A Provably-Efficient Model-Free Algorithm for Infinite-Horizon Average-Reward Constrained Markov Decision Processes

Honghao Wei,[1] **Xin Liu,** [2] **Lei Ying** [1]

[1] University of Michigan, Ann Arbor
[2] ShanghaiTech University
honghaow@umich.edu, liuxin7@shanghaitech.edu.cn, leiying@umich.edu

Abstract

This paper presents a model-free reinforcement learning (RL) algorithm for infinite-horizon average-reward Constrained Markov Decision Processes (CMDPs). Considering a learning horizon K, which is sufficiently large, the proposed algorithm achieves $\tilde{O}\left(\frac{\sqrt{SA}\kappa}{\delta}K^{\frac{5}{6}}\right)$ regret and zero constraint violation, where S is the number of states, A is the number of actions, and κ and δ are two constants independent of the learning horizon K.

Introduction

Reinforcement Learning has gained significant attention because of its successes in board games and video games such as Go (Silver et al. 2017) and StarCraft (Vinyals et al. 2019), and in highly-complex robotics systems (Andrychowicz et al. 2020). An agent's objective in a typical RL problem is to maximize the cumulative reward through interacting with an unknown environment. In board games or video games, the outcomes of a random action are not consequential to the users (e.g. not life-threatening). However, a careless action in an engineering system might have catastrophic outcomes such as collisions and fatalities in robotics and autonomous driving (Ono et al. 2015; Garcia and Fernández 2012; Fisac et al. 2018) or surgical robotics (Richter, Orosco, and Yip 2019). Therefore, it is critical to strike a balance between reward maximization and safety in real-world applications. A standard formulation for RL with constraints is the Constrained Markov Decision Processes framework (Altman 1999), in which the agent aims at learning a policy that maximizes the expected cumulative reward under safety constraints during and after learning.

A CMDP problem can be solved using linear programming (LP) or the dual approach (Altman 1999) when the model (the transition kernel) is given. For example, (Zheng and Ratliff 2020) proposed a LP-based algorithm which learns the optimal policy while satisfying the constraints for a CMDP with a known model. Recent model-based algorithms (Singh, Gupta, and Shroff 2020; Brantley et al. 2020; Kalagarla, Jain, and Nuzzo 2021; Efroni, Mannor, and Pirotta 2020; Qiu et al. 2020) follow a similar approach but learn the models from the data samples collected. This approach has also been utilized for CMDPs with linear function approximation (Ding et al. 2021) under the assumption that the transition kernel is linear. Leveraging the estimated model, the CMDPs can be solved approximately as long as the estimate becomes more and more accurately. The works mentioned above are proved to achieve sublinear constraint violation. A detailed discussion on the sample complexity of CMDPs of model-based approaches can be found in (HasanzadeZonuzy, Kalathil, and Shakkottai 2021). While model-based RL algorithms are sample efficient, they need to continuously solve LPs when the estimated models are updated, so these algorithms are often computationally inefficient and require a large memory to maintain a large number of model parameters.

Model-free algorithms, on the other hand, learn state or action value functions, instead of transition kernels, so require significantly less memory space and have lower computational complexity. In (Borkar 2005), the author proposes an actor-critic RL algorithm and shows its asymptotic global convergence using multi-timescale stochastic approximation theory for infinite-horizon average-reward CMDPs when the model is unknown. Policy gradient approaches have also been developed (Tessler, Mankowitz, and Mannor 2018; Stooke, Achiam, and Abbeel 2020; Yang et al. 2020) and seen successes in practice for solving constrained RL problems, though they lack regret and constraint violation analysis. (Ding et al. 2020; Xu, Liang, and Lan 2020; Chen, Dong, and Wang 2021) show that sublinear regrets and constraint violations are achievable when policy "simulators" (or generative models) are given. Two very recent works (Liu et al. 2021a; Wei, Liu, and Ying 2021) show that sublinear regret bound and zero violation are possible for episodic CMDPs without simulators. In particular, (Liu et al. 2021a) proposes a model-based algorithm and (Wei, Liu, and Ying 2022) presents a model-free algorithm, and (Bura et al. 2021) proves that it is possible to achieve zero violation during training given a safe baseline policy based on a model-based approach. Despite these significant developments, the following question is still open:

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

We answer this question affirmatively and present a

	Algorithm	**Regret**	**Constraint Violations**
Known Model	C-UCRL (Zheng and Ratliff 2020)	$\tilde{\mathcal{O}}(SA\sqrt{K^{1.5}})$	0
Model-based	UCRL-CMDP (Singh, Gupta, and Shroff 2020)	$\tilde{\mathcal{O}}(S\sqrt{A}K^{\frac{2}{3}})$	$\tilde{\mathcal{O}}(S\sqrt{A}K^{\frac{2}{3}})$
Known Model	CMDP-PSRL (Agarwal, Bai, and Aggarwal 2021)	$\tilde{\mathcal{O}}(\text{poly}(SAD)\sqrt{K})$	$\tilde{\mathcal{O}}(\text{poly}(SA)\sqrt{K})$
Model-based*	OptPess-LP (Liu et al. 2021a)	$\tilde{\mathcal{O}}(H^3\sqrt{S^3AK})$	0
Model-based*	OptPess–PrimalDual (Liu et al. 2021a)	$\tilde{\mathcal{O}}(H^3\sqrt{S^3AK})$	$\mathcal{O}(1)$
Model-based*	OPSRL(Bura et al. 2021)	$\tilde{\mathcal{O}}(\sqrt{S^4H^7AK})$	0
Model-free	**This Paper**	$\tilde{\mathcal{O}}\left(\frac{\sqrt{SA}}{\delta}K^{\frac{5}{6}}\right)$	0

Table 1: Regrets and constraint violations of RL algorithms for infinite-horizon average-reward CMDPs ($*$ These algorithms are designed for episodic CMDPs). S is the number of states, A is the number of actions, K is the number of steps, D is the diameter of the CMDP whose definition can be found in the supplementary material, δ is the slackness that will be defined later (Eq. (10)), and $\text{poly}(X)$ denotes a polynomial function of x. Throughout the paper, we use the notation $\tilde{\mathcal{O}}$ to suppress log terms. $\tilde{O}(f(K))$ denotes $\mathcal{O}(f(K)\log^n K)$ with $n > 0$.

model-free RL algorithm that achieves sub-linear regret and zero constraint violation. Table 1 compares the results in this paper with those in the literature. We remark that the proposed algorithm synthesizes the Triple-Q algorithm in (Wei, Liu, and Ying 2022) for episodic CMDPs and Optimistic Q-Learning (Wei et al. 2020) that reduces the average-reward problem to a discounted reward problem.

Preliminaries

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

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

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

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

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

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

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

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

$$\max_\pi J_r^\pi \quad \text{s.t.} \quad J_g^\pi \geq \rho. \tag{1}$$

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

$$\text{Regert}(K) = \mathbb{E}\left[\sum_{k=1}^{K}\left(J_r^{\pi^*} - r(s_k, a_k)\right)\right], \tag{2}$$

$$\text{Violation}(K) = \mathbb{E}\left[\sum_{k=1}^{K}\left(\rho - g(s_k, a_k)\right)\right]. \tag{3}$$

When the transition kernel $p(s'|s, a)$ is known, the optimal stationary policy that solves problem (1) can be obtained by solving the following LP problem (Altman 1999):

$$\max_{\{q(s,a):(s,a)\in\mathcal{S}\times\mathcal{A}\}} \sum_{s,a} q(s, a)r(s, a) \tag{4}$$

$$\text{s.t.} \sum_{s,a} q(s, a)g(s, a) \geq \rho, \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{5}$$

$$q(s, a) \geq 0, \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{6}$$

$$\sum_{s,a} q(s, a) = 1 \tag{7}$$

$$\sum_{a} q(s, a) = \sum_{s',a'} p(s|s', a')q(s', a'), \tag{8}$$

where the $q(s, a)$ is called the occupancy measure, which is defined as the set of distributions generated by executing the associated induced policy π in the infinite-horizon CMDP. $\sum_a q(s, a)$ represents the probability the system is in state s, and $\frac{q(s,a)}{\sum_{a'} q(s,a')}$ is the probability of taking action a in state s. The utility constraint is represented in (5). More details can be found in (Altman 1999).

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

$$\max_{\{q(s,a):(s,a)\in\mathcal{S}\times\mathcal{A}\}} \sum_{s,a} q(s,a)r(s,a) \tag{9}$$

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

$$(6) - (8),$$

where $\epsilon > 0$ is called a tightness constant. As in previous works (Ding et al. 2021, 2020; Efroni, Mannor, and Pirotta 2020; Paternain et al. 2019), we make the following standard assumption of Slater's condition.

