Provably Efficient Algorithms for Safe Reinforcement Learning

by

Honghao Wei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in The University of Michigan
2023

Doctoral Committee:

Professor Lei Ying, Chair
Associate Professor Achilleas Anastasopoulos
Associate Professor Hun-Seok Kim
Professor Henry Liu

Honghao Wei

honghaow@umich.edu

ORCID iD: 0000-0002-1131-326X

To my dear family.

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude and appreciation to my advisor, Prof. Lei Ying. His unwavering support, insightful guidance, and constant encouragement have been instrumental throughout my Ph.D. journey. His enthusiasm, knowledge, and rigorous attitude toward research have had a profound impact on me, spurring me toward excellence. He has always been a great mentor, dedicating immeasurable time and exceptional effort to guide my doctoral study. I am very fortunate to have him as my advisor, who made my Ph.D. journey rewarding.

I am immensely grateful to the members of my dissertation committee, Prof. Achilleas Anastasopoulos, Prof. Hun-Seok Kim, and Prof. Henry Liu, for their insightful comments, constructive suggestions, and continuous encouragement. I truly appreciate the opportunity to interact with and learn from them. I also want to thank Prof. Ness B. Shroff, Prof. Weina Wang, Prof. Xin Liu, Prof. Xingyu Zhou, and Dr. Arnob Ghosh for their helpful advice and inspiring discussions.

The past few years would not have been as fulfilling without the remarkable collaboration and support from my colleagues and friends. Thank you all for the understanding and joy brought to me along this long journey. A special thanks to my cat, Mochi, for the joy and comfort he brought to me during the COVID-19 pandemic.

Finally, I would like to express my deepest appreciation to my family for their endless love and support. I want to express my deepest gratitude to my wife, Lijian, for her always love, support, understanding, and scarification, without whom I would never have come this far.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Safe reinforcement learning (RL) is an area of research focused on developing algorithms and methods that ensure the safety of RL agents during learning and decision-making processes. The goal is to enable RL agents to interact with their environments and learn optimal decisions while avoiding actions that can lead to harmful or undesirable outcomes. This dissertation provides a comprehensive study of *model-free*, *simulator-free* reinforcement learning algorithms for Constrained Markov Decision Processes (CMDPs) with sublinear regret and zero constraint violation, with the focus on three settings: (1) episodic CMDPs; (2) infinite-horizon average-reward CMDPs and (3) non-stationary episodic CMDPs.

The first part provides the first model-free, simulator-free safe-RL algorithm with sublinear regret and zero constraint violation. The algorithm is named Triple-Q because it includes three key components: a Q-function (also called action value function) for the cumulative reward, a Q-function for the cumulative utility of the constraint, and a virtual Queue that (over)-estimates the cumulative constraint violation. Under Triple-Q, at each step, an action is chosen based on the pseudo-Q-value that is a combination of the three "Q" values. The algorithm updates the reward and utility Qvalues with learning rates that depend on the visit counts to the corresponding (state, action) pairs and are periodically reset. In the episodic CMDP setting, Triple-Q achieves $\tilde{\mathcal{O}}\left(\frac{1}{\delta}H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{\frac{4}{5}}\right)$ regret when K is large enrough, where K is the total number of episodes, H is the number of steps in each episode, S is the number of states, A is the number of actions, and δ is Slater's constant. Furthermore, Triple-Q guarantees

zero constraint violation, both on expectation and with a high probability, when K is sufficiently large. Finally, the computational complexity of Triple-Q is similar to SARSA for unconstrained MDPs, and is computationally efficient. In Chapter III, the results are extended to infinite-horizon average-reward Constrained Markov Decision Processes (CMDPs). The proposed algorithm guarantees $\tilde{O}\left(\frac{\sqrt{SA}\kappa}{\delta}K^{\frac{5}{6}}\right)$ regret and zero constraint violation, where κ and δ are two constants independent of the learning horizon K.

Then in Chapter IV the dissertation studies safe-RL in a more challenging setting, non-stationary CMDPs, where the rewards/utilities and dynamics are time-varying and likely unknown a priori. In the nonstationary 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 to both tabular and linear approximation settings.

CHAPTER I

Introduction

Reinforcement learning (RL), with its success in gaming [1; 2] and robotics [3], has been widely viewed as one of the most important technologies for next-generation, AI-driven complex systems such as autonomous driving, digital healthcare, and smart cities. However, despite the significant advances (such as deep RL) over the last few decades, a major obstacle in applying RL in practice is the lack of "safety" guarantees. Here "safety" refers to a broad range of operational constraints. The objective of a traditional RL problem is to maximize the expected cumulative reward, but in practice, many applications need to be operated under a variety of constraints, such as collision avoidance in robotics and autonomous driving [4; 5; 6], legal and business restrictions in financial engineering [7], and resource and budget constraints in healthcare systems [8]. These applications with operational constraints can often be modeled as Constrained Markov Decision Processes (CMDPs), in which the agent's goal is to learn a policy that maximizes the expected cumulative reward subject to the constraints. CMDPs have had an important impact on many other real-world applications: [9] has used CMDPs to solve a hospital admission scheduling problem, [10] developed a pavement management system for the state of Arizona to produce optimal maintenance policies for a 7400−mile network of highways, which saved 14 million dollars in the first year. A standard formulation for RL with constraints is the Constrained Markov Decision

Processes framework [11], in which the agent aims at learning a policy that maximizes the expected cumulative reward under safety constraints during and after learning.

Model-based Solutions for CMDPS: On model-based approaches, the model is assumed to be known or can be predicted. This model captures the agent's understanding of how the environment behaves and predicts the consequences of different actions. The first model-based approach, based on Linear Programming (LP), was first introduced by [12]. The optimal policy can be induced from the decision variables which correspond to the occupancy measure, and the objective of the LP is equivalent to the optimal value of the CMDP. [13] proposed an LP-based algorithm that learns the optimal policy while satisfying the constraints for a CMDP with a known model. [14; 15; 16; 17; 18] follow a similar approach but learn the models from the data samples collected. This approach has also been utilized for CMDPs with linear function approximation [19] under the assumption that the transition kernel is linear. Leveraging the estimated model, the CMDPs can be solved approximately as long as the estimate becomes more and more accurate. The works mentioned above are proven to achieve sublinear constraint violation.

Another approach is to learn the model and find the solution using primal-dual methods, [18; 17; 20] adopt the principle of optimism in the face of uncertainty to design a primal-dual approach which achieves a sublinear regret and constraint violation. [19] extends the studies to constrained episodic MDPs with a linear structure via a primal-dual type policy optimization algorithm.

While model-based RL algorithms are sample efficient, they need to solve LPs when the estimated models are updated continuously, so these algorithms are often computationally inefficient and require a large memory to maintain a large number of model parameters.

Model-free Solutions for CMDPs: Model-free algorithms, on the other hand, learn state or action value functions, instead of transition kernels, so they require

Table 1.1: Existing Approaches in CMDPs.

	Methodology	**Paper**	**Setting**
Model-based	LP-based	[12; 15; 16] [17; 18; 20; 31]	Episodic CMDPs
	LP-based	[13; 14; 32]	Infinite-Horizon Average CMDPs
	LP-based Primial-Dual	[17; 18; 20]	Episodic CMDPs
	Primial-Dual	[19]	Episodic Linear Kernel CMDPs
Model-free	Primal-Dual	[29; 33]	Episodic CMDPs
	Primal-Dual	[30]	Infinite-Horizon Average CMDPs

significantly less memory space and have lower computational complexity. In [21], the author proposes an actor-critic RL algorithm and shows its asymptotic global convergence using multi-timescale stochastic approximation theory for infinite-horizon average-reward CMDPs when the model is unknown. Policy gradient approaches have also been developed [22; 23; 24] and seen successes in practice for solving constrained RL problems, though they lack regret and constraint violation analysis. [25; 26; 27] show that sublinear regrets and constraint violations are achievable when policy "simulators" (or generative models) are given. Some very recent works [20; 28] show that sublinear regret bound and zero violation are possible for episodic CMDPs without simulators. In particular, [20] proposes a model-based algorithm, and [29] presents a model-free algorithm for episodic CMDPs. [30] extends the model-free approach to infinite-horizon average reward CMDPs. For safety concerns, [31] proves that it is possible to achieve zero violation during training given a safe baseline policy based on a model-based approach.

This dissertation investigates designing computationally efficient, model-free RL algorithms with provable regret and constraint violation guarantees under a variety of settings. We summarize the mentioned approaches with provable guarantees in CMDPs in Table 1.1 to make a detailed comparison.

1.1 Linear Programming approach for CMDPs

In this section, we present the LP approach for CMDPs which serves the fundamental role of designing LP-based algorithms. We consider finite-horizon CMDPs with time-dependent dynamics. A finite-horizon CMDPs is denoted by the tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g, \mu_0)$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, H is the horizon. The transition $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is a collection of transition kernels (transition probability matrices). The initial state x_1 at the beginning of each episode is sampled from the distribution μ_0. Then at step h, the immediate reward and utility for taking an action a_h at state x_h are random variables $R_h(x_h, a_h)$ and $G_h(x_h, a_h)$, with expectation $\mathbb{E}[R_h(x, a)] = r_h(x_h, a_h)$ and $\mathbb{E}[G_h(x_h, a_h] = g_h(x_h, a_h)$ respectively. The environment then moves to a new state x_{h+1} sampled from distribution $\mathbb{P}_h(\cdot|x_h, a_h)$. Similar to [34], we assume that $r_h(x, a)(g_h(x, a)) : \mathcal{S} \times \mathcal{A} \to [0, 1]$.

A Markov randomized policy $\pi = (\pi_1, \ldots, \pi_H), \pi_i : \mathcal{S} \to \Delta_A$ maps states to a simplex Δ_A on the action set $\mathcal{A}$. We only consider Markov policies here since they are rich enough to cover all the behavioral policies (policies that depend on all the history). We will discuss more later in Section 1.2. Then at step h, the action $a_h \sim \pi_h(\cdot|x_h)$ is taken at state x_h according to a policy π. Then for any $h \in [H], (s, a) \in \mathcal{S} \times \mathcal{A}$, the reward $Q-$value function is the expected cumulative reward when an agent starts from a state-action pair (x, a) at step h and then follows policy π :

$$Q_h^\pi(x, a) = r_h(x, a) + \mathbb{E}\left[\sum_{i=h+1}^{H} r_i(x_i, \pi_i(x_i)) \middle| \; x_h = x, a_h = a \right], \tag{1.1}$$

where the expectation is over the environment, the initial state distribution, and policy randomness. The value function is expected cumulative rewards from step h to the end of the episode under policy π, which is defined as:

$$V_h^\pi(x_h) = \sum_a \pi_h(a|x_h) Q_h^\pi(x, a) \tag{1.2}$$

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

$$C_h^\pi(x,a) = g_h(x,a) + \mathbb{E}\left[\sum_{i=h+1}^{H} g_i(x_i, \pi_i(x_i)) \middle| \; x_h = x, a_h = a \right] \tag{1.3}$$

$$W_h^\pi(x) = \sum_a \pi_h(a|x_h) C_h^\pi(x,a). \tag{1.4}$$

Given the model defined above, the objective of the agent is to find a policy that maximizes the expected cumulative reward subject to a constraint on the expected utility:

$$\max_{\pi \in \Pi} \mathbb{E}\left[V_1^\pi(x_1)\right] \quad \text{subject to: } \mathbb{E}\left[W_1^\pi(x_1)\right] \geq \rho, \tag{1.5}$$

where we assume $\rho \in [0, H]$ to avoid triviality since all the value functions are in $[0, H]$ due to the assumptions on the bound on reward/utility functions. We also assume that $V_{H+1}(x) = W_{H+1} = 0$ for any $x \in \mathcal{S}$. We can reformulate the optimization problem by using the occupancy measure [35; 11]. The occupancy measure q^π of a policy π is defined as the set of distributions generated by executing the policy π :

$$q_h^\pi(x,a;\mathbb{P}) := \mathbb{E}[\mathbb{I}_{\{x_h = x, a_h = a | x_1 \sim \mu_0, \mathbb{P}, \pi\}}] = \Pr\{x_h = x, a_h = a | x_1 \sim \mu_0, \mathbb{P}, \pi\}. \tag{1.6}$$

For ease of notation, We define the matrix notation $q^\pi \in \mathbb{R}^{HSA}$, where its $(x,a,h), \mathbb{P}$ and the initial distribution of x_1 is given by $q_h^\pi(x,a;\mathbb{P})$ are given and clear. This implies the value function can be rewritten as the occupancy measure:

$$\begin{aligned} V_1^\pi(x_1) =& \mathbb{E}\left[\sum_{h=1}^{H} r_h(x_h, a_h | x_1 \sim \mu_0, \pi)\right] = \sum_{h=1}^{H} \mathbb{E}\left[r_h(x_h, a_h) | x_1 \sim \mu_0, \pi\right] \\ =& \sum_{h=1}^{H} \sum_{x,a} r_h(x,a) \Pr\{x_h = x, a_h = a | x_1 \sim \mu_0, \pi\} \\ =& \sum_{h,x,a} q_h^\pi(x,a;\mathbb{P}) r_h(x,a) := r^\top q^\pi, \end{aligned} \tag{1.7}$$

where the second equality holds due to the linearity of expectation, and $r \in \mathbb{R}^{HSA}$ such that the element x, a, h element is given by $r_h(x, a)$. Similar, we can have $W_1^\pi(x_1) = g^\top q^\pi$. Thus the objective of a CMDP can be formulated as :

$$\pi^* \in \arg\min r^\top q^\pi \tag{1.8}$$

$$s.t. g^\top q^\pi \geq \rho. \tag{1.9}$$

For the occupancy measure$q$, it is easy to say that for any given policy π, it satisfies that [36]:

$$\sum_a q_h^\pi(x, a) = \sum_{x', a'} \mathbb{P}_{h-1}(x|x', a') q_{h-1}^\pi(x', a'), \qquad \forall x \in \mathcal{S} \tag{1.10}$$

$$q_h^\pi(x, a) \geq 0, \qquad \forall x, a, \tag{1.11}$$

for all $h \in [H]/\{1\}$, for $h = 1$ we have that $q_1(x, a) = \pi_1(a|x) \cdot \mu_0(x), \forall x, a$. We remove the dependence on the model $\mathbb{P}$ in q_h^π here for ease of notation. Also the occupancy measure satisfies that $\sum_{x,a} q_h^\pi(x, a) = 1, \forall h \in [H]$. Also since the occupancy measure satisfies the affine constraint, the we can state the following property [37]:

Proposition I.1. *The occupancy measure is convex.*

Then due to the linearity of the constraint function and of the structure of the occupancy measure, the original control problem can be reduced to an LP problem where the optimization variables are occupancy measures. The Markov policy π_h^q can be constructed through the occupancy measure q_h:

$$\pi_h^q(a|x) = \frac{q_h(x, a)}{\sum_{a'} q_h(x, a')}, \quad \forall (x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{1.12}$$

Thus we are ready to state the formal LP problem as:

$$
\begin{aligned}
&\max_{q_h} \sum_{h,x,a} q_h(x,a) r_h(x,a) \\
&\text{s.t.:} \sum_{h,x,a} q_h(x,a) g_h(x,a) \geq \rho \\
&\quad \sum_{a} q_h(x,a) = \sum_{x',a'} \mathbb{P}_{h-1}(x|x',a') q_{h-1}(x',a') \\
&\quad \sum_{x,a} q_h(x,a) = 1, \forall h \in [H] \qquad (1.13) \\
&\quad \sum_{a} q_1(x,a) = \mu_0(x) \\
&\quad q_h(x,a) \geq 0, \forall x \in \mathcal{S}, \forall a \in \mathcal{A}, \forall h \in [H].
\end{aligned}
$$

Therefore, we can observe that we can obtain the optimal policy by solving the LP problem when the reward functions, utility functions, initial distribution, and transition probabilities are given.

1.2 The dominance of Markov policies

In order to show the dominance of Markov policies, we first define a history at time h to be a sequence of previous states and actions, as well as the current state: $b_h = (x_1, a_1, \ldots, x_{h-1}, a_{h-1}, x_h)$. Let B_h be the set of all possible histories of length h.

Definition I.2. A policy π is called a behavioral policy if the agent chooses an action based on the history b_h, i.e., $a \sim \pi(\cdot|b_h)$. The class of all policies defined above is denoted by U, and is called the class of behavioral policies.

Let U_M denote all the Markov policies, for any $\pi \in U_M$, a_h is only a function of x_h. Next we define the dominating policies formally:

Definition I.3. A class of policies $\bar{U}$ is said to be a dominating class of policies for LP optimization problem 1.13. If for any policy $\pi \in U$ there exists a policy $\bar{\pi} \in \bar{U}$

such that

$$V_1^{\pi}(x_1) \leq V_1^{\bar{\pi}}(x_1) \quad \text{and} \quad W_1^{\pi}(x_1) \leq W_1^{\bar{\pi}}(x_1) \tag{1.14}$$

Then we have the following theorem:

Theorem I.4. *The Markov policies are dominating for any utility function which is a function of the marginal distribution of state and actions.*

Thus the class of Markov policies turns out to be rich enough that for any policy in U, there exists an equivalent policy in U_M that induces the same marginal probability measure, i.e., the same probability distribution of the pairs $(x_h, a_h), h \in [H]$. Therefore we only consider Markov policies for the CMDPs. We refer readers to the detailed proofs in Theorem 6.1 in [11].

1.3 The structure of this dissertation

The structure of this dissertation is as follows. Chapter II addresses the episodic CMDP setting and introduces the Triple-Q algorithm, the first model-free RL algorithm for CMDPs with sublinear regret and zero constraint violation. The algorithm is named Triple-Q because it includes three key components: a Q-function for the cumulative reward, a Q-function for the cumulative utility (cost) for the constraint, and a virtual queue that estimates the cumulative constraint violation. Under Triple-Q, at each step, an action is chosen based on the pseudo-Q-value that is a combination of the three "Q" values. Triple-Q combines the principle of "optimism in the face of uncertainty" to overestimate the Q-values and the principle of "pessimism in the face of constraints" to pessimistically track the constraint violation. Triple-Q is designed as a two-time-scale algorithm, which is critical for an accurate estimation of the two Q values while maintaining a small constraint violation. Triple-Q is proven to achieve sublinear reward regret and guarantees zero constraint violation.

Chapter III tackles the more challenging setting of infinite-horizon average-reward CMDPs, where the agent aims for an optimal long-term average reward under constraints. The dissertation presents the first model-free RL algorithm for this setting, based on the primal-dual approach. Through Lyapunov drift analysis, the algorithm also achieves the best existing sublinear regret and zero constraint violation results.

Learning in non-stationary CMDPs, where rewards/utilities and dynamics change over time, presents additional challenges. Chapter IV focuses on designing model-free algorithms with sublinear regret and zero constraint violation guarantees for non-stationary CMDPs, particularly when the total variation budget is unknown. The dissertation presents different algorithms designed for both tabular CMDPs and linear function approximation in large state and action spaces. These algorithms employ periodic restart strategies, optimism bonuses, and a general double restart method based on the "bandit over bandit" idea. Our results [33] 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 to both tabular and linear approximation settings.

In the end, Chapter V concludes this dissertation.

CHAPTER II

Triple-Q: A Model-Free Algorithm for Episodic CMDP with Sublinear Regret and Zero Constraint Violation

2.1 Introduction

Reinforcement Learning has gained significant attention because of its successes in board games and video games such as Go [1] and StarCraft [2], and in highly-complex robotics systems [3]. An agent's objective in a typical RL problem is to maximize the cumulative reward through interacting with an unknown environment. In board games or video games, the outcomes of a random action are not consequential to the users (e.g. not life-threatening). However, a careless action in an engineering system might have catastrophic outcomes such as collisions and fatalities in robotics and autonomous driving [4; 5; 6] or surgical robotics [38]. We consider cumulative constraints in episodic CMDPs in this chapter, which include budget constraints, energy constraints, or structural fatigue in flexible UAVs.

Related Work

Earlier studies on CMDPs assume the model is known. A comprehensive study of these early results can be found in [11]. RL for unknown CMDPs has been a topic

of great interest recently because of its importance in Artificial Intelligence (AI) and Machine Learning (ML). The most noticeable advances recently are *model-based* RL for CMDPs, where the transition kernels are learned and used to solve the linear programming (LP) problem for the CMDP [14; 15; 16; 17; 20; 39], or the LP problem in the primal component of a primal-dual algorithm [18; 17; 20]. If the transition kernel is linear, then it can be learned in a sample-efficient manner even for infinite state and action spaces and then be used in the policy evaluation and improvement in a primal-dual algorithm [19]. [19] also proposes a model-based algorithm for the tabular setting (without assuming a linear transition kernel).

The performance of a model-based RL algorithm depends on how accurately a model can be estimated. For some complex environments, building accurate models is challenging computationally and data-wise [40]. For such environments, model-free RL algorithms often are more desirable. However, there has been little development on model-free RL algorithms for CMDPs with provable optimality or regret guarantees, with the exceptions [25; 41; 27], all of which require simulators. In particular, the sample-based NPG-PD algorithm in [25] requires a simulator that can simulate the MDP from any initial state x, and the algorithms in [41; 27] both require a simulator for policy evaluation. It has been argued in [42; 43; 34] that with a perfect simulator, exploration is not needed, and sample efficiency can be easily achieved because the agent can query any (state, action) pair as it wishes. Unfortunately, for complex environments, building a perfect simulator often is as difficult as deriving the model for the CMDP. For those environments, sample efficiency and the exploration-exploitation trade-off are critical and become one of the most important considerations of RL algorithm design.

Table 2.1: The Exploration-Exploitation Tradeoff in Episodic CMDPs.

	Algorithm	**Regret**	**Constraint Violation**
Model-based	OPDOP [19]	$\tilde{\mathcal{O}}(H^3\sqrt{S^2AK})$	$\tilde{\mathcal{O}}(H^3\sqrt{S^2AK})$
	OptDual-CMDP [17]	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$
	OptPrimalDual-CMDP [17]	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$
	CONRL [15]	$\tilde{\mathcal{O}}(H^3\sqrt{S^3A^2K})$	$\tilde{\mathcal{O}}(H^3\sqrt{S^3A^2K})$
	OptPess-LP [20]	$\tilde{\mathcal{O}}(H^3\sqrt{S^3AK})$	0
	OptPess–PrimalDual [20]	$\tilde{\mathcal{O}}(H^3\sqrt{S^3AK})$	$\mathcal{O}(1)$
	OPSRL[39]	$\tilde{\mathcal{O}}(\sqrt{S^4H^7AK})$	0
Model-free	Triple-Q	$\tilde{\mathcal{O}}\left(\frac{1}{\delta}H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{\frac{4}{5}}\right)$	0

Main Contributions

We consider the online learning problem of an episodic CMDP with a model-free approach *without* a simulator. We develop the first *model-free* RL algorithm for CMDPs with sublinear regret and *zero* constraint violation (for large K). The algorithm is named Triple-Q because it has three key components: (i) a Q-function (also called action-value function) for the expected cumulative reward, denoted by $Q_h(x,a)$ where h is the step index and (x,a) denotes a state-action pair, (ii) a Q-function for the expected cumulative utility for the constraint, denoted by $C_h(x,a)$, and (iii) a virtual-Queue, denoted by Z, which overestimates the cumulative constraint violation so far. At step h in the current episode, when observing state x, the agent selects action a^* based on a *pseudo-Q-value* that is a combination of the three "Q" values:

$$a^* \in \underbrace{\arg\max_a Q_h(x,a) + \frac{Z}{\eta}C_h(x,a)}_{\text{pseudo-Q-value of state } (x,a) \text{ at step } h}, \tag{2.1}$$

where η is a constant. Triple-Q uses UCB-exploration when learning the Q-values, where the UCB bonus and the learning rate at each update both depend on the visit counts to the corresponding (state, action) pair as in [34]). Different from the optimistic Q-learning for unconstrained MDPs (e.g. [34; 44; 45]), the learning rates

in Triple-Q need to be periodically reset at the beginning of each frame, where a frame consists of K^α consecutive episodes. The value of the virtual Queue (the dual variable) is updated once in every frame. So Triple-Q can be viewed as a two-time-scale algorithm where virtual-Queue is updated at a slow time scale, and Triple-Q learns the pseudo-Q-value for fixed Z at a fast time scale within each frame. Furthermore, it is critical to update the two Q-functions ($Q_h(x,a)$ and $C_h(x,a)$) following a rule similar to SARSA [46] instead of Q-learning [47], in other words, using the Q-functions of the action that is taken instead of using the max function.

We prove Triple-Q achieves $\tilde{\mathcal{O}}\left(\frac{1}{\delta}H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{\frac{4}{5}}\right)$ reward regret and guarantees *zero* constraint violation when the total number of episodes $K \geq \left(\frac{16\sqrt{SAH^6\iota^3}}{\delta}\right)^5$, where ι is logarithmic in K. Therefore, in terms of the constraint violation, our bound is sharp for large K. To the best of our knowledge, this is the first *model-free*, *simulator-free* RL algorithm with sublinear regret and *zero* constraint violation. For model-based approaches, it has been shown that a model-based algorithm achieves both $\tilde{\mathcal{O}}(\sqrt{H^4SAK})$ regret and constraint violation (see, e.g. [17]). Two concurrent papers [20; 39] developed model-based approaches that achieve zero constraint violation assuming a strictly safe policy is known a-prior. It remains open that what is the fundamental lower bound on the regret under model-free algorithms for CMDPs and whether the regret bound under Triple-Q is order-wise sharp or can be further improved. Table 2.1 summarizes the key results on the exploration-exploitation tradeoff of CMDPs in the literature. We note that it is technically more challenging to bound regret and constraint violation of model-free algorithms for CMDPs than model-based algorithms. Under a model-based algorithm, the regret and constraint violation are determined by the accuracy of the estimated model (transition kernels, reward functions, etc). The accuracy of the estimated model improves as the number of data samples increases, and so does the performance of the learned policy. Without maintaining a model, the learning target (the pseudo-Q values) varies over time, depending on the dual

variables, which becomes a key difficulty in bounding regret and constraint violation of model-free algorithms for CMDPs. Furthermore, the optimal policy for a CMDP is stochastic in general, so a greedy policy based on fixed pseudo-Q-values will not be optimal, which makes it much more challenging than bounding regret of model-free algorithms for unconstrained MDPs like the optimistic Q-learning [34; 44; 45].

As with many other model-free RL algorithms, a major advantage of Triple-Q is its low computational complexity. The computational complexity of Triple-Q is similar to SARSA for unconstrained MDPs, so it retains both its effectiveness and efficiency while solving a much harder problem. While we consider a tabular setting in this chapter, Triple-Q can easily incorporate function approximations (linear function approximations or neural networks) by replacing the $Q(x, a)$ and $C(x, a)$ with their function approximation versions, making the algorithm a very appealing approach for solving complex CMDPs in practice.

We note that safe exploration is an active topic in reinforcement learning and several heuristic methods, without provable guarantees, have been developed over the past last years (see e.g., [48; 49; 50; 51; 52]). We will compare the performance of Triple-Q and Deep Triple-Q (Triple-Q with neural networks) with some of these algorithms in Section 2.4, and demonstrate significant performance improvements (higher rewards, lower costs, and faster convergence) under Triple-Q.

2.2 Problem Formulation

We consider an episodic CMDP, denoted by $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \{r_h\}_{h=1}^H, \{g_h\}_{h=1}^H, \mu_0)$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, H is the number of steps in each episode, and $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is a collection of transition kernels (transition probability matrices). At the beginning of each episode, an initial state x_1 is sampled from the distribution μ_0. Then at step h, the agent takes action a_h after observing state x_h. Then the agent receives a reward $r_h(x_h, a_h)$ and incurs a utility

$g_h(x_h, a_h)$. The environment then moves to a new state x_{h+1} sampled from distribution $\mathbb{P}_h(\cdot|x_h, a_h)$. Similar to [34], we assume that $r_h(x, a)(g_h(x, a)) : \mathcal{S} \times \mathcal{A} \to [0, 1]$, are deterministic for convenience.

Given a Markovian policy π, which is a collection of H functions $\{\pi_h : \mathcal{S} \to \Delta_A\}_{h=1}^H$, the reward value function V_h^π at step h is the expected cumulative rewards from step h to the end of the episode under policy π :

$$V_h^\pi(x) = \mathbb{E}\left[\sum_{i=h}^{H} r_i(x_i, \pi_i(x_i))\middle| x_h = x\right]. \tag{2.2}$$

The (reward) Q-function $Q_h^\pi(x, a)$ at step h is the expected cumulative reward when an agent starts from a state-action pair (x, a) at step h and then follows policy π :

$$Q_h^\pi(x, a) = r_h(x, a) + \mathbb{E}\left[\sum_{i=h+1}^{H} r_i(x_i, \pi_i(x_i))\middle| \ x_h = x, a_h = a \ \right]. \tag{2.3}$$

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

$$W_h^\pi(x) = \mathbb{E}\left[\sum_{i=h}^{H} g_i(x_i, \pi_i(x_i))\middle| x_h = x\right], \tag{2.4}$$

$$C_h^\pi(x, a) = g_h(x, a) + \mathbb{E}\left[\sum_{i=h+1}^{H} g_i(x_i, \pi_i(x_i))\middle| \ x_h = x, a_h = a \ \right]. \tag{2.5}$$

Given the model defined above, the objective of the agent is to find a policy that maximizes the expected cumulative reward subject to a constraint on the expected utility:

$$\max_{\pi \in \Pi} \mathbb{E}\left[V_1^\pi(x_1)\right] \quad \text{subject to: } \mathbb{E}\left[W_1^\pi(x_1)\right] \geq \rho, \tag{2.6}$$

where we assume $\rho \in [0, H]$ to avoid triviality since all the value functions are in $[0, H]$ due to the assumptions on the bound on reward/utility functions. The expectation is taken with respect to the initial distribution $x_1 \sim \mu_0$. We remark here that the

optimal Markovian policy is proven to exist, we refer readers to the book [11].

For simplicity, we adopt the following notation (some used in [34; 19]):

$$\mathbb{P}_h V_{h+1}^{\pi}(x,a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V_{h+1}^{\pi}(x'), Q_h^{\pi}(x, \pi_h(x)) = \sum_a Q_h^{\pi}(x,a)\mathbb{P}(\pi_h(x) = a), \quad (2.7)$$

$$\mathbb{P}_h W_{h+1}^{\pi}(x,a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} W_{h+1}^{\pi}(x'), C_h^{\pi}(x, \pi_h(x)) = \sum_a C_h^{\pi}(x,a)\mathbb{P}(\pi_h(x) = a). \quad (2.8)$$

From the definitions above, we have

$$V_h^{\pi}(x) = Q_h^{\pi}(x, \pi_h(x)),\ Q_h^{\pi}(x,a) = r_h(x,a) + \mathbb{P}_h V_{h+1}^{\pi}(x,a), \quad (2.9)$$

$$W_h^{\pi}(x) = C_h^{\pi}(x, \pi_h(x)),\ C_h^{\pi}(x,a) = g_h(x,a) + \mathbb{P}_h W_{h+1}^{\pi}(x,a). \quad (2.10)$$

The results in the chapter can be directly applied to a constraint in the form of $\mathbb{E}\left[W_1^{\pi}(x_1)\right] \leq \rho$. Without loss of generality, assume $\rho \leq H$. We define $\tilde{g}_h(x,a) = 1 - g_h(x,a) \in [0,1]$ and $\tilde{\rho} = H - \rho \geq 0$, $\mathbb{E}\left[W_1^{\pi}(x_1)\right] \leq \rho$ can be written as $\mathbb{E}\left[\tilde{W}_1^{\pi}(x_1)\right] \geq \tilde{\rho}$, where

$$\mathbb{E}\left[\tilde{W}_1^{\pi}(x_1)\right] = \mathbb{E}\left[\sum_{i=1}^{H} \tilde{g}_i(x_i, \pi_i(x_i))\right] = H - \mathbb{E}\left[W_1^{\pi}(x_1)\right]. \quad (2.11)$$

Let π^* denote the optimal solution to the CMDP problem defined in (2.6). We evaluate our model-free RL algorithm using regret and constraint violation defined below:

$$\text{Regert}(K) = \mathbb{E}\left[\sum_{k=1}^{K} \left(V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1})\right)\right], \quad (2.12)$$

$$\text{Violation}(K) = \mathbb{E}\left[\sum_{k=1}^{K} \left(\rho - W_1^{\pi_k}(x_{k,1})\right)\right], \quad (2.13)$$

where $V_1^*(x) = V_1^{\pi^*}(x)$, π_k is the policy used in episode k and the expectation is taken with respect to the distribution of the initial state $x_{k,1} \sim \mu_0$ and the randomness of

π_k. We further make the following assumption.

Assumption II.1. *(Slater's Condition). Given initial distribution μ_0, there exist $\delta > 0$ and policy π such that $\mathbb{E}\left[W_1^{\pi}(x_1)\right] - \rho \geq \delta$.*

In this research, Slater's condition simply means there exists a feasible policy that can satisfy the constraint with a slackness δ. This has been commonly used in the literature [19; 25; 17; 53]. We call δ Slater's constant. While the regret and constraint violation bounds depend on δ, our algorithm does not need to know δ under the assumption that K is large (the exact condition can be found in Theorem II.2). This is a noticeable difference from some of the works in CMDPs in which the agent needs to know the value of this constant (e.g. [19]) or alternatively a feasible policy (e.g. [54]) .

2.3 Algorithm

We now formally introduce the algorithm Triple-Q. The design of our algorithm is based on the primal-dual approach in optimization. While RL algorithms based on the primal-dual approach have been developed for CMDPs (see. e.g. [19; 25; 18; 17; 20]), a model-free RL algorithm with sublinear regrets and *zero* constraint violation is new.

The design of Triple-Q is based on the primal-dual approach in optimization. Given Lagrange multiplier λ, we consider the Lagrangian of problem (2.6) from a given initial state x_1 :

$$\begin{aligned} &\max_{\pi} V_1^{\pi}(x_1) + \lambda \left(W_1^{\pi}(x_1) - \rho\right) \\ =&\max_{\pi} \mathbb{E}\left[\sum_{h=1}^{H} r_h(x_h, \pi_h(x_h)) + \lambda g_h(x_h, \pi_h(x_h))\right] - \lambda\rho, \end{aligned} \tag{2.14}$$

which is an unconstrained MDP with reward $r_h(x_h, \pi_h(x_h)) + \lambda g_h(x_h, \pi_h(x_h))$ at step h. Assuming we solve the unconstrained MDP and obtain the optimal policy, denoted

by π^*_λ, we can then update the dual variable (the Lagrange multiplier) using a gradient method:

$$\lambda \leftarrow \left(\lambda + \rho - \mathbb{E}\left[W_1^{\pi^*_\lambda}(x_1)\right]\right)^+. \tag{2.15}$$

While primal-dual is a standard approach, analyzing the finite-time performance such as regret or sample complexity is particularly challenging. For example, over a finite learning horizon, we will not be able to exactly solve the unconstrained MDP for given λ. Therefore, we need to carefully design how often the Lagrange multiplier should be updated. If we update it too often, then the algorithm may not have sufficient time to solve the unconstrained MDP, which leads to divergence; on the other hand, if we update it too slowly, then the solution will converge slowly to the optimal solution, and will lead to large regret and constraint violation. Another challenge is that when λ is given, the primal-dual algorithm solves a problem with an objective different from the original objective and does not consider any constraint violation. Therefore, even when the asymptotic convergence may be established, establishing the finite-time regret is still difficult because we need to evaluate the difference between the policy used at each step and the optimal policy.

Next, we will show that a low-complexity primal-dual algorithm can converge and have sublinear regret and *zero* constraint violation when carefully designed. In particular, Triple-Q includes the following key ideas:

- A sub-gradient algorithm for estimating the Lagrange multiplier, which is updated at the beginning of each frame (recall that a frame consists of K^α consecutive episodes) as follows: $Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}}{K^\alpha}\right)^+$, where $(x)^+ = \max\{x, 0\}$ and $\bar{C}$ is the summation of all $C_1(x_1, a_1)$s of the episodes in the previous frame. We call Z a virtual queue because it is terminology that has been widely used in stochastic networks (see e.g. [55; 56]). If we view $\rho + \epsilon$ as the number of jobs that arrive at a queue within each frame and $\bar{C}$ as the number of jobs that leave the queue

within each frame, then Z is the number of jobs that are waiting at the queue. Note that we added extra utility ϵ to ρ. By choosing $\epsilon = \frac{8\sqrt{SAH^6\iota^3}}{K^{0.2}}$, the virtual queue pessimistically estimates constraint violation so Triple-Q achieves *zero* constraint violation when the number of episodes is large.

- A carefully chosen parameter $\eta = K^{0.2}$ so that when $\frac{Z}{\eta}$ is used as the estimated Lagrange multiplier, it balances the trade-off between maximizing the cumulative reward and satisfying the constraint.

- Carefully chosen learning rate α_t and Upper Confidence Bound (UCB) bonus b_t to guarantee that the estimated Q-value does not significantly deviate from the actual Q-value. We remark that the learning rate and UCB bonus proposed for unconstrained MDPs [34] does not work here. Our learning rate is chosen to be $\frac{K^{0.2}+1}{K^{0.2}+t}$, where t is the number of visits to a given (state, action) pair in a particular step. This decays much slower than the classic learning rate $\frac{1}{t}$ or $\frac{H+1}{H+t}$ used in [34]. The learning rate is further reset from frame to frame, so Triple-Q can continue to learn the pseudo-Q-values that vary from frame to frame due to the change of the virtual-Queue (the Lagrange multiplier).

The detailed description of Triple-Q is presented in Algorithm 1. The algorithm only needs to know the values of H, A, S, K,, and no other problem-specific values are needed. Furthermore, Triple-Q includes updates of two Q-functions per step: one for Q_h and one for C_h; and one simple virtual queue update per frame. So its computational complexity is similar to SARSA.

2.3.1 Main Results

The next theorem summarizes the regret and constraint violation bounds guaranteed under Triple-Q.

Algorithm 1: Triple-Q

Choose $\chi = \eta = K^{0.2}$, $\iota = 128 \log\left(\sqrt{2SAH}K\right)$, $\alpha = 0.6$, and $\epsilon = \frac{8\sqrt{SAH^6\iota^3}}{K^{0.2}}$;

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 : $x_1 \sim \mu_0$;

  for *step* $h = 1, \ldots, H+1$ **do**

    if $h \le H$; // take a greedy action based on the pseudo-Q-function

    then

      Take action $a_h \leftarrow \arg\max_a \left(Q_h(x_h,a) + \frac{Z}{\eta} C_h(x_h,a)\right)$;

      Observe $r_h(x_h,a_h), g_h(x_h,a_h)$, and x_{h+1} ;

      $N_h(x_h,a_h) \leftarrow N_h(x_h,a_h) + 1, V_h(x_h) \leftarrow Q_h(x_h,a_h), W_h(x_h) \leftarrow C_h(x_h,a_h)$;

    if $h \ge 2$; // update the Q-values for (x_{h-1}, a_{h-1}) after observing (s_h, a_h)

    then

      Set $t = N_{h-1}(x_{h-1},a_{h-1}), b_t = \frac{1}{4}\sqrt{\frac{H^2\iota(\chi+1)}{\chi+t}}, \alpha_t = \frac{\chi+1}{\chi+t}$;

      Update the reward Q-value: $Q_{h-1}(x_{h-1},a_{h-1}) \leftarrow (1-\alpha_t)Q_{h-1}(x_{h-1},a_{h-1}) + \alpha_t\left(r_{h-1}(x_{h-1},a_{h-1}) + V_h(x_h) + b_t\right)$;

      Update the utility Q-value: $C_{h-1}(x_{h-1},a_{h-1}) \leftarrow (1-\alpha_t)C_{h-1}(x_{h-1},a_{h-1}) + \alpha_t\left(g_{h-1}(x_{h-1},a_{h-1}) + W_h(x_h) + b_t\right)$;

    if $h = 1$ **then**

      $\bar{C} \leftarrow \bar{C} + C_1(x_1,a_1)$; // add $C_1(x_1,a_1)$ to $\bar{C}$

  if $k \mod (K^\alpha) = 0$; // reset visit counts and Q-functions

  then

    $N_h(x,a) \leftarrow 0$, $Q_h(x,a) = C_h(x,a) \leftarrow H$, $Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}}{K^\alpha}\right)^+$, and $\bar{C} \leftarrow 0$; // update the virtual-queue length

Theorem II.2. *Assume* $K \ge \left(\frac{16\sqrt{SAH^6\iota^3}}{\delta}\right)^5$, *where* $\iota = 128\log(\sqrt{2SAH}K)$. *Triple-Q achieves the following regret and constraint violation bounds:*

$$\text{Regret}(K) \le \frac{13}{\delta} H^4 \sqrt{SA\iota^3} K^{0.8} + \frac{4H^4\iota}{K^{1.2}} \tag{2.16}$$

$$\text{Violation}(K) \le \frac{54H^4\iota K^{0.6}}{\delta} \log\frac{16H^2\sqrt{\iota}}{\delta} + \frac{4\sqrt{H^2\iota}}{\delta} K^{0.8} - 5\sqrt{SAH^6\iota^3}K^{0.8}. \tag{2.17}$$

If we further have $K \geq e^{\frac{1}{\delta}}$, *then* $\text{Violation}(K) \leq 0$ *and*

$$\Pr\left(\sum_{k=1}^{K} \rho - W_1^{\pi_k}(x_{k,1}) \leq 0\right) = 1 - \tilde{\mathcal{O}}\left(e^{-K^{0.2+}} + \frac{1}{K^2}\right), \tag{2.18}$$

in other words, Triple-Q guarantees zero constraint violation both on expectation and with a high probability. □

We note that the theorem holds when K is sufficiently large, and how large K needs to depend on the slackness δ.

Novelty of the Proof Technique

We remark that a key difference between our analysis and the analysis of the optimistic Q-learning for unconstrained MDPs [34; 44; 45; 57; 58] is that our proof relies heavily on the Lyapunov-drift analysis of virtual-Queue Z. The drift analysis on the Lyapunov function Z^2 relates the difference between the optimal reward Q-function and the learned reward Q-function to the difference between the optimal pseudo-Q-function and the learned pseudo-Q-function. For fixed Z, Triple-Q can be regarded as optimistic SARSA for the pseudo-Q-function, so the relationship enables us to establish the regret bound by analyzing the pseudo-Q-function. Furthermore, the Lyapunov-drift analysis on the moment generating function of Z, i.e. $\mathbb{E}[e^{rZ}]$ yields an upper bound on Z that holds uniformly over the entire learning horizon. This upper bound, together with a fundamental relationship between Z and constraint violation, leads to the constraint violation bound. The Lyapunov drift analysis has been used to establish sublinear regret and zero constraint violation in constrained linear bandits [59]. Some of the proofs were inspired by [59]. Compared with bandit problems, CMDPs, however, is a much more challenging problem due to their sequential nature.

