsymbol{k}}, \boldsymbol{g}\rangle^2 \mid g, \pi_k, P_k] \\
&\leq 2\langle \widehat{\boldsymbol{q}}_{\boldsymbol{k}}, \vec{\boldsymbol{h}} \odot \boldsymbol{g}\rangle \\
&\leq 2H^2
\end{aligned}$$

as desired. ■

Having proved Lemma 3.3, we are ready to prove Theorem 1 which is the crucial part of deducing our tighter function estimators.

Proof. [**Proof of Theorem 1**] We assume that the good event $\mathcal{E}$ holds, which holds with probability at least $1-14\delta$ according to Lemma D.2. We observe that $|V_1^{\pi_k}(g,P) - V_1^{\pi_k}(g,P_k)|$ can be rewritten by $|\langle \boldsymbol{g}, \boldsymbol{q}_{\boldsymbol{k}} - \widehat{\boldsymbol{q}}_{\boldsymbol{k}}\rangle|$ using occupancy measures. By Lemma H.3, it follows that

$$
\begin{aligned}
|\langle \boldsymbol{g}, \boldsymbol{q_k} - \widehat{\boldsymbol{q}}_{\boldsymbol{k}}\rangle| &= \left| \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} \widehat{q}_k(s,a,h)\left(P-P_k\right)\left(s' \mid s,a,h\right) V^{\pi_k}_{h+1}(s';g,P) \right| \\
&\leq \underbrace{\left| \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} \widehat{q}_k(s,a,h)(P-P_k)(s' \mid s,a,h) V^{\pi_k}_{h+1}(s';g,P_k) \right|}_{\text{Term 1}} \\
&\quad + \underbrace{\left| \sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} \widehat{q}_k(s,a,h)(P-P_k)(s' \mid s,a,h)\left(V^{\pi_k}_{h+1}(s';g,P) - V^{\pi_k}_{h+1}(s';g,P_k)\right) \right|}_{\text{Term 2}}
\end{aligned}
$$

where $(P-P_k)(s' \mid s,a,h) = P(s' \mid s,a,h) - P_k(s' \mid s,a,h)$.
To bound Term 2, we use bound

$$
P(s' \mid s,a,h) - P_k(s' \mid s,a,h) \leq 6\sqrt{\frac{P_k(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + 94\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}
$$

as explained in (25). This is because $\widehat{q}_k = q^{P_k,\pi_k}$ is an occupancy measure with respect to $P_k \in \mathcal{P}_k$, not P. Then we can apply Lemma H.6 and obtain

$$
\text{Term 2} \leq 10^4 H^2 S^2 \left(\ln\frac{HSAK}{\delta}\right)^2 \sum_{(s,a,h)} \frac{\widehat{q}_k(s,a,h)}{\max\{1, N_k(s,a,h)\}}.
$$

Next, we bound Term 1. Note that $\sum_{s'}\left(P(s' \mid s,a,h) - P_k(s' \mid s,a,h)\right) = 0$. Then it follows that

$$
\begin{aligned}
\text{Term 1} &= \left| \sum_{(s,a,s',h)} \widehat{q}_k(s,a,h)(P-P_k)(s' \mid s,a,h)(V^{\pi_k}_{h+1}(g,P_k) - \widehat{\mu}_k(s,a,h)) \right| \\
&\leq 2 \sum_{(s,a,s',h)} \widehat{q}_k(s,a,h)\epsilon_k(s' \mid s,a,h)\left|V^{\pi_k}_{h+1}(g,P_k) - \widehat{\mu}_k(s,a,h)\right| \\
&= 4 \underbrace{\sum_{(s,a,s',h)} \widehat{q}_k(s,a,h)\sqrt{\frac{\bar{P}_k(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)-1\}}}\left|V^{\pi_k}_{h+1}(s';g,P_k) - \widehat{\mu}_k(s,a,h)\right|}_{\text{Term 3}} \\
&\quad + \frac{28}{3}\underbrace{\sum_{(s,a,s',h)} \widehat{q}_k(s,a,h)\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)-1\}}\left|V^{\pi_k}_{h+1}(s';g,P) - \widehat{\mu}_k(s,a,h)\right|}_{\text{Term 4}}
\end{aligned}
$$

where $\widehat{\mu}_k(s,a,h) = \mathbb{E}_{s'\sim P_k(\cdot|s,a,h)}[V^{\pi_k}_{h+1}(s';g,P_k)]$. The first inequality is from $|(P-P_k)(s' \mid s,a,h)| \leq |(P-\bar{P}_k)(s' \mid s,a,h)| + |(\bar{P}_k - P_k)(s' \mid s,a,h)| \leq 2\epsilon_k(s' \mid s,a,h)$ for any $(s,a,s',h) \in$

$\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$ under the good event $\mathcal{E}$. We note that $\bar{P}_k(s' \mid s,a,h) \le P_k(s' \mid s,a,h)+\epsilon_k(s' \mid s,a,h)$ and define

$$\widehat{\mathbb{V}}_k(s,a,h) = \sum_{s'} P_k(s' \mid s,a,h) \left|V_{h+1}^{\pi_k}(s';g,P_k) - \widehat{\mu}_k(s,a,h)\right|^2.$$

Then we can bound Term 3 as the following.

$$\begin{aligned}
&\frac{\text{Term 3}}{\sqrt{\ln(HSAK/\delta)}} \\
&\le \sum_{(s,a,s',h)} \widehat{q}_k(s,a,h)\sqrt{\frac{(P_k+\epsilon_k)(s' \mid s,a,h)}{\max\{1,N_k(s,a,h)-1\}}}\left|V_{h+1}^{\pi_k}(s';g,P_k) - \widehat{\mu}_k(s,a,h)\right| \\
&\le \sqrt{\sum_{(s,a,s',h)} \widehat{q}_k(s,a,h)(P_k+\epsilon_k)(s' \mid s,a,h)\left|V_{h+1}^{\pi_k}(s';g,P_k) - \widehat{\mu}_k(s,a,h)\right|^2}\sqrt{\sum_{(s,a,s',h)}\frac{\widehat{q}_k(s,a,h)}{\max\{1,N_k(s,a,h)-1\}}} \\
&\le \sqrt{\sum_{(s,a,h)} \widehat{q}_k(s,a,h)\widehat{\mathbb{V}}_k(s,a,h) + 4H^2\sum_{(s,a,s',h)}\widehat{q}_k(s,a,h)\epsilon_k(s' \mid s,a,h)}\sqrt{\sum_{(s,a,s',h)}\frac{\widehat{q}_k(s,a,h)}{\max\{1,N_k(s,a,h)-1\}}}
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and the last inequality is due to $\left|V_{h+1}^{\pi_k}(s';g,P_k) - \widehat{\mu}_k(s,a,h)\right| \le 2H$.
By Lemma 3.3, we deduce that

$$\sum_{(s,a,h)} \widehat{q}_k(s,a,h)\widehat{\mathbb{V}}_k(s,a,h) \le 2H^2.$$

Due to AM-GM inequality, we have

$$\begin{aligned}
&\sqrt{2H^2 + 4H^2\sum_{(s,a,s',h)}\widehat{q}_k(s,a,h)\epsilon_k(s' \mid s,a,h)}\sqrt{\sum_{(s,a,s',h)}\frac{\widehat{q}_k(s,a,h)}{\max\{1,N_k(s,a,h)-1\}}} \\
&\le \left(\sqrt{2H^2} + \sqrt{4H^2\sum_{(s,a,s',h)}\widehat{q}_k(s,a,h)\epsilon_k(s' \mid s,a,h)}\right)\sqrt{\sum_{(s,a,s',h)}\frac{\widehat{q}_k(s,a,h)}{\max\{1,N_k(s,a,h)-1\}}} \\
&\le \frac{H^2}{\alpha_1} + \frac{2H^2}{\alpha_2}\sum_{(s,a,s',h)}\widehat{q}_k(s,a,h)\epsilon_k(s' \mid s,a,h) + \frac{\alpha_1+\alpha_2}{2}\sum_{(s,a,h)}\frac{S\cdot\widehat{q}_k(s,a,h)}{\max\{1,N_k(s,a,h)-1\}}
\end{aligned}$$

for any $\alpha_1,\alpha_2>0$. By taking $\alpha_1 = \frac{\sqrt{HK\ln(HSAK/\delta)}}{S\sqrt{A}}$, $\alpha_2 = \sqrt{H^3\ln(HSAK/\delta)}$, we obtain

$$\begin{aligned}
\text{Term 3} &\le \sum_{(s,a,h)}\widehat{q}_k(s,a,h)\left(\frac{S\sqrt{HA}}{\sqrt{K}} + 2\sqrt{H}\sum_{s'}\epsilon_k(s' \mid s,a,h) + \frac{\sqrt{HK}+\sqrt{H^3S^2A}}{2\sqrt{A}}\frac{\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)-1\}}\right) \\
&\le \sum_{(s,a,h)}\widehat{q}_k(s,a,h)\left(\frac{S\sqrt{HA}}{\sqrt{K}} + 2\sqrt{H}\varepsilon_k(s,a,h) + \frac{\sqrt{HK}+\sqrt{H^3S^2A}}{2\sqrt{A}}\frac{\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)-1\}}\right).
\end{aligned}$$

Note that the last inequality follows from

$$\begin{aligned}
\sum_{s'} \epsilon_k(s' \mid s,a,h) &= \sum_{s'} \left(\sqrt{\frac{4\bar{P}_k(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)-1\}}} + \frac{14\ln(HSAK/\delta)}{3\max\{1, N_k(s,a,h)-1\}} \right) \\
&\le \sqrt{S}\sqrt{\frac{4\sum_{s'}\bar{P}_k(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)-1\}}} + \frac{14S\ln(HSAK/\delta)}{3\max\{1, N_k(s,a,h)-1\}} \\
&= \sqrt{\frac{4S\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)-1\}}} + \frac{14S\ln(HSAK/\delta)}{3\max\{1, N_k(s,a,h)-1\}} \\
&= \varepsilon_k(s,a,h)
\end{aligned}$$

where the inequality is due to the Cauchy-Schwarz inequality and the second equality is due to $\sum_{s'} \bar{P}_k(s' \mid s,a,h) \le 1$.
Since $\left|V^{\pi_k}_{h+1}(s'; g, P) - \widehat{\mu}_k(s,a,h)\right| \le 2H$, Term 4 can be bounded as follows.

$$\text{Term } 4 \le 2HS\ln(HSAK/\delta) \sum_{(s,a,h)} \frac{\widehat{q}_k(s,a,h)}{\max\{1, N_k(s,a,h)-1\}}.$$

Finally, we proved that

$$\begin{aligned}
&|\langle \boldsymbol{g}, \boldsymbol{q_k} - \widehat{\boldsymbol{q}}_{\boldsymbol{k}} \rangle| \\
&\le 4 \cdot (\text{Term } 3) + \frac{28}{3} \cdot (\text{Term } 4) + (\text{Term } 2) \\
&\le \sum_{(s,a,h)} \widehat{q}_k(s,a,h) \left(\frac{4S\sqrt{HA}}{\sqrt{K}} + 8\sqrt{H}\varepsilon_k(s,a,h) + \frac{2\sqrt{HK}\ln(HSAK/\delta)}{\sqrt{A}\max\{1, N_k(s,a,h)-1\}} \right) \\
&\quad + \left(\left(\frac{56}{3}HS + 2H^{1.5}S \right) \ln(HSAK/\delta) + 10^4 H^2 S^2 (\ln(HSAK/\delta))^2 \right) \sum_{(s,a,h)} \frac{\widehat{q}_k(s,a,h)}{\max\{1, N_k(s,a,h)-1\}}
\end{aligned}$$

as required. ■

Appendix F. Missing Proofs for Section 4: Safe Exploration

In this section, we prove Lemma 4.1 that provides an asymptotic upper bound on a sufficient number of episodes executing π_b, which is denoted by K_0, for feasibility of (8).

