Triple-Q: A Model-Free Algorithm for Constrained Reinforcement Learning with Sublinear Regret and Zero Constraint Violation

Honghao Wei
University of Michigan, Ann Arbor

Xin Liu
ShanghaiTech University

Lei Ying
University of Michigan, Ann Arbor

Abstract

This paper presents the first *model-free*, *simulator-free* reinforcement learning algorithm for Constrained Markov Decision Processes (CMDPs) 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 for 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 Q-values 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[1], 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.

[1]**Notation:** $f(n) = \tilde{\mathcal{O}}(g(n))$ denotes $f(n) = \mathcal{O}(g(n)\log^k n)$ with $k > 0$. The same applies to $\tilde{\Omega}$. $\mathbb{R}^+$ denotes non-negative real numbers. $[H]$ denotes the set $\{1, 2, \cdots, H\}$.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 INTRODUCTION

Reinforcement learning (RL), with its success in gaming and robotics, 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 Ono et al. (2015); Garcia and Fernández (2012); Fisac et al. (2018), legal and business restrictions in financial engineering Abe et al. (2010), and resource and budget constraints in healthcare systems Yu et al. (2021). 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. We consider cumulative constraints in episodic CMDPs in this paper, which include budget constraints, energy constraints, or structural fatigue in flexible UAVs.

Earlier studies on CMDPs assume the model is known. A comprehensive study of these early results can be found in Altman (1999). 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 Singh et al. (2020); Brantley et al. (2020); Kalagarla et al. (2021); Efroni et al. (2020); Liu et al. (2021b); Bura et al. (2021), or the LP problem in the primal component of a primal-dual algorithm Qiu et al. (2020); Efroni et al. (2020); Liu et al. (2021b). 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 Ding et al. (2021). Ding et al. (2021) also proposes a model-based algorithm (Algorithm 3) 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 Sutton and Barto (2018). 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 Ding et al. (2020); Xu et al. (2021); Chen et al. (2021), all of which require simulators. In particular, the sample-based NPG-PD algorithm in Ding et al. (2020) requires a simulator which can simulate the MDP from any initial state x, and the algorithms in Xu et al. (2021); Chen et al. (2021) both require a simulator for policy evaluation. It has been argued in Azar et al. (2012, 2013); Jin et al. (2018) 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 tradeoff are critical and become one of the most important considerations of RL algorithm design.

1.1 Main Contributions

In this paper, 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:

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

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 Jin et al. (2018)). Different from the optimistic Q-learning for unconstrained MDPs (e.g. Jin et al. (2018); Wang et al. (2020); Wei et al. (2020)), 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 Rummery and Niranjan (1994) instead of Q-learning Watkins (1989), in other words, using the Q-functions of the action that is taken instead of using the max function.

We prove Triple-Q achieves $\tilde{\mathcal{O}}\left(\frac{1}{\delta} H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}\right)$ 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 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. Efroni et al. (2020)). Two concurrent papers Liu et al. (2021b); Bura et al. (2021) developed model-based approaches that achieve zero constraint violation assuming a strictly safe policy is known a-prior. It remains open 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 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 estimated model improves as the number of data samples increases, 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

Table 1: The Exploration-Exploitation Tradeoff in Episodic CMDPs.

	Algorithm	Regret	Constraint Violation
Model-based	OPDOP (Ding et al., 2021)	$\tilde{\mathcal{O}}(H^3\sqrt{S^2AK})$	$\tilde{\mathcal{O}}(H^3\sqrt{S^2AK})$
	OptDual-CMDP (Efroni et al., 2020)	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$
	OptPrimalDual-CMDP (Efroni et al., 2020)	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$
	CONRL (Brantley et al., 2020)	$\tilde{\mathcal{O}}(H^3\sqrt{S^3A^2K})$	$\tilde{\mathcal{O}}(H^3\sqrt{S^3A^2K})$
	OptPess-LP (Liu et al., 2021b)	$\tilde{\mathcal{O}}(H^3\sqrt{S^3AK})$	0
	OptPess–PrimalDual (Liu et al., 2021b)	$\tilde{\mathcal{O}}(H^3\sqrt{S^3AK})$	$\mathcal{O}(1)$
	OPSRL(Bura et al., 2021)	$\tilde{\mathcal{O}}(\sqrt{S^4H^7AK})$	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

more challenging than bounding regret of model-free algorithms for unconstrained MDPs like the optimistic Q-learning Jin et al. (2018); Wang et al. (2020); Wei et al. (2020).

As 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 paper, 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 provably guarantees, have been developed over the past last years (see e.g., Wachi et al. (2018); Dalal et al. (2018); Cheng et al. (2019); Grbic and Risi (2021); Liu et al. (2021a)). We will compare the performance of Triple-Q and Deep Triple-Q (Triple-Q with neural networks) with some of these algorithms in Section F, and demonstrate significant performance improvements (higher rewards, lower costs, and faster convergence) under Triple-Q.

2 PROBLEM FORMULATION

We consider an episodic CMDP, 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 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 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 Jin et al. (2018), 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 policy π, which is a collection of H functions $\{\pi_h:\mathcal{S}\to\mathcal{A}\}_{h=1}^H$, 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 :

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

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

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

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:

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

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

For simplicity, we adopt the following notation (some used in Jin et al. (2018); Ding et al. (2021)):

$$\begin{aligned}\mathbb{P}_hV_{h+1}^\pi(x,a)=&\mathbb{E}_{x'\sim\mathbb{P}_h(\cdot|x,a)}V_{h+1}^\pi(x'),\\ Q_h^\pi(x,\pi_h(x))=&\sum_a Q_h^\pi(x,a)\mathbb{P}(\pi_h(x)=a),\\ \mathbb{P}_hW_{h+1}^\pi(x,a)=&\mathbb{E}_{x'\sim\mathbb{P}_h(\cdot|x,a)}W_{h+1}^\pi(x'),\\ C_h^\pi(x,\pi_h(x))=&\sum_a C_h^\pi(x,a)\mathbb{P}(\pi_h(x)=a).\end{aligned}$$

From the definitions above, we have

$$\begin{aligned} V_h^\pi(x) =& Q_h^\pi(x, \pi_h(x)), \\ Q_h^\pi(x,a) =& r_h(x,a) + \mathbb{P}_h V_{h+1}^\pi(x,a), \\ W_h^\pi(x) =& C_h^\pi(x, \pi_h(x)), \\ C_h^\pi(x,a) =& g_h(x,a) + \mathbb{P}_h W_{h+1}^\pi(x,a). \end{aligned}$$

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:

$$\max_{\pi \in \Pi} \mathbb{E}\left[V_1^\pi(x_1)\right] \quad \text{subject to: } \mathbb{E}\left[W_1^\pi(x_1)\right] \geq \rho, \quad (1)$$

where we assume $\rho \in [0, H]$ to avoid triviality and the expectation is taken with respect to the initial distribution $x_1 \sim \mu_0$.

Remark 1. *The results in the paper 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*

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

Let π^* denote the optimal solution to the CMDP problem defined in (1). We evaluate our model-free RL algorithm using regret and constraint violation defined below:

$$\text{Regert}(K) = \mathbb{E}\left[\sum_{k=1}^K \left(V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1})\right)\right], \quad (2)$$

$$\text{Violation}(K) = \mathbb{E}\left[\sum_{k=1}^K \left(\rho - W_1^{\pi_k}(x_{k,1})\right)\right], \quad (3)$$

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 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 paper, 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 Ding et al. (2021, 2020); Efroni et al. (2020); Paternain et al. (2019). 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 1). This is a noticeable difference from some of works in CMDPs in which the agent needs to know the value of this constant (e.g. Ding et al. (2021)) or alternatively a feasible policy (e.g. Achiam et al. (2017)) .

3 TRIPLE-Q

In this section, we introduce Triple-Q for CMDPs. 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. Ding et al. (2021, 2020); Qiu et al. (2020); Efroni et al. (2020); Liu et al. (2021b)), 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 (1) from a given initial state x_1 :

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

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:

$$\lambda \leftarrow \left(\lambda + \rho - \mathbb{E}\left[W_1^{\pi_\lambda^*}(x_1)\right]\right)^+. \quad (4)$$

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; and 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. Neely (2010); Srikant and Ying (2014)). 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 Jin et al. (2018) do 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 Jin et al. (2018). 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).

We now formally introduce Triple-Q. A detailed description is presented in Algorithm 1. The algorithm only needs to know the values of H, A, S and 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.

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 1. *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:*

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

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

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

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

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 be depends on the slackness δ.

4 PROOF OUTLINE

Due to the page limit, we will only present an outline of the proof and the key intuitions in this section. The complete proof can be found in the supplementary material. **Notation:** In the proof, we explicitly include the episode index in our notation. In particular, $x_{k,h}$ and $a_{k,h}$ are the state and the action taken at step h of episode k; $Q_{k,h}$, $C_{k,h}$, Z_k, and $\bar{C}_k$ are the reward Q-function, the utility Q-function, the virtual-Queue, and the value of $\bar{C}$ *at the beginning* of episode k; $N_{k,h}$, $V_{k,h}$ and $W_{k,h}$ are the visit count, reward value-function, and utility value-function *after* they are updated at step h of episode k (i.e. after line 10 of Triple-Q). We also use shorthand notation $\{f-g\}(x) = f(x)-g(x)$, when $f(\cdot)$ and $g(\cdot)$ take the same argument value. Similarly $\{(f-g)q\}(x) = (f(x)-g(x))q(x)$. In this shorthand notation, we put functions inside { }, and the common argument(s) outside. A summary of notations used throughout this paper can be found in Table 2 in the supplementary material.

