Towards Achieving Sub-linear Regret and Hard Constraint Violation in Model-free RL

Arnob Ghosh
New Jersey Institute of Technology

Xingyu Zhou
Wayne State University

Ness Shroff
The Ohio State University

Abstract

We study the constrained Markov decision processes (CMDPs), in which an agent aims to maximize the expected cumulative reward subject to a constraint on the expected total value of a utility function. Existing approaches have primarily focused on *soft* constraint violation, which allows compensation across episodes, making it easier to satisfy the constraints. In contrast, we consider a stronger *hard* constraint violation metric, where only positive constraint violations are accumulated. Our main result is the development of the *first model-free*, *simulator-free* algorithm that achieves a sub-linear regret and a sub-linear hard constraint violation simultaneously, even in *large-scale* systems. In particular, we show that $\tilde{\mathcal{O}}(\sqrt{d^3H^4K})$ regret and $\tilde{\mathcal{O}}(\sqrt{d^3H^4K})$ hard constraint violation bounds can be achieved, where K is the number of episodes, d is the dimension of the feature mapping, H is the length of the episode. Our results are achieved via novel adaptations of the primal-dual LSVI-UCB algorithm, i.e., it searches for the dual variable that balances between regret and constraint violation *within* every episode, rather than updating it at the end of each episode. This turns out to be crucial for our theoretical guarantees when dealing with hard constraint violations.

1 Introduction

In many practical applications of online reinforcement learning (RL) (e.g., safety, resource constraints), there exist additional constraints on the learned policy in the sense that it also needs to ensure that the expected total utility (cost, resp.) exceeds a given threshold (is below a threshold, resp.). Such problems are formulated as constrained Markov Decision Processes (CMDPs) (Altman, 1999; Efroni et al., 2020) where the agent gets a reward (r) and utility (g) depending on the state and action. In an episodic CMDP, starting from initial state x_1, the goal is to

$$\text{maximize}_\pi V_{r,1}^\pi(x_1) \quad \text{subject to } V_{g,1}^\pi(x_1) \geq b,$$

where $V_{r,1}^\pi(x_1)$ is the cumulative reward value function (defined in (3)) and $V_{g,1}^\pi(x_1)$ is the cumulative utility value function respectively when the agent follows the policy π.

To develop provably efficient model-free algorithms for CMDPs, most of the prior works (Wei et al., 2021b; Ghosh et al., 2022; Liu et al., 2021a; Ding et al., 2020, 2021) seek to minimize the following metrics

$$\begin{aligned}\text{Regret } (K) &= \sum_{k=1}^{K}(V_{r,1}^*(x_1) - V_{r,1}^{\pi_k}(x_1)) \\ \text{Violation } (K) &= \sum_{k=1}^{K}(b - V_{g,1}^{\pi_k}(x_1)),\end{aligned} \tag{1}$$

where $V_{r,1}^*(x_1)$ is the optimal reward value function. Prior works (Ghosh et al., 2022; Liu et al., 2021a) have shown that $\tilde{\mathcal{O}}(\sqrt{T})$ regret and zero constraint violation are achievable with high-probability. An astute reader may note that in the violation metric defined in (1), a large violation ($V_{g,1}^{\pi_k}(x_1) < b$) at an episode can be offset by a strictly feasible policy ($V_{g,1}^{\pi_k}(x_1) > b$) at another episode. In particular, consider the sequence of policies $\{\pi_k\}_{k=1}^K$, $V_{g,1}^{\pi_k}(x_1) = b+1$ at odd episode k, and $V_{g,1}^{\pi_k}(x_1) = b-1$ at even episode. Then $\sum_k(b - V_{g,1}^{\pi_k}(x_1)) \leq 0$, even though such a sequence of policies violates the constraints in half of the episodes. Thus, the number of episodes where employed policies are not close to satisfying the constraint can grow sub-linearly even when they achieve zero violation by (1).

Clearly, the above is not a desired setting. Thus, in this paper, we seek to minimize the regret along with

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

the following *hard* constraint violation:

$$\text{Violation}_H(K) \coloneqq \sum_{k=1}^{K} \left(b - V_{g,1}^{\pi_k}(x_1)\right)_+ . \qquad (2)$$

Subsequently, we denote the violation metric in (1) as *soft* violation. The example we provided in the previous paragraph shows that even zero soft violation may lead to hard constraint violation that grows linearly with the number of episodes. However, if the hard constraint violation grows sub-linearly, then the number of episodes where the policy violates the constraint by any *fixed* non-zero amount which is independent of K can only grow sub-linearly with the number of episodes.

Efroni et al. (2020) seeks to minimize such hard constraints defined in (2). However, they consider the tabular set-up and proposed algorithms which are linear programming (LP)-based and model-based. The regret and violation scale polynomially with the number of states in the paper. Thus, those results *would not be useful* for large-scale RL applications, where the number of states could be infinite. To address this curse of dimensionality, modern RL has adopted *function approximation* techniques to approximate the (action-)value function of a policy, which greatly expands the potential reach of RL, especially via deep neural networks. Model-based and LP-based algorithms are computationally difficult to extend to large-scale systems (Chen et al., 2021). Motivated by this, we aim to address the following open question:

Can we design a model-free algorithm with sub-linear regret and sub-linear hard constraint violation for CMDPs with function approximation?

Contribution: To answer the above question, we consider CMDPs with linear function approximation, where the transition dynamics and the reward function can be represented as a linear function of some known feature mapping. Our main contributions are as follows.

- We show that our proposed algorithm achieves $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ regret and $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ (hard) constraint violation bounds with a high probability, where d is the dimension of the feature mapping, H is the length of the episode, and T is the total number of steps.
- Our bounds are attained without explicitly estimating the unknown transition model or *requiring a simulator*, and they depend on the state space only through the dimension of the feature mapping. *To the best of knowledge*, these sub-linear bounds for regret and hard constraints are the first such results for model-free online RL algorithms for CMDPs with function approximations. Since linear CMDP contains tabular setup, as a by-product, our result also provides the *first sub-linear regret* and *sub-linear hard constraint bounds* even for the tabular setup under the model-free setup or using primal-dual approach.
- *This is the first result that shows that $\tilde{\mathcal{O}}(\sqrt{T})$ regret and hard violation can be achieved using primal-dual-based approach* since all the existing primal-dual approaches give soft constraint violation bound. Our main results are achieved by a novel approach of tuning the dual variable within each episode rather than updating at the end of the episodes (as done in existing approaches). In particular, we tune the dual variable within each episode to achieve a policy such that the estimated utility value function would exceed b by a small amount indicating the perfect trade-off between reward and utility maximization. This turns out to be the key to achieving hard-constraint violation bound.

1.1 Related Work

Model-based RL algorithms have been proposed for the CMDP (Efroni et al., 2020; Singh et al., 2020; Brantley et al., 2020; Zheng and Ratliff, 2020; Kalagarla et al., 2020; Liu et al., 2021a; Ding et al., 2021). Apart from Efroni et al. (2020) and Liu et al. (2021a) (OptPess-LP), none of the other papers considered the hard constraint bound in (2). Both the papers assumed finite state-space. Naturally, the proposed algorithms there achieved regret bound which scales polynomially with the cardinality of the state-space. Hence, such results cannot cope with the large state space observed in many MDP problems. Moreover, both Efroni et al. (2020) and Liu et al. (2021a) consider LP-based approaches. Further, Liu et al. (2021a) assumes the knowledge of a strictly feasible policy in order to bound which we do not assume. In Section 3, we detail the limitations of existing LP-based approaches for *unconstrained* linear MDP (Neu and Pike-Burke, 2020; Neu and Okolo, 2023; Lakshminarayanan et al., 2017; Bas-Serrano et al., 2021) and the advantages of our approaches compared to a potential extension of those LP-based approaches to linear CMDP. In the bandit setup, Chen et al. (2022); Pacchiano et al. (2021) consider hard violation for linear bandit setup. Note that the bandit setting can be viewed as a degenerate single-state RL, and bandit settings do not have a state transition kernel associated with them. Hence, the approaches for the linear bandit setup cannot be extended to the linear MDP setup. Furthermore, the approaches in the above paper are not primal-dual based. Recently, Guo et al. (2022) proposed a primal-dual algorithm that obtains sublinear regret and hard constraint violation bound in the bandit setup. However, such an approach cannot be extended to the episodic RL setup (Appendix I).

Ding et al. (2020); Xu et al. (2021) proposed policy-gradient based model-free approaches. However, they require 'simulator' or generative model Azar et al. (2012). Recently, model-free RL algorithms without simulators have also been proposed (Wei et al., 2021b; Ghosh et al., 2022) to solve CMDP. Only Ghosh et al. (2022) considered the large state-space scenario in the linear CMDP setting. However, all the aforementioned works consider soft constraint violation (cf.(1)) rather the hard constraint violation (cf.(2)). Since the focus is different our algorithm and analysis are significantly different. Please see Section 4 for more details. We do not assume Slater's condition (i.e., a stirctly feasible policy exists), unlike all the existing approaches. Thus, our analysis does not rely on strong duality. Amani et al. (2021) proposed a RL algorithm for the scenario where a constraint needs to be satisfied at each step of an episode. We consider a constraint where the cumulative utility over an episode must exceed a threshold. Hence, the set of constraints is fundamentally different. Further, unlike in Amani et al. (2021), we do not assume that a safe policy is known.

2 Problem Formulation

We consider an episodic constrained MDP, denoted by $(\mathcal{S}, \mathcal{A}, \mathbb{P}, H, r, g)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the fixed length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is a collection of transition probability measures, $r = \{r_h\}_{h=1}^H$ is a collection of reward functions, and $g = \{g_h\}_{h=1}^H$ is a collection of utility functions. We assume that $\mathcal{S}$ is a measurable space with possibly infinite number of elements, $\mathcal{A}$ is a finite action set. $\mathbb{P}_h(\cdot|x,a)$ is the transition probability kernel which denotes the probability to reach a state when action a is taken at state x. $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$, and $g_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ and are assumed to be deterministic. However, one can readily extend to settings when r_h and g_h are random.

Each episode $k \in [K]$ starts with a fixed state x_1. It can be readily generalized to the setting where x_1 is drawn from a distribution. At each step $h \in [H]$ in episode k, the agent observes state $x_h^k \in \mathcal{S}$, picks an action $a_h^k \in \mathcal{A}$, receives a reward $r_h(x_h^k, a_h^k)$, and a utility $g_h(x_h^k, a_h^k)$. The MDP evolves to x_{h+1}^k drawn from $\mathbb{P}_h(\cdot|x_h^k, a_h^k)$. The episode terminates at step $H+1$. Without loss of generality, we assume that $r_{H+1} = g_{H+1} = 0$. In this paper, we consider the *challenging* scenario where the agent only observes the bandit information $r_h(x_h^k, a_h^k)$ and $g_h(x_h^k, a_h^k)$ at the visited state-action pair (x_h^k, a_h^k). The policy-space of an agent is $\Delta(\mathcal{A}|\mathcal{S}, H)$; $\{\{\pi_h(\cdot|\cdot)\}_{h=1}^H : \pi_h(\cdot|x) \in \Delta(\mathcal{A}), \forall x \in \mathcal{S}, h \in [H]\}$. Here $\Delta(\mathcal{A})$ is the probability simplex over the action space. For any $x_h^k \in \mathcal{S}$, $k \in [K]$, and $h \in [H]$, $\pi_{h,k}(a_h^k|x_h^k)$ denotes the probability that action $a_h^k \in \mathcal{A}$ is taken at episode k at the state x_h^k.

Let $V_{r,h}^\pi(x)$ denote the expected value of the total reward function starting from step h and state x when the agent selects action using the policy $\pi = \{\pi_h\}_{h=1}^H$

$$V_{r,h}^\pi(x) = \mathbb{E}_\pi \left[\sum_{i=h}^H r_i(x_i, a_i) | x_h = x \right], \tag{3}$$

where $\mathbb{E}$ is taken with respect to the policy π and the transition probability kernel $\mathbb{P}$. Let $Q_{r,h}^\pi(x,a)$ denote the expected value of the total reward starting from step h and the state-action pair (x,a) and follows the policy π as $Q_{r,h}^\pi(x,a) = \mathbb{E}_\pi \left[\sum_{i=h}^H r_i(x_i, a_i) | x_h = x, a_h = a \right]$.

