Provably Efficient Model-Free Algorithms for Non-stationary CMDPs

Honghao Wei
University of Michigan

Arnob Ghosh
The Ohio State University

Ness Shroff
The Ohio State University

Lei Ying
University of Michigan

Xingyu Zhou
Wayne State University

Abstract

We study model-free reinforcement learning (RL) algorithms in episodic non-stationary constrained Markov Decision Processes (CMDPs), in which an agent aims to maximize the expected cumulative reward subject to a cumulative constraint on the expected utility (cost). In the non-stationary environment, reward, utility functions, and transition kernels can vary arbitrarily over time as long as the cumulative variations do not exceed certain variation budgets. We propose the first model-free, simulator-free RL algorithms with sublinear regret and zero constraint violation for non-stationary CMDPs in both tabular and linear function approximation settings with provable performance guarantees. Our results on regret bound and constraint violation for the tabular case match the corresponding best results for stationary CMDPs when the total budget is known. Additionally, we present a general framework for addressing the well-known challenges associated with analyzing non-stationary CMDPs, without requiring prior knowledge of the variation budget. We apply the approach for both tabular and linear approximation settings.

1 INTRODUCTION

Safe reinforcement learning (RL) studies how to apply RL algorithms in real-world applications (Amodei et al., 2016; Garcıa and Fernández, 2015; Brunke et al., 2022) that can operate under safety-related constraints. A standard approach for modeling applications with safety constraints is Constrained Markov Decision Processes (CMDPs) (Altman, 1999), where an agent seeks to learn a policy that maximizes the expected total reward under safety constraints on the expected total utility. In classical safe-RL and CMDPs problems, an agent is assumed to interact with a stationary environment. However, stationary models cannot capture the time-varying real-world applications where safety is critical such that the transition functions and reward/utility functions are non-stationary. For example, in autonomous driving (Kiran et al., 2021), collisions must be avoided while modeling and tracking time-varying environments such as traffic conditions; in an automated medical system (Coronato et al., 2020), it is essential to guarantee patient safety under varying patients' behavior.

Learning in a stationary CMDP is a long-standing topic and has been heavily studied recently, including using both model-based and model-free approaches (Brantley et al., 2020; Efroni et al., 2020; Wei et al., 2022b,a; Liu et al., 2021; Bura et al., 2021; Singh et al., 2020; Ding et al., 2021; Chen et al., 2022). RL in non-stationary CMDPs is more challenging since the rewards/utilities and dynamics are time-varying and probably unknown a priori. On the one hand, an agent has to handle the non-stationarity properly to guarantee a sublinear regret and a small or zero constraint violation. On the other hand, the agent also needs to forget the past data samples since they become less useful due to the dynamic of the system. The only existing work of which we are aware that studies non-stationary CMDPs is Ding and Lavaei (2022), via a model-based approach assuming a priori knowledge of the total variation budgets, which is far less computationally efficient compared with model-free approaches and where knowing the variation budgets is less desirable in practice.

In this work, we manage to overcome these challenges and focus on designing model-free algorithms with sublinear regret and zero constraint violation guarantees for non-stationary CMDPs, especially for the scenario when the total variation budget is unknown. Our contributions are as follows:

- Our work contributes to the theoretical understanding of

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Table 1: Dynamic regret and constraint violations comparisons for RL in non-stationary CMDPs. S and A are the number of states and actions, H is the horizon of each episode, K is the total number of episodes and B is the variation budgets. d is the dimension of the feature in linear CMDP. Algorithms 1,2 are for tabular setting, Ding and Lavaei (2022) is for Linear kernel CMDP setting, and Algorithm 3 is for linear CMDP setting. (* : zero constraint holds holds when K is large enough, $\dagger$: we can further the regret order to $\tilde{\mathcal{O}}(K^{3/4})$, see section 7.)

Algorithm	Model-Free?	Regret	Constraint Violation	Known Budget?
Ding and Lavaei (2022)	✗	$\tilde{\mathcal{O}}(S^{\frac{2}{3}}A^{\frac{1}{3}}H^{\frac{5}{3}}K^{\frac{3}{4}}B^{\frac{1}{3}})$	$\tilde{\mathcal{O}}(S^{\frac{2}{3}}A^{\frac{1}{3}}H^{\frac{5}{3}}K^{\frac{3}{4}}B^{\frac{1}{3}})$	✓
Algorithm 1	✓	$\tilde{\mathcal{O}}\left(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{\frac{4}{5}}B^{\frac{1}{3}}\right)$	0^*	✓
Algorithm 2	✓	$\tilde{\mathcal{O}}\left(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{\frac{8}{9}}B^{\frac{1}{3}}\right)$	0^*	✗
Algorithm 3	✓	$\tilde{\mathcal{O}}\left(K^{3/4}H^{9/4}d^{5/4}B^{1/4}\right)$	0^*	✓
Algorithm 4	✓	$\tilde{\mathcal{O}}\left(K^{7/8}H^{9/4}d^{5/4}B^{1/4}\right)^{\dagger}$	0^*	✗

non-stationary episodic CMDPs. We develop different type of model-free algorithms for non-stationary CMDP settings– one is tailored for tabular CMDPs and has low memory and computational complexity, another one is computationally more intensive, however, can be applied to linear function approximation for large, possibly infinite, state and action spaces.

- For the tabular setting, our algorithm adopts a periodic restart strategy and utilizes an extra optimism bonus term to counteract the non-stationarity of the CMDP that an over estimate of the combined objective is guaranteed during learning and exploration. For the case when the budget variation is known, our theoretical result $\tilde{O}(K^{4/5})$ matches the best existing result for stationary CMDPs in terms of the total number of episodes K, and non-stationary MDPs in term of the variation budget B. For linear CMDP, we propose the first model-free, value-based algorithm which obtains $\tilde{\mathcal{O}}(K^{3/4})$ regret and zero constraint violation using the same strategy. Our result in fact improves the dependency with respect to the budget variation and the episode length H compared to Ding and Lavaei (2022).
- We develop, for the first time, a general *double restart* method for non-stationary CMDPs based on the "bandit over bandit" idea. This method can be used for other non-stationary constrained learning problems which aims to achieve zero constraint violation. The method removes the need to have a priori knowledge of the variation budget, an open-problem raised in Ding and Lavaei (2022) for non-stationary CMDPs. While the "bandit over bandit" has been widely used and studied for unconstrained MDPs, adopting it for CMDPs is nontrivial due to multiple challenges that do not exist in unconstrained setting. For example, one needs to account for the constraints. We overcome these difficulties by a new design of the bandit reward function for each arm. We show that the approach can be used in conjunction with the algorithms for the tabular and linear function approximation cases.

Our results are summarized in Table 1.

2 RELATED WORK

Non-stationary MDP. Non-stationary unconstrained MDPs have been mostly studied recently (Auer et al., 2008; Cheung et al., 2020; Domingues et al., 2021; Fei et al., 2020; Ortner et al., 2020; Touati and Vincent, 2020; Wei and Luo, 2021; Zhong et al., 2021; Zhou et al., 2020; Mao et al., 2020). Auer et al. (2008) consider a setting where the MDP is allowed to change for fixed number of times. When the variation budget is known a priori, Fei et al. (2020) propose a policy-based algorithm in the setting where they assume stationary transitions and adversarial full-information rewards. Zhong et al. (2021); Mao et al. (2020); Touati and Vincent (2020); Zhou et al. (2020) consider a more general setting that both transitions and rewards are time-varying. A more recent work Wei and Luo (2021) introduce a procedure that can be used to convert any upper-confidence-bound-type stationary RL problem to a non-stationary RL algorithm to relax the assumption of having a priori knowledge on the variation budget.

CMDP. Stationary CMDPs with provable guarantees have been heavily studied in recent years. In particular, Brantley et al. (2020); Efroni et al. (2020); Singh et al. (2020) propose model-based approaches for tabular CMDPS. Ghosh et al. (2022); Ding et al. (2021) extend the results to the linear and linear kernel CMDPs. Liu et al. (2021); Bura et al. (2021) also provide efficient algorithms with a zero constraint violation guarantee. Besides using an estimated model, Ding et al. (2020); Chen et al. (2021) leverage a simulator for policy evaluation to achieve provable regret guarantees. Moreover, Wei et al. (2022b,a) propose the first model-free and simulator-free algorithms for CMDPs with sublinear regret and zero constraint violation. However, the studies on non-stationary CMDPs are limited. For non-stationary CMDPs, Qiu et al. (2020) consider CMDPs that assume that only the rewards vary over episodes. A concurrent work (Ding

and Lavaei, 2022), which is most related to ours, focuses on the same setting where the transitions and rewards/utilities vary over episodes under a linear kernel CMDP assumption. They also assume that the budget is known a priori. The method proposed is a model-based approach, but we instead consider a more challenge setting where the algorithm is model-free and the budget is not known. Fortunately, we answer the open-problem affirmatively raised in Ding and Lavaei (2022).

3 PROBLEM FORMULATION

We consider an episodic CMDP where an agent interacts with a non-stationary system for K episodes. The CMDP is denoted by $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g)$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, H is the fixed length of each episode, $\mathbb{P} = \{\mathbb{P}_{k,h}\}_{k\in[K],h\in[H]}$ is a collection of transition kernels, and $r = \{r_{k,h}\}_{k\in[K],h\in[H]} (g = \{g_{k,h}\}_{k\in[K],h\in[H]})$ is the set of reward (utility) functions. In Section 7, we extend our analysis to potentially infinite state-space.

At the beginning of an episode k, an initial state $x_{k,1}$ is sampled from the distribution μ_0. Then at step h, the agent takes action $a_{k,h} \in \mathcal{A}$ after observing state $x_{k,h} \in \mathcal{S}$. Then the agent receives a reward $r_{k,h}(x_{k,h}, a_{k,h})$ and incurs a utility $g_{k,h}(x_{k,h}, a_{k,h})$. The environment transitions to a new state $x_{k,h+1}$ following from the distribution $\mathbb{P}_{k,h}(\cdot|x_{k,h}, a_{k,h})$. It is worth emphasizing that the transition kernels, reward functions, and utility functions all depend on the episode index k and time h, and hence the system is non-stationary. For simplicity of notation, we assume that $r_{k,h}(x,a)(g_{k,h}(x,a)) : \mathcal{S}\times\mathcal{A} \to [0,1]$, are deterministic for convenience. Our results generalize to the setting where the reward and utility functions are random. Given a policy π, which is a collection of H functions $\pi : [H]\times\mathcal{S} \to \mathcal{A}$, where $[H]$ represents the set $\{1, 2, \ldots, H\}$. Define the reward value function $V^{\pi}_{k,h}(x) : \mathcal{S} \to \mathbb{R}^+$ at episode k and step h to be the expected cumulative rewards from step h to the end under the policy π :

$$V^{\pi}_{k,h}(x) = \mathbb{E}\left[\sum_{i=h}^{H} r_{k,i}(x_{k,i}, \pi(x_{k,i}))\middle| x_{k,h} = x\right]. \quad (1)$$

The (reward) Q-function $Q^{\pi}_{k,h}(x,a) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$ is the expected cumulative reward when an agent starts from a state-action pair (x,a) at episode k and step h following the policy π :

$$Q^{\pi}_{k,h}(x,a) = r_{k,h}(x,a) + \mathbb{E}\left[\sum_{i=h+1}^{H} r_{k,i}(x_{k,i}, \pi(x_{k,i}))\middle| x_{k,h} = x, a_{k,h} = a\right]. \quad (2)$$

Similarly, we use $W^{\pi}_{k,h}(x) : \mathcal{S} \to \mathbb{R}^+$ to denote the utility value function

$$W^{\pi}_{k,h}(x) = \mathbb{E}\left[\sum_{i=h}^{H} g_{k,i}(x_{k,i}, \pi(x_{k,i}))\middle| x_{k,h} = x\right], \quad (3)$$

and we use $C^{\pi}_{k,h}(x,a) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$ to denote the utility Q-function at episode k, step h:

$$C^{\pi}_{k,h}(x,a) = g_{k,h}(x,a) + \mathbb{E}\left[\sum_{i=h+1}^{H} g_{k,i}(x_{k,i}, \pi(x_{k,i}))\middle| x_{k,h} = x, a_{k,h} = a\right]. \quad (4)$$

For simplicity, we adopt the following notations:

$$\mathbb{P}_{k,h}V^{\pi}_{k,h+1}(x,a) = \mathbb{E}_{x'\sim\mathbb{P}_{k,h}(\cdot|x,a)}V^{\pi}_{k,h+1}(x'), \quad (5)$$

$$\mathbb{P}_{k,h}W^{\pi}_{h+1}(x,a) = \mathbb{E}_{x'\sim\mathbb{P}_{k,h}(\cdot|x,a)}W^{\pi}_{k,h+1}(x'). \quad (6)$$

We also denote the empirical counterparts as

$$\hat{\mathbb{P}}_{k,h}V^{\pi}_{k,h+1}(x,a) = V^{\pi}_{k,h+1}(x_{k+1,h}), \quad (7)$$

$$\hat{\mathbb{P}}_{k,h}W^{\pi}_{k,h+1}(x,a) = W^{\pi}_{k,h+1}(x_{k+1,h}), \quad (8)$$

and is only defined for $(x,a) = (x_{k,h}, a_{k,h})$. Given the model defined above, the objective of the episode k is to find a policy that maximizes the expected cumulative reward subject to a constraint on the expected utility:

$$\max_{\pi_k\in\Pi} \mathbb{E}\left[V^{\pi_k}_{k,1}(x_1)\right] \quad \text{subject to: } \mathbb{E}\left[W^{\pi_k}_{k,1}(x_1)\right] \geq \rho, \quad (9)$$

where we assume $\rho \in [0, H]$ to avoid triviality, and the expectation is taken with respect to the initial distribution and the randomness of π. Let π^*_k denote the optimal solution to the CMDP problem defined in (9) for episode k. We evaluate our model-free RL algorithms using dynamic regret $\mathcal{R}(K)$ and constraint violation $\mathcal{V}(K)$ defined below:

$$\mathcal{R}(K) = \mathbb{E}\left[\sum_{k=1}^{K}\left(V^{\pi^*_k}_{k,1}(x_{k,1}) - V^{\pi_k}_{k,1}(x_{k,1})\right)\right], \quad (10)$$

$$\mathcal{V}(K) = \mathbb{E}\left[\sum_{k=1}^{K}\left(\rho - W^{\pi_k}_{k,1}(x_{k,1})\right)\right], \quad (11)$$

where π_k is the policy used in episode k. Note that here we use dynamic regret concept as the optimal policy may be different. We further make the following standard assumption (Efroni et al., 2020; Ding et al., 2021; Qiu et al., 2020; Wei et al., 2022b).

Assumption 1. *(Slater's Condition). Given initial distribution μ_0, for any episode $k \in [K]$, there exist $\delta > 0$ and at least a policy π such that $\mathbb{E}\left[W^{\pi}_{k,1}(x_{k,1})\right] - \rho \geq \delta$.*

Variation: The non-stationary of the CMDP is measured according to the variation budgets in the reward/utility functions and the transition kernels:

$$B_r := \sum_{k=1}^{K-1}\sum_{h=1}^{H}\max_{x,a}|r_{k,h}(x,a) - r_{k+1,h}(x,a)|$$

$$B_g := \sum_{k=1}^{K-1}\sum_{h=1}^{H}\max_{x,a}|g_{k,h}(x,a)-g_{k+1,h}(x,a)|$$
$$B_p := \sum_{k=1}^{K-1}\sum_{h=1}^{H}\max_{x,a}\|\mathbb{P}_{k,h}(\cdot|x,a)-\mathbb{P}_{k+1,h}(\cdot|x,a)\|_1.$$

We further let $B = B_r + B_g + B_p$ to represent the total variation. To bound the regret, we consider the following offline optimization problem at episode k as our regret baseline:

$$\max_{q_{k,h}} \sum_{h,x,a} q_{k,h}(x,a)r_{k,h}(x,a) \tag{12}$$
$$\text{s.t.:} \sum_{h,x,a} q_{k,h}(x,a)g_{k,h}(x,a) \geq \rho \tag{13}$$
$$\sum_a q_{k,h}(x,a) = \sum_{x',a'}\mathbb{P}_{k,h-1}(x|x',a')q_{k,h-1}(x',a') \tag{14}$$
$$\sum_{x,a} q_{k,h}(x,a) = 1, \forall h \in [H] \tag{15}$$
$$\sum_a q_{k,1}(x,a) = \mu_0(x) \tag{16}$$
$$q_{k,h}(x,a) \geq 0, \forall x \in \mathcal{S}, \forall a \in \mathcal{A}, \forall h \in [H]. \tag{17}$$

To analyze the performance, we need to consider a tightened version of the LP, which is defined below:

$$\max_{q_{k,h}} \sum_{h,x,a} q_{k,h}(x,a)r_{k,h}(x,a) \tag{18}$$
$$\text{s.t.:} \sum_{h,x,a} q_{k,h}(x,a)g_{k,h}(x,a) \geq \rho + \epsilon, \text{ and (14)} - \text{(17)},$$

where $\epsilon > 0$ is called a tightness constant. When $\epsilon \leq \delta$, this problem has a feasible solution due to Slater's condition. We use superscript * to denote the optimal value/policy related to the original CMDP (9) or the solution to the corresponding LP (12) and superscript $^{\epsilon,*}$ to denote the optimal value/policy related to the ϵ-tightened version of CMDP.

4 ALGORITHM FOR TABULAR CMDPs

Next we will start with presenting our algorithm Non-stationary Triple-Q in Algorithm 1 for the scenario when the variation budget is known. Our algorithm uses a restart strategy that divides the total episode K into frames, which is commonly used in both non-stationary bandits and RL to address non-stationarity. We remark that in unconstrained RL, this restarting often results in a worse regret, for example, the regret bound is $\tilde{\mathcal{O}}(\sqrt{K})$ (Jin et al., 2018) in the stationary setting but becomes $\tilde{\mathcal{O}}(K^{\frac{2}{3}})$ (Mao et al., 2020) when the system is non-stationary. However, the order of regret achieved by our Algorithm 1 matches the best existing result in stationary CMDPs obtained by the model-free algorithm Triple-Q (Wei et al., 2022b) under the same setting. That is because Triple-Q itself is built on top of a two-time-scale scheme for balancing the estimation error and tracking the constraint violation, which shares the same insights as the restart strategy for dealing with non-stationarity. Therefore, by appropriately designing the frame size (restarting period), Algorithm 1 can achieve the same order as that in unconstrained CDMPs as well as the optimal order in terms of variation budget.

We first divide the total K episodes into frames, where each frame contains K^α/B^c episodes. Define $B_r^{(T)}, B_g^{(T)}, B_p^{(T)}$ to be the local variation budget of the reward functions, utility functions and transition kernels within the Tth frame, let $\mathcal{N}_T$ denote the set of all the episodes in frame T, then

$$B_r^{(T)} := \sum_{k\in\mathcal{N}_T}\sum_{h=1}^{H}\max_{x,a}|r_{k,h}(x,a)-r_{k+1,h}(x,a)|$$
$$B_g^{(T)} := \sum_{k\in\mathcal{N}_T}\sum_{h=1}^{H}\max_{x,a}|g_{k,h}(x,a)-g_{k+1,h}(x,a)|$$
$$B_p^{(T)} := \sum_{k\in\mathcal{N}_T}\sum_{h=1}^{H}\max_{x,a}\|\mathbb{P}_{k,h}(\cdot|s,a)-\mathbb{P}_{k+1,h}(\cdot|x,a)\|_1.$$

Let the total local variation budget $B^{(T)} = B_r^{(T)} + B_g^{(T)} + B_p^{(T)}$, then by definition we have $\sum_{T=1}^{K^{1-\alpha}B^c} B^{(T)} \leq B$. Our algorithm uses two bonus terms b_t and $\tilde{b}$ to update Q values (Line $10-11$ in Algorithm 1), where b_t is the standard Hoeffding-based bonus in upper confidence bounds, and $\tilde{b}$ is the extra bonus to take into account the non-stationarity of the environment. We assume that $\tilde{b}$ is a uniform upper bound on the total local variation budget B^T for any T, and satisfies $K^{1-\alpha}B^c\tilde{b} \leq B$ which is an assumption commonly seen in the literature on non-stationary RL (Ortner et al., 2020; Mao et al., 2020; Zhou et al., 2020).

5 RESULTS OF TABULAR CMDPs

We now present our main results of the Non-stationary Triple-Q.

Theorem 1. *Assume* $K \geq \max\left\{\left(\frac{16\sqrt{SAH^6\iota^3}}{\delta}\right)^5, e^{\frac{1}{\delta}}\right\}$, *where* $\iota = 128\log(\sqrt{2SAH}K)$. *Algorithm 1 achieves the following regret and constraint violation bounds:*

$$\mathcal{R}(K) = \tilde{\mathcal{O}}(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}B^{\frac{1}{3}}K^{\frac{4}{5}}) \qquad \mathcal{V}(K) = 0$$

Due to the page limit, we only outline some of the key intuitions behind Theorem 1. The detailed proofs are deferred to Section E in the supplementary materials.

5.1 Dynamic Regret

As shown in Algorithm 1, let $Q_{k,h}(x,a), C_{k,h}(x,a)$ denote the estimate Q values at the beginning of the $k-$th episode.

Algorithm 1: Non-stationary Triple-Q

Input: Total Budget B;
Choose $\alpha = 0.6, \eta = K^{\frac{1}{5}}B^{\frac{1}{3}}, \chi = K^{\frac{1}{5}}, c = \frac{2}{3}, \epsilon = \frac{8\sqrt{SAH^6\iota^3}B^{1/3}}{K^{0.2}}$, and $\iota = 128\log(\sqrt{2SAH}K)$;
Initialize $Q_h(x,a) = C_h(x,a) \leftarrow H$ and $Z = \bar{C} = N_h(x,a) = V_{H+1}(x) = W_{H+1}(x) \leftarrow 0$ for all $(x,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$;
for *episode* $k = 1, \ldots, K$ **do**
  Sample the initial state for episode k : $x_{k,1} \sim \mu_0$;
  for *step* $h = 1, \ldots, H$ **do**
    Take action $a_h \leftarrow \arg\max_a \left(Q_h(x_{k,h},a) + \frac{Z}{\eta}C_h(x_{k,h},a)\right)$;
    Observe $r_{k,h}(x_{k,h},a_{k,h})$, $g_{k,h}(x_{k,h},a_{k,h})$, and $x_{k,h+1}$, $N_h(x_{k,h},a_{k,h}) \leftarrow N_h(x_{k,h},a_{k,h}) + 1$;
    Set $t = N_h(x_{k,h},a_{k,h})$, $b_t = \frac{1}{4}\sqrt{\frac{H^2\iota(\chi+1)}{\chi+t}}$, $\alpha_t = \frac{\chi+1}{\chi+t}$;
    $Q_h(x_{k,h},a_{k,h}) \leftarrow (1-\alpha_t)Q_h(x_{k,h},a_{k,h}) + \alpha_t\left(r_{k,h}(x_{k,h},a_{k,h}) + V_{h+1}(x_{k,h+1}) + b_t + 2H\tilde{b}\right)$;
    $C_h(x_{k,h},a_{k,h}) \leftarrow (1-\alpha_t)C_h(x_{k,h},a_{k,h}) + \alpha_t\left(g_{k,h}(x_{k,h},a_{k,h}) + W_{h+1}(x_{k,h+1}) + b_t + 2H\tilde{b}\right)$;
    $a' = \arg\max_a \left(Q_h(x_{k,h},a) + \frac{Z}{\eta}C_h(x_{k,h},a)\right)$, $V_h(x_{k,h}) \leftarrow Q_h(x_{k,h},a')$ $W_h(x_{k,h}) \leftarrow C_h(x_{k,h},a')$;
    if $h = 1$ **then**
      $\bar{C} \leftarrow \bar{C} + C_1(x_{k,1},a_{k,1})$
  if $k \mod (K^\alpha/B^c) = 0$; // reset visit counts and Q-functions
  then
    $N_h(x,a) \leftarrow 0$, $Q_h(x,a) = C_h(x,a) = Q_h(x,a) = C_h(x,a) \leftarrow H$, $Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}\cdot B^c}{K^\alpha}\right)^+$, $\bar{C} \leftarrow 0$

The dynamic regret can be decoupled as:

$$\mathcal{R}(K) = \mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q^*_{k,1}q^*_{k,1} - Q^{\epsilon,*}_{k,1}q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a)\right)\right] + \tag{19}$$

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q^{\epsilon,*}_{k,1}q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right] + \tag{20}$$

$$\mathbb{E}\left[\sum_{k=1}^{K}\left\{Q_{k,1} - Q^{\pi_k}_{k,1}\right\}(x_{k,1},a_{k,1})\right], \tag{21}$$

here we use the shorthand notation $\{f-g\}(x) = f(x) - g(x)$. Before bounding each term, we first show that for any triple (x,a,h), the difference of two different reward/utility Q-value functions within the same frame are bounded by the local variation bound in that frame.