2.3.2 The choices of the Hyper-parameters in Triple-Q

Recall that the regret upper bound in (2.58) and the constraint violation bound in (2.60):

$$\text{Regret}(K) = \mathcal{O}\left(K\epsilon + K^{1-\alpha} + \frac{K}{\chi} + \sqrt{K^{2-\alpha}\chi} + \frac{K}{\eta}\right) \tag{2.19}$$

$$\text{Violation}(K) \leq -K\epsilon + \mathcal{O}\left(K^{\alpha}\eta + K^{1-\alpha} + \frac{K}{\chi} + \sqrt{K^{2-\alpha}\chi}\right). \tag{2.20}$$

Note that we simplify the bounds above by keeping only K and the hyper-parameters χ, α, ϵ and η, which should be chosen as functions of K. Letting $\chi = K^{\beta}$, in order to have $\mathcal{O}(K/\chi)$ and $\mathcal{O}(\sqrt{K^{2-\alpha}\chi})$ be of the same order, we should choose $\alpha = 3\beta$. Therefore,

$$\begin{aligned}\text{Regret}(K) =& \mathcal{O}\left(K\epsilon + K^{1-3\beta} + K^{1-\beta} + K^{1-\beta} + \frac{K}{\eta}\right) \\ =& \mathcal{O}\left(K\epsilon + K^{1-\beta} + \frac{K}{\eta}\right) \end{aligned} \tag{2.21}$$

$$\begin{aligned}\text{Violation}(K) \leq& -K\epsilon + \mathcal{O}\left(K^{3\beta}\eta + K^{1-3\beta} + K^{1-\beta} + K^{1-\beta}\right) = \\ & -K\epsilon + \mathcal{O}\left(K^{3\beta}\eta + K^{1-\beta}\right). \end{aligned} \tag{2.22}$$

To guarantee zero constraint violation, we need to have $K\epsilon$, $K^{\beta}\eta$ and $K^{1-\beta}$ of the same order, so we set

$$\epsilon = \mathcal{O}\left(K^{-\beta}\right) \quad \text{and} \quad \eta = \mathcal{O}\left(K^{1-4\beta}\right).$$

To minimize the regret upper bound, $K^{1-\beta}$ and $\frac{K}{\eta} = K^{4\beta}$ should be of the same order, so $\beta = 0.2$, which leads to the choices of $\alpha = 0.6$, $\chi = K^{0.2}$, $\epsilon = \mathcal{O}(K^{-0.2})$, and $\eta = \mathcal{O}(K^{0.2})$.

2.4 Simulation

We remark that when implementing Triple-Q, we do not need to reset all the $Q_h(x,a)$ and $C_h(s,a)$ to H. Instead, we added extra "bonuses" to the learned values at the beginning of each frame to ensure overestimation. This allows Triple-Q continues to learn across frames. In particular, at the beginning of each frame, we update all Q-values as follows to replace lines 18-20.

Algorithm 2: Replacing Lines 18-20 of Triple-Q

1 **if** $k \mod (K^\alpha) = 0$; // reset visit counts and add bonuses to Q-functions
2 **then**
3 $N_h(x,a) \leftarrow 0$ and $Q_h(x,a) \leftarrow Q_h(x,a) + \frac{2H^3\sqrt{\iota}}{\eta}, \forall (x,a,h)$. **if** $Q_h(x,a) \geq H$ *or* $C_h(x,a) \geq H$ **then**
4 $Q_h(x,a) \leftarrow H$ and $C_h(x,a) \leftarrow H$; $Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}}{K^\alpha}\right)^+$, and $\bar{C} \leftarrow 0$; // update the virtual-queue length

Consider frame $T+1$. Note that if $Q^+_{TK^\alpha+1,h}(x,a) = C^+_{TK^\alpha+1,h}(x,a) = H$, then condition (i) in the proof of Lemma II.5 holds. Otherwise, with the extra bonus, we have $Q^+_{TK^\alpha+1,h}(x,a) = Q^-_{TK^\alpha+1,h}(x,a) + \frac{2H^3\sqrt{\iota}}{\eta} < H$ and $C^+_{TK^\alpha+1,h}(x,a) = C^-_{TK^\alpha+1,h}(x,a) < H$. Here, we use superscript $-$ and $+$ to indicate the Q-values before and after adding the extra bonus and thresholding. Suppose that the overestimation holds at the end of frame T, i.e. $\{F^-_{TK^\alpha+1,h} - F^\pi_{TK^\alpha+1,h}\}(x,a) \geq 0$ for any π, h and (x,a). Then, at the beginning of frame $T+1$, we have

$$
\begin{aligned}
&\{F_{TK^\alpha+1,h} - F^\pi_h\}(x,a) \\
=&Q^-_{TK^\alpha+1,h}(x,a) + \frac{Z_{TK^\alpha}}{\eta} C^-_{TK^\alpha+1,h}(x,a) - Q^\pi_h(x,a) - \frac{Z_{TK^\alpha}}{\eta} C^\pi_h(x,a) \\
+&\frac{2H^3\sqrt{\iota}}{\eta} + \frac{Z_{TK^\alpha+1} - Z_{TK^\alpha}}{\eta} C^-_{TK^\alpha+1,h}(x,a) - \frac{Z_{TK^\alpha+1} - Z_{TK^\alpha}}{\eta} C^\pi_h(x,a) \\
\geq&\frac{2H^3\sqrt{\iota}}{\eta} - 2\frac{|Z_{TK^\alpha+1} - Z_{TK^\alpha}|}{\eta} H \\
\geq&0,
\end{aligned}
\tag{2.23}
$$

where the last inequality holds because according to Lemma A.2,

$$|Z_{TK^\alpha+1} - Z_{TK^\alpha}| \leq \max\left\{\rho + \epsilon, \frac{\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} C_{k,1}(x_{k,1}, a_{k,1})}{K^\alpha}\right\} \leq H^2\sqrt{\iota}.$$

In summary, condition (i) in the proof of Lemma II.5 continues to hold under this modified algorithm, assuming the overestimation result holds in the previous frame, so does the overestimation result in frame $T+1$. The advantage of this method is that the algorithm does need to learn the Q-functions from scratch in each frame.

2.4.1 A Tabular Case

We first evaluated our algorithm using a grid-world environment studied in [60]. The environment is shown in Figure 2.1-(a). 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 100, and for other locations, the Euclidean distance between them and the destination is subtracted from the longest distance. The cost constraint is set to be 6 (we transferred utility to cost as we discussed in the chapter), which means the agent is only allowed to hit the obstacles at most six times. To account for the statistical significance, all results were averaged over 25 trials, the same for later simulations.

The result is shown in Figure 2.2, from which we can observe that Triple-Q can quickly learn a well-performed policy (with about $20,000$ episodes) while satisfying the safety constraint. Triple-Q-stop is a stationary policy obtained by stopping learning (i.e. fixing the Q tables) at $40,000$ training steps (note the virtual-Queue continues to be updated so the policy is a stochastic policy). We can see that Triple-Q-stop has a similar performance as Triple-Q and show that Triple-Q yields a near-optimal, stationary policy after the learning stops.

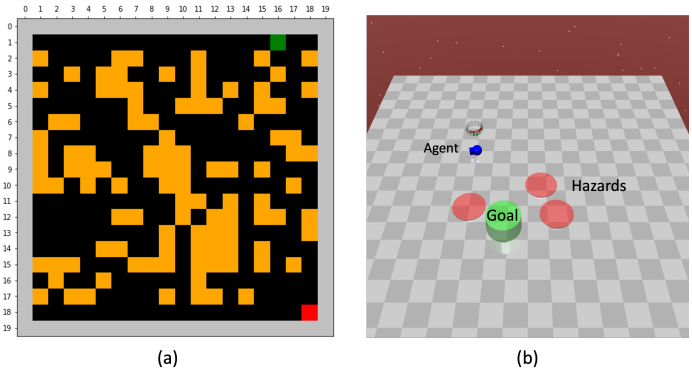

Figure 2.1: Grid World and DynamicEnv with Safety Constraints

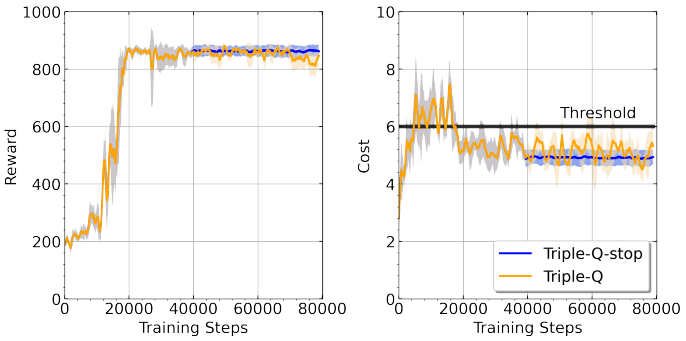

Figure 2.2: The average reward and cost under Triple-Q. The shaded region represents the 95% confidence interval.

2.4.2 Ablation Study

We investigate Triple-Q's sensitivity to hyperparameter η via an ablation study. As shown in Figure 2.3, a smaller η,, which represents a higher weight on the constraint, results in a lower cost while maintaining a similar performance in terms of reward.

2.4.3 Triple-Q with Neural Networks

We also evaluated our algorithm on the Dynamic Gym benchmark (DynamicEnv) [61] as shown in Figure. 2.1-(b). In this environment, a point agent (one actuator for turning and another for moving) navigates on a 2D map to reach the goal position while trying to avoid reaching hazardous areas. The initial state of the agent, the goal position, and the hazards are randomly generated in each episode. At each

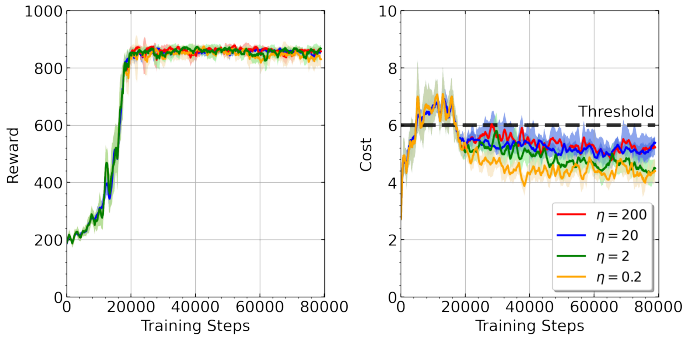

Figure 2.3: Performance of Triple-Q under different choices of η in the Grid World

step, the agents get a cost of 1 if it stays in the hazardous area; otherwise, there is no cost. The constraint is that the expected cost should not exceed 15. In this environment, both the states and action spaces are continuous, we implemented the key ideas of Triple-Q with neural network approximations and the actor-critic method. In particular, two Q functions are trained simultaneously, the virtual queue is updated slowly every few episodes, and the policy network is trained by optimizing the combined three "Q"s (Triple-Q). The implementation details can be found in Table 2.2. These hyperparameters are used in two later environments, pendulum and Ball-1D, as well. We call this algorithm Deep Triple-Q. The simulation results in Figure 2.4 show that Deep Triple-Q learns a safe policy with a high reward much faster than WCSAC [61]. In particular, it took around 0.45 million training steps under Deep Triple-Q, but it took 4.5 million training steps under WCSAC.

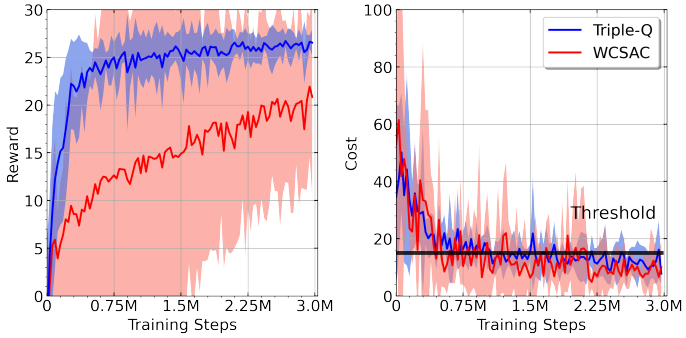

Figure 2.4: The rewards and costs of Deep Triple-Q versus WCSAC during Training

Table 2.2: Hyperparameters

Parameter	Value
optimizer	Adam
learning rate	3×1^{-3}
discount	0.99
replay buffer size	10^6
number of hidden layers (all networks)	2
batch Size	256
nonlinearity	ReLU
number of hidden units per layer (Critic and Actor)	256
virtual queue update frequency	3 episode

We further compared Deep Triple-Q with several existing safe exploration RL algorithms. We first compared Triple-Q with CBF [50] on the Pendulum environment[1]. In this environment, the constraint is that the maximum angle (rad) of the pendulum cannot exceed 1 radian, otherwise, the episode ends. Since Triple-Q was designed to address cumulative constraints, we set the threshold of the angel to be 0.5 so that the angel will not exceed 1 radian with a high probability. The result was averaged over 25 trials. As shown in Figure 2.5, we observed that Triple-Q achieved a higher reward. Although Triple-Q violated the constraint at the early stage and cannot guarantee a strictly safe policy during learning, it can learn a relatively safe policy very quickly without violating the hard constraint. We remark that CBF requires the physical model of the pendulum as prior knowledge, while Deep Triple-Q does not.

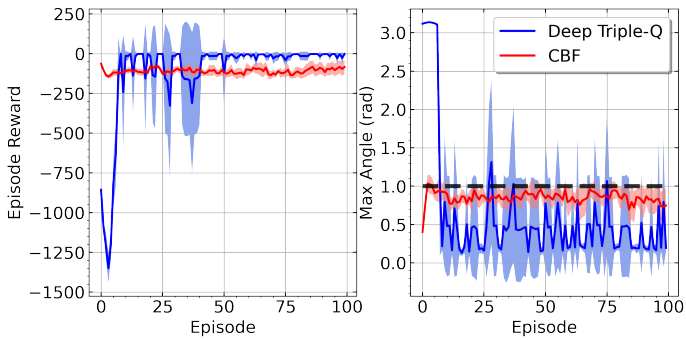

Figure 2.5: Comparison with CBF

[1]https://gym.openai.com/envs/Pendulum-v0/

Finally, we compared Triple-Q with DDPG+Safety Layer in [49] on Ball-1D environment (Figure 2.6), where the goal of the RL agent is to keep the green ball as close to the target (pink ball) as possible by controlling its velocity. The safe region is $[0, 1]$. If the green ball steps out of it, the episode terminates. The threshold in this environment was set to be 0.3. We can observed that Deep Triple-Q converged much faster than DDPG+Safety Layer as shown in Figure 2.7.

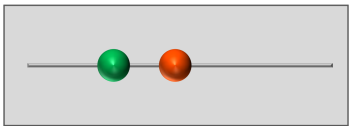

Figure 2.6: Ball1d Environment

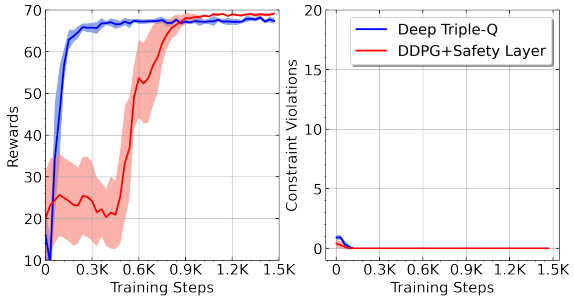

Figure 2.7: Performance during training

2.5 Details of the Proofs

In this chapter, we present the complete proof of the main theorem. A notation table and some supporting lemmas can be found in Appendix A.

2.5.1 Regret

To bound the regret, we consider the following offline optimization problem as our regret baseline [11; 35]:

$$\max_{q_h} \sum_{h,x,a} q_h(x,a) r_h(x,a) \tag{2.24}$$

$$\text{s.t.:} \sum_{h,x,a} q_h(x,a) g_h(x,a) \geq \rho \tag{2.25}$$

$$\sum_a q_h(x,a) = \sum_{x',a'} \mathbb{P}_{h-1}(x|x',a') q_{h-1}(x',a') \tag{2.26}$$

$$\sum_{x,a} q_h(x,a) = 1, \forall h \in [H] \tag{2.27}$$

$$\sum_a q_1(x,a) = \mu_0(x) \tag{2.28}$$

$$q_h(x,a) \geq 0, \forall x \in \mathcal{S}, \forall a \in \mathcal{A}, \forall h \in [H]. \tag{2.29}$$

Recall that $\mathbb{P}_{h-1}(x|x',a')$ is the probability of transitioning to state x upon taking action a' in state x' at step $h-1$. This optimization problem is linear programming (LP), where $q_h(x,a)$ is the probability of (state, action) pair (x,a) occurs in step h, $\sum_a q_h(x,a)$ is the probability the environment is in state x in step h, and

$$\frac{q_h(x,a)}{\sum_{a'} q_h(x,a')} \tag{2.30}$$

is the probability of taking action a in state x at step h, which defines the policy. We can see that (2.25) is the utility constraint, (2.26) is the global-balance equation for the MDP, (2.27) is the normalization condition so that q_h is a valid probability distribution, and (2.28) states that the initial state is sampled from μ_0. Therefore, the optimal solution to this LP solves the CMDP (if the model is known), so we use the optimal solution to this LP as our baseline.

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

$$\max_{q_h} \sum_{h,x,a} q_h(x,a) r_h(x,a) \tag{2.31}$$

$$\text{s.t.:} \sum_{h,x,a} q_h(x,a) g_h(x,a) \geq \rho + \epsilon$$

$$(2.26) - (2.29),$$

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 (2.6) or the solution to the corresponding LP (2.24) and superscript $^{\epsilon,*}$ to denote the optimal value/policy related to the ϵ-tightened version of CMDP (defined in (2.31)).

Following the definition of regret in, we have

$$\begin{aligned}\text{Regret}(K) =& \mathbb{E}\left[\sum_{k=1}^{K} V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1})\right] \\ =& \mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q_1^* q_1^*\right\}(x_{k,1}, a)\right) - Q_1^{\pi_k}(x_{k,1}, a_{k,1})\right]. \end{aligned} \tag{2.32}$$

Now by adding and subtracting the corresponding terms, we obtain

$$\begin{aligned}&\text{Regret}(K) \\ =& \mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q_1^* q_1^* - Q_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right)\right] + && (2.33)\\ & \mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1})\right)\right] + && (2.34)\\ & \mathbb{E}\left[\sum_{k=1}^{K}\left\{Q_{k,1} - Q_1^{\pi_k}\right\}(x_{k,1}, a_{k,1})\right]. && (2.35)\end{aligned}$$

Next, we establish the regret bound by analyzing the three terms above. We first present a brief outline.

Outline of the Regret Analysis

- **Step 1:** First, by comparing the LP associated with the original CMDP (2.24) and

the tightened LP (2.31), Lemma II.3 will show

$$\mathbb{E}\left[\sum_{a}\left\{Q_1^* q_1^* - Q_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right] \leq \frac{H\epsilon}{\delta},$$

which implies that under our choices of ϵ, δ, and ι,

$$(2.33) \leq \frac{KH\epsilon}{\delta} = \tilde{\mathcal{O}}\left(\frac{1}{\delta} H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}\right).$$

- **Step 2:** Note that $Q_{k,h}$ is an estimate of $Q_h^{\pi_k}$, and the estimation error (2.35) is controlled by the learning rates and the UCB bonuses. In Lemma II.4, we will show that the cumulative estimation error over one frame is upper bounded by

$$H^2 SA + \frac{H^3\sqrt{\iota} K^{\alpha}}{\chi} + \sqrt{H^4 SA\iota K^{\alpha}(\chi+1)}.$$

Therefore, under our choices of α, χ, and ι, the cumulative estimation error over K episodes satisfies

$$(2.35) \leq H^2 SAK^{1-\alpha} + \frac{H^3\sqrt{\iota} K}{\chi} + \sqrt{H^4 SA\iota K^{2-\alpha}(\chi+1)} = \tilde{\mathcal{O}}\left(H^3 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}\right).$$

The proof of Lemma II.4 is based on a recursive formula that relates the estimation error at step h to the estimation error at step $h+1$, similar to the one used in [34], but with different learning rates and UCB bonuses.

- **Step 3:** Bounding (2.34) is the most challenging part of the proof. For unconstrained MDPs, the optimistic Q-learning in [34] guarantees that $Q_{k,h}(x,a)$ is an overestimate of $Q_h^*(x,a)$ (so also an overestimate of $Q_h^{\epsilon,*}(x,a)$) for all (x,a,h,k) simultaneously with a high probability. However, this result does not hold under Triple-Q because Triple-Q takes greedy actions with respect to the pseudo-Q-function instead of the reward Q-function. To overcome this challenge, we first add and subtract additional

terms to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_{a}\left\{Q_1^{\epsilon,*}q_1^{\epsilon,*}\right\}(x_{k,1},a)-Q_{k,1}(x_{k,1},a_{k,1})\right)\right]\\
=&\mathbb{E}\left[\sum_{k}\sum_{a}\left(\left\{Q_1^{\epsilon,*}q_1^{\epsilon,*}+\frac{Z_k}{\eta}C_1^{\epsilon,*}q_1^{\epsilon,*}\right\}(x_{k,1},a)-\left\{Q_{k,1}q_1^{\epsilon,*}+\frac{Z_k}{\eta}C_{k,1}q_1^{\epsilon,*}\right\}(x_{k,1},a)\right)\right] \qquad (2.36)\\
&+\mathbb{E}\left[\sum_{k}\left(\sum_{a}\left\{Q_{k,1}q_1^{\epsilon,*}\right\}(x_{k,1},a)-Q_{k,1}(x_{k,1},a_{k,1})\right)\right]\\
&+\mathbb{E}\left[\sum_{k}\frac{Z_k}{\eta}\sum_{a}\left\{\left(C_{k,1}-C_1^{\epsilon,*}\right)q_1^{\epsilon,*}\right\}(x_{k,1},a)\right]. \qquad (2.37)
\end{aligned}$$

We can see (2.36) is the difference of two pseudo-Q-functions. Using a two-dimensional induction on step and episode, we will prove in Lemma II.5 that $\left\{Q_{k,h}+\frac{Z_k}{\eta}C_{k,h}\right\}(x,a)$ is an overestimate of $\left\{Q_h^{\epsilon,*}+\frac{Z_k}{\eta}C_h^{\epsilon,*}\right\}(x,a)$ (i.e. (2.36) ≤ 0) for all (x,a,h,k) simultaneously with a high probability. To guarantee this overestimation, Triple-Q resets all Q-values to H at the beginning of each frame.

Finally, to bound (2.37), 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 will show in Lemma II.6 that the Lyapunov-drift satisfies

$$\mathbb{E}[L_{T+1}-L_T]\leq \text{a negative drift}+H^4\iota+\epsilon^2-\frac{\eta}{K^\alpha}\sum_{k=TK^\alpha+1}^{(T+1)K^\alpha}\Phi_k, \qquad (2.38)$$

where

$$\begin{aligned}
\Phi_k=&\mathbb{E}\left[\left(\sum_{a}\left\{Q_{k,1}q_1^{\epsilon,*}\right\}(x_{k,1},a)-Q_{k,1}(x_{k,1},a_{k,1})\right)\right]\\
&+\mathbb{E}\left[\frac{Z_k}{\eta}\sum_{a}\left\{\left(C_{k,1}-C_1^{\epsilon,*}\right)q_1^{\epsilon,*}\right\}(x_{k,1},a)\right],
\end{aligned}$$

and we note that (2.37) $= \sum_k \Phi_k$. Inequality (2.38) will be established by showing that Triple-Q takes actions to *almost* greedily reduce virtual-Queue Z when Z is large, which results in the negative drift in (2.38). From (2.38), we observe that

$$\mathbb{E}[L_{T+1} - L_T] \leq H^4\iota + \epsilon^2 - \frac{\eta}{K^\alpha} \sum_{k=TK^\alpha+1}^{(T+1)K^\alpha} \Phi_k. \tag{2.39}$$

So we can bound (2.37) by applying the telescoping sum over the $K^{1-\alpha}$ frames on the inequality above:

$$(2.37) = \sum_k \Phi_k \leq \frac{K^\alpha \mathbb{E}\left[L_1 - L_{K^{1-\alpha}+1}\right]}{\eta} + \frac{K(H^4\iota + \epsilon^2)}{\eta} \leq \frac{K(H^4\iota + \epsilon^2)}{\eta},$$

where the last inequality holds because $L_1 = 0$ and $L_T \geq 0$ for all T. Combining the bounds on (2.36) and (2.37), we conclude that under our choices of ι, ϵ and η,

$$(2.34) = \tilde{\mathcal{O}}(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}).$$

Combining the results in the three steps above, we obtain the regret bound in Theorem II.2.

Detailed Proof

We next present detailed proof. The first lemma bounds the difference between the original CMDP and its ϵ-tightened version. The result is intuitive because the ϵ-tightened version is a perturbation of the original problem and $\epsilon \leq \delta$.

Lemma II.3. *Given $\epsilon \leq \delta$, we have*

$$\mathbb{E}\left[\sum_a \left\{Q_1^* q_1^* - Q_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right] \leq \frac{H\epsilon}{\delta}.$$

□

Proof. Given $q_h^*(x,a)$ is the optimal solution, we have

$$\sum_{h,x,a} q_h^*(x,a) g_h(x,a) \geq \rho.$$

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

$$\sum_{h,x,a} q_h^{\xi_1}(x,a) g_h(x,a) \geq \rho + \delta.$$

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

$$\begin{aligned}
\sum_{h,x,a} q_h^{\xi_2}(x,a) g_h(x,a) &= \sum_{h,x,a} \left((1-\frac{\epsilon}{\delta}) q_h^*(x,a) + \frac{\epsilon}{\delta} q_h^{\xi_1}(x,a) \right) g_h(x,a) \geq \rho + \epsilon, \\
\sum_{h,x,a} q_h^{\xi_2}(x,a) &= \sum_{x',a'} p_{h-1}(x|x',a') q_{h-1}^{\xi_2}(x',a'), \\
\sum_{h,x,a} q_h^{\xi_2}(x,a) &= 1.
\end{aligned}$$

Also we have $q_h^{\xi_2}(x,a) \geq 0$ for all (h,x,a). Thus $\{q_h^{\xi_2}(x,a)\}_{h=1}^H$ is a feasible solution to the ϵ-tightened optimization problem (2.31). Then given $\{q_h^{\epsilon,*}(x,a)\}_{h=1}^H$ is the optimal solution to the ϵ-tightened optimization problem, we have

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

where the last inequality holds because $0 \leq r_h(x,a) \leq 1$ under our assumption. Therefore the result follows because

$$\sum_a Q_1^*(x_{k,1},a)q_1^*(x_{k,1},a) = \sum_{h,x,a} q_h^*(x,a)r_h(x,a)$$
$$\sum_a Q_1^{\epsilon,*}(x_{k,1},a)q_1^{\epsilon,*}(x_{k,1},a) = \sum_{h,x,a} q_h^{\epsilon,*}(x,a)r_h(x,a).$$

□

The next lemma bounds the difference between the estimated Q-functions and actual Q-functions in a frame. The bound on (2.35) is an immediate result of this lemma.

Lemma II.4. *Under Triple-Q, we have for any $T \in [K^{1-\alpha}]$,*

$$\mathbb{E}\left[\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left\{Q_{k,1} - Q_1^{\pi_k}\right\}(x_{k,1},a_{k,1})\right] \leq H^2SA + \frac{H^3\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)},$$
$$\mathbb{E}\left[\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left\{C_{k,1} - C_1^{\pi_k}\right\}(x_{k,1},a_{k,1})\right] \leq H^2SA + \frac{H^3\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)}.$$

Proof. We will prove the result on the reward Q-function. The proof for the utility Q-function is almost identical. We first establish a recursive equation between a Q-function with the value-functions in the earlier episodes in the same frame. Recall that under Triple-Q, $Q_{k+1,h}(x,a)$, where k is an episode in frame T, is updated as follows:

$$Q_{k+1,h}(x,a)$$
$$= \begin{cases} (1-\alpha_t)Q_{k,h}(x,a) + \alpha_t\left(r_h(x,a) + V_{k,h+1}(x_{k,h+1}) + b_t\right) & \text{if } (x,a) = (x_{k,h},a_{k,h}) \\ Q_{k,h}(x,a) & \text{otherwise} \end{cases},$$

where $t = N_{k,h}(x,a)$. Define k_t to be the index of the episode in which the agent visits (x,a) in step h for the tth time in the current frame. The updated equation above can be written as:

$$Q_{k,h}(x,a) = (1-\alpha_t) Q_{k_t,h}(x,a) + \alpha_t \left(r_h(x,a) + V_{k_t,h+1}(x_{k_t,h+1}) + b_t \right).$$

Repeatedly using the equation above, we obtain

$$\begin{aligned}
Q_{k,h}(x,a) =& (1-\alpha_t)(1-\alpha_{t-1}) Q_{k_{t-1},h}(x,a) \\
&+ (1-\alpha_t)\alpha_{t-1} \left(r_h(x,a) + V_{k_{t-1},h+1}(x_{k_{t-1},h+1}) + b_{t-1} \right) \\
&+ \alpha_t \left(r_h(x,a) + V_{k_t,h+1}(x_{k_t,h+1}) + b_t \right) \\
=& \cdots \\
=& \alpha_t^0 Q_{(T-1)K^\alpha+1,h}(x,a) + \sum_{i=1}^{t} \alpha_t^i \left(r_h(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i \right) \qquad (2.40) \\
\leq& \alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i \left(r_h(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i \right), \qquad (2.41)
\end{aligned}$$

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

$$\begin{aligned}
\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} Q_{k,h}(x,a) \leq & \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \alpha_t^0 H + \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \sum_{i=1}^{N_{k,h}(x,a)} \alpha_{N_{k,h}}^i \left(r_h(x,a) \right. \\
& \left. + V_{k_i,h+1}(x_{k_i,h+1}) + b_i \right). \qquad (2.42)
\end{aligned}$$

The notation becomes rather cumbersome because for each $(x_{k,h}, a_{k,h})$, we need to consider a corresponding sequence of episode indices in which the agent sees $(x_{k,h}, a_{k,h})$. Next we will analyze a given sample path (i.e. a specific realization of the episodes in

a frame), so we simplify our notation in this proof and use the following notations:

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

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. Furthermore, we will replace $\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha}$ with $\sum_k$ because we only consider episodes in frame T in this proof.

We note that

$$\begin{aligned}\sum_k \sum_{i=1}^{N_{k,h}} \alpha^i_{N_{k,h}} V_{k_i^{(k,h)},h+1}\left(x_{k_i^{(k,h)},h+1}\right) &\leq \sum_k V_{k,h+1}(x_{k,h+1}) \sum_{t=N_{k,h}}^{\infty} \alpha_t^{N_{k,h}} \\ &\leq \left(1+\frac{1}{\chi}\right) \sum_k V_{k,h+1}(x_{k,h+1}), \end{aligned} \tag{2.43}$$

where the first inequality holds because because $V_{k,h+1}(x_{k,h+1})$ appears in the summation on the left-hand side each time when in episode $k' > k$ in the same frame, the environment visits $(x_{k,h}, a_{k,h})$ again, i.e. $(x_{k',h}, a_{k',h}) = (x_{k,h}, a_{k,h})$, and the second inequality holds due to the property of the learning rate proved in Lemma A.1-(d). By substituting (2.43) into (2.42) and noting that $\sum_{i=1}^{N_{k,h}(x,a)} \alpha^i_{N_{k,h}} = 1$ according to Lemma A.1-(b), we obtain

$$\begin{aligned}&\sum_k Q_{k,h}(x_{k,h}, a_{k,h}) \\ \leq &\sum_k \alpha_t^0 H + \sum_k \left(r_h(x_{k,h}, a_{k,h}) + V_{k,h+1}(x_{k,h+1})\right) \\ &+ \frac{1}{\chi} \sum_k V_{k,h+1}(x_{k,h+1}) + \sum_k \sum_{i=1}^{N_{k,h}} \alpha^i_{N_{k,h}} b_i \\ \leq &\sum_k \left(r_h(x_{k,h}, a_{k,h}) + V_{k,h+1}(x_{k,h+1})\right) + HSA + \frac{H^2\sqrt{\iota K^\alpha}}{\chi} + \frac{1}{2}\sqrt{H^2SA\iota K^\alpha(\chi+1)},\end{aligned}$$

where the last inequality holds because (i) we have

$$\sum_k \alpha^0_{N_{k,h}} H = \sum_k H \mathbb{I}_{\{N_{k,h}=0\}} \leq HSA,$$

(ii) $V_{k,h+1}(x_{k,h+1}) \leq H^2\sqrt{\iota}$ by using Lemma A.2, and (iii) we know that

$$\begin{aligned}
&\sum_k \sum_{i=1}^{N_{k,h}} \alpha^i_{N_{k,h}} b_i = \frac{1}{4} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \sum_{i=1}^{N_{k,h}} \alpha^i_{N_{k,h}} \sqrt{\frac{H^2\iota(\chi+1)}{\chi+i}} \\
\leq &\frac{1}{2} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \sqrt{\frac{H^2\iota(\chi+1)}{\chi+N_{k,h}}} \\
= &\frac{1}{2} \sum_{x,a} \sum_{n=1}^{N_{TK^\alpha,h}(x,a)} \sqrt{\frac{H^2\iota(\chi+1)}{\chi+n}} \leq \frac{1}{2} \sum_{x,a} \sum_{n=1}^{N_{TK^\alpha,h}(x,a)} \sqrt{\frac{H^2\iota(\chi+1)}{n}} \quad (1) \\
\leq &\sqrt{H^2SA\iota K^\alpha(\chi+1)},
\end{aligned}$$

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

Therefore, it is maximized when $N_{TK^\alpha,h} = K^\alpha/SA$ for all x, a, i.e. by picking the largest K^α terms. Thus we can obtain

$$\begin{aligned}
&\sum_k Q_{k,h}(x_{k,h}, a_{k,h}) - \sum_k Q^{\pi_k}_h(x_{k,h}, a_{k,h}) \\
\leq &\sum_k \left(V_{k,h+1}(x_{k,h+1}) - \mathbb{P}_h V^{\pi_k}_{h+1}(x_{k,h}, a_{k,h})\right) + HSA + \frac{H^2\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)} \\
\leq &\sum_k \left(V_{k,h+1}(x_{k,h+1}) - \mathbb{P}_h V^{\pi_k}_{h+1}(x_{k,h}, a_{k,h}) + V^{\pi_k}_{h+1}(x_{k,h+1}) - V^{\pi_k}_{h+1}(x_{k,h+1})\right) \\
&+ HSA + \frac{H^2\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)} \\
= &\sum_k \left(V_{k,h+1}(x_{k,h+1})) - V^{\pi_k}_{h+1}(x_{k,h+1}) - \mathbb{P}_h V^{\pi_k}_{h+1}(x_{k,h}, a_{k,h}) + \hat{\mathbb{P}}^k_h V^{\pi_k}_{h+1}(x_{k,h}, a_{k,h})\right) \\
&+ HSA + \frac{H^2\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)}
\end{aligned}$$

$$= \sum_k \big(Q_{k,h+1}(x_{k,h+1}, a_{k,h+1}) - Q^{\pi_k}_{h+1}(x_{k,h+1}, a_{k,h+1}) - \mathbb{P}_h V^{\pi_k}_{h+1}(x_{k,h}, a_{k,h})$$
$$+ \hat{\mathbb{P}}^k_h V^{\pi_k}_{h+1}(x_{k,h}, a_{k,h})\Big) + HSA + \frac{H^2\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)}.$$

Taking the expectation on both sides yields

$$\begin{aligned}
&\mathbb{E}\left[\sum_k Q_{k,h}(x_{k,h}, a_{k,h}) - \sum_k Q^{\pi_k}_h(x_{k,h}, a_{k,h})\right] \\
\leq &\mathbb{E}\left[\sum_k \big(Q_{k,h+1}(x_{k,h+1}, a_{k,h+1})) - Q^{\pi_k}_{h+1}(x_{k,h+1}, a_{k,h+1})\big)\right] + HSA \\
&+ \frac{H^2\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\sum_k Q_{k,h}(x_{k,h}, a_{k,h}) - \sum_k Q^{\pi_k}_h(x_{k,h}, a_{k,h})\right] \\
\leq &H^2SA + \frac{H^3\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^4SA\iota K^\alpha(\chi+1)},
\end{aligned}$$

so the lemma holds.

□

From the lemma above, we can immediately conclude:

$$\mathbb{E}\left[\sum_{k=1}^K \left\{Q_{k,1} - Q^{\pi_k}_1\right\}(x_{k,1}, a_{k,1})\right] \leq H^2SAK^{1-\alpha} + \frac{H^3\sqrt{\iota}K}{\chi} + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)}$$
$$\mathbb{E}\left[\sum_{k=1}^K \left\{C_{k,1} - C^{\pi_k}_1\right\}(x_{k,1}, a_{k,1})\right] \leq H^2SAK^{1-\alpha} + \frac{H^3\sqrt{\iota}K}{\chi} + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)}.$$

We now focus on (2.34), and further expand it as follows:

$$(2.34) = \mathbb{E}\left[\sum_{k=1}^K \left(\sum_a \left\{Q^{\epsilon,*}_1 q^{\epsilon,*}_1\right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1})\right)\right]$$

$$
\begin{aligned}
=&\mathbb{E}\left[\sum_k \sum_a \left\{\left(F_{k,1}^{\epsilon,*} - F_{k,1}\right) q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right] && (2.44)\\
&+ \mathbb{E}\left[\sum_k \left(\sum_a \left\{Q_{k,1} q_1^{\epsilon,*}\right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1})\right)\right]\\
&+ \mathbb{E}\left[\sum_k \frac{Z_k}{\eta} \sum_a \left\{\left(C_{k,1} - C_1^{\epsilon,*}\right) q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right], && (2.45)
\end{aligned}
$$

where

$$
\begin{aligned}
F_{k,h}(x,a) &= Q_{k,h}(x,a) + \frac{Z_k}{\eta} C_{k,h}(x,a)\\
F_h^{\epsilon,*}(x,a) &= Q_h^{\epsilon,*}(x,a) + \frac{Z_k}{\eta} C_h^{\epsilon,*}(x,a).
\end{aligned}
$$

We first show (2.44) can be bounded using the following lemma. This result holds because the choices of the UCB bonuses and the reset at the beginning of each frame guarantee that $F_{k,h}(x,a)$ is an over-estimate of $F_h^{\epsilon,*}(x,a)$ for all k, h and (x,a) with a high probability.

Lemma II.5. *With probability at least* $1 - \frac{1}{K^3}$, *the following inequality holds simultaneously for all* $(x,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$:

$$
\left\{F_{k,h} - F_h^{\pi}\right\}(x,a) \geq 0, \tag{2.46}
$$

which further implies that

$$
\mathbb{E}\left[\sum_{k=1}^{K} \sum_a \left\{\left(F_{k,1}^{\epsilon,*} - F_{k,1}\right) q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right] \leq \frac{4H^4\iota}{\eta K}. \tag{2.47}
$$

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

the following notations:

$$F_{k,h}(x,a) = Q_{k,h}(x,a) + \frac{Z}{\eta} C_{k,h}(x,a), \quad U_{k,h}(x) = V_{k,h}(x) + \frac{Z}{\eta} W_{k,h}(x),$$
$$F_h^\pi(x,a) = Q_h^\pi(x,a) + \frac{Z}{\eta} C_h^\pi(x,a), \quad U_h^\pi(x) = V_h^\pi(x) + \frac{Z}{\eta} W_h^\pi(x).$$

According to Lemma A.3 in the appendix, we have

$$\begin{aligned}
&\{F_{k,h} - F_h^\pi\}(x,a) \\
=&\alpha_t^0 \left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) \\
&+ \sum_{i=1}^t \alpha_t^i \left(\left\{U_{k_i,h+1} - U_{h+1}^\pi\right\}(x_{k_i,h+1}) + \{(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) U_{h+1}^\pi\}(x,a) + \left(1 + \frac{Z}{\eta}\right) b_i \right) \\
\geq_{(a)} &\alpha_t^0 \left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) + \sum_{i=1}^t \alpha_t^i \left\{U_{k_i,h+1} - U_{h+1}^\pi\right\}(x_{k_i,h+1}) \\
=_{(b)} &\alpha_t^0 \left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) \\
&+ \sum_{i=1}^t \alpha_t^i \left(\max_a F_{k_i,h+1}(x_{k_i,h+1}, a) - F_{h+1}^\pi(x_{k_i,h+1}, \pi(x_{k_i,h+1})) \right) \\
\geq &\alpha_t^0 \left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) + \sum_{i=1}^t \alpha_t^i \left\{F_{k_i,h+1} - F_{h+1}^\pi\right\}(x_{k_i,h+1}, \pi(x_{k_i,h+1})), \quad (2.48)
\end{aligned}$$

where inequality (a) holds because of the concentration result in Lemma A.4 in the appendix and

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

by using Lemma A.1-(c), and equality (b) holds because Triple-Q selects the action that maximizes $F_{k_i,h+1}(x_{k_i,h+1}, a)$ so $U_{k_i,h+1}(x_{k_i,h+1}) = \max_a F_{k_i,h+1}(x_{k_i,h+1}, a)$.

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

$$\left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) \geq 0,$$

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

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

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 (line 20 in Algorithm 1).

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 (2.48) becomes

$$\{F_{k,H} - F^{\pi}_{H}\}(x,a) \geq \alpha^0_t \left\{H + \frac{Z_1}{\eta}H - F^{\pi}_{H}\right\}(x,a) \geq 0, \tag{2.49}$$

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

$$\{F_{k,h} - F^{\pi}_{h}\}(x,a) \geq 0. \tag{2.50}$$

holds for any k, h, and (x,a), which completes the proof of (2.46).