Assumption 1. *(Slater's Condition). There exist $\delta > 0$ and a feasible solution $q(s,a)$ to the LP such that*

$$\sum_{s,a} q(s,a)g(s,a) - \rho \geq \delta. \tag{10}$$

It is obvious that when $\epsilon < \delta$ the problem (9) has a feasible solution due to Slater's condition. The Slater's condition is commonly assumed in previous works to ensure the LP problem has strong duality, see proofs in (Paternain et al. 2019, 2022). Unlike (Ding et al. 2021; Efroni, Mannor, and Pirotta 2020), which assume δ is known, and (Achiam et al. 2017; Liu et al. 2021a; Bura et al. 2021), where a strictly feasible policy is given, our assumption is less restrictive. Let

$$J_r^* = \sum_{s,a} q^*(s,a)r(s,a), \tag{11}$$

$$J_g^* = \sum_{s,a} q^*(s,a)g(s,a). \tag{12}$$

be the optimal reward rate and utility rate, where $q^*(s,a)$ is the optimal solution obtained by solving the LP problem (4). Moreover it is obvious that J_r^* and J_g^* are independent of the initial state and we have $J_r^* = J_r^{\pi^*}$ and $J_g^* = J_g^{\pi^*}$.

In the following, we use superscript * to denote the optimal policy achieved by solving the LP (4) of the original CMDP, and superscript $^{\epsilon,*}$ to denote the optimal policy related to the ϵ-tightened version of LP (9).

Primal-Dual-Based, Two-Time-Scale Optimistic SARSA (Triple-QA)

In this section, we introduce our algorithm Triple-QA (see Algorithm 1 for pseudo-code) which achieves sub-linear regret and zero constraint violation. The algorithm is inspired by Triple-Q, an algorithm for episodic CMDPs in (Wei, Liu, and Ying 2021). Triple-QA is for the infinite-horizon average-reward CMDPs with a different update rule. The algorithm further solves the discounted CMDPs with the discount factor γ close to one, an idea used in (Wei et al. 2020). The discounted CMDP is defined on the same state space, action space, reward/utility functions, the transition kernel. The intuition is that the reward of the discounted problem (scaled by $1-\gamma$)) approaches to that of the average reward problem as γ goes to 1.

Algorithm 1: Triple-QA

```
Initialize $Q_1(s,a) = \hat{Q}_1(s,a) \leftarrow H = K^{\frac{1}{6}}$ and $n_1(s,a) \leftarrow 0, \forall(s,a) \in \mathcal{S}\times\mathcal{A}, \gamma = 1 - \frac{1}{H}, \hat{V}_1(s) = H, \forall s \in \mathcal{S}$
Choose $\chi = K^{\frac{1}{3}}, \eta = K^{\frac{1}{6}}, \iota = 8\log(\sqrt{2}K), \beta = \frac{2}{3}$.
Choose $\epsilon = \frac{9\kappa\sqrt{SA\iota}}{K^{\frac{1}{6}}}, \kappa(Eq.(17))\}$.
Initialize $\bar{C} \leftarrow 0, Z_1 \leftarrow 0$.
Define , $\alpha_\tau = \frac{\chi+1}{\chi+\tau}, b_\tau = \kappa\sqrt{\frac{(\chi+1)\iota}{\chi+\tau}}$.
for episode $k = 1, \dots, K$ do
    Take $a_k = \arg\max_a \left(\hat{Q}_k(s_k,a) + \frac{Z}{\eta}\hat{C}_k(s_k,a)\right)$.
    Observe $s_{k+1}$.
    $n_{k+1}(s_k,a_k) \leftarrow n_k(s_k,a_k)+1, \tau \leftarrow n_{k+1}(s_k,a_k)$.
    Update $Q_{k+1}(s_k,a_k) \leftarrow (1-\alpha_\tau)Q_k(s_k,a_k)$
                $+\alpha_\tau[r(s_k,a_k) + \gamma\hat{V}_k(s_{k+1}) + b_\tau]$,
    Update $C_{k+1}(s_k,a_k) \leftarrow (1-\alpha_\tau)C_k(s_k,a_k)$
                $+\alpha_\tau[g(s_k,a_k) + \gamma\hat{W}_k(s_{k+1}) + b_\tau]$.
    if $Q_{k+1}(s_k,a_k) \leq \hat{Q}_k(s_k,a_k)$ and $C_{k+1}(s_k,a_k) \leq \hat{C}_k(s_k,a_k)$ then
        $\hat{Q}_{k+1}(s_k,a_k) \leftarrow Q_{k+1}(s_k,a_k)$
        $\hat{C}_{k+1}(s_k,a_k) \leftarrow C_{k+1}(s_k,a_k)$
    else
        $\hat{Q}_{k+1}(s_k,a_k) \leftarrow \hat{Q}_k(s_k,a_k)$
        $\hat{C}_{k+1}(s_k,a_k) \leftarrow \hat{C}_k(s_k,a_k)$
    $\bar{C} \leftarrow \bar{C} + (1-\gamma)\hat{C}_k(s_k,a_k)$
    $a' = \arg\max_a \left(\hat{Q}_{k+1}(s_k,a) + \frac{Z}{\eta}\hat{C}_{k+1}(s_k,a)\right)$
    $\hat{V}_{k+1}(s_k) \leftarrow \hat{Q}_{k+1}(s_k,a')$
    $\hat{W}_{k+1}(s_k) \leftarrow \hat{C}_{k+1}(s_k,a')$
    if $k$ mod $K^\beta = 0$ then
        $Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}}{K^\beta}\right)$
        Reset $\bar{C} \leftarrow 0, n_t(s,a) \leftarrow 0$.
        Reset $\hat{Q}_{k+1}(s,a), Q_{k+1}(s,a), V_{k+1}(s)$ to $H$
        Reset $\hat{C}_{k+1}(s,a), C_{k+1}(s,a), W_{k+1}(s)$ to $H$
```

Under the discounted CMDP setting, given a policy π, the reward value function V_k^π at step k is the expected cumulative rewards from step k under policy π :

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

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

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

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

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

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

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

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

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

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

The algorithm is an extension of Triple-Q (Wei, Liu, and Ying 2021) for episodic CMDPs by including the discount factor and replacing episode-by-episode updates with step-by-step updates. We adopt the same notations used in (Wei et al. 2020). Same as Triple-Q, the algorithm maintains an estimate $\hat{V}_k(s)$ $(\hat{W}_k(s))$ for the optimal value function $V^*(s)(W^*(s))$ and $\hat{Q}_k(s,a)$ $(\hat{C}_k(s,a))$ for the optimal Q-function $Q^*(s,a)$ $(C^*(s,a))$. At each step k, after observing state s, the agent selects action a_k^* based on the combined Q-value:

$$a_k^* \in \arg\max_{a} \hat{Q}_k(s,a) + \frac{Z}{\eta}\hat{C}_k(s,a), \tag{13}$$

where $\frac{Z}{\eta}$ can be treated as an estimate of the Lagrange multiplier μ. Similar to (Wei, Liu, and Ying 2021), we need to carefully tune the frequency of updating the Lagrange multiplier to balance the convergence and optimality. Updating it too frequent would lead to divergence and too infrequent would result in a large regret and large constraint violation. The algorithm tackles this difficulty by updating Z at a slow time-scale, i.e., every K^{β} steps in line $25-26$ in Algorithm 1, with the following update

$$Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}}{K^{\beta}}\right)^{+}, \tag{14}$$

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

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

It is important to note that for a CMDP,

$$V^*(s) \neq \max_{a} Q^*(s,a). \tag{15}$$

This means optimistic Q-learning algorithms for unconstrained MDPs (e.g. (Jin et al. 2018; Wei et al. 2020; Dong et al. 2019)) cannot be used for estimating the optimal value functions of CMDPs. Instead, Triple-Q (Wei, Liu, and Ying 2021) and this algorithms use a SARSA-type updating rule, as shown in line $11-14$.