Similarly, we define the value function for the utility $V_{g,h}^\pi(x)$, and the action-value function for the utility $Q_{g,h}^\pi(x,a)$. We denote $V_{j,h}^\pi(x)$, and $Q_{j,h}^\pi(x,a)$ for $j = r, g$. We observe $V_{j,h}^\pi(x) = \langle \pi_h(\cdot|x), Q_{j,h}^\pi(x,\cdot) \rangle_{\mathcal{A}}$, where $\langle \pi_h(\cdot|x), Q_{j,h}^\pi(x,\cdot) \rangle_{\mathcal{A}} = \sum_{a \in \mathcal{A}} \pi_h(a|x) Q_{j,h}^\pi(x,a)$.

Definition 1. *For brevity, we denote* $\mathbb{P}_h V_{j,h+1}^\pi(x,a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V_{j,h+1}^\pi(x')$ *for* $j = r, g$.

Using this notation, Bellman's equation associated with the policy π becomes

$$Q_{j,h}^\pi(x,a) = (r_h + \mathbb{P}_h V_{j,h+1}^\pi)(x,a). \tag{4}$$

The objective of the learning agent is to find an optimal solution to the following problem

$$\text{maximize}_\pi V_{r,1}^\pi(x_1), \quad \text{subject to } V_{g,1}^\pi(x_1) \geq b. \tag{5}$$

Note that even though we have only one constraint, it can be readily generalized to the scenario with multiple constraints. Further, constraints like $V_{g,1}^\pi(x_1) \leq b$ can also be accommodated. In order to avoid trivial solutions, we consider $b \in (0, H]$. We denote the optimal policy as π^* which solves the above optimization problem. Since π^* is obtained by having complete information, it is denoted as *the best policy in hindsight*. The CMDP setup is standard (Efroni et al., 2020).

Without any constraint information a priori, an agent cannot know the policies that satisfy the constraint. Instead, we allow the policy to violate the constraint and minimize the regret while minimizing the total constraint violations over the K episodes. We now define the performance metric that we seek to minimize.

Performance Metric: Let the policy employed by the agent at episode k be $\pi_k = [\pi_{1,k}, \ldots, \pi_{h,k}, \ldots, \pi_{H,k}]^T$. The performance metric we consider is the following

$$\text{Regret}(K) = \sum_{k=1}^K V_{r,1}^{\pi^*}(x_1) - V_{r,1}^{\pi_k}(x_1),$$

$$\text{Violation}_H(K) = \sum_{k=1}^K \left(b - V_{g,1}^{\pi_k}(x_1) \right)_+, \tag{6}$$

where $[z]_+ = \max\{z, 0\}$. There is a violation of $b - V_{g,1}^{\pi_k}(x_1)$ at episode k if $V_{g,1}^{\pi_k}(x_1)$ is less than b. On the other hand, if $V_{g,1}^{\pi_k}(x_1) \geq b$, then there is no violation. Thus, we define the total violation as the cumulative sum of all constraint violations over all the episodes. Note the difference with the constraint violation metric (cf. (1)) considered in the existing literature (Ghosh et al., 2022; Ding et al., 2020, 2021; Wei et al., 2021a). In (1), if $V_{g,1}^{\pi_k}(x_1) \geq b$ it can negate the violation $V_{g,1}^{\pi_k}(x_1) < b$ in metric (1). As we have shown in the introduction, a sub-linear (even zero) violation defined in (1) does not guarantee that policies that violate the constraint by a fixed amount (independent of K) are selected for a sub-linear number of episodes. On the other hand, if the violation metric defined in (6) grows only sub-linearly with K, it implies that the number of episodes where such policies are selected scales at most sub-linearly with K.

Linear Function Approximation: To handle a possible large number of states, we consider the following linear MDP.

Assumption 1. *The CMDP is a linear MDP with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any h, there exists d* unknown *signed measures $\mu_h = \{\mu_h^1, \ldots, \mu_h^d\}$ over $\mathcal{S}$ such that for any $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $\mathbb{P}_h(x'|x, a) = \langle \phi(x, a), \mu_h(x') \rangle$ and there exists vectors $\theta_{r,h}$, $\theta_{g,h} \in \mathbb{R}^d$ such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$, $r_h(x, a) = \langle \phi(x, a), \theta_{r,h} \rangle \quad g_h(x, a) = \langle \phi(x, a), \theta_{g,h} \rangle$.*

This is similar to the setting considered in Ghosh et al. (2022) and is based on the definition of linear MDP (Jin et al., 2020; Yang and Wang, 2019). By the above definition, the transition model, the reward, and the utility functions are linear in terms of the feature map ϕ. We remark that despite being linear, $\mathbb{P}_h(\cdot|x, a)$ can still have infinite degrees of freedom since $\mu_h(\cdot)$ is unknown. Note that tabular MDP is a subset of linear MDP (Jin et al., 2020). A recent study (Zhang et al., 2022) showed that policies for linear MDPs can achieve better results than state-of-the-art approaches for benchmark databases. Linear MDPs are also viewed as a critical step toward studying large-scale RL problems, particularly those with infinite state space. Further, as demonstrated in numerous other settings, analyzing linear MDPs can provide insights that can be used to generalize to other settings. Thus, unconstrained linear MDP is extensively studied (Jin et al., 2020, 2021; He et al., 2021b,a; Wang et al., 2020; Hu et al., 2022).

Note that Ding et al. (2021); Zhou et al. (2021) studied another related concept known as linear kernel MDP. In the linear kernel MDP, the transition probability is given by $\mathbb{P}_h(x'|x, a) = \langle \psi(x', x, a), \theta_h \rangle$. In general, linear MDP and linear kernel MDPs are two different classes of MDP (Zhou et al., 2021).

Similar to Proposition 1 in Jin et al. (2020), we can show that for a linear MDP and for any policy π there exists $\{w_{j,h}^\pi\}_{h=1}^H$ such that $Q_{j,h}^\pi(x, a) = \langle w_{j,h}^\pi, \phi(x, a) \rangle$ for any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. We, thus, focus on the linear action-value function.

Dual Variable: We also use dual variable to consider a composite function.

Definition 2. *$V_h^{\pi,Y}(\cdot) = V_{h,r}^\pi(\cdot) + Y V_{h,g}^\pi(\cdot)$, and $Q_h^{\pi,Y}(x, a) = Q_{r,h}^\pi(x, a) + Y Q_{g,h}^\pi(x, a)$, where Y is the dual variable.*

3 Our Approach

We now describe our proposed Algorithm 1.

Algorithm 1 Model Free Algorithm with Linear Function Approximation for hard-constraint violation

1: **Initialization:** $\alpha = (\log(|\mathcal{A}|)\sqrt{K})/(4H)$, $\eta = 1/(d^{H-1}K^{1.5H}\log(|\mathcal{A}|)^H)$, $\beta = C_1 dH\sqrt{\log(4\log|\mathcal{A}|dT/p)}$, $w_{r,h}^1 = 0$, $w_{g,h}^1 = 0$.
2: **for** episodes $k = 1, \ldots, K$ **do**
3: Initialize: $Y_k = 0$, Receive the initial state x_1.
4: **while** $Y_k \leq \sqrt{K}$ **do**
5: **for** step $h = H, H-1, \ldots, 1$ **do**
6: $\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^T + \lambda \mathbf{I}$
7: $w_{r,h}^k = (\Lambda_h^k)^{-1}[\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + V_{r,h+1}^k(x_{h+1}^\tau)]]$
8: $w_{g,h}^k = (\Lambda_h^k)^{-1}[\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)[g_h(x_h^\tau, a_h^\tau) + V_{g,h+1}^k(x_{h+1}^\tau)]]$
9: $Q_{r,h}^k(\cdot, \cdot) = \min\{\langle w_{r,h}^k, \phi(\cdot, \cdot)\rangle + \beta(\phi(\cdot, \cdot)^T(\Lambda_h^k)^{-1}\phi(\cdot, \cdot))^{1/2}, H\}$
10: $Q_{g,h}^k(\cdot, \cdot) = \min\{\langle w_{g,h}^k, \phi(\cdot, \cdot)\rangle + \beta(\phi(\cdot, \cdot)^T(\Lambda_h^k)^{-1}\phi(\cdot, \cdot))^{1/2}, H\}$
11: $\pi_{h,k}(a|\cdot) = \textsc{Soft-Max}_\alpha^a(Q_{r,h}^k + Y_k Q_{g,h}^k)$ (see (8)),
12: $V_{r,h}^k(\cdot) = \sum_a \pi_{h,k}(a|\cdot)Q_{r,h}^k(\cdot, a)$, $V_{g,h}^k(\cdot) = \sum_a \pi_{h,k}(a|\cdot)Q_{g,h}^k(\cdot, a)$
13: **if** $V_{g,h}^k \geq b$ **then**
14: **break** //Found the ideal dual variable
15: $Y_k = Y_k + \eta$ //Increase the dual variable
16: **if** $Y_k > \sqrt{K}$ **then**
17: Set $Y_k = \sqrt{K}$,
18: **for** step $h = 1, \ldots, H$ **do**
19: Receive x_h^k; compute $Q_{j,h}^k(x_h^k, a)$, $\pi(a|x_h^k)$ for all a using $w_{j,h}^k$ and Y_k.
20: Take action $a_h^k \sim \pi_{h,k}(\cdot|x_h^k)$ and observe x_{h+1}^k.

Note that at each episode k, Algorithm 1 can be divided into two steps: i) the dual variable finding step (Lines 4–17), and ii) the policy execution step (Lines 18–20). The dual variable finding step can be divided into two main steps: i) the policy finding step (Lines 5– 12),

and ii) the constraint checking step for a given value of Y_k (Lines 13–15). We now describe the steps in detail.

In order to obtain $V^k_{g,1}(x_1)$, we need to first find the policy and Q-functions. For a given dual variable Y_k, lines 6-12 consist of updating the parameters $w^k_{r,h}, w^k_{g,h}$ and Λ^k_h which are used to update the $Q^k_{j,h}$ and $V^k_{j,h}$ at episode k. Λ^k_h is the Gram-matrix for the regularized least square problem (see (7), later). Note that the lines 9-12 (i.e., Q, V, and policy) are not evaluated for each state, rather, they are evaluated only for the encountered states till episode $k-1$. Hence, we do not need to iterate over a potentially infinite number of states. For the first episode, since $k-1=0$ and $\tau=1$, we have $w^k_{j,h}=0$, $\forall j$ and $\Lambda^k_h=\lambda\mathbf{I}$. We note that $Q^k_{j,H+1}(\cdot,\cdot)=0$ for $j=r,g$.

Q function and Value function Estimation: We need to estimate the value-function and Q-function with respect to the policy π_k for a given value of Y_k. However, there are challenges. We do not know $\mathbb{P}_h$ in Bellman's equation (4), rather $\mathbb{P}_h V^{\pi_k}_{j,h+1}$ should be replaced by the empirical samples. Further, since $Q^\pi_{j,h}(x,a)$ is linear in $\phi(x,a)$, we parameterize $Q^\pi_{j,h}(\cdot,\cdot)$ by a linear form $\langle w^k_{j,h}, \phi(\cdot,\cdot)\rangle$. The intuition is to obtain $w^k_{j,h}$ from Bellman's equation (cf.(4)) using the regularized least-square regression. We obtain $w^k_{j,h}$ for $j=r,g$ according to the following equation

$$w^k_{j,h} \leftarrow \arg\min_{w\in\mathbb{R}^d} \sum_{\tau=1}^{k-1} [j_h(x^\tau_h,a^\tau_h) + V^k_{j,h+1}(x^\tau_{h+1}) - w^T\phi(x^\tau_h,a^\tau_h)]^2 + \lambda||w||_2^2 \tag{7}$$

Then, an additional bonus term $\beta(\phi(\cdot,\cdot)^T(\Lambda^k_h)^{-1}\phi(\cdot,\cdot))^{1/2}$ is added as in Jin et al. (2020), where β is a constant which we will characterize in the next section. Such an additional term is used for the upper confidence bound in LSVI-UCB (Jin et al., 2020). The same bonus term is used for both $Q^k_{r,h}$ and $Q^k_{g,h}$.