Let $\mathcal{E}$ denote the event that (2.46) holds for all k, h and (x,a). Then based on Lemma A.2, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right)q^{\epsilon,*}_1\right\}(x_{k,1},a)\right] \\
=&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right)q^{\epsilon,*}_1\right\}(x_{k,1},a)\middle|\mathcal{E}\right]\Pr(\mathcal{E}) \\
&+\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right)q^{\epsilon,*}_1\right\}(x_{k,1},a)\middle|\mathcal{E}^c\right]\Pr(\mathcal{E}^c) \\
\leq&2K\left(1 + \frac{K^{1-\alpha}H^2\sqrt{\iota}}{\eta}\right)H^2\sqrt{\iota}\frac{1}{K^3} \leq \frac{4H^4\iota}{\eta K}.
\end{aligned} \tag{2.51}$$

□

Next, we bound (2.45) using the Lyapunov drift analysis on virtual queue Z. Since the virtual queue is updated every frame, we abuse the notation and define Z_T to be the virtual queue used in frame T. In particular, $Z_T = Z_{(T-1)K^\alpha+1}$. We further define

$$\bar{C}_T = \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} C_{k,1}(x_{k,1}, a_{k,1}). \tag{2.52}$$

Therefore, under Triple-Q, we have

$$Z_{T+1} = \left(Z_T + \rho + \epsilon - \frac{\bar{C}_T}{K^\alpha} \right)^+ \tag{2.53}$$

Define the Lyapunov function to be

$$L_T = \frac{1}{2} Z_T^2. \tag{2.54}$$

The next lemma bounds the expected Lyapunov drift conditioned on Z_T.

Lemma II.6. *Assume $\epsilon \leq \delta$. The expected Lyapunov drift satisfies*

$$\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\
\leq &\frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(-\eta \mathbb{E}\left[\sum_a \left\{ Q_{k,1} q_1^{\epsilon,*} \right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1}) \middle| Z_T = z \right] \right. \\
&\left. + z \mathbb{E}\left[\sum_a \left\{ (C_1^{\epsilon,*} - C_{k,1}) \, q_1^{\epsilon,*} \right\}(x_{k,1}, a) \middle| Z_T = z \right] \right) + H^4 \iota + \epsilon^2.
\end{aligned} \tag{2.55}$$

Proof. Based on the definition of L_T, the Lyapunov drift is

$$\begin{aligned}
L_{T+1} - L_T \leq &Z_T \left(\rho + \epsilon - \frac{\bar{C}_T}{K^\alpha} \right) + \frac{\left(\frac{\bar{C}_T}{K^\alpha} + \epsilon - \rho \right)^2}{2} \\
\leq &Z_T \left(\rho + \epsilon - \frac{\bar{C}_T}{K^\alpha} \right) + H^4 \iota + \epsilon^2
\end{aligned}$$

$$\leq \frac{Z_T}{K^\alpha} \sum_{k=TK^\alpha+1}^{(T+1)K^\alpha} (\rho + \epsilon - C_{k,1}(x_{k,1}, a_{k,1})) + H^4\iota + \epsilon^2$$

where the first inequality is a result of the upper bound on $|C_{k,1}(x_{k,1}, a_{k,1})|$ in Lemma A.2.

Let $\{q_h^\epsilon\}_{h=1}^H$ be a feasible solution to the tightened LP (2.31). Then the expected Lyapunov drift conditioned on $Z_T = z$ is

$$\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\
\leq &\frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} (\mathbb{E}\left[z\left(\rho + \epsilon - C_{k,1}(x_{k,1}, a_{k,1})\right) - \eta Q_{k,1}(x_{k,1}, a_{k,1})\right| Z_T = z] \\
&+ \eta \mathbb{E}\left[Q_{k,1}(x_{k,1}, a_{k,1})\right| Z_T = z]) + H^4\iota + \epsilon^2. \qquad (2.56)
\end{aligned}$$

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

$$\begin{aligned}
&(\mathbb{E}\left[z\left(\rho + \epsilon - C_{k,1}(x_{k,1}, a_{k,1})\right) - \eta Q_{k,1}(x_{k,1}, a_{k,1})\right| Z_T = z] \\
&+ \eta \mathbb{E}\left[Q_{k,1}(x_{k,1}, a_{k,1})\right| Z_T = z]) \\
\leq_{(a)} &z(\rho + \epsilon) - \mathbb{E}\left[\eta\left(\sum_a \left\{\frac{z}{\eta} C_{k,1} q_1^\epsilon + Q_{k,1} q_1^\epsilon\right\}(x_{k,1}, a)\right)\middle| Z_T = z\right] \\
&+ \eta \mathbb{E}\left[Q_{k,1}(x_{k,1}, a_{k,1})\right| Z_T = z] \\
= &\mathbb{E}\left[z\left(\rho + \epsilon - \sum_a C_{k,1}(x_{k,1}, a) q_1^\epsilon(x_{k,1}, a)\right)\middle| Z_T = z\right] \\
&- \mathbb{E}\left[\eta \sum_a Q_{k,1}(x_{k,1}, a) q_1^\epsilon(x_{k,1}, a) - \eta Q_{k,1}(x_{k,1}, a_{k,1})\middle| Z_T = z\right] \\
= &\mathbb{E}\left[z\left(\rho + \epsilon - \sum_a C_1^\epsilon(x_{k,1}, a) q_1^\epsilon(x_{k,1}, a)\right)\middle| Z_T = z\right] \\
&- \mathbb{E}\left[\eta \sum_a Q_{k,1}(x_{k,1}, a) q_1^\epsilon(x_{k,1}, a) - \eta Q_{k,1}(x_{k,1}, a_{k,1})\middle| Z_T = z\right] \\
&+ \mathbb{E}\left[z \sum_a \left\{(C_1^\epsilon - C_{k,1}) q_1^\epsilon\right\}(x_{k,1}, a)\middle| Z_T = z\right]
\end{aligned}$$

$$\leq -\eta\mathbb{E}\left[\sum_a Q_{k,1}(x_{k,1},a)q_1^\epsilon(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\middle| Z_T = z\right]$$
$$+\mathbb{E}\left[z\sum_a \left\{(C_1^\epsilon - C_{k,1})q_1^\epsilon\right\}(x_{k,1},a)\middle| Z_T = z\right],$$

where inequality (a) holds because $a_{k,h}$ is chosen to maximize $Q_{k,h}(x_{k,h},a)+\frac{Z_T}{\eta}C_{k,h}(x_{k,h},a)$ under Triple-Q, and the last equality holds due to that $\{q_h^\epsilon(x,a)\}_{h=1}^H$ is a feasible solution to the optimization problem (2.31), so

$$\left(\rho+\epsilon-\sum_a C_1^\epsilon(x_{k,1},a)q_1^\epsilon(x_{k,1},a)\right) = \left(\rho+\epsilon-\sum_{h,x,a} g_h(x,a)q_h^\epsilon(x,a)\right)\leq 0.$$

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

After taking expectation with respect to Z, dividing η on both sides, and then applying the telescoping sum, we obtain

$$\mathbb{E}\left[\sum_{k=1}^K\left(\sum_a\left\{Q_{k,1}q_1^{\epsilon,*}\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right]$$
$$+\mathbb{E}\left[\sum_{k=1}^K\frac{Z_k}{\eta}\sum_a\left\{(C_{k,1}-C_1^{\epsilon,*})\,q_1^{\epsilon,*}\right\}(x_{k,1},a)\right]$$
$$\leq\frac{K^\alpha\mathbb{E}\left[L_1 - L_{K^{1-\alpha}+1}\right]}{\eta}+\frac{K\left(H^4\iota+\epsilon^2\right)}{\eta}\leq\frac{K\left(H^4\iota+\epsilon^2\right)}{\eta}, \tag{2.57}$$

where the last inequality holds because that $L_1 = 0$ and L_{T+1} is non-negative.

Now combining Lemma II.5 and inequality (2.57), we conclude that

$$(2.34)\leq\frac{K\left(H^4\iota+\epsilon^2\right)}{\eta}+\frac{4H^4\iota}{\eta K}.$$

Further combining inequality above with Lemma II.3 and Lemma II.4,

$$\begin{aligned}\text{Regret}(K) \leq & \frac{KH\epsilon}{\delta} + H^2SAK^{1-\alpha} + \frac{H^3\sqrt{\iota}K}{\chi} \\ & + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)} + \frac{K\left(H^4\iota+\epsilon^2\right)}{\eta} + \frac{4H^4\iota}{\eta K}. \end{aligned} \tag{2.58}$$

By choosing $\alpha = 0.6$, i.e each frame has $K^{0.6}$ episodes, $\chi = K^{0.2}$, $\eta = K^{0.2}$, and $\epsilon = \frac{8\sqrt{SAH^6\iota^3}}{K^{0.2}}$, we conclude that when $K \geq \left(\frac{8\sqrt{SAH^6\iota^3}}{\delta}\right)^5$, which guarantees that $\epsilon < \delta/2$, we have

$$\text{Regret}(K) \leq \frac{13}{\delta}H^4\sqrt{SA\iota^3}K^{0.8} + \frac{4H^4\iota}{K^{1.2}} = \tilde{\mathcal{O}}\left(\frac{1}{\delta}H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{0.8}\right). \tag{2.59}$$

2.5.2 Constraint Violation

Outline of the Constraint Violation Analysis

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

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

which implies that

$$\begin{aligned}\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(-C_1^{\pi_k}(x_{k,1},a_{k,1}) + \rho\right) \leq & K^\alpha\left(Z_{T+1} - Z_T\right) \\ & + \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(\left\{C_{k,1} - C_1^{\pi_k}\right\}(x_{k,1},a_{k,1}) - \epsilon\right).\end{aligned}$$

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

$$\mathbb{E}\left[\sum_{k=1}^{K}\rho - C_1^{\pi_k}(x_{k,1},a_{k,1})\right] \leq -K\epsilon + K^{\alpha}\mathbb{E}\left[Z_{K^{1-\alpha}+1}\right] + \mathbb{E}\left[\sum_{k=1}^{K}\left\{C_{k,1}-C_1^{\pi_k}\right\}(x_{k,1},a_{k,1})\right], \tag{2.60}$$

where we used the fact $Z_1 = 0$.

In Lemma II.4, we already established an upper bound on the estimation error of $C_{k,1}$:

$$\mathbb{E}\left[\sum_{k=1}^{K}\left\{C_{k,1}-C_1^{\pi_k}\right\}(x_{k,1},a_{k,1})\right] \leq H^2SAK^{1-\alpha} + \frac{H^3\sqrt{\iota}K}{\chi} + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)}. \tag{2.61}$$

Next, we study the moment generating function of Z_T, i.e. $\mathbb{E}\left[e^{rZ_T}\right]$ for some $r > 0$. 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 $1 \leq T \leq K^{1-\alpha}+1$

$$\mathbb{E}[Z_T] \leq \frac{54H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right) + \frac{16H^2\iota}{K^2\delta} + \frac{4\eta\sqrt{H^2\iota}}{\delta}. \tag{2.62}$$

Under our choices of ϵ, α, χ, η and ι, it can be easily verified that $K\epsilon$ dominates the upper bounds in (2.61) and (2.62), which leads to the conclusion that the constraint violation because zero when K is sufficiently large in Theorem II.2.

Detailed Proof

To complete the proof, we need to establish the following upper bound on $\mathbb{E}[Z_{T+1}]$ based on a bound on the moment generating function.

Lemma II.7. *Assuming $\epsilon \leq \frac{\delta}{2}$, we have for any $1 \leq T \leq K^{1-\alpha}$*

$$\mathbb{E}[Z_T] \leq \frac{54H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right) + \frac{16H^2\iota}{K^2\delta} + \frac{4\eta\sqrt{H^2\iota}}{\delta}. \tag{2.63}$$

The proof will also use the following lemma from [62].

Lemma II.8. *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 γ and v with $0 < \gamma \leq v$, 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 II.7. We apply Lemma II.8 to a new Lyapunov function:

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

To verify condition (1) in Lemma II.8, consider $\bar{L}_T = Z_T \geq \theta_T = \frac{4(\frac{4H^2\iota}{K^2} + \eta\sqrt{H^2\iota} + H^4\iota + \epsilon^2)}{\delta}$ and $2\epsilon \leq \delta$. The conditional expected drift of

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

$$
\begin{aligned}
&\leq -\frac{\delta}{2} + \frac{\frac{4H^2\iota}{K^2} + \eta\sqrt{H^2\iota} + H^4\iota + \epsilon^2}{\theta_T} \\
&= -\frac{\delta}{4},
\end{aligned}
$$

where inequality (a) is obtained according to Lemma A.5; and the last inequality holds given $z \geq \theta_T$.

To verify condition (2) in Lemma II.8, we have

$$
Z_{T+1} - Z_T \leq |Z_{T+1} - Z_T| \leq \left|\rho + \epsilon - \bar{C}_T\right| \leq (H^2 + \sqrt{H^4\iota}) + \epsilon \leq 2\sqrt{H^4\iota},
$$

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

Now choose $\gamma = \frac{\delta}{4}$ and $v = 2\sqrt{H^4\iota}$. From Lemma II.8, we obtain

$$
\mathbb{E}\left[e^{rZ_T}\right] \leq e^{rZ_1} + \frac{2e^{r(v+\theta_T)}}{r\gamma}, \quad \text{where} \quad r = \frac{\gamma}{v^2 + v\gamma/3}. \tag{2.64}
$$

By Jensen's inequality, we have

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

which implies that

$$
\begin{aligned}
\mathbb{E}[Z_T] \leq& \frac{1}{r}\log\left(1 + \frac{2e^{r(v+\theta_T)}}{r\gamma}\right) \\
=& \frac{1}{r}\log\left(1 + \frac{6v^2 + 2v\gamma}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq& \frac{1}{r}\log\left(1 + \frac{8v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq& \frac{1}{r}\log\left(\frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq& \frac{4v^2}{3\gamma}\log\left(\frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq& \frac{3v^2}{\gamma}\log\left(\frac{2v}{\gamma}\right) + v + \theta_T
\end{aligned}
$$

$$
\begin{aligned}
\leq & \frac{3v^2}{\gamma}\log\left(\frac{2v}{\gamma}\right) + v + \frac{4(\frac{4H^2\iota}{K^2} + \eta\sqrt{H^2\iota} + H^4\iota + \epsilon^2)}{\delta} \\
= & \frac{48H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right) + 2\sqrt{H^4\iota} + \frac{4(\frac{4H^2\iota}{K^2} + \eta\sqrt{H^2\iota} + H^4\iota + \epsilon^2)}{\delta} \\
\leq & \frac{54H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right) + \frac{16H^2\iota}{K^2\delta} + \frac{4\eta\sqrt{H^2\iota}}{\delta} = \tilde{\mathcal{O}}\left(\frac{\eta H}{\delta}\right), \qquad (2.65)
\end{aligned}
$$

which completes the proof of Lemma II.7. □

Substituting the results from Lemmas II.4 and II.7 into (2.60), under assumption $K \geq \left(\frac{16\sqrt{SAH^6\iota^3}}{\delta}\right)^5$, which guarantees $\epsilon \leq \frac{\delta}{2}$. Then by using the facts that $\epsilon = \frac{8\sqrt{SAH^6\iota^3}}{K^{0.2}}$, we can easily verify that

$$
\text{Violation}(K) \leq \frac{54H^4\iota K^{0.6}}{\delta}\log\frac{16H^2\sqrt{\iota}}{\delta} + \frac{4\sqrt{H^2\iota}}{\delta}K^{0.8} - 5\sqrt{SAH^6\iota^3}K^{0.8}.
$$

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

$$
\text{Violation}(K) \leq \frac{54H^4\iota K^{0.6}}{\delta}\log\frac{16H^2\sqrt{\iota}}{\delta} - \sqrt{SAH^6\iota^3}K^{0.8} = 0.
$$

Now to prove the high probability bound, recall that from inequality (2.53), we have

$$
\sum_{k=1}^{K}\rho - C_1^{\pi_k}(x_{k,1}, a_{k,1}) \leq -K\epsilon + K^{\alpha}Z_{K^{1-\alpha}+1} + \sum_{k=1}^{K}\left\{C_{k,1} - C_1^{\pi_k}\right\}(x_{k,1}, a_{k,1}). \qquad (2.66)
$$

According to inequality (2.64), we have

$$
\mathbb{E}\left[e^{rZ_T}\right] \leq e^{rZ_1} + \frac{2e^{r(v+\theta_T)}}{r\gamma} \leq \frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)},
$$

which implies that

$$
\Pr\left(Z_T \geq \frac{1}{r}\log\left(\frac{11v^2}{3\gamma^2}\right) + 2(v+\theta_T)\right)
$$

$$
\begin{aligned}
&= \Pr(e^{rZ_T} \geq e^{\log\left(\frac{11v^2}{3\gamma^2}\right)+2r(v+\theta_T)}) \\
&\leq \frac{\mathbb{E}[e^{rZ_T}]}{\frac{11v^2}{3\gamma^2}e^{2r(v+\theta_T)}} \\
&\leq \frac{1}{e^{r(v+\theta_T)}} = \tilde{\mathcal{O}}\left(e^{-\eta}\right), \qquad (2.67)
\end{aligned}
$$

where the first inequality is from the Markov inequality.

In the proof of Lemma II.4, we have shown

$$
\begin{aligned}
&\left|\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} C_{k,h}(x_{k,h},a_{k,h}) - C_h^{\pi_k}(x_{k,h},a_{k,h})\right| \\
\leq &\left|\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} C_{k,h+1}(x_{k,h+1},a_{k,h+1}) - C_{h+1}^{\pi_k}(x_{k,h+1},a_{k,h+1})\right| \\
&+ \left|\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} (\hat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^{\pi_k}(x_{k,h},a_{k,h})\right| \\
&+ HSA + \frac{H^2\sqrt{\iota}K^\alpha}{\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)} \qquad (2.68)
\end{aligned}
$$

Following a similar proof as the proof of Lemma A.4, we can prove that

$$
\left|\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} (\hat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^{\pi_k}(x_{k,h},a_{k,h})\right| \leq \frac{1}{4}\sqrt{H^2\iota K^\alpha}
$$

holds with probability at least $1-\frac{1}{K^3}$. By iteratively using inequality (2.68) over h and by summing it over all frames, we conclude that with probability at at least $1-\frac{1}{K^2}$,

$$
\begin{aligned}
&\left|\sum_{k=1}^{K}\{C_{k,1} - C_1^{\pi_k}\}(x_{k,1},a_{k,1})\right| \\
\leq &K^{1-\alpha}H^2SA + \frac{H^3\sqrt{\iota}K}{\chi} + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)} + \frac{1}{4}\sqrt{H^4\iota K^{2-\alpha}} \\
\leq &4\sqrt{H^4SA\iota}K^{0.8}, \qquad (2.69)
\end{aligned}
$$

where the last inequality holds because $\alpha = 0.6$ and $\chi = K^{0.2}$.

Now, by combining inequalities (2.67) and (2.69), and using the union bound, we can show that when $K \geq \max\left\{\left(\frac{8\sqrt{SAH^6\iota^3}}{\delta}\right)^5, e^{\frac{1}{\delta}}\right\}$, with probability at least $1 - \tilde{\mathcal{O}}\left(e^{-K^{0.2}} + \frac{1}{K^2}\right)$

$$\begin{aligned}
&\sum_{k=1}^{K} \rho - C_1^{\pi_k}(x_{k,1}, a_{k,1}) \\
&\leq -K\epsilon + K^{\alpha}\left(\frac{1}{r}\log\left(\frac{11v^2}{3\gamma^2}\right) + 2(v + \theta_T)\right) + 4\sqrt{H^4SA\iota}K^{0.8} \\
&\leq -\sqrt{SAH^6\iota^3}K^{0.8} \leq 0, \qquad (2.70)
\end{aligned}$$

which completes the proof of our main result.

2.6 Summary

In this chapter, we considered CMDPs and proposed a model-free RL algorithm without a simulator, named Triple-Q. From a theoretical perspective, *Triple-Q* achieves sublinear regret and *zero* constraint violation. We believe it is the first *model-free* RL algorithm for CMDPs with provable sublinear regret without a simulator. From an algorithmic perspective, Triple-Q has similar computational complexity with SARSA, and can easily incorporate recent deep Q-learning algorithms to obtain a deep *Triple-Q* algorithm, which makes our method particularly appealing for complex and challenging CMDPs in practice. While we only considered a single constraint in the chapter, it is straightforward to extend the algorithm and the analysis to multiple constraints. Assuming there are J constraints in total, Triple-Q can maintain a virtual queue and a utility Q-function for each constraint, and then selects an action at each step by

solving the following problem:

$$\max_{a} \left(Q_h(x_h, a) + \frac{1}{\eta} \sum_{j=1}^{J} Z^{(j)} C_h^{(j)}(x_h, a) \right).$$

CHAPTER III

A Provably-Efficient Model-Free Algorithm for Infinite-Horizon Average-Reward CMDPs

3.1 Introduction

In the previous chapter, we discussed designing model-free algorithms in episodic finite-horizon CMDPs. In general, learning CMDPs in the infinite horizon average-reward setting, where the learner-environment interaction never ends or resets and the goal is to achieve optimal long-term average-reward under constraints, appears to be much more challenging.

For the episodic CMDPs, two very recent works [20; 28] show that sublinear regret bound and zero violation are possible for episodic CMDPs without simulators. In particular, [20] proposes a model-based algorithm and [29] presents a model-free algorithm, and [31] proves that it is possible to achieve zero violation during training given a safe baseline policy based on a model-based approach. Despite these significant developments, we seek to answer the following question:

Can we design efficient RL algorithms for infinite-horizon, average-reward CMDPs with provable regret and constraint violation guarantees?

We answer this question affirmatively in this chapter and present a model-free RL algorithm *Tiple-QA* that achieves sub-linear regret and zero constraint violation.

	Algorithm	Regret	Constraint Violations
Known Model	C-UCRL [13]	$\tilde{\mathcal{O}}(SA\sqrt{K^{1.5}})$	0
Model-based	UCRL-CMDP [14]	$\tilde{\mathcal{O}}(S\sqrt{A}K^{\frac{2}{3}})$	$\tilde{\mathcal{O}}(S\sqrt{A}K^{\frac{2}{3}})$
Known Model	CMDP-PSRL [32]	$\tilde{\mathcal{O}}(\text{poly}(SAD)\sqrt{K})$	$\tilde{\mathcal{O}}(\text{poly}(SA)\sqrt{K})$
Model-free	**Triple-QA**	$\tilde{\mathcal{O}}\left(\frac{\sqrt{SA}}{\delta}K^{\frac{5}{6}}\right)$	0

Table 3.1: Regrets and constraint violations of RL algorithms for infinite-horizon average-reward CMDPs. S is the number of states, A is the number of actions, K is the number of steps, D is the diameter of the CMDP whose definition can be found in the appendix, δ is the slackness that will be defined later (Eq. (3.13)), and poly(X) denotes a polynomial function of x. Throughout this chapter, we use the notation $\tilde{\mathcal{O}}$ to suppress log terms. $\tilde{O}(f(K))$ denotes $\mathcal{O}(f(K)\log^n K)$ with $n > 0$.

Table 3.1 compares the results in this chapter with those in the literature. We remark that the proposed algorithm synthesizes the Triple-Q algorithm in [29] for episodic CMDPs and Optimistic Q-Learning [45] that reduces the average-reward problem to a discounted reward problem.

3.2 Preliminaries

An infinite-horizon average-reward CMDP can be defined as $(\mathcal{S}, \mathcal{A}, r, g, p)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $r(g) : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the unknown reward (utility) function, and $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel such that $p(s'|s, a) := \mathbb{P}(s_{k+1} = s'|s_k = s, a_k = a)$ for $s_k \in \mathcal{S}, a_k \in \mathcal{A}$. A stationary policy is a mapping $\pi : \mathcal{S} \to \mathcal{A}$, the long-term average reward (reward rate) of a stationary policy π with initial state $s \in \mathcal{S}$ is defined as

$$J_r^\pi(s) := \lim_{K\to\infty} \frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K} r(s_k, \pi(s_k))\middle| s_1 = s\right], \tag{3.1}$$

and the long-term average utility (utility rate) is defined as

$$J_g^\pi(s) := \lim_{K\to\infty} \frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K} g(s_k, \pi(s_k))\middle| s_1 = s\right]. \tag{3.2}$$

We assume that under any stationary policy, s_k is an irreducible aperiodic Markov chain, so it has a unique stationary distribution, and the limits above are well-defined. Letting s_∞^π denote the Markov chain at steady-state under policy π, we have

$$J_r^\pi = \mathbb{E}\left[r(s_\infty^\pi, \pi(s_\infty^\pi))\right] \text{ and } \; J_g^\pi = \mathbb{E}\left[g(s_\infty^\pi, \pi(s_\infty^\pi))\right], \tag{3.3}$$

where we removed the dependence on the initial condition s because the stationary distribution is independent of the initial condition for a finite-state, irreducible, and aperiodic Markov chain.

An optimal stationary policy π^* is defined to be the solution to the following problem:

$$\max_\pi J_r^\pi \quad \text{s.t.} \quad J_g^\pi \geq \rho. \tag{3.4}$$

We consider a constrained RL problem with K steps. At each step k, the agent observes state s_k, takes an action a_k, and receives reward $r(s_k, a_k)$ and utility $g(s_k, a_k)$. The next state s_{k+1} is sampled according to the probability distribution $p(\cdot|s_k, a_k)$. Our goal is to develop an online RL algorithm, which may be nonstationary, that minimizes both the regret and the constraint violation defined below.

$$\text{Regert}(K) = \mathbb{E}\left[\sum_{k=1}^{K}\left(J_r^{\pi^*} - r(s_k, a_k)\right)\right], \tag{3.5}$$

$$\text{Violation}(K) = \mathbb{E}\left[\sum_{k=1}^{K}\left(\rho - g(s_k, a_k)\right)\right]. \tag{3.6}$$

When the transition kernel $p(s'|s, a)$ is known, the optimal stationary policy that

solves problem (3.4) can be obtained by solving the following LP problem [11]:

$$\max_{\{q(s,a):(s,a)\in\mathcal{S}\times\mathcal{A}\}} \sum_{s,a} q(s,a)r(s,a) \tag{3.7}$$

$$\text{s.t.} \sum_{s,a} q(s,a)g(s,a) \geq \rho, \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{3.8}$$

$$q(s,a) \geq 0, \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{3.9}$$

$$\sum_{s,a} q(s,a) = 1 \tag{3.10}$$

$$\sum_{a} q(s,a) = \sum_{s',a'} p(s|s',a')q(s',a'), \tag{3.11}$$

where the $q(s,a)$ is called the occupancy measure, which is defined as the set of distributions generated by executing the associated induced policy π in the infinite-horizon CMDP. $\sum_a q(s,a)$ represents the probability the system is in state s, and $\frac{q(s,a)}{\sum_{a'} q(s,a')}$ is the probability of taking action a in state s. The utility constraint is represented in (3.8). More details can be found in [11].

To analyze the performance of our algorithm, we need to consider a tightened version of the above LP problem later, which is defined below:

$$\max_{\{q(s,a):(s,a)\in\mathcal{S}\times\mathcal{A}\}} \sum_{s,a} q(s,a)r(s,a) \tag{3.12}$$

$$\text{s.t.} \sum_{s,a} q(s,a)g(s,a) \geq \rho + \epsilon, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

$$(3.9) - (3.11),$$

where $\epsilon > 0$ is called a tightness constant. As in previous works [19; 25; 17; 53], we make the following standard assumption of Slater's condition.

Assumption III.1. *(Slater's Condition). There exist $\delta > 0$ and a feasible solution*

$q(s,a)$ *to the LP such that*

$$\sum_{s,a} q(s,a)g(s,a) - \rho \geq \delta. \tag{3.13}$$

It is obvious that when $\epsilon < \delta$ the problem (3.12) has a feasible solution due to Slater's condition. The Slater's condition is commonly assumed in previous works to ensure the LP problem has strong duality, see proofs in [53; 63]. Unlike [19; 17], which assume δ is known, and [54; 20; 31], where a strictly feasible policy is given, our assumption is less restrictive. Let

$$J_r^* = \sum_{s,a} q^*(s,a)r(s,a), \tag{3.14}$$

$$J_g^* = \sum_{s,a} q^*(s,a)g(s,a). \tag{3.15}$$

be the optimal reward rate and utility rate, where $q^*(s,a)$ is the optimal solution obtained by solving the LP problem (3.7). Moreover it is obvious that J_r^* and J_g^* are independent of the initial state and we have $J_r^* = J_r^{\pi^*}$ and $J_g^* = J_g^{\pi^*}$.

In the following, we use superscript * to denote the optimal policy achieved by solving the LP (3.7) of the original CMDP, and superscript $^{\epsilon,*}$ to denote the optimal policy related to the ϵ-tightened version of LP (3.12).

3.3 Algorithm

In this section, we introduce our algorithm Triple-QA (see Algorithm 3 for pseudo-code) which achieves sub-linear regret and zero constraint violation. The algorithm is inspired by Triple-Q, an algorithm for episodic CMDPs in [28]. Triple-QA is for the infinite-horizon average-reward CMDPs with a different update rule. The algorithm further solves the discounted CMDPs with the discount factor γ close to one, an idea used in [45]. The discounted CMDP is defined on the same state space, action space,

reward/utility functions, the transition kernel. The intuition is that the reward of the discounted problem (scaled by $1-\gamma$)) approaches to that of the average reward problem as γ goes to 1.

Algorithm 3: Triple-QA

Initialize $Q_1(s,a) = \hat{Q}_1(s,a) \leftarrow H = K^{\frac{1}{6}}$
$n_1(s,a) \leftarrow 0, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \gamma = 1 - \frac{1}{H}, \hat{V}_1(s) = H, \forall s \in \mathcal{S}$;
Choose $\chi = K^{\frac{1}{3}}, \eta = K^{\frac{1}{6}}, \iota = 8\log(\sqrt{2}K), \beta = \frac{2}{3}$.;
Choose $\epsilon = \frac{9\kappa\sqrt{SA\iota}}{K^{\frac{1}{6}}}, \kappa(Eq.(3.29))\}$. ;
Initialize $\bar{C} \leftarrow 0, Z_1 \leftarrow 0$. ;
Define , $\alpha_\tau = \frac{\chi+1}{\chi+\tau}, b_\tau = \kappa\sqrt{\frac{(\chi+1)\iota}{\chi+\tau}}$. ;
for *episode* $k = 1, \ldots, K$ **do**
  Take $a_k = \arg\max_a \left(\hat{Q}_k(s_k,a) + \frac{Z}{\eta}\hat{C}_k(s_k,a)\right)$.;
  Observe s_{k+1}.;
  $n_{k+1}(s_k,a_k) \leftarrow n_k(s_k,a_k) + 1, \tau \leftarrow n_{k+1}(s_k,a_k)$.;
  Update $Q_{k+1}(s_k,a_k) \leftarrow (1-\alpha_\tau)Q_k(s_k,a_k) + \alpha_\tau[r(s_k,a_k) + \gamma\hat{V}_k(s_{k+1}) + b_\tau]$,;
  Update $C_{k+1}(s_k,a_k) \leftarrow (1-\alpha_\tau)C_k(s_k,a_k) + \alpha_\tau[g(s_k,a_k) + \gamma\hat{W}_k(s_{k+1}) + b_\tau]$.;
  if $Q_{k+1}(s_k,a_k) \leq \hat{Q}_k(s_k,a_k)$ *and* $C_{k+1}(s_k,a_k) \leq \hat{C}_k(s_k,a_k)$ **then**
    $\hat{Q}_{k+1}(s_k,a_k) \leftarrow Q_{k+1}(s_k,a_k)$;
    $\hat{C}_{k+1}(s_k,a_k) \leftarrow C_{k+1}(s_k,a_k)$;
  else
    $\hat{Q}_{k+1}(s_k,a_k) \leftarrow \hat{Q}_k(s_k,a_k)$;
    $\hat{C}_{k+1}(s_k,a_k) \leftarrow \hat{C}_k(s_k,a_k)$;
  $\bar{C} \leftarrow \bar{C} + (1-\gamma)\hat{C}_k(s_k,a_k)$;
  $a' = \arg\max_a \left(\hat{Q}_{k+1}(s_k,a) + \frac{Z}{\eta}\hat{C}_{k+1}(s_k,a)\right)$;
  $\hat{V}_{k+1}(s_k) \leftarrow \hat{Q}_{k+1}(s_k,a')$;
  $\hat{W}_{k+1}(s_k) \leftarrow \hat{C}_{k+1}(s_k,a')$;
  if k *mod* $K^\beta = 0$ **then**
    $Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}}{K^\beta}\right)$;
    Reset $\bar{C} \leftarrow 0, n_t(s,a) \leftarrow 0$.;
    Reset $\hat{Q}_{k+1}(s,a), Q_{k+1}(s,a), V_{k+1}(s)$ to H ;
    Reset $\hat{C}_{k+1}(s,a), C_{k+1}(s,a), W_{k+1}(s)$ to H;

Under the discounted CMDP setting, given a policy π, the reward value function

V_k^π at step k is the expected cumulative rewards from step k under policy π :

$$V_k^\pi(s) = \mathbb{E}\left[\sum_{i=k}^{\infty} \gamma^{i-k} r(s_i, \pi(s_i)) \middle| s_k = s\right]. \tag{3.16}$$

The reward Q-function $Q_k^\pi(s,a)$ at step k is the expected cumulative reward when an agent starts from a state-action pair (s,a) at step k and then follows policy π :

$$Q_k^\pi(s,a) = r(s,a) + \mathbb{E}\left[\sum_{i=k}^{\infty} \gamma^{i-k} r(s_i, \pi(s_i)) \middle| \ s_k = s, a_k = a \ \right]. \tag{3.17}$$

Similarly, we use $W_k^\pi(s) : \mathcal{S} \to \mathbb{R}^+$ and $C_k^\pi(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ to denote the utility value function and utility Q-function at step k:

$$W_k^\pi(x) = \mathbb{E}\left[\sum_{i=k}^{\infty} \gamma^{i-k} g(s_i, \pi_i(s_i)) \middle| s_k = s\right], \tag{3.18}$$

$$C_k^\pi(s,a) = g(s,a) + \mathbb{E}\left[\sum_{i=k}^{\infty} \gamma^{i-k} g(s_i, \pi(s_i)) \middle| \ s_k = s, a_k = a \ \right]. \tag{3.19}$$

It is obvious that all the reward and utility value (Q-value) functions are bounded by $\frac{1}{1-\gamma}$ because the reward and utility are bounded by 1. We define $H = \frac{1}{1-\gamma}$. Then given a state-action pair (s,a) at step k, our algorithm updates the estimate of reward (utility) $Q-$value functions of the discounted CMDP setting instead.

The design of the algorithm is based on the primal-dual approach for constrained optimization problems. Suppose that $V^\pi(s)$ $(W^\pi(s))$ is an accurate estimate of $\frac{J_r^\pi}{1-\gamma}$ $\left(\frac{J_g^\pi}{1-\gamma}\right)$. The formal proof is deferred to the next section. Given Lagrangian multiplier μ, we consider the following problem:

$$\max_\pi J_r^\pi(s) + \mu(J_g^\pi(s) - \rho) \approx \max_\pi (1-\gamma)(V^\pi(s) + \mu W^\pi(s)) - \mu\rho \tag{3.20}$$

which can be interpreted as an unconstrained MDP with a modified reward function $(1-\gamma)(r+\mu g)$.

The algorithm is an extension of Triple-Q [28] for episodic CMDPs by including the discount factor and replacing episode-by-episode updates with step-by-step updates. We adopt the same notations used in [45]. Same as Triple-Q, the algorithm maintains an estimate $\hat{V}_k(s)$ $(\hat{W}_k(s))$ for the optimal value function $V^*(s)(W^*(s))$ and $\hat{Q}_k(s,a)$ $(\hat{C}_k(s,a))$ for the optimal Q-function $Q^*(s,a)$ $(C^*(s,a))$. At each step k, after observing state s, the agent selects action a_k^* based on the combined Q-value:

$$a_k^* \in \arg\max_a \hat{Q}_k(s,a) + \frac{Z}{\eta}\hat{C}_k(s,a), \tag{3.21}$$

where $\frac{Z}{\eta}$ can be treated as an estimate of the Lagrange multiplier μ. Similar to [28], we need to carefully tune the frequency of updating the Lagrange multiplier to balance the convergence and optimality. Updating it too frequently would lead to divergence and too infrequent would result in a large regret and large constraint violation. The algorithm tackles this difficulty by updating Z at a slow time-scale, i.e., every K^β steps in line $25-26$ in Algorithm 3, with the following update

$$Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}}{K^\beta}\right)^+, \tag{3.22}$$

where $(x)^+ = \max\{x, 0\}$, and $\bar{C}$ is the summation of all $(1-\gamma)\hat{C}_k(s_k,a_k)$ of the steps in the previous frame, where each frame consists of K^β consecutive steps.

During each frame, the algorithm learns the combined Q functions for fixed Z at a fast time scale. The estimates of reward and utility value functions are updated after observing a new state-action pair.

It is important to note that for a CMDP,

$$V^*(s) \neq \max_a Q^*(s,a). \tag{3.23}$$

This means optimistic Q-learning algorithms for unconstrained MDPs (e.g. [34; 45; 64]) cannot be used for estimating the optimal value functions of CMDPs. Instead, Algorithm 3 uses a SARSA-type updating rule, as shown in lines $11-14$.

We note that the optimal policy for a CMDP is stochastic in general. The policy under our algorithm is a stochastic policy because the virtual queue Z varies during and after the learning process, which results in a stochastic policy.

We further introduce additional notations before presenting our main theorem. Let $v^\pi(s)$ and $w^\pi(s)$ denote the reward and utility relative value functions for state s under average-reward setting, and $q^\pi(s,a), c^\pi(s,a)$ be the reward and utility Q value functions for any state-action pair (s,a). Based on the Bellman equation, we have

$$J_r^\pi + q^\pi(s,a) = r(s,a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[v^\pi(s')] \tag{3.24}$$

$$v^\pi(s) = \sum_a q^\pi(s,a)\mathbb{P}(\pi(s)=a) \tag{3.25}$$

$$J_g^\pi + c^\pi(s,a) = g(s,a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[w^\pi(s')] \tag{3.26}$$

$$w^\pi(s) = \sum_a c^\pi(s,a)\mathbb{P}(\pi(s)=a) \tag{3.27}$$

Define

$$sp(f) = \max_{s\in\mathcal{S}} f(s) - \min_{s\in\mathcal{S}} f(s) \tag{3.28}$$

to be the span of the function f. It is well known that the span of the optimal reward relative value function $sp(v^*)$ and utility relative value function $sp(w^*)$ are bounded for weakly communication or ergodic MDPs. In particular, they are bounded by the diameter of the MDP [65].

Let

$$\kappa = \max_{0\le\epsilon\le\rho/2} (\max\{sp(v^{\epsilon,*}), sp(w^{\epsilon,*}), 1\}) \tag{3.29}$$

and assume that κ which is used in the algorithm is known beforehand as in [45; 66].

We can always substitute them with any upper bound (e.g. the diameter) when it is unknown.

3.3.1 Main Results

We now state the main results in the following theorem.

Theorem III.2. *Assume* $K \geq \left(\frac{18\kappa\sqrt{SA\iota}}{\delta}\right)^6$ *and let* $\epsilon = \frac{9\kappa\sqrt{SA\iota}}{K^{\frac{1}{6}}}$ *such that* $\epsilon \leq \frac{\delta}{2}$. *By choosing* $m = K^{\frac{1}{6}} \log K$, $H = K^{\frac{1}{6}}$, $\eta = K^{\frac{1}{6}}$, $\chi = K^{\frac{1}{3}}$, *and* $\beta = \frac{2}{3}$, *Algorithm 3 guarantees*

$$\text{Regret}(K) \leq \tilde{O}\left(\frac{\sqrt{SA}\kappa}{\delta} K^{\frac{5}{6}}\right) \tag{3.30}$$

$$\text{Violation}(K) \leq \frac{92K^{\frac{2}{3}}}{\delta} \log\left(\frac{24}{\delta}\right) - \sqrt{SA\iota} K^{\frac{5}{6}} = 0, \tag{3.31}$$

where $\iota = 32\log(\sqrt{2}K)$. □

3.3.2 The Choices of the Hyper-parameters

The regret bound and constraint violation bound are

$$\text{Regret}(K) = \tilde{\mathcal{O}}\left(K\epsilon + \gamma^m K + \frac{Km}{\chi} + \sqrt{K^{2-\beta}\chi} + mK^{1-\beta} + \frac{K}{\eta} + \frac{K}{H}\right) \tag{3.32}$$

$$\text{Violation}(K) = -K\epsilon + \tilde{\mathcal{O}}\left(K^{\beta}\eta + \frac{Km}{\chi} + \sqrt{K^{2-\beta}\chi} + mK^{1-\beta}\right). \tag{3.33}$$

We need to choose all the parameters ϵ, η, m, β, χ, and H carefully in order to balance each term and all the parameters should be functions of K. Let $\chi = K^{\zeta}$ and $m = \tilde{O}(K^{\nu})$. We have $\beta = 3\zeta - 2\nu$ in order to ensure $\frac{Km}{\chi}$ and $\sqrt{K^{2-\beta}\chi}$ are of the same order. Since m and H are of the same order, substituting ζ and ν yields

$$\text{Regret}(K) = \tilde{\mathcal{O}}\left(K\epsilon + K^{1-\zeta+\nu} + K^{1-3\zeta+3\nu} + \frac{K}{\eta} + K^{1-\nu}\right) \tag{3.34}$$

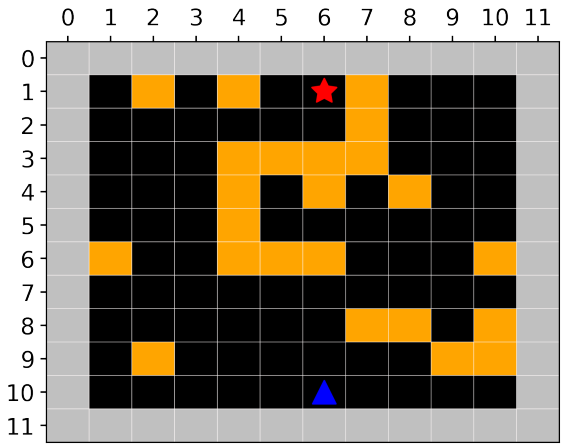

Figure 3.1: A Grid World with Safety Constraints

$$\text{Violation}(K) = -K\epsilon + \tilde{\mathcal{O}}\left(K^{3\zeta-2\nu}\eta + K^{1-\zeta+\nu} + K^{1-3\zeta+3\nu}\right), \tag{3.35}$$

where the term $\gamma^m K$ is omitted because the choice of m ensures $\gamma^m \leq \frac{1}{K}$ (Eq.(3.66)). To make sure $1 > 1-\zeta+\nu > 0$, we need to have $\nu < \zeta$. Then the bounds become

$$\text{Regret}(K) = \tilde{\mathcal{O}}\left(K\epsilon + K^{1-\zeta+\nu} + \frac{K}{\eta} + K^{1-\nu}\right) \tag{3.36}$$

$$\text{Violation}(K) = -K\epsilon + \tilde{\mathcal{O}}\left(K^{3\zeta-2\nu}\eta + K^{1-\zeta+\nu}\right). \tag{3.37}$$

To guarantee zero violation, $K\epsilon$, $K^{3\zeta-2\nu}\eta$, and $K^{1-\zeta+\nu}$ should be of the same order, which means $\epsilon = \tilde{O}(K^{-\zeta+\nu})$ and $\eta = K^{1-4\zeta+3\nu}$. To optimize the regret bound, we need to balance $K^{1-\zeta+\nu}$, $\frac{K}{\eta} = K^{4\zeta-3\nu}$ and $K^{1-\nu}$. Solving the equations we finally have $\zeta = \frac{1}{3}$, $\nu = \frac{1}{6}$, $\beta = 3\zeta - 2\nu = \frac{2}{3}$, which leads to the choices of $\chi = K^{\frac{1}{3}}$, $m = \tilde{O}(K^{\frac{1}{6}})$, $H = K^{\frac{1}{6}}$, $\epsilon = \tilde{O}(K^{-\frac{1}{6}})$, and $\eta = \tilde{O}(K^{\frac{1}{6}})$.