Lemma 1. *Given any frame T, for any (x,a,h), and $(T-1)K^\alpha/B^c \le k_1 \le k_2 \le TK^\alpha/B^c$, we have*

$$|Q^\pi_{k_1,h}(x,a) - Q^{\pi'}_{k_2,h}(x,a)| \le H\tilde{b} \tag{22}$$

$$|C^\pi_{k_1,h}(x,a) - C^{\pi'}_{k_2,h}(x,a)| \le H\tilde{b} \tag{23}$$

Then we show that in Lemma 9 in supplementary materials the first term (19) can be bounded by comparing the original LP associated with the tightened LP such that (19) $\le \frac{KH\epsilon}{\delta}$. The term (21) is the estimation error between $Q_{k,h}$ and the true Q value under policy π_k at episode k. This estimation error can be bounded by our choice of the learning rate (Line 8 in Algorithm 1) and the added bonus. Then (21) $\le H^2SAK^{1-\alpha}B^c + \frac{2(H^3\sqrt{\iota}+2H^2\tilde{b})K}{\chi} + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)B^c} + 2\tilde{b}H^2K$.

For the remaining term (20), we need to add and subtract additional terms to construct an difference between the optimal combined Q value $\{Q^*_{k,h} + \frac{Z}{\eta}\}C^*_{k,h}(x,a)$ and the estimated counterpart $\{Q_{k,h} + \frac{Z}{\eta}C_{k,h}\}(x,a)$. We will show in Lemma 7 that the estimation is always an overestimation for all (x,a,h,k) due to the added bonus when the virtual "queue" Z_T is fixed with high probability, which implies that the difference is negative with high probability. Then in Lemma 10 we leverage Lyapunov-drift method and consider Lyapunov function $L_T = \frac{1}{2}Z_T^2$ to show that the redundant term can also be bounded. Combining the bounds on the estimation and the redundant term we can obtain (20) $\le \frac{K(2H^4\iota+4H^2\tilde{b}^2+\epsilon^2)}{\eta} + \frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}$. Then combining inequalities (19),(20),(21) above we can obtain for $K \ge \left(\frac{16\sqrt{SAH^6\iota^3}B^{1/3}}{\delta}\right)^5$, applying the condition $K^{1-\alpha}B^c\tilde{b} \le B$, along with our choices of parameters (Line 2 in Algorithm 1) for balancing each terms, we conclude that $\mathcal{R}(K) = \tilde{\mathcal{O}}(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}B^{\frac{1}{3}}K^{\frac{4}{5}})$.

5.2 Constraint Violation

According to the virtual-Queue update, we have

$$\begin{aligned} Z_{T+1} =& \left(Z_T + \rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}\right)^+ \\ \geq& Z_T + \rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}, \end{aligned} \tag{24}$$

which implies that for $(T-1)K^\alpha/B^c \leq k \leq TK^\alpha/B^c$,

$$\begin{aligned} &\sum_k \left(-C^{\pi_k}_{k,1}(x_{k,1}, a_{k,1}) + \rho\right) \leq \frac{K^\alpha}{B^c}\left(Z_{T+1} - Z_T\right) \\ &+ \sum_k \left(\left\{C_{k,1} - C^{\pi_k}_{k,1}\right\}(x_{k,1}, a_{k,1}) - \epsilon\right). \end{aligned}$$

Summing the inequality above over all frames and taking expectation on both sides, we obtain the following upper bound on the constraint violation:

$$\begin{aligned} \mathbb{E}\left[\sum_{k=1}^K \rho - C^{\pi_k}_{k,1}(x_{k,1}, a_{k,1})\right] \leq& -K\epsilon + \frac{K^\alpha}{B^c}\mathbb{E}\left[Z_{K^{1-\alpha}B^c+1}\right] \\ &+ \mathbb{E}\left[\sum_{k=1}^K \left\{C_{k,1} - C^{\pi_k}_{k,1}\right\}(x_{k,1}, a_{k,1})\right], \end{aligned} \tag{25}$$

where the inequality is true due to the fact $Z_1 = 0$. In Lemma 8, we will establish an upper bound on the estimation error of $\mathbb{E}\left[\sum_{k=1}^K \left\{C_{k,1} - C^{\pi_k}_1\right\}(x_{k,1}, a_{k,1})\right]$. Next, we study the moment generating function of Z_T, i.e. $\mathbb{E}\left[e^{rZ_T}\right]$ for some $r > 0$. In Lemma 11, based on a Lyapunov drift analysis of this moment generating function and Jensen's inequality, we will establish the following upper bound on Z_T that holds for any $1 \leq T \leq K^{1-\alpha}B^c$,

$$\begin{aligned} \mathbb{E}[Z_T] \leq& \frac{100(H^4\iota + \tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota} + \tilde{b}H)}{\delta}\right) \\ &+ \frac{4H^2B^c}{K\delta} + \frac{4H^2B^c}{\eta K^\alpha \delta} + \frac{4\eta(\sqrt{H^2\iota} + 2H^2\tilde{b})}{\delta}. \end{aligned} \tag{26}$$

Substituting the results from Lemma 8 and (26) into (25), using the choice that $\epsilon = \frac{8\sqrt{SAH^6\iota^3}B^{1/3}}{K^{0.2}}$, we can easily verify that when $K \geq \max\left\{\left(\frac{16\sqrt{SAH^6\iota^3}B^{1/3}}{\delta}\right)^5, e^{\frac{1}{\delta}}\right\}$, we have

$$\begin{aligned} \mathcal{V}(K) \leq& \frac{100(H^4\iota + \tilde{b}^2H^2)K^{0.6}}{\delta B^{2/3}}\log\frac{16(H^2\sqrt{\iota} + H\tilde{b})}{\delta} \\ &- \sqrt{SAH^6\iota^3}K^{0.8}B^{\frac{1}{3}} \leq 0. \end{aligned} \tag{27}$$

6 UNKNOWN VARIATION BUDGETS

The design of the Algorithm 1 relies on the knowledge of the total variation budget B to set the frame size to be K^α/B^c. When an upper bound on the total variation budget is not given, we propose the Algorithm 2 that adaptively learns the variation budget B based on the "Bandit over Bandit" algorithm (Cheung et al., 2022). Algorithm 2 uses an outer loop "bandit algorithm" as a master to learn the true value B, and use the inner loop Algorithm 1 to learn the optimal policy. We first need to divide total K episodes into $\frac{K}{W}$ epochs, which contain $W = K^\zeta$ episodes. Each epoch contains multiple frames. In each epoch, we run an instance of Algorithm 1. Given a candidate set $\mathcal{J}$ of the total budget B, we choose "arms" (estimated budget) using the bandit adversarial bandit algorithm Exp3 (Auer et al., 2003). If the optimal "arm" from the candidate $\mathcal{J}$ can be learned efficiently, we expect that the cumulative reward and utility collected under that arm should be close to the performance of using the best-fixed candidate (closest to true Budget) from $\mathcal{J}$ in hindsight. We remark that although the "Bandit over Bandit" approach is well studied in both unconstrained non-stationary bandit and RL, however, adopting it in CMDPs is nontrivial and new. We now describe the main challenge in adapting the idea to the constrained scenario and how we overcome the challenge.

In particular, given a choice of arm B_i in the unconstrained version, one considers the cumulative reward $R_i(B_i)$ over the epoch W to guide the EXP-3 algorithm towards selecting the optimal arm. The cumulative reward proves to be enough for the unconstrained case, as the optimal arm would correspond to close to the true budget. This can be reflected as the following regret decomposition,

$$\begin{aligned} \mathcal{R}(K) =& \mathbb{E}\left[\sum_{k=1}^K V^{\pi^*_k}_{k,1}(x_{k,1}) - \sum_{i=1}^{K/W} R_i(\hat{B})\right] \qquad (28) \\ &+ \mathbb{E}\left[\sum_{i=1}^{K/W} R_i(\hat{B}) - \sum_{i=1}^{K/W} R_i(B_i)\right], \qquad (29) \end{aligned}$$

where $\hat{B}$ is the optimal candidate from $\mathcal{J}$ (i.e., the true budget). We can show that the term (28) can be bounded since this corresponds to regret when the true budget is known (which we have already bounded). However, the problem becomes that how to bound the term (29). In the unconstrained case, one can employ the result of the EXP-3 algorithm to bound that. The main challenge in extending the above approach to the CMDP is that considering only the reward may lead to a larger violation, since we need to balance both the reward and utility. Thus, one needs to judiciously select the reward based on the total observed reward and utility corresponding to a drawn arm so that the EXP-3 algorithm can choose the arm closest to the optimal one. The natural idea is to set the reward to zero if the observed utility over the epoch does not satisfy the constraint, i.e., if $G_i(B_i)$ is the cumulative utility received

Algorithm 2: Double Restart Non-stationary Triple-Q

Choose $W = K^{5/9}, \mathcal{J}$ defined in Eq. (32) , $\gamma_0 = \min\left\{1, \sqrt{\frac{(K/W)\log(K/W)}{(e-1)KH}}\right\}, \lambda = 1/9$;
Initialize weights of the bandit arms $s_1(j) = 1, \forall j = 0, 1, \ldots, J$;
for *epoch* $i = 1, \ldots, \frac{K}{W}$ **do**
  Update $p_i(j) \leftarrow (1-\gamma_0)\frac{s_i(j)}{\sum_{j'=0}^{J} s_i(j')} + \frac{\gamma_0}{J+1}, \forall j = 0, 1, \ldots, J$;
  Draw an arm $A_i \in [J]$ randomly according to the probabilities $p_i(0), \ldots, p_i(J)$;
  Set the estimated budget $B_i \leftarrow \frac{K^{1/3}W^{\frac{A_i}{J}}}{\Delta^{3/2}W}$;
  Run a new instance of Algorithm 1 for W episodes with parameter value $B \leftarrow B_i, \tilde{b} = B_i^{1-c}K^{\alpha-1}$;
  Observe the cumulative reward R_i and utility G_i.;
  for *arm j=0,1,…,J* **do**
    $\hat{R}_i(j) = \begin{cases} (G_i/K^\lambda) I_{\{j=A_i\}}/(WH(1+1/K^\lambda)p_i(j)) & \text{if } G_i < W\rho \\ (R_i + G_i/K^\lambda) I_{\{j=A_i\}}/(WH(1+1/K^\lambda)p_i(j)) & \text{if } G_i \geq W\rho \end{cases}$; `// normalization`
    $s_{i+1} \leftarrow s_i(j)\exp(\gamma_0 \hat{R}_i(j)/(J+1))$;

after selecting the arm B_i, then one can set

$$\begin{cases} \hat{R}_i(B_i) = 0 & \text{if } G_i(B_i) < W\rho \\ \hat{R}_i(B_i) = R_i(B_i) & \text{if } G_i(B_i) \geq W\rho. \end{cases} \tag{30}$$

Even though it is intuitive, it is not sufficient as it does not distinguish between small and large violation. Thus, we consider the following bandit reward function

$$\begin{cases} \hat{R}_i(B_i) = \dfrac{G_i(B_i)}{K^\lambda} & \text{if } G_i(B_i) < W\rho \\ \hat{R}_i(B_i) = R_i(B_i) + \dfrac{G_i(B_i)}{K^\lambda} & \text{if } G_i(B_i) \geq W\rho. \end{cases} \tag{31}$$

If $G_i(B_i) < W\rho$, then choosing the arm B_i may lead to violating the constraint, hence, we penalize such arm. On the other hand, if $G_i(B_i) \geq W\rho$, the arm may lead to a feasible policy. We thus consider the reward as $R_i(B_i) + G_i(B_i)/K^\lambda$, i.e., the reward is dominated by the accumulated reward. However, the accumulated utility is also considered (albeit with a weight $1/K^\lambda$). Note that since $\lambda > 0$, the weight factor is small as the main focus is to maximize the reward when the constraint is satisfied. Later, we show that how we select λ to balance the regret and the violation. Hence, the weight factor is critical in obtaining sub-linear regret and zero violation.

Next we present a lemma to show the upper bound of the bandit algorithm using our designing of the bandit reward function (31). The proofs can be found in the supplementary materials (Section D).

Lemma 2. *Let $R_i(B_i)(G_i(B_i))$ be the cumulative reward(utility) collected in epoch i by any learning algorithm after running for W episodes with the estimate value B_i chosen using the Exp3 bandit algorithm. If we have $\mathbb{E}[G_i(\hat{B})] \geq W\rho$ then we can obtain*

$$\mathbb{E}\left[\sum_{i=1}^{K/W}(R_i(\hat{B}) - R_i(B_i))\right] = \tilde{\mathcal{O}}(H\sqrt{KW} + HK^{1-\lambda})$$

$$\mathbb{E}\left[\sum_{i=1}^{K/W} G_i(\hat{B}) - G_i(B_i)\right] = \tilde{\mathcal{O}}(HK^\lambda\sqrt{KW}).$$

Note that the above lemma bounds (29). Further, it also bounds the utilities for the choice of $\hat{B}$ and B_i which will be useful to obtain violation.

Next, we will formally define the set $\mathcal{J}$. Subsequently, we will present the results of using "bandit over bandit" with our designing bandit reward function on the Algorithm 1 for the tabular setting. Then we will discuss how to apply it to the linear function approximation setting. We define set $\mathcal{J}$ as

$$\mathcal{J} = \left\{\frac{K^{1/3}}{\Delta^{3/2}W}, \frac{K^{1/3}W^{\frac{1}{J}}}{\Delta^{3/2}W}, \ldots, \frac{K^{1/3}W}{\Delta^{3/2}W}\right\}, \tag{32}$$

as the candidate value for B and we can see that $|\mathcal{J}| = \log(W) + 1 = J + 1$, where $\Delta = \left(\frac{40\sqrt{SAH^6\iota^3}}{\delta}\right)^2$. After an estimated budget B_i for each epoch i is selected. Then we run a new instance of Algorithm 1 for consecutive $W = K^\zeta$ episodes. Each epoch contains $WB_i^c/K^{\alpha\zeta}$ frames. We remark here that when using the Algorithm 1 we need a local budget information, but under assumption $K^{1-\alpha}B^c\tilde{b} \leq B$, we can simply choose $\tilde{b} = B_i^{1-c}K^{\alpha-1}$ with an estimated B_i. The following Theorem states that the Algorithm 2 achieves a sublinear regret and zero constraint violation without the knowledge of the total variation budget B. Detailed proofs are deferred to supplementary materials (Section F).

Theorem 2. *Algorithm 2 achieves the following regret and constraint violation bounds with no prior knowledge of the total variation budget B when $K = \Omega\left((\frac{40\sqrt{SAH^6\iota^3}B^{1/3}}{\delta})^9\right)$, and $K \geq e^{\frac{1}{\delta}}$:*

$$\mathcal{R}(K) = \tilde{\mathcal{O}}(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}B^{\frac{1}{3}}K^{8/9}) \qquad \mathcal{V}(K) = 0$$

7 LINEAR CMDPs

In this section, we consider linear CMDP which can potentially model infinite state space. In particular, we consider reward, utility, and transition probability can be modeled as linear in known feature space Ghosh et al. (2022). The formal definition is given below

Definition 1. *The CMDP is a linear MDP with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any h and k, there exists d* unknown *signed measures $\mu_{k,h} = \{\mu^1_{k,h}, \ldots, \mu^d_{k,h}\}$ over $\mathcal{S}$ such that any $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$\mathbb{P}_{k,h}(x'|x, a) = \langle \phi(x, a), \mu_{k,h}(x') \rangle \tag{33}$$

and there exists (unknown) vectors $\theta_{k,r,h}$, $\theta_{k,g,h} \in \mathbb{R}^d$ such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$,

$$r_{k,h}(x, a) = \langle \phi(x, a), \theta_{k,r,h} \rangle,$$
$$g_{k,h}(x, a) = \langle \phi(x, a), \theta_{k,g,h} \rangle$$

Without loss of generality, we assume $||\phi(x, a)||_2 \leq 1$, $\max\{||\mu_{k,h}||_2, ||\theta_{k,r,h}||_2, ||\theta_{k,g,h}||_2\} \leq \sqrt{d}$.

We adapt the stationary version of the linear CMDP in the non-stationary setup by considering time-varying $\mu_{k,h}$, and $\theta_{k,j,h}$. It extends the non-stationary unconstrained linear MDP Zhou et al. (2020) to the constrained case. We remark that despite being linear, $\mathbb{P}_{k,h}(\cdot|x, a)$ can still have infinite degrees of freedom since $\mu_{k,h}(\cdot)$ is unknown. Note that Ding et al. (2021); Ding and Lavaei (2022) studied another related concept known as linear kernel MDP. In general, linear MDP and linear kernel MDPs are two different classes of MDP Zhou et al. (2021).

Similar to budget variations in the tabular case, we define the total (global) variations on $\mu_{k,h}$ and $\theta_{k,j,h}$ for $j = r, g$ and the total variations as

$$B_j = \sum_{k=2}^{K} \sum_{h=1}^{H} ||\theta_{k,j,h} - \theta_{k-1,j,h}||_2, \tag{34}$$

$$B_p = \sum_{k=2}^{K} \sum_{h=1}^{H} ||\mu_{k,h} - \mu_{k-1,h}||_F, \tag{35}$$

and $B = B_r + B_g + B_p$ is the global budget variation.

Algorithm: Ghosh et al. (2022) proposed an algorithm for the stationary setup. It is a primal-dual adaptation of the unconstrained version Ding and Lavaei (2022) . However, there are some key differences with respect to the unconstrained case. For example, instead of greedy policy with respect to the combined state-action value function one needs the soft-max policy. We adapt the algorithm in the non-stationary case (Algorithm 3 in the supplementary materials G). In particular, we employ the restart strategy to adapt to the non-stationary environment. We divide the total episodes K in K/D frames where each frame consists of D episodes. We employ the algorithm proposed in Ghosh et al. (2022) at each frame. Note that such type of restart strategy is already proposed for the unconstrained version as well (Zhou et al., 2020). However, the algorithm for the constrained linear MDP differs from the unconstrained version, thus, the analysis also differs.

Tabular v.s. Linear Approximation: We remark that although linear CMDPs include tabular CMDPs as a special case (Jin et al., 2018). Directly applying the algorithm to a tabular CMDP will result in higher memory and computational complexity than Nonstationary Triple-Q.

We now flesh out Algorithm 3 for the tabular case which will clarify the memory and computational requirement. We can revert back to the tabular case by setting $\phi(s, a) = e_{s,a}$ where $e_{s,a}$ is a d-dimensional (here $d = |\mathcal{S}||\mathcal{A}|$) vector where $e_{s,a} = 1$ for state-action pair (s, a) and zero for other values of state and action. The $w_{r,h}$ vector update becomes as the following

$$w^k_{r,h}(x, a) = \frac{1}{(n^k_h(x, a) + \lambda)} \sum_{\tau=1}^{n^k_h(x,a)} (r_h(x^\tau_h, a^\tau_h) + V^k_{r,h+1}(x^\tau_{h+1}))$$

where $n^k_h(x, a)$ is the number of times the state-action pair (x, a) has been encountered at step h till episode k. The $Q^k_{r,h}$ update will be

$$Q^k_{r,h}(x, a) = \min\{\langle w^k_{r,h}(x, a), \phi(x, a) \rangle + \beta\sqrt{1/(n^k_h(x, a) + \lambda)}, H\}.$$

In a similar manner, we can update $Q^k_{g,h}$. Note that we need to update this table for every state-action pair at each step and use all the samples generated so far. Using this, one can update $V^k_{r,h}$, and $V^k_{g,h}$ using the soft-max policy.

We further remark that if we maintain $n^k_h(x, a, \tilde{x})$ to be the number of times the state-action-next state $(x, a, \tilde{x})$ has been encountered at step h till episode k. Then

$$w^k_{r,h}(x, a) = \frac{1}{(n^k_h(x, a) + \lambda)} \cdot \left(n^k_h(x, a) r_h(x, a) + \sum_{\tilde{x}} n^k_h(x, a, \tilde{x}) V^k_{r,h+1}(\tilde{x}) \right).$$

In this case, we do not need to go through all samples at each iteration and do not even need to store the old samples. The memory complexity of maintaining the counts $\{n_h(x, a, \tilde{x})\}$ is $O\left(H|\mathcal{S}|^2|\mathcal{A}|\right)$, which is higher than the memory complexity and computational complexity of non-stationary Triple-Q, which are $O\left(H|\mathcal{S}||\mathcal{A}|\right)$, but matches model-based algorithms for tabular settings.

7.1 Main Results

Theorem 3. *With $D = B^{-1/2}d^{1/2}K^{1/2}H^{-1/2}$, Algorithm 3 achieves the following regret and constraint violation bounds:*

$$\mathcal{R}(K) = \mathcal{O}(\frac{1+\delta}{\delta}K^{3/4}H^{9/4}d^{5/4}B^{1/4}\iota)$$
$$\mathcal{V}(K) = \frac{2(1+\xi)}{\xi}\mathcal{O}(K^{3/4}H^{9/4}d^{5/4}B^{1/4}\iota)$$

where $\iota = \log(2\log(|\mathcal{A}|)dT/p)$, and $\xi = 2H/\delta$.

Our algorithm provides a regret guarantee of $\tilde{\mathcal{O}}(d^{5/4}K^{3/4}H^{9/4}B^{1/4})$ and the same order on violation. ξ arises since we truncate the dual variable at ξ in Algorithm 3. Note that regret and violation only scale with d rather than the cardinality of the state space.

Compared to Ding and Lavaei (2022), which also considers linear function approximation (however, it considers linear kernel CMDP rather linear CMDP), we improve their result by a factor of $H^{\frac{1}{4}}$. We also improve the dependence on B and d. Further, we do not need to know the total variations in the optimal solution (B_*), unlike in Ding and Lavaei (2022). The algorithm proposed in Ding and Lavaei (2022) is a model-based policy-based algorithm; ours is a model-free value-based algorithm. Thus, our algorithm enjoys an easy implementation and improved computation efficiency since it does not estimate the next step expected value function as in Ding and Lavaei (2022), which requires an integration oracle to compute a d-dimensional integration at every step.

Zero Violation: Similar to the tabular setup, we obtain zero violation by considering a tighter optimization problem. In particular, if we consider ϵ-tighter constraint where $\epsilon = \min\{\frac{2(1+\xi)}{\xi}\tilde{O}(d^{5/4}B^{1/4}H^{9/4}K^{3/4})/K, \delta/2\}$, the violation is 0. Thus, if $K^{1/4} \geq \frac{4(1+\xi)}{\xi\delta}\tilde{O}(d^{5/4}B^{1/4}H^{9/4})$, we could obtain zero violation while maintaining the same order of regret with respect to K.

Remark 1. *Our algorithm 3 doesn't require the information of the local budget. In the unconstrained version Zhou et al. (2020) achieves $\tilde{\mathcal{O}}(T^{2/3})$ regret if local budget variation is known. We can also achieve $\tilde{\mathcal{O}}(T^{2/3})$ regret and $\tilde{\mathcal{O}}(T^{2/3})$ violation if we assume local budget variation is known.*

7.2 Without knowing the variation budget

Our idea of designing the "bandit over bandit" algorithm can still be applied to the linear CMDPs, We propose an algorithm (see Algorithm 4 in supplementary materials), which can achieve the following result. Details proofs can be found in supplementary materials (Section H).

Theorem 4. *Let $D = B^{-1/2}d^{1/2}K^{1/2}H^{-1/2}, W = \sqrt{K}$, Algorithm 4 achieves the following regret and constraint violation bounds:*

$$\mathcal{R}(K) = \mathcal{O}(\frac{1+\delta}{\delta}K^{7/8}H^{9/4}d^{5/4}B^{1/4}\iota)$$
$$\mathcal{V}(K) = \frac{2(1+\xi)}{\xi}\mathcal{O}(\frac{1+\delta}{\delta}K^{7/8}H^{9/4}d^{5/4}B^{1/4}\iota)$$

We can further achieve zero constraint violation by choosing $\epsilon = \min\{\frac{3(1+\xi)}{\xi}\tilde{O}((1+1/\delta)d^{5/4}\hat{B}^{1/4}H^{9/4}K^{1-\varsigma/4})/K, \delta/2\}$, when $K^8 \geq \frac{6(1+\xi)}{\xi\delta}\tilde{O}(d^{5/4}B^{1/4}H^{9/4})$.

We also provide an approach based on convex optimization to further reduce the order from $\tilde{\mathcal{O}}(K^{7/8})$ to $\tilde{\mathcal{O}}(K^{3/4})$, for both regret and violation see Section I in the supplementary materials for details.

8 SIMULATION

We compare Algorithm 1 with two baseline algorithms: an algorithm (Mao et al., 2020) for *non-stationary* MDPs, and an algorithm (Wei et al., 2022b) for *stationary* constrained MDPs for a grid-world environment. From the simulation results, we observe that our Algorithm 1 can quickly learn a well-performed policy while satisfying the safety constraint even when the MDP varies, while other methods all fail to satisfy the constraints. All the details can be found in supplementary materials (Section J).

9 CONCLUSION

We have studied model-free reinforcement learning algorithms in non-stationary episodic CMDPs. In particular, we consider two settings – one is computationally less intensive for the tabular setting, and another one is computationally more intensive but can be applied to a more general linear approximation setup. We have further presented a general framework for applying any algorithms with zero constraint violation to a more practical scenario where the total variation budget is unknown. Whether we can tighten the bounds for model-free algorithms remains an important future research direction. Whether we can design an approach for using any learning algorithms for CMDPs in a non-stationary environment without the knowledge of the budget also constitutes a future research direction.

Acknowledgements

We thank the anonymous paper reviewers for their insightful comments. The work of Honghao Wei and Lei Ying is supported in part by NSF under grants 2001687, 2112471, 2134081, and 2228974. This work of Ness Shroff and Arnob Ghosh has been partly supported by NSF grants NSF AI Institute (AI-EDGE) 2112471, CNS-2106933, 2007231,

CNS-1955535, and CNS-1901057, and in part by Army Research Office under Grant W911NF-21-1-0244. The work of Xingyu Zhou is supported in part by NSF under grants NSF CNS-2153220.