Policy: The value functions are updated based on the Q function and the policy (line 13). The policy is based on a soft-max policy (line 12) unlike the greedy one in the unconstrained case Jin et al. (2020). Soft-max policy $\textsc{Soft-Max}_\alpha(\mathbf{X}) = \{\textsc{Soft-Max}^i_\alpha(\mathbf{X})\}_{i=1}^{|\mathcal{A}|}$ for any vector $\mathbf{X}\in\mathbb{R}^{|\mathcal{A}|}$ is a $|\mathcal{A}|$-dimensional vector with parameter α where the i-th component

$$\textsc{Soft-Max}^i_\alpha(\mathbf{X}) = \frac{\exp(\alpha X_i)}{\sum_{n=1}^{|\mathcal{A}|}\exp(\alpha X_n)}. \tag{8}$$

At step h, $\pi_{h,k}(a|x)$ is computed based on the soft-max policy with the composite Q-function vector $\{Q^k_{r,h}(x,a)+Y_kQ^k_{g,h}(x,a)\}_{a\in\mathcal{A}}$. When $\alpha=\infty$, this becomes equal to the greedy policy. As shown in Ghosh et al. (2022), the greedy policy is not Lipschitz. Hence, it does not provide a uniform concentration bound for each individual value function, an essential step in proving the regret and violation bound. Note that for different values of Y_k, the policy is different, hence $V^k_{j,h}$ would also be different. We also use the Lipschitz property of soft-max to obtain Y_k that satisfies certain characteristics which we discuss in the following.

Finding Y_k: Once we compute $Q^k_{g,1}$ and compute policy according to soft-max, we obtain $V^k_{g,1}(x_1)$. If $V^k_{g,1}(x_1)<b$, we increase the dual variable by η and repeat the steps (lines 5-12). We continue this process till Y_k reaches $\sqrt{K}$ or we obtain $V^k_{g,1}(x_1)\geq b$. Once we are out of the inner loop, we use Y_k, and $w^k_{j,h}$ to compute the policy in the execution phase.

In summary, at every episode, we identify Y_k such that one of the following three cases holds:

- $Y_k=0$, $V^k_{g,1}(x_1)\geq b$, thus, one can focus on maximizing the reward only.
- $\sqrt{K}\geq Y_k>0$, and $V^k_{g,1}(x_1)\geq b$. We show that in this case $V^k_{g,1}(x_1)\leq b+\mathcal{O}(K^{-1})$ by proper choice of step-size η and α (temp. co-efficient) of the soft-max policy ensuring the right balance. Intuitively, if we select a higher Y_k such that $V^k_{g,1}(x_1)$ exceeds b by a large amount, it would put a smaller weight on reward maximization. Hence, the regret increase.
- $Y_k=\sqrt{K}$, and $V^k_{g,1}(x_1)<b$ which means that we reach the upper bound of the dual variable. We show that an upper bound of $\sqrt{K}$ is the enough to obtain $\tilde{O}(\sqrt{K})$ regret and $\tilde{O}(\sqrt{K})$ hard constraint violation.

Execution: The last part includes the execution of the policy for episode k (lines 22-24). The policy is again based on the soft-max policy with $w^k_{j,h}$ and Y_k obtained in the dual variable finding step.

Difference with other approaches: Ghosh et al. (2022); Ding et al. (2021, 2020); Liu et al. (2021a) (OptPess-PrimalDual) also proposed primal-dual type algorithm. However, their focus was on minimizing the soft violation (cf.(1)) rather than the hard violation. The major difference is that at every episode, we find Y_k such that it has the property characterized above. Since in the soft-constraint violation, one can negate violation at one episode by selecting a strictly feasible policy at the other episode, hence searching for such a dual-variable *at every episode* was not required there. Rather, updating the dual variable at the end of the episode was enough since if the current policy is infeasible one can increase the dual variable to choose a feasible policy in the subsequent episode to negate the violation. However, because of stricter requirements, we need to find the perfect dual variable at

every episode. Our analysis is also different compared to those (see Section 4.2).

Space complexities: We remark that Algorithm 1 only needs to store $r_h(x_h^k, a_h^k), g_h(x_h^k, a_h^k), \Lambda_h^k$, and $\{\phi(x_h^k, a)\}_{a\in\mathcal{A}}$ for all $(h,k) \in [H] \times [K]$, hence, it takes $\mathcal{O}(d^2H + d\mathcal{A}T)$ space which is the same as the unconstrained setup Jin et al. (2020). For the tabular setup, using the approach presented in Appendix H, the space complexity is at most $\mathcal{O}(|S|^2|A|H)$ which is the same as the model-based approach in Efroni et al. (2020). We do not need to invert $(\Lambda_h^k)^{-1}$ for the tabular case.

Comparison with LP-based approach: Efroni et al. (2020) and Liu et al. (2021a) proposed LP-based approach to obtain the state-action occupancy measure for tabular setup in order to bound the hard constraint violation. From the state-action occupancy measure, they obtain the policy. However, solving LP for large state-space is challenging as the decision variables scale with the state-space. Neu and Okolo (2023); Lakshminarayanan et al. (2017); Neu and Pike-Burke (2020); Bas-Serrano et al. (2021) proposed an LP-based approach for linear MDP for the unconstrained problem. One may ask why not extending that approach to the constrained case. However, apart from Neu and Okolo (2023), the above formulations still rely on the finite state space assumption and the decision variables still consist of state-action occupancy measure which can be large for large state space. Neu and Okolo (2023) reduces the number of decision variables by assuming that a core set of state-action pairs is known which essentially means that a low dimensional state-action occupancy measure is enough to represent the rewards for all state-action pairs. However, finding the core set is difficult in general Neu and Okolo (2023). Also, all the above still rely on a model-based approach and need to estimate the transition probability. Neu and Okolo (2023) also proposed a model-free algorithm, however, that relies on the simulator. In summary, all the above approaches either rely on finite state-space assumptions or rely on a simulator and a core-set assumption to obtain a policy.

Instead, we propose a primal-dual-based algorithm and thus, we do not rely on any LP solver. Our algorithm is also model-free and does not need any simulator. Further, our algorithm works for infinite state space and we do not rely on core-set assumption as in Neu and Okolo (2023). Moreover, we would like to point out that since our approach is model-free, our algorithmic approach can be certainly extended to a larger class of RL problems. As described in our approach, we need to estimate the value function (we can use neural-network to estimate the value function) and then tune the dual variable till we achieve $V_{g,1}^k(x_1) \geq b$ or Y_k reaches $\sqrt{K}$.

Assumption of strict feasibility: All the primal-dual-based approaches which focus on minimizing the soft-constraint violation relies on strict feasibility assumption (aka Slater's condition). In particular, all the primal-dual based approaches (model-free or model-based) rely on strong duality (Paternain et al., 2019) in order to achieve soft-constraint violation. *This is the first result that shows that sub-linear regret and constraint violation (even soft) is achievable without assuming Slater's condition for a primal-dual-based approach.* Naturally, our analysis is significantly different (Section 4.2). If we assume the existence of a strictly feasible policy, we can set a lower upper bound for the dual-variable (Appendix J).

Value of η: We obtain theoretical results for $\eta = \mathcal{O}(1/(K^{1.5H}d^{H-1}))$. However, in practice, we observe that $\eta = \mathcal{O}(1/(\sqrt{KH}))$ works well, hence, the algorithm can be faster in practice. The characterization of a more computationally efficient algorithm with theoretical bound is left for the future work.

4 Analysis

We now state the main result. We prove that Algorithm 1 achieves the regret and hard constraint violation which are sublinear in $T = KH$ where T is the total number of steps.

4.1 Main Results

Theorem 1. *Fix any $p \in (0,1)$. If we set $\lambda = 1$, $\beta = C_1 dH\sqrt{\iota}$ in Algorithm 1 where $\iota = \log(\log(|\mathcal{A}|)4dT/p)$ for some absolute constant C_1. With probability $1-2p$, we have*

$$\text{Regret}(K) \leq C\sqrt{d^3H^3T\iota^2},$$
$$\text{Violation}_H(K) \leq C'\sqrt{d^3H^3T\iota^2}$$

for some absolute constants C, and C'.

To the best of our knowledge, this is the first result that achieves $\tilde{\mathcal{O}}(\sqrt{T})$ regret and hard constraint violation bound for linear CMDP. Since linear CMDP contains a tabular setup, as a by-product, our result also provides the hard-constraint violation bound for tabular setup using the model-free algorithm. As we mentioned earlier, existing primal-dual type algorithms only consider soft-constraint violation, rather, *we show that it is possible to achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret and hard constraint violation bound using primal-dual type approach.* Our bound with respect to T matches the bounds attained in the tabular model-based and LP-based approach in Efroni et al. (2020) and Liu et al. (2021a) (OptPess-LP).*

*OptPess-LP achieves zero violation, however, they assume that a safe policy is known.

Note that our bounds are (nearly) optimal. It is shown that $\Omega(d\sqrt{H^2T})$ regret is unavoidable for the unconstrained setup. We can easily construct an example where only one policy is feasible. For example, consider the setup where $g = r$, and $b = V^*_{r,1}(x_1)$, (in this case, we are forcing an unconstrained problem to be constrained) then only the optimal policy is feasible. Thus, $\Omega(d\sqrt{H^2T})$ is also a lower bound for hard constraint violation as well when there is no strictly feasible policy.

Note that our regret bound matches the same order (with respect to T) as in the unconstrained case (Jin et al., 2020) and in the linear CMDP setup with soft-constraint violation (Ghosh et al., 2022). Ghosh et al. (2022) achieves zero soft-constraint violation when Slater's condition holds. Whether it is possible to reduce our violation bound further for the setup when the Slater's condition holds remains open.

Tabular Case: If we consider the tabular case with the following representation $\phi(x, a) = e_{x,a}$ where $e_{x,a}$ is a $|S||A|$ dimensional vector and $e_{x,a} = 1$ when $(s, a) = (x, a)$ and 0 otherwise, then plugging in $d = |S||A|$ we obtain the regret and hard constraint violation bound as $\tilde{O}(\sqrt{|S|^3|A|^3H^3T})$. However, since we can trivially obtain ϵ-covering number for a value function when the state-space is bounded, we can obtain a tighter result by modifying the bonus term β. In particular, in Theorem 3 (Appendix H) we obtain the regret and hard constraint violation bounds as $\tilde{O}(\sqrt{|S|^2|A|H^3T})$. This matches the result of Efroni et al. (2020). This is the first work which shows that it is possible to achieve $\tilde{O}(\sqrt{|S|^2|A|H^3T})$ regret and hard constraint violation bound using *primal-dual* based approach matching the result from model-based LP-based approach (Efroni et al., 2020).

4.2 Proof Outline

We start by highlighting the main differences with the existing approaches.

Novelty in Analysis techniques: Existing primal-dual approaches that focus on bounding the soft-constraint violation (Ghosh et al., 2022; Ding et al., 2021; Efroni et al., 2020) seek to bound for any $Y \geq 0$

$$\sum_{k=1}^{K}(V^*_{r,1}(x_1) - V^{\pi_k}_{r,1}(x_1)) + Y(b - V^{\pi_k}_{g,1}(x_1)) \qquad (9)$$

In order to prove regret, they then bound $(Y - Y_k)(b - V^k_{g,1}(x_1)) \leq \mathcal{O}(\sqrt{K})$ using the fact that dual variable is updated based on the gradient descent step in the dual direction. Since our dual update is different, the regret analysis is significantly different. Further, we can not rely on strong duality result to bound the hard constraint violation unlike obtaining the soft constraint violation bound.

Rather, to prove the regret and the hard constraint violation bound, we first show that if the dual variable differs by ϵ amount, the estimated value function for utility $(V^k_{g,1})$ can also differ by at most $O(K^{1.5H}d^H\epsilon)$ amount. Thus, by incrementing Y_k by $\epsilon = O(K^{-1.5H+1}d^{-H})$, the maximum increment of $V^k_{g,1}$ for two different dual-variables would be bounded by $O(K^{-1})$. Hence, one can find Y_k such that $b \leq V^k_{g,1}(x_1) \leq b + O(K^{-1})$ (if it is achievable within upper bound $\sqrt{K}$). This turns out to be essential to bound the regret and hard constraint violation. Such a guarantee is not required to obtain an upper bound for the soft-constraint violation as considered in the other papers. To bound the total soft constraint violation, one only needs to update the dual variable at the end of the episode k.