3.4 Simulation

In this section, we present simulation results that evaluate our algorithm using the 2D safety grid-world exploration problem [13; 67]. Figure 3.1 shows the map of a 10×10 grid-world with a total of 100 states. We chose an error probability 0.03 which

means with probability 0.03 the agent will choose an action uniformly at random to make the environment stochastic. The objective of the agent is to travel to the destination (the red star) from the original position (the blue triangle) as quickly as possible while limiting the number of times hitting the obstacles (the yellow squares). Hitting an obstacle incurs a cost 1 and otherwise, there is no cost. The reward for the destination is 1, and for others the normalized Euclidean distance between them and the destination times a scaled factor of 0.1. We set the constraint limit as 0.15 through the simulation which means the expected cost rate should be below the limit. To account for statistical significance, the results of each experiment are averaged over 5 trials. We remark that in the simulation we consider the following constraint

$$\liminf_{K\to\infty} \frac{1}{K} \mathbb{E}_\pi \left[\sum_{k=1}^{K} g(s_k, a_k) \right] \leq \rho, \tag{3.38}$$

which is similar to the constraint that the average utility needs to be above a threshold.

Figure 3.2 shows the performance comparison of our algorithm in terms of average reward and average cost during training compared with the algorithm in [45]. We can see that our algorithm is able to learn a policy that achieves a high reward while satisfying the safety constraint very quickly. The optimistic Q-learning algorithm [45] was for unconstrained MDPs, so it yields a higher reward but also violates the safety constraint.

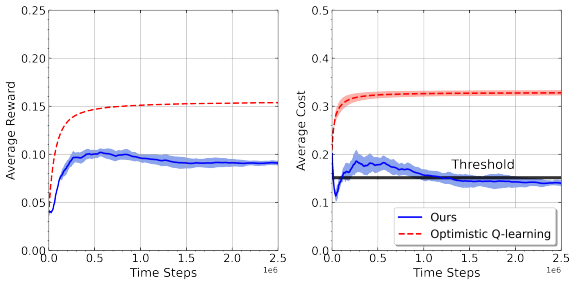

Figure 3.2: Average reward and cost of our algorithm and Optimistic Q-learning during training. The shaded region represents the standard deviations.

3.5 Proof of the Main Theorem

In the following, we use shorthand notation

$$\{f - g\}(x) = f(x) - g(x),$$

where $f(\cdot)$ and $g(\cdot)$ the the same argument value. Similarly,

$$\{(f - g)q\}(x) = (f(x) - g(x))q(x).$$

3.5.1 Regret Analysis

We start the proof by adding and subtracting the corresponding terms to the regret defined in (3.5), and we obtain

$$\begin{aligned}
&\text{Regret}(K) = \mathbb{E}\left[\sum_{k=1}^{K}(J_r^* - r(s_k, a_k))\right] \\
=&\mathbb{E}\left[\sum_{k=1}^{K}(J_r^* - J_r^{\epsilon,*})\right] && (3.39)\\
+&\mathbb{E}\left[\sum_{k=1}^{K}(J_r^{\epsilon,*} - (1-\gamma)V^{\epsilon,*}(s_k))\right] && (3.40)\\
+&\mathbb{E}\left[\sum_{k=1}^{K}(1-\gamma)\left(V^{\epsilon,*}(s_k) - \hat{Q}_k(s_k, a_k)\right)\right] && (3.41)\\
+&\mathbb{E}\left[\sum_{k=1}^{K}\left((1-\gamma)\hat{Q}_k(s_k, a_k) - r(s_k, a_k)\right)\right]. && (3.42)
\end{aligned}$$

We will bound each of the four terms above in the following sequence of lemmas.

The term (3.39) is the difference between the original CMDP and its corresponding $\epsilon-$tighten version, which is a perturbation of the original problem. We establish a bound by using the following lemma.

Lemma III.3. *Under assumption III.1, given $\epsilon \leq \delta$, we have*

$$\sum_{t=1}^{K}(J_r^* - J_r^{\epsilon,*}) \leq \frac{\epsilon K}{\delta} \tag{3.43}$$

For the second term (3.40), we establish a bound by using Lemma III.4, which shows the difference between value functions of the average-reward problem and the value functions of the discounted setting problem is small. The proof is based on the Bellman equations under the two settings. The proof follows Lemma 2 in [45] closely.

Lemma III.4. *For an arbitrary policy π, we have*

$$J_r^\pi - (1-\gamma)V^\pi(s) \leq (1-\gamma)sp(v^\pi(s)), \tag{3.44}$$

$$|V^\pi(s_1) - V^\pi(s_2)| \leq 2sp(v^\pi(s)); \tag{3.45}$$

$$J_g^\pi - (1-\gamma)W^\pi(s) \leq (1-\gamma)sp(w^\pi(s)), \tag{3.46}$$

$$|W^\pi(s_1) - W^\pi(s_2)| \leq 2sp(w^\pi(s)), \tag{3.47}$$

where $V^\pi(s)$ is the value function for the discounted setting under policy π, and $J_r^\pi(J_g^\pi)$ is the reward (utility) rate under policy π.

Then it is easy to obtain

$$J_r^{\epsilon,*} - (1-\gamma)V^{\epsilon,*}(s) \leq (1-\gamma)\kappa, \tag{3.48}$$

Next, we establish a bound on term (3.41) by using the Lyapunov-drift analysis. In unconstrained MDPs, the bound is established by showing that optimistic Q-learning guarantees that $\hat{Q}_k(s,a)$ is an overestimate of $Q^*(s,a)$. However this does not hold in CMDPs because the algorithm needs to consider reward and utility simultaneously so $\hat{Q}_k(s,a)$ is not necessarily an overestimate of $Q^*(s,a)$. To bound this term, we first

add and subtract some additional terms to obtain

$$\begin{aligned}
&\sum_{k=1}^{K}(1-\gamma)\left(V^{\epsilon,*}(s_k)-\hat{Q}_k(s_k,a_k)\right)\\
=&\sum_{k=1}^{K}(1-\gamma)\sum_a\left\{Q^{\epsilon,*}q^{\epsilon,*}+\frac{Z_k}{\eta}C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a) &(3.49)\\
-&\sum_{k=1}^{K}(1-\gamma)\sum_a\left\{\hat{Q}_k q^{\epsilon,*}+\frac{Z_k}{\eta}\hat{C}_k q^{\epsilon,*}\right\}(s_k,a) &(3.50)\\
+&\sum_{k=1}^{K}(1-\gamma)\left(\sum_a\left\{\hat{Q}_k q^{\epsilon,*}\right\}(s_k,a)-\hat{Q}_k(s_k,a_k)\right. &(3.51)\\
&\left.+\frac{Z_k}{\eta}\sum_a\left\{\hat{C}_k q^{\epsilon,*}-C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a)\right). &(3.52)
\end{aligned}$$

We can see (3.49) + (3.50) is the difference between the two combined Q functions. We will show that $\left\{\hat{Q}_k+\frac{Z_k}{\eta}\hat{C}_{k,h}\right\}(s,a)$ is always an over-estimate of $\left\{Q^{\epsilon,*}+\frac{Z_k}{\eta}C^{\epsilon,*}\right\}(s,a)$ (i.e. (3.49) + (3.50) ≤ 0) for all (s,a,k) simultaneously with a high probability in Lemma III.5. This result further implies an upper bound in expectation

$$\mathbb{E}\left[(3.49)+(3.50)\right]\leq(1-\gamma)\frac{3H}{\eta K}. \tag{3.53}$$

Lemma III.5. *With probability at least $1-\frac{1}{K^3}$, the following inequality holds simultaneously for all $(s,a,k)\in\mathcal{S}\times\mathcal{A}\times[K]$:*

$$\left\{\left(\hat{Q}_k-Q^{\epsilon,*}\right)+\frac{Z_k}{\eta}\left(\hat{C}_k-C^{\epsilon,*}\right)\right\}(s,a)\geq 0, \tag{3.54}$$

Then for the term (3.51) + (3.52), we can bound it by using the following lemma.

Lemma III.6. *Assuming $\epsilon<\delta$, we have*

$$\mathbb{E}\left[\sum_{k=1}^{K}(1-\gamma)\left(\sum_a\left\{\hat{Q}_k q^{\epsilon,*}\right\}(s_k,a)-\hat{Q}_k(s_k,a_k)\right.\right.$$

$$+\frac{Z_k}{\eta}\sum_a\left\{\hat{C}_k q^{\epsilon,*}-C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a)\Bigg)\Bigg]$$
$$\leq\frac{2K}{\eta}+\sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{\eta}. \tag{3.55}$$

To see the idea behind Lemma III.6, we need to consider the Lyapunov function $L_T=\frac{1}{2}Z_T^2$, where T is the frame index and Z_T is the virtual-queue length at the beginning of Tth frame. Recall that each frame contains K^β consecutive steps. In the proof of Lemma III.6, we will show that the Lyapunov-drift satisfies

$$\mathbb{E}[L_{T+1}-L_T]\leq \text{a negative drift}$$
$$+2+\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{K^\beta}-\frac{\eta}{K^\beta}\sum_{k=TK^\beta+1}^{(T+1)K^\beta}\Phi_k, \tag{3.56}$$

where

$$\Phi_k=(1-\gamma)\left(\sum_a\left\{\hat{Q}_k q^{\epsilon,*}\right\}(s_k,a)-\hat{Q}_k(s_k,a_k)\right.$$
$$\left.+\frac{Z_k}{\eta}\sum_a\left\{\hat{C}_k q^{\epsilon,*}-C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a)\right). \tag{3.57}$$

Then summing both sides of the equation overall $K^{1-\beta}$ frames, we can obtain

$$\mathbb{E}[L_1-L_{K^{1-\beta}+1}]$$
$$\leq 2K^{1-\beta}+\sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{K^\beta}-\frac{\eta}{K^\beta}\sum_k\Phi_k. \tag{3.58}$$

Therefore

$$(3.51)+(3.52)=\sum_k\Phi_k$$

$$\leq \frac{K^\beta \mathbb{E}[L_1 - L_{K^{1-\beta}+1}]}{\eta} + \frac{2K}{\eta} + \sum_{T=1}^{K^{1-\beta}} \mathbb{E}\left[Z_T\right] \frac{(1-\gamma)\kappa}{\eta}$$
$$\leq \frac{2K}{\eta} + \sum_{T=1}^{K^{1-\beta}} \mathbb{E}\left[Z_T\right] \frac{(1-\gamma)\kappa}{\eta}, \tag{3.59}$$

where the last inequality holds because $L_1 = 0$ and $L_T \geq 0$ for all T.

Then combining the result form (3.53) and Lemma III.6, we can obtain

$$\sum_{k=1}^{K}((1-\gamma)\left(V^{\epsilon,*}(s_k) - \hat{Q}_k(s_k, a_k)\right))$$
$$\leq \frac{2K}{\eta} + \sum_{T=1}^{K^{1-\beta}} \mathbb{E}\left[Z_T\right] \frac{(1-\gamma)\kappa}{\eta} + \frac{3H}{\eta K}. \tag{3.60}$$

The term $\mathbb{E}[Z_T]$ is proved uniformly bounded in Lemma III.7 by using the Lyapunov-drift analysis on the moment generating function of Z i.e. $\mathbb{E}[e^r Z]$ can be bounded by a constant uniformly over the entire learning horizon. The reason is that when the virtual queue Z is large, our algorithm takes actions to almost greedily reduce the virtual queue.

Lemma III.7. *Assuming $\epsilon \leq \frac{\delta}{2}$ and $H \geq \frac{6\kappa}{\delta}$, we have for any $1 \leq T \leq K^{1-\beta}$,*

$$\mathbb{E}[Z_T] \leq \frac{92}{\delta} \log\left(\frac{24}{\delta}\right) + \frac{6\eta}{\delta}. \tag{3.61}$$

We apply the following lemma to bound the last term (3.42).

Lemma III.8. *For any $T \in [K^{1-\beta}]$ and any $m \in \mathbb{Z}^+$,*

$$\mathbb{E}\left[\sum_{k=(T-1)K^\beta+1}^{TK^\beta} \left(\left\{(1-\gamma)\hat{Q}_k - r\right\}(s_k, a_k)\right)\right] \leq 2mS$$
$$+ \gamma^m K^\beta + \frac{K^\beta m}{\chi} + 4(1-\gamma)m\kappa\sqrt{(\chi+1)SAK^\beta\iota} \tag{3.62}$$
$$\mathbb{E}\left[\sum_{k=(T-1)K^\beta+1}^{TK^\beta} \left(\left\{(1-\gamma)\hat{C}_k - g\right\}(s_k, a_k)\right)\right] \leq 2mS$$

$$+\gamma^m K^\beta + \frac{K^\beta m}{\chi} + 4(1-\gamma)m\kappa\sqrt{(\chi+1)SAK^\beta\iota}. \tag{3.63}$$

This lemma is one of our key technical contributions, which shows that the cumulative estimation error over one frame (K^β consecutive episodes) between weighted reward(utility) Q-value functions and average reward (utility) is upper bounded. From the lemma above, we can immediately conclude that:

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(\left\{(1-\gamma)\hat{Q}_t - r\right\}(s_k,a_k)\right)\right] \leq \gamma^m K + \frac{Km}{\chi}$$
$$+ 4(1-\gamma)m\kappa\sqrt{(\chi+1)SAK^{2-\beta}\iota} + 2mSK^{1-\beta} \tag{3.64}$$

To balance the terms in regret, we carefully select that

$$m = H\log K = K^{\frac{1}{6}}\log K,\ \chi = K^{\frac{1}{3}},\ \beta = \frac{2}{3}. \tag{3.65}$$

Then we have

$$\gamma^m = \left(1-\frac{1}{H}\right)^{H\log K} \leq \frac{1}{K}, \tag{3.66}$$

and the order of the second and third terms in the above equation (3.64) is $\tilde{O}(K^{\frac{5}{6}})$, which is also the dominant term in our regret bound.

Then by appropriately choosing other parameters ϵ, ι and η, to balance the terms and combining the results from (3.60), (3.64), Lemma III.3, Lemma III.4, and Lemma III.7, we finish the proof for the regret bound.

3.5.2 Constraint Violation Analysis

Recall that we use Z_T to denote the value of the virtual queue in frame T. According to the update of virtual-queue length, we have

$$Z_{T+1} = \left(Z_T + \rho + \epsilon - \frac{\bar{C}_T}{T^\beta} \right)^+ \geq Z_T + \rho + \epsilon - \frac{\bar{C}_T}{K^\beta}, \tag{3.67}$$

which implies that

$$\begin{aligned} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} (-g(s_k, a_k) + \rho) \leq K^\beta \left(Z_{T+1} - Z_T \right) \\ + \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \left((1-\gamma)\hat{C}_k(s_k, a_k) - g(s_k, a_k) - \epsilon \right). \end{aligned} \tag{3.68}$$

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

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T} \rho - g(s_k, a_k) \right] \leq -K\epsilon + K^\beta \mathbb{E}\left[Z_{K^{1-\beta}+1} \right] \\ + \mathbb{E}\left[\sum_{k=1}^{K} (1-\gamma)\hat{C}_k(s_k, a_k) - g(s_k, a_k) \right], \end{aligned} \tag{3.69}$$

where we used the fact $Z_1 = 0$. Combining the upper bound on the estimation error of $\hat{C}_k$ in Lemma III.8 and the upper bound on $\mathbb{E}[Z_T]$ in Lemma III.7 yields the constraint violation bound. Furthermore, under our careful choices of $m, \gamma, \epsilon, \eta, \alpha, \beta$ and ι, it can be easily verified that $K\epsilon$ dominates the upper bounds in (3.69), which leads to the fact that constraint violation because zero when K is sufficiently large. In particular, under our assumption on K, which implies that $\epsilon \leq \frac{\delta}{2}$, and leads to

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

3.5.3 Detailed Proofs

We provide the detailed proof in this section. A notation table and some supporting lemmas can be found in Appendix B.

3.5.3.1 Proof of Lemma III.3.

Proof. Given $q^*(x,a)$ is the optimal solution, we have

$$\sum_{s,a} q^*(s,a)g(s,a) \geq \rho.$$

Under Assumption 1, we know that there exists a feasible solution $q^{\xi_1}(s,a)$ such that

$$\sum_{s,a} q^{\xi_1}(s,a)g(s,a) \geq \rho + \delta.$$

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

$$\begin{aligned}
\sum_{x,a} q^{\xi_2}(s,a)g(s,a) &= \sum_{s,a} \left((1-\frac{\epsilon}{\delta})q^*(s,a) + \frac{\epsilon}{\delta}q^{\xi_1}(s,a)\right) g(s,a) \geq \rho + \epsilon, \\
\sum_{s,a} q^{\xi_2}(s,a) &= \sum_{x',a'} p(s|s',a')q^{\xi_2}(s',a'), \\
\sum_{s,a} q^{\xi_2}(s,a) &= 1. \qquad (3.70)
\end{aligned}$$

Also we have $q^{\xi_2}(s,a) \geq 0$ for all (s,a). Thus $q^{\xi_2}(s,a)$ is a feasible solution to the ϵ-tightened optimization problem. Then given $q^{\epsilon,*}(s,a)$ is the optimal solution to the ϵ-tightened optimization problem, we have

$$\begin{aligned}
&\sum_{s,a} \left(q^*(x,a) - q^{\epsilon,*}(s,a)\right) r(s,a) \\
\leq &\sum_{s,a} \left(q^*(s,a) - q^{\xi_2}(s,a)\right) r(s,a)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{s,a} \left(q^*(s,a) - \left(1 - \frac{\epsilon}{\delta}\right) q^*(s,a) - \frac{\epsilon}{\delta} q^{\xi_1}(s,a) \right) r(s,a) \\
&\leq \sum_{s,a} \left(q^*(s,a) - \left(1 - \frac{\epsilon}{\delta}\right) q^*(s,a) \right) r(s,a) \\
&\leq \frac{\epsilon}{\delta} \sum_{s,a} q^*(s,a) r(s,a) \\
&\leq \frac{\epsilon}{\delta}, \qquad (3.71)
\end{aligned}
$$

where the last inequality holds because $0 \leq r(s,a) \leq 1$ under our assumption. Therefore the result follows because

$$
J_r^* = \sum_{s,a} q^*(s,a) r(s,a), \qquad (3.72)
$$

$$
J_r^{\epsilon,*} = \sum_{s,a} q^{\epsilon,*}(s,a) r(s,a). \qquad (3.73)
$$

□

3.5.3.2 Proof of Lemma III.4

Proof. We only prove the result for the reward value functions. The proof for the utility function is almost identical. Let π be an arbitrary policy. The proof follows Lemma 2 in [45] closely. According to the Bellman equation, we have

$$
\begin{aligned}
V^\pi(s) =& \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} r(s_k, \pi(s_k)) | s_1 = s, \pi \right] \\
=& \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} \left(J_r^\pi + v^\pi(s_k) - \mathbb{E}_{s' \sim p(\cdot|s_k, \pi(s_k))} v^\pi(s') \right) | s_1 = s, \pi \right] \\
=& \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} \left(J_r^\pi + v^\pi(s_k) - v^\pi(s_{k+1}) \right) | s_1 = s, \pi \right] \\
=& \frac{J_r^\pi}{1-\gamma} + v^\pi(s) - \mathbb{E}\left[\sum_{k=2}^{\infty} (\gamma^{k-2} - \gamma^{k-1}) v^\pi(s_k) | s_1 = s, \pi \right]. \qquad (3.74)
\end{aligned}
$$

Then

$$V^{\pi}(s) \geq \frac{J_r^{\pi}}{1-\gamma} + \min_s v^{\pi}(s) - \max_s v^{\pi}(s) \sum_{k=2}^{\infty} (\gamma^{k-2} - \gamma^{k-1})$$
$$= \frac{J_r^{\pi}}{1-\gamma} - sp(v^{\pi}), \tag{3.75}$$

and

$$V^{\pi}(s) \leq \frac{J_r^{\pi}}{1-\gamma} + \max_s v^{\pi}(s) - \min_s v^{\pi}(s) \sum_{k=2}^{\infty} (\gamma^{k-2} - \gamma^{k-1})$$
$$= \frac{J_r^{\pi}}{1-\gamma} + sp(v^{\pi}). \tag{3.76}$$

Therefore we can conclude that

$$J_r^{\pi} - (1-\gamma) V^{\pi}(s) \leq (1-\gamma) sp(v^{\pi}). \tag{3.77}$$

For any $s_1, s_2 \in \mathcal{S}$, we have

$$|V^{\pi}(s_1) - V^{\pi}(s_2)| \leq \left| V^{\pi}(s_1) - \frac{J_r^{\pi}}{1-\gamma} \right| + \left| V^{\pi}(s_2) - \frac{J_r^{\pi}}{1-\gamma} \right| \leq 2sp(v^{\pi}) \tag{3.78}$$

□

3.5.3.3 Proof of Lemma III.5

Proof. The proof follows Lemma 3 in [28] but for the discounted case. Consider frame T and episodes in frame T. Define $Z = Z_{(T-1)K^{\beta}+1}$ because the value of the virtual queue does not change during each frame. We further define/recall the following notations:

$$F_k(s,a) = Q_k(s,a) + \frac{Z}{\eta} C_k(s,a), \quad U_k(s) = V_k(s) + \frac{Z}{\eta} W_k(s),$$

$$\hat{F}_k(s,a) = \hat{Q}_k(s,a) + \frac{Z}{\eta}\hat{C}_k(s,a), \quad \hat{U}_k(s) = \hat{V}_k(s) + \frac{Z}{\eta}\hat{W}_k(s),$$
$$F^\pi(s,a) = Q^\pi(s,a) + \frac{Z}{\eta}C^\pi(s,a), \quad U^\pi(s) = V^\pi(s) + \frac{Z}{\eta}W^\pi(s).$$

In the following, we use π to denote the policy $\epsilon, *$ without obscurity. Then following a similar proof of Lemma B.3, we have

$$\begin{aligned}
&\{F_{k+1} - F^\pi\}(s,a) \\
=&\alpha_t^0 \left\{F_{(T-1)K^\beta+1} - F^\pi\right\}(s,a) \\
&+ \sum_{i=1}^t \alpha_t^i \left(\left\{\hat{U}_{k_i} - U^\pi\right\}(s_{k_i+1}) + \gamma\left(U^\pi(s_{k_i+1}) - \mathbb{E}_{s'\sim p(\cdot|s,a)}U^\pi(s')\right) + \left(1+\frac{Z}{\eta}\right)b_i \right) \\
\geq_{(a)}&\alpha_t^0 \left\{\hat{F}_{(T-1)K^\beta+1} - F^\pi\right\}(s,a) + \sum_{i=1}^t \alpha_t^i \left\{\hat{U}_{k_i} - U^\pi\right\}(s_{k_i+1}) \\
=_{(b)}&\alpha_t^0 \left\{\hat{F}_{(T-1)K^\beta+1} - F^\pi\right\}(s,a) + \sum_{i=1}^t \alpha_t^i \left(\max_a \hat{F}_{k_i}(s_{k_i+1},a) - F^\pi(s_{k_i},\pi(s_{k_i}))\right) \\
\geq&\alpha_t^0 \left\{\hat{F}_{(T-1)K^\beta+1} - F^\pi\right\}(s,a) + \sum_{i=1}^t \alpha_t^i \left\{\hat{F}_{k_i} - F^\pi\right\}(s_{k_i+1},\pi(s_{k_i+1})),
\end{aligned} \tag{3.79}$$

where inequality (a) holds because of the concentration result in Lemma B.4 and

$$\sum_{i=1}^t \alpha_t^i(1+\frac{Z}{\eta})b_i = \sum_{i=1}^t \alpha_t^i(1+\frac{Z}{\eta})\kappa\sqrt{\frac{(\chi+1)\iota}{\chi+i}} = \frac{\eta+Z}{\eta}\kappa\sqrt{\frac{(\chi+1)\iota}{\chi+t}}, \tag{3.80}$$

where the last equality comes from the properties of the learning rate (Lemma A.1). Equality (b) holds because our algorithm selects the action that maximizes $\hat{F}_{k_i}(s_{k_i+1},a)$ so $\hat{U}_{k_i}(s_{k_i+1}) = \max_a \hat{F}_{k_i}(s_{k_i+1},a)$. The inequality above suggests that we can prove $\{F_{k+1} - F^\pi\}(s,a)$ is an overestimation for any (s,a) if (i)

$$\left\{\hat{F}_{(T-1)K^\beta+1} - F^\pi\right\}(s,a) \geq 0,$$

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

$$\left\{\hat{F}_{k'} - F^{\pi}\right\}(s,a) \geq 0 \qquad \forall k' \leq k$$

i.e. the result holds for all aforementioned steps in the same frame. Furthermore, because that

$$\hat{F}_{k+1}(s,a) = \hat{Q}_{k+1}(s,a) + \frac{Z}{\eta}\hat{C}_{k+1}(s,a) \tag{3.81}$$

and the update rule of $\hat{Q}_{k+1}, \hat{C}_{k+1}$ in Line $12-17$ in Algorithm 3, we have

$$\hat{F}_{k+1}(s,a) - F^{\pi}(s,a) \geq 0.$$

Then we only need to prove at the beginning of each frame, $\left\{\hat{F}_{(T-1)K^{\beta}+1} - F^{\pi}\right\}(s,a) \geq 0$, which is obviously true because all reward and cost Q-functions are reset to H at the beginning of each frame (line 27,28 in Algorithm 3). Let $\mathcal{E}$ denote the event that $\{\hat{F}_k - F^{\epsilon,*}\}(s,a) \geq 0$ for all k. Then we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\pi} - \hat{F}_k\right)q^{\pi}\right\}(s_k,a)\right] \\
=&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\pi} - \hat{F}_k\right)q^{\pi}\right\}(s_k,a)\middle|\mathcal{E}\right]\Pr(\mathcal{E}) \\
&+\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\pi} - \hat{F}_k\right)q^{\pi}\right\}(s_k,a)\middle|\mathcal{E}^c\right]\Pr(\mathcal{E}^c) \\
\leq_{(a)}&K\left(1 + \frac{2K^{1-\beta}}{\eta}\right)H\frac{1}{K^3} \leq \frac{3H}{\eta K},
\end{aligned} \tag{3.82}$$

where inequality (a) holds because at any timestep k, we have

$$(F^{\pi} - \hat{F}_k) \leq \left(1 + \frac{2K^{1-\beta}}{\eta}\right)H.$$

□

3.5.3.4 Porrf of Lemma III.6

Proof. Consider Lyapunov function $L_T = \frac{1}{2}Z_T^2$, where T is the frame index and Z_T is the length of the virtual queue at the beginning of the Tth frame. Firstly, we have

$$\begin{aligned} L_{T+1} - L_T \leq & Z_T \left(\rho + \epsilon - \frac{\bar{C}_T}{K^\beta}\right) + \frac{\left(\rho + \epsilon - \frac{\bar{C}_T}{K^\beta}\right)^2}{2} \\ \leq & \frac{Z_T}{K^\beta} \sum_{k=TK^\beta+1}^{(T+1)K^\beta} (\rho + \epsilon - (1-\gamma)\hat{C}_k(s_k, a_k)) + 2. \end{aligned} \tag{3.83}$$

Then adding and subtracting additional terms,

$$\begin{aligned} & \mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\ \leq & \frac{1}{K^\beta} \sum_{k=TK^\beta+1}^{(T+1)K^\beta} \Big(\mathbb{E}[z(\rho + \epsilon - (1-\gamma)\hat{C}_k(s_k, a_k)) - \eta(1-\gamma)\hat{Q}_k(s_k, a_k) | Z_T = z] \\ & + \eta(1-\gamma)\mathbb{E}[\hat{Q}_k(s_k, a_k) | Z_T = z]\Big) + 2. \end{aligned} \tag{3.84}$$

Specifically, for the term inside the summation, we have

$$\begin{aligned} & \Big(\mathbb{E}[z(\rho + \epsilon - (1-\gamma)\hat{C}_k(s_k, a_k)) - \eta(1-\gamma)\hat{Q}_k(s_k, a_k) | Z_T = z] \\ & + \eta(1-\gamma)\mathbb{E}[\hat{Q}_k(s_k, a_k) | Z_T = z]\Big) \\ \leq & z(\rho + \epsilon) - \mathbb{E}\left[\eta(1-\gamma)\left(\sum_a \left\{\frac{z}{\eta}\hat{C}_k q^\epsilon + \hat{Q}_k q^\epsilon\right\}(s_k, a)\right) \middle| Z_T = z\right] \\ & + \eta(1-\gamma)\mathbb{E}[\hat{Q}_k(s_k, a_k) | Z_T = z] \\ = & \mathbb{E}\left[z\left(\rho + \epsilon - \sum_{s,a} g(s,a) q^\epsilon(s,a)\right) \middle| Z_T = z\right] \\ & + \mathbb{E}\left[z\left(\sum_{s,a} g(s,a) q^\epsilon(s,a) - (1-\gamma)\sum_a C^\epsilon(s_k, a) q^\epsilon(s_k, a)\right) \middle| Z_T = z\right] \end{aligned}$$

$$
\begin{aligned}
&- \eta(1-\gamma)\mathbb{E}\left[\sum_a \hat{Q}_k(s_k,a)q^\epsilon(s_k,a) - \hat{Q}_k(s_k,a_k)\middle| Z_T = z\right] \\
&+ (1-\gamma)\mathbb{E}\left[z\sum_a \left\{(C^\epsilon - \hat{C}_k)q^\epsilon\right\}(s_k,a)\middle| Z_T = z\right] \\
\leq &- \eta(1-\gamma)\mathbb{E}\left[\sum_a \hat{Q}_k(s_k,a)q^\epsilon(s_k,a) - \hat{Q}_k(s_k,a_k)\middle| Z_T = z\right] \\
&+ (1-\gamma)\mathbb{E}\left[z\sum_a \left\{(C^\epsilon - \hat{C}_k)q^\epsilon\right\}(s_k,a)\middle| Z_T = z\right] \\
&+ \mathbb{E}\left[z\left(J_g^\epsilon - (1-\gamma)\sum_a C^\epsilon(s_k,a)q^\epsilon(s_k,a)\right)\middle| Z_T = z\right], \qquad (3.85)
\end{aligned}
$$

where the first inequality holds because a_k is chosen to maximize $\hat{Q}_k(s_k,a)+\frac{Z_k}{\eta}\hat{C}_k(s_k,a)$, and the last inequality is true because $q^\epsilon(s,a)$ is a feasible solution to the optimization problem (3.12) such that

$$
\rho + \epsilon - \sum_{s,a} g(s,a)q^\epsilon(s,a) \leq 0
$$

Therefore by replacing $q^\epsilon(s,a)$ with the optimal solution $q^{\epsilon,*}(s,a)$, we have

$$
\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\
\leq &- \eta(1-\gamma)\mathbb{E}\left[\sum_a \hat{Q}_k(s_k,a)q^{\epsilon,*}(s_k,a) - \hat{Q}_k(s_k,a_k)\middle| Z_T = z\right] \\
&+ (1-\gamma)\mathbb{E}\left[z\sum_a \left\{(C^{\epsilon,*} - \hat{C}_k)q^{\epsilon,*}\right\}(s_k,a)\middle| Z_T = z\right] \\
&+ \mathbb{E}\left[z\left(J_g^{\epsilon,*} - (1-\gamma)\sum_a C^{\epsilon,*}(s_k,a)q^{\epsilon,*}(s_k,a)\right)\middle| Z_T = z\right] + 2 \qquad (3.86)
\end{aligned}
$$

After taking expectation with respect to Z, dividing η on both sides, reorganizing the terms, and then applying the telescoping sum, we get

$$
\mathbb{E}\left[\sum_{k=1}^{K}(1-\gamma)\left(\sum_a \left\{\hat{Q}_k q^{\epsilon,*}\right\}(s_k,a) - \hat{Q}_k(s_k,a_k)\right.\right.
$$

$$
\begin{aligned}
&+\frac{Z_k}{\eta}\sum_a\left\{\hat{C}_k q^{\epsilon,*}-C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a)\Big)\Big]\\
\leq&\frac{2K}{\eta}+\sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)}{\eta}sp(w^{\epsilon,*})+\frac{K^{\beta}\mathbb{E}\left[L_1-L_{K^{1-\beta}+1}\right]}{\eta}\\
\leq&\frac{2K}{\eta}+\sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{\eta},
\end{aligned}
\tag{3.87}
$$

where the first inequality comes from Lemma III.4, and the last inequality comes from the fact that $\kappa=\max_{0\leq\epsilon\leq\rho/2}\left(\max\{sp(v^{\epsilon,*}),sp(w^{\epsilon,*}),1\}\right)$ is non-negative. □

3.5.3.5 Proof of Lemma:III.7

Proof. The proof will also use the following lemma from [59].

Lemma III.9. *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 γ and v with $0<\gamma\leq v$, 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 $r=\frac{\gamma}{v^2+v\gamma/3}$. □

We apply Lemma III.9 to a new Lyapunov function:

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

To verify condition (1) in Lemma III.9, consider $\bar{L}_T=Z_T\geq\theta_T=\frac{6(\eta+2+\frac{3H}{K^2})}{\delta}$ and

$2\epsilon \leq \delta$. The conditional expected drift of $\bar{L}_T$ is

$$\begin{aligned}
&\mathbb{E}\left[Z_{T+1} - Z_T | Z_T = z\right] \\
=&\mathbb{E}\left[\sqrt{Z_{T+1}^2} - \sqrt{z^2}\middle| Z_T = z\right] \\
\leq&\frac{1}{2z}\mathbb{E}\left[\, Z_{T+1}^2 - z^2\middle|\, Z_T = z\right] \\
\leq_{(a)}&-\frac{\delta}{2} + \frac{(\eta + 2 + \frac{3H}{K^2})}{z} \\
\leq&-\frac{(\eta + 2 + \frac{3H}{K^2})}{\theta_T} \qquad (3.88) \\
=&-\frac{\delta}{6}, \qquad (3.89)
\end{aligned}$$

where inequality (a) is obtained according to Lemma A.5; and the last inequality holds given $z \geq \theta_T$.

To verify condition (2) in Lemma III.9, we have

$$Z_{T+1} - Z_T \leq |Z_{T+1} - Z_T| \leq \left|\rho + \epsilon - \bar{C}_T\right| \leq 2. \qquad (3.90)$$

Now choose $\gamma = \frac{\delta}{6}$ and $v = 2$. From Lemma III.9, we obtain

$$\mathbb{E}\left[e^{rZ_T}\right] \leq e^{rZ_1} + \frac{2e^{r(v+\theta_T)}}{r\gamma}, \quad \text{where} \quad r = \frac{\gamma}{v^2 + v\gamma/3}. \qquad (3.91)$$

By Jensen's inequality, we have

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

which implies that

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

$$
\begin{aligned}
&\leq \frac{3v^2}{\lambda}\log\left(\frac{2v}{\lambda}\right) + v + \theta_T \\
&\leq \frac{72}{\delta}\log\left(\frac{24}{\delta}\right) + 2 + \frac{6(\eta+3)}{\delta} \\
&\leq \frac{92}{\delta}\log\left(\frac{24}{\delta}\right) + \frac{6\eta}{\delta}
\end{aligned} \tag{3.92}
$$

□

3.5.3.6 Proof of Lemma III.8

Proof. we have for any K' within a frame,

$$
\begin{aligned}
&\sum_{k=1}^{K'}\Big((1-\gamma)\hat{Q}_k(s_k,a_k) - r(s_k,a_k)\Big) \\
=&\gamma\sum_{k=1}^{K'}\sum_{i=1}^{n_k}\alpha_{n_k}^i\left[-\gamma r(s_k,a_k) + (1-\gamma)\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1})\right] \\
&+ 2(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}},
\end{aligned} \tag{3.93}
$$

where the equality comes from the updating rule of $\hat{Q}_k(s,a)$ and the fact $\sum_{i=1}^{\tau}\alpha_\tau^i = 1$. We use n_k denotes $n_{k+1}(s_k,a_k)$ for short, that is the number of visits to state-action pair (s_k,a_k) by timestep k (including k) within the same frame. Note that $\alpha_{n_k}^0 = 0$ by definition since $n_k \geq 1$. For the second term, we further have

$$
\begin{aligned}
&\gamma(1-\gamma)\sum_{k=1}^{K'}\sum_{i=1}^{n_k}\alpha_{n_k}^i\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1}) \\
=&\gamma(1-\gamma)\sum_{k=1}^{K'}\sum_{s,a}\mathbb{I}_{\{s_k=s,a_k=a\}}\sum_{i=1}^{n_{k+1}(s,a)}\alpha_{n_{k+1}(s,a)}^i\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1}) \\
=&\gamma(1-\gamma)\sum_{s,a}\sum_{j=1}^{n_{K'+1}(s,a)}\sum_{i=1}^{j}\alpha_j^i\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1}) \\
=_{(a)}&\gamma(1-\gamma)\sum_{s,a}\sum_{i=1}^{n_{K'+1}(s,a)}\sum_{j=i}^{n_{K'+1}(s,a)}\alpha_j^i\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1})
\end{aligned}
$$

$$=\gamma(1-\gamma)\sum_{s,a}\sum_{i=1}^{n_{K'+1}(s,a)}\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1})\sum_{j=i}^{n_{K'+1}(s,a)}\alpha_j^i, \tag{3.94}$$

where the equality (a) is true due to the changing of the order of summation on i and j. Since we have a upper bound that $\sum_{j=i}^{n_{T'+1}(s,a)}\alpha_j^i \leq \sum_{j=i}^{\infty}\alpha_j^i = 1+\frac{1}{\chi}$ and $\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1}) >= 0$, then we can obtain

$$\begin{aligned}
&\gamma(1-\gamma)\sum_{k=1}^{K'}\sum_{i=1}^{n_k}\alpha^i_{n_{k+1}(s,a)}\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1})\\
\leq&\gamma(1-\gamma)\sum_{s,a}\sum_{i=1}^{n_{K'+1}(s,a)}\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1})\sum_{j=i}^{\infty}\alpha_j^i\\
=&\gamma(1-\gamma)\sum_{s,a}\sum_{i=1}^{n_{K'+1}(s,a)}\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1})\left(1+\frac{1}{\chi}\right)\\
=&\gamma(1-\gamma)\left(1+\frac{1}{\chi}\right)\sum_{k=1}^{K'}\hat{V}_k(s_{k+1}).
\end{aligned} \tag{3.95}$$

Substituting in (3.93), we have

$$\begin{aligned}
&\sum_{k=1}^{K'}\Big((1-\gamma)\hat{Q}_k(s_k,a_k)-r(s_k,a_k)\Big)\\
\leq&-\gamma\sum_{k=1}^{K'}r(s_k,a_k)+\gamma(1-\gamma)\left(1+\frac{1}{\chi}\right)\sum_{k=1}^{K'}\hat{V}_k(s_{k+1})\\
&+2(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}}\\
\leq_{(a)}&-\gamma\sum_{k=1}^{K'}r(s_k,a_k)+\gamma(1-\gamma)\sum_{k=1}^{K'}\hat{V}_k(s_{k+1})+\frac{K'}{\chi}(1-\gamma)\gamma H\\
&+2(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}}\\
=&-\gamma\sum_{k=1}^{K'}r(s_k,a_k)+\gamma(1-\gamma)\sum_{k=1}^{K'}\Big(\hat{V}_k(s_{k+1})-\hat{V}_{k+1}(s_{k+1})\Big)+\gamma(1-\gamma)\sum_{k=1}^{K'}\hat{V}_{k+1}(s_{k+1})\\
&+\frac{K'}{\chi}(1-\gamma)\gamma H
\end{aligned}$$

$$+ 2(1-\gamma)\kappa \sum_{k=1}^{K'} \sqrt{\frac{(\chi+1)\iota}{\chi+n_k}} \tag{3.96}$$

$$\leq_{(b)} -\gamma \sum_{k=1}^{K'} r(s_k, a_k) + \gamma(1-\gamma) \sum_{k=2}^{K'+1} \hat{V}_k(s_k) + \frac{K'}{\chi}(1-\gamma)\gamma H$$
$$+ 2(1-\gamma)\kappa \sum_{k=1}^{K'} \sqrt{\frac{(\chi+1)\iota}{\chi+n_k}} + \gamma(1-\gamma)SH$$
$$\leq \gamma \sum_{k=1}^{K'} \Big((1-\gamma)\hat{Q}_k(s_k, a_k) - r(s_k, a_k) \Big) + \frac{K'}{\chi}(1-\gamma)\gamma H + 2(1-\gamma)\kappa \sum_{k=1}^{K'} \sqrt{\frac{(\chi+1)\iota}{\chi+n_k}}$$
$$+ 2\gamma(1-\gamma)SH \tag{3.97}$$
$$\leq \gamma^m K'(1-\gamma)H + \frac{K'm}{\chi}(1-\gamma)\gamma H + 2m(1-\gamma)\kappa \sum_{k=1}^{K'} \sqrt{\frac{(\chi+1)\iota}{\chi+n_k}} + 2m\gamma(1-\gamma)SH$$

(repeatedly use the inequality m times)

$$\leq_{(c)} \gamma^m K'(1-\gamma)H + \frac{K'm}{\chi}(1-\gamma)\gamma H$$
$$+ 4\gamma(1-\gamma)m\kappa\sqrt{(\chi+1)SAT'\iota} + 2m\gamma(1-\gamma)SH, \tag{3.98}$$

where the inequality (a) holds because $\hat{V}_k(s)$ is bounded by H, inequality (b) is true because that for any state s, $\hat{V}_k(s) \geq \hat{V}_{t+1}(s)$ and the value can decrease by at most H. Inequality (c) is by nothing but that

$$\sum_{k=1}^{K'} \sqrt{\frac{(\chi+1)\iota}{\chi+n_k}} = \sum_{k=1}^{K'} \sum_{s,a} \sqrt{\frac{\mathbb{I}_{\{s_k=s, a_k=a\}}(\chi+1)\iota}{\chi+n_k}} = \sum_{s,a} \sum_{j=1}^{n_{T'+1}(s,a)} \sqrt{\frac{(\chi+1)\iota}{\chi+j}}$$
$$\leq \sum_{s,a} \sum_{j=1}^{n_{K'+1}(s,a)} \sqrt{\frac{(\chi+1)\iota}{j}} \leq 2 \sum_{s,a} \sqrt{(\chi+1)\iota n_{T'+1}(s,a)} \overset{(1)}{\leq} 2\sqrt{(\chi+1)SAK'\iota},$$

where the last inequality above holds because the left-hand side of (1) is the summation of K' terms and it is maximized when $n_{K'+1} = K'/SA$ for all s, a, i.e. by picking the largest K' terms. We finish the proof by substituting $1-\gamma$ with $\frac{1}{H}$. □

3.6 Summary

In this chapter, we proposed the first model-free RL algorithm for infinite-horizon average-reward CMDPs. The design of the algorithm is based on the primal-dual approach. By using the Lyapunov drift analysis, we proved that our algorithm achieves sublinear regret and zero constraint violation. Our regret bound scales as $\tilde{O}(K^{\frac{5}{6}})$ and is suboptimal compared to model-based approaches. However, this is the first model-free and simulator-free algorithm with sub-linear regret and optimal constraint violation. It is still an interesting open problem how to achieve $\tilde{O}(\sqrt{K})$ regret bound via model-free algorithms. The algorithm is also computationally efficient from an algorithmic perspective because it is model-free. The simulation result also demonstrates the good performance of our algorithm.