References

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*.

Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (2003). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Auer, P., Jaksch, T., and Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. *NeurIPS*, 21.

Brantley, K., Dudik, M., Lykouris, T., Miryoosefi, S., Simchowitz, M., Slivkins, A., and Sun, W. (2020). Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 33, pages 16315–16326. Curran Associates, Inc.

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. (2022). Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444.

Bura, A., HasanzadeZonuzy, A., Kalathil, D., Shakkottai, S., and Chamberland, J.-F. (2021). Safe exploration for constrained reinforcement learning with provable guarantees. *arXiv preprint arXiv:2112.00885*.

Chen, L., Jain, R., and Luo, H. (2022). Learning infinite-horizon average-reward markov decision process with constraints. In *icml*, pages 3246–3270. PMLR.

Chen, Y., Dong, J., and Wang, Z. (2021). A primal-dual approach to constrained Markov decision processes. *arXiv preprint arXiv:2101.10895*.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2020). Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *icml*, pages 1843–1854. PMLR.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2022). Hedging the drift: Learning to optimize under nonstationarity. *Management Science*, 68(3):1696–1713.

Coronato, A., Naeem, M., De Pietro, G., and Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. (2021). Provably efficient safe exploration via primal-dual policy optimization. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 3304–3312. PMLR.

Ding, D., Zhang, K., Basar, T., and Jovanovic, M. (2020). Natural policy gradient primal-dual method for constrained markov decision processes. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 33, pages 8378–8390. Curran Associates, Inc.

Ding, Y. and Lavaei, J. (2022). Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints. *arXiv preprint arXiv:2201.11965*.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. (2021). A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *aistats*, pages 3538–3546. PMLR.

Efroni, Y., Mannor, S., and Pirotta, M. (2020). Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. (2020). Dynamic regret of policy optimization in non-stationary environments. *NeurIPS*, 33:6743–6754.

Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *jmlr*, 16(1):1437–1480.

Ghosh, A., Zhou, X., and Shroff, N. (2022). Provably efficient model-free constrained rl with linear function approximation. In *NeurIPS*.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? In *Advances Neural Information Processing Systems (NeurIPS)*, volume 31, pages 4863–4873.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. (2021). Learning policies with zero or bounded constraint violation for constrained MDPs. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 34.

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., and Başar, T. (2020). Model-free non-stationary rl: Near-optimal regret and applications in multi-agent rl and inventory control. *arXiv preprint arXiv:2010.03161*.

Neely, M. J. (2016). Energy-aware wireless scheduling with near-optimal backlog and convergence time tradeoffs.

IEEE/ACM Transactions on Networking, 24(4):2223–2236.

Ortner, R., Gajane, P., and Auer, P. (2020). Variational regret bounds for reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 81–90. PMLR.

Pan, L., Cai, Q., Meng, Q., Chen, W., and Huang, L. (2021). Reinforcement learning with dynamic boltzmann softmax updates. In *ijcai*, IJCAI'20.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. (2020). Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15277–15287. Curran Associates, Inc.

Singh, R., Gupta, A., and Shroff, N. B. (2020). Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*.

Touati, A. and Vincent, P. (2020). Efficient learning in non-stationary linear markov decision processes. *arXiv preprint arXiv:2010.12870*.

Wei, C.-Y. and Luo, H. (2021). Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *colt*, pages 4300–4354. PMLR.

Wei, H., Liu, X., and Ying, L. (2022a). A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In *AAAI Conf. Artificial Intelligence*.

Wei, H., Liu, X., and Ying, L. (2022b). Triple-Q: a model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*.

Zhong, H., Yang, Z., and Szepesvári, Z. W. C. (2021). Optimistic policy optimization is provably efficient in non-stationary mdps. *arXiv preprint arXiv:2110.08984*.

Zhou, D., He, J., and Gu, Q. (2021). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *icml*, pages 12793–12802. PMLR.

Zhou, H., Chen, J., Varshney, L. R., and Jagmohan, A. (2020). Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244*.

A NOTATION TABLE

The notations used throughout this paper are summarized in Table 2.

Table 2: Notation Table

Notation	Definition
K	total number of episodes
S	number of states
A	number of actions
H	length of each episode
B	total variation budget
W	number of episodes in one epoch.
D	number of episodes in one frame.
B_i	arm selected by the bandit algorithm.
α_t	learning rate
$R_i(B_i)(G_i(B_i))$	reward/utility collected at the epoch i under selected estimate value B_i
$Q_{k,h}(x,a)(C_{k,h}(x,a))$	estimated reward (utility) Q-function at step h in episode k
$Q^{\pi}_{k,h}(x,a)(C^{\pi}_{k,h}(x,a))$	reward (utility) Q-function at step h in episode k under policy π.
$V_{k,h}(x)(W_{k,h}(x))$	estimated reward (utility) value-function at step h in episode k
$V^{\pi}_{k,h}(x)(W^{\pi}_{k,h}(x))$	reward (utility) value-function at step h in episode k under policy π
$F_{k,h}(x,a)$	$F_{k,h}(x,a) = Q_{k,h}(x,a) + \frac{Z_k}{\eta} C_{k,h}(x,a)$.
$U_{k,h}(x)$	$U_{k,h}(x) = V_{k,h}(x) + \frac{Z_k}{\eta} W_{k,h}(x)$.
$r_{k,h}(x,a)(g_{k,h}(x,a))$	reward (utility) of (state, action) pair (x,a) at step h in episode k
$N_{k,h}(x,a)$	number of visits to (x,a) when at step h in episode k (not including k)
Z_k	dual estimation (virtual queue) in episode k.
$q^*_{k,h}$	The optimal solution to the LP (12) in episode k
$q^{\epsilon,*}_{k,h}$	optimal solution to the tightened LP (18) in episode k
π^*_k	optimal policy in episode k
δ	Slater's constant.
d	dimension of the feature vector.
b_t	the UCB bonus for given t
$\mathbb{I}(\cdot)$	indicator function
$\mathbb{P}_{k,h}$	transition kernel at step h in episode k
$\hat{\mathbb{P}}_{k,h}$	empirical transition kernel at step h in episode k
B_r, B_g, B_p	variation budget for reward, utility, and transition
$B_r^{(T)}, B_g^{(T)}, B_p^{(T)}$	variation budget for reward, utility, and transition in frame T
$\phi(x,a)$	feature map for the linear MDP
$\theta_{k,r,h}, \theta_{k,g,h}, \mu_{k,h}$	underlying parameters for the linear MDP

B AUXILIARY LEMMAS

In this section, we state several lemmas that used in our analysis. The first lemma establishes some key properties of the learning rates used in Non-stationary Triple-Q. The proof closely follows the proof of Lemma 4.1 in Jin et al. (2018).

Lemma 3. *Recall that the learning rate used in Triple-Q is* $\alpha_t = \frac{\chi+1}{\chi+t}$, *and*

$$\alpha_t^0 = \prod_{j=1}^{t}(1-\alpha_j) \quad \textit{and} \quad \alpha_t^i = \alpha_i \prod_{j=i+1}^{t}(1-\alpha_j). \tag{36}$$

The following properties hold for α_t^i :

(a) $\alpha_t^0 = 0$ *for* $t \geq 1, \alpha_t^0 = 1$ *for* $t = 0$.

(b) $\sum_{i=1}^t \alpha_t^i = 1$ *for* $t \geq 1$, $\sum_{i=1}^t \alpha_t^i = 0$ *for* $t = 0$.

(c) $\frac{1}{\sqrt{\chi+t}} \leq \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{\chi+i}} \leq \frac{2}{\sqrt{\chi+t}}$.

(d) $\sum_{t=i}^\infty \alpha_t^i = 1 + \frac{1}{\chi}$ *for every* $i \geq 1$.

(e) $\sum_{i=1}^t (\alpha_t^i)^2 \leq \frac{\chi+1}{\chi+t}$ *for every* $t \geq 1$.

□

Proof. The proof of (a) and (b) are straightforward by using the definition of α_t^i. The proof of (d) is the same as that in Jin et al. (2018).

(c): We next prove (c) by induction.

For $t = 1$, we have $\sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{\chi+i}} = \frac{\alpha_1^1}{\sqrt{\chi+1}} = \frac{1}{\sqrt{\chi+1}}$, so (c) holds for $t = 1$.

Now suppose that (c) holds for $t-1$ for $t \geq 2$, i.e.

$$\frac{1}{\sqrt{\chi+t-1}} \leq \sum_{i=1}^{t-1} \frac{\alpha_t^i}{\sqrt{\chi+i-1}} \leq \frac{2}{\sqrt{\chi+t-1}}.$$

From the relationship $\alpha_t^i = (1-\alpha_t)\alpha_{t-1}^i$ for $i = 1, 2, \ldots, t-1$, we have

$$\sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{\chi+i}} = \frac{\alpha_t}{\sqrt{\chi+t}} + (1-\alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{\chi+i}}.$$

Now we apply the induction assumption. To prove the lower bound in (c), we have

$$\frac{\alpha_t}{\sqrt{\chi+t}} + (1-\alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{\chi+i}} \geq \frac{\alpha_t}{\sqrt{\chi+t}} + \frac{1-\alpha_t}{\sqrt{\chi+t-1}} \geq \frac{\alpha_t}{\sqrt{\chi+t}} + \frac{1-\alpha_t}{\sqrt{\chi+t}} \geq \frac{1}{\sqrt{\chi+t}}.$$

To prove the upper bound in (c), we have

$$\begin{aligned}
\frac{\alpha_t}{\sqrt{\chi+t}} + (1-\alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{\chi+i}} \leq & \frac{\alpha_t}{\sqrt{\chi+t}} + \frac{2(1-\alpha_t)}{\sqrt{\chi+t-1}} = \frac{\chi+1}{(\chi+t)\sqrt{\chi+t}} + \frac{2(t-1)}{(\chi+t)\sqrt{\chi+t-1}}, \\
= & \frac{1-\chi-2t}{(\chi+t)\sqrt{\chi+t}} + \frac{2(t-1)}{(\chi+t)\sqrt{\chi+t-1}} + \frac{2}{\sqrt{\chi+t}} \\
\leq & \frac{-\chi-1}{(\chi+t)\sqrt{\chi+t-1}} + \frac{2}{\sqrt{\chi+t}} \leq \frac{2}{\sqrt{\chi+t}}.
\end{aligned} \tag{37}$$

(e) According to its definition, we have

$$\begin{aligned}
\alpha_t^i = & \frac{\chi+1}{i+\chi} \cdot \left(\frac{i}{i+1+\chi} \frac{i+1}{i+2+\chi} \cdots \frac{t-1}{t+\chi} \right) \\
= & \frac{\chi+1}{t+\chi} \cdot \left(\frac{i}{i+\chi} \frac{i+1}{i+1+\chi} \cdots \frac{t-1}{t-1+\chi} \right) \leq \frac{\chi+1}{\chi+t}.
\end{aligned} \tag{38}$$

Therefore, we have

$$\sum_{i=1}^{t} (\alpha_t^i)^2 \leq [\max_{i \in [t]} \alpha_t^i] \cdot \sum_{i=1}^{t} \alpha_t^i \leq \frac{\chi+1}{\chi+t},$$

because $\sum_{i=1}^t \alpha_t^i = 1$. □

Lemma 4. *For any $(x,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, we have the following bounds on $Q_{k,h}(x,a)$ and $C_{k,h}(x,a)$:*

$$0 \le Q_{k,h}(x,a) \le H^2(\sqrt{\iota}+2\tilde{b})$$
$$0 \le C_{k,h}(x,a) \le H^2(\sqrt{\iota}+2\tilde{b}).$$

Proof. We first consider the last step of an episode, i.e. $h=H$. Recall that $V_{k,H+1}(x)=0$ for any k and x by its definition and $Q_{0,H}=H\le H(\sqrt{\iota}+2\tilde{b})$. Suppose $Q_{k',H}(x,a)\le H(\sqrt{\iota}+2\tilde{b})$ for any $k'\le k-1$ and any (x,a). Then,

$$Q_{k,H}(x,a) = (1-\alpha_t)Q_{k_t,H}(x,a) + \alpha_t\left(r_{k,H}(x,a)+b_t+2H\tilde{b}\right) \tag{39}$$
$$\le \max\left\{H\sqrt{\iota}+2\tilde{b}H, 1+\frac{H\sqrt{\iota}}{4}+2H\tilde{b}\right\} \le H\sqrt{\iota}+2\tilde{b}H, \tag{40}$$

where $t=N_{k,H}(x,a)$ is the number of visits to state-action pair (x,a) when in step H by episode k (but not include episode k) and k_t is the index of the episode of the most recent visit. Therefore, the upper bound holds for $h=H$. Note that $Q_{0,h}=H\le H(H-h+1)(\sqrt{\iota}+2\tilde{b})$. Now suppose the upper bound holds for $h+1$, and also holds for $k'\le k-1$. Consider step h in episode k :

$$Q_{k,h}(x,a) = (1-\alpha_t)Q_{k_t,h}(x,a) + \alpha_t\left(r_{k,h}(x,a)+V_{k_t,h+1}(x_{k_t,h+1})+b_t+2\tilde{b}H\right),$$

where $t=N_{k,h}(x,a)$ is the number of visits to state-action pair (x,a) when in step h by episode k (but not include episode k) and k_t is the index of the episode of the most recent visit. We also note that $V_{k,h+1}(x)\le\max_a Q_{k,h+1}(x,a)\le H(H-h)(\sqrt{\iota}+2\tilde{b})$. Therefore, we obtain

$$\begin{aligned}Q_{k,h}(x,a) \le& \max\left\{H(H-h+1)(\sqrt{\iota}+2\tilde{b}), 1+H(H-h)(\sqrt{\iota}+2\tilde{b})+\frac{H\sqrt{\iota}}{4}+2\tilde{b}H\right\}\\ \le& H(H-h+1)(\sqrt{\iota}+2\tilde{b}).\end{aligned}$$

Therefore, we can conclude that $Q_{k,h}(x,a)\le H^2(\sqrt{\iota}+2\tilde{b})$ for any k,h and (x,a). The proof for $C_{k,h}(x,a)$ is identical. □

Lemma 5. *Consider any frame T, any episode k'. Let t=$N_{k,h}(x,a)$ be the number of visits to (x,a) at step h before episode k in the current frame and let $k_1,\dots,k_t<k$ be the indices of these episodes. Under any policy π, with probability at least $1-\frac{1}{K^3}$, the following inequalities hold simultaneously for all $(x,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$,*

$$\left|\sum_{i=1}^{t}\alpha_t^i\left\{(\hat{\mathbb{P}}_{k_i,h}-\mathbb{P}_{k_i,h})V_{k,h+1}^{\pi}\right\}(x,a)\right| \le \frac{1}{4}\sqrt{\frac{H^2\iota(\chi+1)}{(\chi+t)}},$$
$$\left|\sum_{i=1}^{t}\alpha_t^i\left\{(\hat{\mathbb{P}}_{k_i,h}-\mathbb{P}_{k_i,h})W_{k,h+1}^{\pi}\right\}(x,a)\right| \le \frac{1}{4}\sqrt{\frac{H^2\iota(\chi+1)}{(\chi+t)}}.$$

Proof. Without loss of generality, we consider $T=1$. Fix any $(x,a,h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{H}$, a fixed episode k, and any $n\in[K^{\alpha}/B^c]$, define

$$X(n)=\sum_{i=1}^{n}\alpha_\tau^i\cdot\mathbb{I}_{\{k_i\le K\}}\left\{(\hat{\mathbb{P}}_{k_i,h}-\mathbb{P}_{k_i,h})V_{k,h+1}^{\pi}\right\}(x,a).$$

Let $\mathcal{F}_i$ be the $\sigma-$algebra generated by all the random variables until step h in episode k_i. Then

$$\mathbb{E}[X(n+1)|\mathcal{F}_n]=X(n)+\mathbb{E}\left[\alpha_\tau^{n+1}\mathbb{I}_{\{k_{n+1}\le K\}}\left\{(\hat{\mathbb{P}}_{k_{n+1},h}-\mathbb{P}_{k_{n+1},h})V_{k,h+1}^{\pi}\right\}(x,a)|\mathcal{F}_n\right]=X(n),$$

which shows that $X(n)$ is a martingale. We also have for $1\le m\le n$,

$$|X(m)-X(m-1)|\le\alpha_\tau^m\left|\left\{(\hat{\mathbb{P}}_{k_m,h}-\mathbb{P}_{k_m,h})V_{k,h+1}^{\pi}\right\}(x,a)\right|\le\alpha_\tau^m H$$

Let $k_i = K+1$ if it is taken for fewer than i times, and let $\sigma = \sqrt{8\log\left(\sqrt{2SAH}K\right)\sum_{i=1}^{\tau}(\alpha_\tau^i H)^2}$. Then by applying the Azuma-Hoeffding inequality, we have with probability at least $1-2\exp\left(-\frac{\sigma^2}{2\sum_{i=1}^{\tau}(\alpha_\tau^i H)^2}\right) \geq 1-\frac{1}{2S^2A^2H^2K^4}$,

$$|X(\tau)| \leq \sqrt{8\log\left(\sqrt{2SAH}K\right)\sum_{i=1}^{\tau}(\alpha_\tau^i H)^2} \leq \sqrt{\frac{\iota}{16}H^2\sum_{i=1}^{\tau}(\alpha_\tau^i)^2} \leq \frac{1}{4}\sqrt{\frac{H^2\iota(\chi+1)}{\chi+\tau}},$$

Because this inequality holds for any $\tau \in [K]$, it also holds for $\tau = t = N_{k,h}(x,a) \leq K$. Applying the union bound, we obtain that with probability at least $1-\frac{1}{2SAHK^3}$ the following inequality holds simultaneously for all $(x,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$,:

$$\left|\sum_{i=1}^{t}\alpha_t^i\left\{(\hat{\mathbb{P}}_{k_i,h}-\mathbb{P}_{k_i,h})V_{k,h+1}^{\pi}\right\}(x,a)\right| \leq \frac{1}{4}\sqrt{\frac{H^2\iota(\chi+1)}{(\chi+t)}}.$$

Following a similar analysis, we also have that with probability at least $1-\frac{1}{2SAHK^3}$ the following inequality holds simultaneously for all $(x,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$,:

$$\left|\sum_{i=1}^{t}\alpha_t^i\left\{(\hat{\mathbb{P}}_{k_i,h}-\mathbb{P}_{k_i,h})W_{k,h+1}^{\pi}\right\}(x,a)\right| \leq \frac{1}{4}\sqrt{\frac{H^2\iota(\chi+1)}{(\chi+t)}}.$$

Therefore applying a union bound on the two events we finish proving the lemma.

□

C PROOFS OF THE TECHNICAL LEMMAS

Lemma 6. *For any frame T, any x,a,h and any $(T-1)K^\alpha/B^c \leq k_1 \leq k_2 \leq TK^\alpha/B^c$, we have*

$$|Q_{k_1,h}^{\pi}(x,a) - Q_{k_2,h}^{\pi'}(x,a)| \leq H\tilde{b}$$

$$|C_{k_1,h}^{\pi}(x,a) - C_{k_2,h}^{\pi'}(x,a)| \leq H\tilde{b}$$

Proof. First define B_h^r, B_h^g, B_h^p to be the variation of reward, utility functions and transitions at step h within frame T.

$$B_h^r = \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\sup_{x,a}|r_{k,h}(x,a)-r_{k+1,h}(x,a)| \tag{41}$$

$$B_h^g = \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\sup_{x,a}|g_{k,h}(x,a)-g_{k+1,h}(x,a)| \tag{42}$$

$$B_h^p = \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\sup_{x,a}\|\mathbb{P}_{k,h}(\cdot|x,a)-\mathbb{P}_{k+1,h}(\cdot|x,a)\|_1 \tag{43}$$

We will prove the following statement by induction.

$$|Q_{k_1,h}^{\pi}(x,a) - Q_{k_2,h}^{\pi'}(x,a)| \leq \sum_{h'=h}^{H}B_{h'}^r + H\sum_{h'=h}^{H}B_{h'}^p$$

For step H, the statement holds because for any (x,a),

$$\begin{aligned}|Q_{k_1,H}^{\pi}(x,a) - Q_{k_2,H}^{\pi'}(x,a)| =& |r_{k_1,H}(x,a)-r_{k_2,H}(x,a)| \\ \leq& \sum_{k=k_1}^{k_2-1}|r_{k,H}(x,a)-r_{k+1,H}(x,a)| \leq B_H^r\end{aligned}$$

Now suppose the statement holds for $h+1$, then

$$\begin{aligned}
&Q^{\pi}_{k_1,h}(x,a) - Q^{\pi'}_{k_2,h}(x,a)\\
=&\mathbb{P}_{k_1,h}V^{\pi}_{k_1,h+1}(x,a) - \mathbb{P}_{k_2,h}V^{\pi'}_{k_2,h+1}(x,a) + r_{k_1,h}(x,a) - r_{k_2,h}(x,a)\\
\leq&\mathbb{P}_{k_1,h}V^{\pi}_{k_1,h+1}(x,a) - \mathbb{P}_{k_2,h}V^{\pi'}_{k_2,h+1}(x,a) + B^r_h\\
=&\sum_{x'}\mathbb{P}_{k_1,h}(x'|x,a)V^{\pi}_{k_1,h+1}(x') - \sum_{x'}\mathbb{P}_{k_2,h}(x'|x,a)V^{\pi'}_{k_2,h+1}(x') + B^r_h\\
=&\sum_{x'}\mathbb{P}_{k_1,h}(x'|x,a)Q^{\pi}_{k_1,h+1}(x',\pi_{h+1}(x')) - \sum_{x'}\mathbb{P}_{k_2,h}(x'|x,a)Q^{\pi'}_{k_2,h+1}(x',\pi'_{h+1}(x')) + B^r_h
\end{aligned}$$

According to the hypothesis on $h+1$, we have

$$Q^{\pi}_{k_1,h+1}(x',\pi_{h+1}(x')) \leq Q^{\pi'}_{k_2,h+1}(x',\pi'_{h+1}(x')) + \sum_{h'=h+1}^{H} B^r_{h'} + H\sum_{h'=h+1}^{H} B^p_{h'} \tag{44}$$

Therefore

$$\begin{aligned}
&Q^{\pi}_{k_1,h}(x,a) - Q^{\pi'}_{k_2,h}(x,a)\\
\leq&\sum_{x'}\left(\mathbb{P}_{k_1,h}(x'|x,a) - \mathbb{P}_{k_2,h}(x'|x,a)\right)Q^{\pi'}_{k_2,h+1}(x',\pi_{h+1}(x')) + \sum_{h'=h}^{H} B^r_{h'} + H\sum_{h'=h+1}^{H} B^p_{h'}\\
\leq&\|\mathbb{P}_{k_1,h}(\cdot|x,a) - \mathbb{P}_{k_2,h}(\cdot|x,a)\|_1 \cdot H + \sum_{h'=h}^{H} B^r_{h'} + H\sum_{h'=h+1}^{H} B^p_{h'}\\
\leq&B^p_h H + \sum_{h'=h}^{H} B^r_{h'} + H\sum_{h'=h+1}^{H} B^p_{h'}\\
\leq&\sum_{h'=h}^{H} B^r_{h'} + H\sum_{h'=h}^{H} B^p_{h'}
\end{aligned}$$

where the last inequality comes from the assumption on $\tilde{b}$. The same analysis can be applied to $|C^{\pi}_{k_1,h}(x,a) - C^{\pi}_{k_2,h}(x,a)|$. We finish the proof by using the fact that $\sum_{h'=h}^{H} B^r_{h'} + H\sum_{h'=h}^{H} B^p_{h'} \leq H\tilde{b}$. □

Lemma 7. *With probability at least* $1 - \frac{1}{K^3}$, *the following inequality holds simultaneously for all* $(x,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$:

$$\left\{F_{k,h} - F^{\pi}_{k,h}\right\}(x,a) \geq 0, \tag{45}$$

Let π *be a joint policy such that* π *is the optimal policy for the* ϵ*-tight problem at episode* k, *whose reward (utility)* Q *value functions at step* h *are denoted by* $Q^{\epsilon,*}_{k,h}(C^{\epsilon,*}_{k,h})$. *Then we can further obtain*

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right)q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a)\right] \leq \frac{(\eta + K^{1-\alpha})H^2B^c}{\eta K}. \tag{46}$$

The function F *will be defined in Eq.*(47).