Lemma F.1 *Assume that the good event $\mathcal{E}$ holds. Let π_k be any policy for episode k, and let P be the true transition kernel. Let q_k denote the occupancy measure q^{P,π_k} associated with π_k and P. For R_k, U_k, we have*

$$\sum_{k=1}^{K} \langle \boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q_k} \rangle = \mathcal{O}\left(\left(H^{1.5}S\sqrt{AK} + H^3S^3A \right) \left(\ln \frac{HSAK}{\delta} \right)^3 \right).$$

Proof.
Note that $\sum_{k=1}^{K}\langle \boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_k}\rangle$ can be rewritten as

$$\begin{aligned}
&\sum_{k=1}^{K}\langle \boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_k}\rangle\\
&=\sum_{k=1}^{K}\sum_{(s,a,h)} q_k(s,a,h)\sqrt{\frac{\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)\}}}\\
&\quad+\sum_{k=1}^{K}\sum_{(s,a,h)} q_k(s,a,h)\left(\frac{4S\sqrt{HA}}{\sqrt{K}}+8\sqrt{H}\varepsilon_k(s,a,h)+\frac{2(\sqrt{HK}+\sqrt{H^3S^2A})\ln(HSAK/\delta)}{\sqrt{A}\max\{1,N_k(s,a,h)-1\}}\right)\\
&\quad+\left(\frac{56}{3}HS\ln(HSAK/\delta)+10^4H^2S^2(\ln(HSAK/\delta))^2\right)\sum_{k=1}^{K}\sum_{(s,a,h)}\frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)-1\}}.
\end{aligned}$$

Since $\sum_{(s,a,h)}\widehat{q}_k(s,a,h)=H$, we have

$$\sum_{k=1}^{K}\sum_{(s,a,h)} q_k(s,a,h)\cdot\frac{4S\sqrt{HA}}{\sqrt{K}}=\mathcal{O}(H^{1.5}S\sqrt{AK}).$$

Furthermore, Lemma D.1 implies that

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{(s,a,h)}\frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}&=\mathcal{O}(HSA\ln K+H\ln(H/\delta)),\\
\sum_{k=1}^{K}\sum_{(s,a,h)}\frac{q_k(s,a,h)}{\sqrt{\max\{1,N_k(s,a,h)\}}}&=\mathcal{O}(H\sqrt{SAK}+HSA\ln K+H\ln(H/\delta)).
\end{aligned}$$

Then it follows that

$$\sum_{k=1}^{K}\sum_{(s,a,h)} q_k(s,a,h)\sqrt{\frac{\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)\}}}=\mathcal{O}\left((H\sqrt{SAK}+HSA)\left(\ln\frac{HSAK}{\delta}\right)^2\right).$$

Since $\max\{1,N_k(s,a,h)-1\}\geq\frac{1}{2}\max\{1,N_k(s,a,h)\}$, we have

$$\sum_{k=1}^{K}\sum_{(s,a,h)} q_k(s,a,h)\frac{(\sqrt{HK}+\sqrt{H^3S^2A})\ln(HSAK/\delta)}{\sqrt{A}\max\{1,N_k(s,a,h)-1\}}=\mathcal{O}\left((H^{1.5}S\sqrt{AK}+H^{2.5}S^2A)\left(\ln\frac{HSAK}{\delta}\right)^2\right),$$

and moreover,

$$\left(HS\ln(HSAK/\delta)+H^2S^2(\ln(HSAK/\delta))^2\right)\sum_{k=1}^{K}\sum_{(s,a,h)}\frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)-1\}}=\mathcal{O}\left(H^3S^3A\left(\ln\frac{HSAK}{\delta}\right)^3\right).$$

Next, by Lemma D.1, $\sum_{k=1}^{K}\sum_{(s,a,h)} q_k(s,a,h)\left(\sqrt{H}\varepsilon_k(s,a,h)\right)$ can be bounded as follows.

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{(s,a,h)} q_k(s,a,h)\left(\sqrt{H}\varepsilon_k(s,a,h)\right)\\
&=\sqrt{H}\sum_{k=1}^{K}\sum_{(s,a,h)} q_k(s,a,h)\left(\sqrt{\frac{4S\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)-1\}}}+\frac{14S\ln(HSAK/\delta)}{3\max\{1,N_k(s,a,h)-1\}}\right)\\
&=\mathcal{O}\left(\left(H^{1.5}S\sqrt{AK}+H^{1.5}S^2A\right)\left(\ln\frac{HSAK}{\delta}\right)^2\right).
\end{aligned}$$

As a result, we have proved that

$$\sum_{k=1}^{K}\langle \boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_k}\rangle=\mathcal{O}\left((H^{1.5}S\sqrt{AK}+H^3S^3A)\left(\ln\frac{HSAK}{\delta}\right)^3\right),$$

as required. ■

We are ready to prove Lemma 4.1 based on Lemma F.1.

Proof. [**Proof of Lemma 4.1**] We closely follow the proof of (Bura et al., 2022, Proposition 4). We assume that the good event $\mathcal{E}$ holds, which holds with probability at least $1-14\delta$. Let $q_b=q^{P,\pi_b}$ be the occupancy measure associated with the safe baseline policy π_b and the true transition kernel P. Then q_b is a feasible solution of (11) if $\langle \widehat{\boldsymbol{g}}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle\le\bar{C}$ holds. To find a sufficient condition, we deduce that

$$\begin{aligned}
\langle \widehat{\boldsymbol{g}}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle&=\langle \bar{\boldsymbol{g}}_{\boldsymbol{k}}+\boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_b}\rangle\\
&\le\langle \boldsymbol{g}+2\boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_b}\rangle\\
&=\bar{C}_b+\langle 2\boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_b}\rangle
\end{aligned}$$

where the first equality is from the definition of $\widehat{g}_k$, the inequality is from Lemma 3.2, and the last equality follows from $\langle \boldsymbol{g},\boldsymbol{q_b}\rangle=\bar{C}_b$. This implies that a sufficient condition for $\langle \widehat{\boldsymbol{g}}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle\le\bar{C}$ is given by

$$\langle 2\boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_b}\rangle<\bar{C}-\bar{C}_b. \tag{26}$$

Note that $\langle 2\boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_b}\rangle$ decreases as k increases because

$$\frac{1}{\max\{1,N_k(s,a,h)\}},\quad \frac{1}{\sqrt{\max\{1,N_k(s,a,h)\}}}$$

can only decrease as k increases. Then suppose that K_0 is the last episode where (26) does not hold. By definition, K_0+1 is the first episode satisfying $\langle \widehat{\boldsymbol{g}}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle<\bar{C}$. Due to the strict inequality, occupancy measures other than q_b can be potentially feasible to (11). This implies that DOPE+ can sufficiently explore policies other than π_b after K_0 episodes. Then we have

$$K_0(\bar{C}-\bar{C}_b)<\sum_{k=1}^{K_0}\langle 2\boldsymbol{R_k}+\boldsymbol{U_k},\boldsymbol{q_b}\rangle.$$

Since q_b induces the true transition kernel, we can apply Lemma F.1. Then the right-hand side is bounded as follows.

$$\sum_{k=1}^{K_0} \langle 2\boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q_b} \rangle = \widetilde{\mathcal{O}}\left(H^{1.5} S \sqrt{AK_0}\right).$$

Hence, K_0 satisfies

$$K_0 = \widetilde{\mathcal{O}}\left(\frac{H^3 S^2 A}{(\bar{C} - \bar{C}_b)^2}\right).$$

Then we have

$$\langle 2\boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q_b} \rangle \leq \langle 2\boldsymbol{R_{K_0+1}} + \boldsymbol{U_{K_0+1}}, \boldsymbol{q_b} \rangle \leq \bar{C} - \bar{C}_b \quad \forall k = K_0 + 1, \ldots, K.$$

This implies that (8) is feasible after episode K_0 when (π_b, P) becomes a feasible solution in episode K_0. ■

Appendix G. Detailed Proofs for the Regret Analysis

In this section, we prove Theorem 2 that guarantees zero constraint violation for DOPE+. Next, we provide the proofs of Lemmas 5.1, 5.2 and 5.3. Lastly, we show Theorem 3 that gives us the regret upper bound.

G.1 Details of Constraint Violation Analysis

Proof. [**Proof of Theorem 2**] We assume that the good event $\mathcal{E}$ holds, which is the case with probability at least $1 - 14\delta$. Let π_k, P_k denote the policy and the transition kernel obtained from DOPE+ for episode k, respectively. Let $q_k = q^{P,\pi_k}, \widehat{q}_k = q^{P_k,\pi_k}$. We know that the constraint is satisfied if $V_1^{\pi_k}(g, P) = \langle \boldsymbol{g}, \boldsymbol{q_k} \rangle \leq \bar{C}$ for each $k \in [K]$. For $k \leq K_0$, there is no constraint violation because we take $\pi_k = \pi_b$. Now we consider the case when $k > K_0$. We have

$$\begin{aligned}
\langle \boldsymbol{g}, \boldsymbol{q_k} \rangle &= \langle \boldsymbol{g}, \boldsymbol{\widehat{q}_k} \rangle + \langle \boldsymbol{g}, \boldsymbol{q_k} - \boldsymbol{\widehat{q}_k} \rangle \\
&\leq \langle \boldsymbol{\bar{g}_k} + \boldsymbol{R_k}, \boldsymbol{\widehat{q}_k} \rangle + \langle \boldsymbol{g}, \boldsymbol{q_k} - \boldsymbol{\widehat{q}_k} \rangle \\
&\leq \langle \boldsymbol{\bar{g}_k} + \boldsymbol{R_k}, \boldsymbol{\widehat{q}_k} \rangle + \langle \boldsymbol{U_k}, \boldsymbol{\widehat{q}_k} \rangle \\
&= \langle \boldsymbol{\widehat{g}_k}, \boldsymbol{\widehat{q}_k} \rangle \\
&\leq \bar{C}
\end{aligned}$$

where the first inequality follows from Lemma 3.2, the second inequality is from Theorem 1, and the last inequality is due to the update rule of DOPE+. This implies that π_k holds $\langle \boldsymbol{g}, \boldsymbol{q_k} \rangle \leq \bar{C}$ for $k > K_0$. Thus, we showed that Violation$(\vec{\pi}) = 0$ with probability at least $1 - 14\delta$. ■

G.2 Details of Regret Analysis

Proof. [**Proof of Lemma 5.1**] We closely follow the proof of (Bura et al., 2022, Lemma 18). We assume that the good event $\mathcal{E}$ holds, which is the case with probability at least $1-14\delta$. We observe that

$$\sum_{k=K_0+1}^{K}\left(V_1^{\pi^*}(f,P)-V_1^{\pi_k}(\widehat{f}_k,P_k)\right)=\sum_{k=K_0+1}^{K}\langle \boldsymbol{f},\boldsymbol{q^*}\rangle-\sum_{k=K_0+1}^{K}\langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}},\widehat{\boldsymbol{q}}_{\boldsymbol{k}}\rangle.$$

By Lemma C.1, there exist $\bar{q}_b(s,a,s',h)$ and $\bar{q}^*(s,a,s',h)$ such that $q_b(s,a,h)=\sum_{s'\in\mathcal{S}}\bar{q}_b(s,a,s',h)$ and $q^*(s,a,h)=\sum_{s'\in\mathcal{S}}\bar{q}^*(s,a,s',h)$, respectively. Then we define the new occupancy measure $q_{\alpha_k}(s,a,h)$ satisfying $q_{\alpha_k}(s,a,h)=\sum_{s'\in\mathcal{S}}\bar{q}_{\alpha_k}(s,a,s',h)$ where

$$\bar{q}_{\alpha_k}(s,a,s',h)=(1-\alpha_k)\bar{q}_b(s,a,s',h)+\alpha_k\bar{q}^*(s,a,s',h) \tag{27}$$

for $(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$ and $\alpha_k\in[0,1]$. Now we verify (C1),(C2) and (C3) in Lemma C.1 to say q_{α_k} is a valid occupancy measure. Since $\bar{q}_{\alpha_k}$ is a convex combination of $\bar{q}_b$ and $\bar{q}^*$, (C1),(C2) hold. For (C3), we can show that q_{α_k} induces the true transition kernel P as follows. Since we know q_b and q^* induce P, it follows that $\bar{q}_b(s,a,s',h)=P(s'\mid s,a,h)\sum_{s''\in\mathcal{S}}\bar{q}_b(s,a,s'',h)$ and $\bar{q}^*(s,a,s',h)=P(s'\mid s,a,h)\sum_{s''\in\mathcal{S}}\bar{q}^*(s,a,s'',h)$ for $(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$. Then $\bar{q}_{\alpha_k}(s,a,s',h)=P(s'\mid s,a,h)\sum_{s''\in\mathcal{S}}\bar{q}_{\alpha_k}(s,a,s'',h)$ can be derived from (27), which implies that q_{α_k} induces the true transition kernel P. Hence, q_{α_k} is a valid occupancy measure inducing the true transition kernel P.