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

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

$$J_r^{\pi} + q^{\pi}(s,a) = r(s,a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[v^{\pi}(s')]$$
$$v^{\pi}(s) = \sum_{a} q^{\pi}(s,a)\mathbb{P}(\pi(s)=a)$$
$$J_g^{\pi} + c^{\pi}(s,a) = g(s,a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[w^{\pi}(s')]$$
$$w^{\pi}(s) = \sum_{a} c^{\pi}(s,a)\mathbb{P}(\pi(s)=a)$$

Define

$$sp(f) = \max_{s\in\mathcal{S}} f(s) - \min_{s\in\mathcal{S}} f(s) \tag{16}$$

to be the span of the function f. It is well known that the span of the optimal reward relative value function $sp(v^*)$ and utility relative value function $sp(w^*)$ are bounded for weakly communication or ergodic MDPs. In particular, they are bounded by the diameter of the MDP (Lattimore and Szepesvári 2020).

Let

$$\kappa = \max_{0\le\epsilon\le\rho/2}(\max\{sp(v^{\epsilon,*}), sp(w^{\epsilon,*}), 1\}) \tag{17}$$

and assume that κ which is used in the algorithm is known beforehand as in (Wei et al. 2020, 2021). We can always substitute them with any upper bound (e.g. the diameter) when it is unknown. We now state the main results in the following theorem.

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

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

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

where $\iota = 32\log(\sqrt{2}K)$. □

Proof of the Main Theorem

Notations

Throughout the paper, we use shorthand notation

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

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

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

Due to the page limit, we will only present several key lemmas and the key intuitions in this section. The complete proof can be found in the appendix.

Regret Analysis

We start the proof by adding and subtracting the corresponding terms to the regret defined in (2), we obtain

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

$$=\mathbb{E}\left[\sum_{k=1}^{K}(J_r^* - J_r^{\epsilon,*})\right] \tag{18}$$

$$+\mathbb{E}\left[\sum_{k=1}^{K}(J_r^{\epsilon,*} - (1-\gamma)V^{\epsilon,*}(s_k))\right] \tag{19}$$

$$+\mathbb{E}\left[\sum_{k=1}^{K}(1-\gamma)\left(V^{\epsilon,*}(s_k) - \hat{Q}_k(s_k,a_k)\right)\right] \tag{20}$$

$$+\mathbb{E}\left[\sum_{k=1}^{K}\left((1-\gamma)\hat{Q}_k(s_k,a_k) - r(s_k,a_k)\right)\right]. \tag{21}$$

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

Term (18) is the difference between the original CMDP and its corresponding $\epsilon-$tighten version which is a perturbation of the original problem. We establish a bound by using the following lemma. The proof is deferred to supplementary material.

Lemma 2. *Under assumption 1, given $\epsilon \leq \delta$, we have*

$$\sum_{t=1}^{K}(J_r^* - J_r^{\epsilon,*}) \leq \frac{\epsilon K}{\delta} \tag{22}$$

For the second term (19), we establish a bound by using Lemma 3, which shows the difference between value functions of the average-reward problem and the value functions of the discounted setting problem is small. The proof is based on the Bellman equations under teh two settings. The proof follows Lemma 2 in (Wei et al. 2020) closely.

Lemma 3. *For an arbitrary policy π, we have*

$$J_r^{\pi} - (1-\gamma)V^{\pi}(s) \leq (1-\gamma)sp(v^{\pi}(s)), \tag{23}$$

$$|V^{\pi}(s_1) - V^{\pi}(s_2)| \leq 2sp(v^{\pi}(s)); \tag{24}$$

$$J_g^{\pi} - (1-\gamma)W^{\pi}(s) \leq (1-\gamma)sp(w^{\pi}(s)), \tag{25}$$

$$|W^{\pi}(s_1) - W^{\pi}(s_2)| \leq 2sp(w^{\pi}(s)), \tag{26}$$

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

Then it is easy to obtain

$$J_r^{\epsilon,*} - (1-\gamma)V^{\epsilon,*}(s) \leq (1-\gamma)\kappa, \tag{27}$$

Next we establish a bound on term (20) by using the Lyapunov-drift analysis. In unconstrained MDPs, the bound is established by showing that optimistic Q-learning guarantees that $\hat{Q}_k(s,a)$ is an overestimate of $Q^*(s,a)$. However this does not hold in CMDPs because the algorithm needs to consider reward and utility simultaneously so $\hat{Q}_k(s,a)$ is not necessarily an overestimate of $Q^*(s,a)$. To bound this term, we first add and subtract some additional terms to obtain

$$\sum_{k=1}^{K}(1-\gamma)\left(V^{\epsilon,*}(s_k) - \hat{Q}_k(s_k,a_k)\right)$$

$$=\sum_{k=1}^{K}(1-\gamma)\sum_a\left\{Q^{\epsilon,*}q^{\epsilon,*} + \frac{Z_k}{\eta}C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a) \tag{28}$$

$$-\sum_{k=1}^{K}(1-\gamma)\sum_a\left\{\hat{Q}_k q^{\epsilon,*} + \frac{Z_k}{\eta}\hat{C}_k q^{\epsilon,*}\right\}(s_k,a) \tag{29}$$

$$+\sum_{k=1}^{K}(1-\gamma)\left(\sum_a\left\{\hat{Q}_k q^{\epsilon,*}\right\}(s_k,a) - \hat{Q}_k(s_k,a_k)\right. \tag{30}$$

$$\left.+\frac{Z_k}{\eta}\sum_a\left\{\hat{C}_k q^{\epsilon,*} - C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a)\right). \tag{31}$$

We can see (28) + (29) is the difference of the two combined Q functions. We will show that $\left\{\hat{Q}_k + \frac{Z_k}{\eta}\hat{C}_{k,h}\right\}(s,a)$ is always an over-estimate of $\left\{Q^{\epsilon,*} + \frac{Z_k}{\eta}C^{\epsilon,*}\right\}(s,a)$ (i.e. (28) + (29) $\leq$ 0) for all (s,a,k) simultaneously with a high probability in Lemma 4. This result further implies an upper bound in expectation

$$\mathbb{E}\left[(28) + (29)\right] \leq (1-\gamma)\frac{3H}{\eta K}. \tag{32}$$

Lemma 4. *With probability at least $1 - \frac{1}{K^3}$, the following inequality holds simultaneously for all $(s,a,k) \in \mathcal{S}\times\mathcal{A}\times[K]$:*

$$\left\{\left(\hat{Q}_k - Q^{\epsilon,*}\right) + \frac{Z_k}{\eta}\left(\hat{C}_k - C^{\epsilon,*}\right)\right\}(s,a) \geq 0, \tag{33}$$

Then for the term (30) + (31), we can bound it by using the following lemma.

Lemma 5. *Assuming $\epsilon < \delta$, we have*

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

$$\left.\left.+\frac{Z_k}{\eta}\sum_a\left\{\hat{C}_k q^{\epsilon,*} - C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a)\right)\right]$$

$$\leq\frac{2K}{\eta} + \sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{\eta}. \tag{34}$$

To see the idea behind Lemma 5, we need to consider the Lyapunov function $L_T = \frac{1}{2}Z_T^2$, where T is the frame index and Z_T is the virtual-queue length at the beginning of Tth frame. Recall that each frame contains K^{β} consecutive steps. In the proof of Lemma 5, we will show that the Lyapunov-drift satisfies

$$\begin{aligned}&\mathbb{E}[L_{T+1}-L_T] \leq \text{a negative drift}\\ &+2+\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{K^{\beta}}-\frac{\eta}{K^{\beta}}\sum_{k=TK^{\beta}+1}^{(T+1)K^{\beta}}\Phi_k, \qquad (35)\end{aligned}$$

where

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

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

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

Therefore

$$\begin{aligned}&(30)+(31)=\sum_k\Phi_k\\ \leq&\frac{K^{\beta}\mathbb{E}[L_1-L_{K^{1-\beta}+1}]}{\eta}+\frac{2K}{\eta}+\sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{\eta}\\ \leq&\frac{2K}{\eta}+\sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{\eta},\end{aligned}$$

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

Then combining the result form (32) and Lemma 5, we can obtain

$$\begin{aligned}&\sum_{k=1}^{K}((1-\gamma)\left(V^{\epsilon,*}(s_k)-\hat{Q}_k(s_k,a_k))\right)\\ \leq&\frac{2K}{\eta}+\sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{\eta}+\frac{3H}{\eta K}. \qquad (36)\end{aligned}$$

The term $\mathbb{E}[Z_T]$ is proved uniformly bounded in Lemma 6 by using the Lyapunov-drift analysis on the moment generating function of Z i.e. $\mathbb{E}[e^rZ]$ can be bounded by a constant uniformly over the entire learning horizon. The reason is that when the virtual queue Z is large, our algorithm takes actions to almost greedily reduce the virtual-queue.

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

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

We apply the following lemma to bound the last term (21).