To prove that if the dual variable differs by ϵ amount, the estimated value function for utility $(V^k_{g,1})$ can also differ by at most $O(K^{1.5H}d^H\epsilon)$ amount; we use the fact that our policy is a soft-max to show that if the dual variable differs by an ϵ amount, then the value function at the h-th step only differs by at most $O(\epsilon\sqrt{K}H)$ (Lemma 15). However, as the policy for the h-th step changes, the parameter for the $h-1$-th step would also change to fit Bellman's equation. Here, we use the linearity property to show that the parameter value $w^k_{j,h-1}$ (obtained via solving the linear regression problem) also differs by $O(dK\epsilon)$ (Lemma 16). Using the above, we show that the value function at the $h-1$-th step would differ by at most $O(d\epsilon K\sqrt{K}H)$ (Lemma 14). We obtain the final result by induction. Please see Appendix C for details. We now provide the main ideas behind bounding Regret and Hard constraint violation.

Regret Bound: We decompose the regret in the following manner:

$$\text{Regret}(K) = \underbrace{\sum_{k=1}^{K}(V^*_{r,1}(x_1) - V^k_{r,1}(x_1))}_{\mathcal{T}_1} + \underbrace{\sum_{k=1}^{K}(V^k_{r,1}(x_1) - V^{\pi_k}_{r,1}(x_1))}_{\mathcal{T}_2} \qquad (10)$$

In order to bound both $\mathcal{T}_1$ and $\mathcal{T}_2$ we need to obtain uniform concentration bound for each individual value function. In particular, we need to show that the log ϵ-covering number of (estimated) reward and utility value function must scale as $\log(K)$. As discussed in Ghosh et al. (2022), the greedy policy with respect to the composite state-action value function fails to achieve such bound since the greedy policy is not Lipschitz. Instead, the soft-max policy based on the composite state-action policy function achieves the above (Lemma 10).

We, first, discuss how to bound $\mathcal{T}_1$. We observe that

$$\mathcal{T}_1 \leq \underbrace{\sum_{k=1}^{K}(V^*_{r,1}(x_1) + Y_k V^*_{g,1}(x_1) - V^k_{r,1}(x_1) - Y_k V^k_{g,1}(x_1))}_{\mathcal{T}_3} + \underbrace{\sum_{k=1}^{K} Y_k (V^k_{g,1}(x_1) - b)}_{\mathcal{T}_4} \quad (11)$$

where we have used the fact that $V^*_{g,1}(x_1) \geq b$ and $Y_k \geq 0$. We now bound $\mathcal{T}_3$ and $\mathcal{T}_4$.

Readers should note that $\mathcal{T}_3$ is similar to the optimism term with respect to the composite value function. However, since we use soft-max instead of the greedy policy, we cannot bound the above by zero. Rather, using the property of soft-max, we obtain

Lemma 1. *For any k, with probability $1-p$, $(V^*_{r,1}(x_1) - V^k_{r,1}(x_1) + Y_k V^*_{g,1}(x_1) - Y_k V^k_{g,1}(x_1)) \leq \log(|\mathcal{A}|)H/\alpha$.*

When $\alpha = \log(|\mathcal{A}|)\sqrt{K}/(4H)$, the above can be bounded by $\mathcal{O}(H^2/K^{1/2})$. Also, note that it shows that a large value of α would degrade the regret. Now, we provide an upper bound for $\mathcal{T}_4$.

We, first, define the set of episodes where $Y_k = \sqrt{K}$, and $V^k_{g,1}(x_1) < b$, $\mathcal{I}_b = \{k : Y_k = \sqrt{K}, V^k_{g,1}(x_1) < b\}$. Thus, for these episodes $V^k_{g,1}(x_1) - b \leq 0$. Hence, for these episodes, $\mathcal{T}_4$ is trivially upper bounded by 0. Hence, we now need to obtain an upper bound on $\mathcal{T}_4$ on the set of episodes which are not in $\mathcal{I}_b$, i.e., when they belong to the set $\mathcal{I}_b^C$.

As we have discussed, by the choice of η, when we obtain $Y_k > 0$ such that $V^k_{g,1}(x_1) \geq b$, we also obtain $V^k_{g,1}(x_1) - b \leq \mathcal{O}(K^{-1})$. Formally,

Lemma 2. *For any $k \in \mathcal{I}_b^C$, $Y_k(V^k_{g,1}(x_1) - b) \leq \mathcal{O}(HK^{-1/2})$.*

Intuitively, we strike the balance between reward maximization and utility maximization for $V^k_{r,1}(x_1)$ and $V^k_{g,1}(x_1)$ (as even if $V^k_{g,1}(x_1) > b$, it would only exceed by $\mathcal{O}(HK^{-1})$). Hence, we can upper bound $\mathcal{T}_4$ by $\mathcal{O}(HK^{-1/2})$. Thus, summing over the expressions in Lemmas 1 and 2 we obtain with probability $1-p$, $\mathcal{T}_1 \leq \mathcal{O}(HK^{1/2})$.

We obtain the bound of $\mathcal{T}_2$ using Azuma-Hoeffding inequality and the uniform concentration bound,

Lemma 3. *With probability $1-p$, $\mathcal{T}_2 \leq \mathcal{O}(\sqrt{d^3\iota^2 K H^4})$*

Hard constraint Violation Bound: We decompose the violation as the following:

$$\sum_{k=1}^{K}(b - V^{\pi_k}_{g,1}(x_1))_+ \leq \underbrace{\sum_{k=1}^{K}(b - V^k_{g,1}(x_1))_+}_{\mathcal{T}_5} + \underbrace{\sum_{k=1}^{K}(V^k_{g,1}(x_1) - V^{\pi_k}_{g,1}(x_1))_+}_{\mathcal{T}_6} \quad (12)$$

We, further, decompose $\mathcal{T}_5$ as the following

$$\mathcal{T}_5 \leq \sum_{k\in\mathcal{I}_b}(b - V^k_{g,1}(x_1))_+ + \sum_{k\in\mathcal{I}_b^C}(b - V^k_{g,1}(x_1))_+ \quad (13)$$

By the definition of $\mathcal{I}_b$, for episodes in $\mathcal{I}_b$, $Y_k \geq \sqrt{K}$, and $b > V^k_{g,1}(x_1)$, yet since $\alpha = \log(|\mathcal{A}|)K^{1/2}/(4H)$, we obtain from Lemma 1

$$\sum_{k\in\mathcal{I}_b}(b - V^k_{g,1}(x_1))_+ \leq \mathcal{O}(H^2\sqrt{K}) \quad (14)$$

In order to bound the second term in the right-hand side of (13) note that $V^k_{g,1}(x_1) \geq b$ for $k \in \mathcal{I}_b^C$. Hence, $(b - V^k_{g,1}(x_1))_+ = 0$. Thus, we have

Lemma 4. *With probability $1-p$, $\mathcal{T}_5 \leq \mathcal{O}(H^2\sqrt{K})$.*

Finally, we bound $\mathcal{T}_6$. Since we add bonus term to obtain $Q^k_{g,h}$, thus,

Lemma 5. *With probability $1-p$, $V^k_{g,1}(x_1) \geq V^{\pi_k}_{g,1}(x_1)$.*

Thus, we can rewrite $\sum_k (V^k_{g,1}(x_1) - V^{\pi_k}_{g,1}(x_1))_+$ as $\sum_k (V^k_{g,1}(x_1) - V^{\pi_k}_{g,1}(x_1))$. Thus, from Azuma-Hoeffding inequality and uniform concentration bound, we obtain

Lemma 6. *With probability $1-p$, $\mathcal{T}_6 \leq \mathcal{O}(\sqrt{\iota^2 d^3 H^4 K})$.*

From Lemma 4 and 6 we obtain the bound on violation in Theorem 1.

5 Experiments

We evaluate Algorithm 1 on a simulated model (same as in Ghosh et al. (2022), details in Appendix K.1) to validate our theoretical results. We run Algorithm 1 for 3×10^5 episodes (K). We use $\eta = 1/\sqrt{KH}$. Thus, we are using a larger η proposed in Algorithm 1 as it decreases the time complexity. We observe that such η is enough to achieve sub-linear regret and hard constraint violation. We use the feature-space representation similar to tabular setup (Appendix H). We compare our algorithm with two state-of-the-art algorithm: i) the algorithm proposed in Ghosh et al. (2022), and ii) OptPess-PrimalDual proposed in Liu

Figure 1: Comparison of our approach with Ghosh et al. (2022) and OptPess-PrimalDual (Liu et al., 2021a). Each plot is an average of 10 trials. The length of each episode (H) is 10.

et al. (2021a). Our empirical results (Figure 1) suggest that our algorithm significantly reduces the hard constraint violations as compared to both the algorithms. As predicted by our theory, the hard constraint violation scales much smaller than $O(\sqrt{K})$ (Figure 1). In fact, our algorithm selects feasible policy after 1.5×10^5 episodes. However, the constraint violations grow for the other two algorithms which indicate that those algorithms are unable to find feasible policy. Obviously, infeasible policies can give higher rewards, thus, the average reward achieved by the other two algorithms are *slightly* higher compared to our approach. Nevertheless, our algorithm indeed achieves optimal reward. Thus, the empirical result shows the efficacy of our approach in achieving sub-linear hard constraint violation and regret. In Appendix K, we observe similar traits in our empirical results on the OpenAIGym control suite (Brockman et al., 2016) and other CMDP setups.

6 Conclusion and Future Work

We propose a model-free RL-based algorithm for linear CMDP which achieved $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ regret and $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ hard constraint violation bound. To the best of our knowledge, this is the first result which shows $\tilde{\mathcal{O}}(\sqrt{T})$ regret and $\tilde{\mathcal{O}}(\sqrt{T})$ constraint violation bound using primal-dual model-free setup. We achieve our result by finding the dual variable that balances between regret and constraint violation within an episode, rather than only updating it at the end of each episode.

Whether we can tighten the dependence on d and H remains an important future research direction. Extending the work to the setup where the feature space needs to be learnt or non-linear MDP setup is also important. Recent works (Modi et al., 2021; Zhang et al., 2022; Agarwal et al., 2020) on feature-space learning for unconstrained MDPs may provide some insights.

Acknowledgment

A part of this work was done when AG was at the Ohio State University. This work has been supported in part by NJIT Start up fund indexed number 172884, NSF grants: CNS-2153220, CNS-2312835, CNS-2312836, CNS- 2223452, CNS-2225561, CNS-2112471, CNS-2106933, a grant from the Army Research Office: W911NF-21-1-0244, and was sponsored by the Army Research Laboratory under Cooperative Agreement Number W911NF-23-2-0225. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. (2020). Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107.

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

Amani, S., Thrampoulidis, C., and Yang, L. F. (2021). Safe reinforcement learning with linear function approximation. *arXiv preprint arXiv:2106.06239*.

Azar, M. G., Munos, R., and Kappen, B. (2012). On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*.

Bas-Serrano, J., Curi, S., Krause, A., and Neu, G. (2021). Logistic q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3610–3618. PMLR.

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. *arXiv preprint arXiv:2006.05051*.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Chen, T., Gangrade, A., and Saligrama, V. (2022). A

doubly optimistic strategy for safe linear bandits. *arXiv preprint arXiv:2209.13694*.

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 *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR.

Ding, D., Zhang, K., Basar, T., and Jovanovic, M. R. (2020). Natural policy gradient primal-dual method for constrained markov decision processes. In *NeurIPS*.

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

Epasto, A., Mahdian, M., Mirrokni, V., and Zampetakis, M. (2020). Optimal approximation–smoothness tradeoffs for soft-max functions. *arXiv preprint arXiv:2010.11450*.

Ghosh, A., Zhou, X., and Shroff, N. (2022). Provably efficient model-free constrained rl with linear function approximation. *arXiv preprint arXiv:2206.11889*.

Guo, H., Zhu, Q., and Liu, X. (2022). Rectified pessimistic-optimistic learning for stochastic continuum-armed bandit with constraints. *arXiv preprint arXiv:2211.14720*.

He, J., Zhou, D., and Gu, Q. (2021a). Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR.

He, J., Zhou, D., and Gu, Q. (2021b). Uniform-pac bounds for reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:14188–14199.

Hu, P., Chen, Y., and Huang, L. (2022). Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 8971–9019. PMLR.

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

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR.

Kalagarla, K. C., Jain, R., and Nuzzo, P. (2020). A sample-efficient algorithm for episodic finite-horizon mdp with constraints. *arXiv preprint arXiv:2009.11348*.