To use the optimality of $\widehat{q}_k$ in our analysis, we expect that q_{α_k} is a feasible solution for (11). Under the good event $\mathcal{E}$, we know that $q_{\alpha_k}\in\Delta(P,k)$ due to $P\in\mathcal{P}_k$. Then it is sufficient to find a condition for α_k satisfying $\langle\widehat{\boldsymbol{g}}_{\boldsymbol{k}},\boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle\le\bar{C}$. We deduce that

$$\begin{aligned}
\langle\widehat{\boldsymbol{g}}_{\boldsymbol{k}},\boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle&=\langle\bar{\boldsymbol{g}}_{\boldsymbol{k}}+\boldsymbol{R}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle\\
&\le\langle\boldsymbol{g}+2\boldsymbol{R}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle\\
&=(1-\alpha_k)\langle\boldsymbol{g}+2\boldsymbol{R}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle+\alpha_k\langle\boldsymbol{g}+2\boldsymbol{R}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q^*}\rangle\\
&\le(1-\alpha_k)(\bar{C}_b+\langle\boldsymbol{2R_k}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle)+\alpha_k(\bar{C}+\langle\boldsymbol{2R_k}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q^*}\rangle)
\end{aligned}$$

where the first inequality is from Lemma 3.2 and the last inequality is from $\langle\boldsymbol{g},\boldsymbol{q_b}\rangle=\bar{C}_b$ and $\langle\boldsymbol{g},\boldsymbol{q^*}\rangle\le\bar{C}$. Furthermore, the second equality is true because (27) implies that $q_{\alpha_k}(s,a,h)=(1-\alpha_k)q_b(s,a,h)+\alpha_k q^*(s,a,h)$. Hence, a sufficient condition of α_k for $\langle\widehat{\boldsymbol{g}}_{\boldsymbol{k}},\boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle\le\bar{C}$ is given by

$$\alpha_k\le\frac{\bar{C}-\bar{C}_b-\langle 2\boldsymbol{R}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle}{\bar{C}-\bar{C}_b+\langle 2\boldsymbol{R}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q^*}\rangle-\langle 2\boldsymbol{R}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle}.$$

Remember that, in the proof of Lemma 4.1, we defined K_0 so that K_0+1 is the first episode satisfying $\langle 2\boldsymbol{R}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q_b}\rangle\le\bar{C}-\bar{C}_b$. This guarantees that there exists some $\alpha_k\in[0,1]$ satisfying the above inequality for $k>K_0$.

Now, for some α_k, we claim that

$$\langle\boldsymbol{f},\boldsymbol{q^*}\rangle\le\langle\bar{\boldsymbol{f}}_{\boldsymbol{k}}+\frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R}_{\boldsymbol{k}}+\frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U}_{\boldsymbol{k}},\boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle. \tag{28}$$

To show (28), we first take for $\beta \geq 1$,

$$\boldsymbol{f_\beta} = \boldsymbol{\bar{f}_k} + 3\beta \boldsymbol{R_k} + \beta \boldsymbol{U_k}.$$

Then we find α_k, β satisfying $\langle \boldsymbol{f}, \boldsymbol{q^*} \rangle \leq \langle \boldsymbol{f_\beta}, \boldsymbol{q_{\alpha_k}} \rangle$. By Lemma 3.2, we have

$$\begin{aligned}
\langle \boldsymbol{f_\beta}, \boldsymbol{q_{\alpha_k}} \rangle &= \langle \boldsymbol{\bar{f}_k} + 3\beta \boldsymbol{R_k} + \beta \boldsymbol{U_k}, \boldsymbol{q_{\alpha_k}} \rangle \\
&\geq \langle \boldsymbol{f} + 2\beta \boldsymbol{R_k} + \beta \boldsymbol{U_k}, \boldsymbol{q_{\alpha_k}} \rangle \\
&= (1-\alpha_k)\langle \boldsymbol{f} + 2\beta \boldsymbol{R_k} + \beta \boldsymbol{U_k}, \boldsymbol{q_b} \rangle + \alpha_k \langle \boldsymbol{f} + 2\beta \boldsymbol{R_k} + \beta \boldsymbol{U_k}, \boldsymbol{q^*} \rangle.
\end{aligned}$$

We have $\langle \boldsymbol{f}, \boldsymbol{q^*} \rangle \leq \langle \boldsymbol{f_\beta}, \boldsymbol{q_{\alpha_k}} \rangle$ if β satisfies

$$\beta \geq \frac{(1-\alpha_k)(\langle \boldsymbol{f}, \boldsymbol{q^*} \rangle - \langle \boldsymbol{f}, \boldsymbol{q_b} \rangle)}{(1-\alpha_k)\langle 2\boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q_b} \rangle + \alpha_k \langle 2\boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q^*} \rangle}.$$

By taking

$$\alpha_k = \frac{\bar{C} - \bar{C}_b - \langle 2\boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q_b} \rangle}{\bar{C} - \bar{C}_b + \langle 2\boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q^*} \rangle - \langle 2\boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q_b} \rangle}, \tag{29}$$

it follows that

$$\beta \geq \frac{\langle \boldsymbol{f}, \boldsymbol{q^*} \rangle - \langle \boldsymbol{f}, \boldsymbol{q_b} \rangle}{\bar{C} - \bar{C}_b}.$$

Since $\langle \boldsymbol{f}, \boldsymbol{q^*} \rangle - \langle \boldsymbol{f}, \boldsymbol{q_b} \rangle \leq H$, it is sufficient to take

$$\beta = \frac{H}{\bar{C} - \bar{C}_b}. \tag{30}$$

For α_k satisfying (29), we showed that q_{α_k} is a feasible solution for (11). Then it follows $\langle \boldsymbol{\widehat{f}_k}, \boldsymbol{q_{\alpha_k}} \rangle \leq \langle \boldsymbol{\widehat{f}_k}, \boldsymbol{\widehat{q}_k} \rangle$ due to optimality of $\widehat{q}_k$. Furthermore, for β satisfying (30), we have (28). Hence, we deduce that

$$\begin{aligned}
\langle \boldsymbol{f}, \boldsymbol{q^*} \rangle - \langle \boldsymbol{\widehat{f}_k}, \boldsymbol{\widehat{q}_k} \rangle &\leq \langle \boldsymbol{f}, \boldsymbol{q^*} \rangle - \langle \boldsymbol{\widehat{f}_k}, \boldsymbol{q_{\alpha_k}} \rangle \\
&= \langle \boldsymbol{f}, \boldsymbol{q^*} \rangle - \langle \boldsymbol{\bar{f}_k} + \frac{3H}{\bar{C} - \bar{C}_b}\boldsymbol{R_k} + \frac{H}{\bar{C} - \bar{C}_b}\boldsymbol{U_k}, \boldsymbol{q_{\alpha_k}} \rangle \\
&\quad + \langle \boldsymbol{\bar{f}_k} + \frac{3H}{\bar{C} - \bar{C}_b}\boldsymbol{R_k} + \frac{H}{\bar{C} - \bar{C}_b}\boldsymbol{U_k}, \boldsymbol{q_{\alpha_k}} \rangle - \langle \vec{\boldsymbol{B}} \wedge (\boldsymbol{\bar{f}_k} + \frac{3H}{\bar{C} - \bar{C}_b}\boldsymbol{R_k} + \frac{H}{\bar{C} - \bar{C}_b}\boldsymbol{U_k}), \boldsymbol{q_{\alpha_k}} \rangle \\
&\leq \langle \boldsymbol{\bar{f}_k} + \frac{3H}{\bar{C} - \bar{C}_b}\boldsymbol{R_k} + \frac{H}{\bar{C} - \bar{C}_b}\boldsymbol{U_k}, \boldsymbol{q_{\alpha_k}} \rangle - \langle \vec{\boldsymbol{B}} \wedge (\boldsymbol{\bar{f}_k} + \frac{3H}{\bar{C} - \bar{C}_b}\boldsymbol{R_k} + \frac{H}{\bar{C} - \bar{C}_b}\boldsymbol{U_k}), \boldsymbol{q_{\alpha_k}} \rangle
\end{aligned}$$

where the last inequality is from (28). Furthermore, under the good event $\mathcal{E}$, we know that $f_k(s,a,h) \leq B$ for $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k \in [K]$, where $B = 1 + \sqrt{\ln(HSAK/\delta)}$. This implies that $\bar{f}_k(s,a,h) \leq B$. Thus, we have

$$\langle \boldsymbol{\bar{f}_k}, \boldsymbol{q_{\alpha_k}} \rangle \leq \langle \vec{\boldsymbol{B}} \wedge (\boldsymbol{\bar{f}_k} + \frac{3H}{\bar{C} - \bar{C}_b}\boldsymbol{R_k} + \frac{H}{\bar{C} - \bar{C}_b}\boldsymbol{U_k}), \boldsymbol{q_{\alpha_k}} \rangle.$$

Then it follows that

$$
\begin{aligned}
&\langle \bar{\boldsymbol{f}}_{\boldsymbol{k}} + \frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R}_{\boldsymbol{k}} + \frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U}_{\boldsymbol{k}}, \boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle - \langle \vec{\boldsymbol{B}} \wedge (\bar{\boldsymbol{f}}_{\boldsymbol{k}} + \frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R}_{\boldsymbol{k}} + \frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U}_{\boldsymbol{k}}), \boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle \\
&\leq \langle \bar{\boldsymbol{f}}_{\boldsymbol{k}} + \frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R}_{\boldsymbol{k}} + \frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U}_{\boldsymbol{k}}, \boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle - \langle \bar{\boldsymbol{f}}_{\boldsymbol{k}}, \boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle \\
&= \langle \frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R}_{\boldsymbol{k}} + \frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U}_{\boldsymbol{k}}, \boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle.
\end{aligned}
$$

Finally, we proved that

$$
\langle \boldsymbol{f}, \boldsymbol{q}^{\boldsymbol{*}}\rangle - \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}}, \widehat{\boldsymbol{q}}_{\boldsymbol{k}}\rangle \leq \langle \frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R}_{\boldsymbol{k}} + \frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U}_{\boldsymbol{k}}, \boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle.
$$

By Lemma F.1, we have

$$
\begin{aligned}
\sum_{k=K_0+1}^{K} \langle \boldsymbol{f}, \boldsymbol{q}^{\boldsymbol{*}}\rangle - \sum_{k=K_0+1}^{K} \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}}, \widehat{\boldsymbol{q}}_{\boldsymbol{k}}\rangle &\leq \sum_{k=K_0+1}^{K} \langle \frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R}_{\boldsymbol{k}} + \frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U}_{\boldsymbol{k}}, \boldsymbol{q}_{\boldsymbol{\alpha_k}}\rangle \\
&= \mathcal{O}\left(\left(\frac{H^{2.5}}{\bar{C}-\bar{C}_b}S\sqrt{AK} + \frac{H^4}{\bar{C}-\bar{C}_b}S^3A\right)\left(\ln\frac{HSAK}{\delta}\right)^3\right)
\end{aligned}
$$

as desired. ∎

Proof. [**Proof of Lemma 5.2**] The lemma is a direct consequence of Lemma H.7 with $B = \mathcal{O}(\ln(HSAK/\delta))$. Hence, we have

$$
\sum_{k=K_0+1}^{K} \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}}, \widehat{\boldsymbol{q}}_{\boldsymbol{k}} - \boldsymbol{q}_{\boldsymbol{k}}\rangle = \mathcal{O}\left(\left(H^{1.5}S\sqrt{AK} + H^3S^3A\right)\left(\ln\frac{HSAK}{\delta}\right)^4\right)
$$

with probability at least $1-2\delta$ under the good event $\mathcal{E}$. By taking the union bound, the statement holds with probability at least $1-16\delta$. ∎

Proof. [**Proof of Lemma 5.3**] We assume that the good event $\mathcal{E}$ holds, which is the case with probability at least $1-14\delta$. The left-hand side of Lemma 5.3 can be rewritten as

$$
\sum_{k=K_0+1}^{K} \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}} - \boldsymbol{f}, \boldsymbol{q}_{\boldsymbol{k}}\rangle.
$$