4.1 Regret

To bound the regret, we consider an offline optimization problem as our regret baseline Altman (1999); Puterman (1994):

$$\max_{q_h} \sum_{h,x,a} q_h(x,a) r_h(x,a) \tag{5}$$

$$\text{s.t.:} \sum_{h,x,a} q_h(x,a) g_h(x,a) \geq \rho \tag{6}$$

$$\sum_a q_h(x,a) = \sum_{x',a'} \mathbb{P}_{h-1}(x|x',a') q_{h-1}(x',a'), \tag{7}$$

$$\sum_a q_1(x,a) = \mu_0(x), q_h(x,a) \geq 0 \tag{8}$$

$$\sum_{x,a} q_h(x,a) = 1, \forall h \in [H], \forall x \in \mathcal{S}. \tag{9}$$

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 $\frac{q_h(x,a)}{\sum_{a'} q_h(x,a')}$ is the probability of taking action a in state x at step h, which defines the policy. We can see that (6) is the utility constraint, (7) represents the global-balance equation for the MDP and the initial state is sampled from μ_0, (8)-(9) indicate the normalization condition so that q_h is a valid probability distribution.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:

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

$$\text{s.t.:} \sum_{h,x,a} q_h(x,a) g_h(x,a) \geq \rho + \epsilon, \text{ and } (7)-(9), \tag{10}$$

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 (1) or the solution to the corresponding LP (5) and superscript $^{\epsilon,*}$ to denote the optimal value/policy related to the ϵ-tightened version of CMDP (defined in (10)).

Following the definition of the regret in (2), we have

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

$$= \mathbb{E}\left[\sum_{k=1}^{K} \left(\sum_a \{Q_1^* q_1^*\}(x_{k,1}, a)\right) - Q_1^{\pi_k}(x_{k,1}, a_{k,1})\right].$$

Now by adding and subtracting the corresponding terms, we obtain

$$\text{Regret}(K) =$$

$$\mathbb{E}\left[\sum_{k=1}^{K} \left(\sum_a \{Q_1^* q_1^* - Q_1^{\epsilon,*} q_1^{\epsilon,*}\}(x_{k,1}, a)\right)\right] + \tag{11}$$

$$\mathbb{E}\left[\sum_{k=1}^{K} \left(\sum_a \{Q_1^{\epsilon,*} q_1^{\epsilon,*}\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1})\right)\right] + \tag{12}$$

$$\mathbb{E}\left[\sum_{k=1}^{K} \{Q_{k,1} - Q_1^{\pi_k}\}(x_{k,1}, a_{k,1})\right]. \tag{13}$$

Next, we establish the regret bound by analyzing the three terms above. We first present a brief outline. The complete proof is presented in the supplementary material.

- **Step 1:** First, by comparing the LP associated with the original CMDP (5) and the tightened LP (10), Lemma 1 will show $\mathbb{E}\left[\sum_a \left\{Q_1^* q_1^* - Q_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right] \le \frac{H\epsilon}{\delta}$, which implies that under our choices of ϵ, δ, and ι,

$$(11) \le \frac{KH\epsilon}{\delta} = \tilde{\mathcal{O}}\left(\frac{1}{\delta} H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}\right).$$

- **Step 2:** Note that $Q_{k,h}$ is an estimate of $Q_h^{\pi_k}$, and the estimation error (13) is controlled by the learning rates and the UCB bonuses. In Lemma 2, we will show that the cumulative estimation error over one frame(K^α consecutive episodes) is upper bounded by

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

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

$$\begin{aligned}(13) \le& H^2 SAK^{1-\alpha} + \frac{H^3\sqrt{\iota}K}{\chi} + \sqrt{H^4 SA\iota K^{2-\alpha}(\chi+1)}\\ =& \tilde{\mathcal{O}}\left(H^3 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}\right).\end{aligned}$$

The proof of Lemma 2 is based on a recursive formula that relates the estimation error at step h to the estimation error at step $h+1$.

- **Step 3:** Bounding (12) is the most challenging part of the proof. For unconstrained MDPs, the optimistic Q-learning in Jin et al. (2018) 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

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^K \left(\sum_a \left\{Q_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1})\right)\right]\\ &= \mathbb{E}\left[\sum_k \sum_a \left(\left\{Q_1^{\epsilon,*} q_1^{\epsilon,*} + \frac{Z_k}{\eta} C_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right.\right.\\ &\left.\left. - \left\{Q_{k,1} q_1^{\epsilon,*} + \frac{Z_k}{\eta} C_{k,1} q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right)\right] \qquad (14)\\ &+ \mathbb{E}\left[\sum_k \left(\sum_a \left\{Q_{k,1} q_1^{\epsilon,*}\right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1})\right)\right] \qquad (15)\\ &+ \mathbb{E}\left[\sum_k \frac{Z_k}{\eta} \sum_a \left\{\left(C_{k,1} - C_1^{\epsilon,*}\right) q_1^{\epsilon,*}\right\}(x_{k,1}, a)\right]. \qquad (16)\end{aligned}$$

We can see (14) is the difference of two pseudo-Q-functions. Using a two-dimensional induction on step and episode, we will prove in Lemma 3 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. $(14) \le 0$) for all (x,a,h,k) simultaneously with a high probability. To guarantee this overestimation, Triple-Q resets the reward and cost Q-functions to H at the beginning of each frame. Finally, to bound (15)+(16), 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 4 that the Lyapunov-drift satisfies

$$\begin{aligned}\mathbb{E}[L_{T+1} - L_T] \le& \text{ a negative drift}\\ &+ H^4\iota + \epsilon^2 - \frac{\eta}{K^\alpha} \sum_{k=TK^\alpha+1}^{(T+1)K^\alpha} \Phi_k, \qquad (17)\end{aligned}$$

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 $(15)+(16) = \sum_k \Phi_k$. Inequality (17) 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 (17). From (17), we observe that

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

So we can bound (15)+(16) by applying the telescoping sum over the $K^{1-\alpha}$ frames

$$\begin{aligned}(15)+(16) =& \sum_k \Phi_k \le \frac{K^\alpha \mathbb{E}\left[L_1 - L_{K^{1-\alpha}+1}\right]}{\eta}\\ &+ \frac{K(H^4\iota + \epsilon^2)}{\eta} \le \frac{K(H^4\iota + \epsilon^2)}{\eta},\end{aligned}$$

where the last inequality holds because $L_1 = 0$ and $L_T \ge 0$ for all T. Combining the bounds on (14) and (15)+(16), we conclude that under our choices of ι, ϵ and η,

$$(12) = \tilde{\mathcal{O}}(H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}).$$

Combining the results in the three steps above, we obtain the regret bound in Theorem 1,

$$\text{Regret}(K) \le \frac{KH\epsilon}{\delta} + H^2 SAK^{1-\alpha} + \frac{H^3\sqrt{\iota}K}{\chi}$$

$$+\sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)}+\frac{K\left(H^4\iota+\epsilon^2\right)}{\eta}+\frac{4H^4\iota}{\eta K}.$$

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

$$\begin{aligned}\text{Regret}(K) \leq & \frac{13}{\delta}H^4\sqrt{SA\iota^3}K^{0.8}+\frac{4H^4\iota}{K^{1.2}}\\ =&\tilde{\mathcal{O}}\left(\frac{1}{\delta}H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{0.8}\right).\end{aligned} \tag{18}$$

4.2 Constraint Violation

We present a brief outline of the proof. The details are in the supplementary material. 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

$$\begin{aligned}Z_{T+1} =& \left(Z_T+\rho+\epsilon-\frac{\bar{C}_T}{K^\alpha}\right)^+\\ \geq& Z_T+\rho+\epsilon-\frac{\bar{C}_T}{K^\alpha},\end{aligned} \tag{19}$$

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, taking expectation on both sides, and using the fact $Z_1 = 0$, we obtain the following upper bound on the constraint violation:

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^K\rho-C_1^{\pi_k}(x_{k,1},a_{k,1})\right]\leq& -K\epsilon+K^\alpha\mathbb{E}\left[Z_{K^{1-\alpha}+1}\right]\\ &+\mathbb{E}\left[\sum_{k=1}^K\left\{C_{k,1}-C_1^{\pi_k}\right\}(x_{k,1},a_{k,1})\right],\end{aligned} \tag{20}$$

Lemma 2 in the supplementary material will establish an upper bound on the estimation error of $C_{k,1}$:

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^K\left\{C_{k,1}-C_1^{\pi_k}\right\}(x_{k,1},a_{k,1})\right]\leq& H^2SAK^{1-\alpha}\\ &+\frac{H^3\sqrt{\iota}K}{\chi}+\sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)}.\end{aligned} \tag{21}$$

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 $1 \leq T \leq K^{1-\alpha}+1$

$$\begin{aligned}\mathbb{E}[Z_T]\leq& \frac{54H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right)\\ &+\frac{16H^2\iota}{K^2\delta}+\frac{4\eta\sqrt{H^2\iota}}{\delta}.\end{aligned} \tag{22}$$

Under our choices of ϵ, α, χ, η and ι, it can be easily verified that $K\epsilon$ dominates the upper bounds in (21) and (22), which leads to the conclusion that the constraint violation becomes zero when K is sufficiently large in Theorem 1.