Lemma 7. *For any $T\in[K^{1-\beta}]$ and any $m\in\mathbb{Z}^+$,*

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

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

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\left(\left\{(1-\gamma)\hat{Q}_t-r\right\}(s_k,a_k)\right)\right]\leq\gamma^mK+\frac{Km}{\chi}\\ &+4(1-\gamma)m\kappa\sqrt{(\chi+1)SAK^{2-\beta}\iota}+2mSK^{1-\beta} \qquad (37)\end{aligned}$$

To balance the terms in regret, we carefully select that

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

Then we have

$$\gamma^m=\left(1-\frac{1}{H}\right)^{H\log K}\leq\frac{1}{K}, \qquad (38)$$

and the order of the second and third terms in the above equation (37) is $\tilde{O}(K^{\frac{5}{6}})$, which is also the dominate term in our regret bound.

Then by appropriately choosing other parameters ϵ, ι and η, to balance the terms and combining the results from (36), (37), Lemma 2, Lemma 3, and Lemma 6, we finish the proof for the regret bound. The details can be found in the supplementary material.

Constraint Violation Analysis

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

$$\begin{aligned}Z_{T+1}=&\left(Z_T+\rho+\epsilon-\frac{\bar{C}_T}{T^{\beta}}\right)^+\\ \geq&Z_T+\rho+\epsilon-\frac{\bar{C}_T}{K^{\beta}}, \qquad (39)\end{aligned}$$

which implies that

$$\sum_{k=(T-1)K^{\beta}+1}^{TK^{\beta}}(-g(s_k,a_k)+\rho)\leq K^{\beta}\left(Z_{T+1}-Z_T\right)$$

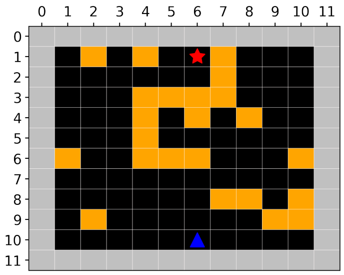


Figure 1: A Grid World with Safety Constraints

$$+\sum_{k=(T-1)K^{\beta}+1}^{TK^{\beta}}\left((1-\gamma)\hat{C}_k(s_k,a_k)-g(s_k,a_k)-\epsilon\right).$$

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

$$\mathbb{E}\left[\sum_{t=1}^{T}\rho-g(s_k,a_k)\right]\leq -K\epsilon+K^{\beta}\mathbb{E}\left[Z_{K^{1-\beta}+1}\right]$$
$$+\,\mathbb{E}\left[\sum_{k=1}^{K}(1-\gamma)\hat{C}_k(s_k,a_k)-g(s_k,a_k)\right], \qquad (40)$$

where we used the fact $Z_1 = 0$. Combining the upper bound on the estimation error of $\hat{C}_k$ in Lemma 7 and the upper bound on $\mathbb{E}[Z_T]$ in Lemma 6 yields the constraint violation bound. Furthermore, under our carefully choices of $m, \gamma, \epsilon, \eta, \alpha, \beta$ and ι, it can be easily verified that $K\epsilon$ dominates the upper bounds in (40), which leads to fact constraint violation because zero when K is sufficiently large. In particular, under our assumption on K, which implies that $\epsilon \leq \frac{\delta}{2}$, and leads to

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

The details can be found in the supplementary material.

Simulations

In this section, we present simulation results that evaluate our algorithm using the 2D safety grid-world exploration problem (Zheng and Ratliff 2020; Leike et al. 2017). Figure 1 shows the map of a 10×10 grid-world with a total of 100 states. We chose an error probability 0.03 which means with probability 0.03 the agent will choose an action uniformly at random to make the environment stochastic. The objective of the agent is to travel to the destination (the red star) from the original position (the blue triangle) as quickly as possible while limiting the number of times hitting the obstacles (the yellow squares). Hitting an obstacle incurs cost 1 and otherwise, there is no cost. The reward for the destination is 1, and for others are the normalized Euclidean distance between them and the destination times a scaled factor 0.1. We set constraint limit as 0.15 through the simulation which means the expected cost rate should below the limit. To account for statistical significance, the results of each experiment are averaged over 5 trials. We remark that in the simulation we consider the following constraint

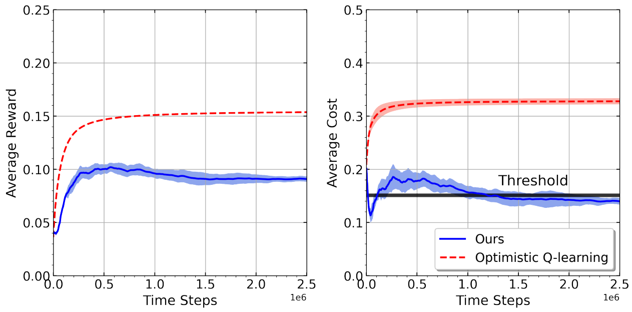


Figure 2: Average reward and cost of our algorithm and Optimistic Q-learning during training. The shaded region represents the standard deviations.

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

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

Figure 2 shows the performance comparison of our algorithm in terms of average reward and average cost during training compared with the algorithm in (Wei et al. 2020). We can see that our algorithm is able to learn a policy that achieves a high reward while satisfying the safety constraint very quickly. The optimistic Q-learning algorithm (Wei et al. 2020) was for unconstrained MDPs, so it yields a higher reward but also violates the safety constraint.

We further generalized Triple-QA for environments with continuous state and action spaces by incorporating with neural network function approximations. The details are included in the supplementary material.

Conclusion

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

The algorithm is also computationally efficient from algorithmic perspective because it is model-free, which means that it is potential to apply our method for complex and challenging CMDPs in practice. Simulation result also demonstrates the good performance of our algorithm.

Acknowledgments

The work of Honghao Wei and Lei Ying is supported in part by NSF under grants 2002608, 2001687, and 2112471.

References

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

Agarwal, M.; Bai, Q.; and Aggarwal, V. 2021. Markov Decision Processes with Long-Term Average Constraints. *arXiv preprint arXiv:2106.06680*.

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

Andrychowicz, O. M.; Baker, B.; Chociej, M.; Jozefowicz, R.; McGrew, B.; Pachocki, J.; Petron, A.; Plappert, M.; Powell, G.; Ray, A.; et al. 2020. Learning dexterous in-hand manipulation. *The Int. Journal of Robotics Research*, 39(1): 3–20.

Borkar, V. S. 2005. An actor-critic algorithm for constrained Markov decision processes. *Systems & control letters*, 54(3): 207–213.

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, 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*.

Ding, D.; Wei, X.; Yang, Z.; Wang, Z.; and Jovanovic, M. 2021. Provably Efficient Safe Exploration via Primal-Dual Policy Optimization. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, volume 130, 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, 8378–8390. Curran Associates, Inc.

Dong, K.; Wang, Y.; Chen, X.; and Wang, L. 2019. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*.

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.

Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. 2018. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

HasanzadeZonuzy, A.; Kalathil, D. M.; and Shakkottai, S. 2021. Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdps. In *AAAI Conf. Artificial Intelligence*, volume 35, 7667–7674.

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, 4863–4873.

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, 8030–8037.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Leike, J.; Martic, M.; Krakovna, V.; Ortega, P. A.; Everitt, T.; Lefrancq, A.; Orseau, L.; and Legg, S. 2017. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*.

Liu, T.; Zhou, R.; Kalathil, D.; Kumar, P.; and Tian, C. 2021a. 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. 2021b. An Efficient Pessimistic-Optimistic Algorithm for Stochastic Linear Bandits with General Constraints. In *Advances Neural Information Processing Systems (NeurIPS)*.

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.; Calvo-Fullana, M.; Chamon, L. F. O.; and Ribeiro, A. 2022. Safe Policies for Reinforcement Learning via Primal-Dual Methods. *IEEE Trans. Autom. Control*, 1–1.

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.

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, 15277–15287. Curran Associates, Inc.

Richter, F.; Orosco, R. K.; and Yip, M. C. 2019. Open-sourced reinforcement learning environments for surgical robotics. *arXiv preprint arXiv:1903.02090*.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359.

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

Stooke, A.; Achiam, J.; and Abbeel, P. 2020. Responsive safety in reinforcement learning by pid lagrangian methods. In *Int. Conf. Machine Learning (ICML)*, 9133–9143. PMLR.

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.

Wei, C.-Y.; Jafarnia Jahromi, M.; Luo, H.; and Jain, R. 2021. Learning Infinite-horizon Average-reward MDPs with Linear Function Approximation. In Banerjee, A.; and Fukumizu, K., eds., *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, volume 130, 3007–3015. PMLR.