Under the good event $\mathcal{E}$, we have $\bar{f}_k(s,a,h) \le f(s,a,h) + R_k(s,a,h)$ for $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ and $k \in [K]$. Furthermore, $H/(\bar{C}-\bar{C}_b) \ge 1$ due to $\bar{C}-\bar{C}_b \le H$. Then it follows that

$$\begin{aligned}
\sum_{k=K_0+1}^{K} \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}} - \boldsymbol{f}, \boldsymbol{q_k} \rangle &= \sum_{k=K_0+1}^{K} \langle \vec{\boldsymbol{B}} \wedge (\bar{\boldsymbol{f}}_{\boldsymbol{k}} + \frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R_k} + \frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U_k}) - \boldsymbol{f}, \boldsymbol{q_k} \rangle \\
&\le \sum_{k=K_0+1}^{K} \langle \bar{\boldsymbol{f}}_{\boldsymbol{k}} + \frac{3H}{\bar{C}-\bar{C}_b}\boldsymbol{R_k} + \frac{H}{\bar{C}-\bar{C}_b}\boldsymbol{U_k} - \boldsymbol{f}, \boldsymbol{q_k} \rangle \\
&\le \frac{H}{\bar{C}-\bar{C}_b} \sum_{k=K_0+1}^{K} \langle 4\boldsymbol{R_k} + \boldsymbol{U_k}, \boldsymbol{q_k} \rangle \\
&= \mathcal{O}\left(\left(\frac{H^{2.5}}{\bar{C}-\bar{C}_b} S\sqrt{AK} + \frac{H^4}{\bar{C}-\bar{C}_b} S^3 A \right) \left(\ln \frac{HSAK}{\delta} \right)^3 \right)
\end{aligned}$$

where the last equality is due to Lemma F.1. ∎

Proof. [**Proof of Theorem 3**] We assume that the good event $\mathcal{E}$ holds, which is the case with probability at least $1 - 14\delta$. We decompose the regret as follows using occupancy measures.

$$\begin{aligned}
&\text{Regret}\,(\vec{\pi}) \\
&= \underbrace{\sum_{k=1}^{K_0} \langle \boldsymbol{f}, \boldsymbol{q}^* \rangle - \sum_{k=1}^{K_0} \langle \boldsymbol{f}, \boldsymbol{q_k} \rangle}_{\text{(I)}} + \underbrace{\sum_{k=K_0+1}^{K} \langle \boldsymbol{f}, \boldsymbol{q}^* \rangle - \sum_{k=K_0+1}^{K} \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}}, \widehat{\boldsymbol{q}}_{\boldsymbol{k}} \rangle}_{\text{(II)}} + \underbrace{\sum_{k=K_0+1}^{K} \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}}, \widehat{\boldsymbol{q}}_{\boldsymbol{k}} - \boldsymbol{q_k} \rangle}_{\text{(III)}} + \underbrace{\sum_{k=K_0+1}^{K} \langle \widehat{\boldsymbol{f}}_{\boldsymbol{k}} - \boldsymbol{f}, \boldsymbol{q_k} \rangle}_{\text{(IV)}}.
\end{aligned}$$

As explained in Section 5.2, we can upper bound term (I) as

$$\widetilde{\mathcal{O}}\left(\frac{H^4 S^2 A}{(\bar{C}-\bar{C}_b)^2} \right).$$

because $K_0 = \widetilde{\mathcal{O}}\left(\frac{H^3 S^2 A}{(\bar{C}-\bar{C}_b)^2} \right)$ due to Lemma 4.1 and $\langle \boldsymbol{f}, \boldsymbol{q}^* \rangle \le H$.
By Lemma 5.1, we have

$$\text{Term (II)} = \mathcal{O}\left(\left(\frac{H^{2.5}}{\bar{C}-\bar{C}_b} S\sqrt{AK} + \frac{H^4}{\bar{C}-\bar{C}_b} S^3 A \right) \left(\ln \frac{HSAK}{\delta} \right)^3 \right).$$

By Lemma 5.2, with probability at least $1 - 2\delta$, it follows that

$$\text{Term (III)} = \mathcal{O}\left(\left(H^{1.5} S\sqrt{AK} + H^3 S^3 A \right) \left(\ln \frac{HSAK}{\delta} \right)^4 \right).$$

Moreover, it follows from Lemma 5.3 that

$$\text{Term (IV)} = \mathcal{O}\left(\left(\frac{H^{2.5}}{\bar{C}-\bar{C}_b} S\sqrt{AK} + \frac{H^4}{\bar{C}-\bar{C}_b} S^3 A \right) \left(\ln \frac{HSAK}{\delta} \right)^3 \right).$$

Hence, by taking the union bound,

$$\text{Regret}\left(\vec{\pi}\right) = \widetilde{\mathcal{O}}\left(\frac{H}{\bar{C}-\bar{C}_b}\left(H^{1.5}S\sqrt{AK} + \frac{H^4S^3A}{\bar{C}-\bar{C}_b}\right)\right)$$

with probability at least $1-16\delta$. ∎

Appendix H. Technical Lemmas

In this section, we provide technical lemmas that are crucial for our regret and constraint violation analysis. The following lemma is from (Chen and Luo, 2021) with a few modifications, and it is useful to bound the variance of $\langle \boldsymbol{n_k}, \boldsymbol{f_k}\rangle$.

Lemma H.1 (Chen and Luo, 2021, Lemma 2) *Let π_k be any policy for episode k, and let q_k denote the occupancy measure q^{P,π_k}. Let $\ell : \mathcal{S}\times\mathcal{A}\times[H]\to[-B,B]$ be an arbitrary function, and let P be an arbitrary transition kernel. Then*

$$\mathbb{E}\left[\langle \boldsymbol{n_k}, \boldsymbol{\ell}\rangle^2 \mid \ell, \pi_k, P\right] \le 2B\langle \boldsymbol{q_k}, \vec{\boldsymbol{h}}\odot\boldsymbol{\ell}\rangle$$

where $\boldsymbol{q_k}, \boldsymbol{n_k}, \boldsymbol{\ell}$ are the vector representations of q_k, n_k, ℓ.

Proof. For ease of notation, let $\mathbb{E}_k[\cdot]$ denotes $\mathbb{E}[\cdot \mid \ell, \pi_k, P]$, and let s_h and a_h denote s_h^{P,π_k} and a_h^{P,π_k}, respectively for $h\in[H]$. Note that

$$\begin{aligned}
\mathbb{E}_k\left[\langle \boldsymbol{n_k}, \boldsymbol{\ell}\rangle^2\right] &= \mathbb{E}_k\left[\left(\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} n_k(s,a,h)\ell(s,a,h)\right)^2\right]\\
&= \mathbb{E}_k\left[\left(\sum_{h=1}^{H}\ell(s_h,a_h,h)\right)^2\right]\\
&\le 2\mathbb{E}_k\left[\sum_{h=1}^{H}\ell(s_h,a_h,h)\left(\sum_{m=h}^{H}\ell(s_m,a_m,m)\right)\right]\\
&= 2\mathbb{E}_k\left[\sum_{h=1}^{H}\mathbb{E}_k\left[\ell(s_h,a_h,h)\left(\sum_{m=h}^{H}\ell(s_m,a_m,m)\right) \mid s_h, a_h\right]\right]\\
&= 2\mathbb{E}_k\left[\sum_{h=1}^{H}\ell(s_h,a_h,h)\mathbb{E}_k\left[\sum_{m=h}^{H}\ell(s_m,a_m,m) \mid s_h, a_h\right]\right]\\
&= 2\mathbb{E}_k\left[\sum_{h=1}^{H}\ell(s_h,a_h,h)Q_h^{\pi_k}(s_h,a_h;\ell,P)\right]\\
&= 2\mathbb{E}_k\left[\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} n_k(s,a,h)\ell(s,a,h)Q_h^{\pi_k}(s,a;\ell,P)\right]
\end{aligned}$$

where the first inequality holds because $(\sum_{h=1}^{H} x_h)^2 \leq 2\sum_{h=1}^{H} x_h(\sum_{m=h}^{H} x_m)$. Moreover,

$$\begin{aligned}
\mathbb{E}_k\left[\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} n_k(s,a,h)\ell(s,a,h)Q_h^{\pi_k}(s,a;\ell,P)\right] &= \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \ell(s,a,h)Q_h^{\pi_k}(s,a;\ell,P)\mathbb{E}_k\left[n_k(s,a,h)\right] \\
&= \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \ell(s,a,h)Q_h^{\pi_k}(s,a;\ell,P)q_k(s,a,h) \\
&= \langle \boldsymbol{q_k}, \boldsymbol{\ell}\odot \boldsymbol{Q^{P,\pi_k,\ell}}\rangle.
\end{aligned}$$

Therefore, it follows that

$$\mathbb{E}_k\left[\langle \boldsymbol{n_k}, \boldsymbol{\ell}\rangle^2\right] \leq 2\langle \boldsymbol{q_k}, \boldsymbol{\ell}\odot \boldsymbol{Q^{P,\pi_k,\ell}}\rangle.$$

Next, observe that

$$\begin{aligned}
\langle \boldsymbol{q_k}, \boldsymbol{\ell}\odot \boldsymbol{Q^{P,\pi_k,\ell}}\rangle &\leq B\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} Q_h^{\pi_k}(s,a;\ell,P)q_k(s,a,h) \\
&= B\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_k(a\mid s,h)Q_h^{\pi_k}(s,a;\ell,P)\left(\sum_{a'\in\mathcal{A}} q_k(s,a',h)\right) \\
&= B\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} V_h^{\pi_k}(s;\ell,P)\left(\sum_{a'\in\mathcal{A}} q_k(s,a',h)\right) \\
&= B\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left(\sum_{m=h}^{H}\sum_{(s'',a'')\in\mathcal{S}\times\mathcal{A}} q_k(s'',a'',m\mid s,h)\ell(s'',a'',m)\right)\left(\sum_{a'\in\mathcal{A}} q_k(s,a',h)\right) \\
&= B\sum_{h=1}^{H}\sum_{m=h}^{H}\sum_{(s'',a'')\in\mathcal{S}\times\mathcal{A}}\sum_{s\in\mathcal{S}} q_k(s'',a'',m\mid s,h)\left(\sum_{a'\in\mathcal{A}} q_k(s,a',h)\right)\ell(s'',a'',m) \\
&= B\sum_{h=1}^{H}\sum_{m=h}^{H}\sum_{(s'',a'')\in\mathcal{S}\times\mathcal{A}} q_k(s'',a'',m)\ell(s'',a'',m) \\
&= B\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} h\cdot q_k(s,a,h)\ell(s,a,h) \\
&= B\langle \boldsymbol{q_k}, \vec{\boldsymbol{h}}\odot\boldsymbol{\ell}\rangle
\end{aligned}$$

where the first inequality holds because $\ell(s,a,h)\leq B$ for any (s,a,h), the first equality holds because

$$q_k(s,a,h) = \pi_k(a\mid s,h)\sum_{a'\in\mathcal{A}} q_k(s,a',h),$$

the fifth equality follows from

$$\sum_{s\in\mathcal{S}} q_k(s'',a'',m\mid s,h)\left(\sum_{a'\in\mathcal{A}} q_k(s,a',h)\right) = q_k(s'',a'',m).$$

Therefore, we get that $\langle \boldsymbol{q_k}, \boldsymbol{\ell} \odot \boldsymbol{Q}^{\boldsymbol{P},\boldsymbol{\pi_k},\boldsymbol{\ell}} \rangle \le B \langle \boldsymbol{q_k}, \vec{\boldsymbol{h}} \odot \boldsymbol{\ell} \rangle$ as required. ■

The following lemma is from the first statement of (Chen and Luo, 2021, Lemma 7) with a few modifications to adapt the proof to our setting.