Substituting the results from (21) and (22) into (20), 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 if further we have $K \geq e^{\frac{1}{\delta}}$, $\text{Violation}(K) = 0$. This shows zero constraint violation on expectation. We can further prove that under the same assumptions, $\sum_{k=1}^K \rho - W_1^{\pi_k}(x_{k,1}) \leq 0$ with a high probability. The details can be found at the end of Section B in the supplementary material.

4.3 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 Jin et al. (2018); Wang et al. (2020); Wei et al. (2020); Vial et al. (2021); Jin et al. (2020) is that our proof relies heavily on the Lyapunov-drift analysis of virtual-Queue Z. The drift analysis on 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 Liu et al. (2021c). Some of the proofs were inspired by Liu et al. (2021c). Compared with bandit problems, CMDPs, however, is a much more challenging problem due to its sequential nature.

5 CONVERGENCE and ϵ-OPTIMAL POLICY

The Triple-Q algorithm is an online learning algorithm and is not a stationary policy. In theory, we can obtain a near-optimal, stationary policy following the idea proposed in Jin et al. (2018). Assume the agent stores all the policies used during learning. Note that each policy is defined by the two Q tables and the value of the virtual queue. At the end of learning horizon K, the agent constructs a stochastic policy $\bar{\pi}$ such that at the beginning of each episode, the agent uniformly and randomly selects a policy from the K policies, i.e. $\bar{\pi} = \pi_k$ with probability $\frac{1}{K}$. We note that given any initial state x,

$$\frac{1}{K}\sum_{k=1}^{K} V_1^{\pi_k}(x) = V_1^{\bar{\pi}}(x), \tag{23}$$

$$\frac{1}{K}\sum_{k=1}^{K} W_1^{\pi_k}(x) = W_1^{\bar{\pi}}(x). \tag{24}$$

Therefore, under policy $\bar{\pi}$, we have

$$\begin{aligned}
&\mathbb{E}\left[V_1^*(x_{k,1}) - V_1^{\bar{\pi}}(x_{k,1})\right]\\
=&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(V_1^*(x_{k,1}) - V_1^{\bar{\pi}}(x_{k,1})\right)\right]\\
=&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1})\right)\right]\\
=&\tilde{O}\left(\frac{H^4\sqrt{SA}}{\delta K^{0.2}}\right),
\end{aligned} \tag{25}$$

and

$$\begin{aligned}
\mathbb{E}\left[\rho - W_1^{\bar{\pi}}(x_{k,1})\right] &= \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(\rho - W_1^{\bar{\pi}}(x_{k,1})\right)\right]\\
&= \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(\rho - W_1^{\pi_k}(x_{k,1})\right)\right] \le 0.
\end{aligned} \tag{26}$$

Therefore, given any ϵ, $\bar{\pi}$ is an ϵ-optimal policy when K is sufficiently large.

While $\bar{\pi}$ is a near-optimal policy, in practice, it may not be possible to store all policies during learning due to memory constraint, so extracting the stationary policy based on output randomization is not ideal. A heuristic approach to obtain a near optimal, stationary policy is to fix the two Q functions (reward and utility) after learning horizon K and continue to adapt the virtual queue with frame size $\sqrt{K}$. The virtual queue needs to be continuously updated because an optimal policy for a CMDP is stochastic in general and a greedy policy based on a fixed virtual queue will not be optimal. By adapting the virtual queue with fixed Q functions, Triple-Q outputs a stochastic policy. When the virtual queue reaches its steady state, we obtain a stationary policy. The resulted policy has great performance in our experiment (see Section F in the supplementary material).

6 EVALUATION

We evaluated Triple-Q using a grid-world environment (Chow et al., 2018). We further implemented Triple-Q with neural network approximations for environments with continuous state and action spaces, called Deep Triple-Q. We compared the performance of Deep Triple-Q with existing safe RL algorithms, including WCSAC in the Dynamic Gym benchmark (Yang et al., 2021), CBF in the Pendulum environment Cheng et al. (2019), and DDPG+Safety Layter in the Ball-1d environment Dalal et al. (2018). The details can be found in section F in the supplementary material. The code for the simulations is available found at: `https://github.com/honghaow/Triple-Q`.

7 CONCLUSIONS

This paper 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 paper, 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:

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

Acknowledgements: The authors sincerely thank Arnob Ghosh, Prof. Ness Shroff, and Prof. Xingyu Zhou for their great comments. The work of Honghao Wei and Lei Ying is supported in part by NSF under grants 2002608, 2001687, and 2112471.

References

Abe, N., Melville, P., Pendus, C., Reddy, C. K., Jensen, D. L., Thomas, V. P., Bennett, J. J., Anderson, G. F., Cooley, B. R., Kowalczyk, M., Domick, M., and Gardinier, T. (2010). Optimizing debt collections using constrained reinforcement learning. In *Proc. Ann. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pages 75–84. ACM.

Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *Int. Conf. Machine Learning (ICML)*, volume 70, pages 22–31. JMLR.

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

Azar, M. G., Munos, R., and Kappen, H. J. (2012). On the sample complexity of reinforcement learning with a generative model. In *Int. Conf. Machine Learning (ICML)*, page 1707–1714. Omnipress.

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.*, 91(3):325–349.

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

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

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

Cheng, R., Orosz, G., Murray, R. M., and Burdick, J. W. (2019). End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *AAAI Conf. Artificial Intelligence*, volume 33, pages 3387–3395.

Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. (2018). A lyapunov-based approach to safe reinforcement learning. In *Advances Neural Information Processing Systems (NeurIPS)*, page 8103–8112. Curran Associates Inc.

Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., and Tassa, Y. (2018). Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. (2021). Provably efficient safe exploration via primal-dual policy optimization. In *Proc. of The 24th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 3304–3312. PMLR.

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

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

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

Garcia, J. and Fernández, F. (2012). Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 45:515–564.

Grbic, D. and Risi, S. (2021). Safer reinforcement learning through transferable instinct networks. *arXiv preprint arXiv:2107.06686*.

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

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

Kalagarla, K. C., Jain, R., and Nuzzo, P. (2021). A sample-efficient algorithm for episodic finite-horizon MDP with constraints. In *AAAI Conf. Artificial Intelligence*, volume 35, pages 8030–8037.

Liu, P., Tateo, D., Ammar, H. B., and Peters, J. (2021a). Robot reinforcement learning on the constraint manifold. In *Conference on Robot Learning (CoRL)*, pages 1357–1366. PMLR.

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

Liu, X., Li, B., Shi, P., and Ying, L. (2021c). An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 34.

Neely, M. J. (2010). *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan and Claypool Publishers.

Neely, M. J. (2016). Energy-aware wireless scheduling with near-optimal backlog and convergence time tradeoffs. *IEEE/ACM Transactions on Networking*, 24(4):2223–2236.

Ono, M., Pavone, M., Kuwata, Y., and Balaram, J. (2015). Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4):555–571.

Paternain, S., Chamon, L., Calvo-Fullana, M., and Ribeiro, A. (2019). Constrained reinforcement learning has zero duality gap. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, Inc.

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

Rummery, G. A. and Niranjan, M. (1994). *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK.

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

Srikant, R. and Ying, L. (2014). *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*. Cambridge University Press.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Vial, D., Parulekar, A., Shakkottai, S., and Srikant, R. (2021). Regret bounds for stochastic shortest path problems with linear function approximation. *CoRR*, abs/2105.01593.

Wachi, A., Sui, Y., Yue, Y., and Ono, M. (2018). Safe exploration and optimization of constrained MDPs using Gaussian processes. In *AAAI Conf. Artificial Intelligence*, volume 32, page 6548–6555.

Wang, Y., Dong, K., Chen, X., and Wang, L. (2020). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *Int. Conf. on Learning Representations (ICLR)*.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College, King's College, Cambridge United Kingdom.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *Int. Conf. Machine Learning (ICML)*, pages 10170–10180. PMLR.

Xu, T., Liang, Y., and Lan, G. (2021). Crpo: A new approach for safe reinforcement learning with convergence guarantee. In Meila, M. and 0001, T. Z., editors, *Int. Conf. Machine Learning (ICML)*, volume 139, pages 11480–11491. PMLR.

Yang, Q., Simão, T. D., Tindemans, S. H., and Spaan, M. T. (2021). Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. In *AAAI Conf. Artificial Intelligence*, volume 35, pages 10639–10646.

Yu, C., Liu, J., Nemati, S., and Yin, G. (2021). Reinforcement learning in healthcare: a survey. *ACM Comput. Surv.*, 55(1).

Supplementary Material

Part 1: Proof of Theorem 1

In this part of the supplementary material, we present the complete proof of the main theorem. For the convenience of the reader, we restate Theorem 1 and the proof outlines.