CHAPTER IV

Provably Efficient Model-Free Algorithms for Non-stationary CMDPs

4.1 Introduction

Safe reinforcement learning (RL) studies how to apply RL algorithms in real-world applications [68; 69; 70] that can operate under safety-related constraints. In classical safe-RL and CMDP 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 [71], collisions must be avoided while modeling and tracking time-varying environments such as traffic conditions; in an automated medical system [72], 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 [15; 17; 29; 30; 20; 39; 14; 19; 73]. 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 [74], 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 chapter, 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 non-stationary episodic CMDPs. We develop different types 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 overestimate 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 [74].

- 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 aim to achieve zero constraint violation. The method removes the need to have a priori knowledge of the variation budget, an open problem raised in [74] 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 settings. For example, one needs to account for the constraints. We overcome these difficulties with 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.

Related Work

Non-stationary MDP. Non-stationary unconstrained MDPs have been mostly studied recently [75; 76; 77; 78; 79; 80; 81; 82; 83; 84]. [75] consider a setting where the MDP is allowed to change for fixed number of times. When the variation budget is known a priori, [78] propose a policy-based algorithm in the setting where they assume stationary transitions and adversarial full-information rewards. [82; 84; 80; 83] consider a more general setting that both transitions and rewards are time-varying. A more recent work [81] introduces 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, [15; 17; 14] propose model-based approaches for tabular CMDPS. [85; 19] extend the results to the linear and linear kernel CMDPs. [20; 39] also provide efficient algorithms with a zero constraint violation guarantee. Besides using an estimated model, [25; 27] leverage a simulator for policy evaluation to

achieve provable regret guarantees. Moreover, [29; 30] 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, [18] consider CMDPs that assume that only the rewards vary over episodes. A concurrent work [74], 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 challenging setting where the algorithm is model-free and the budget is not known. Fortunately, we answer the open problem affirmatively raised in [74].

4.2 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 4.4, 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 :

$$V^{\pi}_{k,h}(x) = \mathbb{E}\left[\sum_{i=h}^{H} r_{k,i}(x_{k,i}, \pi(x_{k,i}))\middle| x_{k,h} = x\right]. \tag{4.1}$$

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 :

$$Q^{\pi}_{k,h}(x, a) = r_{k,h}(x, a) + \mathbb{E}\left[\sum_{i=h+1}^{H} r_{k,i}(x_{k,i}, \pi(x_{k,i}))\middle| x_{k,h} = x, a_{k,h} = a\right]. \tag{4.2}$$

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

$$W^{\pi}_{k,h}(x) = \mathbb{E}\left[\sum_{i=h}^{H} g_{k,i}(x_{k,i}, \pi(x_{k,i}))\middle| x_{k,h} = x\right], \tag{4.3}$$

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:

$$C^{\pi}_{k,h}(x, a) = g_{k,h}(x, a) + \mathbb{E}\left[\sum_{i=h+1}^{H} g_{k,i}(x_{k,i}, \pi(x_{k,i}))\middle| x_{k,h} = x, a_{k,h} = a\right]. \tag{4.4}$$

For simplicity, we adopt the following notations:

$$\mathbb{P}_{k,h} V^{\pi}_{k,h+1}(x, a) = \mathbb{E}_{x' \sim \mathbb{P}_{k,h}(\cdot|x,a)} V^{\pi}_{k,h+1}(x'), \tag{4.5}$$

$$\mathbb{P}_{k,h} W^{\pi}_{h+1}(x, a) = \mathbb{E}_{x' \sim \mathbb{P}_{k,h}(\cdot|x,a)} W^{\pi}_{k,h+1}(x'). \tag{4.6}$$

We also denote the empirical counterparts as

$$\hat{\mathbb{P}}_{k,h} V^{\pi}_{k,h+1}(x,a) = V^{\pi}_{k,h+1}(x_{k+1,h}), \tag{4.7}$$

$$\hat{\mathbb{P}}_{k,h} W^{\pi}_{k,h+1}(x,a) = W^{\pi}_{k,h+1}(x_{k+1,h}), \tag{4.8}$$

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:

$$\max_{\pi_k \in \Pi} \mathbb{E}\left[V^{\pi_k}_{k,1}(x_1)\right] \quad \text{subject to: } \mathbb{E}\left[W^{\pi_k}_{k,1}(x_1)\right] \geq \rho, \tag{4.9}$$

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 (4.9) for episode k. We evaluate our model-free RL algorithms using dynamic regret $\mathcal{R}(K)$ and constraint violation $\mathcal{V}(K)$ defined below:

$$\mathcal{R}(K) = \mathbb{E}\left[\sum_{k=1}^{K} \left(V^{\pi^*_k}_{k,1}(x_{k,1}) - V^{\pi_k}_{k,1}(x_{k,1})\right)\right], \tag{4.10}$$

$$\mathcal{V}(K) = \mathbb{E}\left[\sum_{k=1}^{K} \left(\rho - W^{\pi_k}_{k,1}(x_{k,1})\right)\right], \tag{4.11}$$

where π_k is the policy used in episode k. Note that here we use the dynamic regret concept as the optimal policy may be different. We further make the following standard assumption [17; 19; 18; 29].

Assumption IV.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:

$$B_r := \sum_{k=1}^{K-1} \sum_{h=1}^{H} \max_{x,a} |r_{k,h}(x,a) - r_{k+1,h}(x,a)| \tag{4.12}$$

$$B_g := \sum_{k=1}^{K-1} \sum_{h=1}^{H} \max_{x,a} |g_{k,h}(x,a) - g_{k+1,h}(x,a)| \tag{4.13}$$

$$B_p := \sum_{k=1}^{K-1} \sum_{h=1}^{H} \max_{x,a} \|\mathbb{P}_{k,h}(\cdot|x,a) - \mathbb{P}_{k+1,h}(\cdot|x,a)\|_1. \tag{4.14}$$

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:

$$\max_{q_{k,h}} \sum_{h,x,a} q_{k,h}(x,a) r_{k,h}(x,a) \tag{4.15}$$

$$\text{s.t.:} \sum_{h,x,a} q_{k,h}(x,a) g_{k,h}(x,a) \geq \rho \tag{4.16}$$

$$\sum_{a} q_{k,h}(x,a) = \sum_{x',a'} \mathbb{P}_{k,h-1}(x|x',a') q_{k,h-1}(x',a') \tag{4.17}$$

$$\sum_{x,a} q_{k,h}(x,a) = 1, \forall h \in [H] \tag{4.18}$$

$$\sum_{a} q_{k,1}(x,a) = \mu_0(x) \tag{4.19}$$

$$q_{k,h}(x,a) \geq 0, \forall x \in \mathcal{S}, \forall a \in \mathcal{A}, \forall h \in [H]. \tag{4.20}$$

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

$$\max_{q_{k,h}} \sum_{h,x,a} q_{k,h}(x,a) r_{k,h}(x,a) \tag{4.21}$$

$$\text{s.t.:} \sum_{h,x,a} q_{k,h}(x,a) g_{k,h}(x,a) \geq \rho + \epsilon, \text{ and } (4.17) - (4.20),$$

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 (4.9) or the solution to the corresponding LP (4.15) and superscript $^{\epsilon,*}$ to denote the optimal value/policy related to the ϵ-tightened version of CMDP.

4.3 Model-free Algorithms for the Tabular CMDP Setting

Next, we will start with presenting our algorithm Non-stationary Triple-Q in Algorithm 4 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, the restarting results in a worse regret. For example, the regret bound is $\tilde{\mathcal{O}}(\sqrt{K})$ [34] in the stationary setting but becomes $\tilde{\mathcal{O}}(K^{\frac{2}{3}})$ [84] 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 [29] 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

$$B_r^{(T)} := \sum_{k \in \mathcal{N}_T} \sum_{h=1}^{H} \max_{x,a} |r_{k,h}(x,a) - r_{k+1,h}(x,a)| \tag{4.22}$$

$$B_g^{(T)} := \sum_{k \in \mathcal{N}_T} \sum_{h=1}^{H} \max_{x,a} |g_{k,h}(x,a) - g_{k+1,h}(x,a)| \tag{4.23}$$

$$B_p^{(T)} := \sum_{k \in \mathcal{N}_T} \sum_{h=1}^{H} \max_{x,a} \|\mathbb{P}_{k,h}(\cdot|s,a) - \mathbb{P}_{k+1,h}(\cdot|x,a)\|_1. \tag{4.24}$$

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 4), 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 [79; 84; 83].

4.3.1 Results of Tabular CMDPs

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

Theorem IV.2. *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:*

$$\mathcal{R}(K) = \tilde{\mathcal{O}}(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} B^{\frac{1}{3}} K^{\frac{4}{5}}) \tag{4.25}$$

$$\mathcal{V}(K) = 0 \tag{4.26}$$

Algorithm 4: 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 $\iota = 128\log(\sqrt{2SAHK})$;

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 $t = N_h(x_{k,h}, a_{k,h}), b_t = \frac{1}{4}\sqrt{\frac{H^2\iota(\chi+1)}{\chi+t}}, \alpha_t = \frac{\chi+1}{\chi+t}$;

    $Q_h(x_{k,h}, a_{k,h}) \leftarrow$
    $(1-\alpha_t)Q_h(x_{k,h}, a_{k,h}) + \alpha_t \left(r_{k,h}(x_{k,h}, a_{k,h}) + V_{h+1}(x_{k,h+1}) + b_t + 2H\tilde{b}\right)$;

    $C_h(x_{k,h}, a_{k,h}) \leftarrow$
    $(1-\alpha_t)C_h(x_{k,h}, a_{k,h}) + \alpha_t \left(g_{k,h}(x_{k,h}, a_{k,h}) + W_{h+1}(x_{k,h+1}) + b_t + 2H\tilde{b}\right)$;

    $a' = \arg\max_a \left(Q_h(x_{k,h}, a) + \frac{Z}{\eta} C_h(x_{k,h}, a)\right)$,
    $V_h(x_{k,h}) \leftarrow Q_h(x_{k,h}, a') \quad W_h(x_{k,h}) \leftarrow C_h(x_{k,h}, a')$;

    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

    $N_h(x,a) \leftarrow 0$, $Q_h(x,a) = C_h(x,a) = Q_h(x,a) = C_h(x,a) \leftarrow H$,
    $Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C} \cdot B^c}{K^\alpha}\right)^+, \bar{C} \leftarrow 0$

Dynamic Regret

As shown in Algorithm 4, let $Q_{k,h}(x,a), C_{k,h}(x,a)$ denote the estimate Q values at the beginning of the $k-$th episode. The dynamic regret can be decoupled as:

$$\mathcal{R}(K) = \mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q^*_{k,1}q^*_{k,1} - Q^{\epsilon,*}_{k,1}q^{\epsilon,*}_{k,1}\right\}(x_{k,1}, a)\right)\right] + \tag{4.27}$$

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q^{\epsilon,*}_{k,1}q^{\epsilon,*}_{k,1}\right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1})\right)\right] + \tag{4.28}$$

$$\mathbb{E}\left[\sum_{k=1}^{K}\left\{Q_{k,1}-Q_{k,1}^{\pi_k}\right\}(x_{k,1},a_{k,1})\right], \tag{4.29}$$

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 IV.3. *Given any frame T, for any (x,a,h), and $(T-1)K^{\alpha}/B^{c}\leq k_1\leq k_2\leq TK^{\alpha}/B^{c}$, we have*

$$|Q_{k_1,h}^{\pi}(x,a)-Q_{k_2,h}^{\pi'}(x,a)|\leq H\tilde{b} \tag{4.30}$$

$$|C_{k_1,h}^{\pi}(x,a)-C_{k_2,h}^{\pi'}(x,a)|\leq H\tilde{b} \tag{4.31}$$

Then we will show that in Lemma IV.12 the first term (4.27) can be bounded by comparing the original LP associated with the tightened LP such that (4.27) $\leq \frac{KH\epsilon}{\delta}$. The term (4.29) 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 4) and the added bonus. Then (4.29) $\leq H^2SAK^{1-\alpha}B^{c}+\frac{2(H^3\sqrt{\iota}+2H^2\tilde{b})K}{\chi}+\sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)B^{c}}+2\tilde{b}H^2K$.

For the remaining term (4.28), we need to add and subtract additional terms to construct a 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 IV.10 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 IV.13 we leverage Lyapunov-drift method and consider the 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 (4.28) $\leq \frac{K(2H^4\iota+4H^2\tilde{b}^2+\epsilon^2)}{\eta}+$

$\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}$. Then combining inequalities (4.27),(4.28),(4.29) above we can obtain for $K \geq \left(\frac{16\sqrt{SAH^6\iota^3}B^{1/3}}{\delta}\right)^5$, applying the condition $K^{1-\alpha}B^c\tilde{b} \leq B$, along with our choices of parameters (Line 2 in Algorithm 4) for balancing each terms, we conclude that $\mathcal{R}(K) = \tilde{\mathcal{O}}(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}B^{\frac{1}{3}}K^{\frac{4}{5}})$.

Constraint Violation

According to the virtual-Queue update, we have

$$\begin{aligned} Z_{T+1} =& \left(Z_T + \rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}\right)^+ \\ \geq& Z_T + \rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}, \end{aligned} \tag{4.32}$$

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

$$\begin{aligned} &\sum_k \left(-C_{k,1}^{\pi_k}(x_{k,1}, a_{k,1}) + \rho\right) \leq \frac{K^\alpha}{B^c}\left(Z_{T+1} - Z_T\right) \\ &+ \sum_k \left(\left\{C_{k,1} - C_{k,1}^{\pi_k}\right\}(x_{k,1}, a_{k,1}) - \epsilon\right). \end{aligned} \tag{4.33}$$

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

$$\begin{aligned} \mathbb{E}\left[\sum_{k=1}^K \rho - C_{k,1}^{\pi_k}(x_{k,1}, a_{k,1})\right] \leq& -K\epsilon + \frac{K^\alpha}{B^c}\mathbb{E}\left[Z_{K^{1-\alpha}B^c+1}\right] \\ &+ \mathbb{E}\left[\sum_{k=1}^K \left\{C_{k,1} - C_{k,1}^{\pi_k}\right\}(x_{k,1}, a_{k,1})\right], \end{aligned} \tag{4.34}$$

where the inequality is true due to the fact $Z_1 = 0$. In Lemma IV.11, we will establish an upper bound on the estimation error of $\mathbb{E}\left[\sum_{k=1}^K \left\{C_{k,1} - C_1^{\pi_k}\right\}(x_{k,1}, a_{k,1})\right]$.

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 IV.14, 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 $1 \leq T \leq K^{1-\alpha}B^c$,

$$\mathbb{E}[Z_T] \leq \frac{100(H^4\iota + \tilde{b}^2H^2)}{\delta} \log\left(\frac{16(H^2\sqrt{\iota} + \tilde{b}H)}{\delta}\right) + \frac{4H^2B^c}{K\delta} + \frac{4H^2B^c}{\eta K^\alpha \delta} + \frac{4\eta(\sqrt{H^2\iota} + 2H^2\tilde{b})}{\delta}. \tag{4.35}$$

Substituting the results from Lemma IV.11 and (4.35) into (4.34), 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

$$\mathcal{V}(K) \leq \frac{100(H^4\iota + \tilde{b}^2H^2)K^{0.6}}{\delta B^{2/3}} \log\frac{16(H^2\sqrt{\iota} + H\tilde{b})}{\delta} - \sqrt{SAH^6\iota^3}K^{0.8}B^{\frac{1}{3}} \leq 0. \tag{4.36}$$

4.3.2 Unknown Variation Budgets

The design of the Algorithm 4 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 5 that adaptively learns the variation budget B based on the "Bandit over Bandit" algorithm [86]. Algorithm 5 uses an outer loop "bandit algorithm" as a master to learn the true value B, and use the inner loop Algorithm 4 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 4. Given a candidate set $\mathcal{J}$ of the total budget B, we choose "arms" (estimated budget) using the bandit adversarial bandit algorithm Exp3 [87]. 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

Algorithm 5: Double Restart Non-stationary Triple-Q

Choose $W = K^{5/9}, \mathcal{J}$ defined in Eq. (4.43)
$, \gamma_0 = \min\left\{1, \sqrt{\frac{(K/W)\log(K/W)}{(e-1)KH}}\right\}, \lambda = 1/9$;
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 4 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
    $s_{i+1} \leftarrow s_i(j)\exp(\gamma_0\hat{R}_i(j)/(J+1))$;

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,

$$\mathcal{R}(K) = \mathbb{E}\left[\sum_{k=1}^{K} V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_{i=1}^{K/W} R_i(\hat{B})\right] \tag{4.37}$$

$$+\mathbb{E}\left[\sum_{i=1}^{K/W} R_i(\hat{B}) - \sum_{i=1}^{K/W} R_i(B_i)\right], \tag{4.38}$$

where $\hat{B}$ is the optimal candidate from $\mathcal{J}$ (i.e., the true budget). We can show that

the term (4.37) can be bounded since this corresponds to regret when the true budget is known (which we have already bounded). However, the problem becomes how to bound the term (4.38). 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 after selecting the arm B_i, then one can set

$$\begin{cases} \hat{R}_i(B_i) = 0 & \text{if } G_i(B_i) < W\rho \\ \hat{R}_i(B_i) = R_i(B_i) & \text{if } G_i(B_i) \geq W\rho. \end{cases} \tag{4.39}$$

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

$$\begin{cases} \hat{R}_i(B_i) = \dfrac{G_i(B_i)}{K^\lambda} & \text{if } G_i(B_i) < W\rho \\ \hat{R}_i(B_i) = R_i(B_i) + \dfrac{G_i(B_i)}{K^\lambda} & \text{if } G_i(B_i) \geq W\rho. \end{cases} \tag{4.40}$$

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 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 design of the bandit reward function (4.40).

Lemma IV.4. *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 estimated 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*

$$\mathbb{E}\left[\sum_{i=1}^{K/W}(R_i(\hat{B}) - R_i(B_i))\right] = \tilde{\mathcal{O}}(H\sqrt{KW} + HK^{1-\lambda}) \tag{4.41}$$

$$\mathbb{E}\left[\sum_{i=1}^{K/W} G_i(\hat{B}) - G_i(B_i)\right] = \tilde{\mathcal{O}}(HK^{\lambda}\sqrt{KW}). \tag{4.42}$$

Note that the above lemma bounds (4.38). 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 4 for the tabular setting. Then we will discuss how to apply it to the linear function approximation setting. We define set $\mathcal{J}$ as

$$\mathcal{J} = \left\{\frac{K^{1/3}}{\Delta^{3/2}W}, \frac{K^{1/3}W^{\frac{1}{J}}}{\Delta^{3/2}W}, \ldots, \frac{K^{1/3}W}{\Delta^{3/2}W}\right\}, \tag{4.43}$$

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 4 for consecutive $W = K^{\zeta}$ episodes. Each epoch contains $WB_i^c/K^{\alpha\zeta}$ frames. We remark here that when using the Algorithm 4 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 5 achieves a sublinear regret and zero constraint violation without the knowledge of the total variation budget B.

Theorem IV.5. *Algorithm IV.5 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 $K \geq e^{\frac{1}{\delta}}$:*

$$\mathcal{R}(K) = \tilde{\mathcal{O}}(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} B^{\frac{1}{3}} K^{8/9}) \tag{4.44}$$

$$\mathcal{V}(K) = 0 \tag{4.45}$$

4.3.3 Simulation

We compare Algorithm 4 with two baseline algorithms: an algorithm [84] for non-stationary MDPs, and an algorithm [29] for stationary constrained MDPs using a grid-world environment, which is shown in Figure. 4.1. 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. 4.2, we can observe that our Algorithm 4 can quickly learn a well-performed policy while satisfying the safety constraint (below the threshold), while other methods all fail to satisfy the constraint.

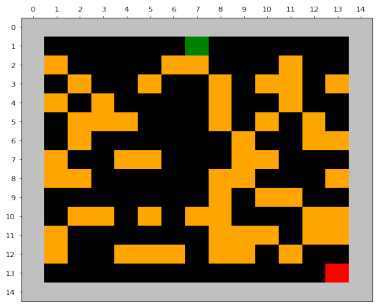

Figure 4.1: Grid World

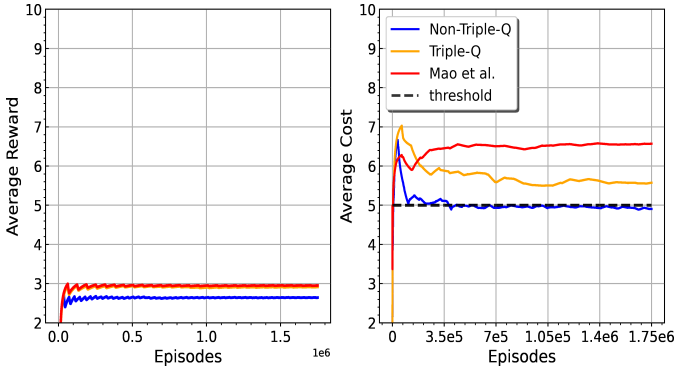

Figure 4.2: Average Reward and Cost during training

Figure 4.3: Performance of the three algorithms under a non-stationary environment

4.4 Model-free Algorithms For the Linear CMDP Setting

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 [85]. The formal definition is given below

Definition IV.6. 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 $\mu_{k,h} = \{\mu_{k,h}^1, \ldots, \mu_{k,h}^d\}$ over

$\mathcal{S}$ such that any $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

$$\mathbb{P}_{k,h}(x'|x, a) = \langle \phi(x, a), \mu_{k,h}(x') \rangle \tag{4.46}$$

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

$$r_{k,h}(x, a) = \langle \phi(x, a), \theta_{k,r,h} \rangle, \tag{4.47}$$

$$g_{k,h}(x, a) = \langle \phi(x, a), \theta_{k,g,h} \rangle \tag{4.48}$$

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 [83] 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 [19; 74] studied another related concept known as linear kernel MDP. In general, linear MDP and linear kernel MDPs are two different classes of MDP [88].

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

$$B_j = \sum_{k=2}^{K} \sum_{h=1}^{H} ||\theta_{k,j,h} - \theta_{k-1,j,h}||_2, \tag{4.49}$$

$$B_p = \sum_{k=2}^{K} \sum_{h=1}^{H} ||\mu_{k,h} - \mu_{k-1,h}||_F, \tag{4.50}$$

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

Algorithm: [85] proposed an algorithm for the stationary setup. It is a primal-dual adaptation of the unconstrained version [74]. However, there are some key differences with respect to the unconstrained case. For example, instead of a greedy policy with

Algorithm 6: Model Free Primal-Dual Algorithm for Linear Function Approximation for Non-stationary Setting

Initialization: $Y_1 = 0$, $w_{j,h} = 0$, $\alpha = \dfrac{\log(|\mathcal{A}|)K}{2(1+\xi+H)}$, $\eta = \xi/\sqrt{KH^2}$, $\beta = dH\sqrt{\log(2\log|\mathcal{A}|dT/p)}$, $D = B^{-1/2}H^{-1/2}d^{1/2}K^{1/2}$.

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**

      $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^T + \lambda\mathbb{I}$;

      $w_{r,h}^k \leftarrow (\Lambda_h^k)^{-1}[\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + V_{r,h+1}^k(x_{h+1}^\tau)]]$;

      $w_{g,h}^k \leftarrow (\Lambda_h^k)^{-1}[\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)[g_h(x_h^\tau, a_h^\tau) + V_{g,h+1}^k(x_{h+1}^\tau)]]$;

      $Q_{r,h}^k(\cdot,\cdot) \leftarrow \min\{\langle w_{r,h}^k, \phi(\cdot,\cdot)\rangle + \beta(\phi(\cdot,\cdot)^T(\Lambda_h^k)^{-1}\phi(\cdot,\cdot))^{1/2}, H\}$;

      $Q_{g,h}^k(\cdot,\cdot) \leftarrow \min\{\langle w_{g,h}^k, \phi(\cdot,\cdot)\rangle + \beta(\phi(\cdot,\cdot)^T(\Lambda_h^k)^{-1}\phi(\cdot,\cdot))^{1/2}, H\}$;

      $\pi_{h,k}(a|\cdot) = \dfrac{\exp(\alpha(Q_{r,h}^k(\cdot,a) + Y_k Q_{g,h}^k(\cdot,a)))}{\sum_a \exp(\alpha(Q_{r,h}^k(\cdot,a) + Y_k Q_{g,h}^k(\cdot,a)))}$;

      $V_{r,h}^k(\cdot) = \sum_a \pi_{h,k}(a|\cdot)Q_{r,h}^k(\cdot,a)$;

      $V_{g,h}^k(\cdot) = \sum_a \pi_{h,k}(a|\cdot)Q_{g,h}^k(\cdot,a)$;

    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 ;

    $Y_{k+1} = \max\{\min\{Y_k + \eta(\rho - V_{g,1}^k(x_1)), \xi\}, 0\}$

respect to the combined state-action value function one needs the soft-max policy. We adapt the algorithm in the non-stationary case (Algorithm 6. 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 [85] at each frame. Note that such type of restart strategy is already proposed for the unconstrained version as well [83]. 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 [34]. 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 6 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

$$w^k_{r,h}(x,a) = \frac{1}{(n^k_h(x,a)+\lambda)} \sum_{\tau=1}^{n^k_h(x,a)} (r_h(x^\tau_h, a^\tau_h) + V^k_{r,h+1}(x^\tau_{h+1})) \tag{4.51}$$

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

$$Q^k_{r,h}(x,a) = \min\{\langle w^k_{r,h}(x,a), \phi(x,a)\rangle + \beta\sqrt{1/(n^k_h(x,a)+\lambda)}, H\}. \tag{4.52}$$

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

$$w^k_{r,h}(x,a) = \frac{1}{(n^k_h(x,a)+\lambda)} \cdot \left(n^k_h(x,a) r_h(x,a) + \sum_{\tilde{x}} n^k_h(x,a,\tilde{x}) V^k_{r,h+1}(\tilde{x}) \right). \tag{4.53}$$

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.

4.4.1 Main Results

Theorem IV.7. *With $D = B^{-1/2}d^{1/2}K^{1/2}H^{-1/2}$, Algorithm 6 achieves the following regret and constraint violation bounds:*

$$\mathcal{R}(K) = \mathcal{O}(\frac{1+\delta}{\delta}K^{3/4}H^{9/4}d^{5/4}B^{1/4}\iota) \tag{4.54}$$

$$\mathcal{V}(K) = \frac{2(1+\xi)}{\xi}\mathcal{O}(K^{3/4}H^{9/4}d^{5/4}B^{1/4}\iota) \tag{4.55}$$

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 6. Note that regret and violation only scale with d rather than the cardinality of the state space.

Compared to [74], 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 [74]. The algorithm proposed in [74] 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 [74], 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 IV.8. Our algorithm 6 doesn't require the information of the local budget. In the unconstrained version [83] 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.

4.4.2 Unknown Variation Budgets

Our idea of designing the "bandit over bandit" algorithm can still be applied to the linear CMDPs, We propose an Algorithm 7, which can achieve the following result.

Theorem IV.9. *Let $D = B^{-1/2}d^{1/2}K^{1/2}H^{-1/2}, W = \sqrt{K}$, Algorithm 7 achieves the following regret and constraint violation bounds:*

$$\begin{aligned}\mathcal{R}(K) &= \mathcal{O}(\frac{1+\delta}{\delta}K^{7/8}H^{9/4}d^{5/4}B^{1/4}\iota)\\ \mathcal{V}(K) &= \frac{2(1+\xi)}{\xi}\mathcal{O}(\frac{1+\delta}{\delta}K^{7/8}H^{9/4}d^{5/4}B^{1/4}\iota)\end{aligned} \tag{4.56}$$

We can further achieve zero constraint violation by choosing

$$\epsilon = \min\{\frac{3(1+\xi)}{\xi}\tilde{O}((1+1/\delta)d^{5/4}\hat{B}^{1/4}H^{9/4}K^{1-\zeta/4})/K, \delta/2\},$$

when

$$K^8 \geq \frac{6(1+\xi)}{\xi\delta}\tilde{O}(d^{5/4}B^{1/4}H^{9/4}).$$

4.4.3 Another approach for unknown budget

We 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. 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 7 is replaced

Algorithm 7: 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. (4.148)),
$\gamma_0 = \min\left\{1, \sqrt{\frac{(K/W)\log(K/W)}{(e-1)KH}}\right\}, \lambda = 1/8$;
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 6 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
    $s_{i+1} \leftarrow s_i(j)\exp(\gamma_0 \hat{R}_i(j)/(J+1))$;

with

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

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

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

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 $D^{\dagger}$

$$\begin{aligned} &\sum_m (R_m(D^{\dagger}) + Y_m G_m(D^{\dagger})) - (R_m(D_m) + Y_m G_m(D_m)) \\ \leq &2\sqrt{e-1}WH(1+\xi)\sqrt{(K/W)(J+1)\ln(J+1)} = \tilde{\mathcal{O}}(H\xi\sqrt{KW}), \end{aligned} \tag{4.57}$$

Now, from the dual domain analysis, we obtain a similar to (Lemma IV.17)

$$\sum_m (Y - Y_m)(W\rho - G_m(D_m)) \leq \frac{Y^2 W}{2\eta} + \frac{\eta H^2 K}{2} \tag{4.58}$$

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 IV.7, we have for the optimal choice of $D^\dagger$ from $\mathcal{J}$

$$\sum_m (W\rho - G_m(D^\dagger)) \leq \tilde{\mathcal{O}}(K\sqrt{d^3H^4/D^\dagger} + D^\dagger\sqrt{dD^\dagger}H^2B). \tag{4.59}$$

$$\sum_k^K V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_m R_i(D^\dagger) \leq \tilde{\mathcal{O}}(K\sqrt{d^3H^4/D^\dagger} + D^\dagger\sqrt{dD^\dagger}H^2B). \tag{4.60}$$

Hence, we have

$$\begin{aligned}
&\sum_m -Y_m(G_m(D^\dagger) - G_m(D_m)) \\
=&\sum_m -Y_m(G_m(D^\dagger) - W\rho) + \sum_m -Y_m(W\rho - G_m(D_m)) \\
\leq&\tilde{\mathcal{O}}\left(K\sqrt{d^3H^4/D^\dagger}\xi + D^\dagger\sqrt{dD^\dagger}H^2B\xi + \xi\sqrt{WKH^2}\right)
\end{aligned} \tag{4.61}$$

where we use (4.58) (with $Y = 0$) for the first inequality, and (4.59) (where we use $|Y_m| \leq \xi$) for the second term.

Hence, from (4.57)

$$\begin{aligned}
&\sum_m (R_m(D^\dagger) - R_m(D_m)) \\
\leq&\tilde{\mathcal{O}}(H\xi 2\sqrt{e-1}WH(1+\xi)\sqrt{(K/W)(J+1)\ln(J+1)} \\
&+ \sum_m -Y_m(G_m(D^\dagger) - G_m(D_m)) \\
\leq&\tilde{\mathcal{O}}\left(K\sqrt{d^3H^4/D^\dagger}\xi + D^\dagger\sqrt{dD^\dagger}H^2B\xi + \xi\sqrt{WKH^2} + H\xi\sqrt{KW}\right)
\end{aligned} \tag{4.62}$$

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 (4.60) 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 (4.62). 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

$$\begin{aligned}
&\sum_k^K V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_m R_m(D_m) + Y\sum_m (W\rho - G_m(D_m)) \\
=&\sum_k^K V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_m R_m(D^\dagger) + \sum_m Y_m(W\rho - G_m(D^\dagger)) \\
&+ \sum_m Y_m(G_m(D^\dagger) - G_m(D_m)) + \sum_m (R_m(D^\dagger) - R_m(D_m)) \\
&+ \sum_m (Y - Y_m)(W\rho - G_m(D_m)) \\
\leq&\tilde{\mathcal{O}}\left(K\sqrt{d^3H^4/D^\dagger}\xi + D^\dagger\sqrt{dD^\dagger}H^2B\xi + \xi\sqrt{WKH^2} + H\xi\sqrt{KW}\right) \qquad (4.63)
\end{aligned}$$

where we use (4.60), (4.59), (4.57), and (4.58) to bound each term in the right-hand side respectively.

By using lemma C.6, we can have

$$\begin{aligned}
\sum_m W\rho - G_m(D_m) \leq&\tilde{\mathcal{O}}\left(K\sqrt{d^3H^4/D^\dagger} + D^\dagger\sqrt{dD^\dagger}H^2B + \sqrt{WKH^2} + H\sqrt{KW}\right. \\
&\left. + \frac{1}{\xi}(K\sqrt{d^3H^4/D^\dagger}) + D^\dagger\sqrt{dD^\dagger}H^2B)\right) \qquad (4.64)
\end{aligned}$$

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 $\xi = \frac{2H}{\gamma}$.

4.5 Proofs for the Tabular Setting

4.5.1 Proof of Theorem IV.2

Dynamic Regret

Recall that the regret can be decoupled as

$$\begin{aligned}
&\text{Regret}(K) \\
=&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_{a}\left\{Q_{k,1}^{*}q_{k,1}^{*}-Q_{k,1}^{\epsilon,*}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\right)\right]+ &(4.65)\\
&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_{a}\left\{Q_{1}^{\epsilon,*}q_{1}^{\epsilon,*}\right\}(x_{k,1},a)-Q_{k,1}(x_{k,1},a_{k,1})\right)\right]+ &(4.66)\\
&\mathbb{E}\left[\sum_{k=1}^{K}\left\{Q_{k,1}-Q_{k,1}^{\pi_k}\right\}(x_{k,1},a_{k,1})\right]. &(4.67)
\end{aligned}$$

Firstly, in lemma IV.12, we will show that the first term can be bounded by comparing the original LP associated with the tightened LP such that

$$(4.65) \leq \frac{KH\epsilon}{\delta}. \tag{4.68}$$

By using Lemma IV.11, we can show that:

$$(4.67) \leq H^2SAK^{1-\alpha}B^c + \frac{2(H^3\sqrt{\iota}+2H^4\tilde{b})K}{\chi} + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)B^c} + 2\tilde{b}H^2K$$

For the last term 4.66, we first add and subtract additional terms to obtain

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_{a}\left\{Q_{k,1}^{\epsilon,*}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)-Q_{k,1}(x_{k,1},a_{k,1})\right)\right]$$

$$=\mathbb{E}\left[\sum_k \sum_a \left(\left\{Q^{\epsilon,*}_{k,1} q^{\epsilon,*}_{k,1} + \frac{Z_k}{\eta} C^{\epsilon,*}_{k,1} q^{\epsilon,*}_{k,1}\right\}(x_{k,1}, a) - \left\{Q_{k,1} q^{\epsilon,*}_{k,1} + \frac{Z_k}{\eta} C_{k,1} q^{\epsilon,*}_{k,1}\right\}(x_{k,1}, a)\right)\right] \tag{4.69}$$

$$+\mathbb{E}\left[\sum_k \left(\sum_a \left\{Q_{k,1} q^{\epsilon,*}_{k,1}\right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1})\right)\right]$$

$$+\mathbb{E}\left[\sum_k \frac{Z_k}{\eta} \sum_a \left\{\left(C_{k,1} - C^{\epsilon,*}_{k,1}\right) q^{\epsilon,*}_{k,1}\right\}(x_{k,1}, a)\right]. \tag{4.70}$$

We can see (4.69) is the difference between two combined Q functions. In Lemma IV.10 we show that $\left\{Q_{k,h} + \frac{Z_k}{\eta} C_{k,h}\right\}(x, a)$ is an overestimate of $\left\{Q^{\epsilon,*}_{k,h} + \frac{Z_k}{\eta} C^{\epsilon,*}_{k,h}\right\}(x, a)$ (i.e. $(4.69) \leq 0$) with high probability. To bound (4.70), 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 IV.13 that the Lyapunov-drift satisfies

$$\mathbb{E}[L_{T+1} - L_T] \leq \text{a negative drift} + 2H^4\iota + 4H^4\tilde{b}^2 + \epsilon^2 - \frac{\eta B^c}{K^\alpha} \sum_{k=TK^\alpha/B^c+1}^{(T+1)K^\alpha/B^c} \Phi_k, \tag{4.71}$$

where

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

So we can bound (4.70) by applying the telescoping sum over the $K^{1-\alpha}$ frames on the inequality above:

$$(4.70) = \sum_k \Phi_k \leq \frac{K^\alpha B^c \mathbb{E}\left[L_1 - L_{K^{1-\alpha}+1}\right]}{\eta} + \frac{K(2H^4\iota + 4H^4\tilde{b}^2 + \epsilon^2)}{\eta} \leq \frac{K(2H^4\iota + 4H^4\tilde{b}^2 + \epsilon^2)}{\eta}, \tag{4.72}$$

where the last inequality holds because $L_1 = 0$ and $L_T \geq 0$ for all T. Now combining Lemma IV.10 and inequality (4.72), we conclude that

$$(4.66) \leq \frac{K(2H^4\iota + 4H^4\tilde{b}^2 + \epsilon^2)}{\eta} + \frac{(\eta + K^{1-\alpha})H^2B^c}{\eta K}. \tag{4.73}$$

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

$$\begin{aligned}\text{Regret}(K) \leq & \frac{KH\epsilon}{\delta} + H^2SAK^{1-\alpha}B^c + \frac{2(H^3\sqrt{\iota} + 2H^4\tilde{b})K}{\chi} \\ & + \sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)B^c} + 2\tilde{b}H^2K \\ & + \frac{K(2H^4\iota + 4H^4\tilde{b}^2 + \epsilon^2)}{\eta} + \frac{(\eta + K^{1-\alpha})H^2B^c}{\eta K}. \end{aligned} \tag{4.74}$$

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 $K^{1-\alpha}B^c\tilde{b} \leq B$,

$$\text{Regret}(K) = \tilde{\mathcal{O}}(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}B^{\frac{1}{3}}K^{\frac{4}{5}}). \tag{4.75}$$

Constraint Violation

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

$$Z_{T+1} = \left(Z_T + \rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}\right)^+ \geq Z_T + \rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}, \tag{4.76}$$

which implies that

$$\begin{aligned} & \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \left(-C_{k,1}^{\pi_k}(x_{k,1}, a_{k,1}) + \rho\right) \\ \leq & \frac{K^\alpha}{B^c}\left(Z_{T+1} - Z_T\right) + \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \left(\left\{C_{k,1} - C_{k,1}^{\pi_k}\right\}(x_{k,1}, a_{k,1}) - \epsilon\right). \end{aligned} \tag{4.77}$$

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

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\rho-C_{k,1}^{\pi_k}(x_{k,1},a_{k,1})\right]\\&\leq-K\epsilon+\frac{K^{\alpha}}{B^c}\mathbb{E}\left[Z_{K^{1-\alpha}B^c+1}\right]+\mathbb{E}\left[\sum_{k=1}^{K}\left\{C_{k,1}-C_{k,1}^{\pi_k}\right\}(x_{k,1},a_{k,1})\right],\end{aligned}\tag{4.78}$$

where we used the fact $Z_1=0$.