Proof. Consider frame T and episodes in frame T. Define $Z = Z_{(T-1)K^{\alpha}/B^c+1}$ because the value of the virtual queue does not change during each frame. We further define/recall the following notations:

$$\begin{aligned}
F_{k,h}(x,a) &= Q_{k,h}(x,a) + \frac{Z}{\eta}C_{k,h}(x,a), \quad U_{k,h}(x) = V_{k,h}(x) + \frac{Z}{\eta}W_{k,h}(x)\\
F^{\pi}_{k,h}(x,a) &= Q^{\pi}_{k,h}(x,a) + \frac{Z}{\eta}C^{\pi}_{k,h}(x,a), \quad U^{\pi}_{k,h}(x) = V^{\pi}_{k,h}(x) + \frac{Z}{\eta}W^{\pi}_{k,h}(x).
\end{aligned} \tag{47}$$

From the updating rule of Q functions, we first know that

$$\{Q_{k,h} - Q^{\pi}_{k,h}\}(x,a) = \alpha^0_t\{Q_{(T-1)K^{\alpha}/B^c+1,h} - Q^{\pi}_{k,h}\}(x,a)$$

$$+\sum_{i=1}^{t}\alpha_t^i\left(\{V_{k_i,h+1}-V^\pi_{k,h+1}\}(x_{k_i,h+1})+\{(\hat{\mathbb{P}}^{k_i}_{k,h}-\mathbb{P}_{k,h})V^\pi_{k,h+1}\}(x,a)+b_i+2H\tilde{b}\right) \quad (48)$$

Then we have with probability at least $1-\frac{1}{k^3}$

$$\begin{aligned}
&\{F_{k,h}-F^\pi_{k,h}\}(x,a)\\
=&\alpha_t^0\left\{F_{(T-1)K^\alpha/B^c+1,h}-F^\pi_{k,h}\right\}(x,a)\\
&+\sum_{i=1}^{t}\alpha_t^i\left(\left\{U_{k_i,h+1}-U^\pi_{k,h+1}\right\}(x_{k_i,h+1})+\{(\hat{\mathbb{P}}^{k_i}_{k,h}-\mathbb{P}_{k,h})U^\pi_{k,h+1}\}(x,a)+\left(1+\frac{Z}{\eta}\right)(b_i+2H\tilde{b})\right)\\
=&\alpha_t^0\left\{F_{(T-1)K^\alpha/B^c+1,h}-F^\pi_{k,h}\right\}(x,a)+\sum_{i=1}^{t}\alpha_t^i\left(\{(\hat{\mathbb{P}}^{k_i}_{k,h}-\mathbb{P}_{k_i,h})U^\pi_{k,h+1}\}\right)\\
&+\sum_{i=1}^{t}\alpha_t^i\left(\left\{U_{k_i,h+1}-U^\pi_{k,h+1}\right\}(x_{k_i,h+1})+\{(\mathbb{P}_{k_i,h}-\mathbb{P}_{k,h})U^\pi_{k,h+1}\}(x,a)+\left(1+\frac{Z}{\eta}\right)(b_i+2H\tilde{b})\right)\\
\geq_{(a)}&\alpha_t^0\left\{F_{(T-1)K^\alpha/B^c+1,h}-F^\pi_{k,h}\right\}(x,a)\\
&+\sum_{i=1}^{t}\alpha_t^i\left(\left\{U_{k_i,h+1}-U^\pi_{k,h+1}\right\}(x_{k_i,h+1})+\{(\mathbb{P}_{k_i,h}-\mathbb{P}_{k,h})U^\pi_{k,h+1}\}(x,a)+\left(1+\frac{Z}{\eta}\right)(b_i+H\tilde{b})\right)\\
\geq_{(b)}&\alpha_t^0\left\{F_{(T-1)K^\alpha/B^c+1,h}-F^\pi_{k,h}\right\}(x,a)+\sum_{i=1}^{t}\alpha_t^i\left(\left\{U_{k_i,h+1}-U^\pi_{k,h+1}\right\}(x_{k_i,h+1})+\left(1+\frac{Z}{\eta}\right)H\tilde{b}\right)\\
=&\alpha_t^0\left\{F_{(T-1)K^\alpha/B^c+1,h}-F^\pi_{k,h}\right\}(x,a)+\sum_{i=1}^{t}\alpha_t^i\left\{U_{k_i,h+1}-U^\pi_{k_i,h+1}\right\}(x_{k_i,h+1})\\
&+\sum_{i=1}^{t}\alpha_t^i\left\{U^\pi_{k_i,h+1}-U^\pi_{k,h+1}\right\}(x_{k_i,h+1})+\left(1+\frac{Z}{\eta}\right)H\tilde{b}\\
=_{(c)}&\alpha_t^0\left\{F_{(T-1)K^\alpha/B^c+1,h}-F^\pi_{k,h}\right\}(x,a)+\sum_{i=1}^{t}\alpha_t^i\left(\max_a F_{k_i,h+1}(x_{k_i,h+1},a)-F^\pi_{k_i,h+1}(x_{k_i,h+1},\pi(x_{k_i,h+1}))\right)\\
&+\sum_{i=1}^{t}\alpha_t^i\left\{U^\pi_{k_i,h+1}-U^\pi_{k,h+1}\right\}(x_{k_i,h+1})+\left(1+\frac{Z}{\eta}\right)H\tilde{b}\\
\geq_{(d)}&\alpha_t^0\left\{F_{(T-1)K^\alpha/B^c+1,h}-F^\pi_{k,h}\right\}(x,a)+\sum_{i=1}^{t}\alpha_t^i\left(\max_a F_{k_i,h+1}(x_{k_i,h+1},a)-F^\pi_{k_i,h+1}(x_{k_i,h+1},\pi(x_{k_i,h+1}))\right)\\
&-\sum_{i=1}^{t}\alpha_t^i|(1+\frac{Z}{\eta})H\tilde{b}|+(1+\frac{Z}{\eta})H\tilde{b}\\
\geq&\alpha_t^0\left\{F_{(T-1)K^\alpha/B^c+1,h}-F^\pi_{k_i,h}\right\}(x,a)+\sum_{i=1}^{t}\alpha_t^i\left\{F_{k_i,h+1}-F^\pi_{k_i,h+1}\right\}(x_{k_i,h+1},\pi(x_{k_i,h+1})),
\end{aligned} \quad (49)$$

where inequality (a) holds because that

$$\left|\sum_{i=1}^{t}\alpha_t^i\left\{(\mathbb{P}_{k_i,h}-\mathbb{P}_{k,h})V^\pi_{k,h+1}\right\}(x,a)\right|=\left|\sum_{i=1}^{t}\sum_{j=k_i}^{k-1}\alpha_t^i\left\{(\mathbb{P}_{j,h}-\mathbb{P}_{j+1,h})V^\pi_{k,h+1}\right\}(x,a)\right|\leq\tilde{b}H,$$

and the same analysis can be applied to $\left|\sum_{i=1}^{t}\alpha_t^i\left\{(\mathbb{P}_{k_i,h}-\mathbb{P}_{k,h})W^\pi_{k,h+1}\right\}(x,a)\right|$. The inequality (b) is true due to the concentration result in Lemma 5 and

$$\sum_{i=1}^{t}\alpha_t^i(1+\frac{Z}{\eta})b_i=\frac{1}{4}\sum_{i=1}^{t}\alpha_t^i(1+\frac{Z}{\eta})\sqrt{\frac{H^2\iota(\chi+1)}{\chi+t}}\geq\frac{\eta+Z}{4\eta}\sqrt{\frac{H^2\iota(\chi+1)}{\chi+t}}.$$

Equality (c) holds because our algorithm selects the action that maximizes $F_{k_i,h+1}(x_{k_i,h+1},a)$ so $U_{k_i,h+1}(x_{k_i,h+1}) = \max_a F_{k_i,h+1}(x_{k_i,h+1},a)$, and inequality (c) is obtained by using Lemma 6 and the property (d) of the learning rate.

The inequality above suggests that we can prove $\{F_{k,h} - F^{\pi}_{k,h}\}(x,a)$ for any (x,a) if (i)

$$\left\{F_{(T-1)K^{\alpha}/B^c+1,h} - F^{\pi}_{k,h}\right\}(x,a) \geq 0,$$

i.e. the result holds at the beginning of the frame and (ii)

$$\left\{F_{k',h+1} - F^{\pi}_{k',h+1}\right\}(x,a) \geq 0 \quad \text{for any} \quad k' \leq k$$

and (x,a), i.e. the result holds for step $h+1$ in all the episodes in the *same* frame.

It is straightforward to see that (i) holds because all reward and cost Q-functions are set to H at the beginning of each frame.

We now prove condition (ii) using induction, and consider the first frame, i.e. $T=1$. The proof is identical for other frames. Consider $h=H$ i.e. the last step. In this case, inequality (49) becomes

$$\{F_{k,H} - F^{\pi}_{k,H}\}(x,a) \geq \alpha^0_t \left\{H + \frac{Z_1}{\eta}H - F^{\pi}_{k,H}\right\}(x,a) \geq 0, \tag{50}$$

i.e. condition (ii) holds for any k in the first frame and $h=H$. By applying induction on h, we conclude that

$$\{F_{k,h} - F^{\pi}_{k,h}\}(x,a) \geq 0. \tag{51}$$

holds for any k, h, and (x,a), which completes the proof of (45). Since Eq. (45) can only be applied to a single policy, in order to have a bound on $\sum_{k=1}^{K}\sum_a \left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right) q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a)$, we first need to substitute $F^{\pi}_{k,1}$ with $F^{\epsilon,*}_{k,1}$ in Eq. (45), and use a union bound over all the episodes, which means with probability at least $1-\frac{1}{K^2}$ that $F_{k,1} - F^{\epsilon,*}_{k,1} \geq 0$. Let $\mathcal{E}$ denote such event that $F_{k,h} - F^{\epsilon,*}_{k,h} \geq 0$ holds for all k, h and (x,a). Then we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\sum_a \left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right) q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a)\right] \\
=&\mathbb{E}\left[\sum_{k=1}^{K}\sum_a \left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right) q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a)\middle|\mathcal{E}\right]\Pr(\mathcal{E}) + \mathbb{E}\left[\sum_{k=1}^{K}\sum_a \left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right) q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a)\middle|\mathcal{E}^c\right]\Pr(\mathcal{E}^c) \\
\leq& KH\left(1 + \frac{K^{1-\alpha}B^cH}{\eta}\right)\frac{1}{K^2} \leq \frac{(\eta + K^{1-\alpha})H^2B^c}{\eta K}.
\end{aligned} \tag{52}$$

□

Lemma 8. *Under our algorithm, we have for any $T \in [K^{1-\alpha}\cdot B^c]$,*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=(T-1)K^{\alpha}/B^c+1}^{TK^{\alpha}/B^c}\left\{Q_{k,1} - Q^{\pi_k}_{k,1}\right\}(x_{k,1},a_{k,1})\right] \\
\leq& H^2SA + \frac{2(H^3\sqrt{\iota} + 2H^3\tilde{b})K^{\alpha}}{B^c\chi} + \sqrt{\frac{H^4SA\iota K^{\alpha}(\chi+1)}{B^c}} + \frac{2K^{\alpha}H^2\tilde{b}}{B^c} \\
&\mathbb{E}\left[\sum_{k=(T-1)K^{\alpha}/B^c+1}^{TK^{\alpha}/B^c}\left\{C_{k,1} - C^{\pi_k}_{k,1}\right\}(x_{k,1},a_{k,1})\right] \\
\leq& H^2SA + \frac{2(H^3\sqrt{\iota} + 2H^3\tilde{b})K^{\alpha}}{B^c\chi} + \sqrt{\frac{H^4SA\iota K^{\alpha}(\chi+1)}{B^c}} + \frac{2K^{\alpha}H^2\tilde{b}}{B^c}.
\end{aligned}$$

Proof. We prove this lemma for the first frame such that $1 \leq k \leq k^{\alpha}/B^c$. By using the update rule recursively, we have

$$Q_{k,h}(x,a) \leq \alpha^0_t H + \sum_{i=1}^{t}\alpha^i_t\left(r_{k_i,h}(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i + 2H\tilde{b}\right), \tag{53}$$

where $\alpha_t^0 = \prod_{j=1}^t (1-\alpha_j)$ and $\alpha_t^i = \alpha_i \prod_{j=i+1}^t (1-\alpha_j)$. From the inequality above, we further obtain

$$\sum_{k=1}^{K^\alpha/B^c} Q_{k,h}(x,a) \le \sum_{k=1}^{K^\alpha/B^c} \alpha_t^0 H + \sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}(x,a)} \alpha_{N_{k,h}}^i \left(r_{k_i,h}(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i + 2H\tilde{b} \right). \tag{54}$$

We simplify our notation in this proof and use the following notations:

$$N_{k,h} = N_{k,h}(x_{k,h}, a_{k,h}), \quad k_i^{(k,h)} = k_i(x_{k,h}, a_{k,h}),$$

where $k_i^{(k,h)}$ is the index of the episode in which the agent visits state-action pair $(x_{k,h}, a_{k,h})$ for the ith time. Since in a given sample path, (k,h) can uniquely determine $(x_{k,h}, a_{k,h})$, this notation introduces no ambiguity. We note that

$$\sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}} \alpha_{N_{k,h}}^i V_{k_i^{(k,h)},h+1}\left(x_{k_i^{(k,h)},h+1}\right) \le \sum_{k=1}^{K^\alpha/B^c} V_{k,h+1}(x_{k,h+1}) \sum_{t=N_{k,h}}^{\infty} \alpha_t^{N_{k,h}} \le \left(1+\frac{1}{\chi}\right) \sum_k V_{k,h+1}(x_{k,h+1}), \tag{55}$$

Then we obtain

$$\begin{aligned}
&\sum_{k=1}^{K^\alpha/B^c} Q_{k,h}(x_{k,h}, a_{k,h}) \\
&\le \sum_{k=1}^{K^\alpha/B^c} \alpha_t^0 H + (1+\frac{1}{\chi}) \sum_{k=1}^{K^\alpha/B^c} (r_{k,h}(x_{k,h}, a_{k,h}) + V_{k,h+1}(x_{k,h+1})) + \sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}} \alpha_{N_{k,h}}^i b_i + K^\alpha \tilde{b}/B^c \\
&\le \sum_{k=1}^{K^\alpha/B^c} (r_{k,h}(x_{k,h}, a_{k,h}) + V_{k,h+1}(x_{k,h+1})) + HSA + \frac{2(H^2\sqrt{\iota} + 2H^2\tilde{b})K^\alpha}{B^c\chi} \\
&\quad + \frac{1}{2}\sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c,
\end{aligned}$$

where the last inequality holds because (i) we have

$$\sum_{k=1}^{K^\alpha/B^c} \alpha_{N_{k,h}}^0 H = \sum_k H \mathbb{I}_{\{N_{k,h}=0\}} \le HSA,$$

(ii) $V_{k,h+1}(x_{k,h+1}) \le (H^2\sqrt{\iota} + \tilde{b})$, $r_{k,h}(x_{k,h}, a_{k,h}) \le 1$, and (iii) we know that

$$\begin{aligned}
&\sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}} \alpha_{N_{k,h}}^i b_i = \frac{1}{4} \sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}} \alpha_{N_{k,h}}^i \sqrt{\frac{H^2\iota(\chi+1)}{\chi+i}} \le \frac{1}{2} \sum_{k=1}^{K^\alpha/B^c} \sqrt{\frac{H^2\iota(\chi+1)}{\chi+N_{k,h}}} \\
&= \frac{1}{2} \sum_{x,a} \sum_{n=1}^{N_{K^\alpha/B^c,h}(x,a)} \sqrt{\frac{H^2\iota(\chi+1)}{\chi+n}} \le \frac{1}{2} \sum_{x,a} \sum_{n=1}^{N_{K^\alpha/B^c,h}(x,a)} \sqrt{\frac{H^2\iota(\chi+1)}{n}} \overset{(1)}{\le} \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c},
\end{aligned}$$

where the last inequality above holds because the left hand side of (1) is the summation of K^α/B^c terms and $\sqrt{\frac{H^2\iota(\chi+1)}{\chi+n}}$ is a decreasing function of n.

Therefore, it is maximized when $N_{K^\alpha/B^c,h} = K^\alpha/B^c SA$ for all x, a. Thus we can obtain

$$\begin{aligned}
&\sum_{k=1}^{K^\alpha/B^c} Q_{k,h}(x_{k,h}, a_{k,h}) - \sum_k Q_{k,h}^{\pi_k}(x_{k,h}, a_{k,h}) \\
&\le \sum_{k=1}^{K^\alpha/B^c} \left(V_{k,h+1}(x_{k,h+1}) - \mathbb{P}_{k,h} V_{k,h+1}^{\pi_k}(x_{k,h}, a_{k,h}) \right) + HSA + \frac{2(H^2\sqrt{\iota} + 2H^2\tilde{b})K^\alpha}{B^c\chi} \\
&\quad + \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + \frac{2K^\alpha H\tilde{b}}{B^c}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K^\alpha/B^c} \Big(V_{k,h+1}(x_{k,h+1}) - \mathbb{P}_{k,h}V_{h+1}^{\pi_k}(x_{k,h},a_{k,h}) + V_{k,h+1}^{\pi_k}(x_{k,h+1}) - V_{k,h+1}^{\pi_k}(x_{k,h+1})\Big) \\
&\quad + HSA + \frac{2(H^2\sqrt{\iota}+2H^2\tilde{b})K^\alpha}{B^c\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c \\
&= \sum_{k=1}^{K^\alpha/B^c} \Big(V_{k,h+1}(x_{k,h+1})) - V_{k,h+1}^{\pi_k}(x_{k,h+1}) - \mathbb{P}_{k,h}V_{k,h+1}^{\pi_k}(x_{k,h},a_{k,h}) + \hat{\mathbb{P}}_{k,h}V_{k,h+1}^{\pi_k}(x_{k,h},a_{k,h})\Big) \\
&\quad + HSA + \frac{2(H^2\sqrt{\iota}+2H^2\tilde{b})K^\alpha}{B^c\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c \\
&= \sum_{k=1}^{K^\alpha/B^c} \Big(Q_{k,h+1}(x_{k,h+1},a_{k,h+1}) - Q_{k,h+1}^{\pi_k}(x_{k,h+1},a_{k,h+1}) - \mathbb{P}_hV_{k,h+1}^{\pi_k}(x_{k,h},a_{k,h}) + \hat{\mathbb{P}}_{k,h}V_{k,h+1}^{\pi_k}(x_{k,h},a_{k,h}\Big) \\
&\quad + HSA + \frac{2(H^2\sqrt{\iota}+2H^2\tilde{b})K^\alpha}{B^c\chi} \\
&\quad + \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c.
\end{aligned}
$$

Taking the expectation on both sides yields

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K^\alpha/B^c} Q_{k,h}(x_{k,h},a_{k,h}) - \sum_k Q_{k,h}^{\pi_k}(x_{k,h},a_{k,h})\right] \\
\leq&\mathbb{E}\left[\sum_{k=1}^{K^\alpha/B^c} \Big(Q_{k,h+1}(x_{k,h+1},a_{k,h+1}) - Q_{k,h+1}^{\pi_k}(x_{k,h+1},a_{k,h+1})\Big)\right] \\
&\quad + HSA + \frac{2(H^2\sqrt{\iota}+2H^2\tilde{b})K^\alpha}{B^c\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c.
\end{aligned}
$$

Then by using the inequality repeatably, we obtain for any $h \in [H]$,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K^\alpha/B^c} \Big(Q_{k,h}(x_{k,h},a_{k,h}) - Q_{k,h}^{\pi_k}(x_{k,h},a_{k,h})\Big)\right] \\
\leq&H^2SA + \frac{2(H^3\sqrt{\iota}+2H^3\tilde{b})K^\alpha}{B^c\chi} + \sqrt{H^4SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H^2\tilde{b}/B^c.
\end{aligned}
$$

We finish the proof.

□

Lemma 9. *Given $\epsilon \leq \delta$, we have*

$$
\mathbb{E}\left[\sum_a \left\{Q_{k,1}^* q_{k,1}^* - Q_{k,1}^{\epsilon,*} q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\right] \leq \frac{H\epsilon}{\delta}.
$$

□

Proof. Given $q_{k,h}^*(x,a)$ is the optimal solution for episode k, we have

$$
\sum_{h,x,a} q_{k,h}^*(x,a) g_{k,h}(x,a) \geq \rho.
$$

Under Assumption 1, we know that there exists a feasible solution $\{q_{k,h}^{\xi_1}(x,a)\}_{h=1}^H$ such that

$$
\sum_{h,x,a} q_{k,h}^{\xi_1}(x,a) g_{k,h}(x,a) \geq \rho + \delta.
$$

We construct $q^{\xi_2}_{k,h}(x,a) = (1-\frac{\epsilon}{\delta})q^*_{k,h}(x,a) + \frac{\epsilon}{\delta}q^{\xi_1}_{k,h}(x,a)$, which satisfies that

$$\begin{aligned}
\sum_{h,x,a} q^{\xi_2}_{k,h}(x,a) g_{k,h}(x,a) &= \sum_{h,x,a} \left((1-\frac{\epsilon}{\delta})q^*_{k,h}(x,a) + \frac{\epsilon}{\delta}q^{\xi_1}_{k,h}(x,a)\right) g_{k,h}(x,a) \geq \rho + \epsilon,\\
\sum_{h,x,a} q^{\xi_2}_{k,h}(x,a) &= \sum_{x',a'} \mathbb{P}_{k,h-1}(x|x',a') q^{\xi_2}_{k,h-1}(x',a'),\\
\sum_{h,x,a} q^{\xi_2}_{k,h}(x,a) &= 1.
\end{aligned}$$

Also we have $q^{\xi_2}_{k,h}(x,a) \geq 0$ for all (h,x,a). Thus $\{q^{\xi_2}_{k,h}(x,a)\}_{h=1}^H$ is a feasible solution to the ϵ-tightened optimization problem (18). Then given $\{q^{\epsilon,*}_{k,h}(x,a)\}_{h=1}^H$ is the optimal solution to the ϵ-tightened optimization problem, we have

$$\begin{aligned}
&\sum_{h,x,a} \left(q^*_{k,h}(x,a) - q^{\epsilon,*}_{k,h}(x,a)\right) r_{k,h}(x,a)\\
\leq &\sum_{h,x,a} \left(q^*_{k,h}(x,a) - q^{\xi_2}_{k,h}(x,a)\right) r_{k,h}(x,a)\\
\leq &\sum_{h,x,a} \left(q^*_{k,h}(x,a) - \left(1-\frac{\epsilon}{\delta}\right) q^*_{k,h}(x,a) - \frac{\epsilon}{\delta}q^{\xi_1}_{k,h}(x,a)\right) r_{k,h}(x,a)\\
\leq &\sum_{h,x,a} \left(q^*_{k,h}(x,a) - \left(1-\frac{\epsilon}{\delta}\right) q^*_{k,h}(x,a)\right) r_{k,h}(x,a)\\
\leq &\frac{\epsilon}{\delta}\sum_{h,x,a} q^*_{k,h}(x,a) r_{k,h}(x,a) \leq \frac{H\epsilon}{\delta},
\end{aligned}$$

where the last inequality holds because $0 \leq r_{k,h}(x,a) \leq 1$ under our assumption. Therefore the result follows because

$$\begin{aligned}
\sum_a Q^*_{k,1}(x_{k,1},a) q^*_{k,1}(x_{k,1},a) &= \sum_{h,x,a} q^*_{k,h}(x,a) r_{k,h}(x,a)\\
\sum_a Q^{\epsilon,*}_{k,1}(x_{k,1},a) q^{\epsilon,*}_{k,1}(x_{k,1},a) &= \sum_{h,x,a} q^{\epsilon,*}_{k,h}(x,a) r_{k,h}(x,a).
\end{aligned}$$

□

Lemma 10. *Assume $\epsilon \leq \delta$. The expected Lyapunov drift satisfies*

$$\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right]\\
\leq &\frac{B^c}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(-\eta \mathbb{E}\left[\sum_a \left\{\hat{Q}_{k,1} q_1^{\epsilon,*}\right\}(x_{k,1},a) - \hat{Q}_{k,1}(x_{k,1},a_{k,1}) \middle| Z_T = z\right]\right.\\
&\left. + z\mathbb{E}\left[\sum_a \left\{\left(C^{\epsilon,*}_{k,1} - C_{k,1}\right) q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a) \middle| Z_T = z\right]\right) + 2H^4\iota + 4H^4\tilde{b} + \epsilon^2. \qquad (56)
\end{aligned}$$

Proof. Assume $\epsilon \leq \delta$. The expected Lyapunov drift satisfies

$$\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right]\\
\leq &\frac{B^c}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(-\eta \mathbb{E}\left[\sum_a \left\{\hat{Q}_{k,1} q_1^{\epsilon,*}\right\}(x_{k,1},a) - \hat{Q}_{k,1}(x_{k,1},a_{k,1}) \middle| Z_T = z\right]\right.\\
&\left. + z\mathbb{E}\left[\sum_a \left\{\left(C^{\epsilon,*}_{k,1} - C_{k,1}\right) q^{\epsilon,*}_{k,1}\right\}(x_{k,1},a) \middle| Z_T = z\right]\right) + 2H^4\iota + 4H^4\tilde{b} + \epsilon^2. \qquad (57)
\end{aligned}$$

Based on the definition of $L_T = \frac{1}{2}Z_T^2$, the Lyapunov drift is

$$L_{T+1} - L_T \leq Z_T\left(\rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}\right) + \frac{\left(\frac{\bar{C}_T B^c}{K^\alpha} + \epsilon - \rho\right)^2}{2}$$

$$\leq Z_T\left(\rho+\epsilon-\frac{\bar{C}_T B^c}{K^\alpha}\right)+2H^4\iota+4H^4\tilde{b}+\epsilon^2$$
$$\leq \frac{Z_T B^c}{K^\alpha}\sum_{k=TK^\alpha/B^c+1}^{(T+1)K^\alpha/B^c}\left(\rho+\epsilon-\hat{C}_{k,1}(x_{k,1},a_{k,1})\right)+2H^4\iota+4H^4\tilde{b}+\epsilon^2$$

where the first inequality is because the upper bound on $|\hat{C}_{k,1}(x_{k,1},a_{k,1})|$ is $H^2(\sqrt{\iota}+2\tilde{b})$ from Lemma 4. Let $\{q^\epsilon_{k,h}\}_{h=1}^H$ be a feasible solution to the tightened LP (18) at episode k. Then the expected Lyapunov drift conditioned on $Z_T=z$ is

$$\begin{aligned}
&\mathbb{E}\left[L_{T+1}-L_T|Z_T=z\right]\\
\leq&\frac{B^c}{K^\alpha}\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha/B^c}\left(\mathbb{E}\left[z\left(\rho+\epsilon-\hat{C}_{k,1}(x_{k,1},a_{k,1})\right)-\eta\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]+\eta\mathbb{E}\left[\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]\right)\\
&+2H^4\iota+4H^4\tilde{b}+\epsilon^2. \qquad (58)
\end{aligned}$$