Lemma H.2 (Chen and Luo, 2021, Lemma 7) *Let π be a policy, and let $\widetilde{P}, \widehat{P}$ be two different transition kernels. We denote by $\widetilde{q}$ the occupancy measure $q^{\widetilde{P},\pi}$ associated with $\widetilde{P}$ and π, and we denote by $\widehat{q}$ the occupancy measure $q^{\widehat{P},\pi}$ associated with $\widehat{P}$ and π. Then*

$$\widehat{q}(s,a,h) - \widetilde{q}(s,a,h) = \sum_{(s',a',s'') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \sum_{m=1}^{h-1} \widetilde{q}(s',a',m) \left(\widehat{P}(s'' \mid s',a',m) - \widetilde{P}(s'' \mid s',a',m) \right) \widehat{q}(s,a,h \mid s'',m+1).$$

Proof. We prove the first statement by induction on h. When $h=1$, note that

$$\widehat{q}(s,a,h) = \widetilde{q}(s,a,h) = \pi(a \mid s,1) \cdot p(s).$$

Hence, both the left-hand side and right-hand side are equal to 0. Next, assume that the equality holds with $h-1 \ge 1$. Then we consider h. By the definition of occupancy measures,

$$\begin{aligned}
\widehat{q}(s,a,h) - \widetilde{q}(s,a,h) &= \pi(a \mid s,h) \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} (\widehat{P}(s \mid s',a',h-1)\widehat{q}(s',a',h-1) - \widetilde{P}(s \mid s',a',h-1)\widetilde{q}(s',a',h-1)) \\
&= \underbrace{\pi(a \mid s,h) \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \widehat{P}(s \mid s',a',h-1)(\widehat{q}(s',a',h-1) - \widetilde{q}(s',a',h-1))}_{\text{Term 1}} \\
&\quad \underbrace{+ \pi(a \mid s,h) \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \widetilde{q}(s',a',h-1)(\widehat{P}(s \mid s',a',h-1) - \widetilde{P}(s \mid s',a',h-1))}_{\text{Term 2}}.
\end{aligned}$$

To provide an upper bound on Term 1, we use the induction hypothesis for $h-1$:

$$\begin{aligned}
&\widehat{q}(s',a',h-1) - \widetilde{q}(s',a',h-1) \\
&= \sum_{(s'',a'',s''') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \sum_{m=1}^{h-2} \widetilde{q}(s'',a'',m) \left((\widehat{P} - \widetilde{P})(s''' \mid s'',a'',m) \right) \widehat{q}(s',a',h-1 \mid s''',m+1)
\end{aligned}$$

where

$$(\widehat{P} - \widetilde{P})(s''' \mid s'',a'',m) = \widehat{P}(s''' \mid s'',a'',m) - \widetilde{P}(s''' \mid s'',a'',m).$$

In addition, observe that

$$\pi(a \mid s,h) \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \widehat{P}(s \mid s',a',h-1)\widehat{q}(s',a',h-1 \mid s''',m+1) = \widehat{q}(s,a,h \mid s''',m+1).$$

Therefore, it follows that Term 1 is equal to

$$\begin{aligned}
&\sum_{(s'',a'',s''') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \sum_{m=1}^{h-2} \widetilde{q}(s'',a'',m) \left((\widehat{P} - \widetilde{P})(s''' \mid s'',a'',m) \right) \widehat{q}(s,a,h \mid s''',m+1) \\
&= \sum_{(s',a',s'') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \sum_{m=1}^{h-2} \widetilde{q}(s',a',m) \left(\widehat{P}(s'' \mid s',a',m) - \widetilde{P}(s'' \mid s',a',m) \right) \widehat{q}(s,a,h \mid s'',m+1).
\end{aligned}$$

Next, we upper bound Term 2. Note that

$$\widehat{q}(s, a, h \mid s'', h) = \pi(a \mid s'', h) \cdot \mathbf{1}\left[s'' = s\right].$$

Then it follows that

$$\begin{aligned}
&\pi(a \mid s, h)(\widehat{P}(s \mid s', a', h-1) - \widetilde{P}(s \mid s', a', h-1)) \\
&= \sum_{s'' \in \mathcal{S}} \mathbf{1}\left[s'' = s\right] \cdot \pi(a \mid s'', h)(\widehat{P}(s'' \mid s', a', h-1) - \widetilde{P}(s'' \mid s', a', h-1)) \\
&= \sum_{s'' \in \mathcal{S}} \widehat{q}(s, a, h \mid s'', h)(\widehat{P}(s'' \mid s', a', h-1) - \widetilde{P}(s'' \mid s', a', h-1)),
\end{aligned}$$

implying in turn that Term 2 equals

$$\sum_{(s', a', s'') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \widetilde{q}(s', a', h-1)(\widehat{P}(s'' \mid s', a', h-1) - \widetilde{P}(s'' \mid s', a', h-1))\widehat{q}(s, a, h \mid s'', h).$$

Adding the equivalent expression of Term 1 and that of Term 2 that we have obtained, we get the right-hand side of the statement. ■

The following lemma is called value difference lemma (Dann et al., 2017). Based on Lemma D.3 and Lemma H.2, we show the following lemma, which is a modification of (Chen and Luo, 2021, Lemma 7, the second statement).

Lemma H.3 *Let π be a policy, and let $\widetilde{P}, \widehat{P}$ be two different transition kernels. We denote by $\widetilde{q}$ the occupancy measure $q^{\widetilde{P},\pi}$ associated with $\widehat{P}$ and π, and we denote by $\widehat{q}$ the occupancy measure $q^{\widehat{P},\pi}$ associated with $\widehat{P}$ and π. Let $\ell : \mathcal{S} \times \mathcal{A} \times [H] \to [-B, B]$ be an arbitrary function. If $\widetilde{P}, \widehat{P} \in \mathcal{P}_k$, then we have*

$$\begin{aligned}
|\langle \boldsymbol{\ell}, \widehat{\boldsymbol{q}} - \widetilde{\boldsymbol{q}} \rangle| &= \left| \sum_{(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]} \widetilde{q}(s, a, h) \left(\widehat{P}(s' \mid s, a, h) - \widetilde{P}(s' \mid s, a, h) \right) V^{\pi}_{h+1}(s'; \ell, \widehat{P}) \right| \\
&\leq BH \sum_{(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]} \widetilde{q}(s, a, h) \epsilon^{\star}_k(s' \mid s, a, h)
\end{aligned}$$

where $\widehat{\boldsymbol{q}}, \widetilde{\boldsymbol{q}}, \boldsymbol{\ell}$ are the vector representations of $\widehat{q}, \widetilde{q}, \ell$.

Proof. First, observe that

$$\langle \boldsymbol{\ell}, \widehat{\boldsymbol{q}} - \widetilde{\boldsymbol{q}} \rangle = \sum_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]} \left(\widehat{q}(s, a, h) - \widetilde{q}(s, a, h) \right) \ell(s, a, h).$$

By Lemma H.2, the right-hand side can be rewritten so that we obtain the following.

$$
\begin{aligned}
\langle \boldsymbol{\ell}, \widehat{\boldsymbol{q}} - \widetilde{\boldsymbol{q}} \rangle &= \sum_{(s,a,h)} \sum_{(s',a',s'')} \sum_{m=1}^{h-1} \widetilde{q}(s',a',m) \left((\widehat{P} - \widetilde{P})(s'' \mid s',a',m) \right) \widehat{q}(s,a,h \mid s'', m+1) \ell(s,a,h) \\
&= \sum_{m=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',m) \left((\widehat{P} - \widetilde{P})(s'' \mid s',a',m) \right) \sum_{(s,a,h):h>m} \widehat{q}(s,a,h \mid s'', m+1) \ell(s,a,h) \\
&= \sum_{m=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',m) \left((\widehat{P} - \widetilde{P})(s'' \mid s',a',m) \right) V^{\pi}_{m+1}(s''; \ell, \widehat{P}) \\
&= \sum_{h=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',h) \left(\widehat{P}(s'' \mid s',a',h) - \widetilde{P}(s'' \mid s',a',h) \right) V^{\pi}_{h+1}(s''; \ell, \widehat{P}).
\end{aligned}
$$

Since $\widetilde{P}, \widehat{P} \in \mathcal{P}_k$, Lemma D.3 implies that

$$
\begin{aligned}
|\langle \boldsymbol{\ell}, \widehat{\boldsymbol{q}} - \widetilde{\boldsymbol{q}} \rangle| &\leq \sum_{h=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',h) \left| \widehat{P}(s'' \mid s',a',h) - \widetilde{P}(s'' \mid s',a',h) \right| V^{\pi}_{h+1}(s''; \ell, \widehat{P}) \\
&\leq \sum_{h=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',h) \left(2\epsilon_k(s'' \mid s',a',h) \right) V^{\pi}_{h+1}(s''; \ell, \widehat{P}) \\
&\leq \sum_{h=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',h) \epsilon^{\star}_k(s'' \mid s',a',h) V^{\pi}_{h+1}(s''; \ell, \widehat{P}) \\
&\leq BH \sum_{h=1}^{H} \sum_{(s',a',s'')} \widetilde{q}(s',a',h) \epsilon^{\star}_k(s'' \mid s',a',h) \\
&= BH \sum_{(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]} \widetilde{q}(s,a,h) \epsilon^{\star}_k(s' \mid s,a,h)
\end{aligned}
$$

where the third inequality holds because $V^{\pi}_{h+1}(s''; \ell, \widehat{P}) \leq BH$, as required. ■

Lemma H.4 *Let π be a policy, and let $\widetilde{P}, \widehat{P}$ be two different transition kernels. We denote by $\widetilde{q}$ the occupancy measure $q^{\widetilde{P},\pi}$ associated with $\widetilde{P}$ and π, and we denote by $\widehat{q}$ the occupancy measure $q^{\widehat{P},\pi}$ associated with $\widehat{P}$ and π. Let $(s,h) \in \mathcal{S} \times [H]$, and consider $\widetilde{q}(\cdot \mid s,h), \widehat{q}(\cdot \mid s,h) : \mathcal{S} \times \mathcal{A} \times \{h, \ldots, H\}$. If $\widetilde{P}, \widehat{P} \in \mathcal{P}_k$, then we have*

$$
\left| \langle \boldsymbol{\ell}_{(h)}, \widehat{\boldsymbol{q}}_{(s,h)} - \widetilde{\boldsymbol{q}}_{(s,h)} \rangle \right| \leq BH \sum_{(s',a',s'',m) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \{h,\ldots,H\}} \widetilde{q}(s',a',m \mid s,h) \epsilon^{\star}_k(s'' \mid s',a',m)
$$

where $\widetilde{\boldsymbol{q}}_{(s,h)}, \widehat{\boldsymbol{q}}_{(s,h)}, \boldsymbol{\ell}_{(h)}$ are the vector representations of $\widehat{q}(\cdot \mid s,h), \widetilde{q}(\cdot \mid s,h) : \mathcal{S} \times \mathcal{A} \times \{h, \ldots, H\} \to [0,1]$ and $\ell_{(h)} : \mathcal{S} \times \mathcal{A} \times [H] \to [-B, B]$.

Proof. The proof follows the same argument used to prove Lemmas H.2 and H.3. ∎

The following lemma is called a Bellman-type law of total variance lemma (Azar et al., 2017; Chen and Luo, 2021). We follow the proof of (Chen and Luo, 2021, Lemma 4) after some changes to adapt to our setting.

Lemma H.5 (Chen and Luo, 2021, Lemma 4) *Let π_k be the policy for episode k, P be an arbitrary transition kernel, and let q_k denote the occupancy measure q^{P,π_k}. Let $\ell : \mathcal{S} \times \mathcal{A} \times [H] \to [-B, B]$ be an arbitrary reward function, and define $\mathbb{V}_k(s,a,h) = \mathrm{Var}_{s' \sim P(\cdot|s,a,h)} \left[V^{\pi_k}_{h+1}(s'; \ell, P) \right]$. Then*

$$\langle \boldsymbol{q_k}, \mathbb{V}_{\boldsymbol{k}} \rangle \leq \mathrm{Var} \left[\langle \boldsymbol{n_k}, \boldsymbol{\ell} \rangle \mid \ell, \pi_k, P \right]$$

where $\boldsymbol{q_k}, \mathbb{V}_{\boldsymbol{k}}, \boldsymbol{n_k}, \boldsymbol{\ell}$ are the vector representations of $q_k, \mathbb{V}_k, n_k, \ell$.

Proof. For ease of notation, let s_h and a_h denote s_h^{P,π_k} and a_h^{P,π_k}, respectively for $h \in [H]$. Moreover, let $V(s,h)$ denote $V_h^{\pi}(s; \ell, P)$ for $(s,h) \in \mathcal{S} \times [H]$. Note that

$$\langle \boldsymbol{n_k}, \boldsymbol{\ell} \rangle = \sum_{(s,a,h) \mathcal{S} \times \mathcal{A} \times [H]} \ell(s,a,h) n_k(s,a,h) = \sum_{h=1}^{H} \ell\left(s_h, a_h, h\right).$$

For ease of notation, let $\mathbb{E}_k\left[\cdot\right]$ and $\mathrm{Var}_k\left[\cdot\right]$ denote $\mathbb{E}\left[\cdot \mid \ell, \pi_k, P\right]$ and $\mathrm{Var}\left[\cdot \mid \ell, \pi_k, P\right]$, respectively. Then

$$\mathbb{E}_k\left[\langle \boldsymbol{n_k}, \boldsymbol{\ell} \rangle\right] = \mathbb{E}_k\left[\sum_{h=1}^{H} \ell\left(s_h, a_h, h\right)\right] = \mathbb{E}_k\left[\mathbb{E}\left[\sum_{h=1}^{H} \ell\left(s_h, a_h, h\right) \mid \ell, \pi_k, P, s_1\right]\right] = \mathbb{E}_k\left[V(s_1, 1)\right].$$