In Lemma IV.11, we established an upper bound on the estimation error of $C_{k,1}$:

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\{C_{k,1}-C_1^{\pi_k}\}(x_{k,1},a_{k,1})\right]\\&\leq H^2SAK^{1-\alpha}B^c+\frac{2(H^3\sqrt{\iota}+2H^4\tilde{b})K}{\chi}+\sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)B^c}+2\tilde{b}H^2K.\end{aligned}\tag{4.79}$$

In Lemma IV.14, 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 $1\leq T\leq K^{1-\alpha}B^c+1$

$$\begin{aligned}\mathbb{E}[Z_T]\leq&\frac{100(H^4\iota+\tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota}+\tilde{b}H^2)}{\delta}\right)\\&+\frac{4H^2B^c}{K\delta}+\frac{4H^2B^c}{\eta\delta K^{\alpha}}+\frac{4\eta(\sqrt{H^2\iota}+2H^2\tilde{b})}{\delta}.\end{aligned}\tag{4.80}$$

Substituting the results from Lemmas IV.11 and (4.80) into (4.78), 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

$$\begin{aligned}\text{Violation}(K)\leq&\frac{100(H^4\iota+\tilde{b}^2H^2)K^{0.6}}{\delta B^{2/3}}\log\frac{16(H^2\sqrt{\iota}+\tilde{b}H^2)}{\delta}+\frac{4(H^2\sqrt{\iota}+2H^2\tilde{b})}{\delta B^{1/3}}K^{0.8}\\&-5\sqrt{SAH^6\iota^3}K^{0.8}B^{\frac{1}{3}}.\end{aligned}\tag{4.81}$$

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

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

4.5.2 Proof of Theorem IV.5

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 IV.4 for any choice of $\hat{B}$, the second term is upper bounded:

$$\mathbb{E}\left[\left(\sum_{i=1}^{K/W} R_i(\hat{B}) - \sum_{i=1}^{K/W} R_i(B_i)\right)\right] \leq \tilde{\mathcal{O}}(H\sqrt{KW} + HK^{1-\lambda}). \tag{4.82}$$

For the first term, according to the regret bound analysis of Algorithm 4, we have that

$$E\left[\sum_{k=1}^{K}\left(V_{k,1}^{\pi_k^*}(x_{k,1}) - \sum_{i=1}^{K/W} R_i(\hat{B})\right)\right] \leq \tilde{\mathcal{O}}\left(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{1-0.2\zeta}\left(\hat{B}\right)^{\frac{1}{3}}\right). \tag{4.83}$$

We need to consider whether B is covered in the range of $\mathcal{J}$ to further obtain the bound of (4.83). 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 (4.83) by

$$\begin{aligned}\tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(\hat{B}\right)^{\frac{1}{3}}\right) \leq & \tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(BW^{1/J}\right)^{\frac{1}{3}}\right) \\ \leq & \tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} B^{\frac{1}{3}} K^{1-0.2\zeta}\right), \end{aligned} \tag{4.84}$$

  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 (4.83) by

$$\begin{aligned}\tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(\hat{B}\right)^{\frac{1}{3}}\right) \leq & \tilde{\mathcal{O}}\left(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{1-0.2\zeta}\left(\frac{K^{1/3}}{\Delta^{3/2}W}\right)^{\frac{1}{3}}\right) \\ \leq & \tilde{\mathcal{O}}\left(HK^{10/9-0.2\zeta}\frac{1}{K^{\zeta/3}}\right). \end{aligned} \tag{4.85}$$

  For the constraint violation, according to Lemma IV.4 we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\rho - C_{k,1}^{\pi_k}(x_{k,1}, a_{k,1})\right] = \mathbb{E}\left[\sum_{i=1}^{K/W}\left(W\rho - G_i(B_i)\right)\right] \\ =&\mathbb{E}\left[\sum_{i=1}^{K/W}\left(W\rho - G_i(\hat{B})\right)\right] + \mathbb{E}\left[\sum_{i=1}^{K/W}\left(G_i(\hat{B}) - G_i(B_i)\right)\right] \end{aligned} \tag{4.86}$$

For the first term, according to Theorem IV.2, by selecting ϵ as $\epsilon = \frac{20\sqrt{SAH^6\iota^3}\hat{B}^{1/3}}{K^{0.2\zeta}}$. we

have

$$\mathbb{E}\left[\sum_{i=1}^{K/W}\left(W\rho - G_i(\hat{B})\right)\right] \leq \frac{100(H^4\iota + \tilde{b}^2H^2)K^{0.6\zeta}}{\delta\hat{B}^{2/3}}\log\frac{16(H^2\sqrt{\iota}+H^2\tilde{b})}{\delta} - 13\sqrt{SAH^6\iota^3}K^{1-0.2\zeta}\hat{B}^{\frac{1}{3}}. \tag{4.87}$$

For the second term, we are able to obtain an upper bound by using Lemma IV.4

$$\mathbb{E}\left[\sum_{i=1}^{K/W}(G_i(\hat{B}) - G_i(B_i))\right] \leq 12K^{\lambda}H\sqrt{K^{1+\zeta}(J+1)\ln(J+1)} \tag{4.88}$$

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 $K \geq e^{\frac{1}{\delta}}$,

$$\text{Violation}(K) \leq \frac{100(H^4\iota + \tilde{b}^2H^2)K^{1/3}}{\delta\hat{B}^{2/3}}\log\frac{16(H^2\sqrt{\iota}+H^2\tilde{b})}{\delta} - \sqrt{SAH^6\iota^3}K^{8/9}\hat{B}^{\frac{1}{3}} \leq 0. \tag{4.89}$$

We finish the proof of Theorem IV.5.

4.5.3 Detailed Proofs

We provide the detailed proof in this section. A notation table and some supporting lemmas can be found in Appendix C.

4.5.3.1 Proof of Lemma IV.3

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.

$$B_h^r = \sum_{k=(T-1)K^{\alpha}/B^c+1}^{TK^{\alpha}/B^c}\sup_{x,a}|r_{k,h}(x,a) - r_{k+1,h}(x,a)| \tag{4.90}$$

$$B_h^g = \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \sup_{x,a} |g_{k,h}(x,a) - g_{k+1,h}(x,a)| \tag{4.91}$$

$$B_h^p = \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \sup_{x,a} \|\mathbb{P}_{k,h}(\cdot|x,a) - \mathbb{P}_{k+1,h}(\cdot|x,a)\|_1 \tag{4.92}$$

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

$$\begin{aligned} & Q_{k_1,h}^\pi(x,a) - Q_{k_2,h}^{\pi'}(x,a) \\ =& \mathbb{P}_{k_1,h} V_{k_1,h+1}^\pi(x,a) - \mathbb{P}_{k_2,h} V_{k_2,h+1}^{\pi'}(x,a) + r_{k_1,h}(x,a) - r_{k_2,h}(x,a) \\ \leq& \mathbb{P}_{k_1,h} V_{k_1,h+1}^\pi(x,a) - \mathbb{P}_{k_2,h} V_{k_2,h+1}^{\pi'}(x,a) + B_h^r \\ =& \sum_{x'} \mathbb{P}_{k_1,h}(x'|x,a) V_{k_1,h+1}^\pi(x') - \sum_{x'} \mathbb{P}_{k_2,h}(x'|x,a) V_{k_2,h+1}^{\pi'}(x') + B_h^r \\ =& \sum_{x'} \mathbb{P}_{k_1,h}(x'|x,a) Q_{k_1,h+1}^\pi(x', \pi_{h+1}(x')) - \sum_{x'} \mathbb{P}_{k_2,h}(x'|x,a) Q_{k_2,h+1}^{\pi'}(x', \pi'_{h+1}(x')) + B_h^r \end{aligned}$$

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

$$Q_{k_1,h+1}^\pi(x', \pi_{h+1}(x')) \leq Q_{k_2,h+1}^{\pi'}(x', \pi'_{h+1}(x')) + \sum_{h'=h+1}^{H} B_{h'}^r + H \sum_{h'=h+1}^{H} B_{h'}^p \tag{4.93}$$

Therefore

$$
\begin{aligned}
&Q^{\pi}_{k_1,h}(x,a) - Q^{\pi'}_{k_2,h}(x,a) \\
\leq &\sum_{x'} \left(\mathbb{P}_{k_1,h}(x'|x,a) - \mathbb{P}_{k_2,h}(x'|x,a)\right) Q^{\pi'}_{k_2,h+1}(x', \pi_{h+1}(x')) + \sum_{h'=h}^{H} B^r_{h'} + H \sum_{h'=h+1}^{H} B^p_{h'} \\
\leq &\|\mathbb{P}_{k_1,h}(\cdot|x,a) - \mathbb{P}_{k_2,h}(\cdot|x,a)\|_1 \cdot H + \sum_{h'=h}^{H} B^r_{h'} + H \sum_{h'=h+1}^{H} B^p_{h'} \\
\leq &B^p_h H + \sum_{h'=h}^{H} B^r_{h'} + H \sum_{h'=h+1}^{H} B^p_{h'} \\
\leq &\sum_{h'=h}^{H} B^r_{h'} + H \sum_{h'=h}^{H} B^p_{h'} \qquad (4.94)
\end{aligned}
$$

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 IV.10. *With probability at least $1 - \frac{1}{K^3}$, the following inequality holds simultaneously for all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$:*

$$\left\{F_{k,h} - F^{\pi}_{k,h}\right\}(x,a) \geq 0, \qquad (4.95)$$

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

$$\mathbb{E}\left[\sum_{k=1}^{K} \sum_{a} \left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right) q^{\epsilon,*}_{k,1}\right\}(x_{k,1}, a)\right] \leq \frac{(\eta + K^{1-\alpha}) H^2 B^c}{\eta K}. \qquad (4.96)$$

The function F will be defined in Eq.(4.97)*.*

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:

$$\begin{aligned} F_{k,h}(x,a) &= Q_{k,h}(x,a) + \frac{Z}{\eta} C_{k,h}(x,a), \quad U_{k,h}(x) = V_{k,h}(x) + \frac{Z}{\eta} W_{k,h}(x) \\ F^{\pi}_{k,h}(x,a) &= Q^{\pi}_{k,h}(x,a) + \frac{Z}{\eta} C^{\pi}_{k,h}(x,a), \quad U^{\pi}_{k,h}(x) = V^{\pi}_{k,h}(x) + \frac{Z}{\eta} W^{\pi}_{k,h}(x). \end{aligned} \tag{4.97}$$

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

$$\begin{aligned} &\{Q_{k,h} - Q^{\pi}_{k,h}\}(x,a) = \alpha_t^0 \{Q_{(T-1)K^{\alpha}/B^c+1,h} - Q^{\pi}_{k,h}\}(x,a) \\ &+ \sum_{i=1}^{t} \alpha_t^i \left(\{V_{k_i,h+1} - V^{\pi}_{k,h+1}\}(x_{k_i,h+1}) + \{(\hat{\mathbb{P}}^{k_i}_{k,h} - \mathbb{P}_{k,h}) V^{\pi}_{k,h+1}\}(x,a) + b_i + 2H\tilde{b} \right) \end{aligned} \tag{4.98}$$

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

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

$$
\begin{aligned}
&\geq_{(b)} \alpha_t^0 \left\{F_{(T-1)K^\alpha/B^c+1,h} - F^\pi_{k,h}\right\}(x,a) + \sum_{i=1}^{t} \alpha_t^i \left(\left\{U_{k_i,h+1} - U^\pi_{k,h+1}\right\}(x_{k_i,h+1}) \right. \\
&\quad \left. + \left(1+\frac{Z}{\eta}\right) H\tilde{b}\right) \\
&= \alpha_t^0 \left\{F_{(T-1)K^\alpha/B^c+1,h} - F^\pi_{k,h}\right\}(x,a) + \sum_{i=1}^{t} \alpha_t^i \left\{U_{k_i,h+1} - U^\pi_{k_i,h+1}\right\}(x_{k_i,h+1}) \\
&\quad + \sum_{i=1}^{t} \alpha_t^i \left\{U^\pi_{k_i,h+1} - U^\pi_{k,h+1}\right\}(x_{k_i,h+1}) + \left(1+\frac{Z}{\eta}\right) H\tilde{b} \\
&=_{(c)} \alpha_t^0 \left\{F_{(T-1)K^\alpha/B^c+1,h} - F^\pi_{k,h}\right\}(x,a) \\
&\quad + \sum_{i=1}^{t} \alpha_t^i \left(\max_a F_{k_i,h+1}(x_{k_i,h+1},a) - F^\pi_{k_i,h+1}(x_{k_i,h+1},\pi(x_{k_i,h+1}))\right) \\
&\quad + \sum_{i=1}^{t} \alpha_t^i \left\{U^\pi_{k_i,h+1} - U^\pi_{k,h+1}\right\}(x_{k_i,h+1}) + \left(1+\frac{Z}{\eta}\right) H\tilde{b} \\
&\geq_{(d)} \alpha_t^0 \left\{F_{(T-1)K^\alpha/B^c+1,h} - F^\pi_{k,h}\right\}(x,a) \\
&\quad + \sum_{i=1}^{t} \alpha_t^i \left(\max_a F_{k_i,h+1}(x_{k_i,h+1},a) - F^\pi_{k_i,h+1}(x_{k_i,h+1},\pi(x_{k_i,h+1}))\right) \\
&\quad - \sum_{i=1}^{t} \alpha_t^i |(1+\frac{Z}{\eta})H\tilde{b}| + (1+\frac{Z}{\eta})H\tilde{b} \\
&\geq \alpha_t^0 \left\{F_{(T-1)K^\alpha/B^c+1,h} - F^\pi_{k_i,h}\right\}(x,a) + \sum_{i=1}^{t} \alpha_t^i \left\{F_{k_i,h+1} - F^\pi_{k_i,h+1}\right\}(x_{k_i,h+1},\pi(x_{k_i,h+1})),
\end{aligned}
\tag{4.99}
$$

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 C.2 and

$$
\sum_{i=1}^{t} \alpha_t^i (1+\frac{Z}{\eta}) b_i = \frac{1}{4}\sum_{i=1}^{t} \alpha_t^i (1+\frac{Z}{\eta}) \sqrt{\frac{H^2\iota(\chi+1)}{\chi+t}} \geq \frac{\eta+Z}{4\eta}\sqrt{\frac{H^2\iota(\chi+1)}{\chi+t}}. \tag{4.100}
$$

Equality (c) holds an action is selected by maximizing $F_{k_i,h+1}(x_{k_i,h+1}, a)$, 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 IV.3 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 to other frames.

Consider $h = H$ i.e. the last step. In this case, inequality (4.99) becomes

$$\{F_{k,H} - F^{\pi}_{k,H}\}(x, a) \geq \alpha_t^0 \left\{H + \frac{Z_1}{\eta}H - F^{\pi}_{k,H}\right\}(x, a) \geq 0, \tag{4.101}$$

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

$$\{F_{k,h} - F^{\pi}_{k,h}\}(x, a) \geq 0. \tag{4.102}$$

holds for any k, h, and (x, a), which completes the proof of (4.95). Since Eq. (4.95)

can only be applied to a single policy, in order to have a bound on

$$\sum_{k=1}^{K}\sum_{a}\left\{\left(F_{k,1}^{\epsilon,*}-F_{k,1}\right)q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a),$$

we first need to substitute $F_{k,1}^{\pi}$ with $F_{k,1}^{\epsilon,*}$ in Eq. (2.46), and use a union bound over all the episodes, which means with probability at least $1-\frac{1}{K^2}$ that $F_{k,1}-F_{k,1}^{\epsilon,*}\geq 0$. Let $\mathcal{E}$ denote such event that $F_{k,h}-F_{k,h}^{\epsilon,*}\geq 0$ holds for all k, h and (x,a). Then we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F_{k,1}^{\epsilon,*}-F_{k,1}\right)q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\right]\\
=&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F_{k,1}^{\epsilon,*}-F_{k,1}\right)q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\middle|\mathcal{E}\right]\Pr(\mathcal{E})\\
&+\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F_{k,1}^{\epsilon,*}-F_{k,1}\right)q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\middle|\mathcal{E}^c\right]\Pr(\mathcal{E}^c)\\
\leq&KH\left(1+\frac{K^{1-\alpha}B^cH}{\eta}\right)\frac{1}{K^2}\leq\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}.
\end{aligned}\tag{4.103}$$

□

Lemma IV.11. *Under Algorithm 4, we have for any $T\in[K^{1-\alpha}\cdot B^c]$,*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\left\{Q_{k,1}-Q_{k,1}^{\pi_k}\right\}(x_{k,1},a_{k,1})\right]\\
\leq&H^2SA+\frac{2(H^3\sqrt{\iota}+2H^3\tilde{b})K^\alpha}{B^c\chi}+\sqrt{\frac{H^4SA\iota K^\alpha(\chi+1)}{B^c}}+\frac{2K^\alpha H^2\tilde{b}}{B^c}
\end{aligned}\tag{4.104}$$

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\left\{C_{k,1}-C_{k,1}^{\pi_k}\right\}(x_{k,1},a_{k,1})\right]\\
\leq&H^2SA+\frac{2(H^3\sqrt{\iota}+2H^3\tilde{b})K^\alpha}{B^c\chi}+\sqrt{\frac{H^4SA\iota K^\alpha(\chi+1)}{B^c}}+\frac{2K^\alpha H^2\tilde{b}}{B^c}.
\end{aligned}\tag{4.105}$$

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

$$Q_{k,h}(x,a) \leq \alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i \left(r_{k_i,h}(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i + 2H\tilde{b} \right), \tag{4.106}$$

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

$$\begin{aligned} \sum_{k=1}^{K^\alpha/B^c} Q_{k,h}(x,a) \leq & \sum_{k=1}^{K^\alpha/B^c} \alpha_t^0 H \\ & + \sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}(x,a)} \alpha_{N_{k,h}}^i \left(r_{k_i,h}(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i + 2H\tilde{b} \right). \end{aligned} \tag{4.107}$$

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

$$\begin{aligned} \sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}} \alpha_{N_{k,h}}^i V_{k_i^{(k,h)},h+1} \left(x_{k_i^{(k,h)},h+1} \right) &\leq \sum_{k=1}^{K^\alpha/B^c} V_{k,h+1}(x_{k,h+1}) \sum_{t=N_{k,h}}^{\infty} \alpha_t^{N_{k,h}} \\ &\leq \left(1 + \frac{1}{\chi}\right) \sum_k V_{k,h+1}(x_{k,h+1}), \end{aligned} \tag{4.108}$$

Then we obtain

$$\begin{aligned} & \sum_{k=1}^{K^\alpha/B^c} Q_{k,h}(x_{k,h}, a_{k,h}) \\ \leq & \sum_{k=1}^{K^\alpha/B^c} \alpha_t^0 H + (1 + \frac{1}{\chi}) \sum_{k=1}^{K^\alpha/B^c} \left(r_{k,h}(x_{k,h}, a_{k,h}) + V_{k,h+1}(x_{k,h+1}) \right) \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}} \alpha^i_{N_{k,h}} b_i + K^\alpha \tilde{b}/B^c \\
&\leq \sum_{k=1}^{K^\alpha/B^c} \left(r_{k,h}(x_{k,h}, a_{k,h}) + V_{k,h+1}(x_{k,h+1})\right) + HSA + \frac{2(H^2\sqrt{\iota} + 2H^2\tilde{b})K^\alpha}{B^c\chi} \\
&+ \frac{1}{2}\sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c, \qquad (4.109)
\end{aligned}
$$

where the last inequality holds because (i) we have

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

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

$$
\begin{aligned}
&\sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}} \alpha^i_{N_{k,h}} b_i = \frac{1}{4} \sum_{k=1}^{K^\alpha/B^c} \sum_{i=1}^{N_{k,h}} \alpha^i_{N_{k,h}} \sqrt{\frac{H^2\iota(\chi+1)}{\chi+i}} \leq \frac{1}{2} \sum_{k=1}^{K^\alpha/B^c} \sqrt{\frac{H^2\iota(\chi+1)}{\chi+N_{k,h}}} \\
=&\frac{1}{2} \sum_{x,a} \sum_{n=1}^{N_{K^\alpha/B^c,h}(x,a)} \sqrt{\frac{H^2\iota(\chi+1)}{\chi+n}} \\
\leq&\frac{1}{2} \sum_{x,a} \sum_{n=1}^{N_{K^\alpha/B^c,h}(x,a)} \sqrt{\frac{H^2\iota(\chi+1)}{n}} \overset{(1)}{\leq} \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c}, \qquad (4.110)
\end{aligned}
$$

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 $N_{K^\alpha/B^c,h} = K^\alpha/B^cSA$ for all x, a. Thus we can obtain

$$
\begin{aligned}
&\sum_{k=1}^{K^\alpha/B^c} Q_{k,h}(x_{k,h}, a_{k,h}) - \sum_k Q^{\pi_k}_{k,h}(x_{k,h}, a_{k,h}) \\
&\leq \sum_{k=1}^{K^\alpha/B^c} \left(V_{k,h+1}(x_{k,h+1}) - \mathbb{P}_{k,h} V^{\pi_k}_{k,h+1}(x_{k,h}, a_{k,h})\right) + HSA + \frac{2(H^2\sqrt{\iota} + 2H^2\tilde{b})K^\alpha}{B^c\chi} \\
&+ \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + \frac{2K^\alpha H\tilde{b}}{B^c} \qquad (4.111)
\end{aligned}
$$

$$\leq \sum_{k=1}^{K^\alpha/B^c} \left(V_{k,h+1}(x_{k,h+1}) - \mathbb{P}_{k,h}V_{h+1}^{\pi_k}(x_{k,h},a_{k,h}) + V_{k,h+1}^{\pi_k}(x_{k,h+1}) - V_{k,h+1}^{\pi_k}(x_{k,h+1})\right) \tag{4.112}$$

$$+ HSA + \frac{2(H^2\sqrt{\iota}+2H^2\tilde{b})K^\alpha}{B^c\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c$$

$$= \sum_{k=1}^{K^\alpha/B^c} \left(V_{k,h+1}(x_{k,h+1})) - V_{k,h+1}^{\pi_k}(x_{k,h+1})\right.$$

$$\left. -\mathbb{P}_{k,h}V_{k,h+1}^{\pi_k}(x_{k,h},a_{k,h}) + \hat{\mathbb{P}}_{k,h}V_{k,h+1}^{\pi_k}(x_{k,h},a_{k,h})\right)$$

$$+ HSA + \frac{2(H^2\sqrt{\iota}+2H^2\tilde{b})K^\alpha}{B^c\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c$$

$$= \sum_{k=1}^{K^\alpha/B^c} \left(Q_{k,h+1}(x_{k,h+1},a_{k,h+1}) - Q_{k,h+1}^{\pi_k}(x_{k,h+1},a_{k,h+1})\right.$$

$$\left. - \mathbb{P}_hV_{k,h+1}^{\pi_k}(x_{k,h},a_{k,h}) + \hat{\mathbb{P}}_{k,h}V_{k,h+1}^{\pi_k}(x_{k,h},a_{k,h}\right)$$

$$+ HSA + \frac{2(H^2\sqrt{\iota}+2H^2\tilde{b})K^\alpha}{B^c\chi}$$

$$+ \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c. \tag{4.113}$$

Taking the expectation on both sides yields

$$\mathbb{E}\left[\sum_{k=1}^{K^\alpha/B^c} Q_{k,h}(x_{k,h},a_{k,h}) - \sum_k Q_{k,h}^{\pi_k}(x_{k,h},a_{k,h})\right]$$

$$\leq \mathbb{E}\left[\sum_{k=1}^{K^\alpha/B^c} \left(Q_{k,h+1}(x_{k,h+1},a_{k,h+1}) - Q_{k,h+1}^{\pi_k}(x_{k,h+1},a_{k,h+1})\right)\right]$$

$$+ HSA + \frac{2(H^2\sqrt{\iota}+2H^2\tilde{b})K^\alpha}{B^c\chi} + \sqrt{H^2SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H\tilde{b}/B^c.$$

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

$$\mathbb{E}\left[\sum_{k=1}^{K^\alpha/B^c} \left(Q_{k,h}(x_{k,h},a_{k,h}) - Q_{k,h}^{\pi_k}(x_{k,h},a_{k,h})\right)\right]$$

$$\leq H^2SA + \frac{2(H^3\sqrt{\iota}+2H^3\tilde{b})K^\alpha}{B^c\chi} + \sqrt{H^4SA\iota K^\alpha(\chi+1)/B^c} + 2K^\alpha H^2\tilde{b}/B^c. \tag{4.114}$$

We finish the proof. □

Lemma IV.12. *Given $\epsilon \leq \delta$, we have*

$$\mathbb{E}\left[\sum_{a}\left\{Q_{k,1}^{*}q_{k,1}^{*}-Q_{k,1}^{\epsilon,*}q_{k,1}^{\epsilon,*}\right\}(x_{k,1},a)\right]\leq\frac{H\epsilon}{\delta}. \tag{4.115}$$

□

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 IV.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 $q_{k,h}^{\xi_2}(x,a)=(1-\frac{\epsilon}{\delta})q_{k,h}^{*}(x,a)+\frac{\epsilon}{\delta}q_{k,h}^{\xi_1}(x,a)$, which satisfies that

$$\begin{aligned}
\sum_{h,x,a}q_{k,h}^{\xi_2}(x,a)g_{k,h}(x,a)&=\sum_{h,x,a}\left((1-\frac{\epsilon}{\delta})q_{k,h}^{*}(x,a)+\frac{\epsilon}{\delta}q_{k,h}^{\xi_1}(x,a)\right)g_{k,h}(x,a)\geq\rho+\epsilon,\\
\sum_{h,x,a}q_{k,h}^{\xi_2}(x,a)&=\sum_{x',a'}\mathbb{P}_{k,h-1}(x|x',a')q_{k,h-1}^{\xi_2}(x',a'),\\
\sum_{h,x,a}q_{k,h}^{\xi_2}(x,a)&=1.
\end{aligned} \tag{4.116}$$

Also we have $q_{k,h}^{\xi_2}(x,a)\geq 0$ for all (h,x,a). Thus $\{q_{k,h}^{\xi_2}(x,a)\}_{h=1}^{H}$ is a feasible solution to the ϵ-tightened optimization problem (4.21). Then given $\{q_{k,h}^{\epsilon,*}(x,a)\}_{h=1}^{H}$ is the optimal solution to the ϵ-tightened optimization problem, we have

$$\begin{aligned}
&\sum_{h,x,a}\left(q_{k,h}^{*}(x,a)-q_{k,h}^{\epsilon,*}(x,a)\right)r_{k,h}(x,a)\\
\leq&\sum_{h,x,a}\left(q_{k,h}^{*}(x,a)-q_{k,h}^{\xi_2}(x,a)\right)r_{k,h}(x,a)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{h,x,a} \left(q^*_{k,h}(x,a) - \left(1 - \frac{\epsilon}{\delta}\right) q^*_{k,h}(x,a) - \frac{\epsilon}{\delta} q^{\xi_1}_{k,h}(x,a) \right) r_{k,h}(x,a) \\
&\leq \sum_{h,x,a} \left(q^*_{k,h}(x,a) - \left(1 - \frac{\epsilon}{\delta}\right) q^*_{k,h}(x,a) \right) r_{k,h}(x,a) \\
&\leq \frac{\epsilon}{\delta} \sum_{h,x,a} q^*_{k,h}(x,a) r_{k,h}(x,a) \leq \frac{H\epsilon}{\delta}, \qquad (4.117)
\end{aligned}
$$

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

$$
\sum_a Q^*_{k,1}(x_{k,1},a) q^*_{k,1}(x_{k,1},a) = \sum_{h,x,a} q^*_{k,h}(x,a) r_{k,h}(x,a) \qquad (4.118)
$$

$$
\sum_a Q^{\epsilon,*}_{k,1}(x_{k,1},a) q^{\epsilon,*}_{k,1}(x_{k,1},a) = \sum_{h,x,a} q^{\epsilon,*}_{k,h}(x,a) r_{k,h}(x,a). \qquad (4.119)
$$

□

Lemma IV.13. *Assume $\epsilon \leq \delta$. The expected Lyapunov drift satisfies*

$$
\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\
\leq &\frac{B^c}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(-\eta \mathbb{E}\left[\sum_a \left\{ \hat{Q}_{k,1} q_1^{\epsilon,*} \right\}(x_{k,1},a) - \hat{Q}_{k,1}(x_{k,1},a_{k,1}) \middle| Z_T = z \right] \right. \\
&\left. + z\mathbb{E}\left[\sum_a \left\{ \left(C^{\epsilon,*}_{k,1} - C_{k,1}\right) q^{\epsilon,*}_{k,1} \right\}(x_{k,1},a) \middle| Z_T = z \right] \right) + 2H^4\iota + 4H^4\tilde{b} + \epsilon^2. \qquad (4.120)
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\
\leq &\frac{B^c}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(-\eta \mathbb{E}\left[\sum_a \left\{ \hat{Q}_{k,1} q_1^{\epsilon,*} \right\}(x_{k,1},a) - \hat{Q}_{k,1}(x_{k,1},a_{k,1}) \middle| Z_T = z \right] \right. \\
&\left. + z\mathbb{E}\left[\sum_a \left\{ \left(C^{\epsilon,*}_{k,1} - C_{k,1}\right) q^{\epsilon,*}_{k,1} \right\}(x_{k,1},a) \middle| Z_T = z \right] \right) + 2H^4\iota + 4H^4\tilde{b} + \epsilon^2. \qquad (4.121)
\end{aligned}
$$

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

$$\begin{aligned} L_{T+1} - L_T \leq & Z_T \left(\rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}\right) + \frac{\left(\frac{\bar{C}_T B^c}{K^\alpha} + \epsilon - \rho\right)^2}{2} \\ \leq & Z_T \left(\rho + \epsilon - \frac{\bar{C}_T B^c}{K^\alpha}\right) + 2H^4\iota + 4H^4\tilde{b} + \epsilon^2 \\ \leq & \frac{Z_T B^c}{K^\alpha} \sum_{k=TK^\alpha/B^c+1}^{(T+1)K^\alpha/B^c} \left(\rho + \epsilon - \hat{C}_{k,1}(x_{k,1}, a_{k,1})\right) + 2H^4\iota + 4H^4\tilde{b} + \epsilon^2 \quad (4.122) \end{aligned}$$

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 C.1. Let $\{q^\epsilon_{k,h}\}_{h=1}^H$ be a feasible solution to the tightened LP (4.21) at episode k. Then the expected Lyapunov drift conditioned on $Z_T = z$ is

$$\begin{aligned} & \mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\ \leq & \frac{B^c}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha/B^c} \Big(\mathbb{E}\left[z\left(\rho + \epsilon - \hat{C}_{k,1}(x_{k,1}, a_{k,1})\right) - \eta\hat{Q}_{k,1}(x_{k,1}, a_{k,1}) \middle| Z_T = z\right] \\ & + \eta\mathbb{E}\left[\hat{Q}_{k,1}(x_{k,1}, a_{k,1}) \middle| Z_T = z\right]\Big) + 2H^4\iota + 4H^4\tilde{b} + \epsilon^2. \quad (4.123) \end{aligned}$$

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

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

$$
\begin{aligned}
=&\mathbb{E}\left[z\left(\rho+\epsilon-\sum_a C^{\epsilon}_{k,1}(x_{k,1},a)q^{\epsilon}_{k,1}(x_{k,1},a)\right)\middle| Z_T=z\right]\\
&-\mathbb{E}\left[\eta\sum_a \hat{Q}_{k,1}(x_{k,1},a)q^{\epsilon}_{k,1}(x_{k,1},a)-\eta\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle| Z_T=z\right]\\
&+\mathbb{E}\left[z\sum_a\left\{(C^{\epsilon}_{k,1}-\hat{C}_{k,1})q^{\epsilon}_{k,1}\right\}(x_{k,1},a)\middle| Z_T=z\right]\\
\leq&-\eta\mathbb{E}\left[\sum_a \hat{Q}_{k,1}(x_{k,1},a)q^{\epsilon}_{k,1}(x_{k,1},a)-\hat{Q}_{k,1}(x_{k,1},a_{k,1})\middle| Z_T=z\right]\\
&+\mathbb{E}\left[z\sum_a\left\{(C^{\epsilon}_{k,1}-\hat{C}_{k,1})q^{\epsilon}_{k,1}\right\}(x_{k,1},a)\middle| Z_T=z\right], \qquad (4.124)
\end{aligned}
$$

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 (4.21), 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 IV.14. *Assuming $\epsilon\leq\frac{\delta}{2}$, we have for any $1\leq T\leq K^{1-\alpha}\cdot B^c$*

$$
\begin{aligned}
\mathbb{E}[Z_T]\leq&\frac{100(H^4\iota+\tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota}+\tilde{b}H^2)}{\delta}\right)+\frac{4H^2B^c}{K\delta}\\
&+\frac{4H^2B^c}{\eta\delta K^{\alpha}}+\frac{4\eta(\sqrt{H^2\iota}+2H^2\tilde{b})}{\delta}. \qquad (4.125)
\end{aligned}
$$

The proof will also use the following lemma from [62].

Lemma IV.15. *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 IV.14. We apply Lemma IV.15 to a new Lyapunov function:

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

To verify condition (1) in Lemma IV.15, consider

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

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

$$\begin{aligned}
&\mathbb{E}\left[Z_{T+1} - Z_T | Z_T = z\right] \\
=&\mathbb{E}\left[\left.\sqrt{Z_{T+1}^2} - \sqrt{z^2}\right| Z_T = z\right] \\
\leq& \frac{1}{2z}\mathbb{E}\left[\left. Z_{T+1}^2 - z^2 \right| Z_T = z\right] \\
\leq_{(a)}& -\frac{\delta}{2} + \frac{4(\frac{(\eta + K^{1-\alpha})H^2 B^c}{\eta K} + \eta(\sqrt{H^2\iota} + 2H^2\tilde{b}) + H^4\iota + \epsilon^2 + 2H^4\tilde{b}^2)}{z} \\
\leq& -\frac{\delta}{2} + \frac{4(\frac{(\eta + K^{1-\alpha})H^2 B^c}{\eta K} + \eta(\sqrt{H^2\iota} + 2H^2\tilde{b}) + H^4\iota + \epsilon^2 + 2H^4\tilde{b}^2}{\theta_T} \\
=& -\frac{\delta}{4},
\end{aligned} \tag{4.126}$$

where inequality (a) is obtained according to Lemma IV.16; and the last inequality holds given $z \geq \theta_T$.

To verify condition (2) in Lemma IV.15, we have

$$\begin{aligned} Z_{T+1} - Z_T \leq |Z_{T+1} - Z_T| \leq \left|\rho + \epsilon - \bar{C}_T\right| \\ \leq (H + H^2\sqrt{\iota} + 2\tilde{b}H^2) + \epsilon \leq 2(H^2\sqrt{\iota} + \tilde{b}H^2), \end{aligned} \tag{4.127}$$

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 IV.15, we obtain

$$\mathbb{E}\left[e^{rZ_T}\right] \leq e^{rZ_1} + \frac{2e^{r(v+\theta_T)}}{r\gamma}, \quad \text{where} \quad r = \frac{\gamma}{v^2 + v\gamma/3}. \tag{4.128}$$

By Jensen's inequality, we have

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

which implies that

$$\begin{aligned} &\mathbb{E}[Z_T] \leq \frac{1}{r}\log\left(1 + \frac{2e^{r(v+\theta_T)}}{r\gamma}\right) \\ =&\frac{1}{r}\log\left(1 + \frac{6v^2 + 2v\gamma}{3\gamma^2}e^{r(v+\theta_T)}\right) \\ \leq&\frac{1}{r}\log\left(1 + \frac{8v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\ \leq&\frac{1}{r}\log\left(\frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\ \leq&\frac{4v^2}{3\gamma}\log\left(\frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\ \leq&\frac{3v^2}{\gamma}\log\left(\frac{2v}{\gamma}\right) + v + \theta_T \\ \leq&\frac{3v^2}{\gamma}\log\left(\frac{2v}{\gamma}\right) + v \\ &+ \frac{4(\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K} + \eta(\sqrt{H^2\iota} + 2H^2\tilde{b}) + H^4\iota + \epsilon^2 + 2H^4\tilde{b}^2)}{\delta} \\ =&\frac{96(H^4\iota + \tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota} + \tilde{b}H^2)}{\delta}\right) + 2(H^2\sqrt{\iota} + \tilde{b}H^2) \end{aligned}$$

$$+\frac{4(\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}+\eta(\sqrt{H^2\iota}+2H^2\tilde{b})+H^4\iota+\epsilon^2+2H^4\tilde{b}^2)}{\delta}$$
$$\leq\frac{100(H^4\iota+\tilde{b}^2H^2)}{\delta}\log\left(\frac{16(H^2\sqrt{\iota}+\tilde{b}H^2)}{\delta}\right)+\frac{4H^2B^c}{K\delta}+\frac{4H^2B^c}{\eta\delta K^\alpha}+\frac{4\eta(\sqrt{H^2\iota}+2H^2\tilde{b})}{\delta}, \tag{4.129}$$

which completes the proof of Lemma IV.14. □

Lemma IV.16. *Given $\delta \geq 2\epsilon$, under our algorithm 4, the conditional expected drift is*

$$\mathbb{E}\left[L_{T+1}-L_T|Z_T=z\right]\leq-\frac{\delta}{2}z+\frac{(\eta+K^{1-\alpha})H^2B^c}{\eta K}$$
$$+\eta(\sqrt{H^2\iota}+2H^2\tilde{b})+H^4\iota+\epsilon^2+2H^4\tilde{b}^2 \tag{4.130}$$

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 (4.123), we have

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

$$
\begin{aligned}
\leq & \frac{B^c}{K^\alpha} \sum_{k=(T-1)/B^cK^\alpha+1}^{TK^\alpha/B^c} \mathbb{E}\left[Z_T \left(\rho + \epsilon - \sum_a \left\{ C^\pi_{k,1} q^\pi_{k,1} \right\}(x_{k,1}, a) \right) \right. \\
& \left. -\eta \sum_a \{Q_{k,1} q^\pi_{k,1}\}(x_{k,1}, a) + \eta Q_{k,1}(x_{k,1}, a_{k,1}) | Z_T = z \right] \\
& + \frac{B^c}{K^\alpha} \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \mathbb{E}\left[Z_T \sum_a \left\{ C^\pi_{k,1} q^\pi_{k,1} \right\}(x_{k,1}, a) \right. \\
& \left. - Z_T \sum_a \left\{ C_{k,1} q^\pi_{k,1} \right\}(x_{k,1}, a) | Z_T = z \right] \\
& + \frac{B^c}{K^\alpha} \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \mathbb{E}\left[\eta \sum_a \left\{ Q^\pi_{k,1} q^\pi_{k,1} \right\}(x_{k,1}, a) - \eta \sum_a \left\{ Q^\pi_{k,1} q^\pi_{k,1} \right\}(x_{k,1}, a) | Z_T = z \right] \\
& + H^4 \iota + \epsilon^2 + 2H^4 \tilde{b}^2 \\
\leq_{(b)} & -\frac{\delta}{2} z + \frac{B^c}{K^\alpha} \sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c} \mathbb{E}\left[\eta \sum_a \left\{ (F^\pi_{k,1} - F_{k,1}) q^\pi_{k,1} \right\}(x_{k,1}, a) \right. \\
& \left. + \eta Q_{k,1}(x_{k,1}, a_{k,1}) |\, Z_T = z \right] \\
& + H^4 \iota + \epsilon^2 + 2H^4 \tilde{b}^2 \\
\leq_{(c)} & -\frac{\delta}{2} z + \frac{(\eta + K^{1-\alpha}) H^2 B^c}{\eta K} + \eta(\sqrt{H^2 \iota} + 2H^2 \tilde{b}) + H^4 \iota + \epsilon^2 + 2H^4 \tilde{b}^2. \qquad (4.131)
\end{aligned}
$$

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

$$
\begin{aligned}
& \epsilon + \rho - \mathbb{E}\left[\sum_a C^\pi_{k,1}(x_{k,1}, a) q^\pi_{k,1}(x_{k,1}, a) \right] \\
=& \rho + \epsilon - \mathbb{E}\left[\sum_{h,x,a} q^\pi_{k,h}(x, a) g_{k,h}(x, a) \right] \leq -\delta + \epsilon \leq -\frac{\delta}{2}. \qquad (4.132)
\end{aligned}
$$

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

by using Lemma C.1, and the fact that

$$\mathbb{E}\left[\sum_{k=(T-1)K^\alpha/B^c+1}^{TK^\alpha/B^c}\sum_a\left\{(F_{k,1}^\pi - F_{k,1})q_{k,1}^\pi\right\}(x_{k,1},a)\middle| Z_T=z\right] \tag{4.133}$$

can be bounded as (4.103) (note that the overestimation result and the concentration result in frame T hold regardless of the value of Z_T). □

4.5.3.2 Proof of Lemma IV.4

Proof. Let $R_i(B_i)(G_i(B_i))$ be the cumulative reward(utility) collected in epoch i by the given algorithm with the estimated 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

$$\mathbb{E}\left[\sum_{i=1}^{K/W}(R_i(\hat{B})-R_i(B_i))\right] =\tilde{\mathcal{O}}(H\sqrt{KW}+HK^{1-\lambda}) \tag{4.134}$$

$$\mathbb{E}\left[\sum_{i=1}^{K/W}G_i(\hat{B})-G_i(B_i)\right] =\tilde{\mathcal{O}}(HK^{\lambda}\sqrt{KW}) \tag{4.135}$$

Apply the regret bound of the Exp3 algorithm, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{K/W}(f_r(R_i(\hat{B}))+f_g(G_i(\hat{B}))-\sum_{i=1}^{K/W}(f_r(R_i(B_i))+f_g(G_i(B_i))\right]\\ \leq& 2\sqrt{e-1}WH(1+1/K^\lambda)\sqrt{(K/W)(J+1)\ln(J+1)}=\tilde{\mathcal{O}}(H\sqrt{KW}),\end{aligned} \tag{4.136}$$

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

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

$$
\begin{aligned}
&+\mathbb{E}\left[\sum_{i=1}^{K/W}(f_g(G_i(B_i))-f_g(G_i(\hat{B})))\right]\\
\leq& 2\sqrt{e-1}WH(1+1/K^{\lambda})\sqrt{(K/W)(J+1)\ln(J+1)}+\frac{WH}{K^{\lambda}}\cdot\frac{K}{W}\\
=&\tilde{\mathcal{O}}(H\sqrt{KW}+HK^{1-\lambda}),
\end{aligned}
\tag{4.137}
$$

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

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K/W}G_i(\hat{B})-G_i(B_i)\right]=K^{\lambda}\mathbb{E}\left[\sum_{i=1}^{K/W}\frac{G_i(\hat{B})-G_i(B_i)}{K^{\lambda}}\right]\\
=&K^{\lambda}\mathbb{E}\left[\sum_{i=1}^{K/W}f_g(G_i(\hat{B}))-f_g(G_i(B_i))\right]\\
\leq&K^{\lambda}\left(2\sqrt{e-1}WH(1+1/K^{\lambda})\sqrt{(K/W)(J+1)\ln(J+1)}\right.\\
&\left.+\mathbb{E}\left[\sum_{i=1}^{K/W}(f_r(R_i(B_i))-f_r(R_i(\hat{B})))\right]\right)\\
\leq&K^{\lambda}\left(2\sqrt{e-1}WH(1+1/K^{\lambda})\sqrt{(K/W)(J+1)\ln(J+1)}\right)\\
=&\tilde{\mathcal{O}}(HK^{\lambda}\sqrt{KW}),
\end{aligned}
\tag{4.138}
$$

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 $\mathbb{E}[f_r(R_i(B_i))]=0$. □

4.6 Proofs for the Linear Apprroximation Setting

4.6.1 Proof of Theorem IV.7

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

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

Outline of Proof of Theorem IV.7

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

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

$$\sum_{k=1}^{K}(V^{\pi^*_k}_{k,r,1}(x_1) - V^{\pi_k}_{k,r,1}(x_1)) + Y\sum_{k=1}^{K}(\rho - V^{\pi_k}_{k,g,1}(x_1)) \leq \frac{1}{2\eta}Y^2 + \frac{\eta}{2}H^2K + \underbrace{\sum_{k=1}^{K}\Big(V^{\pi^*_k}_{k,r,1}(x_1) + Y_k V^{\pi^*_k}_{k,g,1}(x_1)\Big) - \big(V^k_{r,1}(x_1) + Y_k V^k_{g,1}(x_1)\big)}_{\mathcal{T}_1} + \underbrace{\sum_{k=1}^{K}\big(V^k_{r,1}(x_1) - V^{\pi_k}_{k,r,1}(x_1)\big) + Y\sum_{k=1}^{K}\big(V^k_{g,1}(x_1) - V^{\pi_k}_{k,g,1}(x_1)\big)}_{\mathcal{T}_2} \tag{4.139}$$

Note that when $Y = 0$, we recover the dynamic regret.