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)*, 10170–10180. PMLR.

Wei, H.; Liu, X.; and Ying, L. 2021. A Provably-Efficient Model-Free Algorithm for Constrained Markov Decision Processes. *arXiv preprint arXiv:2106.01577*.

Wei, H.; Liu, X.; and Ying, L. 2022. Triple-Q: A Model-Free Algorithm for Constrained Reinforcement Learning with Sublinear Regret and Zero Constraint Violation. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*.

Xu, T.; Liang, Y.; and Lan, G. 2020. A Primal Approach to Constrained Policy Optimization: Global Optimality and Finite-Time Analysis. *arXiv preprint arXiv:2011.05869*.

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, 10639–10646.

Yang, T.-Y.; Rosca, J.; Narasimhan, K.; and Ramadge, P. J. 2020. Projection-based constrained policy optimization. In *Int. Conf. on Learning Representations (ICLR)*.

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

Supplementary Materials

Notations

We summarize notations used throughout the paper in Table 2.

Table 2: Notation Table

Notation	Definition
K	The total number of episodes.
J_r^π	The reward rate under policy π.
J_r^π	The utility rate under policy π.
$V^\pi(s)$	The cumulative discounted reward under policy π and initial state s.
$Q^\pi(s)$	The cumulative discounted reward under policy π and initial state action pair (s, a).
$W^\pi(s)$	The cumulative discounted utility under policy π and initial state s.
$C^\pi(s)$	The cumulative discounted utility under policy π and initial state action pair (s, a).
$v^\pi(s)$	The relative reward value function for state s.
$w^\pi(s)$	The relative utility value function for state s.
$sp(v^\pi)$	Span of relative reward value function: $sp(v^\pi) = \max_s v^\pi(s) - \min_s v^\pi(s)$.
$sp(w^\pi)$	Span of relative utility value function: $sp(w^\pi) = \max_s w^\pi(s) - \min_s w^\pi(s)$.
κ	$\max\{sp(v^{\epsilon,*}), sp(w^{\epsilon,*}), 1\}$

Definition 1 (Diameter)**.** *The diameter of an MDP $\mathcal{M}$ is defined as:*

$$D(\mathcal{M}) = \max_{s' \neq s} \min_{\pi} \mathbb{E}[\min\{t \geq 1 : S_t = s'\} | S_1 = s] - 1,$$

where the expectation is taken with respect to the Markov chain $(S_t)_{t=1}^{\infty}$ induced by the policy π and $\mathcal{M}$.

Auxiliary Lemmas

In this section we state several lemmas that will be used in our analysis. the first lemma establishes some key properties of the learning rates used in our algorithm.

Lemma 8. *Recall that the learning rate used in our algorithm is $\alpha_t = \frac{\chi+1}{\chi+t}$, and*

$$\alpha_t^0 = \prod_{j=1}^{t}(1-\alpha_j) \quad \textit{and} \quad \alpha_t^i = \alpha_i \prod_{j=i+1}^{t}(1-\alpha_j). \tag{41}$$

The following properties hold for α_t^i :

1. $\alpha_t^0 = 0$ *for* $t \geq 1, \alpha_t^0 = 1$ *for* $t = 0$.
2. $\sum_{i=1}^{t} \alpha_t^i = 1$ *for* $t \geq 1$, $\sum_{i=1}^{t} \alpha_t^i = 0$ *for* $t = 0$.
3. $\frac{1}{\sqrt{\chi+t}} \leq \sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{\chi+i}} \leq \frac{2}{\sqrt{\chi+t}}$.
4. $\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{\chi}$ *for every* $i \geq 1$.
5. $\sum_{i=1}^{t} (\alpha_t^i)^2 \leq \frac{\chi+1}{\chi+t}$ *for every* $t \geq 1$.

□

The proof of this lemma can be found in (Wei, Liu, and Ying 2021).

Lemma 9. *(Azuma's inequality) Let $X_1, X_2, \dots$ be a martingale difference with $|X_i| \leq c_i$ for all i. Then for any $0 < \delta < 1$,*

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i \geq \sqrt{2\bar{c}_N^2 ln \frac{1}{\delta}}\right) \leq \delta,$$

where $c_N^2 := \sum_{i=1}^{N} c_i^2$.

Lemma 10. *For any $k = 1, \ldots, K^\beta - 1$ in frame T and state-action pair (s, a), the following holds:*

$$\begin{aligned}Q_{k+1}(s,a) - Q^\pi(s,a) = \alpha_\tau^0(\hat{Q}_{(T-1)k^\beta+1}(s,a) - Q^\pi(s,a)) + \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \left[\hat{V}_{k_i}(s_{k_i+1}) - V^\pi(s_{k_i+1})\right] \\ + \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \left[V^\pi(s_{k_i+1}) - \mathbb{E}_{s' \sim p(\cdot|s,a)} V^\pi(s')\right] + \sum_{i=1}^{\tau} \alpha_\tau^i b_i,\end{aligned}$$

where $\tau = n_{k+1}(s,a)$, is the total number of visits to (s, a) for the first k timesteps.

Proof. By recursively using the updating rule for $Q_{k+1}(s, a)$, we have

$$Q_{k+1}(s,a) = \hat{Q}_1(s,a)\alpha_t^0 + \sum_{i=1}^{\tau} \alpha_\tau^i \left[r(s,a) + \gamma \hat{V}_{k_i}(s_{k_i} + 1)\right] + \sum_{i=1}^{\tau} \alpha_\tau^i b_i$$

According to the Bellman equation $Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} V^\pi(s')$ and the fact $\sum_{i=1}^{\tau} \alpha_\tau^i = 1$, we have

$$Q^\pi(s,a) = \alpha_\tau^0 Q^\pi(s,a) + \sum_{i=1}^{\tau} \alpha_\tau^i \left[r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} V^\pi(s')\right],$$

which finishes the proof. □

Lemma 11. *Consider any frame T. Let t=$N_{k,h}(s,a)$ be the number of visits to (s, a) before timestep k in the current frame and let $k_1, \ldots, k_t < k$ be the indices of these steps. Under any policy π, with probability at least $1 - \frac{1}{K^3}$, the following inequalities hold simultaneously for all $(s, a, k) \in \mathcal{S} \times \mathcal{A} \times [K]$*

$$\begin{aligned}\left|\sum_{i=1}^{\tau} \alpha_\tau^i \left[V^{\epsilon,*}(s_{k_i+1}) - \mathbb{E}_{s' \sim p(\cdot|s,a)} V^{\epsilon,*}(s')\right]\right| \leq & \kappa \sqrt{\frac{(\chi+1)\iota}{\chi+\tau}}, \\ \left|\sum_{i=1}^{\tau} \alpha_\tau^i \left[W^{\epsilon,*}(s_{k_i+1}) - \mathbb{E}_{s' \sim p(\cdot|s,a)} W^{\epsilon,*}(s')\right]\right| \leq & \kappa \sqrt{\frac{(\chi+1)\iota}{\chi+\tau}}.\end{aligned}$$

Proof. Note that

$$\sum_{i=1}^{\tau} \alpha_\tau^i \left[V^{\epsilon,*}(s_{k_i+1}) - \mathbb{E}_{s' \sim p(\cdot|s,a)} V^{\epsilon,*}(s')\right]$$

is a martingale, and each term in the summation belongs to $\left[-\alpha_\tau^i sp(V^{\epsilon,*}), \alpha_\tau^i sp(V^{\epsilon,*})\right]$ according to Lemma 3.