Moreover,

$$\begin{aligned}
\mathrm{Var}_k\left[\langle \boldsymbol{n_k}, \boldsymbol{\ell} \rangle\right] &= \mathbb{E}_k\left[\left(\sum_{h=1}^{H} \ell\left(s_h, a_h, h\right) - \mathbb{E}_k\left[V(s_1,1)\right]\right)^2\right] \\
&= \mathbb{E}_k\left[\left(\sum_{h=1}^{H} \ell\left(s_h, a_h, h\right) - V(s_1,1) + V(s_1,1) - \mathbb{E}_k\left[V(s_1,1)\right]\right)^2\right] \\
&= \mathbb{E}_k\left[\left(\sum_{h=1}^{H} \ell\left(s_h, a_h, h\right) - V(s_1,1)\right)^2\right] + \mathbb{E}_k\left[\left(V(s_1,1) - \mathbb{E}_k\left[V(s_1,1)\right]\right)^2\right] \\
&\quad + 2\mathbb{E}_k\left[\left(\sum_{h=1}^{H} \ell\left(s_h, a_h, h\right) - V(s_1,1)\right)\left(V(s_1,1) - \mathbb{E}_k\left[V(s_1,1)\right]\right)\right] \\
&\geq \mathbb{E}_k\left[\left(\sum_{h=1}^{H} \ell\left(s_h, a_h, h\right) - V(s_1,1)\right)^2\right]
\end{aligned}$$

where the inequality is by $\mathbb{E}_k\left[V(s_1,1)-\mathbb{E}_k\left[V(s_1,1)\right]\mid s_1\right]=0$ and $\left(V(s_1,1)-\mathbb{E}_k\left[V(s_1,1)\right]\right)^2\geq 0$. Therefore,

$$\mathrm{Var}_k\left[\langle \boldsymbol{n_k},\boldsymbol{\ell}\rangle\right]\geq \mathbb{E}_k\left[\left(\sum_{h=2}^{H}\ell\left(s_h,a_h,h\right)-V(s_2,2)+\ell\left(s_1,a_1,1\right)+V(s_2,2)-V(s_1,1)\right)^2\right].$$

Note that

$$\mathbb{E}_k\left[\sum_{h=2}^{H}\ell\left(s_h,a_h,h\right)-V(s_2,2)\mid s_1,a_1,s_2\right]=\mathbb{E}_k\left[\sum_{h=2}^{H}\ell\left(s_h,a_h,h\right)\mid s_2\right]-V(s_2,2)=0. \tag{31}$$

Then

$$\begin{aligned}\mathrm{Var}_k\left[\langle \boldsymbol{n_k},\boldsymbol{\ell}\rangle\right]\geq{}& \mathbb{E}_k\left[\left(\sum_{h=2}^{H}\ell\left(s_h,a_h,h\right)-V(s_2,2)\right)^2\right]+\mathbb{E}_k\left[\left(\ell\left(s_1,a_1,1\right)+V(s_2,2)-V(s_1,1)\right)^2\right]\\&+2\mathbb{E}_k\left[\mathbb{E}_k\left[\left(\sum_{h=2}^{H}\ell\left(s_h,a_h,h\right)-V(s_2,2)\right)\left(\ell\left(s_1,a_1,1\right)+V(s_2,2)-V(s_1,1)\right)\mid s_1,a_1,s_2\right]\right]\\={}&\mathbb{E}_k\left[\left(\sum_{h=2}^{H}\ell\left(s_h,a_h,h\right)-V(s_2,2)\right)^2\right]+\mathbb{E}_k\left[\left(\ell\left(s_1,a_1,1\right)+V(s_2,2)-V(s_1,1)\right)^2\right]\\&+2\mathbb{E}_k\left[\left(\ell\left(s_1,a_1,1\right)+V(s_2,2)-V(s_1,1)\right)\mathbb{E}_k\left[\sum_{h=2}^{H}\ell\left(s_h,a_h,h\right)-V(s_2,2)\mid s_1,a_1,s_2\right]\right]\\={}&\mathbb{E}_k\left[\left(\sum_{h=2}^{H}\ell\left(s_h,a_h,h\right)-V(s_2,2)\right)^2\right]+\mathbb{E}_k\left[\left(\ell\left(s_1,a_1,1\right)+V(s_2,2)-V(s_1,1)\right)^2\right]\end{aligned}$$

where the last equality follows from (31). Here, the second term from the right-most side can be bounded from below as follows.

$$
\begin{aligned}
&\mathbb{E}_k\left[\left(\ell\left(s_1, a_1, 1\right)+V\left(s_2, 2\right)-V\left(s_1, 1\right)\right)^2\right] \\
&=\mathbb{E}_k\left[\left(\ell\left(s_1, a_1, 1\right)+\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)-V\left(s_1, 1\right)+V\left(s_2, 2\right)-\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)\right)^2\right] \\
&=\mathbb{E}_k\left[\left(\ell\left(s_1, a_1, 1\right)+\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)-V\left(s_1, 1\right)\right)^2\right] \\
&\quad+\mathbb{E}_k\left[\left(V\left(s_2, 2\right)-\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)\right)^2\right] \\
&\quad+2 \mathbb{E}_k\left[\left(\ell\left(s_1, a_1, 1\right)+\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)-V\left(s_1, 1\right)\right)\left(V\left(s_2, 2\right)-\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)\right)\right] \\
&=\mathbb{E}_k\left[\left(\ell\left(s_1, a_1, 1\right)+\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)-V\left(s_1, 1\right)\right)^2\right] \\
&\quad+\mathbb{E}_k\left[\left(V\left(s_2, 2\right)-\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)\right)^2\right] \\
&\geq \mathbb{E}_k\left[\mathbb{V}_k(s_1, a_1, 1)\right]
\end{aligned}
$$

where third equality holds because

$$
\begin{aligned}
&\mathbb{E}_k\left[\left(\ell\left(s_1, a_1, 1\right)+\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)-V\left(s_1, 1\right)\right)\left(V\left(s_2, 2\right)-\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)\right) \mid s_1, a_1\right] \\
&=\left(\ell\left(s_1, a_1, 1\right)+\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)-V\left(s_1, 1\right)\right) \mathbb{E}_k\left[V\left(s_2, 2\right)-\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2) \mid s_1, a_1\right] \\
&=\left(\ell\left(s_1, a_1, 1\right)+\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)-V\left(s_1, 1\right)\right) \times 0
\end{aligned}
$$

and the last inequality holds because

$$
\mathbb{E}_k\left[\left(V\left(s_2, 2\right)-\sum_{s' \in \mathcal{S}} P(s' \mid s_1, a_1, 1) V(s', 2)\right)^2\right]=\mathbb{E}_k\left[\mathbb{V}_k(s_1, a_1, 1)\right].
$$

Then it follows that

$$
\operatorname{Var}_k\left[\langle \boldsymbol{n}_{\boldsymbol{k}}, \boldsymbol{\ell}\rangle\right] \geq \mathbb{E}_k\left[\left(\sum_{h=1}^H \ell\left(s_h, a_h, h\right)-V\left(s_1, 1\right)\right)^2\right] \geq \mathbb{E}_k\left[\left(\sum_{h=2}^H \ell\left(s_h, a_h, h\right)-V\left(s_2, 2\right)\right)^2\right]+\mathbb{E}_k\left[\mathbb{V}_k(s_1, a_1, 1)\right.
$$

Repeating the same argument, we deduce that

$$\mathrm{Var}_k\left[\langle \boldsymbol{n_k}, \boldsymbol{\ell}\rangle\right] \geq \sum_{h=1}^{H} \mathbb{E}_k\left[\mathbb{V}_k(s_h, a_h, h)\right] = \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} q_k(s,a,h)\mathbb{V}_k(s,a,h) = \langle \boldsymbol{q_k}, \mathbb{V}_{\boldsymbol{k}}\rangle,$$

as required. ■

The following lemma is useful when we prove Theorem 1. The proof is inspired by (Chen and Luo, 2021, Lemma 10) with a few modifications.

Lemma H.6 *Assume that the good event $\mathcal{E}$ holds. Let π_k be any policy for episode $k \in [K]$, and let q_k denote the occupancy measure $q^{P,\pi_k} : \mathcal{S}\times\mathcal{A}\times[H] \to [0,1]$. Let $\ell : \mathcal{S}\times\mathcal{A}\times[H] \to [-B, B]$ be an arbitrary reward function. Then*

$$\begin{aligned}&\left|\sum_{(s,a,s',h)} q_k(s,a,h)\left(P - P_k\right)(s' \mid s,a,h)\left(V_{h+1}^{\pi_k}(s';\ell,P_k) - V_{h+1}^{\pi_k}(s';\ell,P)\right)\right| \\ &\leq 10^4 BH^2S^2\left(\ln\frac{HSAK}{\delta}\right)^2 \sum_{(s,a,h)} \frac{q_k(s,a,h)}{\max\{1, N_k(s,a,h)\}}\end{aligned}$$

for any $P_k \in \mathcal{P}_k$ where $(P - P_k)(s' \mid s,a,h) = P(s' \mid s,a,h) - P_k(s' \mid s,a,h)$.

Proof. Let $\boldsymbol{q^{P_k,\pi_k}_{(s',h+1)}}, \boldsymbol{q^{P,\pi_k}_{(s',h+1)}}, \boldsymbol{\ell}$ be the vector representations of $q^{P_k,\pi_k}(\cdot \mid s', h+1), q^{P,\pi_k}(\cdot \mid s', h+1) : \mathcal{S}\times\mathcal{A}\times\{h+1,\ldots,H\} \to [0,1]$, and $\ell_{(h+1)} : \mathcal{S}\times\mathcal{A}\times\{h+1,\ldots,H\} \to [-B,B]$ respectively. Note that

$$\begin{aligned}&\left|\sum_{(s,a,s',h)} q_k(s,a,h)\left(\left(P - P_k\right)(s' \mid s,a,h)\right)\left(V_{h+1}^{\pi_k}(s';\ell,P_k) - V_{h+1}^{\pi_k}(s';\ell,P)\right)\right| \\ &\leq \sum_{(s,a,s',h)} q_k(s,a,h)\epsilon_k^\star(s' \mid s,a,h)\left|\left(V_{h+1}^{\pi_k}(s';\ell,P_k) - V_{h+1}^{\pi_k}(s';\ell,P)\right)\right| \\ &= \sum_{(s,a,s',h)} q_k(s,a,h)\epsilon_k^\star(s' \mid s,a,h)\left|\langle \boldsymbol{q^{P_k,\pi_k}_{(s',h+1)}} - \boldsymbol{q^{P,\pi_k}_{(s',h+1)}}, \boldsymbol{\ell_{(h+1)}}\rangle\right| \\ &\leq BH\sum_{(s,a,s',h)} q_k(s,a,h)\epsilon_k^\star(s' \mid s,a,h)\sum_{(s'',a'',s'''),m\geq h+1} q_k(s'',a'',m \mid s',h+1)\epsilon_k^\star(s''' \mid s'',a'',m)\end{aligned}$$

where the first inequality is from Lemma D.3, the first equality holds because $V_{h+1}^{\pi_k}(s';\ell,P_k) = \langle \boldsymbol{q^{P_k,\pi_k}_{(s',h+1)}}, \boldsymbol{\ell_{(h+1)}}\rangle$ and $V_{h+1}^{\pi_k}(s';\ell,P) = \langle \boldsymbol{q^{P,\pi_k}_{(s',h+1)}}, \boldsymbol{\ell_{(h+1)}}\rangle$, the second inequality is due to Lemma H.4. Remember that the definition of $\epsilon_k^\star$ is given by

$$\epsilon_k^\star(s' \mid s,a,h) = 6\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} + 94\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}.$$

Then it follows that

$$
\begin{aligned}
&\left(\ln \frac{HSAK}{\delta}\right)^{-2} \sum_{(s,a,s',h)} q_k(s,a,h)\epsilon_k^\star(s' \mid s,a,h) \sum_{(s'',a'',s'''),m\geq h+1} q_k(s'',a'',m \mid s',h+1)\epsilon_k^\star(s''' \mid s'',a'',m) \\
&\leq 36 \underbrace{\sum_{\substack{(s,a,s',h),\\(s'',a'',s'''),\\m\geq h+1}} \sqrt{\frac{q_k(s,a,h)^2 P(s' \mid s,a,h)}{\max\{1,N_k(s,a,h)\}}} \sqrt{\frac{q_k(s'',a'',m \mid s',h+1)^2 P(s''' \mid s'',a'',m)}{\max\{1,N_k(s'',a'',m)\}}}}_{\text{Term 1}} \\
&\quad + 564 \underbrace{\sum_{\substack{(s,a,s',h),\\(s'',a'',s'''),\\m\geq h+1}} \sqrt{\frac{q_k(s,a,h)^2 P(s' \mid s,a,h)}{\max\{1,N_k(s,a,h)\}}} \frac{q_k(s'',a'',m \mid s',h+1)}{\max\{1,N_k(s'',a'',m)\}}}_{\text{Term 2}} \\
&\quad + 564 \underbrace{\sum_{\substack{(s,a,s',h),\\(s'',a'',s'''),\\m\geq h+1}} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}} \sqrt{\frac{q_k(s'',a'',m \mid s',h+1)^2 P(s''' \mid s'',a'',m)}{\max\{1,N_k(s'',a'',m)\}}}}_{\text{Term 3}} \\
&\quad + 8836 \underbrace{\sum_{\substack{(s,a,s',h),\\(s'',a'',s'''),\\m\geq h+1}} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}} \frac{q_k(s'',a'',m \mid s',h+1)}{\max\{1,N_k(s'',a'',m)\}}}_{\text{Term 4}}.
\end{aligned}
$$

Term 1 can be bounded as follows.

$$
\begin{aligned}
\text{Term 1} \leq &\sqrt{\sum_{\substack{(s,a,s',h),\\(s'',a'',s'''),\\m\geq h+1}} \frac{q_k(s,a,h)P(s''' \mid s'',a'',m)q_k(s'',a'',m \mid s',h+1)}{\max\{1,N_k(s,a,h)\}}} \sqrt{\sum_{\substack{(s,a,s',h),\\(s'',a'',s'''),\\m\geq h+1}} \frac{q_k(s'',a'',m \mid s',h+1)P(s' \mid s,a,h)q_k(s,a,h)}{\max\{1,N_k(s'',a'',m)\}}} \\
\leq &\sqrt{HS \sum_{(s,a,h)} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}} \sqrt{HS \sum_{(s'',a'',m)} \frac{q_k(s'',a'',m)}{\max\{1,N_k(s'',a'',m)\}}} \\
= &HS \sum_{(s,a,h)} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}
\end{aligned}
$$

where the first inequality is from the Cauchy-Schwarz inequality.