Now we focus on the term inside the summation and obtain that

$$\begin{aligned}
&\left(\mathbb{E}\left[z\left(\rho+\epsilon-\hat{C}_{k,1}(x_{k,1},a_{k,1})\right)-\eta\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]+\eta\mathbb{E}\left[\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]\right)\\
\leq_{(a)}&z(\rho+\epsilon)-\mathbb{E}\left[\eta\left(\sum_a\left\{\frac{z}{\eta}\hat{C}_{k,1}q^\epsilon_{k,1}+\hat{Q}_{k,1}q^\epsilon_{k,1}\right\}(x_{k,1},a)\right)\middle|Z_T=z\right]+\eta\mathbb{E}\left[\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]\\
=&\mathbb{E}\left[z\left(\rho+\epsilon-\sum_a\hat{C}_{k,1}(x_{k,1},a)q^\epsilon_{k,1}(x_{k,1},a)\right)\middle|Z_T=z\right]\\
&-\mathbb{E}\left[\eta\sum_a\hat{Q}_{k,1}(x_{k,1},a)q^\epsilon_{k,1}(x_{k,1},a)-\eta\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]\\
=&\mathbb{E}\left[z\left(\rho+\epsilon-\sum_a C^\epsilon_{k,1}(x_{k,1},a)q^\epsilon_{k,1}(x_{k,1},a)\right)\middle|Z_T=z\right]\\
&-\mathbb{E}\left[\eta\sum_a\hat{Q}_{k,1}(x_{k,1},a)q^\epsilon_{k,1}(x_{k,1},a)-\eta\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]+\mathbb{E}\left[z\sum_a\left\{(C^\epsilon_{k,1}-\hat{C}_{k,1})q^\epsilon_{k,1}\right\}(x_{k,1},a)\middle|Z_T=z\right]\\
\leq&-\eta\mathbb{E}\left[\sum_a\hat{Q}_{k,1}(x_{k,1},a)q^\epsilon_{k,1}(x_{k,1},a)-\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]+\mathbb{E}\left[z\sum_a\left\{(C^\epsilon_{k,1}-\hat{C}_{k,1})q^\epsilon_{k,1}\right\}(x_{k,1},a)\middle|Z_T=z\right],
\end{aligned}$$

where inequality (a) holds because $a_{k,h}$ is chosen to maximize $\hat{Q}_{k,h}(x_{k,h},a)+\frac{Z_T}{\eta}\hat{C}_{k,h}(x_{k,h},a)$. and the last equality holds due to that $\{q^\epsilon_{k,h}(x,a)\}_{h=1}^H$ is a feasible solution to the optimization problem (18), so

$$\left(\rho+\epsilon-\sum_a C^\epsilon_{k,1}(x_{k,1},a)q^\epsilon_{k,1}(x_{k,1},a)\right)=\left(\rho+\epsilon-\sum_{h,x,a}g_{k,h}(x,a)q^\epsilon_{k,h}(x,a)\right)\leq 0.$$

Therefore, we can conclude the lemma by substituting $q^\epsilon_{k,h}(x,a)$ with the optimal solution $q^{\epsilon,*}_{k,h}(x,a)$. □

Lemma 11. *Assuming $\epsilon\leq\frac{\delta}{2}$, we have for any $1\leq T\leq K^{1-\alpha}\cdot B^c$*

$$\mathbb{E}[Z_T]\leq\frac{100(H^4\iota+\tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota}+\tilde{b}H^2)}{\delta}\right)+\frac{4H^2B^c}{K\delta}+\frac{4H^2B^c}{\eta\delta K^\alpha}+\frac{4\eta(\sqrt{H^2\iota}+2H^2\tilde{b})}{\delta}. \qquad (59)$$

The proof will also use the following lemma from (Neely, 2016).

Lemma 12. *Let S_t be the state of a Markov chain, L_t be a Lyapunov function with $L_0=l_0$, and its drift $\Delta_t=L_{t+1}-L_t$. Given the constant δ and v with $0<\delta\leq v$, suppose that the expected drift $\mathbb{E}[\Delta_t|S_t=s]$ satisfies the following conditions:*

(1) There exists constant $\gamma>0$ and $\theta_t>0$ such that $\mathbb{E}[\Delta_t|S_t=s]\leq-\gamma$ when $L_t\geq\theta_t$.

(2) $|L_{t+1} - L_t| \leq v$ *holds with probability one.*

Then we have

$$\mathbb{E}[e^{rL_t}] \leq e^{rl_0} + \frac{2e^{r(v+\theta_t)}}{r\gamma},$$

where $r = \frac{\gamma}{v^2+v\gamma/3}$. □

Proof of Lemma 11. We apply Lemma 12 to a new Lyapunov function:

$$\bar{L}_T = Z_T.$$

To verify condition (1) in Lemma 12, consider

$$\bar{L}_T = Z_T \geq \theta_T = \frac{4(\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K} + \eta(\sqrt{H^2\iota} + 2H^2\tilde{b}) + H^4\iota + \epsilon^2 + 2H^4\tilde{b}^2)}{\delta}$$

and $2\epsilon \leq \delta$. The conditional expected drift of

$$\begin{aligned}
&\mathbb{E}\left[Z_{T+1} - Z_T | Z_T = z\right] \\
=&\mathbb{E}\left[\sqrt{Z_{T+1}^2} - \sqrt{z^2}\middle| Z_T = z\right] \\
\leq& \frac{1}{2z}\mathbb{E}\left[Z_{T+1}^2 - z^2\middle| Z_T = z\right] \\
\leq_{(a)}& -\frac{\delta}{2} + \frac{4(\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K} + \eta(\sqrt{H^2\iota} + 2H^2\tilde{b}) + H^4\iota + \epsilon^2 + 2H^4\tilde{b}^2)}{z} \\
\leq& -\frac{\delta}{2} + \frac{4(\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K} + \eta(\sqrt{H^2\iota} + 2H^2\tilde{b}) + H^4\iota + \epsilon^2 + 2H^4\tilde{b}^2}{\theta_T} \\
=& -\frac{\delta}{4},
\end{aligned}$$

where inequality (a) is obtained according to Lemma 13; and the last inequality holds given $z \geq \theta_T$.

To verify condition (2) in Lemma 12, we have

$$Z_{T+1} - Z_T \leq |Z_{T+1} - Z_T| \leq \left|\rho + \epsilon - \bar{C}_T\right| \leq (H + H^2\sqrt{\iota} + 2\tilde{b}H^2) + \epsilon \leq 2(H^2\sqrt{\iota} + \tilde{b}H^2),$$

where the last inequality holds because $2\epsilon \leq \delta \leq 1$.

Now choose $\gamma = \frac{\delta}{4}$ and $v = 2(\sqrt{H^4\iota} + \tilde{b}H^2)$. From Lemma 12, we obtain

$$\mathbb{E}\left[e^{rZ_T}\right] \leq e^{rZ_1} + \frac{2e^{r(v+\theta_T)}}{r\gamma}, \quad \text{where} \quad r = \frac{\gamma}{v^2 + v\gamma/3}. \tag{60}$$

By Jensen's inequality, we have

$$e^{r\mathbb{E}[Z_T]} \leq \mathbb{E}\left[e^{rZ_T}\right],$$

which implies that

$$\begin{aligned}
\mathbb{E}[Z_T] \leq& \frac{1}{r}\log\left(1 + \frac{2e^{r(v+\theta_T)}}{r\gamma}\right) \\
=& \frac{1}{r}\log\left(1 + \frac{6v^2 + 2v\gamma}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq& \frac{1}{r}\log\left(1 + \frac{8v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq& \frac{1}{r}\log\left(\frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{4v^2}{3\gamma}\log\left(\frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)}\right)\\
&\leq \frac{3v^2}{\gamma}\log\left(\frac{2v}{\gamma}\right)+v+\theta_T\\
&\leq \frac{3v^2}{\gamma}\log\left(\frac{2v}{\gamma}\right)+v\\
&\quad+\frac{4(\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}+\eta(\sqrt{H^2\iota}+2H^2\tilde{b})+H^4\iota+\epsilon^2+2H^4\tilde{b}^2)}{\delta}\\
&=\frac{96(H^4\iota+\tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota}+\tilde{b}H^2)}{\delta}\right)+2(H^2\sqrt{\iota}+\tilde{b}H^2)\\
&\quad+\frac{4(\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}+\eta(\sqrt{H^2\iota}+2H^2\tilde{b})+H^4\iota+\epsilon^2+2H^4\tilde{b}^2)}{\delta}\\
&\leq \frac{100(H^4\iota+\tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota}+\tilde{b}H^2)}{\delta}\right)+\frac{4H^2B^c}{K\delta}+\frac{4H^2B^c}{\eta\delta K^{\alpha}}+\frac{4\eta(\sqrt{H^2\iota}+2H^2\tilde{b})}{\delta},
\end{aligned}
\tag{61}
$$

which completes the proof of Lemma 11. □

Lemma 13. *Given $\delta \geq 2\epsilon$, under our algorithmsl, the conditional expected drift is*

$$
\mathbb{E}\left[L_{T+1}-L_T|Z_T=z\right] \leq -\frac{\delta}{2}z+\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}+\eta(\sqrt{H^2\iota}+2H^2\tilde{b})+H^4\iota+\epsilon^2+2H^4\tilde{b}^2 \tag{62}
$$

Proof. Recall that $L_T=\frac{1}{2}Z_T^2$, and the virtual queue is updated by using

$$
Z_{T+1}=\left(Z_T+\rho+\epsilon-\frac{\bar{C}_TB^c}{K^{\alpha}}\right)^+.
$$

From inequality (58), we have

$$
\begin{aligned}
&\mathbb{E}\left[L_{T+1}-L_T|Z_T=z\right]\\
\leq&\frac{B^c}{K^{\alpha}}\sum_{k=(T-1)K^{\alpha}/B^c+1}^{TK^{\alpha}/B^c}\mathbb{E}\left[Z_T\left(\rho+\epsilon-C_{k,1}(x_{k,1},a_{k,1})\right)-\eta Q_{k,1}(x_{k,1},a_{k,1})\right.\\
&\left.+\eta Q_{k,1}(x_{k,1},a_{k,1})|Z_T=z\right]+H^4\iota+2H^4\tilde{b}^2+\epsilon^2\\
\leq_{(a)}&\frac{B^c}{K^{\alpha}}\sum_{k=(T-1)K^{\alpha}/B^c+1}^{TK^{\alpha}/B^c}\mathbb{E}\left[Z_T\left(\rho+\epsilon-\sum_a\left\{C_{k,1}q_{k,1}^{\pi}\right\}(x_{k,1},a)\right)\right.\\
&\left.-\eta\sum_a\{Q_{k,1}q_{k,1}^{\pi}\}(x_{k,1},a)+\eta Q_{k,1}(x_{k,1},a_{k,1})|Z_T=z\right]\\
&+\epsilon^2+H^4\iota+2H^4\tilde{b}^2\\
\leq&\frac{B^c}{K^{\alpha}}\sum_{k=(T-1)/B^cK^{\alpha}+1}^{TK^{\alpha}/B^c}\mathbb{E}\left[Z_T\left(\rho+\epsilon-\sum_a\left\{C_{k,1}^{\pi}q_{k,1}^{\pi}\right\}(x_{k,1},a)\right)\right.\\
&\left.-\eta\sum_a\{Q_{k,1}q_{k,1}^{\pi}\}(x_{k,1},a)+\eta Q_{k,1}(x_{k,1},a_{k,1})|Z_T=z\right]\\
&+\frac{B^c}{K^{\alpha}}\sum_{k=(T-1)K^{\alpha}/B^c+1}^{TK^{\alpha}/B^c}\mathbb{E}\left[Z_T\sum_a\left\{C_{k,1}^{\pi}q_{k,1}^{\pi}\right\}(x_{k,1},a)-Z_T\sum_a\left\{C_{k,1}q_{k,1}^{\pi}\right\}(x_{k,1},a)|Z_T=z\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{B^c}{K^\alpha}\sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\mathbb{E}\left[\eta\sum_a\left\{Q^\pi_{k,1}q^\pi_{k,1}\right\}(x_{k,1},a)-\eta\sum_a\left\{Q^\pi_{k,1}q^\pi_{k,1}\right\}(x_{k,1},a)|Z_T=z\right]+H^4\iota+\epsilon^2+2H^4\tilde{b}^2\\
\leq_{(b)}&-\frac{\delta}{2}z+\frac{B^c}{K^\alpha}\sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\mathbb{E}\left[\eta\sum_a\left\{(F^\pi_{k,1}-F_{k,1})q^\pi_{k,1}\right\}(x_{k,1},a)+\eta Q_{k,1}(x_{k,1},a_{k,1})\middle|Z_T=z\right]\\
&+H^4\iota+\epsilon^2+2H^4\tilde{b}^2\\
\leq_{(c)}&-\frac{\delta}{2}z+\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}+\eta(\sqrt{H^2\iota}+2H^2\tilde{b})+H^4\iota+\epsilon^2+2H^4\tilde{b}^2.
\end{aligned}
$$

Inequality (a) holds because of our algorithm. Inequality (b) holds because $\sum_a\left\{Q^\pi_{k,1}q^\pi_{k,1}\right\}(x_{k,1},a)$ is non-negative, and under Slater's condition, we can find policy π such that

$$
\epsilon+\rho-\mathbb{E}\left[\sum_a C^\pi_{k,1}(x_{k,1},a)q^\pi_{k,1}(x_{k,1},a)\right]=\rho+\epsilon-\mathbb{E}\left[\sum_{h,x,a}q^\pi_{k,h}(x,a)g_{k,h}(x,a)\right]\leq-\delta+\epsilon\leq-\frac{\delta}{2}.
$$

Finally, inequality (c) is obtained due to the fact that $Q_{k,1}(x_{k,1},a_{k,1})$ is bounded by using Lemma 4, and the fact that

$$
\mathbb{E}\left[\sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\sum_a\left\{(F^\pi_{k,1}-F_{k,1})q^\pi_{k,1}\right\}(x_{k,1},a)\middle|Z_T=z\right]
$$

can be bounded as (52) (note that the overestimation result and the concentration result in frame T hold regardless of the value of Z_T). □

D Proof of Lemma 2

Lemma 14. *Let*

$$
f_g(G_i)=\begin{cases}G_i/K^\lambda & \text{if } G_i<W\rho\\ G_i/K^\lambda & \text{if } G_i\geq W\rho\end{cases}\tag{63}
$$

$$
f_r(R_i)=\begin{cases}0 & \text{if } G_i<W\rho\\ R_i & \text{if } G_i\geq W\rho\end{cases}\tag{64}
$$

Let $R_i(B_i)(G_i(B_i))$ be the cumulative reward(utility) collected in epoch i by the given algorithm with the estimate value B_i chosen using Exp3 Algorithm. Let $\hat{B}$ be the optimal candidate from $\mathcal{J}$ that leads to the lowest regret while achieving zero constraint violation. Then we have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{K/W}(R_i(\hat{B})-R_i(B_i))\right]=&\tilde{\mathcal{O}}(H\sqrt{KW}+HK^{1-\lambda})\\
\mathbb{E}\left[\sum_{i=1}^{K/W}G_i(\hat{B})-G_i(B_i)\right]=&\tilde{\mathcal{O}}(HK^\lambda\sqrt{KW})
\end{aligned}
$$

Proof. Apply the regret bound of the Exp3 algorithm, we have

$$
\mathbb{E}\left[\sum_{i=1}^{K/W}(f_r(R_i(\hat{B}))+f_g(G_i(\hat{B}))-\sum_{i=1}^{K/W}(f_r(R_i(B_i))+f_g(G_i(B_i))\right]\tag{65}
$$

$$
\leq 2\sqrt{e-1}WH(1+1/K^\lambda)\sqrt{(K/W)(J+1)\ln(J+1)}=\tilde{\mathcal{O}}(H\sqrt{KW}),\tag{66}
$$

Recall that $\mathbb{E}[W\rho - G_i(\hat{B})] \leq 0$. Then it is easy to obtain

$$\mathbb{E}\left[\sum_{i=1}^{K/W}(R_i(\hat{B}) - R_i(B_i))\right] \leq \mathbb{E}\left[\sum_{i=1}^{K/W}(f_r(R_i(\hat{B})) - f_r(R_i(B_i)))\right] \tag{67}$$

$$\leq 2\sqrt{e-1}WH(1+1/K^{\lambda})\sqrt{(K/W)(J+1)\ln(J+1)} + \mathbb{E}\left[\sum_{i=1}^{K/W}(f_g(G_i(B_i)) - f_g(G_i(\hat{B})))\right] \tag{68}$$

$$\leq 2\sqrt{e-1}WH(1+1/K^{\lambda})\sqrt{(K/W)(J+1)\ln(J+1)} + \frac{WH}{K^{\lambda}}\cdot\frac{K}{W} \tag{69}$$

$$= \tilde{\mathcal{O}}(H\sqrt{KW} + HK^{1-\lambda}), \tag{70}$$

where the last inequality due to the fact that the term $\mathbb{E}\left[\sum_{i=1}^{K/W}(-f_g(G_i(\hat{B})))\right]$ is always non-positive. Furthermore, we have

$$\mathbb{E}\left[\sum_{i=1}^{K/W} G_i(\hat{B}) - G_i(B_i)\right] = K^{\lambda}\mathbb{E}\left[\sum_{i=1}^{K/W}\frac{G_i(\hat{B}) - G_i(B_i)}{K^{\lambda}}\right] \tag{71}$$

$$= K^{\lambda}\mathbb{E}\left[\sum_{i=1}^{K/W} f_g(G_i(\hat{B})) - f_g(G_i(B_i))\right] \tag{72}$$

$$\leq K^{\lambda}\left(2\sqrt{e-1}WH(1+1/K^{\lambda})\sqrt{(K/W)(J+1)\ln(J+1)} + \mathbb{E}\left[\sum_{i=1}^{K/W}(f_r(R_i(B_i)) - f_r(R_i(\hat{B})))\right]\right) \tag{73}$$

$$\leq K^{\lambda}\left(2\sqrt{e-1}WH(1+1/K^{\lambda})\sqrt{(K/W)(J+1)\ln(J+1)}\right) \tag{74}$$

$$= \tilde{\mathcal{O}}(HK^{\lambda}\sqrt{KW}), \tag{75}$$

where the last inequality is true because the second term is always non-positive. The reason is that when $\mathbb{E}[G_i(B_i)] \geq W\rho$, $\mathbb{E}[f_r(R_i(B_i))] \leq \mathbb{E}[f_r(R_i(\hat{B}))]$ because $\mathbb{E}[f_r(R_i(\hat{B}))] = \mathbb{E}[R_i(\hat{B})]$ is the largest return, and when $\mathbb{E}[G_i(B_i)] < W\rho$, we have $\mathbb{E}[f_r(R_i(B_i))] = 0$. □

E DETAILS PROOF OF THEOREM 1

E.1 Dynamic Regret

Recall that the regret can be decoupled as

$$\begin{aligned}
&\text{Regret}(K)\\
=&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_{a}\left\{Q_{k,1}^{*}q_{k,1}^{*} - Q_{k,1}^{\epsilon,*}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\right)\right] + &(76)\\
&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_{a}\left\{Q_{1}^{\epsilon,*}q_{1}^{\epsilon,*}\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right] + &(77)\\
&\mathbb{E}\left[\sum_{k=1}^{K}\left\{Q_{k,1} - Q_{k,1}^{\pi_k}\right\}(x_{k,1},a_{k,1})\right]. &(78)
\end{aligned}$$

Firstly, in lemma 6 we show that the first term can be bounded by comparing the original LP associated with the tightened LP such that

$$(76) \leq \frac{KH\epsilon}{\delta}. \tag{79}$$

By using Lemma 8, we can show that:

$$(78) \leq H^2SAK^{1-\alpha}B^c + \frac{2(H^3\sqrt{\iota}+2H^4\tilde{b})K}{\chi} + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)B^c} + 2\tilde{b}H^2K$$

For the last term 77, we first add and subtract additional terms to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a\left\{Q_{k,1}^{\epsilon,*}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right] \\
=&\mathbb{E}\left[\sum_k\sum_a\left(\left\{Q_{k,1}^{\epsilon,*}q_{k,1}^{\epsilon,*} + \frac{Z_k}{\eta}C_{k,1}^{\epsilon,*}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a) - \left\{Q_{k,1}q_{k,1}^{\epsilon,*} + \frac{Z_k}{\eta}C_{k,1}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\right)\right] \qquad (80)\\
&+\mathbb{E}\left[\sum_k\left(\sum_a\left\{Q_{k,1}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right] + \mathbb{E}\left[\sum_k\frac{Z_k}{\eta}\sum_a\left\{\left(C_{k,1}-C_{k,1}^{\epsilon,*}\right)q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\right]. \qquad (81)
\end{aligned}$$

We can see (80) is the difference of two combined Q functions. In Lemma 7 we show that $\left\{Q_{k,h}+\frac{Z_k}{\eta}C_{k,h}\right\}(x,a)$ is an overestimate of $\left\{Q_{k,h}^{\epsilon,*}+\frac{Z_k}{\eta}C_{k,h}^{\epsilon,*}\right\}(x,a)$ (i.e. (80) ≤ 0) with high probability. To bound (81), we use the Lyapunov-drift method and consider Lyapunov function $L_T = \frac{1}{2}Z_T^2$, where T is the frame index and Z_T is the value of the virtual queue at the beginning of the Tth frame. We show that in Lemma 10 that the Lyapunov-drift satisfies

$$\mathbb{E}[L_{T+1}-L_T] \leq \text{a negative drift} + 2H^4\iota + 4H^4\tilde{b}^2 + \epsilon^2 - \frac{\eta B^c}{K^\alpha}\sum_{k=TK^\alpha/B^c+1}^{(T+1)K^\alpha/B^c}\Phi_k, \qquad (82)$$

where

$$\Phi_k = \mathbb{E}\left[\left(\sum_a\left\{Q_{k,1}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right] + \mathbb{E}\left[\frac{Z_k}{\eta}\sum_a\left\{\left(C_{k,1}-C_{k,1}^{\epsilon,*}\right)q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\right],$$

So we can bound (81) by applying the telescoping sum over the $K^{1-\alpha}$ frames on the inequality above:

$$(81) = \sum_k\Phi_k \leq \frac{K^\alpha B^c\mathbb{E}\left[L_1 - L_{K^{1-\alpha}+1}\right]}{\eta} + \frac{K(2H^4\iota+4H^4\tilde{b}^2+\epsilon^2)}{\eta} \leq \frac{K(2H^4\iota+4H^4\tilde{b}^2+\epsilon^2)}{\eta}, \qquad (83)$$

where the last inequality holds because $L_1 = 0$ and $L_T \geq 0$ for all T. Now combining Lemma 7 and inequality (83), we conclude that

$$(77) \leq \frac{K(2H^4\iota+4H^4\tilde{b}^2+\epsilon^2)}{\eta} + \frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}.$$

Further combining inequality above we can obtain for $K \geq \left(\frac{16\sqrt{SAH^6\iota^3}B^{1/3}}{\delta}\right)^5$,

$$\begin{aligned}
\text{Regret}(K) \leq& \frac{KH\epsilon}{\delta} + H^2SAK^{1-\alpha}B^c + \frac{2(H^3\sqrt{\iota}+2H^4\tilde{b})K}{\chi} + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)B^c} + 2\tilde{b}H^2K \\
&+ \frac{K(2H^4\iota+4H^4\tilde{b}^2+\epsilon^2)}{\eta} + \frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}.
\end{aligned}$$

We conclude that under our choices of $\iota = 128\log(\sqrt{2SAH}K), \epsilon = \frac{8\sqrt{SAH^6\iota^3}B^{1/3}}{K^{0.2}}$ and $\alpha = 0.6, \eta = K^{\frac{1}{5}}B^{\frac{1}{3}}, \chi = K^{\frac{1}{5}}, c = \frac{2}{3}$, and $K^{1-\alpha}B^c\tilde{b} \leq B$,

$$\text{Regret}(K) = \tilde{\mathcal{O}}(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}B^{\frac{1}{3}}K^{\frac{4}{5}}).$$

E.2 Constraint Violation

Again, we use Z_T to denote the value of virtual-Queue in frame T. According to the virtual-Queue update, we have

$$Z_{T+1} = \left(Z_T + \rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}\right)^+ \geq Z_T + \rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha},$$

which implies that

$$\sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \left(-C_{k,1}^{\pi_k}(x_{k,1}, a_{k,1}) + \rho\right) \leq \frac{K^\alpha}{B^c}\left(Z_{T+1} - Z_T\right) + \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \left(\left\{C_{k,1} - C_{k,1}^{\pi_k}\right\}(x_{k,1}, a_{k,1}) - \epsilon\right).$$

Summing the inequality above over all frames and taking expectation on both sides, we obtain the following upper bound on the constraint violation:

$$\mathbb{E}\left[\sum_{k=1}^{K} \rho - C_{k,1}^{\pi_k}(x_{k,1}, a_{k,1})\right] \leq -K\epsilon + \frac{K^\alpha}{B^c}\mathbb{E}\left[Z_{K^{1-\alpha}B^c+1}\right] + \mathbb{E}\left[\sum_{k=1}^{K}\left\{C_{k,1} - C_{k,1}^{\pi_k}\right\}(x_{k,1}, a_{k,1})\right], \tag{84}$$

where we used the fact $Z_1 = 0$.