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

Lemma IV.18. *With probability* $1-2p$,

$$\mathcal{T}_1 \leq H^3(1+2/\delta)BD^{3/2}\sqrt{d} + \frac{KH\log(|\mathcal{A}|)}{\alpha} \tag{4.140}$$

$$\mathcal{T}_2 \leq (1+Y)(\mathcal{O}(\sqrt{H^4d^3K^2\iota^2/D}) + \sqrt{d}D^{3/2}BH^2) \tag{4.141}$$

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

Note from Lemma IV.17 we have

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

From Lemma IV.18, we obtain

$$\sum_{k=1}^{K}(V_{k,r,1}^{\pi_k^*}(x_1) - V_{k,r,1}^{\pi_k}(x_1)) + Y(\rho - V_{k,g,1}^{\pi_k}(x_1)) \leq \frac{Y^2}{2\eta} + \frac{\eta KH^2}{2} + \\ \frac{HK\log(|\mathcal{A}|)}{\alpha} + H^3(1+2/\delta)BD^{3/2}\sqrt{d} + (1+Y)(\mathcal{O}(\sqrt{H^4d^3K^2\iota^2/D}) + \sqrt{d}D^{3/2}BH^2) \tag{4.142}$$

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

$$\begin{aligned}\sum_{k=1}^{K}(V_{k,r,1}^{\pi_k^*}(x_1) - V_{k,r,1}^{\pi_k}(x_1)) + Y(b - V_{k,g,1}^{\pi_k}(x_1)) &\leq \xi\sqrt{KH^2} \\ &+ H2(1+\xi+H) + H^{9/4}(1+2/\delta)B^{1/4}K^{3/4}d^{5/4} \\ &+ (Y+1)(\mathcal{O}(H^{9/4}d^{5/4}K^{3/4}B^{1/4}\iota^2) + H^{5/4}d^{5/4}K^{3/4})\end{aligned} \tag{4.143}$$

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

$$\text{Regret}(K) \leq \mathcal{O}(H^{9/4}d^{5/4}K^{3/4}B^{1/4}\iota^2) + \mathcal{O}((1+1/\delta)H^{9/4}d^{5/4}K^{3/4}B^{1/4}) \tag{4.144}$$

For the constraint violation bound, we use the following Lemma IV.19 (Lemma J.10

in [74]).

Lemma IV.19. *Let $\bar{C}^* \geq 2\max_k \mu^{k,*}$, then, if*

$$\sum_{k=1}^{K}(V_{k,r,1}^{\pi_k^*}(x_1) - V_{k,r,1}^{\pi_k}(x_1)) + 2\bar{C}^* \sum_{k=1}^{K}(b_k - V_{k,g,1}^{\pi_k}(x_1)) \leq \delta \tag{4.145}$$

, then

$$\sum_{k=1}^{K}(b_k - V_{k,g,1}^{\pi_k}(x_1)) \leq \frac{2\delta}{\bar{C}^*} \tag{4.146}$$

Note that $\xi \geq 2\max_k \mu^{k,*}$. Thus, we replace $Y = \xi$ in (4.143). Thus, from (4.143) and Lemma IV.19, we obtain

$$\sum_{k=1}^{K}(\rho - V_{k,g,1}^{\pi}(x_1)) \leq \frac{2(1+\xi)}{\xi}(\mathcal{O}(H^{9/4}d^{5/4}K^{3/4}B^{1/4}\iota^2) + \mathcal{O}(H^{5/4}d^{5/4}K^{3/4}B^{1/4})) \tag{4.147}$$

Hence, the result follows. □

4.6.2 Proofs of TheoremIV.9

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

$$\mathcal{J} = \left\{\frac{\sqrt{K}}{\Delta W}, \frac{\sqrt{K}W^{\frac{1}{J}}}{\Delta W}, \frac{\sqrt{K}W^{\frac{2}{J}}}{\Delta W}, \ldots, \frac{\sqrt{K}W}{\Delta W}\right\}, \Delta = \left(\frac{6(1+\xi)}{\xi\delta}\tilde{\mathcal{O}}((1+\delta)d^{5/4}H^{9/4})\right)^4 \tag{4.148}$$

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 IV.4 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 6, we have for the W episodes

$$E\left[\sum_{k=1}^{W}\left(V_{k,1}^{\pi_k^*}(x_{k,1}) - R_i(\hat{D})\right)\right] \leq \tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}\hat{B}^{1/4}\right). \tag{4.149}$$

We need to consider whether $\hat{B}$ is covered in the range of $\mathcal{J}$ to further obtain the bound of (4.149). 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 (4.149) by

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

$$=\tilde{\mathcal{O}}\left(\frac{1+\delta}{\delta}K^{1-\frac{1\zeta}{4}}H^{9/4}d^{5/4}B^{1/4}\right)$$

- 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 (4.149) 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 IV.4 we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\rho - C_{k,1}^{\pi_k}(x_{k,1},a_{k,1})\right] = \mathbb{E}\left[\sum_{i=1}^{K/W}(W\rho - G_i(B_i))\right]\\ =&\mathbb{E}\left[\sum_{i=1}^{K/W}\left(W\rho - G_i(\hat{B})\right)\right] + \mathbb{E}\left[\sum_{i=1}^{K/W}\left(G_i(\hat{B}) - G_i(B_i)\right)\right] \qquad (4.150)\end{aligned}$$

For the first term, according to Theorem IV.7, by selecting $\epsilon = \frac{3(1+\xi)}{\xi}\tilde{O}((1+1/\delta)d^{5/4}\hat{B}^{1/4}H^{9/4}K^{1-\zeta/4})/K$, we have

$$\mathbb{E}\left[\sum_{i=1}^{K/W}\left(W\rho - G_i(\hat{B})\right)\right] \leq -\frac{(1+\xi)}{\xi}\tilde{\mathcal{O}}((1+1/\delta)K^{1-\zeta/4}H^{9/4}d^{5/4}\hat{B}^{1/4}). \qquad (4.151)$$

For the second term, we are able to obtain an upper bound by using Lemma IV.4

$$\mathbb{E}\left[\sum_{i=1}^{K/W}(G_i(\hat{B}) - G_i(B_i))\right] \leq 12K^{\lambda}H\sqrt{K^{1+\zeta}(J+1)\ln(J+1)} \qquad (4.152)$$

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

$$\text{Violation}(K) = 0. \tag{4.153}$$

We finish the proof of Theorem IV.9.

4.6.3 Detailed Proofs

Notations: We describe the specific notations we have used in this section. A detailed notation table and some supporting lemmas can be found in Appendix C. With slight abuse of notations, in this section, we denote $V^{\pi}_{k,r,h}$ as the value function at step h for policy π at episode k. We denote $V^{\pi}_{k,g,h}$ as the utility value function at step h of episode k. We denote $Q^{\pi}_{k,j,h}$, $j = r, g$ as the state-action value function at step j for policy π.

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

4.6.3.1 Proof of Lemma IV.17

We first state and prove the following result which is similar to the one proved in [85].

Lemma IV.20. *For $Y \in [0, \xi]$,*

$$\sum_{k=1}^{K} (Y - Y_k)(\rho - V^k_{g,1}(x_1)) \leq \frac{Y^2}{2\eta} + \frac{\eta H^2 K}{2} \tag{4.154}$$

Proof.

$$|Y_{k+1} - Y|^2 = |Proj_{[0,\xi]}(Y_k + \eta(\rho - V^k_{g,1}(x_1))) - Proj_{[0,\xi]}(Y)|^2$$

$$\leq (Y_k + \eta(\rho - V_{g,1}^k(x_1))) - Y)^2$$
$$\leq (Y_k - Y)^2 + \eta^2 H^2 + 2\eta Y_k(\rho - V_{g,1}^k(x_1)) \tag{4.155}$$

Summing over k, we obtain

$$0 \leq |Y_{K+1} - Y|^2 \leq |Y_1 - Y|^2 + 2\eta \sum_{k=1}^{K} (\rho - V_{g,1}^k(x_1))(Y_k - Y) + \eta^2 H^2 K$$
$$\sum_{k=1}^{K} (Y - Y_k)(\rho - V_{g,1}^k(x_1)) \leq \frac{|Y_1 - Y|^2}{2\eta} + \frac{\eta H^2 K}{2} \tag{4.156}$$

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

Now, we prove Lemma IV.17.

Proof. Note that

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

where the first inequality follows from Lemma IV.20, and the second inequality follows from the fact that $V_{k,g,1}^{\pi_k^*}(x_1) \geq \rho$. Hence, the result simply follows from the above inequality. □

4.6.3.2 Proof of Lemma IV.18

We now move on to bound $\mathcal{T}_1$ and $\mathcal{T}_2$. First, we state and prove Lemmas IV.21, IV.22, IV.23, IV.24,IV.26, and IV.27.

Lemma IV.21. *There exists a constant C_2 such that for any fixed $p \in (0,1)$, if we let E be the event that*

$$\| \sum_{\tau=1}^{k-1} \phi_{j,h}^{\tau} [V_{j,h+1}^{k}(x_{h+1}^{\tau}) - \mathbb{P}_{k,h} V_{j,h+1}^{k}(x_h^{\tau}, a_h^{\tau})] \|_{(\Lambda_h^k)^{-1}} \leq C_2 dH\sqrt{\chi} \tag{4.157}$$

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 [58]. The proof relies on the uniform concentration lemma similar to [58]. However, there is an additional $\log(|\mathcal{A}|)$ in χ. This arises due to the fact that the policy (Algorithm 6) is soft-max, unlike the greedy policy in [58]. [85] shows that greedy policy is unable to prove the uniform concentration lemma. The proof is similar to Lemma 8 in [85], thus, we remove it.

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

$$B_{j,\mathcal{E}}^{k} = \sum_{\tau=2}^{k} \sum_{h=1}^{H} ||\theta_{\tau,j,h} - \theta_{\tau-1,j,h}||, B_j^{\mathcal{E}} = \sum_{\tau=2}^{\mathcal{E}} \sum_{h=1}^{H} ||\theta_{\tau,j,h} - \theta_{\tau-1,j,h}||$$
$$B_{p,\mathcal{E}}^{k} = \sum_{\tau=2}^{k} \sum_{h=1}^{H} ||\mu_{\tau,h} - \mu_{\tau-1,h}||, B_p^{\mathcal{E}} = \sum_{\tau=2}^{\mathcal{E}} \sum_{h=1}^{H} ||\mu_{\tau,h} - \mu_{\tau-1,h}||$$

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

Now, we are bound the difference between our estimated $Q_{j,h}^{k}$ and $Q_{k,j,h}^{\pi}$. Using the Lemma IV.21, we show the following

Lemma IV.22. *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 IV.21, we have*

$$\begin{aligned} &\langle \phi(x,a), w_{j,h}^{k} \rangle - Q_{k,j,h}^{\pi}(x,a) \\ =&\mathbb{P}_{k,h}(V_{j,h+1}^{k} - V_{k,j,h+1}^{\pi})(x,a) + \Delta_h^k(x,a) + + B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD} \end{aligned} \tag{4.158}$$

for some $\Delta_h^k(x,a)$ *that satisfies* $|\Delta_h^k(x,a)| \leq \beta\sqrt{\phi(x,a)^T(\Lambda_h^k)^{-1}\phi(x,a)}$*, for any* $k \in \mathcal{E}$*.*

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

Note that $Q_{k,r,h}^{\pi}(x,a) = \langle \phi(x,a), w_{r,h}^{\pi} \rangle = r_{k,h}(x,a) + \mathbb{P}_{k,h} V_{k,r,h+1}^{\pi}(x,a)$.

Hence, we have

$$\begin{aligned} w_{r,h}^k - w_{k,r,h}^{\pi} &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[r_h^{\tau} + V_{r,h+1}^k(x_{h+1}^{\tau})] - w_{k,r,h}^{\pi} \\ &= -\lambda(\Lambda_h^k)^{-1}(w_{k,r,h}^{\pi}) + (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[r_{\tau,h}(x_h^{\tau}, a_h^{\tau}) + V_{r,h+1}^k - r_{k,h}(x_h^{\tau}, a_h^{\tau}) - \mathbb{P}_{k,h} V_{k,r,h+1}^{\pi}] \end{aligned} \tag{4.159}$$

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

$$\begin{aligned} &(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[r_{\tau,h}(x_h^{\tau}, a_h^{\tau}) + V_{r,h+1}^k - r_{k,h}(x_h^{\tau}, a_h^{\tau}) - \mathbb{P}_{k,h} V_{k,r,h+1}^{\pi}] \\ &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[r_{\tau,h}(x_h^{\tau}, a_h^{\tau}) + V_{r,h+1}^k - r_{k,h}(x_h^{\tau}, a_h^{\tau}) - \mathbb{P}_{k,h} V_{r,h+1}^k] \\ &+ (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[\mathbb{P}_{k,h} V_{r,h+1}^k - \mathbb{P}_{k,h} V_{k,r,h+1}^{\pi}] \\ &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[r_{\tau,h}(x_h^{\tau}, a_h^{\tau}) - r_{k,h}(x_h^{\tau}, a_h^{\tau})] + (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} [V_{r,h+1}^k - \mathbb{P}_{\tau,h} V_{r,h+1}^k] \\ &+ (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} [\mathbb{P}_{\tau,h} V_{r,h+1}^k - \mathbb{P}_{k,h} V_{r,h+1}^k] + (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[\mathbb{P}_{k,h} V_{r,h+1}^k - \mathbb{P}_{k,h} V_{k,r,h+1}^{\pi}] \end{aligned} \tag{4.160}$$

By plugging in the above in (4.159) we obtain

$$\begin{aligned} &w_{r,h}^k - w_{k,r,h}^{\pi} \\ &= \underbrace{-\lambda(\Lambda_h^k)^{-1}(w_{k,r,h}^{\pi})}_{q_1} + \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[r_{\tau,h}(x_h^{\tau}, a_h^{\tau}) - r_{k,h}(x_h^{\tau}, a_h^{\tau})]}_{q_2} \end{aligned} \tag{4.161}$$

$$+\underbrace{(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}[V_{r,h+1}^k - \mathbb{P}_{\tau,h}V_{r,h+1}^k]}_{q_3}$$
$$+\underbrace{(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}[\mathbb{P}_{\tau,h}V_{r,h+1}^k - \mathbb{P}_{k,h}V_{r,h+1}^k]}_{q_4} + \underbrace{(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau[\mathbb{P}_{k,h}V_{r,h+1}^k - \mathbb{P}_{k,h}V_{k,r,h+1}^\pi]}_{q_5} \tag{4.162}$$

For the first term,

$$|\langle\phi(x,a), q_1\rangle| \leq \phi(x,a)^T(\Lambda_h^k)^{-1}\lambda w_{k,r,h}^\pi \leq ||w_{k,r,h}^\pi||||\phi(x,a)||_{(\Lambda_h^k)^{-1}} \tag{4.163}$$

For the second term we have

$$\begin{aligned}
&\phi(x,a)^T(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau[r_{\tau,h}(x_h^\tau,a_h^\tau) - r_{k,h}(x_h^\tau,a_h^\tau)] \\
&\leq \phi(x,a)^T(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau||\phi_h^\tau||||\theta_{\tau,r,h} - \theta_{k,r,h}|| \\
&\leq \phi(x,a)^T(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau||\phi_h^\tau||||\sum_{s=\tau}^{k-1}\theta_{s,r,h} - \theta_{s+1,r,h}|| \\
&\leq B_r^k\sqrt{dk}||\phi(x,a)||_{(\Lambda_h^k)^{-1}}
\end{aligned}$$

The last inequality follows from Lemma C.4 in [58]. Since

$$||\phi(x,a)||_{(\Lambda_h^k)^{-1}} \leq \sqrt{1/\lambda}$$

and $D \geq k$. We have

$$|\langle\phi(x,a), q_2\rangle| \leq B_r^{\mathcal{E}}\sqrt{dD} \tag{4.164}$$

Similarly, we can bound

$$\phi(x,a)^T(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau[\mathbb{P}_{\tau,h}V_{r,h+1}^k-\mathbb{P}_{k,h}V_{r,h+1}^k]\leq HB_p^k\sqrt{dk}||\phi(x,a)||_{(\Lambda_h^k)^{-1}} \tag{4.165}$$

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

$$|\langle\phi(x,a),q_3\rangle|\leq HB_p^{\mathcal{E}}\sqrt{dD} \tag{4.166}$$

From Lemma, the fourth term can be bounded as

$$|\langle\phi(x,a),q_4\rangle|\leq CdH\sqrt{\chi} \tag{4.167}$$

For the fifth term, note that

$$\begin{aligned}
\langle\phi(x,a),q_5\rangle &= \langle\phi(x,a),(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau[\mathbb{P}_h(V_{r,h+1}^k-V_{k,r,h+1}^\pi)(x_h^\tau,a_h^\tau)]\rangle\\
&= \langle\phi(x,a),(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau(\phi_h^\tau)^T\int(V_{r,h+1}^k-V_{k,r,h+1}^\pi)(x')d\mu_{k,h}(x')\rangle\\
&= \langle\phi(x,a),\int(V_{r,h+1}^k-V_{k,r,h+1}^\pi)(x')d\mu_{k,h}(x')\rangle\\
&- \langle\phi(x,a),\lambda(\Lambda_h^k)^{-1}\int(V_{r,h+1}^k-V_{r,h+1}^\pi)(x')d\mu_{k,h}(x')\rangle
\end{aligned} \tag{4.168}$$

The last term in (4.168) can be bounded as the following

$$|\langle\phi(x,a),\lambda(\Lambda_h^k)^{-1}\int(V_{r,h+1}^k-V_{k,r,h+1}^\pi)(x')d\mu_{k,h}(x')\rangle|\leq 2H\sqrt{d\lambda}\sqrt{\phi(x,a)^T(\Lambda_h^k)^{-1}\phi(x,a)} \tag{4.169}$$

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 (4.168) is equal to

$$\mathbb{P}_{k,h}(V^k_{r,h+1} - V^\pi_{r,h+1})(x,a) \tag{4.170}$$

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

$$\langle \phi(x,a), w^k_{j,h}\rangle - Q^\pi_{k,j,h} = \mathbb{P}_{k,h}(V^k_{j,h+1} - V^\pi_{k,j,h+1})(x,a) + \Delta^k_h + B^{\mathcal{E}}_r\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dW} \tag{4.171}$$

where $|\Delta^k_h| \le \beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)}$. □

Using Lemma IV.22, we also bound the difference between the combined Q-function (estimated) and the actual Q-function.

Lemma IV.23. *With probability* $1-2p$,

$$\begin{aligned} Q^\pi_{k,r,h} + Y_k Q^\pi_{k,g,h} \ge & Q^k_{r,h} + Y_k Q^k_{g,h} + \mathbb{P}_{k,h}(V^\pi_{k,r,h+1} + Y_k V^\pi_{k,g,h+1} - V^k_{r,h+1} - Y_k V^k_{g,h+1}) \\ & + B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD} \end{aligned} \tag{4.172}$$

Proof. From Lemma IV.22, we have

$$Q^\pi_{k,r,h} \le \langle \phi(x,a), w^k_{r,h}\rangle + \mathbb{P}_{k,h}(V^\pi_{k,r,h+1} - V^k_{r,h}) + \beta||\phi(x,a)||_{\Lambda^{-1}_{k,h}} + B^{\mathcal{E}}_r\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD} \tag{4.173}$$

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

$$Q^\pi_{k,r,h} \le \mathbb{P}_{k,h}(V^\pi_{k,r,h+1} - V^k_{r,h}) + Q^k_{r,h} + B^{\mathcal{E}}_r\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD} \tag{4.174}$$

Similarly,

$$Y_k Q^{\pi}_{k,g,h} \leq Y_k \mathbb{P}_{k,h}(V^{\pi}_{k,g,h+1} - V^k_{g,h}) + Y_k Q^k_{g,h} + Y_k B^{\mathcal{E}}_g \sqrt{dD} + Y_k H B^{\mathcal{E}}_p \sqrt{dD} \quad (4.175)$$

□

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 IV.24. *Then,* $\bar{V}^k_h(x) - V^k_h(x) \leq \dfrac{\log |\mathcal{A}|}{\alpha}$

where

Definition IV.25. $\bar{V}^k_h(\cdot) = \max_a [Q^k_{r,h}(\cdot, a) + Y_k Q^k_{g,h}(\cdot, a)]$.

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

Proof. Note that

$$V^k_h(x) = \sum_a \pi_{h,k}(a|x)[Q^k_{r,h}(x, a) + Y_k Q^k_{g,h}(x, a)] \quad (4.176)$$

where

$$\pi_{h,k}(a|x) = \frac{\exp(\alpha[Q^k_{r,h}(x, a) + Y_k Q^k_{g,h}(x, a)])}{\sum_a \exp(\alpha[Q^k_{r,h}(x, a) + Y_k Q^k_{g,h}(x, a)])} \quad (4.177)$$

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

Now, recall from Definition IV.25 that $\bar{V}^k_h(x) = [Q^k_{r,h}(x, a_x) + Y_k Q^k_{g,h}(x, a_x)]$. Then,

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

$$
\begin{aligned}
&\leq \left(\frac{\log(\sum_a \exp(\alpha(Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a))))}{\alpha} \right) \\
&- \sum_a \pi_{h,k}(a|x)[Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a)] \\
&\leq \frac{\log(|\mathcal{A}|)}{\alpha} \qquad (4.178)
\end{aligned}
$$

where the last inequality follows from Proposition 1 in [89]. □

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

Lemma IV.26. *With probability* $1 - 2p$,

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

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 IV.21 and from Lemma IV.22, we have for $j = r, g$,

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

Hence, for any (x,a),

$$
\begin{aligned}
Q^\pi_{j,H}(x,a) &\leq \min\{\langle \phi(x,a), w^k_{j,H}\rangle + \beta\sqrt{\phi(x,a)^T(\Lambda^k_H)^{-1}\phi(x,a)} + B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD}, H\} \\
&\leq Q^k_{j,H}(x,a) + B^{\mathcal{E}}_j\sqrt{dD} + HB^{\mathcal{E}}_p\sqrt{dD} \qquad (4.179)
\end{aligned}
$$

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

$$
\bar{V}^k_H(x) = \max_a [Q^k_{r,H}(x,a) + Y_k Q^k_{g,h}(x,a)]
$$

$$
\begin{aligned}
&\geq \sum_a \pi(a|x)[Q^{\pi}_{r,H}(x,a) + Y_k Q^{\pi}_{g,H}(x,a)] \\
&- (B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD}) \\
&\geq V^{\pi,Y_k}_H(x) - (B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + H(1+Y_k)B^{\mathcal{E}}_p\sqrt{dD}) \qquad (4.180)
\end{aligned}
$$

for any policy π. Thus, it also holds for π^*_k, the optimal policy. Hence, from Lemma IV.24, we have

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

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,

$$
\begin{aligned}
\mathbb{P}_{k,h}(V^{\pi,Y_k}_{h+1} - V^k_{h+1})(x,a) \leq &\frac{(H-h)\log(|\mathcal{A}|)}{\alpha} \\
&+ (H-h)(B^{\mathcal{E}}_r\sqrt{dW} + Y_k B^{\mathcal{E}}_g\sqrt{dW} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dW}) \qquad (4.181)
\end{aligned}
$$

From Lemma IV.22 we have for any (x,a)

$$
\begin{aligned}
&Q^{\pi}_{k,r,h}(x,a) + Y_k Q^{\pi}_{k,g,h}(x,a) \leq Q^k_{r,h}(x,a) + Y_k Q^k_{g,h}(x,a) + \frac{(H-h)\log(|\mathcal{A}|)}{\alpha} \\
&+ (H-h+1)(B^{\mathcal{E}}_r\sqrt{dD} + Y_k B^{\mathcal{E}}_g\sqrt{dD} + (1+Y_k)HB^{\mathcal{E}}_p\sqrt{dD}) \qquad (4.182)
\end{aligned}
$$

Hence,

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

Now, again from Lemma IV.24, we have $\bar{V}_h^k(x) - V_h^k(x) \leq \dfrac{\log(|\mathcal{A}|)}{\alpha}$. Thus,

$$V_h^{\pi,Y_k}(x) - V_h^k(x) \leq \frac{(H-h+1)\log(|\mathcal{A}|)}{\alpha} + (H-h+1)(B_r^{\mathcal{E}}\sqrt{dD} + Y_k B_g^{\mathcal{E}}\sqrt{dD} + (1+Y_k)HB_p^{\mathcal{E}}\sqrt{dD}) \tag{4.183}$$

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

$$\left(V_{r,h}^{\pi^*}(x) + Y_k V_{g,h}^{\pi^*}(x)\right) - \left(V_{r,h}^k(x) + Y_k V_{g,h}^k(x)\right) \leq \frac{(H-h+1)\log(|\mathcal{A}|)}{\alpha} + (H-h+1)(B_r^{\mathcal{E}}\sqrt{dD} + Y_k B_g^{\mathcal{E}}\sqrt{dD} + (1+Y_k)HB_p^{\mathcal{E}}\sqrt{dD}) \tag{4.184}$$

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

$$\begin{aligned} D_{j,h,1}^k &= \langle (Q_{j,h}^k(x_h^k, \cdot) - Q_{j,h}^{\pi_k}(x_h^k, \cdot)), \pi_{h,k}(\cdot|x_h^k)\rangle - (Q_{j,h}^k(x_h^k, a_h^k) - Q_{j,h}^{\pi_k}(x_h^k, a_h^k)) \\ D_{j,h,2}^k &= \mathbb{P}_{k,h}(V_{j,h+1}^k - V_{j,h+1}^{\pi_k})(x_h^k, a_h^k) - [V_{j,h+1}^k - V_{j,h+1}^{\pi_k}](x_{h+1}^k) \end{aligned} \tag{4.185}$$

Lemma IV.27. *On the event defined in E in Lemma IV.21, we have*

$$V_{j,1}^k(x_1) - V_{k,j,1}^{\pi_k} \leq \sum_{h=1}^H (D_{j,h,1}^k + D_{j,h,2}^k) + \sum_{h=1}^H 2\beta\sqrt{\phi(x_h^k, a_h^k)^T(\Lambda_h^k)^{-1}\phi(x_h^k, a_h^k)} + H(B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD}) \tag{4.186}$$

Proof. By Lemma IV.22, for any x, h, a, k

$$\begin{aligned} &\langle w_{j,h}^k(x,a), \phi(x,a)\rangle + \beta\sqrt{\phi(x,a)^T(\Lambda_h^k)^{-1}\phi(x,a)} - Q_{j,h}^{\pi_k} \\ &\leq \mathbb{P}_{k,h}(V_{j,h+1}^k - V_{k,j,h+1}^{\pi_k})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda_h^k)^{-1}\phi(x,a)} \end{aligned}$$

$$+ H(B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD})$$

Thus,

$$Q_{j,h}^k(x,a) - Q_{j,h}^{\pi_k}(x,a) \leq \mathbb{P}_{k,h}(V_{j,h+1}^k - V_{k,j,h+1}^{\pi_k})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda_h^k)^{-1}\phi(x,a)}$$
$$+ H(B_r^{\mathcal{E}}\sqrt{dD} + B_g^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD})$$
$$\mathbb{P}_{k,h}(V_{j,h+1}^k - V_{k,j,h+1}^{\pi_k})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda_h^k)^{-1}\phi(x,a)} +$$
$$B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD} - (Q_{j,h}^k(x,a) - Q_{k,j,h}^{\pi_k}(x,a)) \geq 0 \tag{4.187}$$

Since

$$V_{j,h}^k(x) = \sum_a \pi_{h,k}(a|x)Q_{j,h}^k(x,a)$$

and

$$V_{k,j,h}^{\pi_k}(x) = \sum_a \pi_{h,k}(a|x)Q_{k,j,h}^{\pi_k}(x,a)$$

where

$$\pi_{h,k}(a|\cdot) = \textsc{Soft-Max}_\alpha^a(Q_{r,h}^k + Y_k Q_{g,h}^k), \forall a$$

.

Thus, from (4.187),

$$V_{j,h}^k(x_h^k) - V_{k,j,h}^{\pi_k}(x_h^k) = \sum_a \pi_{h,k}(a|x_h^k)[Q_{j,h}^k(x_h^k,a) - Q_{k,j,h}^{\pi_k}(x_h^k,a)]$$
$$\leq \sum_a \pi_{h,k}(a|x_h^k)[Q_{j,h}^k(x_h^k,a) - Q_{k,j,h}^{\pi_k}(x_h^k,a)] + (B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD})$$
$$+ 2\beta\sqrt{\phi(x_h^k,a_h^k)^T(\Lambda_h^k)^{-1}\phi(x_h^k,a_h^k)} + \mathbb{P}_{k,h}(V_{j,h+1}^k - V_{j,h+1}^{\pi_k})(x_h^k,a_h^k)$$
$$- (Q_{j,h}^k(x_h^k,a_h^k) - Q_{k,j,h}^{\pi_k}(x_h^k,a_h^k)) \tag{4.188}$$

Thus, from (4.188), we have

$$V_{j,h}^k(x_h^k) - V_{j,h}^{\pi_k}(x_h^k) \leq D_{j,h,1}^k + D_{j,h,2}^k + [V_{j,h+1}^k - V_{j,h+1}^{\pi_k}](x_{h+1}^k) + 2\beta\sqrt{\phi(x_h^k, a_h^k)^T(\Lambda_h^k)^{-1}\phi(x_h^k, a_h^k)} + (B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD}) \tag{4.189}$$

Hence, by iterating recursively, we have

$$V_{j,1}^k(x_1) - V_{j,1}^{\pi_k} \leq \sum_{h=1}^{H}(D_{j,h,1}^k + D_{j,h,2}^k) + \sum_{h=1}^{H} 2\beta\sqrt{\phi(x_h^k, a_h^k)^T(\Lambda_h^k)^{-1}\phi(x_h^k, a_h^k)} + H(B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD}) \tag{4.190}$$

The result follows. □

Now, we are ready to prove Lemma IV.18.

Proof of Lemma IV.18

Proof. First, from Lemma IV.26,

$$\begin{aligned}&(V_{k,r,1}^{\pi_k^*}(x_1) + Y_k V_{k,g,1}^{\pi_k^*}(x_1)) - (V_{r,1}^k(x_1) + Y_k V_{g,1}^k(x_1)) \\ \leq &\frac{H\log(|\mathcal{A}|)}{\alpha} + H(B_r^{\mathcal{E}}\sqrt{dD} + Y_k B_g^{\mathcal{E}}\sqrt{dD} + (1+Y_k)HB_p^{\mathcal{E}}\sqrt{dD})\end{aligned} \tag{4.191}$$

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

$$\begin{aligned}&\sum_{k=1}^{D}(V_{k,r,1}^{\pi_k^*}(x_1) + Y_k V_{k,g,1}^{\pi_k^*}(x_1)) - (V_{r,1}^k(x_1) + Y_k V_{g,1}^k(x_1)) \leq \\ &\frac{HD\log(|\mathcal{A}|)}{\alpha} + H\sqrt{d}(B_r^{\mathcal{E}}D^{3/2} + 2H/\delta B_g^{\mathcal{E}}D^{3/2} + (1+2H/\delta)HB_p^{\mathcal{E}}D^{3/2})\end{aligned} \tag{4.192}$$

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

$$\sum_{\mathcal{E}=1}^{K/D}\sum_{k=1}^{D}(V_{k,r,1}^{\pi_k^*}(x_1)+Y_kV_{k,g,1}^{\pi_k^*}(x_1))-(V_{r,1}^{k}(x_1)+Y_kV_{g,1}^{k}(x_1))\leq\frac{HK\log(|\mathcal{A}|)}{\alpha}$$
$$+\sum_{\mathcal{E}=1}^{K/D}H\sqrt{d}(B_r^{\mathcal{E}}D^{3/2}+2H/\delta B_g^{\mathcal{E}}D^{3/2}+(1+2H/\delta)HB_p^{\mathcal{E}}D^{3/2})$$
$$\leq\frac{HK\log(|\mathcal{A}|)}{\alpha}+H^2(1+2H/\delta)\sqrt{d}BD^{3/2} \tag{4.193}$$

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 IV.27,

$$\sum_{k=1}^{D}(V_{j,1}^{k}(x_1)-V_{j,1}^{\pi_k}(x_1))$$
$$\leq\sum_{k=1}^{D}\sum_{h=1}^{H}(D_{j,h,1}^{k}+D_{j,h,2}^{k})+\sum_{k=1}^{D}\sum_{h=1}^{H}2\beta\sqrt{\phi(x_h^k,a_h^k)^T(\Lambda_h^k)^{-1}\phi(x_h^k,a_h^k)}$$
$$+\sum_{\mathcal{E}=1}^{K/D}\sum_{k=1}^{D}H(B_j^{\mathcal{E}}\sqrt{dD}+HB_p^{\mathcal{E}}\sqrt{dD}) \tag{4.194}$$

We, now, bound the individual terms of the right-hand side in (4.194). 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

$$t(k,h,m)=2(k-1)H+2(h-1)+m \tag{4.195}$$

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

Note from the definitions in (4.185) 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 $(k,h) \in [K] \times [H]$,

$$\mathbb{E}[D_{j,h,1}^k | \mathcal{F}_{h-1,2}^k] = 0, \quad \mathbb{E}[D_{j,h,2}^k | \mathcal{F}_{h,1}^k] = 0 \tag{4.196}$$

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

$$\begin{aligned} M_{j,h,m}^k &= \sum_{\tau=1}^{k-1} \sum_{i=1}^{H} (D_{j,i,1}^\tau + D_{j,i,2}^\tau) + \sum_{i=1}^{h-1} (D_{j,i,1}^k + D_{j,i,2}^k) + \sum_{l=1}^{m} D_{j,h,l}^k \\ &= \sum_{(\tau,i,l) \in [\mathcal{E}] \times [H] \times [2], t(\tau,i,l) \leq t(k,h,m)} D_{j,i,l}^\tau \end{aligned} \tag{4.197}$$

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 $\{\mathcal{F}_{h,m}^k\}_{(k,h,m) \in [D] \times [H] \times [2]}$, and particularly

$$\sum_{k=1}^{D} \sum_{h=1}^{H} (D_{j,h,1}^k + D_{j,h,2}^k) = M_{j,H,2}^D \tag{4.198}$$

Thus, $M_{j,H,2}^K$ is a Martingale difference satisfying $|M_{j,H,2}^D| \leq 4H$ since $|D_{j,h,1}^k|, |D_{j,h,2}^k| \leq 2H$ From the Azuma-Hoeffding inequality, we have

$$\Pr(M_{j,H,2}^D > s) \leq 2 \exp(-\frac{s^2}{16DH^2}) \tag{4.199}$$

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

$$\sum_k \sum_h M_{j,H,2}^D \leq \sqrt{16DH^2 \log(4/p)} \tag{4.200}$$

Now, we bound the second term of the right-hand side of (4.194). Note that the minimum eigenvalue of Λ_h^k is at least $\lambda = 1$ for all $(k,h) \in [D] \times [H]$. By Lemma C.5

in the appendix,

$$\sum_{k=1}^{K}(\phi_h^k)^T(\Lambda_h^k)^{-1}\phi_h^k \leq 2\log\left[\frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^1)}\right] \tag{4.201}$$

Moreover, note that $||\Lambda_h^{k+1}|| = ||\sum_{\tau=1}^{k}\phi_h^k(\phi_h^k)^T + \lambda\mathbb{I}|| \leq \lambda + k$, hence,

$$\sum_{k=1}^{D}(\phi_h^k)^T(\Lambda_h^k)^{-1}\phi_h^k \leq 2d\log\left[\frac{\lambda+k}{\lambda}\right] \leq 2d\iota \tag{4.202}$$

Now, by Cauchy-Schwartz inequality, we have

$$\begin{aligned}\sum_{k=1}^{D}\sum_{h=1}^{H}\sqrt{(\phi_h^k)^T(\Lambda_h^k)^{-1}\phi_h^k} &\leq \sum_{h=1}^{H}\sqrt{W}[\sum_{k=1}^{K}(\phi_h^k)^T(\Lambda_h^k)^{-1}\phi_h^k]^{1/2} \\ &\leq H\sqrt{2dD\iota}\end{aligned} \tag{4.203}$$

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

$$\mathcal{O}(\sqrt{H^4d^3D\iota^2}) \tag{4.204}$$

The third term of (4.194) is bounded by

$$\sum_{k=1}^{D}H(B_j^{\mathcal{E}}\sqrt{dD} + HB_p^{\mathcal{E}}\sqrt{dD}) = \sqrt{d}D^{3/2}H(B_j^{\mathcal{E}} + HB_p^{\mathcal{E}}) \tag{4.205}$$

Hence, summing (4.194) over the epochs we obtain

$$\sum_{\mathcal{E}=1}^{K/D}\sum_{k=1}^{D}(V_{j,1}^k(x_1) - V_{j,1}^{\pi_k}(x_1)) \leq \sum_{\mathcal{E}=1}^{K/D}\mathcal{O}(\sqrt{H^4d^3D\iota^2}) + \sum_{\mathcal{E}=1}^{K/D}\sqrt{d}D^{3/2}H(B_j^{\mathcal{E}} + HB_p^{\mathcal{E}}) \tag{4.206}$$

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

$$\sum_{\mathcal{E}=1}^{K/D} \sum_{k=1}^{D} (V_{j,1}^{k}(x_1) - V_{k,j,1}^{\pi_k}(x_1)) \leq \mathcal{O}(\sqrt{H^4 d^3 K^2 \iota^2 / D}) + \sqrt{d} D^{3/2} B H^2 \tag{4.207}$$

Thus,

$$\begin{aligned} &\sum_{k=1}^{K} (V_{r,1}^{k}(x_1) - V_{k,r,1}^{\pi_k}(x_1)) + Y(V_{g,1}^{k}(x_1) - V_{k,g,1}^{\pi_k}(x_1)) \\ \leq &(1+Y)(\mathcal{O}(\sqrt{H^4 d^3 K^2 \iota^2 / D}) + \sqrt{d} D^{3/2} B H^2) \end{aligned} \tag{4.208}$$

Hence, the result follows. □

4.7 Summary

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.

CHAPTER V

Conclusion and Future Work

Reinforcement Learning (RL) has gained significant attention due to its successes in several domains. However, applying RL to real-world applications raises concerns regarding safety, e.g., the consequences of actions in engineering systems can be catastrophic. Therefore, it is critical to strike a balance between reward maximization and safety in real-world applications. This dissertation investigated designing efficient model-free, simulator-free algorithms for different CMDP settings with provable safety guarantees. In what follows, we summarize the main contributions.

This dissertation starts with the episodic CMDP setting in Chapter II. We develop the first *model-free* RL algorithm for CMDPs with sublinear regret and *zero* constraint violation. The algorithm is named Triple-Q, and it has three key components: (i) a Q-function for the expected cumulative reward, denoted by $Q_h(x, a)$, (ii) a Q-function for the expected cumulative utility for the constraint, and (iii) a virtual-Queue, denoted by Z, which overestimates the cumulative constraint violation so far. Triple-Q uses UCB exploration when learning the Q-values to ensure an overestimation of the combined objective. Triple-Q is a two-time-scale algorithm where the virtual queue is updated at a slow time scale, and Triple-Q learns the pseudo-Q-value for fixed virtual queue length at a fast time scale within each frame. We prove Triple-Q achieves $\tilde{\mathcal{O}}\left(K^{\frac{4}{5}}\right)$ reward regret and guarantees *zero* constraint violation when the total number of

episodes K is large enough.

In Chapter III, we consider a more advanced and challenging setting, the infinite-horizon average-reward CMDPs. The agent-environment interaction never ends, or resets, and the goal is to achieve optimal long-term average reward under constraints, which appears to be much more challenging. We also proposed the first model-free RL algorithm for infinite-horizon average-reward CMDPs. The design of the algorithm is based on the primal-dual approach. By using the Lyapunov drift analysis, we proved that our algorithm achieves sublinear regret and zero constraint violation. Our regret bound scales as $\tilde{O}\left(K^{\frac{5}{6}}\right)$ and is suboptimal compared to model-based approaches. However, this is the first model-free and simulator-free algorithm with sub-linear regret and optimal constraint violation.

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. 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. In Chapter IV, 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. We develop different types 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, it 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

overestimate 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 CMDPs, 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. 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 aim to achieve zero constraint violation. The method eliminates the need to have a priori knowledge of the variation budget.

All of our algorithms are computationally efficient from an algorithmic perspective because they are model-free, which means that it is possible to apply our method to complex and challenging CMDPs in practice. The simulation results also demonstrate the good performance of our algorithm.

There are still many open problems in safe-RL, for example, how to satisfy stochastic and adversarial hard constraints under both model-based and model-free algorithms, how to leverage the benefits from offline RL to design a more efficient online RL algorithm for CMDPs, how to design efficient and provable algorithms for multi-agent CMDPs, and what are the fundamental sample complexity and regret bounds remain to be developed.

APPENDICES

APPENDIX A

Appendix for Chapter II

A.1 Notation Table for Chapter II

The notations used throughout this report are summarized in Table A.1.

A.2 Supporting Lemmas for Chapter II

In this section, we state several lemmas that are used in our analysis. The first lemma establishes some key properties of the learning rates used in Triple-Q. The proof closely follows the proof of Lemma 4.1 in [34].

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

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

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* $t = 0$.

Table A.1: Notation Table

Notation	Definition
K	The total number of episodes
S	The number of states
A	The number of actions
H	The length of each episode
$[H]$	Set $\{1, 2, \ldots, H\}$
$Q_{k,h}(x,a)$	The estimated reward Q-function at step h in episode k
$Q_h^\pi(x,a)$	The reward Q-function at step h in episode k under policy π
$V_{k,h}(x)$	The estimated reward value-function at step h in episode k
$V_h^\pi(x)$	The value-function at step h in episode k under policy π
$C_{k,h}(x,a)$	The estimated utility Q-function at step h in episode k
$C_h^\pi(x,a)$	The utility Q-function at step h in episode k under policy π
$W_{k,h}(x)$	The estimated utility value-function at step h in episode k
$W_h^\pi(x)$	The utility value-function at step h in episode k under policy π
$F_{k,h}(x,a)$	$F_{k,h}(x,a) = Q_{k,h}(x,a) + \frac{Z_k}{\eta} C_{k,h}(x,a)$
$U_{k,h}(x)$	$U_{k,h}(x) = V_{k,h}(x) + \frac{Z_k}{\eta} W_{k,h}(x)$
$r_h(x,a)$	The reward of (state, action) pair (x,a) at step h.
$g_h(x,a)$	The utility of (state, action) pair (x,a) at step h.
$N_{k,h}(x,a)$	The number of visits to (x,a) when at step h in episode k (not including k)
Z_k	The dual estimation (virtual queue) in episode k.
q_h^*	The optimal solution to the LP of the CMDP (2.24).
$q_h^{\epsilon,*}$	The optimal solution to the tightened LP (2.31).
δ	Slater's constant.
b_t	the UCB bonus for given t
$\mathbb{I}(\cdot)$	The indicator function

(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 [34].