Lakshminarayanan, C., Bhatnagar, S., and Szepesvári, C. (2017). A linearly relaxed approximate linear program for markov decision processes. *IEEE Transactions on Automatic control*, 63(4):1185–1191.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. (2021a). Learning policies with zero or bounded constraint violation for constrained mdps. *arXiv preprint arXiv:2106.02684*.

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

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. (2021). Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*.

Moskovitz, T., O'Donoghue, B., Veeriah, V., Flennerhag, S., Singh, S., and Zahavy, T. (2023). Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. *arXiv preprint arXiv:2302.01275*.

Neu, G. and Okolo, N. (2023). Efficient global planning in large mdps via stochastic primal-dual optimization. In *International Conference on Algorithmic Learning Theory*, pages 1101–1123. PMLR.

Neu, G. and Pike-Burke, C. (2020). A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. (2021). Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR.

Pan, L., Cai, Q., Meng, Q., Chen, W., Huang, L., and Liu, T.-Y. (2019). Reinforcement learning with dynamic boltzmann softmax updates. *arXiv preprint arXiv:1903.05926*.

Paternain, S., Calvo-Fullana, M., Chamon, L. F., and Ribeiro, A. (2019). Safe policies for reinforcement learning via primal-dual methods. *arXiv preprint arXiv:1911.09101*.

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

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wang, R., Du, S. S., Yang, L., and Salakhutdinov, R. R. (2020). On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. (2021a). Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR.

Wei, H., Liu, X., and Ying, L. (2021b). A provably-efficient model-free algorithm for constrained markov decision processes. *arXiv preprint arXiv:2106.01577*.

Xu, T., Liang, Y., and Lan, G. (2021). Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR.

Zhang, X., Song, Y., Uehara, M., Wang, M., Agarwal, A., and Sun, W. (2022). Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR.

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

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable]
   Yes, (please see Section 3).
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable]
   Yes, (please see Section 3).
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]
   Yes. We have described all the hyperparameters required to rerun the simulations. The source code will be published for the camera-ready version.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable]
   Yes, (please see Section 4).
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable]
   Yes, (please see Appendix).
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable]
   Yes.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   Yes, (in Section 5 and Appendix K).
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
   Not applicable.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable]
   Not applicable.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable]
   Not applicable.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable]
   Not applicable.
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable]
   Not applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]
   Not applicable.
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable]
   Not applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable]
   Not applicable.
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable]

Not applicable.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable]

Not applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable]

Not applicable.

Organization of Appendix: In Section A, we state some results which we use throughout. In Section B, we prove Lemma 1. In Section B.1, we state and prove base results Lemmas 10, 11, and 12 which are necessary to prove the Lemma 1. Subsequently, we prove Lemma 1. In Section C, we prove Lemma 2. In Section D, we prove Lemmas 4 and 5. In Section E, we prove Lemmas 3 and 6. In Section F we prove Lemma 10 (the uniform concentration result) which is essential to prove all the previous results. In Section G, we state some results proved in the existing literature which we have used in proving our results. In Section H, we detail our algorithm for tabular setup. In Section I, we describe why the approach for Bandit setup can not be applied to our CMDP setup. In Section J, we show that when a strictly feasible policy exists (aka Slater's condition holds) upper bound of H/γ for the dual variable is enough to obtain $\tilde{\mathcal{O}}(\sqrt{K})$ regret and hard constraint violation instead of an upper bound of $\sqrt{K}$. Finally, in Section K, we provide empirical results of our algorithm for various CMDP setups including OpenAI Gyms suite Brockman et al. (2016).

Notations: Throughout the rest of this paper, we denote $Q^{k,Y}_{r,h}, Q^{k,Y}_{g,h}, V^{k,Y}_{r,h}, V^{k,Y}_{g,h}, w^{k,Y}_{r,h}, w^{k,Y}_{g,h}$, as the Q-value, value-function, and the parameter values estimated respectively at the episode k for a given dual variable Y (inside the while loop in Algorithm 1). Note that policy depends on Y, hence, for different Y, $V^{k,Y}_{j,h}$ would be different. Naturally, $w^{k,Y}_{j,h}$ and $Q^{k,Y}_{j,h}$ would be different (cf.(7)). We denote $V^{k,Y}_{h+1}(\cdot) = V^{k,Y}_{r,h+1}(\cdot) + YV^{k,Y}_{g,h}(\cdot)$.

Further, we denote $Q^k_{j,h}, V^k_{j,h}, w^k_{j,h}$, as the Q- value, value-function, and the parameters chosen for the determined Y_k (i.e., after the While loop in Algorithm 1 terminates). Hence, $V^k_{j,h}(\cdot) = V^{k,Y_k}_{j,h}(\cdot)$, $Q^k_{j,h}(\cdot,\cdot) = Q^{k,Y_k}_{j,h}(\cdot,\cdot)$. $V^{k,Y}_{j,h}(\cdot) = \langle \pi_{h,k}(\cdot|\cdot), Q^{k,Y}_{j,h}(\cdot,\cdot)\rangle_{\mathcal{A}}$. $\pi_{h,k}(\cdot|x)$ is the soft-max policy based on the composite Q-function at the k-th episode as $Q^k_{r,h} + YQ^k_{g,h}$. Here, $\pi_{h,k}(\cdot)$ depends on the dual variable. Thus, the dependence is implicit. Sometimes, we also use the notation π^Y to make the dependence on the dual variable explicit.

To simplify the presentation, we denote $\phi^k_h = \phi(x^k_h, a^k_h)$. Without loss of generality, we assume $||\phi(x,a)||_2 \leq 1$ for all $(x,a) \in \mathcal{S}\times\mathcal{A}$, $||\mu_h(\mathcal{S})||_2 \leq \sqrt{d}$, $||\theta_{j,h}||_2 \leq \sqrt{d}$ for $j = r, g$ and all $h \in [H]$.

A Preliminary Results

Lemma 7. *Under Assumption 1, for any fixed policy π, let w^π_h be the corresponding weights such that $Q^\pi_{j,h} = \langle \phi(x,a), w^\pi_{j,h}\rangle$, for $j \in \{r,g\}$, then we have for all $h \in [H]$,*

$$||w^\pi_{j,h}|| \leq 2H\sqrt{d} \tag{15}$$

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

$$\begin{aligned} Q^\pi_{j,h}(x,a) &= j_h(x,a) + \mathbb{P}_h V^\pi_{j,h}(x,a) \\ &= \langle \phi(x,a), \theta_{j,h}\rangle + \int_{\mathcal{S}} V^\pi_{j,h+1}(x')\langle \phi(x,a), d\mu_h(x')\rangle \\ &= \langle \phi(x,a), w^\pi_{j,h}\rangle \end{aligned} \tag{16}$$

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

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

Lemma 8. *For any (k,h,Y), the weight $w^{k,Y}_{j,h}$ satisfies*

$$||w^{k,Y}_{j,h}|| \leq 2H\sqrt{dk/\lambda} \tag{17}$$

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

$$|v^T w^{k,Y}_{j,h}| = |v^T(\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h(x^\tau_h, a^\tau_h)(j_h(x^\tau_h, a^\tau_h) + \sum_a \pi_{h+1,k}(a|x^\tau_{h+1})Q^{k,Y}_{j,h+1}(x^\tau_{h+1}, a))| \tag{18}$$

here $\pi_{h,k}(\cdot|x)$ is the Soft-max policy. Note that $\pi_{h,k}(\cdot|x)$ implicitly depends on the dual-variable Y.

Note that $Q^{k,Y}_{j,h+1}(x,a) \leq H$ for any (x,a). Hence, from (18) we have

$$\begin{aligned}|v^T w^{k,Y}_{j,h}| &\leq \sum_{\tau=1}^{k-1} |v^T (\Lambda^k_h)^{-1} \phi^\tau_h|.2H \\ &\leq \sqrt{\sum_{\tau=1}^{k-1} v^T (\Lambda^h_k)^{-1} v} \sqrt{\sum_{\tau=1}^{k-1} \phi^\tau_h (\Lambda^k_h)^{-1} \phi^\tau_h}.2H \\ &\leq 2H||v|| \frac{\sqrt{dk}}{\sqrt{\lambda}} \end{aligned} \tag{19}$$

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

B Proof of Lemma 1

We prove a more general result.

Lemma 9. *For any episode k and $0 \leq Y \leq \sqrt{K}$, with probability $1-p$, $(V^*_{r,1}(x_1) - V^{k,Y}_{r,1}(x_1) + YV^*_{g,1}(x_1) - YV^{k,Y}_{g,1}(x_1)) \leq \frac{\log(|\mathcal{A}|)H}{\alpha}$.*

Note that since the above holds for $0 \leq Y \leq \sqrt{K}$, it will hold for Y_k chosen value by Algorithm 1 at episode k when the While loop terminates. Thus, Lemma 1 readily follows.

In order to prove the above result, we prove some base results in Section B.1. Subsequently, we prove Lemma 9 in Section B.2.

B.1 Proof of Base Results

We state and prove Lemmas 10,11, and 12.

First, we state the concentration lemma which is essential in controlling the fluctuations in the least square value iteration for individual value function.

Lemma 10. *There exists a constant C_2 such that for any fixed $p \in (0,1)$, if we let $\mathcal{E}$ be the event that*

$$\left\| \sum_{\tau=1}^{k-1} \phi^\tau_{j,h} [V^{k,Y}_{j,h+1}(x^\tau_{h+1}) - \mathbb{P}_h V^{k,Y}_{j,h+1}(x^\tau_h, a^\tau_h)] \right\|_{(\Lambda^k_h)^{-1}} \leq C_2 dH\sqrt{\chi} \tag{20}$$

for all $j \in \{r,g\}$, $\chi = \log[4(C_1+1)\log(|\mathcal{A}|)dT/p]$, for some constant C_2, then $\Pr(\mathcal{E}) = 1-p$.

The proof of Lemma 10 is technical and relegated to Appendix F. Note from the unconstrained setup Jin et al. (2020), in order to prove the above bound, one needs to rely on the uniform concentration lemma. However, greedy policy fails to provide such bound as shown in Ghosh et al. (2022). Hence, similar to Ghosh et al. (2022), we use soft-max policy. *However, there is a subtle difference with Ghosh et al. (2022).* Ghosh et al. (2022) considered an upper bound which is constant (i.e., $2H/\gamma$, see Appendix J). However, in our setup the upper bound of Y_k is $\sqrt{K}$. Nevertheless, even though the upper bound depends on K, it only scales the constant C_1 while keeping the $\chi = \mathcal{O}(\log(T))$, the same as in Ghosh et al. (2022).

We now, recursively bound the difference between the value function maintained in Algorithm 1 (without the bonus term) and the value function for any policy for both the reward and utility value functions. We bound this using the expected difference at the next step plus an error term. This error term can be upper bounded by the bonus term with a high-probability.

Lemma 11. *There exists an absolute constant $\beta = C_1 dH\sqrt{\iota}$, $\iota = \log(\log(|\mathcal{A}|)4dT/p)$, and for any fixed policy π, on the event $\mathcal{E}$ defined in Lemma 10, we have*

$$\langle \phi(x,a), w^{k,Y}_{j,h} \rangle - Q^\pi_{j,h}(x,a) = \mathbb{P}_h(V^{k,Y}_{j,h+1} - V^\pi_{j,h+1})(x,a) + \Delta^k_h(x,a) \tag{21}$$

for some $\Delta^k_h(x,a)$ that satisfies $|\Delta^k_h(x,a)| \leq \beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)}$.

Proof. We only prove for $j = r$, the proof for $j = g$ is similar. For notational simplicity, we also remove Y from the superscript in $w_{j,h}^{k,Y}$ for the remainder of this proof.

Note that $Q_{r,h}^{\pi}(x,a) = \langle \phi(x,a), w_{r,h}^{\pi} \rangle = r_h(x,a) + \mathbb{P}_h V_{r,h+1}^{\pi}(x,a)$.

Hence, we have

$$\begin{aligned} w_{r,h}^k - w_{r,h}^{\pi} &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[r_h^{\tau} + V_{r,h+1}^k(x_{h+1}^{\tau})] - w_{r,h}^{\pi} \\ &= -\lambda(\Lambda_h^k)^{-1}(w_{r,h}^{\pi}) + (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[V_{r,h+1}^k(x_{h+1}^{\tau}) - \mathbb{P}_h V_{r,h+1}^k(x_h^{\tau}, a_h^{\tau})] \\ &+ (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[\mathbb{P}_h V_{r,h+1}^k(x_h^{\tau}, a_h^{\tau}) - \mathbb{P}_h V_{r,h+1}^{\pi}(x_h^{\tau}, a_h^{\tau})] \end{aligned} \tag{22}$$

Now, we bound each term in the right hand side of expression in (22). We call those terms as $\mathbf{q}_1$, $\mathbf{q}_2$, and $\mathbf{q}_3$ respectively.