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

$$\begin{aligned}\text{Regret}(K) &\leq \frac{13}{\delta}H^4\sqrt{SA\iota^3}K^{0.8} + \frac{4H^4\iota}{K^{1.2}}\\ \text{Violation}(K) &\leq \frac{54H^4\iota K^{0.6}}{\delta}\log\frac{16H^2\sqrt{\iota}}{\delta} + \frac{4\sqrt{H^2\iota}}{\delta}K^{0.8} - 5\sqrt{SAH^6\iota^3}K^{0.8}.\end{aligned}$$

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

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

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

A REGRET

To bound the regret, we consider the following offline optimization problem as our regret baseline Altman (1999); Puterman (1994):

$$\max_{q_h}\sum_{h,x,a} q_h(x,a)r_h(x,a) \tag{27}$$

$$\text{s.t.:}\sum_{h,x,a} q_h(x,a)g_h(x,a) \geq \rho \tag{28}$$

$$\sum_a q_h(x,a) = \sum_{x',a'}\mathbb{P}_{h-1}(x|x',a')q_{h-1}(x',a') \tag{29}$$

$$\sum_{x,a} q_h(x,a) = 1, \forall h \in [H] \tag{30}$$

$$\sum_a q_1(x,a) = \mu_0(x) \tag{31}$$

$$q_h(x,a) \geq 0, \forall x \in \mathcal{S}, \forall a \in \mathcal{A}, \forall h \in [H]. \tag{32}$$

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

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

is the probability of taking action a in state x at step h, which defines the policy. We can see that (28) is the utility constraint, (29) is the global-balance equation for the MDP, (30) is the normalization condition so that q_h is a valid probability distribution, and (31) 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:

$$\max_{q_h}\sum_{h,x,a} q_h(x,a)r_h(x,a) \tag{33}$$

$$\text{s.t.:} \sum_{h,x,a} q_h(x,a) g_h(x,a) \geq \rho + \epsilon$$
$$(29) - (32),$$

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 (1) or the solution to the corresponding LP (27) and superscript $^{\epsilon,*}$ to denote the optimal value/policy related to the ϵ-tightened version of CMDP (defined in (33)).

Following the definition of the regret in, we have

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

Now by adding and subtracting the corresponding terms, we obtain

$$\begin{aligned}
&\text{Regret}(K)\\
=&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q_1^* q_1^* - Q_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1},a)\right)\right] + && (34)\\
&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a \left\{Q_1^{\epsilon,*} q_1^{\epsilon,*}\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right] + && (35)\\
&\mathbb{E}\left[\sum_{k=1}^{K}\left\{Q_{k,1} - Q_1^{\pi_k}\right\}(x_{k,1},a_{k,1})\right]. && (36)
\end{aligned}$$

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

A.1 Outline of the Regret Analysis

- **Step 1:** First, by comparing the LP associated with the original CMDP (27) and the tightened LP (33), Lemma 1 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,
$$(34) \leq \frac{KH\epsilon}{\delta} = \tilde{\mathcal{O}}\left(\frac{1}{\delta} H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}\right).$$

- **Step 2:** Note that $Q_{k,h}$ is an estimate of $Q_h^{\pi_k}$, and the estimation error (36) is controlled by the learning rates and the UCB bonuses. In Lemma 2, 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
$$(36) \leq H^2 SAK^{1-\alpha} + \frac{H^3\sqrt{\iota K}}{\chi} + \sqrt{H^4 SA\iota K^{2-\alpha}(\chi+1)} = \tilde{\mathcal{O}}\left(H^3 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{\frac{4}{5}}\right).$$
The proof of Lemma 2 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 Jin et al. (2018), but with different learning rates and UCB bonuses.

- **Step 3:** Bounding (35) is the most challenging part of the proof. For unconstrained MDPs, the optimistic Q-learning in Jin et al. (2018) 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

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\left(\sum_a\left\{Q_1^{\epsilon,*}q_1^{\epsilon,*}\right\}(x_{k,1},a)-Q_{k,1}(x_{k,1},a_{k,1})\right)\right]\\
=&\mathbb{E}\left[\sum_k\sum_a\left(\left\{Q_1^{\epsilon,*}q_1^{\epsilon,*}+\frac{Z_k}{\eta}C_1^{\epsilon,*}q_1^{\epsilon,*}\right\}(x_{k,1},a)-\left\{Q_{k,1}q_1^{\epsilon,*}+\frac{Z_k}{\eta}C_{k,1}q_1^{\epsilon,*}\right\}(x_{k,1},a)\right)\right] \qquad (37)\\
&+\mathbb{E}\left[\sum_k\left(\sum_a\left\{Q_{k,1}q_1^{\epsilon,*}\right\}(x_{k,1},a)-Q_{k,1}(x_{k,1},a_{k,1})\right)\right]+\mathbb{E}\left[\sum_k\frac{Z_k}{\eta}\sum_a\left\{\left(C_{k,1}-C_1^{\epsilon,*}\right)q_1^{\epsilon,*}\right\}(x_{k,1},a)\right]. \qquad (38)
\end{aligned}$$

We can see (37) is the difference of two pseudo-Q-functions. Using a two-dimensional induction on step and episode, we will prove in Lemma 3 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. (37) ≤ 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 (38), 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 4 that the Lyapunov-drift satisfies

$$\begin{aligned}
&\mathbb{E}[L_{T+1}-L_T]\\
\leq&\text{a negative drift}+H^4\iota+\epsilon^2-\frac{\eta}{K^\alpha}\sum_{k=TK^\alpha+1}^{(T+1)K^\alpha}\Phi_k, \qquad (39)
\end{aligned}$$

where

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

and we note that (38) $=\sum_k\Phi_k$. Inequality (39) 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 (39). From (39), we observe that

$$\mathbb{E}[L_{T+1}-L_T]\leq H^4\iota+\epsilon^2-\frac{\eta}{K^\alpha}\sum_{k=TK^\alpha+1}^{(T+1)K^\alpha}\Phi_k. \qquad (40)$$

So we can bound (38) by applying the telescoping sum over the $K^{1-\alpha}$ frames on the inequality above:

$$(38)=\sum_k\Phi_k\leq\frac{K^\alpha\mathbb{E}\left[L_1-L_{K^{1-\alpha}+1}\right]}{\eta}+\frac{K(H^4\iota+\epsilon^2)}{\eta}\leq\frac{K(H^4\iota+\epsilon^2)}{\eta},$$

where the last inequality holds because $L_1=0$ and $L_T\geq 0$ for all T. Combining the bounds on (37) and (38), we conclude that under our choices of ι, ϵ and η,

$$(35)=\tilde{\mathcal{O}}(H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{\frac{4}{5}}).$$

Combining the results in the three steps above, we obtain the regret bound in Theorem 1.

A.2 Detailed Proof

We next present the 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 1. *Given $\epsilon \le \delta$, we have*

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

□

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

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

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

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

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

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

Also we have $q_h^{\xi_2}(x,a) \ge 0$ 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 (33). Then given $\{q_h^{\epsilon,*}(x,a)\}_{h=1}^H$ is the optimal solution to the ϵ-tightened optimization problem, we have

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

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

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

□

The next lemma bounds the difference between the estimated Q-functions and actual Q-functions in a frame. The bound on (36) is an immediate result of this lemma.

Lemma 2. *Under Triple-Q, we have for any $T \in [K^{1-\alpha}]$,*

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

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 update 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

$$\begin{aligned} Q_{k,h}(x,a) =& (1-\alpha_t)(1-\alpha_{t-1})Q_{k_{t-1},h}(x,a) + (1-\alpha_t)\alpha_{t-1}\left(r_h(x,a) + V_{k_{t-1},h+1}(x_{k_{t-1},h+1}) + b_{t-1}\right) \\ &+ \alpha_t\left(r_h(x,a) + V_{k_t,h+1}(x_{k_t,h+1}) + b_t\right) \\ =& \cdots \\ =& \alpha_t^0 Q_{(T-1)K^\alpha+1,h}(x,a) + \sum_{i=1}^t \alpha_t^i\left(r_h(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i\right) \qquad (41) \\ \leq& \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i\left(r_h(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i\right), \qquad (42) \end{aligned}$$

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

$$\sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} Q_{k,h}(x,a) \leq \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \alpha_t^0 H + \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \sum_{i=1}^{N_{k,h}(x,a)} \alpha_{N_{k,h}}^i\left(r_h(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i\right). \qquad (43)$$

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:

$$\begin{aligned} N_{k,h} =& N_{k,h}(x_{k,h}, a_{k,h}) \\ k_i^{(k,h)} =& k_i(x_{k,h}, a_{k,h}), \end{aligned}$$

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

$$\sum_k \sum_{i=1}^{N_{k,h}} \alpha_{N_{k,h}}^i V_{k_i^{(k,h)},h+1}\left(x_{k_i^{(k,h)},h+1}\right) \leq \sum_k V_{k,h+1}(x_{k,h+1}) \sum_{t=N_{k,h}}^\infty \alpha_t^{N_{k,h}} \leq \left(1 + \frac{1}{\chi}\right)\sum_k V_{k,h+1}(x_{k,h+1}), \qquad (44)$$

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 7-(d). By substituting (44) into (43) and noting that $\sum_{i=1}^{N_{k,h}(x,a)} \alpha^i_{N_{k,h}} = 1$ according to Lemma Lemma 7-(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 8, and (iii) we know that

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

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

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

Taking the expectation on both sides yields

$$\mathbb{E}\left[\sum_k Q_{k,h}(x_{k,h}, a_{k,h}) - \sum_k Q_h^{\pi_k}(x_{k,h}, a_{k,h})\right]$$

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

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

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

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 (35), and further expand it as follows:

$$\begin{aligned}
&(35)\\
=&\mathbb{E}\left[\sum_{k=1}^K \left(\sum_a \left\{Q^{\epsilon,*}_1 q^{\epsilon,*}_1\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right]\\
=&\mathbb{E}\left[\sum_k\sum_a \left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right)q^{\epsilon,*}_1\right\}(x_{k,1},a)\right] &(45)\\
&+\mathbb{E}\left[\sum_k \left(\sum_a \left\{Q_{k,1}q^{\epsilon,*}_1\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right] + \mathbb{E}\left[\sum_k \frac{Z_k}{\eta}\sum_a \left\{\left(C_{k,1} - C^{\epsilon,*}_1\right)q^{\epsilon,*}_1\right\}(x_{k,1},a)\right], &(46)
\end{aligned}$$

where

$$F_{k,h}(x,a) = Q_{k,h}(x,a) + \frac{Z_k}{\eta}C_{k,h}(x,a)$$
$$F^{\epsilon,*}_h(x,a) = Q^{\epsilon,*}_h(x,a) + \frac{Z_k}{\eta}C^{\epsilon,*}_h(x,a).$$

We first show (45) 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 $F^{\epsilon,*}_h(x,a)$ for all k, h and (x,a) with a high probability.