In Lemma 8, we established an upper bound on the estimation error of $C_{k,1}$:

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\left\{C_{k,1} - C_1^{\pi_k}\right\}(x_{k,1}, a_{k,1})\right]\\ \leq& H^2 SAK^{1-\alpha}B^c + \frac{2(H^3\sqrt{\iota} + 2H^4\tilde{b})K}{\chi} + \sqrt{H^4 SA\iota K^{2-\alpha}(\chi+1)B^c} + 2\tilde{b}H^2K.\end{aligned} \tag{85}$$

In Lemma 11, based on a Lyapunov drift analysis of this moment generating function and Jensen's inequality, we establish the following upper bound on Z_T that holds for any $1 \leq T \leq K^{1-\alpha}B^c + 1$

$$\mathbb{E}[Z_T] \leq \frac{100(H^4\iota + \tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota} + \tilde{b}H^2)}{\delta}\right) + \frac{4H^2B^c}{K\delta} + \frac{4H^2B^c}{\eta\delta K^\alpha} + \frac{4\eta(\sqrt{H^2\iota} + 2H^2\tilde{b})}{\delta}. \tag{86}$$

Substituting the results from Lemmas 8 and (86) into (84), under assumption $K \geq \left(\frac{16\sqrt{SAH^6\iota^3}B^{1/3}}{\delta}\right)^5$, which guarantees $\epsilon \leq \frac{\delta}{2}$. Then by using the choice that $\epsilon = \frac{8\sqrt{SAH^6\iota^3}B^{1/3}}{K^{0.2}}$, we can easily verify that

$$\text{Violation}(K) \leq \frac{100(H^4\iota + \tilde{b}^2H^2)K^{0.6}}{\delta B^{2/3}}\log\frac{16(H^2\sqrt{\iota} + \tilde{b}H^2)}{\delta} + \frac{4(H^2\sqrt{\iota} + 2H^2\tilde{b})}{\delta B^{1/3}}K^{0.8} - 5\sqrt{SAH^6\iota^3}K^{0.8}B^{\frac{1}{3}}.$$

If further we have $K \geq e^{\frac{1}{\delta}}$, we can obtain

$$\text{Violation}(K) \leq \frac{100(H^4\iota + \tilde{b}^2H^2)K^{0.6}}{\delta B^{2/3}}\log\frac{16(H^2\sqrt{\iota} + H^2\tilde{b})}{\delta} - \sqrt{SAH^6\iota^3}K^{0.8}B^{\frac{1}{3}} = 0.$$

F PROOF OF THEOREM 2

Let $\hat{B}$ be the optimal candidate value in $\mathcal{J}$ that leads to the lowest regret while achieving zero constraint violation. Let $R_i(B_i)$ be the expected cumulative reward received in epoch i with the estimated budget B_i. Then the regret can be decomposed into:

$$\begin{aligned}\text{Regret}(K) =& \mathbb{E}\left[\sum_{k=1}^{K}\left(V_{k,1}^{\pi_k^*}(x_{k,1}) - V_{k,1}^{\pi_k}(x_{k,1})\right)\right]\\ =& \mathbb{E}\left[\sum_{k=1}^{K} V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_{i=1}^{K/W} R_i(\hat{B})\right] + \mathbb{E}\left[\sum_{i=1}^{K/W} R_i(\hat{B}) - \sum_{i=1}^{K/W} R_i(B_i)\right].\end{aligned}$$

The first term is the regret of using the optimal candidate $\hat{B}$ from $\mathcal{J}$; the second term is the difference between using $\hat{B}$ and B_i which is selected by Exp3 algorithm. Applying the analysis of the Exp3 algorithm, we know that by using Lemma 2 for any choice of $\hat{B}$, the second term is upper bounded:

$$\mathbb{E}\left[\left(\sum_{i=1}^{K/W} R_i(\hat{B}) - \sum_{i=1}^{K/W} R_i(B_i)\right)\right] \leq \tilde{\mathcal{O}}(H\sqrt{KW} + HK^{1-\lambda}).$$

For the first term, according to the regret bound analysis of Algorithm 1, we have that

$$E\left[\sum_{k=1}^{K}\left(V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_{i=1}^{K/W} R_i(\hat{B})\right)\right] \leq \tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(\hat{B}\right)^{\frac{1}{3}}\right). \tag{87}$$

We need to consider whether B is covered in the range of $\mathcal{J}$ to further obtain the bound of (87). First we assume that $K = \Omega\left(\left(\frac{40\sqrt{SAH^6\iota^3}B^{1/3}}{\delta}\right)^9\right)$, which implies $B \leq \frac{K^{1/3}W}{\Delta^{3/2}W}$. Then we need to consider the following two cases:

- The first case is that B is covered in the range of $\mathcal{J}$. Note that two consecutive values in $\mathcal{J}$ only differ from each other by a factor of $W^{\frac{1}{J}}$, then there exists a value $\hat{B} \in \mathcal{J}$ such that $B \leq \hat{B} \leq W^{1/J}B$. Therefore we can bound the RHS of (87) by

$$\begin{aligned}\tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(\hat{B}\right)^{\frac{1}{3}}\right) \leq& \tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(BW^{1/J}\right)^{\frac{1}{3}}\right)\\ \leq& \tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} B^{\frac{1}{3}} K^{1-0.2\zeta}\right),\end{aligned}$$

  where the last step comes from the fact $W^{1/J} = W^{1/(\ln W+1)} \leq e$.

- The second case is that B is not covered in the range of $\mathcal{J}$, i.e., $B < \frac{K^{1/3}}{\Delta^{3/2}W}$. The optimal candidate in $\mathcal{J}$ is the smallest such that one $\hat{B} = \frac{K^{1/3}}{\Delta^{3/2}W}$, then we can bound the RHS of (87) by

$$\begin{aligned}\tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(\hat{B}\right)^{\frac{1}{3}}\right) \leq& \tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(\frac{K^{1/3}}{\Delta^{3/2}W}\right)^{\frac{1}{3}}\right)\\ \leq& \tilde{\mathcal{O}}\left(HK^{10/9-0.2\zeta}\frac{1}{K^{\zeta/3}}\right).\end{aligned}$$

For the constraint violation, according to Lemma 2 we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\rho - C_{k,1}^{\pi_k}(x_{k,1},a_{k,1})\right] = \mathbb{E}\left[\sum_{i=1}^{K/W}(W\rho - G_i(B_i))\right]\\ =&\mathbb{E}\left[\sum_{i=1}^{K/W}\left(W\rho - G_i(\hat{B})\right)\right] + \mathbb{E}\left[\sum_{i=1}^{K/W}\left(G_i(\hat{B}) - G_i(B_i)\right)\right]\end{aligned}$$

For the first term, according to Theorem 1, by selecting ϵ as $\epsilon = \frac{20\sqrt{SAH^6\iota^3}\hat{B}^{1/3}}{K^{0.2\zeta}}$. we have

$$\mathbb{E}\left[\sum_{i=1}^{K/W}\left(W\rho - G_i(\hat{B})\right)\right] \leq \frac{100(H^4\iota + \tilde{b}^2H^2)K^{0.6\zeta}}{\delta\hat{B}^{2/3}}\log\frac{16(H^2\sqrt{\iota}+H^2\tilde{b})}{\delta} - 13\sqrt{SAH^6\iota^3}K^{1-0.2\zeta}\hat{B}^{\frac{1}{3}}. \tag{88}$$

For the second term, we are able to obtain an upper bound by using Lemma 2

$$\mathbb{E}\left[\sum_{i=1}^{K/W}(G_i(\hat{B}) - G_i(B_i))\right] \leq 12K^{\lambda}H\sqrt{K^{1+\zeta}(J+1)\ln(J+1)} \tag{89}$$

By balancing the terms $\tilde{O}(K^{1-0.2\zeta})$, $\tilde{O}(K^{\lambda+(1+\zeta)/2})$ and $K^{1-\lambda}$, the best selection are $\zeta = 5/9$ and $\lambda = 1/9$. Therefore we further obtain when $K \geq e^{\frac{1}{\delta}}$,

$$\text{Violation}(K) \leq \frac{100(H^4\iota + \tilde{b}^2H^2)K^{1/3}}{\delta \hat{B}^{2/3}} \log \frac{16(H^2\sqrt{\iota} + H^2\tilde{b})}{\delta} - \sqrt{SAH^6\iota^3}K^{8/9}\hat{B}^{\frac{1}{3}} \leq 0. \tag{90}$$

We finish the proof of Theorem 2.

Algorithm 3: Model Free Primal-Dual Algorithm for Linear Function Approximation for Non-stationary Setting

Initialization: $Y_1 = 0$, $w_{j,h} = 0$, $\alpha = \frac{\log(|\mathcal{A}|)K}{2(1+\xi+H)}$, $\eta = \xi/\sqrt{KH^2}$, $\beta = dH\sqrt{\log(2\log|\mathcal{A}|dT/p)}$, $D = B^{-1/2}H^{-1/2}d^{1/2}K^{1/2}$.
for *frames* $\mathcal{E} = 1, \ldots, K/D$ **do**
  for *episodes* $k = 1, \ldots, D$ **do**
    Receive the initial state x_1^k.
    for *step* $h = H, H-1, \ldots, 1$ **do**
      $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^T + \lambda \mathbf{I}$;
      $w_{r,h}^k \leftarrow (\Lambda_h^k)^{-1}[\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + V_{r,h+1}^k(x_{h+1}^\tau)]]$;
      $w_{g,h}^k \leftarrow (\Lambda_h^k)^{-1}[\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)[g_h(x_h^\tau, a_h^\tau) + V_{g,h+1}^k(x_{h+1}^\tau)]]$;
      $Q_{r,h}^k(\cdot,\cdot) \leftarrow \min\{\langle w_{r,h}^k, \phi(\cdot,\cdot)\rangle + \beta(\phi(\cdot,\cdot)^T(\Lambda_h^k)^{-1}\phi(\cdot,\cdot))^{1/2}, H\}$;
      $Q_{g,h}^k(\cdot,\cdot) \leftarrow \min\{\langle w_{g,h}^k, \phi(\cdot,\cdot)\rangle + \beta(\phi(\cdot,\cdot)^T(\Lambda_h^k)^{-1}\phi(\cdot,\cdot))^{1/2}, H\}$;
      $\pi_{h,k}(a|\cdot) = \frac{\exp(\alpha(Q_{r,h}^k(\cdot,a) + Y_k Q_{g,h}^k(\cdot,a)))}{\sum_a \exp(\alpha(Q_{r,h}^k(\cdot,a) + Y_k Q_{g,h}^k(\cdot,a)))}$;
      $V_{r,h}^k(\cdot) = \sum_a \pi_{h,k}(a|\cdot)Q_{r,h}^k(\cdot,a)$;
      $V_{g,h}^k(\cdot) = \sum_a \pi_{h,k}(a|\cdot)Q_{g,h}^k(\cdot,a)$;
    for *step* $h = 1, \ldots, H$ **do**
      Compute $Q_{r,h}^k(x_h^k, a)$, $Q_{g,h}^k(x_h^k, a)$, $\pi(a|x_h^k)$ for all a ;
      Take action $a_h^k \sim \pi_{h,k}(\cdot|x_h^k)$ and observe x_{h+1}^k ;
    $Y_{k+1} = \max\{\min\{Y_k + \eta(\rho - V_{g,1}^k(x_1)), \xi\}, 0\}$

G DETAILS PROOF OF THEOREM 3

Notations: We describe the specific notations we have used in this section. With slight abuse of notations, in this section, we denote $V_{k,r,h}^\pi$ as the value function at step h for policy π at episode k. We denote $V_{k,g,h}^\pi$ as the utility value function at step h of episode k. We denote $Q_{k,j,h}^\pi$, $j = r, g$ as the state-action value function at step j for policy π.

Throughout this section, we denote $Q_{r,h}^k, Q_{g,h}^k, w_{r,h}^k, w_{g,h}^k, \Lambda_h^k$ as the Q-value and the parameter values estimated at the episode k. $V_{j,h}^k(\cdot) = \langle \pi_{h,k}(\cdot|\cdot), Q_{j,h}^k(\cdot,\cdot)\rangle_{\mathcal{A}}$. $\pi_{h,k}(\cdot|x)$ is the soft-max policy based on the composite Q-function at the k-th episode as $Q_{r,h}^k + Y_k Q_{g,h}^k$. To simplify the presentation, we denote $\phi_h^k = \phi(x_h^k, a_h^k)$.

G.1 Outline of Proof of Theorem 3

Step 1: The key to prove both the dynamic regret and violation is to show the following

Lemma 15. *For any* $Y \in [0, \xi]$,

$$\sum_{k=1}^K (V_{k,r,1}^{\pi_k^*}(x_1) - V_{k,r,1}^{\pi_k}(x_1)) + Y\sum_{k=1}^K (\rho - V_{k,g,1}^{\pi_k}(x_1)) \leq \frac{1}{2\eta}Y^2 + \frac{\eta}{2}H^2K +$$
$$\underbrace{\sum_{k=1}^K \left(V_{k,r,1}^{\pi_k^*}(x_1) + Y_k V_{k,g,1}^{\pi_k^*}(x_1)\right) - \left(V_{r,1}^k(x_1) + Y_k V_{g,1}^k(x_1)\right) +}_{\mathcal{T}_1}$$

$$\underbrace{\sum_{k=1}^{K}\Big(V_{r,1}^{k}(x_1)-V_{k,r,1}^{\pi_k}(x_1)\Big)+Y\sum_{k=1}^{K}\Big(V_{g,1}^{k}(x_1)-V_{k,g,1}^{\pi_k}(x_1)\Big)}_{\mathcal{T}_2} \tag{91}$$

Note that when $Y=0$, we recover the dynamic regret. The proof is in Appendix G.2.

Step-2: In order to bound $\mathcal{T}_1$, and $\mathcal{T}_2$, we use the following result

Lemma 16. *With probability* $1-2p$,

$$\begin{aligned}\mathcal{T}_1 &\le H^3(1+2/\delta)BD^{3/2}\sqrt{d}+\frac{KH\log(|\mathcal{A}|)}{\alpha}\\ \mathcal{T}_2 &\le (1+Y)(\mathcal{O}(\sqrt{H^4d^3K^2\iota^2/D})+\sqrt{d}D^{3/2}BH^2)\end{aligned} \tag{92}$$

The proof is in Appendix G.3.

Step-3: The final result is obtained by combining all the pieces.

Proof of Theorem 3:

Note from Lemma 15 we have

$$\sum_{k=1}^{K}(V_{k,r,1}^{\pi_k^*}(x_1)-V_{k,r,1}^{\pi_k}(x_1))+Y(\rho-V_{k,g,1}^{\pi_k}(x_1))\le\frac{Y^2}{2\eta}+\frac{\eta KH^2}{2}+\mathcal{T}_1+\mathcal{T}_2$$

From Lemma 16, we obtain

$$\begin{aligned}&\sum_{k=1}^{K}(V_{k,r,1}^{\pi_k^*}(x_1)-V_{k,r,1}^{\pi_k}(x_1))+Y(\rho-V_{k,g,1}^{\pi_k}(x_1))\le\frac{Y^2}{2\eta}+\frac{\eta KH^2}{2}+\\ &\frac{HK\log(|\mathcal{A}|)}{\alpha}+H^3(1+2/\delta)BD^{3/2}\sqrt{d}+(1+Y)(\mathcal{O}(\sqrt{H^4d^3K^2\iota^2/D})+\sqrt{d}D^{3/2}BH^2)\end{aligned} \tag{93}$$

Since $\eta=\dfrac{\xi}{\sqrt{KH^2}}$,$\alpha=\dfrac{\log(|\mathcal{A}|)K}{2(1+\xi+H)}$, $D=B^{-1/2}H^{-1/2}d^{1/2}K^{1/2}$, we obtain

$$\begin{aligned}&\sum_{k=1}^{K}(V_{k,r,1}^{\pi_k^*}(x_1)-V_{k,r,1}^{\pi_k}(x_1))+Y(b-V_{k,g,1}^{\pi_k}(x_1))\le\xi\sqrt{KH^2}\\ &+H2(1+\xi+H)+H^{9/4}(1+2/\delta)B^{1/4}K^{3/4}d^{5/4}+(Y+1)(\mathcal{O}(H^{9/4}d^{5/4}K^{3/4}B^{1/4}\iota^2)+H^{5/4}d^{5/4}K^{3/4})\end{aligned} \tag{94}$$

Since the above expression is true for any $Y\in[0,\xi]$, thus, plugging $Y=0$, we obtain

$$\text{Regret}(K)\le\mathcal{O}(H^{9/4}d^{5/4}K^{3/4}B^{1/4}\iota^2)+\mathcal{O}((1+1/\delta)H^{9/4}d^{5/4}K^{3/4}B^{1/4})$$

For the constraint violation bound, we use Lemma 27. Note that $\xi\ge 2\max_k\mu^{k,*}$. Thus, we replace $Y=\xi$ in (94). Thus, from (94) and Lemma 27, we obtain

$$\sum_{k=1}^{K}(\rho-V_{k,g,1}^{\pi}(x_1))\le\frac{2(1+\xi)}{\xi}(\mathcal{O}(H^{9/4}d^{5/4}K^{3/4}B^{1/4}\iota^2)+\mathcal{O}(H^{5/4}d^{5/4}K^{3/4}B^{1/4})) \tag{95}$$

Hence, the result follows. □

G.2 Proof of Lemma 15

We first state and prove the following result which is similar to the one proved in Ghosh et al. (2022).

Lemma 17. *For* $Y\in[0,\xi]$,

$$\sum_{k=1}^{K}(Y-Y_k)(\rho-V_{g,1}^{k}(x_1))\le\frac{Y^2}{2\eta}+\frac{\eta H^2K}{2} \tag{96}$$

Proof.

$$
\begin{aligned}
|Y_{k+1} - Y|^2 &= |Proj_{[0,\xi]}(Y_k + \eta(\rho - V^k_{g,1}(x_1))) - Proj_{[0,\xi]}(Y)|^2 \\
&\leq (Y_k + \eta(\rho - V^k_{g,1}(x_1))) - Y)^2 \\
&\leq (Y_k - Y)^2 + \eta^2 H^2 + 2\eta Y_k(\rho - V^k_{g,1}(x_1)) \qquad (97)
\end{aligned}
$$

Summing over k, we obtain

$$
\begin{aligned}
0 \leq |Y_{K+1} - Y|^2 &\leq |Y_1 - Y|^2 + 2\eta \sum_{k=1}^{K} (\rho - V^k_{g,1}(x_1))(Y_k - Y) + \eta^2 H^2 K \\
\sum_{k=1}^{K} (Y - Y_k)(\rho - V^k_{g,1}(x_1)) &\leq \frac{|Y_1 - Y|^2}{2\eta} + \frac{\eta H^2 K}{2} \qquad (98)
\end{aligned}
$$

Since $Y_1 = 0$, we have the result. □

Now, we prove Lemma 15.

Proof. Note that

$$
\begin{aligned}
Y \sum_{k=1}^{K} (\rho - V^{\pi_k}_{k,g,1}(x_1)) &= \sum_{k} (Y - Y_k)(\rho - V^k_{g,1}(x_1)) + Y_k(\rho - V^k_{g,1}) + Y(V^k_{g,1}(x_1) - V^{\pi_k}_{k,g,1}(x_1)) \\
&\leq \frac{1}{2\eta} Y^2 + \frac{\eta}{2} H^2 K + \sum_{k=1}^{K} (Y_k \rho - Y_k V^k_{g,1}(x_1)) + Y(V^k_{g,1}(x_1) - V^{\pi_k}_{g,1}(x_1)) \\
&\leq \frac{1}{2\eta} Y^2 + \frac{\eta}{2} H^2 K + \sum_{k=1}^{K} (Y_k V^{\pi^*_k}_{k,g,1}(x_1) - Y_k V^k_{g,1}(x_1)) + \sum_{k=1}^{K} Y(V^k_{g,1}(x_1) - V^{\pi_k}_{k,g,1}(x_1))
\end{aligned}
$$

where the first inequality follows from Lemma 17, and the second inequality follows from the fact that $V^{\pi^*_k}_{k,g,1}(x_1) \geq \rho$. Hence, the result simply follows from the above inequality. □

G.3 Proof of Lemma 16

We now move on to bound $\mathcal{T}_1$ and $\mathcal{T}_2$. First, we state and prove Lemmas 18, 19, 20, 21,22, and 23.

Lemma 18. *There exists a constant C_2 such that for any fixed $p \in (0,1)$, if we let E be the event that*

$$
\| \sum_{\tau=1}^{k-1} \phi^\tau_{j,h} [V^k_{j,h+1}(x^\tau_{h+1}) - \mathbb{P}_{k,h} V^k_{j,h+1}(x^\tau_h, a^\tau_h)] \|_{(\Lambda^k_h)^{-1}} \leq C_2 dH\sqrt{\chi} \qquad (99)
$$

for all $j \in \{r, g\}$, $\chi = \log[2(C_1+1)\log(|\mathcal{A}|)dT/p]$, for some constant C_2, then $\Pr(E) = 1 - 2p$.

This result is similar to the concentration lemma, which is crucial in controlling the fluctuations in least-squares value iteration as done in Jin et al. (2020). The proof relies on the uniform concentration lemma similar to Jin et al. (2020). However, there is an additional $\log(|\mathcal{A}|)$ in χ. This arises due to the fact that the policy (Algorithm 3) is soft-max unlike the greedy policy in Jin et al. (2020). Ghosh et al. (2022) shows that greedy policy is unable to prove the uniform concentration lemma. The proof is similar to Lemma 8 in Ghosh et al. (2022), thus, we remove it.

Now, we introduce some notations which we use throughout this paper.

For any $k \in \mathcal{E}$,i.e., any episode k within the frame $\mathcal{E}$, we define the variation as the following

$$
B^k_{j,\mathcal{E}} = \sum_{\tau=2}^{k} \sum_{h=1}^{H} ||\theta_{\tau,j,h} - \theta_{\tau-1,j,h}||, B^{\mathcal{E}}_j = \sum_{\tau=2}^{\mathcal{E}} \sum_{h=1}^{H} ||\theta_{\tau,j,h} - \theta_{\tau-1,j,h}||
$$

$$B^k_{p,\mathcal{E}} = \sum_{\tau=2}^{k}\sum_{h=1}^{H} ||\mu_{\tau,h} - \mu_{\tau-1,h}||, B^{\mathcal{E}}_p = \sum_{\tau=2}^{\mathcal{E}}\sum_{h=1}^{H} ||\mu_{\tau,h} - \mu_{\tau-1,h}||$$

These are local budget variation. Note that $|\mathcal{E}| = D$.

Now, we are bound the difference between our estimated $Q^k_{j,h}$ and $Q^\pi_{k,j,h}$. Using the Lemma 18, we show the following

Lemma 19. *There exists an absolute constant $\beta = C_1 dH\sqrt{\iota}$, $\iota = \log(\log(|\mathcal{A}|)2dT/p)$, and for any fixed policy π, on the event E defined in Lemma 18, we have*

$$\langle \phi(x,a), w^k_{j,h}\rangle - Q^\pi_{k,j,h}(x,a) = \mathbb{P}_{k,h}(V^k_{j,h+1} - V^\pi_{k,j,h+1})(x,a) + \Delta^k_h(x,a) + + B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD} \tag{100}$$

for some $\Delta^k_h(x,a)$ that satisfies $|\Delta^k_h(x,a)| \le \beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)}$, for any $k \in \mathcal{E}$.

Proof. We only prove for $j = r$, the proof for $j = g$ is similar.

Note that $Q^\pi_{k,r,h}(x,a) = \langle \phi(x,a), w^\pi_{r,h}\rangle = r_{k,h}(x,a) + \mathbb{P}_{k,h}V^\pi_{k,r,h+1}(x,a)$.