We can bound Term 2 as the following argument.

$$\begin{aligned}
\text{Term } 2 \leq & \sqrt{\sum_{\substack{(s,a,s',h),\\(s'',a'',s'''),\\m\geq h+1}} \frac{q_k(s,a,h)q_k(s'',a'',m|s',h+1)}{\max\{1,N_k(s,a,h)\}\max\{1,N_k(s'',a'',m)\}}} \sqrt{\sum_{\substack{(s,a,s',h),\\(s'',a'',s'''),\\m\geq h+1}} \frac{q_k(s'',a'',m|s',h+1)P(s'|s,a,h)q_k(s,a,h)}{\max\{1,N_k(s'',a'',m)\}}} \\
\leq & \sqrt{HS^2 \sum_{(s,a,h)} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}} \sqrt{HS \sum_{(s'',a'',m)} \frac{q_k(s'',a'',m)}{\max\{1,N_k(s'',a'',m)\}}} \\
= & HS^{1.5} \sum_{(s,a,h)} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}.
\end{aligned}$$

Similar to Term 2, we have an upper bound on Term 3 as follows.

$$\text{Term } 3 = HS^{1.5} \sum_{(s,a,h)} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}.$$

Since $1/\max\{1, N_k(s,a,h)\} \leq 1$, we bound Term 4 in the following way.

$$\text{Term } 4 \leq HS^2 \sum_{(s,a,h)} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}.$$

Finally, we deduce that

$$\begin{aligned}
&\left| \sum_{(s,a,s',h)} q_k(s,a,h) \left(P - P_k\right)(s' \mid s,a,h) \left(V^{\pi_k}_{h+1}(s';\ell,P_k) - V^{\pi_k}_{h+1}(s';\ell,P)\right) \right| \\
&\leq 10^4 BH^2S^2 \left(\ln \frac{HSAK}{\delta}\right)^2 \sum_{(s,a,h)} \frac{q_k(s,a,h)}{\max\{1,N_k(s,a,h)\}}
\end{aligned}$$

as desired. ■

Next, we provide Lemma H.7, which is a modification of (Chen and Luo, 2021, Lemma 9) to our finite-horizon MDP setting.

Lemma H.7 *Assume that the good event $\mathcal{E}$ holds. Let π_k be any policy for episode k, let P_k be any transition kernel from $\mathcal{P}_k$ for episode k, and let P be the true transition kernel. Let $q_k, \widehat{q}_k$ denote the occupancy measures $q^{P,\pi_k}, q^{P_k,\pi_k}$, respectively. Let $\ell_k : \mathcal{S} \times \mathcal{A} \times [H] \to [-B, B]$ be an arbitrary reward function for episode k. With probability at least $1 - 2\delta$,*

$$\sum_{k=1}^{K} |\langle \boldsymbol{\ell_k}, \boldsymbol{q_k} - \boldsymbol{\widehat{q}_k} \rangle| = \mathcal{O}\left(B \left(H^{1.5} S\sqrt{AK} + H^3S^3A \right) \left(\ln \frac{HSAK}{\delta} \right)^3 \right).$$

where $\boldsymbol{q_k}, \boldsymbol{\widehat{q}_k}, \boldsymbol{\ell_k}$ are the vector representations of $q_k, \widehat{q}_k, \ell_k$.

Proof. We define ξ_1 as $\xi_1 = \{\ell_1, \pi_1\}$ and for $k \geq 2$, we define ξ_k as

$$\left\{s_1^{P,\pi_{k-1}}, a_1^{P,\pi_{k-1}}, \ldots, s_h^{P,\pi_{k-1}}, a_h^{P,\pi_{k-1}}, \ell_k, \pi_k\right\}$$

where π_{k-1} and π_k denote the policies for episode $k-1$ and episode k, respectively, and

$$\left(s_1^{P,\pi_{k-1}}, a_1^{P,\pi_{k-1}}, \ldots, s_h^{P,\pi_{k-1}}, a_h^{P,\pi_{k-1}}\right)$$

is the trajectory generated under policy π_{k-1} and transition kernel P. Then for $k \in [K]$, let $\mathcal{H}_k$ be defined as the σ-algebra generated by the random variables in $\xi_1 \cup \cdots \cup \xi_k$. Then it follows that $\mathcal{H}_1, \ldots, \mathcal{H}_k$ give rise to a filtration.

Let us define

$$\mu_k(s,a,h) = \mathbb{E}_{s' \sim P(\cdot|s,a,h)}\left[V_{h+1}^{\pi_k}(s'; \ell_k, P)\right].$$

Note that

$$\begin{aligned}
\sum_{k=1}^{K} |\langle \boldsymbol{\ell_k}, \boldsymbol{q_k} - \widehat{\boldsymbol{q}}_{\boldsymbol{k}} \rangle| &= \sum_{k=1}^{K} \left| \sum_{(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_k(s,a,h) \left(P(s' \mid s,a,h) - P_k(s' \mid s,a,h)\right) V_{h+1}^{\pi_k}(s'; \ell_k, P_k) \right| \\
&\leq \sum_{k=1}^{K} \left| \sum_{(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_k(s,a,h) \left(P(s' \mid s,a,h) - P_k(s' \mid s,a,h)\right) V_{h+1}^{\pi_k}(s'; \ell_k, P) \right| \\
&\quad + \mathcal{O}\left(BH^3 S^3 A \left(\ln(HSAK/\delta)\right)^3\right)
\end{aligned}$$

where the equality is due to Lemma H.3 and the inequality is due to Lemmas H.6 and D.1.

Moreover,

$$
\begin{aligned}
&\sum_{k=1}^{K}\left|\sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_k(s,a,h)\left(P(s'\mid s,a,h)-P_k(s'\mid s,a,h)\right)V_{h+1}^{\pi_k}(s';\ell_k,P)\right| \\
&=\sum_{k=1}^{K}\left|\sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_k(s,a,h)\left(\left(P-P_k\right)(s'\mid s,a,h)\right)\left(V_{h+1}^{\pi_k}(s';\ell_k,P)-\mu_k(s,a,h)\right)\right| \\
&\leq\sum_{k=1}^{K}\sum_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]} q_k(s,a,h)\epsilon_k^{\star}(s'\mid s,a,h)\left|V_{h+1}^{\pi_k}(s';\ell_k,P)-\mu_k(s,a,h)\right| \\
&\leq\mathcal{O}\left(\sum_{k=1}^{K}\sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} q_k(s,a,h)\sqrt{\frac{P(s'\mid s,a,h)\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)\}}\left(V_{h+1}^{\pi_k}(s';\ell_k,P)-\mu_k(s,a,h)\right)^2}\right) \\
&\quad+\mathcal{O}\left(BHS\sum_{k=1}^{K}\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}\frac{q_k(s,a,h)\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)\}}\right) \\
&\leq\mathcal{O}\left(\sum_{k=1}^{K}\sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} q_k(s,a,h)\sqrt{\frac{P(s'\mid s,a,h)\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)\}}\left(V_{h+1}^{\pi_k}(s';\ell_k,P)-\mu_k(s,a,h)\right)^2}\right) \\
&\quad+\mathcal{O}\left(BH^2S^2A\left(\ln(HSAK/\delta)\right)^2\right)
\end{aligned}
$$

where the first equality holds because $\sum_{s'\in\mathcal{S}}\left(P-P_k\right)(s'\mid s,a,h)=0$ and $\mu_k(s,a,h)$ is independent of s', the first inequality is due to Lemma D.3, the second inequality is from $\left|V_{h+1}^{\pi_k}(s';\ell_k,P)-\mu_k(s,a,h)\right|\leq 2BH$, and the last inequality is from Lemma D.1. Recall that $q_k(s,a,h)=\mathbb{E}\left[n_k(s,a,h)\mid\pi_k,P\right]$, which implies that

$$
\sum_{k=1}^{K}\sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} q_k(s,a,h)\sqrt{\frac{P(s'\mid s,a,h)\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)\}}\left(V_{h+1}^{\pi_k}(s';\ell_k,P)-\mu_k(s,a,h)\right)^2}=\sum_{k=1}^{K}\mathbb{E}\left[X_k\mid\mathcal{H}_k,P\right]
$$

where

$$
X_k=\sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h)\sqrt{\frac{P(s'\mid s,a,h)\ln(HSAK/\delta)}{\max\{1,N_k(s,a,h)\}}\left(V_{h+1}^{\pi_k}(s';\ell_k,P)-\mu_k(s,a,h)\right)^2}.
$$

Here, we have

$$
0\leq X_k\leq\mathcal{O}\left(BHS\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} n_k(s,a,h)\sqrt{\ln(HSAK/\delta)}\right)=\mathcal{O}(BH^2S\sqrt{\ln(HSAK/\delta)}).
$$

Then it follows from Lemma I.6 that with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}\left[X_k \mid \mathcal{H}_k, P\right] \\
&\leq 2 \sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}} \left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2} \\
&\quad + \mathcal{O}\left(BH^2 S \left(\ln(HSAK/\delta)\right)^{1.5}\right).
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}} \left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2} \\
&\leq \sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_{k+1}(s,a,h)\}} \left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2} \\
&\quad + BH \sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) \left(\sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} - \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_{k+1}(s,a,h)\}}}\right) \\
&\leq \sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_{k+1}(s,a,h)\}} \left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2} \\
&\quad + BH\sqrt{S} \sum_{k=1}^{K} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \left(\sqrt{\frac{\ln(HSAK/\delta)}{\max\{1, N_k(s,a,h)\}}} - \sqrt{\frac{\ln(HSAK/\delta)}{\max\{1, N_{k+1}(s,a,h)\}}}\right) \\
&\leq \sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in\\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_{k+1}(s,a,h)\}} \left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2} \\
&\quad + \mathcal{O}\left(BH^2 S^{1.5} A \sqrt{\ln(HSAK/\delta)}\right).
\end{aligned}$$

where the first inequality holds because $\left|V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right| \leq BH$, the second inequality holds because $n_k(s,a,h) \leq 1$ and the Cauchy-Schwarz inequality implies that

$$\sum_{s'\in\mathcal{S}} \sqrt{P(s' \mid s,a,h)} \leq \sqrt{S \sum_{s'\in\mathcal{S}} P(s' \mid s,a,h)} = \sqrt{S},$$

and the third inequality follows from

$$\sum_{k=1}^{K} \left(\sqrt{\frac{1}{\max\{1, N_k(s,a,h)\}}} - \sqrt{\frac{1}{\max\{1, N_{k+1}(s,a,h)\}}}\right) \leq \sqrt{\frac{1}{\max\{1, N_1(s,a,h)\}}} = 1.$$

Next, the Cauchy-Schwarz inequality implies the following.

$$
\begin{aligned}
&\sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in \\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) \sqrt{\frac{P(s' \mid s,a,h)\ln(HSAK/\delta)}{\max\{1, N_{k+1}(s,a,h)\}} \left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2} \\
&\le \sqrt{\sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in \\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) P(s' \mid s,a,h) \left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2} \\
&\quad \times \sqrt{\sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in \\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) \frac{\ln(HSAK/\delta)}{\max\{1, N_{k+1}(s,a,h)\}}}
\end{aligned}
$$