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

$$\left\{F_{k,h} - F^\pi_h\right\}(x,a) \geq 0, \tag{47}$$

which further implies that

$$\mathbb{E}\left[\sum_{k=1}^K\sum_a \left\{\left(F^{\epsilon,*}_{k,1} - F_{k,1}\right)q^{\epsilon,*}_1\right\}(x_{k,1},a)\right] \leq \frac{4H^4\iota}{\eta K}. \tag{48}$$

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 9 in the appendix, we have

$$\begin{aligned}
&\{F_{k,h} - F_h^\pi\}(x,a) \\
=&\alpha_t^0 \left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) \\
&+ \sum_{i=1}^t \alpha_t^i \left(\left\{U_{k_i,h+1} - U_{h+1}^\pi\right\}(x_{k_i,h+1}) + \{(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) U_{h+1}^\pi\}(x,a) + \left(1 + \frac{Z}{\eta}\right) b_i \right) \\
\geq_{(a)} &\alpha_t^0 \left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) + \sum_{i=1}^t \alpha_t^i \left\{U_{k_i,h+1} - U_{h+1}^\pi\right\}(x_{k_i,h+1}) \\
=_{(b)} &\alpha_t^0 \left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) + \sum_{i=1}^t \alpha_t^i \left(\max_a F_{k_i,h+1}(x_{k_i,h+1},a) - F_{h+1}^\pi(x_{k_i,h+1}, \pi(x_{k_i,h+1})) \right) \\
\geq &\alpha_t^0 \left\{F_{(T-1)K^\alpha+1,h} - F_h^\pi\right\}(x,a) + \sum_{i=1}^t \alpha_t^i \left\{F_{k_i,h+1} - F_{h+1}^\pi\right\}(x_{k_i,h+1}, \pi(x_{k_i,h+1})),
\end{aligned} \tag{49}$$

where inequality (a) holds because of the concentration result in Lemma 10 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 7-(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)

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

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 (49) becomes

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

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

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

holds for any k, h, and (x,a), which completes the proof of (47).

Let $\mathcal{E}$ denote the event that (47) holds for all k, h and (x,a). Then based on Lemma 8, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^K \sum_a \left\{\left(F_{k,1}^{\epsilon,*} - F_{k,1}\right) q_1^{\epsilon,*}\right\}(x_{k,1},a)\right] \\
=&\mathbb{E}\left[\sum_{k=1}^K \sum_a \left\{\left(F_{k,1}^{\epsilon,*} - F_{k,1}\right) q_1^{\epsilon,*}\right\}(x_{k,1},a) \middle| \mathcal{E}\right] \Pr(\mathcal{E}) + \mathbb{E}\left[\sum_{k=1}^K \sum_a \left\{\left(F_{k,1}^{\epsilon,*} - F_{k,1}\right) q_1^{\epsilon,*}\right\}(x_{k,1},a) \middle| \mathcal{E}^c\right] \Pr(\mathcal{E}^c) \\
\leq &2K \left(1 + \frac{K^{1-\alpha} H^2 \sqrt{\iota}}{\eta}\right) H^2 \sqrt{\iota} \frac{1}{K^3} \leq \frac{4H^4 \iota}{\eta K}.
\end{aligned} \tag{52}$$

□

Next we bound (46) 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

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

Therefore, under Triple-Q, we have

$$Z_{T+1} = \left(Z_T + \rho + \epsilon - \frac{\bar{C}_T}{K^\alpha} \right)^+ \tag{53}$$

Define the Lyapunov function to be

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

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

Lemma 4. *Assume $\epsilon \le \delta$. The expected Lyapunov drift satisfies*

$$\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\
\le &\frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(-\eta \mathbb{E}\left[\sum_a \left\{ Q_{k,1} q_1^{\epsilon,*} \right\}(x_{k,1}, a) - Q_{k,1}(x_{k,1}, a_{k,1}) \middle| Z_T = z \right] \right. \\
&\left. + z \mathbb{E}\left[\sum_a \left\{ \left(C_1^{\epsilon,*} - C_{k,1} \right) q_1^{\epsilon,*} \right\}(x_{k,1}, a) \middle| Z_T = z \right] \right) + H^4 \iota + \epsilon^2.
\end{aligned} \tag{54}$$

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

$$\begin{aligned}
L_{T+1} - L_T \le &Z_T \left(\rho + \epsilon - \frac{\bar{C}_T}{K^\alpha} \right) + \frac{\left(\frac{\bar{C}_T}{K^\alpha} + \epsilon - \rho \right)^2}{2} \\
\le &Z_T \left(\rho + \epsilon - \frac{\bar{C}_T}{K^\alpha} \right) + H^4 \iota + \epsilon^2 \\
\le &\frac{Z_T}{K^\alpha} \sum_{k=TK^\alpha+1}^{(T+1)K^\alpha} \left(\rho + \epsilon - C_{k,1}(x_{k,1}, a_{k,1}) \right) + H^4 \iota + \epsilon^2
\end{aligned}$$

where the first inequality is a result of the upper bound on $|C_{k,1}(x_{k,1}, a_{k,1})|$ in Lemma 8.

Let $\{q_h^\epsilon\}_{h=1}^H$ be a feasible solution to the tightened LP (33). Then the expected Lyapunov drift conditioned on $Z_T = z$ is

$$\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\
\le &\frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \left(\mathbb{E}\left[z\left(\rho + \epsilon - C_{k,1}(x_{k,1}, a_{k,1}) \right) - \eta Q_{k,1}(x_{k,1}, a_{k,1}) \middle| Z_T = z \right] + \eta \mathbb{E}\left[Q_{k,1}(x_{k,1}, a_{k,1}) \middle| Z_T = z \right] \right) \\
&+ H^4 \iota + \epsilon^2.
\end{aligned} \tag{55}$$

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

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

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

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 (33), 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

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^K \left(\sum_a \left\{Q_{k,1} q_1^{\epsilon,*}\right\}(x_{k,1},a) - Q_{k,1}(x_{k,1},a_{k,1})\right)\right] + \mathbb{E}\left[\sum_{k=1}^K \frac{Z_k}{\eta} \sum_a \left\{\left(C_{k,1} - C_1^{\epsilon,*}\right) q_1^{\epsilon,*}\right\}(x_{k,1},a)\right] \\
\leq& \frac{K^\alpha \mathbb{E}\left[L_1 - L_{K^{1-\alpha}+1}\right]}{\eta} + \frac{K\left(H^4\iota + \epsilon^2\right)}{\eta} \leq \frac{K\left(H^4\iota + \epsilon^2\right)}{\eta},
\end{aligned} \tag{56}
$$

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

Now combining Lemma 3 and inequality (56), we conclude that

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

Further combining inequality above with Lemma 1 and Lemma 2,

$$
\text{Regret}(K) \leq \frac{KH\epsilon}{\delta} + H^2 SAK^{1-\alpha} + \frac{H^3\sqrt{\iota}K}{\chi} + \sqrt{H^4 SA\iota K^{2-\alpha}(\chi+1)} + \frac{K\left(H^4\iota + \epsilon^2\right)}{\eta} + \frac{4H^4\iota}{\eta K}. \tag{57}
$$

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

$$
\text{Regret}(K) \leq \frac{13}{\delta} H^4 \sqrt{SA\iota^3} K^{0.8} + \frac{4H^4\iota}{K^{1.2}} = \tilde{\mathcal{O}}\left(\frac{1}{\delta} H^4 S^{\frac{1}{2}} A^{\frac{1}{2}} K^{0.8}\right). \tag{58}
$$

B CONSTRAINT VIOLATION

B.1 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

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

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

$$\mathbb{E}\left[\sum_{k=1}^{K}\rho-C_1^{\pi_k}(x_{k,1},a_{k,1})\right] \le -K\epsilon+K^\alpha\mathbb{E}\left[Z_{K^{1-\alpha}+1}\right]+\mathbb{E}\left[\sum_{k=1}^{K}\{C_{k,1}-C_1^{\pi_k}\}(x_{k,1},a_{k,1})\right], \tag{59}$$

where we used the fact $Z_1=0$.

In Lemma 2, we already established an upper bound on the estimation error of $C_{k,1}$:

$$\mathbb{E}\left[\sum_{k=1}^{K}\{C_{k,1}-C_1^{\pi_k}\}(x_{k,1},a_{k,1})\right] \le H^2SAK^{1-\alpha}+\frac{H^3\sqrt{\iota}K}{\chi}+\sqrt{H^4SA\iota K^{2-\alpha}(\chi+1)}. \tag{60}$$

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 $1\le T\le K^{1-\alpha}+1$

$$\mathbb{E}[Z_T] \le \frac{54H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right)+\frac{16H^2\iota}{K^2\delta}+\frac{4\eta\sqrt{H^2\iota}}{\delta}. \tag{61}$$

Under our choices of ϵ, α, χ, η and ι, it can be easily verified that $K\epsilon$ dominates the upper bounds in (60) and (61), which leads to the conclusion that the constraint violation because zero when K is sufficiently large in Theorem 1.