Hence, we have

$$\begin{aligned} w^k_{r,h} - w^\pi_{k,r,h} &= (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[r^\tau_h + V^k_{r,h+1}(x^\tau_{h+1})] - w^\pi_{k,r,h} \\ &= -\lambda(\Lambda^k_h)^{-1}(w^\pi_{k,r,h}) + (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[r_{\tau,h}(x^\tau_h,a^\tau_h) + V^k_{r,h+1} - r_{k,h}(x^\tau_h,a^\tau_h) - \mathbb{P}_{k,h}V^\pi_{k,r,h+1}] \end{aligned} \tag{101}$$

In the above expression, the second term of the right hand-side can be written as

$$\begin{aligned} &(\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[r_{\tau,h}(x^\tau_h,a^\tau_h) + V^k_{r,h+1} - r_{k,h}(x^\tau_h,a^\tau_h) - \mathbb{P}_{k,h}V^\pi_{k,r,h+1}] \\ &= (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[r_{\tau,h}(x^\tau_h,a^\tau_h) + V^k_{r,h+1} - r_{k,h}(x^\tau_h,a^\tau_h) - \mathbb{P}_{k,h}V^k_{r,h+1}] + (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[\mathbb{P}_{k,h}V^k_{r,h+1} - \mathbb{P}_{k,h}V^\pi_{k,r,h+1}] \\ &= (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[r_{\tau,h}(x^\tau_h,a^\tau_h) - r_{k,h}(x^\tau_h,a^\tau_h)] + (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}[V^k_{r,h+1} - \mathbb{P}_{\tau,h}V^k_{r,h+1}] \\ &+ (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}[\mathbb{P}_{\tau,h}V^k_{r,h+1} - \mathbb{P}_{k,h}V^k_{r,h+1}] + (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[\mathbb{P}_{k,h}V^k_{r,h+1} - \mathbb{P}_{k,h}V^\pi_{k,r,h+1}] \end{aligned} \tag{102}$$

By plugging in the above in (101) we obtain

$$\begin{aligned} &w^k_{r,h} - w^\pi_{k,r,h} \\ &= \underbrace{-\lambda(\Lambda^k_h)^{-1}(w^\pi_{k,r,h})}_{q_1} + \underbrace{(\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[r_{\tau,h}(x^\tau_h,a^\tau_h) - r_{k,h}(x^\tau_h,a^\tau_h)]}_{q_2} + \underbrace{(\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}[V^k_{r,h+1} - \mathbb{P}_{\tau,h}V^k_{r,h+1}]}_{q_3} \\ &+ \underbrace{(\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}[\mathbb{P}_{\tau,h}V^k_{r,h+1} - \mathbb{P}_{k,h}V^k_{r,h+1}]}_{q_4} + \underbrace{(\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[\mathbb{P}_{k,h}V^k_{r,h+1} - \mathbb{P}_{k,h}V^\pi_{k,r,h+1}]}_{q_5} \end{aligned} \tag{103}$$

For the first term,

$$|\langle \phi(x,a), q_1\rangle| \le \phi(x,a)^T(\Lambda^k_h)^{-1}\lambda w^\pi_{k,r,h} \le ||w^\pi_{k,r,h}||||\phi(x,a)||_{(\Lambda^k_h)^{-1}} \tag{104}$$

For the second term we have

$$\phi(x,a)^T(\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[r_{\tau,h}(x^\tau_h,a^\tau_h) - r_{k,h}(x^\tau_h,a^\tau_h)]$$

$$\leq \phi(x,a)^T(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau||\phi_h^\tau||||\theta_{\tau,r,h}-\theta_{k,r,h}||$$
$$\leq \phi(x,a)^T(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau||\phi_h^\tau||||\sum_{s=\tau}^{k-1}\theta_{s,r,h}-\theta_{s+1,r,h}||$$
$$\leq B_r^k\sqrt{dk}||\phi(x,a)||_{(\Lambda_h^k)^{-1}}$$

The last inequality follows from Lemma C.4 in Jin et al. (2020). Since $||\phi(x,a)||_{(\Lambda_h^k)^{-1}} \leq \sqrt{1/\lambda}$ and $D \geq k$. We have

$$|\langle\phi(x,a),q_2\rangle| \leq B_r^{\mathcal{E}}\sqrt{dD} \tag{105}$$

Similarly, we can bound

$$\phi(x,a)^T(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau[\mathbb{P}_{\tau,h}V_{r,h+1}^k-\mathbb{P}_{k,h}V_{r,h+1}^k] \leq HB_p^k\sqrt{dk}||\phi(x,a)||_{(\Lambda_h^k)^{-1}} \tag{106}$$

Again since $D \geq k$, and $||\phi(x,a)||_{(\Lambda_h^k)^{-1}} \leq \sqrt{1/\lambda}$, we have

$$|\langle\phi(x,a),q_3\rangle| \leq HB_p^{\mathcal{E}}\sqrt{dD} \tag{107}$$

From Lemma, the fourth term can be bounded as

$$|\langle\phi(x,a),q_4\rangle| \leq CdH\sqrt{\chi} \tag{108}$$

For the fifth term, note that

$$\langle\phi(x,a),q_5\rangle = \langle\phi(x,a),(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau[\mathbb{P}_h(V_{r,h+1}^k-V_{k,r,h+1}^\pi)(x_h^\tau,a_h^\tau)]\rangle$$
$$= \langle\phi(x,a),(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau(\phi_h^\tau)^T\int(V_{r,h+1}^k-V_{k,r,h+1}^\pi)(x')d\mu_{k,h}(x')\rangle$$
$$= \langle\phi(x,a),\int(V_{r,h+1}^k-V_{k,r,h+1}^\pi)(x')d\mu_{k,h}(x')\rangle - \langle\phi(x,a),\lambda(\Lambda_h^k)^{-1}\int(V_{r,h+1}^k-V_{r,h+1}^\pi)(x')d\mu_{k,h}(x')\rangle \tag{109}$$

The last term in (109) can be bounded as the following

$$|\langle\phi(x,a),\lambda(\Lambda_h^k)^{-1}\int(V_{r,h+1}^k-V_{k,r,h+1}^\pi)(x')d\mu_{k,h}(x')\rangle| \leq 2H\sqrt{d\lambda}\sqrt{\phi(x,a)^T(\Lambda_h^k)^{-1}\phi(x,a)} \tag{110}$$

since $||\int(V_{r,h+1}^k-V_{r,h+1}^\pi)(x')d\mu_{k,h}(x')||_2 \leq 2H\sqrt{d}$ as $||\mu_{k,h}(\mathcal{S})|| \leq \sqrt{d}$. The first term in (109) is equal to

$$\mathbb{P}_{k,h}(V_{r,h+1}^k-V_{r,h+1}^\pi)(x,a) \tag{111}$$

Note that $\langle\phi(x,a),w_{r,h}^k\rangle - Q_{k,r,h}^\pi(x,a) = \langle\phi(x,a),w_{r,h}^k-w_{k,r,h}^\pi\rangle = \langle\phi(x,a),q_1+q_2+q_3+q_4+q_5\rangle$, we have

$$\langle\phi(x,a),w_{j,h}^k\rangle - Q_{k,j,h}^\pi = \mathbb{P}_{k,h}(V_{j,h+1}^k-V_{k,j,h+1}^\pi)(x,a)+\Delta_h^k+B_r^{\mathcal{E}}\sqrt{dD}+HB_p^{\mathcal{E}}\sqrt{dW} \tag{112}$$

where $|\Delta_h^k| \leq \beta\sqrt{\phi(x,a)^T(\Lambda_h^k)^{-1}\phi(x,a)}$. □

Using Lemma 19, we also bound the difference between the combined Q-function (estimated) and the actual Q-function.

Lemma 20. *With probability* $1-2p$,

$$Q_{k,r,h}^\pi + Y_kQ_{k,g,h}^\pi \geq Q_{r,h}^k + Y_kQ_{g,h}^k + \mathbb{P}_{k,h}(V_{k,r,h+1}^\pi + Y_kV_{k,g,h+1}^\pi - V_{r,h+1}^k - Y_kV_{g,h+1}^k)$$
$$+ B_r^{\mathcal{E}}\sqrt{dD} + Y_kB_g^{\mathcal{E}}\sqrt{dD} + (1+Y_k)HB_p^{\mathcal{E}}\sqrt{dD} \tag{113}$$

Proof. From Lemma 19, we have

$$Q^{\pi}_{k,r,h} \leq \langle \phi(x,a), w^k_{r,h} \rangle + \mathbb{P}_{k,h}(V^{\pi}_{k,r,h+1} - V^k_{r,h}) + \beta ||\phi(x,a)||_{\Lambda^{-1}_{k,h}} + B^{\mathcal{E}}_r \sqrt{dD} + HB^{\mathcal{E}}_p \sqrt{dD} \tag{114}$$

From the definition of $Q^k_{j,h}$, we have

$$Q^{\pi}_{k,r,h} \leq \mathbb{P}_{k,h}(V^{\pi}_{k,r,h+1} - V^k_{r,h}) + Q^k_{r,h} + B^{\mathcal{E}}_r \sqrt{dD} + HB^{\mathcal{E}}_p \sqrt{dD} \tag{115}$$

Similarly,

$$Y_k Q^{\pi}_{k,g,h} \leq Y_k \mathbb{P}_{k,h}(V^{\pi}_{k,g,h+1} - V^k_{g,h}) + Y_k Q^k_{g,h} + Y_k B^{\mathcal{E}}_g \sqrt{dD} + Y_k HB^{\mathcal{E}}_p \sqrt{dD} \tag{116}$$

□

We now show that using the soft-max parameter α, one can bound the difference between the best estimated value function and the one achieved using the soft-max policy.

Lemma 21. *Then,* $\bar{V}^k_h(x) - V^k_h(x) \leq \dfrac{\log |\mathcal{A}|}{\alpha}$

where

Definition 2. $\bar{V}^k_h(\cdot) = \max_a [Q^k_{r,h}(\cdot, a) + Y_k Q^k_{g,h}(\cdot, a)]$.

$\bar{V}^k_h(\cdot)$ is the value function corresponds to the greedy-policy with respect to the composite Q-function.

Proof. Note that

$$V^k_h(x) = \sum_a \pi_{h,k}(a|x)[Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a)] \tag{117}$$

where

$$\pi_{h,k}(a|x) = \frac{\exp(\alpha[Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a)])}{\sum_a \exp(\alpha[Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a)])} \tag{118}$$

Denote $a_x = \arg\max_a [Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a)]$

Now, recall from Definition 2 that $\bar{V}^k_h(x) = [Q^k_{r,h}(x, a_x) + Y_k Q^k_{g,h}(x, a_x)]$. Then,

$$\begin{aligned}
&\bar{V}^k_h(x) - V^k_h(x) = [Q^k_{r,h}(x, a_x) + Y_k Q^k_{g,h}(x, a_x)] \\
&- \sum_a \pi_{h,k}(a|x)[Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a)] \\
&\leq \left(\frac{\log(\sum_a \exp(\alpha(Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a))))}{\alpha} \right) \\
&- \sum_a \pi_{h,k}(a|x)[Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a)] \\
&\leq \frac{\log(|\mathcal{A}|)}{\alpha}
\end{aligned} \tag{119}$$

where the last inequality follows from Proposition 1 in Pan et al. (2021). □

Using the above result, we bound the difference $\mathcal{T}_1$ (albeit for each episode).

Lemma 22. *With probability* $1 - 2p$,

$$(V^{\pi^*_k}_{k,r,1}(x_1) + Y_k V^{\pi^*_k}_{k,g,1}(x_1)) - (V^k_{r,1}(x_1) + Y_k V^k_{g,1}(x_1)) \leq \frac{H \log(|\mathcal{A}|)}{\alpha} + H(B^{\mathcal{E}}_r \sqrt{D} + Y_k B^{\mathcal{E}}_g \sqrt{D} + (1 + Y_k) HB^{\mathcal{E}}_p \sqrt{D})$$

Proof. First, we prove for the step H.

Note that $Q^k_{j,H+1} = 0 = Q^\pi_{j,H+1}$.

Under the event in E as described in Lemma 18 and from Lemma 19, we have for $j = r, g$,

$$|\langle \phi(x,a), w^k_{j,H}(x,a)\rangle - Q^\pi_{j,H}(x,a)| \le \beta\sqrt{\phi(x,a)^T(\Lambda^k_H)^{-1}\phi(x,a)} + B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD}$$

Hence, for any (x,a),

$$\begin{aligned} Q^\pi_{j,H}(x,a) &\le \min\{\langle \phi(x,a), w^k_{j,H}\rangle + \beta\sqrt{\phi(x,a)^T(\Lambda^k_H)^{-1}\phi(x,a)} + B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD}, H\} \\ &\le Q^k_{j,H}(x,a) + B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD} \end{aligned} \tag{120}$$

Hence, from the definition of $\bar{V}^k_h$,

$$\begin{aligned} \bar{V}^k_H(x) &= \max_a [Q^k_{r,H}(x,a) + Y_k Q^k_{g,h}(x,a)] \\ &\ge \sum_a \pi(a|x)[Q^\pi_{r,H}(x,a) + Y_k Q^\pi_{g,H}(x,a)] \\ &\quad - (B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD}) \\ &\ge V^{\pi,Y_k}_H(x) - (B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + H(1+Y_k)B^{\mathcal{E}}_p\sqrt{dD}) \end{aligned} \tag{121}$$

for any policy π. Thus, it also holds for π^*_k, the optimal policy. Hence, from Lemma 21, we have

$$V^{\pi^*_k,Y_k}_H(x) - V^k_H(x) \le \frac{\log(|\mathcal{A}|)}{\alpha} + (B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD})$$

Now, suppose that it is true till the step $h+1$ and consider the step h.

Since, it is true till step $h+1$, thus, for any policy π,

$$\begin{aligned} \mathbb{P}_{k,h}(V^{\pi,Y_k}_{h+1} - V^k_{h+1})(x,a) \le &\frac{(H-h)\log(|\mathcal{A}|)}{\alpha} \\ &+ (H-h)(B^{\mathcal{E}}_r\sqrt{dW} + Y_k B^{\mathcal{E}}_g\sqrt{dW} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dW}) \end{aligned} \tag{122}$$

From Lemma 19 we have for any (x,a)

$$\begin{aligned} &Q^\pi_{k,r,h}(x,a) + Y_k Q^\pi_{k,g,h}(x,a) \le Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a) + \frac{(H-h)\log(|\mathcal{A}|)}{\alpha} \\ &+ (H-h+1)(B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD}) \end{aligned} \tag{123}$$

Hence,

$$V^{\pi,Y_k}_h(x) \le \bar{V}^k_h(x) + \frac{(H-h)\log(|\mathcal{A}|)}{\alpha} + (H-h+1)(B^{\mathcal{E}}_r\sqrt{dW} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD})$$

Now, again from Lemma 21, we have $\bar{V}^k_h(x) - V^k_h(x) \le \frac{\log(|\mathcal{A}|)}{\alpha}$. Thus,

$$V^{\pi,Y_k}_h(x) - V^k_h(x) \le \frac{(H-h+1)\log(|\mathcal{A}|)}{\alpha} + (H-h+1)(B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD}) \tag{124}$$

Now, since it is true for any policy π, it will be true for π^*_k. From the definition of V^{π,Y_k}, we have

$$\begin{aligned} &\left(V^{\pi^*}_{r,h}(x) + Y_k V^{\pi^*}_{g,h}(x)\right) - \left(V^k_{r,h}(x) + Y_k V^k_{g,h}(x)\right) \le \frac{(H-h+1)\log(|\mathcal{A}|)}{\alpha} \\ &+ (H-h+1)(B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD}) \end{aligned} \tag{125}$$

Hence, the result follows by summing over K and considering $h = 1$. □

We now focus on bounding $\mathcal{T}_2$. First, we introduce some notations.

Let

$$D^k_{j,h,1} = \langle (Q^k_{j,h}(x^k_h,\cdot) - Q^{\pi_k}_{j,h}(x^k_h,\cdot)), \pi_{h,k}(\cdot|x^k_h)\rangle - (Q^k_{j,h}(x^k_h,a^k_h) - Q^{\pi_k}_{j,h}(x^k_h,a^k_h))$$
$$D^k_{j,h,2} = \mathbb{P}_{k,h}(V^k_{j,h+1} - V^{\pi_k}_{j,h+1})(x^k_h,a^k_h) - [V^k_{j,h+1} - V^{\pi_k}_{j,h+1}](x^k_{h+1}) \tag{126}$$

Lemma 23. *On the event defined in E in Lemma 18, we have*

$$V^k_{j,1}(x_1) - V^{\pi_k}_{k,j,1} \le \sum_{h=1}^H (D^k_{j,h,1} + D^k_{j,h,2}) + \sum_{h=1}^H 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} + H(B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD}) \tag{127}$$

Proof. By Lemma 19, for any x, h, a, k

$$\langle w^k_{j,h}(x,a), \phi(x,a)\rangle + \beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)} - Q^{\pi_k}_{j,h}$$
$$\le \mathbb{P}_{k,h}(V^k_{j,h+1} - V^{\pi_k}_{k,j,h+1})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)} + H(B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD})$$

Thus,

$$Q^k_{j,h}(x,a) - Q^{\pi_k}_{j,h}(x,a) \le \mathbb{P}_{k,h}(V^k_{j,h+1} - V^{\pi_k}_{k,j,h+1})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)}$$
$$+ H(B^{\mathcal{E}}_r\sqrt{dD} + B^{\mathcal{E}}_g\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD})$$
$$\mathbb{P}_{k,h}(V^k_{j,h+1} - V^{\pi_k}_{k,j,h+1})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)} +$$
$$B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD} - (Q^k_{j,h}(x,a) - Q^{\pi_k}_{k,j,h}(x,a)) \ge 0 \tag{128}$$

Since $V^k_{j,h}(x) = \sum_a \pi_{h,k}(a|x)Q^k_{j,h}(x,a)$ and $V^{\pi_k}_{k,j,h}(x) = \sum_a \pi_{h,k}(a|x)Q^{\pi_k}_{k,j,h}(x,a)$ where $\pi_{h,k}(a|\cdot) = \text{SOFT-MAX}^a_\alpha(Q^k_{r,h} + Y_k Q^k_{g,h})\ \forall a$.

Thus, from (128),

$$V^k_{j,h}(x^k_h) - V^{\pi_k}_{k,j,h}(x^k_h) = \sum_a \pi_{h,k}(a|x^k_h)[Q^k_{j,h}(x^k_h,a) - Q^{\pi_k}_{k,j,h}(x^k_h,a)]$$
$$\le \sum_a \pi_{h,k}(a|x^k_h)[Q^k_{j,h}(x^k_h,a) - Q^{\pi_k}_{k,j,h}(x^k_h,a)] + (B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD})$$
$$+ 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} + \mathbb{P}_{k,h}(V^k_{j,h+1} - V^{\pi_k}_{j,h+1})(x^k_h,a^k_h) - (Q^k_{j,h}(x^k_h,a^k_h) - Q^{\pi_k}_{k,j,h}(x^k_h,a^k_h)) \tag{129}$$

Thus, from (129), we have

$$V^k_{j,h}(x^k_h) - V^{\pi_k}_{j,h}(x^k_h) \le D^k_{j,h,1} + D^k_{j,h,2} + [V^k_{j,h+1} - V^{\pi_k}_{j,h+1}](x^k_{h+1}) + 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} + (B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD}) \tag{130}$$

Hence, by iterating recursively, we have

$$V^k_{j,1}(x_1) - V^{\pi_k}_{j,1} \le \sum_{h=1}^H (D^k_{j,h,1} + D^k_{j,h,2}) + \sum_{h=1}^H 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} + H(B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD}) \tag{131}$$

The result follows. □

Now, we are ready to prove Lemma 16.

Proof of Lemma 16

Proof. First, from Lemma 22,

$$(V_{k,r,1}^{\pi_k^*}(x_1) + Y_k V_{k,g,1}^{\pi_k^*}(x_1)) - (V_{r,1}^k(x_1) + Y_k V_{g,1}^k(x_1)) \leq \frac{H\log(|\mathcal{A}|)}{\alpha} + H(B_r^{\mathcal{E}}\sqrt{dD} + Y_k B_g^{\mathcal{E}}\sqrt{dD} + (1+Y_k)HB_p^{\mathcal{E}}\sqrt{dD}) \quad (132)$$

Note that $Y_k = 2H/\delta$. Now, summing over k within frame $\mathcal{E}$ we obtain

$$\sum_{k=1}^{D}(V_{k,r,1}^{\pi_k^*}(x_1) + Y_k V_{k,g,1}^{\pi_k^*}(x_1)) - (V_{r,1}^k(x_1) + Y_k V_{g,1}^k(x_1)) \leq \frac{HD\log(|\mathcal{A}|)}{\alpha} + H\sqrt{d}(B_r^{\mathcal{E}}D^{3/2} + 2H/\delta B_g^{\mathcal{E}}D^{3/2} + (1+2H/\delta)HB_p^{\mathcal{E}}D^{3/2}) \quad (133)$$

Now, summing over the epochs $\mathcal{E}$, we obtain

$$\begin{aligned}\sum_{\mathcal{E}=1}^{K/D}\sum_{k=1}^{D}(V_{k,r,1}^{\pi_k^*}(x_1) + Y_k V_{k,g,1}^{\pi_k^*}(x_1)) - (V_{r,1}^k(x_1) + Y_k V_{g,1}^k(x_1)) &\leq \frac{HK\log(|\mathcal{A}|)}{\alpha} \\ + \sum_{\mathcal{E}=1}^{K/D} H\sqrt{d}(B_r^{\mathcal{E}}D^{3/2} + 2H/\delta B_g^{\mathcal{E}}D^{3/2} + (1+2H/\delta)HB_p^{\mathcal{E}}D^{3/2}) &\\ \leq \frac{HK\log(|\mathcal{A}|)}{\alpha} + H^2(1+2H/\delta)\sqrt{d}BD^{3/2} &\end{aligned} \quad (134)$$

where we have used the fact that $\sum_{\mathcal{E}}(B_r^{\mathcal{E}} + B_g^{\mathcal{E}} + B_p^{\mathcal{E}}) = B_r + B_g + B_p = B$. This gives the bound for $\mathcal{T}_1$. Now, we bound $\mathcal{T}_2$.

From Lemma 23,

$$\sum_{k=1}^{D}(V_{j,1}^k(x_1) - V_{j,1}^{\pi_k}(x_1)) \leq \sum_{k=1}^{D}\sum_{h=1}^{H}(D_{j,h,1}^k + D_{j,h,2}^k) + \sum_{k=1}^{D}\sum_{h=1}^{H} 2\beta\sqrt{\phi(x_h^k,a_h^k)^T(\Lambda_h^k)^{-1}\phi(x_h^k,a_h^k)} + \sum_{\mathcal{E}=1}^{K/D}\sum_{k=1}^{D} H(B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD}) \quad (135)$$

We, now, bound the individual terms of the right-hand side in (135). First, we show that the first term corresponds to a Martingale difference.

For any $(k,h) \in [\mathcal{E}] \times [H]$, we define $\mathcal{F}_{h,1}^k$ as σ-algebra generated by the state-action sequences, reward, and constraint values, $\{(x_i^\tau, a_i^\tau)\}_{(\tau,i)\in[k-1]\times[H]} \cup \{(x_i^k, a_i^k)\}_{i\in[h]}$.

Similarly, we define the $\mathcal{F}_{h,2}^k$ as the σ-algebra generated by $\{(x_i^\tau, a_i^\tau)\}_{(\tau,i)\in[k-1]\times[H]} \cup \{(x_i^k, a_i^k)\}_{i\in[h]} \cup \{x_{h+1}^k\}$. x_{H+1}^k is a null state for any $k \in [K]$.

A filtration is a sequence of σ-algebras $\{\mathcal{F}_{h,m}^k\}_{(k,h,m)\in[\mathcal{E}]\times[H]\times[2]}$ in terms of time index

$$t(k,h,m) = 2(k-1)H + 2(h-1) + m \quad (136)$$

which holds that $\mathcal{F}_{h,m}^k \subset \mathcal{F}_{h',m'}^{k'}$ for any $t \leq t'$.