First, note that

$$\begin{aligned} |\langle \phi(x,a), \mathbf{q}_1 \rangle| &= |\lambda \langle \phi(x,a), (\Lambda_h^k)^{-1}(w_{r,h}^{\pi}) \rangle| \\ &\leq \sqrt{\lambda} ||w_{r,h}^{\pi}|| \sqrt{\phi(x,a)^T (\Lambda_h^k)^{-1} \phi(x,a)} \end{aligned} \tag{23}$$

Second, from Lemma 10, for the event in $\mathcal{E}$, we have

$$|\langle \phi(x,a), \mathbf{q}_2 \rangle| \leq CdH\sqrt{\chi} \sqrt{\phi(x,a)^T (\Lambda_h^k)^{-1} \phi(x,a)} \tag{24}$$

where $\chi = \log(4(C_1+1)\log(|\mathcal{A}|)dT/p)$. Third,

$$\begin{aligned} \langle \phi(x,a), \mathbf{q}_3 \rangle &= \langle \phi(x,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}[\mathbb{P}_h(V_{r,h+1}^k - V_{r,h+1}^{\pi})(x_h^{\tau}, a_h^{\tau})] \rangle \\ &= \langle \phi(x,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^{\tau}(\phi_h^{\tau})^T \int (V_{r,h+1}^k - V_{r,h+1}^{\pi})(x') d\mu_h(x') \rangle \\ &= \langle \phi(x,a), \int (V_{r,h+1}^k - V_{r,h+1}^{\pi})(x') d\mu_h(x') \rangle - \langle \phi(x,a), \lambda(\Lambda_h^k)^{-1} \int (V_{r,h+1}^k - V_{r,h+1}^{\pi})(x') d\mu_h(x') \rangle \end{aligned} \tag{25}$$

The last term in (25) can be bounded as the following

$$|\langle \phi(x,a), \lambda(\Lambda_h^k)^{-1} \int (V_{r,h+1}^k - V_{r,h+1}^{\pi})(x') d\mu_h(x') \rangle| \leq 2H\sqrt{d\lambda} \sqrt{\phi(x,a)^T (\Lambda_h^k)^{-1} \phi(x,a)} \tag{26}$$

since $||\int (V_{r,h+1}^k - V_{r,h+1}^{\pi})(x') d\mu_h(x')||_2 \leq 2H\sqrt{d}$ as $||\mu_h(\mathcal{S})|| \leq \sqrt{d}$. The first term in (25) is equal to

$$\mathbb{P}_h(V_{r,h+1}^k - V_{r,h+1}^{\pi})(x,a) \tag{27}$$

Note that $\langle \phi(x,a), w_{r,h}^k \rangle - Q_{r,h}^{\pi}(x,a) = \langle \phi(x,a), w_{r,h}^k - w_{r,h}^{\pi} \rangle = \langle \phi(x,a), \mathbf{q_1} + \mathbf{q_2} + \mathbf{q_3} \rangle$. Since $\lambda = 1$, we have from (23), (24,(26), and (27)

$$|\langle \phi(x,a), w_{r,h}^k \rangle - Q_{r,h}^{\pi}(x,a) - \mathbb{P}_h(V_{r,h+1}^k - V_{r,h+1}^{\pi})(x,a)| \leq C_3 dH\sqrt{\chi} \sqrt{\phi(x,a)^T (\Lambda_h^k)^{-1} \phi(x,a)} \tag{28}$$

for some constant C_3 which is independent of C_1. Finally, note that

$$\begin{aligned} C_3\sqrt{\chi} &= \sqrt{\log(4(C_1+1)\log(|\mathcal{A}|)dT/p)} \\ &= C_3\sqrt{\iota + \log(C_1+1)} \\ &\leq C_1\sqrt{\iota} \end{aligned} \tag{29}$$

where $\iota = \log(4\log(|\mathcal{A}|)dT/p)$. The last inequality follows from the fact that $\iota \in [\log 4, \infty)$ as $|A| \geq 2$, and C_3 is independent of C_1. Hence, we can always pick $C_3\sqrt{\log 4 + \log(C_1+1)} \leq C_1\sqrt{\log 4}$ which satisfies (29) for all values of $\iota \in [\log 4, \infty)$. □

Next, using the above lemma, we bound the difference between the composite value function maintained by the algorithm and the composite value function for a policy with the Lagrangian Y_k.

Lemma 12. *With prob.* $1-p$, *(for the event in* $\mathcal{E}$*) and any* $0 \leq Y \leq \sqrt{K}$,

$$Q^{\pi}_{r,h}(x,a) + YQ^{\pi}_{g,h}(x,a) \leq Q^{k}_{r,h}(x,a) + YQ^{k,Y}_{g,h}(x,a) - \mathbb{P}_h(V^{k,Y}_{h+1} - V^{\pi,Y}_{h+1})(x,a) \tag{30}$$

Proof. Note the fact that $|Q^{\pi}_{r,h}| \leq H$. Thus, from Lemma 11 on the event $\mathcal{E}$ and for any $Y \in [0, \sqrt{K}]$,

$$\begin{aligned} Q^{\pi}_{r,h}(x,a) &\leq \min\{\langle \phi(x,a), w^{k,Y}_{r,h}\rangle + \beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)}, H\} \\ &\quad + \mathbb{P}_h(V^{\pi}_{r,h+1} - V^{k,Y}_{r,h+1})(x,a) \\ &= Q^{k,Y}_{r,h}(x,a) + \mathbb{P}_h(V^{\pi}_{r,h+1} - V^{k,Y}_{r,h+1})(x,a) \end{aligned}$$

where the last equality follows from the definition of $Q^{k,Y}_{r,h}$.

Similarly, for the event $\mathcal{E}$,

$$YQ^{\pi}_{g,h}(x,a) \leq YQ^{k,Y}_{g,h}(x,a) + Y\mathbb{P}_h(V^{\pi}_{g,h+1} - V^{k,Y}_{g,h+1})(x,a)$$

Hence, for the event $\mathcal{E}$,

$$Q^{\pi}_{r,h}(x,a) + YQ^{\pi}_{g,h}(x,a) \leq Q^{k,Y}_{r,h} + YQ^{k,Y}_{g,h}(x,a) + \mathbb{P}_h(V^{\pi,Y}_{h+1} - V^{k,Y}_{h+1})(x,a)$$

□

B.2 Proof of Lemma 1

First, we show that for a given Y, the gap between the maximum value attained by any composite (estimated) value function and our (estimated) composite value function are close for the parameters $w^{k,Y}_{j,h}$ which we use to show Lemma 9 by controlling the parameter α.

Lemma 13. $\bar{V}^{k,Y}_h(x) - V^{k,Y}_h(x) \leq \dfrac{\log|\mathcal{A}|}{\alpha}$

where

Definition 3. $\bar{V}^{k,Y}_h(\cdot) = \max_a[Q^{k,Y}_{r,h}(\cdot,a) + YQ^{k,Y}_{g,h}(\cdot,a)]$.

$\bar{V}^{k,Y}_h(\cdot)$ is the value function corresponds to the greedy-policy with respect to the composite Q-function attained by our estimated value function.

Proof. Note that

$$V^{k,Y}_h(x) = \sum_a \pi_{h,k}(a|x)[Q^{k,Y}_{r,h}(x,a) + YQ^{k,Y}_{g,h}(x,a)] \tag{31}$$

where

$$\pi_{h,k}(a|x) = \frac{\exp(\alpha[Q^{k,Y}_{r,h}(x,a) + YQ^{k,Y}_{g,h}(x,a)])}{\sum_a \exp(\alpha[Q^{k}_{r,h}(x,a) + YQ^{k,Y}_{g,h}(x,a)])} \tag{32}$$

Denote $a_x = \arg\max_a[Q^{k,Y}_{r,h}(x,a) + YQ^{k,Y}_{g,h}(x,a)]$

Now, recall from Definition 3 that $\bar{V}_h^{k,Y}(x) = [Q_{r,h}^{k,Y}(x,a_x) + YQ_{g,h}^{k,Y}(x,a_x)]$. Then,

$$\begin{aligned}
&\bar{V}_h^{k,Y}(x) - V_h^{k,Y}(x) = [Q_{r,h}^{k,Y}(x,a_x) + YQ_{g,h}^{k,Y}(x,a_x)] \\
&- \sum_a \pi_{h,k}(a|x)[Q_{r,h}^{k,Y}(x,a) + YQ_{g,h}^{k,Y}(x,a)] \\
&\leq \left(\frac{\log(\sum_a \exp(\alpha(Q_{r,h}^{k,Y}(x,a) + YQ_{g,h}^{k,Y}(x,a))))}{\alpha} \right) \\
&- \sum_a \pi_{h,k}(a|x)[Q_{r,h}^{k,Y}(x,a) + YQ_{g,h}^{k,Y}(x,a)] \\
&\leq \frac{\log(|\mathcal{A}|)}{\alpha} \qquad (33)
\end{aligned}$$

where the last inequality follows from Proposition 1 in Pan et al. (2019). □

We are now ready to show Lemma 9.

Proof. We prove the lemma by Induction.

First, we prove for the step H.

Note that $Q_{j,H+1}^{k,Y} = 0 = Q_{j,H+1}^{\pi}$.

Under the event in $\mathcal{E}$ as described in Lemma 10 and from Lemma 11, we have for $j = r, g$,

$$|\langle \phi(x,a), w_{j,H}^{k,Y}(x,a) \rangle - Q_{j,H}^{\pi}(x,a)| \leq \beta \sqrt{\phi(x,a)^T (\Lambda_H^k)^{-1} \phi(x,a)}$$

Hence, for any (x,a),

$$\begin{aligned}
Q_{j,H}^{\pi}(x,a) &\leq \min\{\langle \phi(x,a), w_{j,H}^{k,Y} \rangle + \beta \sqrt{\phi(x,a)^T (\Lambda_H^k)^{-1} \phi(x,a)}, H\} \\
&= Q_{j,H}^{k,Y}(x,a) \qquad (34)
\end{aligned}$$

Hence, from the definition of $\bar{V}_h^{k,Y}$ (Recall Definition 3)

$$\begin{aligned}
\bar{V}_H^{k,Y}(x) &= \max_a [Q_{r,H}^k(x,a) + YQ_{g,H}^k(x,a)] \geq \sum_a \pi(a|x)[Q_{r,H}^{\pi}(x,a) + YQ_{g,H}^{\pi}(x,a)] \\
&= V_H^{\pi,Y}(x) \qquad (35)
\end{aligned}$$

for any policy π. Thus, it also holds for π^*, the optimal policy. Hence, from Lemma 13, we have

$$V_H^{\pi^*,Y}(x) - V_{H,Y}^k(x) \leq \frac{\log(|\mathcal{A}|)}{\alpha}$$

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

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

$$\mathbb{P}_h(V_{h+1}^{\pi,Y} - V_{h+1}^{k,Y})(x,a) \leq \frac{(H-h)\log(|\mathcal{A}|)}{\alpha} \qquad (36)$$

From (30) in Lemma 12 and the above result, we have for any (x,a)

$$Q_{r,h}^{\pi}(x,a) + YQ_{g,h}^{\pi}(x,a) \leq Q_{r,h}^{k,Y}(x,a) + YQ_{g,h}^{k,Y}(x,a) + \frac{(H-h)\log(|\mathcal{A}|)}{\alpha} \qquad (37)$$

Hence,

$$V_h^{\pi,Y}(x) \leq \bar{V}_h^{k,Y}(x) + \frac{(H-h)\log(|\mathcal{A}|)}{\alpha} \qquad (38)$$

Now, again from Lemma 13, we have $\bar{V}_h^{k,Y}(x) - V_h^{k,Y}(x) \leq \frac{\log(|\mathcal{A}|)}{\alpha}$. Thus,

$$V_h^{\pi,Y}(x) - V_h^{k,Y}(x) \leq \frac{(H-h+1)\log(|\mathcal{A}|)}{\alpha} \tag{39}$$

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

$$\left(V_{r,h}^{\pi^*}(x) + YV_{g,h}^{\pi^*}(x)\right) - \left(V_{r,h}^{k,Y}(x) + YV_{g,h}^{k,Y}(x)\right) \leq \frac{(H-h+1)\log(|\mathcal{A}|)}{\alpha} \tag{40}$$

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

C Proof of Lemma 2

Before proving Lemma 2, we prove the following result.