B.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 5. *Assuming $\epsilon\le\frac{\delta}{2}$, we have for any $1\le T\le K^{1-\alpha}$*

$$\mathbb{E}[Z_T] \le \frac{54H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right)+\frac{16H^2\iota}{K^2\delta}+\frac{4\eta\sqrt{H^2\iota}}{\delta}. \tag{62}$$

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

Lemma 6. *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 $0<\gamma\le v$, 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 $\mathbb{E}[\Delta_t|S_t=s]\le-\gamma$ when $L_t\ge\theta_t$.

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

Then we have

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

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

Proof of Lemma 5. We apply Lemma 6 to a new Lyapunov function:

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

To verify condition (1) in Lemma 6, 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

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

where inequality (a) is obtained according to Lemma 11; and the last inequality holds given $z \geq \theta_T$.

To verify condition (2) in Lemma 6, 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 6, we obtain

$$\mathbb{E}\left[e^{rZ_T}\right] \leq e^{rZ_1} + \frac{2e^{r(v+\theta_T)}}{r\gamma}, \quad \text{where} \quad r = \frac{\gamma}{v^2 + v\gamma/3}. \tag{63}$$

By Jensen's inequality, we have

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

which implies that

$$\begin{aligned}
&\mathbb{E}[Z_T] \leq \frac{1}{r}\log\left(1 + \frac{2e^{r(v+\theta_T)}}{r\gamma}\right) \\
=&\frac{1}{r}\log\left(1 + \frac{6v^2+2v\gamma}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq&\frac{1}{r}\log\left(1 + \frac{8v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq&\frac{1}{r}\log\left(\frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq&\frac{4v^2}{3\gamma}\log\left(\frac{11v^2}{3\gamma^2}e^{r(v+\theta_T)}\right) \\
\leq&\frac{3v^2}{\gamma}\log\left(\frac{2v}{\gamma}\right) + v + \theta_T \\
\leq&\frac{3v^2}{\gamma}\log\left(\frac{2v}{\gamma}\right) + v + \frac{4(\frac{4H^2\iota}{K^2}+\eta\sqrt{H^2\iota}+H^4\iota+\epsilon^2)}{\delta} \\
=&\frac{48H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right) + 2\sqrt{H^4\iota} + \frac{4(\frac{4H^2\iota}{K^2}+\eta\sqrt{H^2\iota}+H^4\iota+\epsilon^2)}{\delta} \\
\leq&\frac{54H^4\iota}{\delta}\log\left(\frac{16H^2\sqrt{\iota}}{\delta}\right) + \frac{16H^2\iota}{K^2\delta} + \frac{4\eta\sqrt{H^2\iota}}{\delta} = \tilde{\mathcal{O}}\left(\frac{\eta H}{\delta}\right),
\end{aligned} \tag{64}$$

which completes the proof of Lemma 5. □

Substituting the results from Lemmas 2 and 5 into (59), 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 (53), we have

$$\sum_{k=1}^{K}\rho - C_1^{\pi_k}(x_{k,1},a_{k,1}) \leq -K\epsilon + K^{\alpha}Z_{K^{1-\alpha}+1} + \sum_{k=1}^{K}\left\{C_{k,1} - C_1^{\pi_k}\right\}(x_{k,1},a_{k,1}). \tag{65}$$

According to inequality (63), 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

$$\begin{aligned}
&\Pr\left(Z_T \geq \frac{1}{r}\log\left(\frac{11v^2}{3\gamma^2}\right) + 2(v+\theta_T)\right)\\
=&\Pr(e^{rZ_T} \geq e^{\log\left(\frac{11v^2}{3\gamma^2}\right)+2r(v+\theta_T)})\\
\leq&\frac{\mathbb{E}[e^{rZ_T}]}{\frac{11v^2}{3\gamma^2}e^{2r(v+\theta_T)}}\\
\leq&\frac{1}{e^{r(v+\theta_T)}} = \tilde{\mathcal{O}}\left(e^{-\eta}\right),
\end{aligned} \tag{66}$$

where the first inequality is from the Markov inequality.

In the proof of Lemma 2, we have shown

$$\begin{aligned}
&\left|\sum_{k=(T-1)K^{\alpha}+1}^{TK^{\alpha}} C_{k,h}(x_{k,h},a_{k,h}) - C_h^{\pi_k}(x_{k,h},a_{k,h})\right|\\
\leq&\left|\sum_{k=(T-1)K^{\alpha}+1}^{TK^{\alpha}} C_{k,h+1}(x_{k,h+1},a_{k,h+1}) - C_{h+1}^{\pi_k}(x_{k,h+1},a_{k,h+1})\right| + \left|\sum_{k=(T-1)K^{\alpha}+1}^{TK^{\alpha}}(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^{\pi_k}(x_{k,h},a_{k,h})\right|\\
&+ HSA + \frac{H^2\sqrt{\iota}K^{\alpha}}{\chi} + \sqrt{H^2SA\iota K^{\alpha}(\chi+1)}
\end{aligned} \tag{67}$$

Following a similar proof as the proof of Lemma 10, 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 (67) over h and by summing it over all frames, we conclude that with probability at at least $1 - \frac{1}{K^2}$,

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

$$\leq 4\sqrt{H^4SA\iota}K^{0.8}, \tag{68}$$

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

Now, by combining inequalities (66) and (68), 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 $1 - \tilde{\mathcal{O}}\left(e^{-K^{0.2}} + \frac{1}{K^2}\right)$

$$\begin{aligned}
&\sum_{k=1}^{K} \rho - C_1^{\pi_k}(x_{k,1}, a_{k,1}) \\
&\leq -K\epsilon + K^{\alpha}\left(\frac{1}{r}\log\left(\frac{11v^2}{3\gamma^2}\right) + 2(v+\theta_T)\right) + 4\sqrt{H^4SA\iota}K^{0.8} \\
&\leq -\sqrt{SAH^6\iota^3}K^{0.8} \leq 0,
\end{aligned} \tag{69}$$

which completes the proof of our main result.

C THE CHOICES OF THE HYPER-PARAMETERS IN TRIPLE-Q

Recall that the regret upper bound in (57) and the constraint violation bound in (59):

$$\begin{aligned}
\text{Regret}(K) =& \mathcal{O}\left(K\epsilon + K^{1-\alpha} + \frac{K}{\chi} + \sqrt{K^{2-\alpha}\chi} + \frac{K}{\eta}\right) \\
\text{Violation}(K) \leq& -K\epsilon + \mathcal{O}\left(K^{\alpha}\eta + K^{1-\alpha} + \frac{K}{\chi} + \sqrt{K^{2-\alpha}\chi}\right).
\end{aligned}$$

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,

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

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 $\eta = \mathcal{O}(K^{0.2})$.

D NOTATION TABLE

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

E AUXILIARY LEMMAS

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

Lemma 7. *Recall that the learning rate used in Triple-Q is* $\alpha_t = \frac{\chi+1}{\chi+t}$, *and*

$$\alpha_t^0 = \prod_{j=1}^{t}(1-\alpha_j) \quad \textit{and} \quad \alpha_t^i = \alpha_i \prod_{j=i+1}^{t}(1-\alpha_j). \tag{70}$$

The following properties hold for α_t^i :

Table 2: 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 (5).
$q_h^{\epsilon,*}$	The optimal solution to the tightened LP (10).
δ	Slater's constant.
b_t	the UCB bonus for given t
$\mathbb{I}(\cdot)$	The indicator function

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

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

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

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

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

□

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
\frac{\alpha_t}{\sqrt{\chi + t}} + (1-\alpha_t) \sum_{i=1}^{t-1} \frac{\alpha_{t-1}^i}{\sqrt{\chi + i}} \leq & \frac{\alpha_t}{\sqrt{\chi + t}} + \frac{2(1-\alpha_t)}{\sqrt{\chi + t - 1}} = \frac{\chi + 1}{(\chi + t)\sqrt{\chi + t}} + \frac{2(t-1)}{(\chi + t)\sqrt{\chi + t - 1}}, \\
= & \frac{1 - \chi - 2t}{(\chi + t)\sqrt{\chi + t}} + \frac{2(t-1)}{(\chi + t)\sqrt{\chi + t - 1}} + \frac{2}{\sqrt{\chi + t}} \\
\leq & \frac{-\chi - 1}{(\chi + t)\sqrt{\chi + t - 1}} + \frac{2}{\sqrt{\chi + t}} \leq \frac{2}{\sqrt{\chi + t}}.
\end{aligned} \tag{71}$$

(e) According to its definition, we have

$$\begin{aligned}
\alpha_t^i = & \frac{\chi + 1}{i + \chi} \cdot \left(\frac{i}{i + 1 + \chi} \frac{i + 1}{i + 2 + \chi} \cdots \frac{t - 1}{t + \chi} \right) \\
= & \frac{\chi + 1}{t + \chi} \cdot \left(\frac{i}{i + \chi} \frac{i + 1}{i + 1 + \chi} \cdots \frac{t - 1}{t - 1 + \chi} \right) \leq \frac{\chi + 1}{\chi + t}.
\end{aligned} \tag{72}$$

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 $\sum_{i=1}^{t} \alpha_t^i = 1$. □

The next lemma establishes upper bounds on $Q_{k,h}$ and $C_{k,h}$ under Triple-Q.

Lemma 8. *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. □

Next, we present the following lemma from Jin et al. (2018), which establishes a recursive relation between $Q_{k,h}$ and Q_h^π for any π. We include the proof so the paper is self-contained.