Define $\sigma = \sqrt{8 \log\left(\sqrt{2}K\right) \sum_{i=1}^{\tau} (\alpha_\tau^i sp(V^{\epsilon,*}))^2}$. By using Azuma's inequality (Lemma 9), we obtain that the following inequality holds

$$\left|\sum_{i=1}^{\tau} \alpha_\tau^i \left[V^{\epsilon,*}(s_{k_i+1}) - \mathbb{E}_{s' \sim p(\cdot|s,a)} V^{\epsilon,*}(s')\right]\right| \leq \sigma = sp(V^{\epsilon,*}) \sqrt{8 \sum_{i=1}^{\tau} (\alpha_\tau^i)^2 \log \sqrt{2}K} \leq \kappa \sqrt{\frac{(\chi+1)\iota}{\chi+\tau}}$$

with probability at least

$$1 - 2\exp\left(-\frac{\sigma^2}{2\sum_{i=1}^{\tau} (\alpha_\tau^i sp(V^{\epsilon,*}))^2}\right) \geq 1 - \frac{1}{K^3}.$$

□

Lemma 12. *Given $\delta \geq 2\epsilon, H \geq \frac{6\kappa}{\delta}$, under our algorithm, the conditional expected drift of L is*

$$\mathbb{E}\left[L_{T+1} - L_T | Z_T = z\right] \leq -\frac{\delta}{3} Z_T + \frac{3H}{K^2} + \eta + 2. \tag{42}$$

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

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

Then we have

$$\begin{aligned}
&\mathbb{E}\left[L_{K+1} - L_K | Z_T = z\right] \\
\leq & \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[Z_T \left(\rho + \epsilon - (1-\gamma)\hat{C}_k(s_k, a_k)\right) - \eta(1-\gamma)\hat{Q}_k(s_k,a_k) + \eta(1-\gamma)\hat{Q}_k(s_k,a_k) \middle| Z_T = z \right] + 2 \\
\leq_{(a)} & \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[Z_T \left(\rho + \epsilon - (1-\gamma)\sum_a \left\{ \hat{C}_k q^\pi \right\}(s_k, a) \right) - \eta(1-\gamma)\sum_a \{\hat{Q}_k q^\pi\}(s_k,a) \middle| Z_k = z \right] \\
& + E\left[\eta(1-\gamma)\hat{Q}_k(s_k,a_k) \middle| Z_T = z \right] + 2 \\
\leq & \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[Z_T \left(\rho + \epsilon - \sum_{s,a} \{gq^\pi\}(s,a) \right) - \eta(1-\gamma)\sum_a \{\hat{Q}_k q^\pi\}(s_k,a) + \eta(1-\gamma)\hat{Q}_k(s_k,a_k) \middle| Z_T = z \right] \\
& + \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[Z_T \left(\sum_{s,a}\{gq^\pi\}(s,a) - \sum_a (1-\gamma)\{C_k^\pi q^\pi\}(s_k,a) \right) \middle| Z_T = z \right] \\
& + \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[Z_T(1-\gamma)\sum_a \{C_k^\pi q^\pi\}(s_k,a) - Z_T(1-\gamma)\sum_a \left\{\hat{C}_k q^\pi\right\}(s_k,a) \, | Z_T = z \right] \\
& + \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[\eta(1-\gamma)\sum_a \{Q_k^\pi q^\pi\}(s_k,a) - \eta(1-\gamma)\sum_a \{Q_k^\pi q^\pi\}(s_k,a) \middle| Z_T = z \right] + 2 \\
\leq_{(b)} & -\frac{\delta}{2} z + (1-\gamma)sp(v^\pi) + \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[\eta(1-\gamma)\sum_a \left\{(F_k^\pi - \hat{F}_k)q_1^\pi\right\}(s_k,a) \ + \eta(1-\gamma)\hat{Q}_k(s_k,a_k) \middle| Z_T = z \right] + 2
\end{aligned}$$

Inequality (a) holds because of our algorithm. Inequality (b) holds because $\sum_a \{Q_t^\pi q^\pi\}(s_t,a)$ is non-negative, under Slater's condition, we can find policy π such that

$$\epsilon + \rho - \mathbb{E}\left[\sum_{s,a} g(s_k,a) q^\pi(s_k,a) \right] \leq -\delta + \epsilon \leq -\frac{\delta}{2},$$

and according to Lemma 3

$$\begin{aligned}
&\sum_{s,a}\{gq^\pi\}(s,a) - \sum_a (1-\gamma)\{C_k^\pi q^\pi\}(s_k,a) \\
=& J^\pi - (1-\gamma)V^\pi(s_k) \\
\leq& (1-\gamma)sp(v^\pi).
\end{aligned}$$

Note that when K is sufficiently large, $(1-\gamma)sp(v^{\epsilon,*}) \leq \frac{\kappa}{H} \leq \frac{\delta}{6}$. By applying $\pi = \epsilon, *$ we obtain

$$\begin{aligned}
&\mathbb{E}\left[L_{K+1} - L_K | Z_T = z\right] \\
\leq & -\frac{\delta}{2}z + \frac{\delta}{6} + \frac{1}{K^\beta} \sum_{k=(T-1)K^\beta+1}^{TK^\beta} \mathbb{E}\left[\eta(1-\gamma)\sum_a \left\{(F_k^\pi - \hat{F}_k)q_1^\pi\right\}(s_k,a) \ + \eta(1-\gamma)\hat{Q}_k(s_k,a_k) \middle| Z_T = z \right] + 2 \\
\leq & -\frac{\delta}{3}z + \frac{3H}{K^2} + \eta + 2,
\end{aligned}$$

where the last inequality holds due to (i) the overestimation established in Lemma 4 and (ii) $\hat{Q}(\cdot,\cdot)$ is bounded by $\frac{1}{1-\gamma}$. We remark that that the overestimation result and the concentration result in frame T hold regardless of the value of Z_T. □

Detailed Proof

We next present the complete proof of our main theorem.

Recall that the regret can be decomposed as:

$$\begin{aligned}\text{Regret} =& \sum_{k=1}^{K}(J_r^* - r(s_k,a_k)) = \sum_{k=1}^{K}(J_r^* - J_r^{\epsilon,*}) + \sum_{k=1}^{K}(J_r^{\epsilon,*} - r(s_k,a_k))\\ =& \sum_{k=1}^{K}(J_r^* - J_r^{\epsilon,*})\\ &+ \sum_{k=1}^{K}(J_r^{\epsilon,*} - (1-\gamma)V^{\epsilon,*}(s_k))\\ &+ \sum_{k=1}^{K}((1-\gamma)V^{\epsilon,*}(s_k) - (1-\gamma)\hat{Q}_k(s_k,a_k))\\ &+ \sum_{k=1}^{K}\Big((1-\gamma)\hat{Q}_k(s_k,a_k) - r(s_k,a_k)\Big). \end{aligned} \tag{43}$$

For the constraint violation, according to the dual variable updating rule, we have

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

which implies

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

Then by summing the above equation over all the frames, we obtain

$$\text{Violation} = \mathbb{E}\left[\sum_{t=1}^{T}\rho - g(s_k,a_k)\right]_+ \leq -K\epsilon + K^\beta\mathbb{E}\left[Z_{K^{1-\beta}+1}\right] + \mathbb{E}\left[\sum_{k=1}^{K}(1-\gamma)\hat{C}_k(s_k,a_k) - g(s_k,a_k)\right]. \tag{44}$$

Proof of Lemma 2.

Lemma (Restatement of Lemma 2)**.** *Given $\epsilon \leq \delta$, we have*

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

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

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

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

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

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

$$\sum_{x,a}q^{\xi_2}(s,a)g(s,a) = \sum_{s,a}\left((1-\frac{\epsilon}{\delta})q^*(s,a) + \frac{\epsilon}{\delta}q^{\xi_1}(s,a)\right)g(s,a) \geq \rho + \epsilon,$$

$$\sum_{s,a}q^{\xi_2}(s,a) = \sum_{x',a'}p(s|s',a')q^{\xi_2}(s',a'),$$

$$\sum_{s,a} q^{\xi_2}(s,a) = 1.$$

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

$$\begin{aligned}
&\sum_{s,a} \left(q^*(x,a) - q^{\epsilon,*}(s,a)\right) r(s,a) \\
\leq &\sum_{s,a} \left(q^*(s,a) - q^{\xi_2}(s,a)\right) r(s,a) \\
\leq &\sum_{s,a} \left(q^*(s,a) - \left(1-\frac{\epsilon}{\delta}\right) q^*(s,a) - \frac{\epsilon}{\delta} q^{\xi_1}(s,a)\right) r(s,a) \\
\leq &\sum_{s,a} \left(q^*(s,a) - \left(1-\frac{\epsilon}{\delta}\right) q^*(s,a)\right) r(s,a) \\
\leq &\frac{\epsilon}{\delta} \sum_{s,a} q^*(s,a) r(s,a) \\
\leq &\frac{\epsilon}{\delta},
\end{aligned}$$

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

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

□

Porrf of Lemma 3

Lemma (Restatement of Lemma 3)**.** *For an arbitrary policy π obtained from the LP* (8)*, we have*

$$\begin{aligned}
J_r^\pi - (1-\gamma)V^\pi(s) \leq (1-\gamma)sp(v^\pi), \quad |V^\pi(s_1) - V^\pi(s_2)| \leq 2sp(v^\pi); \\
J_g^\pi - (1-\gamma)W^\pi(s) \leq (1-\gamma)sp(w^\pi), \quad |W^\pi(s_1) - W^\pi(s_2)| \leq 2sp(w^\pi),
\end{aligned}$$