Here, the second term can be bounded as follows.

$$
\begin{aligned}
\sum_{k=1}^{K} \sum_{(s,a,s',h)} n_k(s,a,h) \frac{\ln(HSAK/\delta)}{\max\{1, N_{k+1}(s,a,h)\}} &= S\ln\left(HSAK/\delta\right) \sum_{k=1}^{K} \sum_{(s,a,h)} \frac{n_k(s,a,h)}{\max\{1, N_{k+1}(s,a,h)\}} \\
&= S\ln\left(HSAK/\delta\right) \sum_{(s,a,h)} \sum_{k=1}^{K} \frac{n_k(s,a,h)}{\max\{1, N_{k+1}(s,a,h)\}} \\
&= \mathcal{O}\left(HS^2A\left(\ln\left(HSAK/\delta\right)\right)^2\right).
\end{aligned}
$$

For $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, we define

$$
\mathbb{V}_k(s,a,h) = \operatorname*{Var}_{s'\sim P(\cdot|s,a,h)} \left[V_{h+1}^{\pi_k}(s';\ell_k,P)\right].
$$

Then

$$
\begin{aligned}
\mathbb{V}_k(s,a,h) &= \mathbb{E}_{s'\sim P(\cdot|s,a,h)}\left[\left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2\right] \\
&= \sum_{s'\in\mathcal{S}} P(s' \mid s,a,h)\left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2
\end{aligned}
$$

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

$$
\begin{aligned}
&\sum_{k=1}^{K} \sum_{\substack{(s,a,s',h)\in \\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}} n_k(s,a,h) P(s' \mid s,a,h) \left(V_{h+1}^{\pi_k}(s';\ell_k,P) - \mu_k(s,a,h)\right)^2 \\
&= \sum_{k=1}^{K} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} n_k(s,a,h)\mathbb{V}_k(s,a,h) \\
&= \sum_{k=1}^{K} \langle \boldsymbol{q_k}, \mathbb{V}_{\boldsymbol{k}}\rangle + \sum_{k=1}^{K} \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} (n_k(s,a,h) - q_k(s,a,h))\mathbb{V}_k(s,a,h) \\
&\le \sum_{k=1}^{K} \operatorname{Var}\left[\langle n_k, \ell_k\rangle \mid \ell_k, \pi_k, P\right] + \mathcal{O}\left(B^2H^3\sqrt{K\ln(1/\delta)}\right)
\end{aligned}
$$

where $\mathbb{V}_{\boldsymbol{k}} \in \mathbb{R}^{SAH}$ is the vector representation of $\mathbb{V}_k$ and the inequality follows from Lemma H.5, $\mathbb{V}_k(s,a,h) \leq B^2H^2$,

$$\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} (n_k(s,a,h) - q_k(s,a,h))\mathbb{V}_k(s,a,h) \leq \sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} (n_k(s,a,h) + q_k(s,a,h))B^2H^2 \leq 2B^2H^3,$$

and Lemma I.4. Therefore, we finally have proved that

$$\sum_{k=1}^{K} |\langle \boldsymbol{\ell_k}, \boldsymbol{q_k} - \widehat{\boldsymbol{q}}_{\boldsymbol{k}}\rangle| = \mathcal{O}\left(\sqrt{HS^2A\left(\ln\frac{HSAK}{\delta}\right)^2\left(\sum_{k=1}^{K}\mathrm{Var}\left[\langle n_k, \ell_k\rangle \mid \ell_k, \pi_k, P\right] + B^2H^3\sqrt{K\ln\frac{1}{\delta}}\right)}\right) + \mathcal{O}\left(BH^3S^3A\left(\ln\frac{HSAK}{\delta}\right)^3\right).$$

Moreover, we know from Lemma H.1 that

$$\mathrm{Var}\left[\langle \boldsymbol{n_k}, \boldsymbol{\ell_k}\rangle \mid \ell_k, \pi_k, P\right] \leq \mathbb{E}\left[\langle \boldsymbol{n_k}, \boldsymbol{\ell_k}\rangle^2 \mid \ell_k, \pi_k, P\right] \leq 2B\langle \boldsymbol{q_k}, \vec{\boldsymbol{h}} \odot \boldsymbol{\ell_k}\rangle,$$

and therefore, it follows that

$$\begin{aligned}
\sum_{k=1}^{K} |\langle \boldsymbol{\ell_k}, \boldsymbol{q_k} - \widehat{\boldsymbol{q}}_{\boldsymbol{k}}\rangle| &= \mathcal{O}\left(\left(\sqrt{HS^2A\left(B\sum_{k=1}^{K}\langle \boldsymbol{q_k}, \vec{\boldsymbol{h}} \odot \boldsymbol{\ell_k}\rangle + B^2H^3\sqrt{K}\right)} + BH^3S^3A\right)\left(\ln\frac{HSAK}{\delta}\right)^3\right) \\
&= \mathcal{O}\left(\left(\sqrt{B^2H^3S^2AK + B^2H^4S^2A\sqrt{K}} + BH^3S^3A\right)\left(\ln\frac{HSAK}{\delta}\right)^3\right) \\
&= \mathcal{O}\left(\left(\sqrt{B^2H^3S^2AK + B^2H^3S^2AK + B^2H^5S^2A} + BH^3S^3A\right)\left(\ln\frac{HSAK}{\delta}\right)^3\right) \\
&= \mathcal{O}\left(B\left(H^{1.5}S\sqrt{AK} + H^3S^3A\right)\left(\ln\frac{HSAK}{\delta}\right)^3\right)
\end{aligned}$$

where the second equality holds because $\langle \boldsymbol{q_k}, \vec{\boldsymbol{h}} \odot \boldsymbol{\ell_k}\rangle = \mathcal{O}(BH^2)$ and the third equality holds because $B^2H^4S^2A\sqrt{K} = \mathcal{O}\left(B^2\left(H^3S^2AK + H^5S^2A\right)\right)$. ■

Appendix I. Concentration Inequalities

Lemma I.1 (Hoeffding's inequality) *For i.i.d. random variables $Z_1, \ldots, Z_n$ following* $1/2$-*sub-Gaussian with zero mean,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n} Z_j \geq \epsilon\right) \leq \exp\left(-n\epsilon^2\right),$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n} Z_j \leq -\epsilon\right) \leq \exp\left(-n\epsilon^2\right).$$

Lemma I.2 (Maurer and Pontil, 2009, Theorem 4) *Let $Z_1, \dots, Z_n \in [0, 1]$ be i.i.d. random variables with mean z, and let $\delta > 0$. Then with probability at least $1 - \delta$,*

$$z - \frac{1}{n}\sum_{j=1}^{n} Z_j \leq \sqrt{\frac{2V_n \ln(2/\delta)}{n}} + \frac{7\ln(2/\delta)}{3(n-1)}$$

where V_n is the sample variance given by

$$V_n = \frac{1}{n(n-1)} \sum_{1 \leq j < k \leq n} (Z_j - Z_k)^2.$$

Next, we need the following Bernstein-type concentration inequality for martingales due to Beygelzimer et al. (2011). We take the version used in (Jin et al., 2020, Lemma 9).

Lemma I.3 (Beygelzimer et al., 2011, Theorem 1) *Let $Y_1, \dots, Y_n$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1, \dots, \mathcal{F}_n$. Assume that $Y_j \leq R$ almost surely for all $j \in [n]$. Then for any $\delta \in (0, 1)$ and $\lambda \in (0, 1/R]$, with probability at least $1 - \delta$, we have*

$$\sum_{j=1}^{n} Y_j \leq \lambda \sum_{j=1}^{n} \mathbb{E}\left[Y_j^2 \mid \mathcal{F}_j\right] + \frac{\ln(1/\delta)}{\lambda}.$$

Lemma I.4 (Azuma's inequality) *Let $Y_1, \dots, Y_n$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1, \dots, \mathcal{F}_n$. Assume that $|Y_j| \leq B$ for $j \in [n]$. Then with probability at least $1 - \delta$, we have*

$$\left|\sum_{j=1}^{n} Y_j\right| \leq B\sqrt{2n\ln(2/\delta)}.$$

Next, we need the following concentration inequalities due to Cohen et al. (2020).

Lemma I.5 (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. Suppose that $0 \leq X_n \leq B$ holds almost surely for all n. Then with probability at least $1 - \delta$, the following holds for all $n \geq 1$ simultaneously:*

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

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

Lemma I.6 (Cohen et al., 2020, Lemma D.4) *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables adapted to the filtration $\{\mathcal{F}_n\}_{n=1}^{\infty}$. Suppose that $0 \leq X_n \leq B$ holds almost surely for all n. Then with probability at least $1 - \delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\sum_{i=1}^{n} \mathbb{E}\left[X_i \mid \mathcal{F}_i\right] \leq 2\sum_{i=1}^{n} X_i + 4B\ln\left(2n/\delta\right).$$

Appendix J. Experimental Setup Detailes

We evaluate DOPE+ via the following numerical experiment. We first explain the details of our CMDP setting, which is a modification of the three-state CMDP instances of Zheng and Ratliff (2020); Simão et al. (2021); Bura et al. (2022). We define the state space $\{s_1, s_2, s_3\}$ and the action space $\{a_1, a_2\}$. In Figure 2, we illustrate the transition probability. For taking a_1 at s_1, the agent remains in s_1 with probability 0.8, and moves to s_2 with probability 0.2. For taking a_2 at s_1, the agent moves to s_2 with probability 0.8, and remains in s_2 with probability 0.2. Furthermore, the same transition rule is applied to s_2 and s_3.

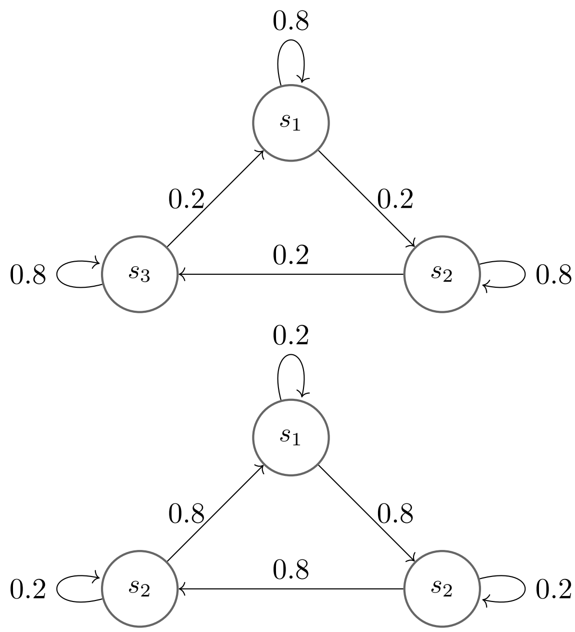

Figure 2: Transition Probability for Taking a_1 and a_2 at Each State: Taking a_1 (Left) and Taking a_2 (Right)

Next, we present the reward function f and the cost function g. When the agent takes a_1, no reward or cost occurs. Then it can be written as $f(s, a_1) = g(s, a_1) = 0$ for $s = s_1, s_2, s_3$. When a_2 is taken, the reward occurs depending on the current state. Specifically, we set $f(s_1, a_2) = 1/3,\ f(s_2, a_2) = 2/3$, and $f(s_3, a_2) = 1$. On the other hand, for any state, the same amount of cost is incurred for a_2, i.e, $g(s_1, a_2) = g(s_2, a_2) = g(s_3, a_2) = 1$. Hence, a_2 is an action with a high reward and a high cost while a_1 is an action with zero reward and zero cost. Furthermore, for taking action a at state s, the agent can observe the noisy reward $f(s, a) + \zeta_1$ and the noisy cost $g(s, a) + \zeta_2$, where ζ_1, ζ_2 are independently drawn from a zero-mean 1/2-sub-Gaussian distribution.

In Figure 1, we compare regret and constraint violation under DOPE+ and DOPE for $200,000$ episodes when $H = 30$. We consider DOPE as a benchmark algorithm because it provides the best-known regret bound among the existing algorithms while ensuring zero hard constraint violation. For the parameters of the experiment, we use $H = 30$, $K = 200,000$, $\bar{C} = 18$, $\bar{C}_b = 15$, $\delta = 0.01$, and the uniform initial distribution of states. To obtain safe baseline policies, we sample a random policy whose expected cost is less than $\bar{C}_b$. Furthermore, we run the safe baseline policies until the LP becomes feasible for both DOPE+ and DOPE. In Figure 1, to observe the learning process easily, we consider the regret and constraint violations incurred after each LP becomes feasible. Our results are averaged across 5 runs with different random seeds, and we display the 95% confidence interval with shaded regions. The experiment was conducted on an Apple M2 Pro.