Lemma 9. *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:*

$$\begin{aligned}(Q_{k,h} - Q_h^\pi)(x,a) =& \alpha_t^0\left\{Q_{(T-1)K^\alpha+1,h} - Q_h^\pi\right\}(x,a) \\ &+ \sum_{i=1}^t \alpha_t^i\left(\left\{V_{k_i,h+1} - V_{h+1}^\pi\right\}(x_{k_i,h+1}) + \left\{\hat{\mathbb{P}}_h^{k_i}V_{h+1}^\pi - \mathbb{P}_h V_{h+1}^\pi\right\}(x,a) + b_i\right), \qquad (73)\\ (C_{k,h} - C_h^\pi)(x,a) =& \alpha_t^0\left\{C_{(T-1)K^\alpha+1,h} - C_h^\pi\right\}(x,a) \\ &+ \sum_{i=1}^t \alpha_t^i\left(\left\{W_{k_i,h+1} - W_{h+1}^\pi\right\}(x_{k_i,h+1}) + \left\{\hat{\mathbb{P}}_h^{k_i}W_{h+1}^\pi - \mathbb{P}_h W_{h+1}^\pi\right\}(x,a) + b_i\right), \qquad (74)\end{aligned}$$

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 (73). The proof for (74) is identical. Recall that under Triple-Q, $Q_{k+1,h}(x,a)$ is updated as follows:

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

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

$$\begin{aligned}Q_{k,h}(x,a) =& (1-\alpha_t)(1-\alpha_{t-1})Q_{k_{t-1},h}(x,a) + (1-\alpha_t)\alpha_{t-1}\left(r_h(x,a) + V_{k_{t-1},h+1}(x_{k_{t-1},h+1}) + b_{t-1}\right) \\ &+ \alpha_t\left(r_h(x,a) + V_{k_t,h+1}(x_{k_t,h+1}) + b_t\right) \\ =& \cdots \\ =& \alpha_t^0 Q_{(T-1)K^\alpha+1,h}(x,a) + \sum_{i=1}^t \alpha_t^i\left(r_h(x,a) + V_{k_i,h+1}(x_{k_i,h+1}) + b_i\right), \qquad (75)\end{aligned}$$

where the last equality holds due to the definition of α_t^i in (70) 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

$$\begin{aligned}Q_h^\pi(x,a) &= \alpha_t^0 Q_h^\pi(x,a) + (1-\alpha_t^0)Q_h^\pi(x,a) \\ &= \alpha_t^0 Q_h^\pi(x,a) + \sum_{i=1}^t \alpha_t^i\left(r(x,a) + \mathbb{P}_h V_{h+1}^\pi(x,a) + V_{h+1}^\pi(x_{k_i,h+1}) - V_{h+1}^\pi(x_{k_i,h+1})\right) \\ &= \alpha_t^0 Q_h^\pi(x,a) + \sum_{i=1}^t \alpha_t^i\left(r_h(x,a) + \mathbb{P}_h V_{h+1}^\pi(x,a) + V_{h+1}^\pi(x_{k_i,h+1}) - \hat{\mathbb{P}}_h^{k_i}V_{h+1}^\pi(x,a)\right)\end{aligned}$$

$$= \alpha_t^0 Q_h^\pi(x,a) + \sum_{i=1}^{t} \alpha_t^i \left(r_h(x,a) + V_{h+1}^\pi(x_{k_i,h+1}) + \left\{ \mathbb{P}_h V_{h+1}^\pi - \hat{\mathbb{P}}_h^{k_i} V_{h+1}^\pi \right\}(x,a) \right). \tag{76}$$

Then subtracting (76) from (75) yields

$$\begin{aligned}(Q_{k,h} - Q_h^\pi)(x,a) =& \alpha_t^0 \left\{ Q_{(T-1)K^\alpha+1,h} - Q_h^\pi \right\}(x,a) \\ &+ \sum_{i=1}^{t} \alpha_t^i \left(\left\{ V_{k_i,h+1} - V_{h+1}^\pi \right\}(x_{k_i,h+1}) + \left\{ \hat{\mathbb{P}}_h^{k_i} V_{h+1}^\pi - \mathbb{P}_h V_{h+1}^\pi \right\}(x,a) + b_i \right).\end{aligned}$$

□

Lemma 10. *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 $k_1,\dots,k_t < k$ be the indices of these episodes. Under any policy π, with probability at least $1-\frac{1}{K^3}$, the following inequalities hold simultaneously for all $(x,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$*

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

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

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

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

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

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

$$|X(i) - X(i-1)| \le \alpha_\tau^i \left| \left\{ (\hat{\mathbb{P}}_h^{k_{n+1}} - \mathbb{P}_h) V_{h+1}^\pi \right\}(x,a) \right| \le \alpha_\tau^i H$$

Then let $\sigma = \sqrt{8\log\left(\sqrt{2SAH}K\right)\sum_{i=1}^{\tau}(\alpha_\tau^i H)^2}$. By applying the Azuma-Hoeffding inequality, we have with probability at least $1-2\exp\left(-\frac{\sigma^2}{2\sum_{i=1}^{\tau}(\alpha_\tau^i H)^2}\right) = 1-\frac{1}{SAHK^4}$,

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

where the last inequality holds due to $\sum_{i=1}^{\tau}(\alpha_\tau^i)^2 \le \frac{\chi+1}{\chi+\tau}$ from Lemma 7.(e). Because this inequality holds for any $\tau\in[K]$, it also holds for $\tau = t = N_{k,h}(x,a) \le K$, Applying the union bound, we obtain that with probability at least $1-\frac{1}{K^3}$ the following inequality holds simultaneously for all $(x,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]$,:

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

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

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

□

This lemma bound the conditional expected Lyapunov drift.

Lemma 11. *Given $\delta \geq 2\epsilon$, under Triple-Q, the conditional expected drift is*

$$\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \leq -\frac{\delta}{2} Z_T + \frac{4H^2\iota}{K^2} + \eta\sqrt{H^2\iota} + H^4\iota + \epsilon^2 \tag{77}$$

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 (55), we have

$$\begin{aligned}
&\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \\
\leq &\frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \mathbb{E}\left[Z_T\left(\rho + \epsilon - C_{k,1}(x_{k,1}, a_{k,1})\right) - \eta Q_{k,1}(x_{k,1}, a_{k,1})\right. \\
&\left. + \eta Q_{k,1}(x_{k,1}, a_{k,1}) | Z_T = z\right] + H^4\iota + \epsilon^2 \\
\leq_{(a)} &\frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \mathbb{E}\left[Z_T\left(\rho + \epsilon - \sum_a \left\{C_{k,1} q_1^\pi\right\}(x_{k,1}, a)\right)\right. \\
&\left. - \eta \sum_a \{Q_{k,1} q_1^\pi\}(x_{k,1}, a) + \eta Q_{k,1}(x_{k,1}, a_{k,1}) | Z_T = z\right] \\
&+ \epsilon^2 + H^4\iota \\
\leq &\frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \mathbb{E}\left[Z_T\left(\rho + \epsilon - \sum_a \left\{C_1^\pi q_1^\pi\right\}(x_{k,1}, a)\right)\right. \\
&\left. - \eta \sum_a \{Q_{k,1} q_1^\pi\}(x_{k,1}, a) + \eta Q_{k,1}(x_{k,1}, a_{k,1}) | Z_T = z\right] \\
&+ \frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \mathbb{E}\left[Z_T \sum_a \left\{C_1^\pi q_1^\pi\right\}(x_{k,1}, a) - Z_T \sum_a \left\{C_{k,1} q_1^\pi\right\}(x_{k,1}, a) | Z_T = z\right] \\
&+ \frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \mathbb{E}\left[\eta \sum_a \left\{Q_1^\pi q_1^\pi\right\}(x_{k,1}, a) - \eta \sum_a \left\{Q_1^\pi q_1^\pi\right\}(x_{k,1}, a) | Z_T = z\right] + H^4\iota + \epsilon^2 \\
\leq_{(b)} &-\frac{\delta}{2} z + \frac{1}{K^\alpha} \sum_{k=(T-1)K^\alpha+1}^{TK^\alpha} \mathbb{E}\left[\eta \sum_a \left\{(F_1^\pi - F_{k,1}) q_1^\pi\right\}(x_{k,1}, a) + \eta Q_{k,1}(x_{k,1}, a_{k,1}) \middle| Z_T = z\right] + H^4\iota + \epsilon^2 \\
\leq_{(c)} &-\frac{\delta}{2} z + \frac{4H^2\iota}{K^2} + \eta\sqrt{H^2\iota} + H^4\iota + \epsilon^2.
\end{aligned}$$

Inequality (a) holds because of our algorithm. Inequality (b) holds because $\sum_a \left\{Q_1^\pi q_1^\pi\right\}(x_{k,1}, a)$ 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 8, 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 (52) (note that the overestimation result and the concentration result in frame T hold regardless of the value of Z_T). □

Part 2: Simulations

F EVALUATION

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^α) = 0 ;                          // reset visit counts and add bonuses to Q-functions
2 then
3     N_h(x,a) ← 0 and Q_h(x,a) ← Q_h(x,a) + 2H^3√ι/η, ∀(x,a,h).
4     if Q_h(x,a) ≥ H or C_h(x,a) ≥ H then
5         Q_h(x,a) ← H and C_h(x,a) ← H;
6         Z ← (Z + ρ + ε − C̄/K^α)^+, and C̄ ← 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 3 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

$$\begin{aligned}
\{F_{TK^\alpha+1,h} - F^\pi_h\}(x,a) =& Q^-_{TK^\alpha+1,h}(x,a) + \frac{Z_{TK^\alpha}}{\eta} C^-_{TK^\alpha+1,h}(x,a) - Q^\pi_h(x,a) - \frac{Z_{TK^\alpha}}{\eta} C^\pi_h(x,a) \\
&+ \frac{2H^3\sqrt{\iota}}{\eta} + \frac{Z_{TK^\alpha+1} - Z_{TK^\alpha}}{\eta} C^-_{TK^\alpha+1,h}(x,a) - \frac{Z_{TK^\alpha+1} - Z_{TK^\alpha}}{\eta} C^\pi_h(x,a) \\
\geq& \frac{2H^3\sqrt{\iota}}{\eta} - 2\frac{|Z_{TK^\alpha+1} - Z_{TK^\alpha}|}{\eta} H \\
\geq& 0,
\end{aligned} \tag{78}$$

where the last inequality holds because according to Lemma 8,

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

In a summary, condition (i) in the proof of Lemma 3 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.