Lemma 14. *Let $V_{g,1}^{k,Y}$ be the estimated value function computed by the Algorithm 1 when the dual-variable is Y, then*

$$|V_{g,1}^{k,Y}(x_1) - V_{g,1}^{k,Y+\eta}(x_1)| \leq \mathcal{O}(H/K)$$

We first show that because of the soft-max property, the difference between $V_{j,h}^{k,Y}(x)$ and $V_{j,h}^{k,Y'}(x)$ is bounded.

Lemma 15. $|V_{j,h}^{k,Y}(x) - V_{j,h}^{k,Y'}(x)| \leq 2\alpha H(H\epsilon'' + \max_Y Y\epsilon' + 2\epsilon')$ *if* $|Y' - Y| \leq \epsilon''$, $|Q_{j,h}^{k,Y}(x,a) - Q_{j,h}^{k,Y'}(x,a)| \leq \epsilon'$ *for all* (x,a), *and the policy is soft-max* π^Y.

Proof. Note that

$$\begin{aligned}
&|Q_{r,h}^{k,Y}(x,a) + Y_k Q_{g,h}^{k,Y}(x,a) - Q_{r,h}^{k,Y'}(x,a) - Y'Q_{g,h}^{k,Y'}(x,a)| \\
&\leq |Q_{r,h}^{k,Y}(x,a) - Q_{r,h}^{k,Y'}(x,a)| + |YQ_{g,h}^{k,Y}(x,a) - Y'Q_{g,h}^{k,Y'}(x,a)| \\
&\leq \epsilon' + Y|Q_{g,h}^{k,Y}(x,a) - Q_{g,h}^{k,Y'}(x,a)| + |Y' - Y|Q_{g,h}^{k,Y'}(x,a) \\
&\leq \epsilon' + Y\epsilon' + H\epsilon''
\end{aligned} \tag{41}$$

Hence, by the property of the soft-max (Theorem 4.4 in Epasto et al. (2020))

$$||\pi^Y - \pi^{Y'}||_1 \leq 2\alpha(\epsilon' + \max Y\epsilon' + H\epsilon'') \tag{42}$$

Now,

$$\begin{aligned}
&|\langle \pi^Y, Q_{j,h}^{k,Y}\rangle - \langle \pi^{Y'}, Q_{j,h}^{k,Y'}\rangle| \\
&= |\langle \pi^Y - \pi^{Y'}, Q_{j,h}^{k,Y}\rangle - \langle \pi^{Y'}, Q_{j,h}^{k,Y} - Q_{j,h}^{k,Y'}\rangle| \\
&\leq ||\pi^Y - \pi^{Y'}||_1 ||Q_{j,h}^{k}||_\infty + ||\pi^{Y'}||_1 ||Q_{j,h}^{k,Y} - Q_{j,h}^{k,Y'}||_\infty \\
&\leq 2H\alpha(\epsilon' + \max Y\epsilon' + H\epsilon'') + \epsilon'
\end{aligned} \tag{43}$$

Since $\alpha H \geq 1$, thus, we have the result. □

Since $\alpha = \log(|\mathcal{A}|)\sqrt{K}$, then, we have from (43), and $Y \leq \sqrt{K}$, thus,

$$|V_{j,h}^{k,Y}(x) - V_{j,h}^{k,Y'}(x)| \leq 4H\sqrt{K}\log(|\mathcal{A}|)(H\epsilon'' + \sqrt{K}\epsilon') \tag{44}$$

We are now ready to prove Lemma 14.

Proof. We prove the above by induction. In particular, we show that if $|Y - Y'| \leq (\log(|\mathcal{A}|))^{-H} K^{-1.5H} (\sqrt{d})^{-H+1}$, then, $|V_{j,h}^{k,Y}(x) - V_{j,h}^{k,Y'}(x)| \leq H \log(|\mathcal{A}|)^{-h+1} (\sqrt{d})^{-h} K^{-(1.5)h-1}$.

First, consider $h = H$. Since $V_{j,H+1}^{k} = 0$, thus, we have

$$Q_{j,H}^{k,Y}(x,a) - Q_{j,H}^{k,Y'}(x,a) = 0 \leq \epsilon'$$

Hence, by Lemma 15 (identifying $\epsilon' = \eta$, and plugging $\alpha = \log(|\mathcal{A}|)/(4H)$) we have

$$\begin{aligned} |V_{j,H}^{k,Y}(x) - V_{j,H}^{k,Y'}(x)| &\leq \log(|\mathcal{A}|)^{-H+1} H (\sqrt{d})^{-H+1} K^{-1.5H+0.5} \\ &= H \log(|\mathcal{A}|)^{-H+1} (\sqrt{d})^{-H+1} K^{-1.5(H-1)-1} \end{aligned} \tag{45}$$

for all x. Hence, the statement is true for $h = H$.

In order to prove this for h, we need to show the following.

Lemma 16. *If $|V_{j,h+1}^{k,Y}(x) - V_{j,h+1}^{k,Y'}(x)| \leq \epsilon$, then $|\phi(x,a)^T w_{j,h}^{k,Y_k} - \phi(x,a)^T w_{j,h}^{k,Y'_k}| \leq \epsilon \sqrt{dk}$*

Proof. We show that for $j = r$. Note that

$$\begin{aligned} w_{r,h}^{k,Y} - w_{r,h}^{k,Y'} &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau [r_h^\tau + V_{r,h+1}^{k,Y}(x_{h+1}^\tau)] \\ &- (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau [r_h^\tau + V_{r,h+1}^{k,Y'}(x_{h+1}^\tau)] \\ &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau [V_{r,h+1}^{k,Y}(x_{h+1}^\tau) - V_{r,h+1}^{k,Y'}(x_{h+1}^\tau)] \end{aligned} \tag{46}$$

Hence,

$$\begin{aligned} |\phi(x,a)^T (w_{r,h}^{k,Y} - w_{r,h}^{k,Y'})| &\leq \phi(x,a)^T (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \epsilon \\ &\leq \epsilon \sqrt{dk} ||\phi(x,a)||_{(\Lambda_h^k)^{-1}} \leq \epsilon \sqrt{dk} \end{aligned} \tag{47}$$

where in the penultimate step we use Lemma 24. In the last inequality, we use the fact that $(\Lambda_h^K) \geq \lambda I$, $||\phi(x,a)|| \leq 1$. Hence, we have

$$|Q_{r,h}^{k,Y}(x,a) - Q_{r,h}^{k,Y'}(x,a)| \leq \epsilon \sqrt{dk} \tag{48}$$

for all (x,a). □

Thus, we have from Lemma 15

$$|V_{j,h}^{k,Y}(x) - V_{j,h}^{k,Y'}(x)| \leq H \log(|\mathcal{A}|) \sqrt{K} (H\epsilon'' + \epsilon' \sqrt{dk} \sqrt{K}) \tag{49}$$

Since the statetemnt is true till $h+1$, thus, $|V_{j,h+1}^{k,Y}(x) - V_{j,h+1}^{k,Y'}(x)| \leq (\log(|\mathcal{A}|)^{-h} H (\sqrt{d})^{-h} K^{-1.5h-1} = \epsilon'$.

$$\epsilon' \sqrt{dk} \sqrt{K} \leq (\log(|\mathcal{A}|)^{-h} (\sqrt{d})^{-h+1} K^{-1.5h} \tag{50}$$

Now, $\epsilon'' = (\log(|\mathcal{A}|))^{-H} (\sqrt{d})^{-H+1} K^{-1.5H}$, Thus, plugging $\alpha = \log(|\mathcal{A}|)/(4H)$, we have for any x,

$$\begin{aligned} &|V_{j,h}^{k,Y}(x) - V_{j,h}^{k,Y'}(x)| \leq \\ &\log(|\mathcal{A}|) \sqrt{K} \left(H (\log(|\mathcal{A}|))^{-H} 4^{-H} H^{-H} (\sqrt{d})^{-H+1} K^{-1.5H} + (\log(|\mathcal{A}|)^{-h} 4^{-h} H^{-h+1} (\sqrt{d})^{-h+1} K^{-1.5h} \right) \\ &= H (\log(|\mathcal{A}|))^{-h+1} (\sqrt{d})^{-h+1} K^{-1.5h+0.5} = H (\log(|\mathcal{A}|))^{-h+1} \sqrt{d}^{-h+1} K^{-1.5(h-1)-1} \end{aligned} \tag{51}$$

Thus, by induction we have the result by plugging in $h = 1$. □

Now, we are ready to prove Lemma 2.

Proof. First, if $Y_k = 0$, the result is trivially true.

Now, consider that the while loop terminates at some value $Y_k > 0$. When $Y_k > 0$ and $k \in \mathcal{I}_b^C$, thus, we know that $V_{g,1}^{k,Y_k}(x_1) \geq b$, however, $V_{g,1}^{k,Y_k-\eta}(x_1) < b$ (otherwise, the loop would have terminated at $Y_k - \eta$). Now, from Lemma 14,

$$\begin{aligned} |V_{g,1}^{k,Y_k}(x_1) - V_{g,1}^{k,Y_k-\eta}(x_1)| &\leq \mathcal{O}(HK^{-1}) \\ V_{g,1}^{k,Y_k}(x_1) &\leq V_{g,1}^{k,Y_k-\eta}(x_1) + \mathcal{O}(HK^{-1}) \\ &\leq b + \mathcal{O}(HK^{-1}) \end{aligned} \tag{52}$$

Since $Y_k \leq \sqrt{K}$, thus,

$$Y_k(b - V_{g,1}^{k,Y_k}(x_1)) \geq \mathcal{O}(HK^{-0.5}) \tag{53}$$

Since Y_k is the selected dual variable at episode k, thus, $V_{g,1}^{k,Y_k}(x) = V_{g,1}^k(x)$ for any x. Hence, the result follows. □

D Proof of Lemmas 4 and 5

D.1 Proof of Lemma 4

Recall the definition of $\mathcal{I}_b$ which is the set of episodes where Algorithm 1 returns $Y_k = \sqrt{K}$ and $V_{g,1}^k(x_1) < b$. Thus, for the episodes k in $\mathcal{I}_b^c$, $(b - V_{g,1}^k(x_1)) \leq 0$. We now, bound $(b - V_{g,1}^k(x_1))_+$ for episodes in $\mathcal{I}_b$.

From Lemma 1 and the value of α, we obtain

$$\begin{aligned} \sum_{k \in \mathcal{I}_b} (V_{r,1}^*(x_1) - V_{r,1}^k(x_1) + Y_k(b - V_{g,1}^k(x_1))_+) &\leq \mathcal{O}(H^2\sqrt{K}) \\ \sum_{k \in \mathcal{I}_b} Y_k(b - V_{g,1}^k(x_1)) &\leq \sum_k (V_{r,1}^k(x_1) - V_{r,1}^*(x_1)) + \mathcal{O}(H^2\sqrt{K}) \\ \sum_{k \in \mathcal{I}_b} (b - V_{g,1}^k(x_1)) &\leq HK/\sqrt{K} + \mathcal{O}(H\sqrt{K})/\sqrt{K} = \mathcal{O}(H^2\sqrt{K}) \end{aligned} \tag{54}$$

where the last inequality follows from the fact that for all the episodes in $\mathcal{I}_b$, $Y_k \geq \sqrt{K}$. Thus,

$$\sum_k (b - V_{g,1}^k(x_1))_+ \leq \mathcal{O}(H^2\sqrt{K}) \tag{55}$$

□

D.2 Proof of Lemma 5

Proof. We prove via induction. Since $V_{j,H+1}^k(x) = 0$ for all x, thus, from Lemma 11

$$Q_{j,H}^{\pi_k}(x,a) \leq \min\{\phi(x,a)^T w_{j,H}^{k,Y_k} + \beta ||\phi(x,a)||_{(\Lambda_H^k)^{-1}}, H\} = Q_{j,H}^{k,Y_k}(x,a) \tag{56}$$

Recall that Y_k is the selected dual variable at episode k. Thus, $Q_{j,h}^{k,Y_k} = Q_{j,h}^k$. Hence,

$$\langle \pi_k, Q_{j,H}^{\pi_k} \rangle \leq \langle \pi_k, Q_{j,H}^k \rangle \tag{57}$$

Hence, $V_{j,H}^{\pi_k}(x) \leq V_{j,H}^k(x)$ for all x.