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

Proof. We only prove the result for the reward value functions. The proof for the utility function is almost identical. Let π be an arbitrary policy. The proof follows Lemma 2 in (Wei et al. 2020) closely. According to the Bellman equation, we have

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

Then

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

and

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

Therefore we can conclude that

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

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

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

□

Proof of Lemma 4

Lemma (Restatement of Lemma 4)**.** *With probability at least* $1 - \frac{1}{K^3}$*, the following inequality holds simultaneously for all* $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times [K]$:

$$\left\{\hat{F}_k - F^{\epsilon,*}\right\}(s,a) \geq 0, \tag{45}$$

which further implies that

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\epsilon,*} - \hat{F}_k\right)q^\pi\right\}(s_k,a)\right] \leq \frac{3H}{\eta K}. \tag{46}$$

Proof. The proof follows Lemma 3 in (Wei, Liu, and Ying 2021) but for the discounted case. Consider frame T and episodes in frame T. Define $Z = Z_{(T-1)K^\beta+1}$ because the value of the virtual queue does not change during each frame. We further define/recall the following notations:

$$\begin{aligned} F_k(s,a) &= Q_k(s,a) + \frac{Z}{\eta}C_k(s,a), \quad U_k(s) = V_k(s) + \frac{Z}{\eta}W_k(s),\\ \hat{F}_k(s,a) &= \hat{Q}_k(s,a) + \frac{Z}{\eta}\hat{C}_k(s,a), \quad \hat{U}_k(s) = \hat{V}_k(s) + \frac{Z}{\eta}\hat{W}_k(s),\\ F^\pi(s,a) &= Q^\pi(s,a) + \frac{Z}{\eta}C^\pi(s,a), \quad U^\pi(s) = V^\pi(s) + \frac{Z}{\eta}W^\pi(s). \end{aligned}$$

In the following, we use π denote the policy $\epsilon, *$ without obscurity. Then following a similar proof of Lemma 10, we have

$$\begin{aligned} &\{F_{k+1} - F^\pi\}(s,a)\\ =&\alpha_t^0\left\{F_{(T-1)K^\beta+1} - F^\pi\right\}(s,a)\\ &+\sum_{i=1}^{t}\alpha_t^i\left(\left\{\hat{U}_{k_i} - U^\pi\right\}(s_{k_i+1}) + \gamma\left(U^\pi(s_{k_i+1}) - \mathbb{E}_{s'\sim p(\cdot|s,a)}U^\pi(s')\right) + \left(1+\frac{Z}{\eta}\right)b_i\right)\\ \geq_{(a)}&\alpha_t^0\left\{\hat{F}_{(T-1)K^\beta+1} - F^\pi\right\}(s,a) + \sum_{i=1}^{t}\alpha_t^i\left\{\hat{U}_{k_i} - U^\pi\right\}(s_{k_i+1})\\ =_{(b)}&\alpha_t^0\left\{\hat{F}_{(T-1)K^\beta+1} - F^\pi\right\}(s,a) + \sum_{i=1}^{t}\alpha_t^i\left(\max_a \hat{F}_{k_i}(s_{k_i+1},a) - F^\pi(s_{k_i},\pi(s_{k_i}))\right)\\ \geq&\alpha_t^0\left\{\hat{F}_{(T-1)K^\beta+1} - F^\pi\right\}(s,a) + \sum_{i=1}^{t}\alpha_t^i\left\{\hat{F}_{k_i} - F^\pi\right\}(s_{k_i+1},\pi(s_{k_i+1})), \end{aligned} \tag{47}$$

where inequality (a) holds because of the concentration result in Lemma 11 and

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

where the last equality comes from the properties of the learning rate (Lemma 8). Equality (b) holds because our algorithm selects the action that maximizes $\hat{F}_{k_i}(s_{k_i+1},a)$ so $\hat{U}_{k_i}(s_{k_i+1}) = \max_a \hat{F}_{k_i}(s_{k_i+1},a)$. The inequality above suggests that we can prove $\{F_{k+1} - F^\pi\}(s,a)$ for any (s,a) if (i)

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

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

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

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

$$\begin{aligned}\hat{F}_{k+1}(s,a) =& \hat{Q}_{k+1}(s,a) + \frac{Z}{\eta}\hat{C}_{k+1}(s,a)\\ =& \min\{\hat{Q}_k(s,a) + \frac{Z}{\eta}\hat{C}_k(s,a), Q_{k+1}(s,a) + \frac{Z}{\eta}C_{k+1}(s,a)\}\\ =& \min\{\hat{F}_k(s,a), F_{k+1}(s,a)\}\end{aligned}$$

we have

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

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

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\pi} - \hat{F}_k\right)q^{\pi}\right\}(s_k,a)\right]\\ =&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\pi} - \hat{F}_k\right)q^{\pi}\right\}(s_k,a)\middle|\mathcal{E}\right]\Pr(\mathcal{E})\\ &+\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\left\{\left(F^{\pi} - \hat{F}_k\right)q^{\pi}\right\}(s_k,a)\middle|\mathcal{E}^c\right]\Pr(\mathcal{E}^c)\\ \leq_{(a)}&K\left(1+\frac{2K^{1-\beta}}{\eta}\right)H\frac{1}{K^3} \leq \frac{3H}{\eta K},\end{aligned} \tag{48}$$

where inequality (a) holds because at any timestep k, we have $(F^{\pi} - \hat{F}_k) \leq \left(1+\frac{2K^{1-\beta}}{\eta}\right)H.$ □

Porrf of Lemma 5

Lemma (Restatement of Lemma 5). *Assuming $\epsilon < \delta$, we have*

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}(1-\gamma)\left(\sum_{a}\left\{\hat{Q}_k q^{\epsilon,*}\right\}(s_k,a) - \hat{Q}_k(s_k,a_k) + \frac{Z_k}{\eta}\sum_{a}\left\{\hat{C}_k q^{\epsilon,*} - C^{\epsilon,*}q^{\epsilon,*}\right\}(s_k,a)\right)\right]\\ \leq&\frac{2K}{\eta} + \sum_{T=1}^{K^{1-\beta}}\mathbb{E}\left[Z_T\right]\frac{(1-\gamma)\kappa}{\eta}.\end{aligned}$$

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

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

Then adding and subtracting additional terms,

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

Specifically, for the term inside the summation, we have

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

where the first inequality holds because a_k is chosen to maximize $\hat{Q}_k(s_k,a)+\frac{Z_k}{\eta}\hat{C}_k(s_k,a)$, and the last inequality is true because $q^\epsilon(s,a)$ is a feasible solution to the optimization problem (9) such that

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

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

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

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

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

where the first inequality comes from Lemma 3, and the last inequality comes from the fact that $\kappa = \max_{0\leq\epsilon\leq\rho/2}\left(\max\{sp(v^{\epsilon,*}),sp(w^{\epsilon,*}),1\}\right)$ is non-negative. □

Proof of Lemma:6

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

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

Proof. The proof will also use the following lemma from (Liu et al. 2021b).

Lemma 13. *Let S_t be the state of a Markov chain, L_t be a Lyapunov function with $L_0=l_0$, and its drift $\Delta_t=L_{t+1}-L_t$. Given the constant γ and v with $0<\gamma\leq v$, suppose that the expected drift $\mathbb{E}[\Delta_t|S_t=s]$ satisfies the following conditions:*

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

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

Then we have

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

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

We apply Lemma 13 to a new Lyapunov function:

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

To verify condition (1) in Lemma 13, consider $\bar{L}_T = Z_T \geq \theta_T = \frac{6(\eta+2+\frac{3H}{K^2})}{\delta}$ and $2\epsilon \leq \delta$. The conditional expected drift of $\bar{L}_T$ is

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

where inequality (a) is obtained according to Lemma 12; and the last inequality holds given $z \geq \theta_T$.

To verify condition (2) in Lemma 13, we have

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

Now choose $\gamma = \frac{\delta}{6}$ and $v = 2$. From Lemma 13, we obtain

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

By Jensen's inequality, we have

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

which implies that

$$\begin{aligned}
\mathbb{E}\left[Z_T\right] \leq&\frac{1}{r}\log\left(1 + \frac{2e^{r(v+\theta_T)}}{r\lambda}\right) \\
\leq&\frac{1}{r}\log\left(\frac{11v^2}{3\lambda^2}e^{r(v+\theta_T)}\right) \\
\leq&\frac{3v^2}{\lambda}\log\left(\frac{2v}{\lambda}\right) + v + \theta_T \\
\leq&\frac{72}{\delta}\log\left(\frac{24}{\delta}\right) + 2 + \frac{6(\eta+3)}{\delta} \\
\leq&\frac{92}{\delta}\log\left(\frac{24}{\delta}\right) + \frac{6\eta}{\delta}
\end{aligned}$$

□

Proof of Lemma 7

Lemma (Restatement of Lemma 7)). *For any* $T \in [K^{1-\beta}]$,

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

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

where m is a positive integer.