Note from the definitions in (126) that $D_{j,h,1}^k \in \mathcal{F}_{h,1}^k$ and $D_{j,h,2}^k \in \mathcal{F}_{h,2}^k$. Thus, for any $(k,h) \in [K] \times [H]$,

$$\mathbb{E}[D_{j,h,1}^k | \mathcal{F}_{h-1,2}^k] = 0, \quad \mathbb{E}[D_{j,h,2}^k | \mathcal{F}_{h,1}^k] = 0 \quad (137)$$

Notice that $t(k,0,2) = t(k-1,H,2) = 2(H-1)k$. Clearly, $\mathcal{F}_{0,2}^k = \mathcal{F}_{H,2}^{k-1}$ for any $k \geq 2$. Let $\mathcal{F}_{0,2}^1$ be empty. We define a Martingale sequence

$$M_{j,h,m}^k = \sum_{\tau=1}^{k-1}\sum_{i=1}^{H}(D_{j,i,1}^\tau + D_{j,i,2}^\tau) + \sum_{i=1}^{h-1}(D_{j,i,1}^k + D_{j,i,2}^k) + \sum_{l=1}^{m} D_{j,h,l}^k$$

$$= \sum_{(\tau,i,l)\in[\mathcal{E}]\times[H]\times[2],t(\tau,i,l)\le t(k,h,m)} D^{\tau}_{j,i,l} \tag{138}$$

where $t(k,h,m) = 2(k-1)H + 2(h-1) + m$ is the time index. Clearly, this martingale is adopted to the filtration $\{\mathcal{F}^k_{h,m}\}_{(k,h,m)\in[D]\times[H]\times[2]}$, and particularly

$$\sum_{k=1}^{D}\sum_{h=1}^{H}(D^k_{j,h,1} + D^k_{j,h,2}) = M^D_{j,H,2} \tag{139}$$

Thus, $M^K_{j,H,2}$ is a Martingale difference satisfying $|M^D_{j,H,2}| \le 4H$ since $|D^k_{j,h,1}|, |D^k_{j,h,2}| \le 2H$ From the Azuma-Hoeffding inequality, we have

$$\Pr(M^D_{j,H,2} > s) \le 2\exp(-\frac{s^2}{16DH^2}) \tag{140}$$

With probability $1 - p/2$ at least for any $j = r, g$,

$$\sum_k \sum_h M^D_{j,H,2} \le \sqrt{16DH^2\log(4/p)} \tag{141}$$

Now, we bound the second term of the right-hand side of (135). Note that the minimum eigen value of Λ^k_h is at least $\lambda = 1$ for all $(k,h) \in [D]\times[H]$. By Lemma 26,

$$\sum_{k=1}^{K}(\phi^k_h)^T(\Lambda^k_h)^{-1}\phi^k_h \le 2\log\left[\frac{\det(\Lambda^{k+1}_h)}{\det(\Lambda^1_h)}\right] \tag{142}$$

Moreover, note that $||\Lambda^{k+1}_h|| = ||\sum_{\tau=1}^{k}\phi^k_h(\phi^k_h)^T + \lambda\mathbf{I}|| \le \lambda + k$, hence,

$$\sum_{k=1}^{D}(\phi^k_h)^T(\Lambda^k_h)^{-1}\phi^k_h \le 2d\log\left[\frac{\lambda+k}{\lambda}\right] \le 2d\iota \tag{143}$$

Now, by Cauchy-Schwartz inequality, we have

$$\begin{aligned}\sum_{k=1}^{D}\sum_{h=1}^{H}\sqrt{(\phi^k_h)^T(\Lambda^k_h)^{-1}\phi^k_h} &\le \sum_{h=1}^{H}\sqrt{W}[\sum_{k=1}^{K}(\phi^k_h)^T(\Lambda^k_h)^{-1}\phi^k_h]^{1/2} \\ &\le H\sqrt{2dD\iota}\end{aligned} \tag{144}$$

Note that $\beta = C_1 dH\sqrt{\iota}$. Hence, the second term is bounded by

$$\mathcal{O}(\sqrt{H^4d^3D\iota^2}) \tag{145}$$

The third term of (135) is bounded by

$$\sum_{k=1}^{D} H(B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD}) = \sqrt{d}D^{3/2}H(B^{\mathcal{E}}_j + HB^{\mathcal{E}}_p) \tag{146}$$

Hence, summing (135) over the epochs we obtain

$$\sum_{\mathcal{E}=1}^{K/D}\sum_{k=1}^{D}(V^k_{j,1}(x_1) - V^{\pi_k}_{j,1}(x_1)) \le \sum_{\mathcal{E}=1}^{K/D}\mathcal{O}(\sqrt{H^4d^3D\iota^2}) + \sum_{\mathcal{E}=1}^{K/D}\sqrt{d}D^{3/2}H(B^{\mathcal{E}}_j + HB^{\mathcal{E}}_p) \tag{147}$$

Replacing $\sum_{\mathcal{E}} B^{\mathcal{E}}_j = B_j$, and $\sum_{\mathcal{E}} B^{\mathcal{E}}_p = B_p$, we obtain

$$\sum_{\mathcal{E}=1}^{K/D}\sum_{k=1}^{D}(V^k_{j,1}(x_1) - V^{\pi_k}_{k,j,1}(x_1)) \le \mathcal{O}(\sqrt{H^4d^3K^2\iota^2/D}) + \sqrt{d}D^{3/2}BH^2 \tag{148}$$

Thus,

$$\sum_{k=1}^{K}(V^k_{r,1}(x_1) - V^{\pi_k}_{k,r,1}(x_1)) + Y(V^k_{g,1}(x_1) - V^{\pi_k}_{k,g,1}(x_1)) \le (1+Y)(\mathcal{O}(\sqrt{H^4d^3K^2\iota^2/D}) + \sqrt{d}D^{3/2}BH^2) \tag{149}$$

Hence, the result follows. □

G.4 Supporting Results

Lemma 24. *Under Definition 1, for any fixed policy π, let $w^{\pi}_{k,j,h}$ be the corresponding weights such that $Q^{\pi}_{k,j,h} = \langle \phi(x,a), w^{\pi}_{k,j,h} \rangle$, for $j \in \{r, g\}$, then we have for all $h \in [H]$ and $k \in [K]$*

$$||w^{\pi}_{k,j,h}|| \leq 2H\sqrt{d} \tag{150}$$

Proof. From the linearity of the action-value function, we have

$$\begin{aligned} Q^{\pi}_{k,j,h}(x,a) &= j_{k,h}(x,a) + \mathbb{P}_{k,h} V^{\pi}_{k,j,h}(x,a) \\ &= \langle \phi(x,a), \theta_{j,h} \rangle + \int_{\mathcal{S}} V^{\pi}_{k,j,h+1}(x') \langle \phi(x,a), d\mu_{k,h}(x') \rangle \\ &= \langle \phi(x,a), w^{\pi}_{k,j,h} \rangle \end{aligned} \tag{151}$$

where $w^{\pi}_{j,h} = \theta_{j,h} + \int_{\mathcal{S}} V^{\pi}_{j,h+1}(x') d\mu_h(x')$.

Now, $||\theta_{j,h}|| \leq \sqrt{d}$, and $|| \int_{\mathcal{S}} V^{\pi}_{j,h+1}(x') d\mu_h(x')|| \leq H\sqrt{d}$. Thus, the result follows. □

Lemma 25. *For any (k,h), the weight $w^{k}_{j,h}$ satisfies*

$$||w^{k}_{j,h}|| \leq 2H\sqrt{dk/\lambda} \tag{152}$$

Proof. For any vector $v \in \mathcal{R}^d$ we have

$$|v^T w^{k}_{j,h}| = |v^T (\Lambda^{k}_{h})^{-1} \sum_{\tau=1}^{k-1} \phi^{\tau}_{h}(x^{\tau}_{h}, a^{\tau}_{h})(j_h(x^{\tau}_{h}, a^{\tau}_{h}) + \sum_{a} \pi_{h+1,k}(a|x^{\tau}_{h+1}) Q^{k}_{j,h+1}(x^{\tau}_{h+1}, a))| \tag{153}$$

here $\pi_{h,k}(\cdot|x)$ is the Soft-max policy.

Note that $Q^{k}_{j,h+1}(x,a) \leq H$ for any (x,a). Hence, from (153) we have

$$\begin{aligned} |v^T w^{k}_{j,h}| &\leq \sum_{\tau=1}^{k-1} |v^T (\Lambda^{k}_{h})^{-1} \phi^{\tau}_{h}| . 2H \\ &\leq \sqrt{\sum_{\tau=1}^{k-1} v^T (\Lambda^{h}_{k})^{-1} v} \sqrt{\sum_{\tau=1}^{k-1} \phi^{\tau}_{h} (\Lambda^{k}_{h})^{-1} \phi^{\tau}_{h}} . 2H \\ &\leq 2H||v|| \frac{\sqrt{dk}}{\sqrt{\lambda}} \end{aligned} \tag{154}$$

Note that $||w^{k}_{j,h}|| = \max_{v:||v||=1} |v^T w^{k}_{j,h}|$. Hence, the result follows. □

The following result is shown in Abbasi-yadkori et al. (2011) and in Lemma D.2 in Jin et al. (2020).

Lemma 26. *Let $\{\phi_t\}_{t \geq 0}$ be a sequence in $\Re^d$ satisfying $\sup_{t \geq 0} ||\phi_t|| \leq 1$. For any $t \geq 0$, we define $\Lambda_t = \Lambda_0 + \sum_{j=0}^{t} \phi_j \phi_j^T \phi_j$. Then if the smallest eigen value of Λ_0 be at least 1, we have*

$$\log \left[\frac{\det(\Lambda^{k+1}_{h})}{\det(\Lambda^{1}_{h})} \right] \leq \sum_{k=1}^{K} (\phi^{k}_{h})^T (\Lambda^{k}_{h})^{-1} \phi^{k}_{h} \leq 2 \log \left[\frac{\det(\Lambda^{k+1}_{h})}{\det(\Lambda^{1}_{h})} \right] \tag{155}$$

We use the following result (Lemma J.10 in Ding and Lavaei (2022)).

Lemma 27. *Let $\bar{C}^* \geq 2 \max_k \mu^{k,*}$, then, if*

$$\sum_{k=1}^{K} (V^{\pi^*_k}_{k,r,1}(x_1) - V^{\pi_k}_{k,r,1}(x_1)) + 2\bar{C}^* \sum_{k=1}^{K} (b_k - V^{\pi_k}_{k,g,1}(x_1)) \leq \delta \tag{156}$$

, then

$$\sum_{k=1}^{K}(b_k - V_{k,g,1}^{\pi_k}(x_1)) \leq \frac{2\delta}{\bar{C}^*} \tag{157}$$

Algorithm 4: Model Free Primal-Dual Algorithm for Linear Function Approximation for Non-stationary Setting without knowing the variation budget

Choose $W = K^{1/2}, \mathcal{J}$ (defined in Eq. (158)), $\gamma_0 = \min\left\{1, \sqrt{\frac{(K/W)\log(K/W)}{(e-1)KH}}\right\}, \lambda = 1/8$;
Initialize weights of the bandit arms $s_1(j) = 1, \forall j = 0, 1, \ldots, J$;
for *epoch* $i = 1, \ldots, \frac{K}{W}$ **do**
  Update $p_i(j) \leftarrow (1-\delta)\frac{s_i(j)}{\sum_{j'=0}^{J} s_i(j')} + \frac{\gamma_0}{J+1}, \forall j = 0, 1, \ldots, J$;
  Draw an arm $A_i \in [J]$ randomly according to the probabilities $p_i(0), \ldots, p_i(J)$;
  Set the estimated budget $B_i \leftarrow \frac{\sqrt{K}W^{\frac{A_i}{J}}}{\Delta W}$;
  Run a new instance of Algorithm 3 for W episodes with parameter value $B \leftarrow B_i$;
  Observe the cumulative reward R_i and utility G_i.;
  for *arm j=0,1,…,J* **do**
    $\hat{R}_i(j) = \begin{cases} (G_i/K^{\lambda})I_{\{j=A_i\}}/(WH(1+1/K^{\lambda})p_i(j)) & \text{if } G_i < W\rho \\ (R_i + G_i/K^{\lambda})I_{\{j=A_i\}}/(WH(1+1/K^{\lambda})p_i(j)) & \text{if } G_i \geq W\rho \end{cases}$; `// normalization`
    $s_{i+1} \leftarrow s_i(j)\exp(\gamma_0 \hat{R}_i(j)/(J+1))$;

H DETAILS PROOF OF THEOREM THEOREM 4

Let $W = K^{\zeta}$ and

$$\mathcal{J} = \left\{\frac{\sqrt{K}}{\Delta W}, \frac{\sqrt{K}W^{\frac{1}{J}}}{\Delta W}, \frac{\sqrt{K}W^{\frac{2}{J}}}{\Delta W}, \ldots, \frac{\sqrt{K}W}{\Delta W}\right\}, \Delta = \left(\frac{6(1+\xi)}{\xi\delta}\tilde{\mathcal{O}}((1+\delta)d^{5/4}H^{9/4})\right)^4 \tag{158}$$

where $J = \log W$ as the candidate sets for B in the linear CMDPs. Under assumption $K^{1/8} \geq \frac{6(1+\xi)}{\xi\delta}\tilde{O}((1 + 1/\delta)d^{5/4}B^{1/4}H^{9/4}$ we know the optimal budget $B \in \mathcal{J}$. Let $\hat{B}$ be any candidate value in $\mathcal{J}$ that leads to the lowest regret while achieving zero constraint violation. Let $R_i(B_i)$ be the expected cumulative reward received in epoch i with the estimated epoch length B. Then the regret can be decomposed into:

$$\begin{aligned}\text{Regret}(K) =& \mathbb{E}\left[\sum_{k=1}^{K}\left(V_{k,1}^{\pi_k^*}(x_{k,1}) - V_{k,1}^{\pi_k}(x_{k,1})\right)\right] \\ =& \mathbb{E}\left[\sum_{k=1}^{K} V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_{i=1}^{K/W} R_i(\hat{B})\right] + \mathbb{E}\left[\sum_{i=1}^{K/W} R_i(\hat{B}) - \sum_{i=1}^{K/W} R_i(B_i)\right].\end{aligned}$$

The first term is the regret of using the candidate $\hat{B}$ from $\mathcal{J}$; the second term is the difference between using $\hat{B}$ and B_i which is selected by Exp3 algorithm. Applying the analysis of the Exp3 algorithm, we know that by using Lemma 2 for any choice of $\hat{B}$, the second term is upper bounded:

$$\mathbb{E}\left[\left(\sum_{i=1}^{K/W} R_i(\hat{B}) - \sum_{i=1}^{K/W} R_i(B_i)\right)\right] \leq \tilde{\mathcal{O}}(H\sqrt{KW} + HK^{1-\lambda}).$$

For the first term, according to the regret bound analysis of Algorithm 3, we have for the W episodes

$$E\left[\sum_{k=1}^{W}\left(V_{k,1}^{\pi_k^*}(x_{k,1}) - R_i(\hat{D})\right)\right] \leq \tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}\hat{B}^{1/4}\right). \tag{159}$$

We need to consider whether $\hat{B}$ is covered in the range of $\mathcal{J}$ to further obtain the bound of (159). We consider the following two cases

- The first case is that optimal B is covered in the range of $\mathcal{J}$. Note that two consecutive values in $\mathcal{J}$ only differ from each other by a factor of $W^{\frac{1}{J}}$, then there exists a value $\hat{B} \in \mathcal{J}$ such that $B \leq \hat{B} \leq W^{1/J}B$. Therefore we can bound the RHS of (159) by

$$\begin{aligned}\tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}\hat{B}^{1/4}\right) \leq& \tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}W^{1/J}B^{1/4}\right)\\ \leq& \tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}eB^{1/4}\right)\\ =& \tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}B^{1/4}\right)\end{aligned}$$

- The second case is that B is not covered in the range of $\mathcal{J}$,i.e., $B \leq \frac{\sqrt{K}}{\Delta W}$, then the optimal candidate value in $\mathcal{J}$ is $\frac{\sqrt{K}}{\Delta W}$,we can bound the RHS of (159) by

$$\begin{aligned}&\tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}\hat{B}^{1/4}\right)\\ \leq& \tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}(\frac{\sqrt{K}}{\Delta W})^{1/4}\right)\end{aligned}$$

For the constraint violation, according to Lemma 2 we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\rho - C_{k,1}^{\pi_k}(x_{k,1},a_{k,1})\right] = \mathbb{E}\left[\sum_{i=1}^{K/W}(W\rho - G_i(B_i))\right]\\ =&\mathbb{E}\left[\sum_{i=1}^{K/W}\Big(W\rho - G_i(\hat{B})\Big)\right] + \mathbb{E}\left[\sum_{i=1}^{K/W}\Big(G_i(\hat{B}) - G_i(B_i)\Big)\right]\end{aligned}$$

For the first term, according to Theorem 3, by selecting $\epsilon = \dfrac{3(1+\xi)}{\xi}\tilde{O}((1+1/\delta)d^{5/4}\hat{B}^{1/4}H^{9/4}K^{1-\zeta/4})/K$, we have

$$\mathbb{E}\left[\sum_{i=1}^{K/W}\Big(W\rho - G_i(\hat{B})\Big)\right] \leq -\frac{(1+\xi)}{\xi}\tilde{\mathcal{O}}((1+1/\delta)K^{1-\zeta/4}H^{9/4}d^{5/4}\hat{B}^{1/4}). \tag{160}$$

For the second term, we are able to obtain an upper bound by using Lemma 2

$$\mathbb{E}\left[\sum_{i=1}^{K/W}(G_i(\hat{B}) - G_i(B_i))\right] \leq 12K^{\lambda}H\sqrt{K^{1+\zeta}(J+1)\ln(J+1)} \tag{161}$$

By balancing the terms $\tilde{O}(K^{1-\zeta/4})$, $\tilde{O}(K^{\lambda+(1+\zeta)/2})$ and $K^{1-\lambda}$, the best selection are $\zeta = 1/2$ and $\lambda = 1/8$. Therefore we further obtain

$$\text{Violation}(K) = 0. \tag{162}$$

We finish the proof of Theorem 4.

I ANOTHER APPROACH FOR UNKNOWN BUDGET

We consider a primal-dual adaptation in the outer loop as well. In particular, after collecting $R_i(B_i)$ and $G_i(B_i)$ under the selected epoch length B_i, the bandit reward is $R_i(B_i)+Y_iG_i(B_i)$, where $Y_i = \min\{\max\{Y_{i-1}+\eta(\rho-G_i(B_i)/W),0\},\xi\}$. Then line 10 in Algorithm 4 is replaced with

$$\hat{R}_i(j) = (R_i(B_i) + Y_iG_i(B_i))/(WH + \xi WH)$$

Let $W = d^{1/2}H^{-1/2}K^{1/2}$ be the epoch length, and

$$\mathcal{J} = \left\{1, W^{\frac{1}{J}}, \dots, W\right\},$$

where $J = \log W$ as the candidate sets for D in the linear CMDPs. We still use Exp-3 to choose an arm. From the Exp-3 analysis we know for any $D^\dagger$

$$\begin{aligned}&\sum_m (R_m(D^\dagger) + Y_m G_m(D^\dagger)) - (R_m(D_m) + Y_m G_m(D_m))\\ \leq& 2\sqrt{e-1}WH(1+\xi)\sqrt{(K/W)(J+1)\ln(J+1)} = \tilde{\mathcal{O}}(H\xi\sqrt{KW}),\end{aligned} \tag{163}$$

Now, from the dual domain analysis, we obtain a similar to (Lemma 15)

$$\sum_m (Y - Y_m)(W\rho - G_m(D_m)) \leq \frac{Y^2 W}{2\eta} + \frac{\eta H^2 K}{2} \tag{164}$$

We note that $\eta = \sqrt{\xi^2 W/(KH^2)}$, then the upper bound is $\xi\sqrt{WKH^2}$. From the results analysis of the constraint violation from Theorem 3, we have for the optimal choice of $D^\dagger$ from $\mathcal{J}$

$$\sum_m (W\rho - G_m(D^\dagger)) \leq \tilde{\mathcal{O}}(K\sqrt{d^3H^4/D^\dagger} + D^\dagger\sqrt{dD^\dagger}H^2B). \tag{165}$$

$$\sum_k^K V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_m R_i(D^\dagger) \leq \tilde{\mathcal{O}}(K\sqrt{d^3H^4/D^\dagger} + D^\dagger\sqrt{dD^\dagger}H^2B). \tag{166}$$

Hence, we have

$$\begin{aligned}&\sum_m -Y_m(G_m(D^\dagger) - G_m(D_m)) = \sum_m -Y_m(G_m(D^\dagger) - W\rho) + \sum_m -Y_m(W\rho - G_m(D_m))\\ \leq& \tilde{\mathcal{O}}\left(K\sqrt{d^3H^4/D^\dagger}\xi + D^\dagger\sqrt{dD^\dagger}H^2B\xi + \xi\sqrt{WKH^2}\right)\end{aligned} \tag{167}$$

where we use (164) (with $Y = 0$) for the first inequality, and (165) (where we use $|Y_m| \leq \xi$) for the second term.

Hence, from (163)

$$\begin{aligned}&\sum_m (R_m(D^\dagger) - R_m(D_m))\\ \leq& \tilde{\mathcal{O}}(H\xi 2\sqrt{e-1}WH(1+\xi)\sqrt{(K/W)(J+1)\ln(J+1)} + \sum_m -Y_m(G_m(D^\dagger) - G_m(D_m))\\ \leq& \tilde{\mathcal{O}}\left(K\sqrt{d^3H^4/D^\dagger}\xi + D^\dagger\sqrt{dD^\dagger}H^2B\xi + \xi\sqrt{WKH^2} + H\xi\sqrt{KW}\right)\end{aligned} \tag{168}$$

Now, suppose that optimal D exists in the range, thus, $D^\dagger \leq D \leq D^\dagger W^{1/J} = eD^\dagger$. Hence, from $D = B^{-1/2}W$, and (166) we have the regret bound of $\tilde{\mathcal{O}}((1+1/\delta)H^{9/4}d^{5/4}B^{1/4}K^{3/4})$.

If D is not covered – if $D < 1$, then $B^{-1/2}d^{1/2}H^{-1/2}K^{1/2} \leq 1$, thus, $B \geq \mathcal{O}(K)$ which will make the regret and violation bound vacuous. Thus, we consider $D > W$. Hence, $B^{-1/2}d^{1/2}H^{-1/2}K^{1/2} > d^{1/2}H^{-1/2}K^{1/2}$, thus, we have $B < 1$. Hence, the optimal $D^\dagger = d^{1/2}H^{-1/2}K^{1/2}$ by balancing the terms in (168). Thus, the regret bound again follows, i.e., the regret bound is $\tilde{\mathcal{O}}((1+1/\delta)H^{9/4}d^{5/4}B^{1/4}K^{3/4})$.

Now, we bound the constraint violation. Note that

$$\begin{aligned}&\sum_k^K V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_m R_m(D_m) + Y\sum_m (W\rho - G_m(D_m))\\ =& \sum_k^K V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_m R_m(D^\dagger) + \sum_m Y_m(W\rho - G_m(D^\dagger))\end{aligned}$$

$$+\sum_m Y_m(G_m(D^\dagger)-G_m(D_m))+\sum_m(R_m(D^\dagger)-R_m(D_m))+\sum_m(Y-Y_m)(W\rho-G_m(D_m))$$
$$\leq\tilde{\mathcal{O}}\left(K\sqrt{d^3H^4/D^\dagger}\xi+D^\dagger\sqrt{dD^\dagger}H^2B\xi+\xi\sqrt{WKH^2}+H\xi\sqrt{KW}\right) \tag{169}$$

where we use (166), (165), (163), and (164) to bound each term in the right-hand side respectively.

By using lemma 27, we can have

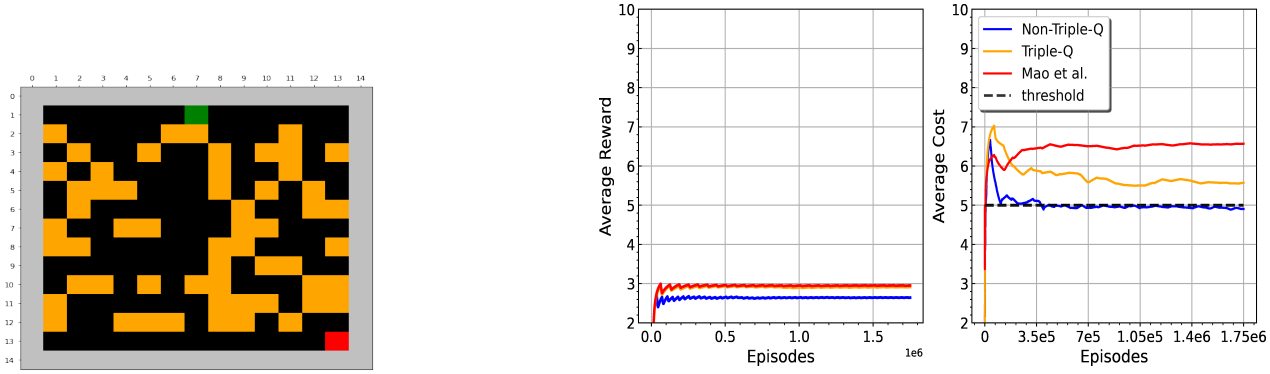
$$\sum_m W\rho-G_m(D_m)\leq\tilde{\mathcal{O}}\left(K\sqrt{d^3H^4/D^\dagger}+D^\dagger\sqrt{dD^\dagger}H^2B+\sqrt{WKH^2}+H\sqrt{KW}\right.$$
$$\left.+\frac{1}{\xi}(K\sqrt{d^3H^4/D^\dagger})+D^\dagger\sqrt{dD^\dagger}H^2B)\right) \tag{170}$$

From a similar argument (for regret) where optimal D is covered within the range or not, we bound $D^\dagger$ and obtain the result for constraint violation. We prove the results by substituting $\xi=\frac{2H}{\gamma}$.

J Simulation

We compare Algorithm 1 with two baseline algorithms: an algorithm (Mao et al., 2020) for non-stationary MDPs, and an algorithm (Wei et al., 2022b) for stationary constrained MDPs using a grid-world environment, which is shown in Figure. 1a. The objective of the agent is to travel to the destination as quickly as possible while avoiding obstacles for safety. Hitting an obstacle incurs a cost of 1. The reward for the destination is 1. Denote the Euclidean distance from the current location x to the destination as $d_0(x)$, the longest Euclidean distance is denoted by $d_{\max}$, then the reward function for a locations x is defined as $\frac{0.1*(d_{\max}-d_0)}{d_{\max}}$. The cost constraint is set to be 5 (we used cost instead of utility in this simulation), which means the agent is only allowed to hit the obstacles at most five times. To account for the statistical significance, all results were averaged over 10 trials. To test the algorithms in a non-stationary environment, we gradually vary the transition probability, reward, and cost functions. In particular, the reward is added an additional variation of $\pm\frac{0.1}{K}$, where the sign is uniformly sampled, the cost varies $\frac{0.1}{K}$ at all the locations. We vary the transitions in a way that the intended transition "succeeds" with probability 0.95 at the beginning; that is, even if the agent takes the correct action at a certain step, there is still a 0.05 probability that it will take an action randomly. The probability is increased with $\frac{0.1}{K}$ at each iteration.

As shown in Figure. 1b, we can observe that our Algorithm 1 can quickly learn a well-performed policy while satisfying the safety constraint (below the threshold), while other methods all fail to satisfy the constraint.



(a) Grid World

(b) Average Reward and Cost during training

Figure 1: Performance of the three algorithms under a non-stationary environment