F.1 A Tabular Case

We first evaluated our algorithm using a grid-world environment studied in Chow et al. (2018). The environment is shown in Figure 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 are the Euclidean distance between them and the destination subtracted from the longest distance. The cost constraint is set to be 6 (we transferred utility to cost as we discussed in the paper), which means the agent is only allowed to hit the obstacles as most six times. To account for the statistical significance, all results were averaged over 25 trials, same for later simulations.

The result is shown in Figure 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 similar performance as Triple-Q, and show that Triple-Q yields a near-optimal, stationary policy after the learning stops.

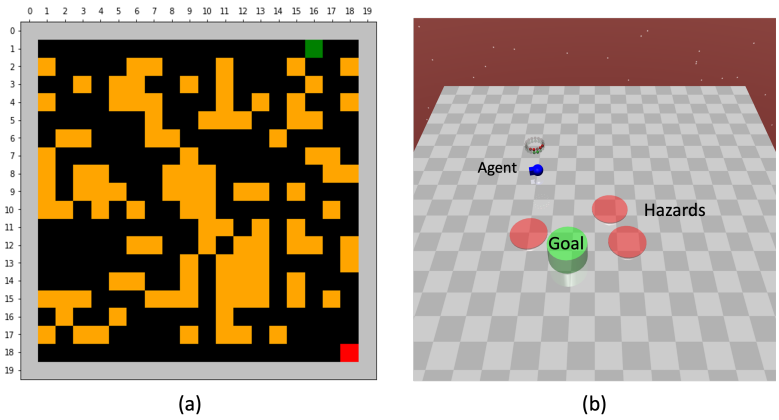


Figure 1: Grid World and DynamicEnv[1] with Safety Constraints

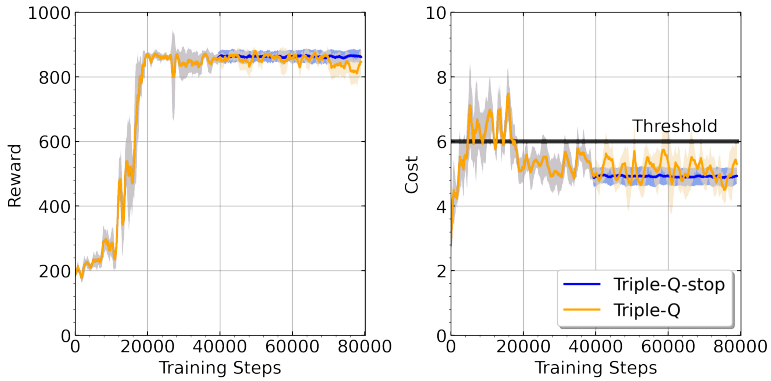


Figure 2: The average reward and cost under Triple-Q during training. The shaded region represents the 95% confidence interval.

Ablation Study

We investigate Triple-Q's sensitivity to hyperparameter η via an ablation study. As shown in Figure 3, a smaller η, which represents a higher weight on constraint, results in a lower cost while maintaining a similar performance in terms of reward.

F.2 Triple-Q with Neural Networks

We also evaluated our algorithm on the Dynamic Gym benchmark (DynamicEnv) Yang et al. (2021) as shown in Figure. 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 hazards are randomly generated in each episode. At each step, the agents get a cost of 1 if it stays in the hazardous area; and 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 3. These hyperparameters are used in two later environments, pendulum and Ball-1D, as well. We call this algorithm Deep

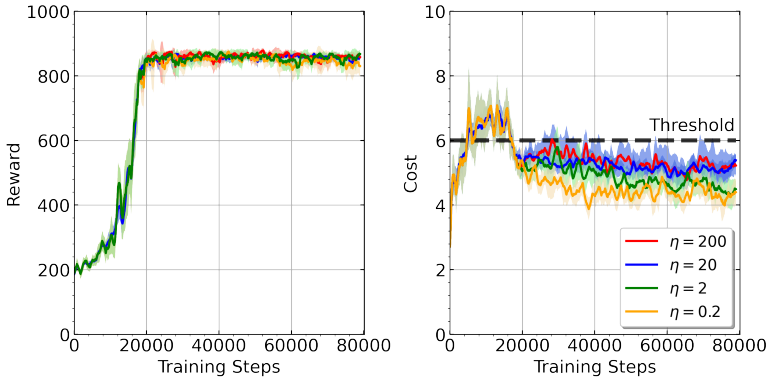


Figure 3: Performance of Triple-Q under different choices of η in the Grid World

Triple-Q. The simulation results in Figure 4 show that Deep Triple-Q learns a safe-policy with a high reward much faster than WCSAC Yang et al. (2021). In particular, it took around 0.45 million training steps under Deep Triple-Q, but it took 4.5 million training steps under WCSAC.

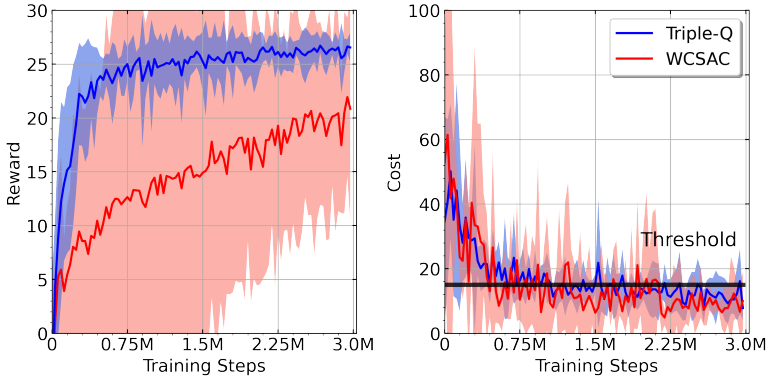


Figure 4: The rewards and costs of Deep Triple-Q versus WCSAC during Training

Table 3: 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 Cheng et al. (2019) on the Pendulum environment.(https://gym.openai.com/envs/Pendulum-v0/) 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 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 5, we observed that Triple-Q achieved a higher reward. Although

[1]**Image Sorce:** The environment is generated using safety-gym: https://github.com/openai/safety-gym.

Triple-Q violated the constrained at the early stage and cannot guarantee an 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 a prior knowledge while Deep Triple-Q does not.

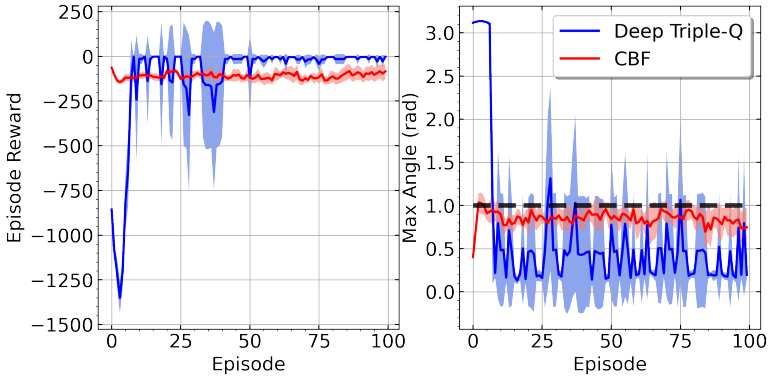
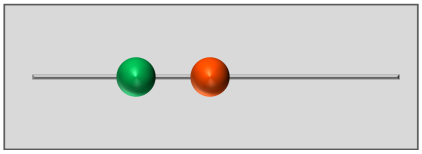
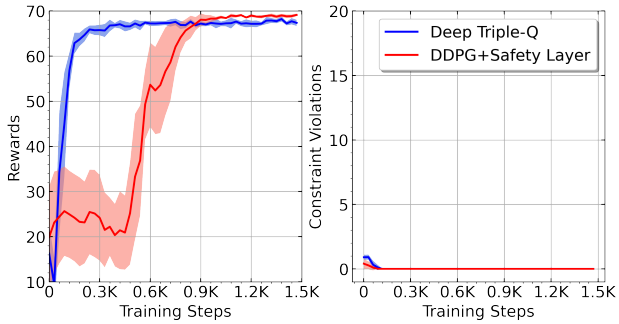


Figure 5: Comparison with CBF

Finally, we compared Triple-Q with DDPG+Safety Layer in Dalal et al. (2018) on Ball-1D environment (Figure 6a), 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 6b.



(a) Ball1d Environment



(b) Performance during training

Figure 6: Comparison with DDPG+Safety Layer