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

$$\begin{aligned}
&\sum_{k=1}^{K'}\Big((1-\gamma)\hat{Q}_k(s_k,a_k)-r(s_k,a_k)\Big)\\
=&\gamma\sum_{k=1}^{K'}\sum_{i=1}^{n_k}\alpha_{n_k}^i\left[-\gamma r(s_k,a_k)+(1-\gamma)\hat{V}_{k_i(s_k,a_k)}(s_{k_i(s_k,a_k)+1})\right]+2(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}},
\end{aligned} \tag{49}$$

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

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

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

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

Substituting in (49), we have

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

$$
\begin{aligned}
&\le_{(a)} -\gamma\sum_{k=1}^{K'} r(s_k,a_k)+\gamma(1-\gamma)\sum_{k=1}^{K'}\hat{V}_k(s_{k+1})+\frac{K'}{\chi}(1-\gamma)\gamma H+2(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}}\\
&=-\gamma\sum_{k=1}^{K'} r(s_k,a_k)+\gamma(1-\gamma)\sum_{k=1}^{K'}\left(\hat{V}_k(s_{k+1})-\hat{V}_{k+1}(s_{k+1})\right)+\gamma(1-\gamma)\sum_{k=1}^{K'}\hat{V}_{k+1}(s_{k+1})\\
&\quad+\frac{K'}{\chi}(1-\gamma)\gamma H+2(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}}\\
&\le_{(b)} -\gamma\sum_{k=1}^{K'} r(s_k,a_k)+\gamma(1-\gamma)\sum_{k=2}^{K'+1}\hat{V}_k(s_k)+\frac{K'}{\chi}(1-\gamma)\gamma H+2(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}}+\gamma(1-\gamma)SH\\
&\le\gamma\sum_{k=1}^{K'}\left((1-\gamma)\hat{Q}_k(s_k,a_k)-r(s_k,a_k)\right)+\frac{K'}{\chi}(1-\gamma)\gamma H+2(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}}+2\gamma(1-\gamma)SH\\
&\le\gamma^m K'(1-\gamma)H+\frac{K'm}{\chi}(1-\gamma)\gamma H+2m(1-\gamma)\kappa\sum_{k=1}^{K'}\sqrt{\frac{(\chi+1)\iota}{\chi+n_k}}+2m\gamma(1-\gamma)SH\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(repeatedly use the inequality } m \text{ times)}\\
&\le_{(c)}\gamma^m K'(1-\gamma)H+\frac{K'm}{\chi}(1-\gamma)\gamma H+4\gamma(1-\gamma)m\kappa\sqrt{(\chi+1)SAT'\iota}+2m\gamma(1-\gamma)SH,
\end{aligned}
$$

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

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

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

The Choices of the Hyper-parameters

The regret bound (43) and constraint violation bound (40) are

$$
\text{Regret}(K)=\tilde{\mathcal{O}}\left(K\epsilon+\gamma^m K+\frac{Km}{\chi}+\sqrt{K^{2-\beta}\chi}+mK^{1-\beta}+\frac{K}{\eta}+\frac{K}{H}\right) \tag{50}
$$

$$
\text{Violation}(K)=-K\epsilon+\tilde{\mathcal{O}}\left(K^{\beta}\eta+\frac{Km}{\chi}+\sqrt{K^{2-\beta}\chi}+mK^{1-\beta}\right). \tag{51}
$$

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

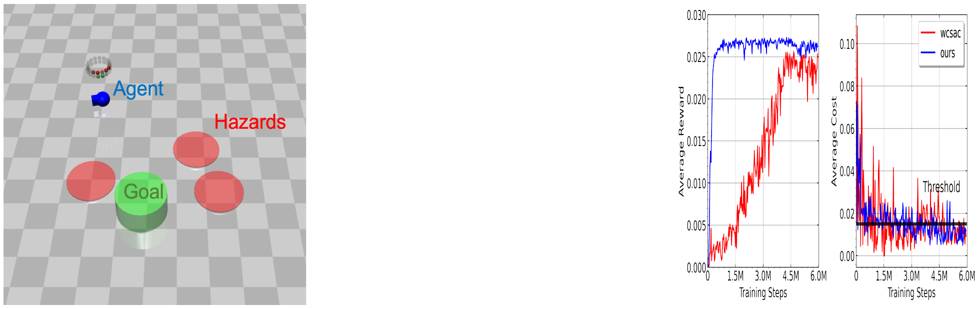
$$
\text{Regret}(K)=\tilde{\mathcal{O}}\left(K\epsilon+K^{1-\zeta+\nu}+K^{1-3\zeta+3\nu}+\frac{K}{\eta}+K^{1-\nu}\right) \tag{52}
$$

$$
\text{Violation}(K)=-K\epsilon+\tilde{\mathcal{O}}\left(K^{3\zeta-2\nu}\eta+K^{1-\zeta+\nu}+K^{1-3\zeta+3\nu}\right), \tag{53}
$$

where the term $\gamma^m K$ is omitted because the choice of m ensures $\gamma^m \le \frac{1}{K}$ (Eq.(38)). To make sure $1 > 1-\zeta+\nu > 0$, we need to have $\nu < \zeta$. Then the bounds become

$$
\text{Regret}(K)=\tilde{\mathcal{O}}\left(K\epsilon+K^{1-\zeta+\nu}+\frac{K}{\eta}+K^{1-\nu}\right) \tag{54}
$$

$$
\text{Violation}(K)=-K\epsilon+\tilde{\mathcal{O}}\left(K^{3\zeta-2\nu}\eta+K^{1-\zeta+\nu}\right). \tag{55}
$$



(a) DynamicEnv with Safety Constraints

(b) The average reward and cost during training

Figure 3: Simulation Results

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

Additional Simulations

We also evaluated our algorithm Triple-QA with neural network approximations on the Dynamic Gym benchmark (DynamicEnv) (Yang et al. 2021) with continuous state and action spaces as shown in Figure 3a. In this environment, the objective for the agent is to reach the goal position on a 2D map while trying to avoid randomly generated hazardous areas. The cost for hazardous positions is 1 and is 0 for other locations. We build an algorithm which incorporated the key ideas of Triple-QA on top of SAC (Haarnoja et al. 2018). In particular, two Q functions are trained simultaneously, one for reward Q function and the other is for cost Q function. The dual variable is updated at a slow timescale (every few episodes). The actor network is trained by optimizing the combined networks (Eq. (13)). The simulation results in Figure. 3b show that Deep Triple-QA learns a safe-policy with a high reward much faster than WCSAC (Yang et al. 2021).

We remark that when implementing our algorithm in practice, we do not need to reset all the $Q_k(s,a)$ and $C_k(s,a)$ to H. Instead, we added extra "bonuses" to the learned values at the beginning of each frame to ensure overestimation like in (Wei, Liu, and Ying 2021).

Algorithm 2: Replacing Lines 27-28 in Triple-QA

```
**if** $k$ mod $K^\beta = 0$ **then**
    $Z \leftarrow \left(Z + \rho + \epsilon - \frac{\bar{C}}{K^\beta}\right)$
    Reset $\bar{C} \leftarrow 0, n_t(s,a) \leftarrow 0.$
    $\hat{Q}_{k+1}(s,a) \leftarrow \hat{Q}_{k+1}(s,a) + \frac{4H}{\eta}, \forall (s,a)$
    $Q_{k+1}(s,a) \leftarrow Q_{k+1}(s,a) + \frac{4H}{\eta}, \forall (s,a)$
    **if** $\hat{Q}_{k+1}(s,a) > H$ or $\hat{C}_{k+1}(s,a) > H$ **then**
        Reset $\hat{Q}_{k+1}(s,a), Q_{k+1}(s,a), V_{k+1}(s)$ to $H$
        Reset $\hat{C}_{k+1}(s,a), C_{k+1}(s,a), W_{k+1}(s)$ to $H$
```

Experimental settings

We run all each experiment independently with five seeds.We used single NVIDIA GeForce RTX 2080 Super with AMD Ryzen 7 3700 8- Core Processor.