(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

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

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

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

(e) According to its definition, we have

$$\begin{aligned}
\alpha_t^i = & \frac{\chi+1}{i+\chi} \cdot \left(\frac{i}{i+1+\chi} \frac{i+1}{i+2+\chi} \cdots \frac{t-1}{t+\chi} \right) \\
= & \frac{\chi+1}{t+\chi} \cdot \left(\frac{i}{i+\chi} \frac{i+1}{i+1+\chi} \cdots \frac{t-1}{t-1+\chi} \right) \leq \frac{\chi+1}{\chi+t}. \qquad \text{(A.3)}
\end{aligned}$$

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

The next lemma establishes upper bounds on $Q_{k,h}$ and $C_{k,h}$ under Triple-Q.

Lemma A.2. *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 $C_{k,h}(x, a)$:*

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

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 $Q_{0,H} = H \leq H\sqrt{\iota}$. Suppose $Q_{k',H}(x, a) \leq H\sqrt{\iota}$ for any $k' \leq k - 1$ and any (x, a). Then,

$$Q_{k,H}(x, a) = (1-\alpha_t)Q_{k_t,H}(x, a)+\alpha_t \left(r_H(x, a) + b_t\right) \leq \max\left\{H\sqrt{\iota}, 1 + \frac{H\sqrt{\iota}}{4}\right\} \leq H\sqrt{\iota},$$

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 $Q_{0,h} = H \leq H(H - h + 1)\sqrt{\iota}$. Now suppose the upper bound holds for $h + 1$, and also holds for $k' \leq k - 1$. Consider step h in episode k :

$$Q_{k,h}(x, a) =(1 - \alpha_t)Q_{k_t,h}(x, a) + \alpha_t \left(r_h(x, a) + V_{k_t,h+1}(x_{k_t,h+1}) + b_t\right),$$

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 $V_{k,h+1}(x) \leq \max_a Q_{k,h+1}(x, a) \leq H(H - h)\sqrt{\iota}$. Therefore, we obtain

$$Q_{k,h}(x, a) \leq \max\left\{H(H - h + 1)\sqrt{\iota}, 1 + H(H - h)\sqrt{\iota} + \frac{H\sqrt{\iota}}{4}\right\} \leq H(H - h + 1)\sqrt{\iota}.$$

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

Next, we present the following lemma from [34], which establishes a recursive relation between $Q_{k,h}$ and Q_h^π for any π. We include the proof so the report is self-contained.

Lemma A.3. *Consider any $(x,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, and any policy π. Let $t{=}N_{k,h}(x,a)$ be the number of visits to (x,a) when at step h in frame T before episode k, and $k_1, \ldots, k_t$ be the indices of the episodes in which these visits occurred. We have the following two equations:*

$$\begin{aligned}
&(Q_{k,h} - Q_h^\pi)(x,a) = \alpha_t^0 \left\{Q_{(T-1)K^\alpha+1,h} - Q_h^\pi\right\}(x,a) \\
&+ \sum_{i=1}^{t} \alpha_t^i \left(\left\{V_{k_i,h+1} - V_{h+1}^\pi\right\}(x_{k_i,h+1}) + \left\{\hat{\mathbb{P}}_h^{k_i} V_{h+1}^\pi - \mathbb{P}_h V_{h+1}^\pi\right\}(x,a) + b_i \right), \qquad \text{(A.4)} \\
&(C_{k,h} - C_h^\pi)(x,a) = \alpha_t^0 \left\{C_{(T-1)K^\alpha+1,h} - C_h^\pi\right\}(x,a) \\
&+ \sum_{i=1}^{t} \alpha_t^i \left(\left\{W_{k_i,h+1} - W_{h+1}^\pi\right\}(x_{k_i,h+1}) + \left\{\hat{\mathbb{P}}_h^{k_i} W_{h+1}^\pi - \mathbb{P}_h W_{h+1}^\pi\right\}(x,a) + b_i \right), \qquad \text{(A.5)}
\end{aligned}$$

where $\hat{\mathbb{P}}_h^k V_{h+1}(x,a) := V_{h+1}(x_{k,h+1})$ is the empirical counterpart of $\mathbb{P}_h V_{h+1}^\pi(x,a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V_{h+1}^\pi(x')$. This definition can also be applied to W_h^π as well.

Proof. We will prove (A.4). The proof for (A.5) is identical. Recall that under Triple-Q, $Q_{k+1,h}(x,a)$ is updated as follows:

$$\begin{aligned}
&Q_{k+1,h}(x,a) \\
&= \begin{cases} (1-\alpha_t) Q_{k,h}(x,a) + \alpha_t \left(r_h(x,a) + V_{k,h+1}(x_{h+1,k}) + b_t\right) & \text{if } (x,a) = (x_{k,h}, a_{k,h}) \\ Q_{k,h}(x,a) & \text{otherwise} \end{cases}.
\end{aligned}$$

From the update equation above, we have in episode k,

$$Q_{k,h}(x,a) = (1-\alpha_t)Q_{k_t,h}(x,a) + \alpha_t \left(r_h(x,a) + V_{k_t,h+1}(x_{k_t,h+1}) + b_t\right).$$

Repeatedly using the equation above, we obtain

$$\begin{aligned}
Q_{k,h}(x,a) =& (1-\alpha_t)(1-\alpha_{t-1})Q_{k_{t-1},h}(x,a) \\
&+(1-\alpha_t)\alpha_{t-1}\left(r_h(x,a) + V_{k_{t-1},h+1}(x_{k_{t-1},h+1}) + b_{t-1}\right) \\
&\quad + \alpha_t \left(r_h(x,a) + V_{k_t,h+1}(x_{k_t,h+1}) + b_t\right) \\
=& \cdots \\
=& \alpha_t^0 Q_{(T-1)K^\alpha+1,h}(x,a) + \sum_{i=1}^{t} \alpha_t^i \left(r_h(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i\right), \qquad \text{(A.6)}
\end{aligned}$$

where the last equality holds due to the definition of α_t^i in (A.1) and the fact that all $Q_{1,h}(x,a)$s are initialized to be H. Now applying the Bellman equation $Q_h^\pi(x,a) = \left\{r_h + \mathbb{P}_h V_{h+1}^\pi\right\}(x,a)$ and the fact that $\sum_{i=1}^t \alpha_t^i = 1$, we can further obtain

$$\begin{aligned}
Q_h^\pi(x,a) &= \alpha_t^0 Q_h^\pi(x,a) + (1-\alpha_t^0)Q_h^\pi(x,a) \\
&= \alpha_t^0 Q_h^\pi(x,a) + \sum_{i=1}^{t} \alpha_t^i \left(r(x,a) + \mathbb{P}_h V_{h+1}^\pi(x,a) + V_{h+1}^\pi(x_{k_i,h+1}) - V_{h+1}^\pi(x_{k_i,h+1})\right) \\
&= \alpha_t^0 Q_h^\pi(x,a) + \sum_{i=1}^{t} \alpha_t^i \left(r_h(x,a) + \mathbb{P}_h V_{h+1}^\pi(x,a) + V_{h+1}^\pi(x_{k_i,h+1}) - \hat{\mathbb{P}}_h^{k_i} V_{h+1}^\pi(x,a)\right) \\
&= \alpha_t^0 Q_h^\pi(x,a) + \sum_{i=1}^{t} \alpha_t^i \left(r_h(x,a) + V_{h+1}^\pi(x_{k_i,h+1}) + \left\{\mathbb{P}_h V_{h+1}^\pi - \hat{\mathbb{P}}_h^{k_i} V_{h+1}^\pi\right\}(x,a)\right). \qquad \text{(A.7)}
\end{aligned}$$

Then subtracting (A.7) from (A.6) yields

$$(Q_{k,h} - Q_h^\pi)(x,a) = \alpha_t^0 \left\{Q_{(T-1)K^\alpha+1,h} - Q_h^\pi\right\}(x,a)$$

$$+\sum_{i=1}^{t}\alpha_t^i\left(\left\{V_{k_i,h+1}-V_{h+1}^{\pi}\right\}(x_{k_i,h+1})+\left\{\hat{\mathbb{P}}_h^{k_i}V_{h+1}^{\pi}-\mathbb{P}_hV_{h+1}^{\pi}\right\}(x,a)+b_i\right).$$

□

Lemma A.4. *Consider any frame T. 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 $k_1,\ldots,k_t<k$ 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 $(x,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$*

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

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

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

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

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

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

$$|X(i)-X(i-1)|\leq\alpha_\tau^i\left|\left\{(\hat{\mathbb{P}}_h^{k_{n+1}}-\mathbb{P}_h)V_{h+1}^{\pi}\right\}(x,a)\right|\leq\alpha_\tau^iH$$

Then let $\sigma=\sqrt{8\log\left(\sqrt{2SAH}K\right)\sum_{i=1}^{\tau}(\alpha_\tau^iH)^2}$. By applying the Azuma-Hoeffding

inequality, we have with probability at least $1 - 2\exp\left(-\frac{\sigma^2}{2\sum_{i=1}^{\tau}(\alpha_\tau^i H)^2}\right) = 1 - \frac{1}{SAHK^4}$,

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

where the last inequality holds due to $\sum_{i=1}^{\tau}(\alpha_\tau^i)^2 \leq \frac{\chi+1}{\chi+\tau}$ from Lemma A.1.(e). 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 $1 - \frac{1}{K^3}$ the following inequality holds simultaneously for all $(x,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$,:

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

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

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

□

This lemma bound the conditional expected Lyapunov drift.

Lemma A.5. *Given $\delta \geq 2\epsilon$, under Triple-Q, the conditional expected drift is*

$$\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \leq -\frac{\delta}{2}Z_T + \frac{4H^2\iota}{K^2} + \eta\sqrt{H^2\iota} + H^4\iota + \epsilon^2 \tag{A.8}$$

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

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

From inequality (2.56), we have

$$
\begin{aligned}
&\mathbb{E}\left[L_{T+1}-L_T|Z_T=z\right]\\
\leq&\frac{1}{K^\alpha}\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha}\mathbb{E}\left[Z_T\left(\rho+\epsilon-C_{k,1}(x_{k,1},a_{k,1})\right)-\eta Q_{k,1}(x_{k,1},a_{k,1})\right.\\
&\left.+\eta Q_{k,1}(x_{k,1},a_{k,1})|Z_T=z\right]+H^4\iota+\epsilon^2\\
\leq_{(a)}&\frac{1}{K^\alpha}\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha}\mathbb{E}\left[Z_T\left(\rho+\epsilon-\sum_a\left\{C_{k,1}q_1^\pi\right\}(x_{k,1},a)\right)\right.\\
&\left.-\eta\sum_a\{Q_{k,1}q_1^\pi\}(x_{k,1},a)+\eta Q_{k,1}(x_{k,1},a_{k,1})|Z_T=z\right]\\
&+\epsilon^2+H^4\iota\\
\leq&\frac{1}{K^\alpha}\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha}\mathbb{E}\left[Z_T\left(\rho+\epsilon-\sum_a\left\{C_1^\pi q_1^\pi\right\}(x_{k,1},a)\right)\right.\\
&\left.-\eta\sum_a\{Q_{k,1}q_1^\pi\}(x_{k,1},a)+\eta Q_{k,1}(x_{k,1},a_{k,1})|Z_T=z\right]\\
&+\frac{1}{K^\alpha}\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha}\mathbb{E}\left[Z_T\sum_a\left\{C_1^\pi q_1^\pi\right\}(x_{k,1},a)-Z_T\sum_a\left\{C_{k,1}q_1^\pi\right\}(x_{k,1},a)|Z_T=z\right]\\
&+\frac{1}{K^\alpha}\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha}\mathbb{E}\left[\eta\sum_a\left\{Q_1^\pi q_1^\pi\right\}(x_{k,1},a)-\eta\sum_a\left\{Q_1^\pi q_1^\pi\right\}(x_{k,1},a)|Z_T=z\right]\\
&+H^4\iota+\epsilon^2\\
\leq_{(b)}&-\frac{\delta}{2}z+\frac{1}{K^\alpha}\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha}\mathbb{E}\left[\eta\sum_a\left\{(F_1^\pi-F_{k,1})q_1^\pi\right\}(x_{k,1},a)\right.\\
&\left.+\eta Q_{k,1}(x_{k,1},a_{k,1})|\,Z_T=z\right]+H^4\iota+\epsilon^2\\
\leq_{(c)}&-\frac{\delta}{2}z+\frac{4H^2\iota}{K^2}+\eta\sqrt{H^2\iota}+H^4\iota+\epsilon^2.
\end{aligned}
$$

Inequality (a) holds because of our algorithm. Inequality (b) holds because the fact that $\sum_a\left\{Q_1^\pi q_1^\pi\right\}(x_{k,1},a)$ is non-negative, and under Slater's condition, we can find

policy π such that

$$\epsilon+\rho-\mathbb{E}\left[\sum_{a} C_1^{\pi}(x_{k,1},a)q_1^{\pi}(x_{k,1},a)\right] = \rho+\epsilon-\mathbb{E}\left[\sum_{h,x,a} q_h^{\pi}(x,a)g_h(x,a)\right] \leq -\delta+\epsilon \leq -\frac{\delta}{2}.$$

Finally, inequality (c) is obtained due to the fact that $Q_{k,1}(x_{k,1},a_{k,1})$ is bounded by using Lemma A.2, and the fact that

$$\mathbb{E}\left[\sum_{k=(T-1)K^{\alpha}+1}^{TK^{\alpha}} \sum_{a} \left\{(F_1^{\pi}-F_{k,1})q_1^{\pi}\right\}(x_{k,1},a) \middle| Z_T = z\right]$$

can be bounded as (2.51) (note that the overestimation result and the concentration result in frame T hold regardless of the value of Z_T). $\square$

APPENDIX B

Appendix for Chapter III

B.1 Notation Table for Chapter III

We summarize notations used throughout this chapter in Table B.1.

Table B.1: Notation Table

Notation	Definition
K	The total number of episodes.
J_r^π	The reward rate under policy π.
J_r^π	The utility rate under policy π.
$V^\pi(s)$	The cumulative discounted reward under policy π and initial state s.
$Q^\pi(s)$	The cumulative discounted reward under policy π and initial state action pair (s,a).
$W^\pi(s)$	The cumulative discounted utility under policy π and initial state s.
$C^\pi(s)$	The cumulative discounted utility under policy π and initial state action pair (s,a).
$v^\pi(s)$	The relative reward value function for state s.
$w^\pi(s)$	The relative utility value function for state s.
$sp(v^\pi)$	Span of relative reward value function: $sp(v^\pi) = \max_s v^\pi(s) - \min_s v^\pi(s)$.
$sp(w^\pi)$	Span of relative utility value function: $sp(w^\pi) = \max_s w^\pi(s) - \min_s w^\pi(s)$.
κ	$\max\{sp(v^{\epsilon,*}), sp(w^{\epsilon,*}), 1\}$

Definition B.1 (Diameter)**.** The diameter of an MDP $\mathcal{M}$ is defined as:

$$D(\mathcal{M}) = \max_{s' \neq s} \min_{\pi} \mathbb{E}[\min\{t \geq 1 : S_t = s'\} | S_1 = s] - 1, \tag{B.1}$$

where the expectation is taken with respect to the Markov chain $(S_t)_{t=1}^{\infty}$ induced by the policy π and $\mathcal{M}$.

B.2 Supporting Lemmas for Chapter III

Lemma B.2. *(Azuma's inequality) Let $X_1, X_2, \ldots$ be a martingale difference with $|X_i| \leq c_i$ for all i. Then for any $0 < \delta < 1$,*

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i \geq \sqrt{2\bar{c}_N^2 ln\frac{1}{\delta}}\right) \leq \delta,$$

where $c_N^2 := \sum_{i=1}^{N} c_i^2$.

Lemma B.3. *For any $k = 1, \ldots, K^{\beta} - 1$ in frame T and state-action pair (s, a), the following holds:*

$$\begin{aligned} Q_{k+1}(s,a) - Q^{\pi}(s,a) =& \alpha_{\tau}^{0}(\hat{Q}_{(T-1)k^{\beta}+1}(s,a) - Q^{\pi}(s,a)) \\ &+\gamma \sum_{i=1}^{\tau} \alpha_{\tau}^{i} \left[\hat{V}_{k_i}(s_{k_i+1}) - V^{\pi}(s_{k_i+1})\right] \\ &+\gamma \sum_{i=1}^{\tau} \alpha_{\tau}^{i} \left[V^{\pi}(s_{k_i+1}) - \mathbb{E}_{s' \sim p(\cdot|s,a)} V^{\pi}(s')\right] + \sum_{i=1}^{\tau} \alpha_{\tau}^{i} b_i, \end{aligned} \tag{B.2}$$

where $\tau = n_{k+1}(s,a)$, is the total number of visits to (s,a) for the first k timesteps.

Proof. By recursively using the updating rule for $Q_{k+1}(s,a)$, we have

$$Q_{k+1}(s,a) = \hat{Q}_1(s,a)\alpha_t^0 + \sum_{i=1}^{\tau} \alpha_{\tau}^{i} \left[r(s,a) + \gamma \hat{V}_{k_i}(s_{k_i}+1)\right] + \sum_{i=1}^{\tau} \alpha_{\tau}^{i} b_i. \tag{B.3}$$

According to the Bellman equation $Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} V^{\pi}(s')$ and the fact $\sum_{i=1}^{\tau} \alpha_{\tau}^{i} = 1$, we have

$$Q^{\pi}(s,a) = \alpha_{\tau}^{0} Q^{\pi}(s,a) + \sum_{i=1}^{\tau} \alpha_{\tau}^{i} \left[r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} V^{\pi}(s')\right], \tag{B.4}$$

which finishes the proof. □

Lemma B.4. *Consider any frame T. Let $t{=}N_{k,h}(s,a)$ be the number of visits to (s,a) before timestep k in the current frame and let $k_1,\dots,k_t<k$ be the indices of these steps. Under any policy π, with probability at least $1-\frac{1}{K^3}$, the following inequalities hold simultaneously for all $(s,a,k)\in\mathcal{S}\times\mathcal{A}\times[K]$*

$$\left|\sum_{i=1}^{\tau}\alpha_\tau^i\left[V^{\epsilon,*}(s_{k_i+1})-\mathbb{E}_{s'\sim p(\cdot|s,a)}V^{\epsilon,*}(s')\right]\right|\leq\kappa\sqrt{\frac{(\chi+1)\iota}{\chi+\tau}},\tag{B.5}$$

$$\left|\sum_{i=1}^{\tau}\alpha_\tau^i\left[W^{\epsilon,*}(s_{k_i+1})-\mathbb{E}_{s'\sim p(\cdot|s,a)}W^{\epsilon,*}(s')\right]\right|\leq\kappa\sqrt{\frac{(\chi+1)\iota}{\chi+\tau}}.\tag{B.6}$$

Proof. Note that

$$\sum_{i=1}^{\tau}\alpha_\tau^i\left[V^{\epsilon,*}(s_{k_i+1})-\mathbb{E}_{s'\sim p(\cdot|s,a)}V^{\epsilon,*}(s')\right]$$

is a martingale, and each term in the summation belongs to $[-\alpha_\tau^i sp(V^{\epsilon,*}),\alpha_\tau^i sp(V^{\epsilon,*})]$ according to Lemma III.4.

Define $\sigma=\sqrt{8\log\left(\sqrt{2}K\right)\sum_{i=1}^{\tau}(\alpha_\tau^i sp(V^{\epsilon,*}))^2}$. By using Azuma's inequality (Lemma B.2), we obtain that the following inequality holds

$$\begin{aligned}&\left|\sum_{i=1}^{\tau}\alpha_\tau^i\left[V^{\epsilon,*}(s_{k_i+1})-\mathbb{E}_{s'\sim p(\cdot|s,a)}V^{\epsilon,*}(s')\right]\right|\leq\sigma\\ =&sp(V^{\epsilon,*})\sqrt{8\sum_{i=1}^{\tau}(\alpha_\tau^i)^2\log\sqrt{2}K}\leq\kappa\sqrt{\frac{(\chi+1)\iota}{\chi+\tau}}\end{aligned}\tag{B.7}$$

with probability at least

$$1-2\exp\left(-\frac{\sigma^2}{2\sum_{i=1}^{\tau}(\alpha_\tau^i sp(V^{\epsilon,*}))^2}\right)\geq 1-\frac{1}{K^3}.\tag{B.8}$$

□

Lemma B.5. *Given $\delta\geq 2\epsilon, H\geq\frac{6\kappa}{\delta}$, under our algorithm, the conditional expected*

drift of L is

$$\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \leq -\frac{\delta}{3} Z_T + \frac{3H}{K^2} + \eta + 2. \tag{B.9}$$

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

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

Then we have

$$\begin{aligned}
&\mathbb{E}\left[L_{K+1} - L_K | Z_T = z\right] \\
\leq &\frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\Big[Z_T \left(\rho + \epsilon - (1-\gamma)\hat{C}_k(s_k, a_k) \right) - \eta(1-\gamma)\hat{Q}_k(s_k, a_k) \\
&+ \eta(1-\gamma)\hat{Q}_k(s_k, a_k) \Big| Z_T = z \Big] + 2 \\
\leq_{(a)} &\frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\Bigg[Z_T \left(\rho + \epsilon - (1-\gamma) \sum_a \left\{ \hat{C}_k q^\pi \right\}(s_k, a) \right) \\
&- \eta(1-\gamma) \sum_a \{\hat{Q}_k q^\pi\}(s_k, a) \Bigg| Z_k = z \Bigg] \\
&+ E\left[\eta(1-\gamma)\hat{Q}_k(s_k, a_k) \Big| Z_T = z \right] + 2 \\
\leq &\frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\Bigg[Z_T \left(\rho + \epsilon - \sum_{s,a} \left\{ g q^\pi \right\}(s, a) \right) \\
&- \eta(1-\gamma) \sum_a \{\hat{Q}_k q^\pi\}(s_k, a) + \eta(1-\gamma)\hat{Q}_k(s_k, a_k) \Bigg| Z_T = z \Bigg] \\
&+ \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[Z_T \left(\sum_{s,a} \{g q^\pi\}(s, a) - \sum_a (1-\gamma) \left\{ C_k^\pi q^\pi \right\}(s_k, a) \right) \Bigg| Z_T = z \right] \\
&+ \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\Bigg[Z_T (1-\gamma) \sum_a \left\{ C_k^\pi q^\pi \right\}(s_k, a)
\end{aligned}$$

$$
\begin{aligned}
&-Z_T(1-\gamma)\sum_a\left\{\hat{C}_k q^\pi\right\}(s_k,a)\,|Z_T=z\Big] \\
&+\frac{1}{K^\beta}\sum_{k=(T-1)K^\beta+1}^{TK^\beta}\mathbb{E}\Big[\eta(1-\gamma)\sum_a\left\{Q_k^\pi q^\pi\right\}(s_k,a)- \\
&\eta(1-\gamma)\sum_a\left\{Q_k^\pi q^\pi\right\}(s_k,a)\Big|\,Z_T=z\Big]+2 \\
\leq_{(b)}&-\frac{\delta}{2}z+(1-\gamma)sp(v^\pi)+\frac{1}{K^\beta}\sum_{k=(T-1)K^\beta+1}^{TK^\beta}\mathbb{E}\Big[\eta(1-\gamma)\sum_a\left\{(F_k^\pi-\hat{F}_k)q_1^\pi\right\}(s_k,a) \\
&+\eta(1-\gamma)\hat{Q}_k(s_k,a_k)\Big|\,Z_T=z\Big]+2 \qquad \text{(B.10)}
\end{aligned}
$$

Inequality (a) holds because of our algorithm. Inequality (b) holds because of the fact that $\sum_a\left\{Q_t^\pi q^\pi\right\}(s_t,a)$ is non-negative, under Slater's condition, we can find policy π such that

$$
\epsilon+\rho-\mathbb{E}\left[\sum_{s,a}g(s_k,a)q^\pi(s_k,a)\right]\leq-\delta+\epsilon\leq-\frac{\delta}{2},
$$

and according to Lemma III.4

$$
\begin{aligned}
&\sum_{s,a}\{gq^\pi\}(s,a)-\sum_a(1-\gamma)\left\{C_k^\pi q^\pi\right\}(s_k,a) \\
=&J^\pi-(1-\gamma)V^\pi(s_k) \\
\leq&(1-\gamma)sp(v^\pi). \qquad \text{(B.11)}
\end{aligned}
$$

Note that when K is sufficiently large, $(1-\gamma)sp(v^{\epsilon,*})\leq\frac{\kappa}{H}\leq\frac{\delta}{6}$. By applying $\pi=\epsilon,*$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[L_{K+1}-L_K|Z_T=z\right] \\
\leq&-\frac{\delta}{2}z+\frac{\delta}{6}+\frac{1}{K^\beta}\sum_{k=(T-1)K^\beta+1}^{TK^\beta}\mathbb{E}\Big[\eta(1-\gamma)\sum_a\left\{(F_k^\pi-\hat{F}_k)q_1^\pi\right\}(s_k,a) \\
&+\ \eta(1-\gamma)\hat{Q}_k(s_k,a_k)\Big|\,Z_T=z\Big]+2
\end{aligned}
$$

$$\leq -\frac{\delta}{3}z + \frac{3H}{K^2} + \eta + 2, \tag{B.12}$$

where the last inequality holds due to (i) the overestimation established in Lemma III.5 and (ii) $\hat{Q}(\cdot,\cdot)$ is bounded by $\frac{1}{1-\gamma}$. We remark that the overestimation result and the concentration result in frame T hold regardless of the value of Z_T. $\square$

APPENDIX C

Appendix for Chapter IV

C.1 Notation Table for Chapter IV

The notations used throughout this chapter are summarized in Table C.1.

C.2 Supporting Lemmas for Chapter IV

Lemma C.1. *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 $C_{k,h}(x, a)$:*

$$0 \leq Q_{k,h}(x, a) \leq H^2(\sqrt{\iota} + 2\tilde{b}) \tag{C.1}$$

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

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 $Q_{0,H} = H \leq H(\sqrt{\iota}+2\tilde{b})$. Suppose $Q_{k',H}(x, a) \leq H(\sqrt{\iota} + 2\tilde{b})$ for any $k' \leq k - 1$ and any (x, a). Then,

$$Q_{k,H}(x, a) = (1 - \alpha_t)Q_{k_t,H}(x, a) + \alpha_t \left(r_{k,H}(x, a) + b_t + 2H\tilde{b}\right) \tag{C.3}$$

Table C.1: 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 (4.15) in episode k
$q^{\epsilon,*}_{k,h}$	optimal solution to the tightened LP (4.21) 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

$$\leq \max\left\{H\sqrt{\iota} + 2\tilde{b}H, 1 + \frac{H\sqrt{\iota}}{4} + 2H\tilde{b}\right\} \leq H\sqrt{\iota} + 2\tilde{b}H, \tag{C.4}$$

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 $Q_{0,h} = H \leq H(H-h+1)(\sqrt{\iota} + 2\tilde{b})$. Now suppose the upper bound holds for $h+1$,

and also holds for $k' \leq k-1$. 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 $V_{k,h+1}(x) \leq \max_a Q_{k,h+1}(x,a) \leq H(H-h)(\sqrt{\iota}+2\tilde{b})$. Therefore, we obtain

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

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

Lemma C.2. *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 $k_1, \ldots, k_t < k$ 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 $(x,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$,*

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

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

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

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

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

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

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

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 $(x,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$,:

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

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.

□

Lemma C.3. *Under Definition IV.6, for any fixed policy π, let $w_{k,j,h}^{\pi}$ be the corresponding weights such that $Q_{k,j,h}^{\pi} = \langle \phi(x, a), w_{k,j,h}^{\pi} \rangle$, for $j \in \{r, g\}$, then we have for all $h \in [H]$ and $k \in [K]$*

$$||w_{k,j,h}^{\pi}|| \leq 2H\sqrt{d} \tag{C.5}$$

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

$$\begin{aligned} Q_{k,j,h}^{\pi}(x, a) &= j_{k,h}(x, a) + \mathbb{P}_{k,h} V_{k,j,h}^{\pi}(x, a) \\ &= \langle \phi(x, a), \theta_{j,h} \rangle + \int_{\mathcal{S}} V_{k,j,h+1}^{\pi}(x') \langle \phi(x, a), d\mu_{k,h}(x') \rangle \\ &= \langle \phi(x, a), w_{k,j,h}^{\pi} \rangle \end{aligned} \tag{C.6}$$

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

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

□

Lemma C.4. *For any (k, h), the weight $w_{j,h}^{k}$ satisfies*

$$||w_{j,h}^{k}|| \leq 2H\sqrt{dk/\lambda} \tag{C.7}$$

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

$$|v^T w_{j,h}^k| = |v^T (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau (x_h^\tau, a_h^\tau)(j_h(x_h^\tau, a_h^\tau) + \sum_a \pi_{h+1,k}(a|x_{h+1}^\tau) Q_{j,h+1}^k(x_{h+1}^\tau, a))| \tag{C.8}$$

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

Note that $Q_{j,h+1}^k(x,a) \leq H$ for any (x,a). Hence, from (C.8) we have

$$\begin{aligned} |v^T w_{j,h}^k| &\leq \sum_{\tau=1}^{k-1} |v^T (\Lambda_h^k)^{-1} \phi_h^\tau|.2H \\ &\leq \sqrt{\sum_{\tau=1}^{k-1} v^T (\Lambda_k^h)^{-1} v} \sqrt{\sum_{\tau=1}^{k-1} \phi_h^\tau (\Lambda_h^k)^{-1} \phi_h^\tau}.2H \\ &\leq 2H||v|| \frac{\sqrt{dk}}{\sqrt{\lambda}} \end{aligned} \tag{C.9}$$

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

The following result is shown in [90] and in Lemma D.2 in [58].

Lemma C.5. *Let $\{\phi_t\}_{t\geq 0}$ be a sequence in $\mathbb{R}^d$ satisfying $\sup_{t\geq 0} ||\phi_t|| \leq 1$. For any $t \geq 0$, we define $\Lambda_t = \Lambda_0 + \sum_{j=0}^t \phi_j \phi_j^T$. Then if the smallest eigen value of Λ_0 be at least 1, we have*

$$\log\left[\frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^1)}\right] \leq \sum_{k=1}^K (\phi_h^k)^T (\Lambda_h^k)^{-1} \phi_h^k \leq 2\log\left[\frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^1)}\right] \tag{C.10}$$

We use the following result (Lemma J.10 in [74]).

Lemma C.6. *Let $\bar{C}^* \geq 2\max_k \mu^{k,*}$, then, if*

$$\sum_{k=1}^K (V_{k,r,1}^{\pi_k^*}(x_1) - V_{k,r,1}^{\pi_k}(x_1)) + 2\bar{C}^* \sum_{k=1}^K (b_k - V_{k,g,1}^{\pi_k}(x_1)) \leq \delta \tag{C.11}$$

, then

$$\sum_{k=1}^{K}(b_k - V_{k,g,1}^{\pi_k}(x_1)) \leq \frac{2\delta}{\bar{C}^*} \tag{C.12}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[2] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[3] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, *et al.*, "Learning dexterous in-hand manipulation," *The Int. Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[4] M. Ono, M. Pavone, Y. Kuwata, and J. Balaram, "Chance-constrained dynamic programming with application to risk-aware robotic space exploration," *Autonomous Robots*, vol. 39, no. 4, pp. 555–571, 2015.

[5] J. Garcia and F. Fernández, "Safe exploration of state and action spaces in reinforcement learning," *Journal of Artificial Intelligence Research*, vol. 45, pp. 515–564, 2012.

[6] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2737–2752, 2018.

[7] N. Abe, P. Melville, C. Pendus, C. K. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk, M. Domick, and T. Gardinier, "Optimizing debt collections using constrained reinforcement learning," in *Proc. Ann. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 75–84, ACM, 2010.

[8] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: a survey," *ACM Comput. Surv.*, vol. 55, nov 2021.

[9] P. Kolesar, "A markovian model for hospital admission scheduling," *Management Science*, vol. 16, no. 6, pp. B–384, 1970.

[10] K. Golabi, R. B. Kulkarni, and G. B. Way, "A statewide pavement management system," *Interfaces*, vol. 12, no. 6, pp. 5–21, 1982.

[11] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.

[12] C. Derman and A. F. Veinott Jr, "Constrained markov decision chains," *Management Science*, vol. 19, no. 4-part-1, pp. 389–390, 1972.

[13] L. Zheng and L. Ratliff, "Constrained upper confidence reinforcement learning," in *Learning for Dynamics and Control*, pp. 620–629, PMLR, 2020.

[14] R. Singh, A. Gupta, and N. B. Shroff, "Learning in markov decision processes under constraints," *arXiv preprint arXiv:2002.12435*, 2020.

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

[16] K. C. Kalagarla, R. Jain, and P. Nuzzo, "A sample-efficient algorithm for episodic finite-horizon MDP with constraints," in *AAAI Conf. Artificial Intelligence*, vol. 35, pp. 8030–8037, 2021.

[17] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained MDPs," *arXiv preprint arXiv:2003.02189*, 2020.

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

[19] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic, "Provably efficient safe exploration via primal-dual policy optimization," in *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, vol. 130, pp. 3304–3312, PMLR, 2021.

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

[21] V. S. Borkar, "An actor-critic algorithm for constrained markov decision processes," *Systems & control letters*, vol. 54, no. 3, pp. 207–213, 2005.

[22] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *arXiv preprint arXiv:1805.11074*, 2018.

[23] A. Stooke, J. Achiam, and P. Abbeel, "Responsive safety in reinforcement learning by pid lagrangian methods," in *Int. Conf. Machine Learning (ICML)*, pp. 9133–9143, PMLR, 2020.

[24] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, "Projection-based constrained policy optimization," in *Int. Conf. on Learning Representations (ICLR)*, 2020.

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

[26] T. Xu, Y. Liang, and G. Lan, "A primal approach to constrained policy optimization: Global optimality and finite-time analysis," *arXiv preprint arXiv:2011.05869*, 2020.

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

[28] H. Wei, X. Liu, and L. Ying, "A provably-efficient model-free algorithm for constrained markov decision processes," *arXiv preprint arXiv:2106.01577*, 2021.

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

[30] H. Wei, X. Liu, and L. Ying, "A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes," in *AAAI Conf. Artificial Intelligence*, Feb. 2022.

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

[32] M. Agarwal, Q. Bai, and V. Aggarwal, "Markov decision processes with long-term average constraints," *arXiv preprint arXiv:2106.06680*, 2021.

[33] H. Wei, A. Ghosh, N. Shroff, L. Ying, and X. Zhou, "Provably efficient model-free algorithms for non-stationary CMDPs," in *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pp. 6527–6570, PMLR, 2023.

[34] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is q-learning provably efficient?," in *Advances Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 4863–4873, 2018.

[35] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

[36] A. Zimin and G. Neu, "Online learning in episodic markovian decision processes by relative entropy policy search," *nips*, vol. 26, 2013.

[37] S. Mannor and J. N. Tsitsiklis, "On the empirical state-action frequencies in markov decision processes under general policies," *Mathematics of Operations Research*, vol. 30, no. 3, pp. 545–561, 2005.

[38] F. Richter, R. K. Orosco, and M. C. Yip, "Open-sourced reinforcement learning environments for surgical robotics," *arXiv preprint arXiv:1903.02090*, 2019.

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

[40] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[41] T. Xu, Y. Liang, and G. Lan, "Crpo: A new approach for safe reinforcement learning with convergence guarantee," in *Int. Conf. Machine Learning (ICML)* (M. Meila and T. Z. 0001, eds.), vol. 139, pp. 11480–11491, PMLR, 2021.

[42] M. G. Azar, R. Munos, and H. J. Kappen, "On the sample complexity of reinforcement learning with a generative model," in *Int. Conf. Machine Learning (ICML)*, p. 1707–1714, Omnipress, 2012.

[43] M. G. Azar, R. Munos, and H. J. Kappen, "Minimax pac bounds on the sample complexity of reinforcement learning with a generative model," *Mach. Learn.*, vol. 91, p. 325–349, June 2013.

[44] Y. Wang, K. Dong, X. Chen, and L. Wang, "Q-learning with UCB exploration is sample efficient for infinite-horizon MDP," in *Int. Conf. on Learning Representations (ICLR)*, 2020.

[45] C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain, "Model-free reinforcement learning in infinite-horizon average-reward markov decision processes," in *Int. Conf. Machine Learning (ICML)*, pp. 10170–10180, PMLR, 2020.

[46] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*, vol. 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[47] C. J. C. H. Watkins, *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, May 1989.

[48] A. Wachi, Y. Sui, Y. Yue, and M. Ono, "Safe exploration and optimization of constrained MDPs using Gaussian processes," in *AAAI Conf. Artificial Intelligence*, vol. 32, p. 6548–6555, 2018.

[49] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa, "Safe exploration in continuous action spaces," *arXiv preprint arXiv:1801.08757*, 2018.

[50] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," in *AAAI Conf. Artificial Intelligence*, vol. 33, pp. 3387–3395, 2019.

[51] D. Grbic and S. Risi, "Safer reinforcement learning through transferable instinct networks," *arXiv preprint arXiv:2107.06686*, 2021.

[52] P. Liu, D. Tateo, H. B. Ammar, and J. Peters, "Robot reinforcement learning on the constraint manifold," in *Conference on Robot Learning (CoRL)*, pp. 1357–1366, PMLR, 2021.

[53] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," in *Advances Neural Information Processing Systems (NeurIPS)*, vol. 32, Curran Associates, Inc., 2019.

[54] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Int. Conf. Machine Learning (ICML)*, vol. 70, pp. 22–31, JMLR, 2017.

[55] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[56] R. Srikant and L. Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*. Cambridge University Press, 2014.

[57] D. Vial, A. Parulekar, S. Shakkottai, and R. Srikant, "Regret bounds for stochastic shortest path problems with linear function approximation," *arXiv preprint arXiv:2105.01593*, 2021.

[58] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Conference on Learning Theory*, pp. 2137–2143, PMLR, 2020.

[59] X. Liu, B. Li, P. Shi, and L. Ying, "An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints," in *Advances Neural Information Processing Systems (NeurIPS)*, 2021.

[60] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A lyapunov-based approach to safe reinforcement learning," in *Advances Neural Information Processing Systems (NeurIPS)*, p. 8103–8112, Curran Associates Inc., 2018.

[61] Q. Yang, T. D. Simão, S. H. Tindemans, and M. T. Spaan, "Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning," in *AAAI Conf. Artificial Intelligence*, vol. 35, pp. 10639–10646, 2021.

[62] M. J. Neely, "Energy-aware wireless scheduling with near-optimal backlog and convergence time tradeoffs," *IEEE/ACM Transactions on Networking*, vol. 24, no. 4, pp. 2223–2236, 2016.

[63] S. Paternain, M. Calvo-Fullana, L. F. O. Chamon, and A. Ribeiro, "Safe policies for reinforcement learning via primal-dual methods," *IEEE Trans. Autom. Control*, pp. 1–1, 2022.

[64] K. Dong, Y. Wang, X. Chen, and L. Wang, "Q-learning with ucb exploration is sample efficient for infinite-horizon mdp," *arXiv preprint arXiv:1901.09311*, 2019.

[65] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[66] C.-Y. Wei, M. Jafarnia Jahromi, H. Luo, and R. Jain, "Learning infinite-horizon average-reward mdps with linear function approximation," in *Int. Conf. Artificial Intelligence and Statistics (AISTATS)* (A. Banerjee and K. Fukumizu, eds.), vol. 130, pp. 3007–3015, PMLR, 13–15 Apr 2021.

[67] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg, "Ai safety gridworlds," *arXiv preprint arXiv:1711.09883*, 2017.

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

[69] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *jmlr*, vol. 16, no. 1, pp. 1437–1480, 2015.

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

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

[72] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," *Artificial Intelligence in Medicine*, vol. 109, p. 101964, 2020.

[73] L. Chen, R. Jain, and H. Luo, "Learning infinite-horizon average-reward markov decision process with constraints," in *icml*, pp. 3246–3270, PMLR, 2022.

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

[75] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," *NeurIPS*, vol. 21, 2008.

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

[77] O. D. Domingues, P. Ménard, M. Pirotta, E. Kaufmann, and M. Valko, "A kernel-based approach to non-stationary reinforcement learning in metric spaces," in *aistats*, pp. 3538–3546, PMLR, 2021.

[78] Y. Fei, Z. Yang, Z. Wang, and Q. Xie, "Dynamic regret of policy optimization in non-stationary environments," *NeurIPS*, vol. 33, pp. 6743–6754, 2020.

[79] R. Ortner, P. Gajane, and P. Auer, "Variational regret bounds for reinforcement learning," in *Uncertainty in Artificial Intelligence*, pp. 81–90, PMLR, 2020.

[80] A. Touati and P. Vincent, "Efficient learning in non-stationary linear markov decision processes," *arXiv preprint arXiv:2010.12870*, 2020.

[81] C.-Y. Wei and H. Luo, "Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach," in *colt*, pp. 4300–4354, PMLR, 2021.

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

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

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

[85] A. Ghosh, X. Zhou, and N. Shroff, "Provably efficient model-free constrained rl with linear function approximation," in *NeurIPS*, 2022.

[86] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Hedging the drift: Learning to optimize under nonstationarity," *Management Science*, vol. 68, no. 3, pp. 1696–1713, 2022.

[87] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2003.

[88] D. Zhou, J. He, and Q. Gu, "Provably efficient reinforcement learning for discounted mdps with feature mapping," in *icml*, pp. 12793–12802, PMLR, 2021.

[89] L. Pan, Q. Cai, Q. Meng, W. Chen, and L. Huang, "Reinforcement learning with dynamic boltzmann softmax updates," in *ijcai*, IJCAI'20, 2021.

[90] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems 24*, 2011.