Suppose that it is true for step $h+1$. Hence, $\mathbb{P}_h(V_{j,h+1}^{\pi_k} - V_{j,h+1}^k)(x,a) \leq 0$. Hence, from Lemma 11,

$$Q_{j,h}^{\pi_k}(x,a) \leq \min\{\phi(x,a)^T w_{j,h}^k + \beta ||\phi(x,a)||_{(\Lambda_h^k)^{-1}}, H\} = Q_{j,h}^k(x,a) \tag{58}$$

Thus,

$$\langle \pi_h^k, Q_{j,h}^{\pi_k} \rangle \leq \langle \pi_h^k, Q_{j,h}^k \rangle \tag{59}$$

Thus, $V_{j,h}^{\pi_k}(x_1) \leq V_{j,h}^k(x_1)$. □

E Proof of Lemmas 3 and 6

In order to prove the Lemma 3 and 6, we state and prove the following result. In this section, we obtain bounds for the selected value Y_k, hence, we use $Q^k_{j,h}$, $w^k_{j,h}$ and $V^k_{j,h}$.

First, we introduce a notation. Let

$$D^k_{j,h,1} = \langle (Q^k_{j,h}(x^k_h,\cdot) - Q^{\pi_k}_{j,h}(x^k_h,\cdot)), \pi_{h,k}(\cdot|x^k_h)\rangle - (Q^k_{j,h}(x^k_h,a^k_h) - Q^{\pi_k}_{j,h}(x^k_h,a^k_h))$$
$$D^k_{j,h,2} = \mathbb{P}_h(V^k_{j,h+1} - V^{\pi_k}_{j,h+1})(x^k_h,a^k_h) - [V^k_{j,h+1} - V^{\pi_k}_{j,h+1}](x^k_{h+1}) \tag{60}$$

Lemma 17. *On the event defined in $\mathcal{E}$ in Lemma 10, we have*

$$V^k_{j,1}(x_1) - V^{\pi_k}_{j,1}(x_1) \le \sum_{h=1}^H (D^k_{j,h,1} + D^k_{j,h,2}) + \sum_{h=1}^H 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} \tag{61}$$

Proof. By Lemma 11, for any x,h,a,k

$$\langle w^k_{j,h}(x,a),\phi(x,a)\rangle + \beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)} - Q^{\pi_k}_{j,h}$$
$$\le \mathbb{P}_h(V^k_{j,h+1} - V^{\pi_k}_{j,h+1})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)} \tag{62}$$

Thus,

$$Q^k_{j,h}(x,a) - Q^{\pi_k}_{j,h}(x,a) \le \mathbb{P}_h(V^k_{j,h+1} - V^{\pi_k}_{j,h+1})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)}$$
$$\mathbb{P}_h(V^k_{j,h+1} - V^{\pi_k}_{j,h+1})(x,a) + 2\beta\sqrt{\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)} - (Q^k_{j,h}(x,a) - Q^{\pi_k}_{j,h}(x,a)) \ge 0 \tag{63}$$

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

Thus, from (63),

$$V^k_{j,h}(x^k_h) - V^{\pi_k}_{j,h}(x^k_h) = \sum_a \pi_{h,k}(a|x^k_h)[Q^k_{j,h}(x^k_h,a) - Q^{\pi_k}_{j,h}(x^k_h,a)]$$
$$\le \sum_a \pi_{h,k}(a|x^k_h)[Q^k_{j,h}(x^k_h,a) - Q^{\pi_k}_{j,h}(x^k_h,a)]$$
$$+ 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} + \mathbb{P}_h(V^k_{j,h+1} - V^{\pi_k}_{j,h+1})(x^k_h,a^k_h) - (Q^k_{j,h}(x^k_h,a^k_h) - Q^{\pi_k}_{j,h}(x^k_h,a^k_h)) \tag{64}$$

Thus, from (64), we have

$$V^k_{j,h}(x^k_h) - V^{\pi_k}_{j,h}(x^k_h) \le D^k_{j,h,1} + D^k_{j,h,2} + [V^k_{j,h+1} - V^{\pi_k}_{j,h+1}](x^k_{h+1}) + 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} \tag{65}$$

Hence, by iterating recursively, we have

$$V^k_{j,1}(x_1) - V^{\pi_k}_{j,1}(x_1) \le \sum_{h=1}^H (D^k_{j,h,1} + D^k_{j,h,2}) + \sum_{h=1}^H 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} \tag{66}$$

The result follows. □

We, are now ready to prove Lemmas 3 and 6.

Proof. Note from Lemma 17, we have

$$\sum_{k=1}^K V^k_{j,1}(x_1) - V^{\pi_k}_{j,1}(x_1) \le \sum_{k=1}^K\sum_{h=1}^H (D^k_{j,h,1} + D^k_{j,h,2}) + \sum_{k=1}^K\sum_{h=1}^H 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)} \tag{67}$$

We, now, bound the individual terms. First, we show that the first term corresponds to a Martingale difference. For any $(k,h) \in [K] \times [H]$, we define $\mathcal{F}^k_{h,1}$ as σ-algebra generated by the state-action sequences, reward, and constraint values, $\{(x^\tau_i, a^\tau_i)\}_{(\tau,i)\in[k-1]\times[H]} \cup \{(x^k_i, a^k_i)\}_{i\in[h]}$.

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

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

$$t(k,h,m) = 2(k-1)H + 2(h-1) + m \tag{68}$$

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

Note from the definitions in (60) that $D^k_{j,h,1} \in \mathcal{F}^k_{h,1}$ and $D^k_{j,h,2} \in \mathcal{F}^k_{h,2}$. Thus, for any $(k,h) \in [K] \times [H]$,

$$\mathbb{E}[D^k_{j,h,1}|\mathcal{F}^k_{h-1,2}] = 0, \quad \mathbb{E}[D^k_{j,h,2}|\mathcal{F}^k_{h,1}] = 0 \tag{69}$$

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

$$\begin{aligned} M^k_{j,h,m} &= \sum_{\tau=1}^{k-1}\sum_{i=1}^{H}(D^\tau_{j,i,1} + D^\tau_{j,i,2}) + \sum_{i=1}^{h-1}(D^k_{j,i,1} + D^k_{j,i,2}) + \sum_{l=1}^{m} D^k_{j,h,l} \\ &= \sum_{(\tau,i,l)\in[K]\times[H]\times[2], t(\tau,i,l)\le t(k,h,m)} D^\tau_{j,i,l} \end{aligned} \tag{70}$$

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

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(D^k_{j,h,1} + D^k_{j,h,2}) = M^K_{j,H,2} \tag{71}$$

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

$$\Pr(M^K_{j,H,2} > s) \le 2\exp(-\frac{s^2}{16TH^2}) \tag{72}$$

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

$$\sum_k \sum_h M^K_{j,H,2} \le \sqrt{16TH^2\log(4/p)} \tag{73}$$

Now, we bound the second term. Note that the minimum eigen value of Λ^k_h is at least $\lambda = 1$ for all $(k,h) \in [K]\times[H]$. By Lemma 22,

$$\sum_{k=1}^{K}(\phi^k_h)^T(\Lambda^k_h)^{-1}\phi^k_h \le 2\log\left[\frac{\det(\Lambda^{k+1}_h)}{\det(\Lambda^1_h)}\right] \tag{74}$$

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

$$\sum_{k=1}^{K}(\phi^k_h)^T(\Lambda^k_h)^{-1}\phi^k_h \le 2d\log\left[\frac{\lambda+k}{\lambda}\right] \le 2d\iota \tag{75}$$

Now, by Cauchy-Schwartz inequality, we have

$$\begin{aligned} \sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{(\phi^k_h)^T(\Lambda^k_h)^{-1}\phi^k_h} &\le \sum_{h=1}^{H}\sqrt{K}[\sum_{k=1}^{K}(\phi^k_h)^T(\Lambda^k_h)^{-1}\phi^k_h]^{1/2} \\ &\le H\sqrt{2dK\iota} \end{aligned} \tag{76}$$

Note that $\beta = C_1 dH\sqrt{\iota}$.

Thus, we have with probability $1 - p/2$,

$$\sum_{k=1}^{K} V_{j,1}^{k}(x_1) - V_{j,1}^{\pi_k}(x_1) \leq [\sqrt{2TH^2 \log(4/p)} + C_4\sqrt{d^3H^3T\iota^2}] \tag{77}$$

Hence, the result follows. □

F Proof of Lemma 10

To simplify the notation, we remove h and Y from the subscript and superscript from $w_{j,h}^{k,Y}$, $Q_{j,h}^{k,Y}$ and $V_{j,h}^{k,Y}$ in this Section.

Note that the proof follows the similar direction as in Ghosh et al. (2022) (Lemma 8 there). However, there is a major difference. In Ghosh et al. (2022), the upper bound of the dual variable Y was $2H/\gamma$ (Appendix J), in our case, the upper bound of Y is $\sqrt{K}$. However, we show that it only adds to the constant term C_1.

In order to prove the Lemma 10, we first compute the ϵ-covering number for the class of value functions (Lemma 18). In order to compute that we first compute the ϵ-covering number of the individual Q-functions (Lemma 19) which is essential to compute the covering number for composite Q-functions (Corollary 1). Subsequently, we show that if the two Q-functions and the Lagrange multipliers are close, the policies are also close (Lemma 20).

We first introduce the set of Q-functions.

Definition 4. *Let* $\mathcal{Q}_j = \{Q | Q(\cdot,\cdot) = \min\{w_j^T \phi(\cdot,\cdot) + \beta\sqrt{\phi^T(\cdot,\cdot)^T \Lambda^{-1} \phi(\cdot,\cdot)}, H\}\}$

The set $\mathcal{Q}$ is parameterized by w_j, and Λ. We have $||w_j|| \leq 2H\sqrt{dk/\lambda}$ (from Lemma 8). The minimum eigen value of Λ satisfies $\lambda_{min} \geq 1$. Hence, the Frobenius norm of Λ^{-1} is bounded. Note that $Q_j^k \in \mathcal{Q}_j$ for $j = r, g$.

We now introduce the class of value function for $j = r, g$.

Definition 5. *Let* $\mathcal{V}_j = \{V_j | V_j(\cdot) = \sum_a \pi(a|\cdot) Q_j(\cdot, a); Q_r \in \mathcal{Q}_r, Q_g \in \mathcal{Q}_j, Y \in [0, \xi]\}$ *for* $j = r, g$, *where*

$$\Pi = \{\pi | \forall a \in \mathcal{A}, \pi(a|\cdot) = \textsc{Soft-Max}_\alpha^a((Q_r(\cdot,\cdot) + YQ_g(\cdot,\cdot))Q_r \in \mathcal{Q}_r, Q_g \in \mathcal{Q}_g, Y \in [0,\xi]\}.$$

where $\xi = \sqrt{K}$.

The class of value function $\mathcal{V}_j$ is parameterized by w_r, w_g, Λ, and $Y \in [0, \xi]$. Note that even the individual value function depends on the Q-functions for both the reward and utility since the policy depends on the composite Q-function.

First, we need to see whether $V_j^k \in \mathcal{V}_j$. Recall the definition of V_j^k at the k-th episode $V_j^k(\cdot) = \sum_a \pi_k(a|\cdot) Q_j^k(\cdot, a)$ where

$$\pi_k(a|\cdot) = \textsc{Soft-Max}_\alpha^a((Q_r(\cdot,\cdot) + Y_k Q_g(\cdot,\cdot)).$$

Since $Q_j \in \mathcal{Q}_j$ for all j, and $0 \leq Y_k \leq \xi$, thus, $V_j \in \mathcal{V}_j$.

We now bound the ϵ-covering number for the class of value function

Lemma 18. *There exists a* $\tilde{V}_j \in \mathcal{V}_j$ *parameterized by* $(\tilde{w}_r, \tilde{w}_g, \tilde{\beta}, \Lambda, \tilde{Y})$ *such that DIST* $(V_j, \tilde{V}_j) \leq \epsilon$ *where*

$$\text{DIST}(V_j, \tilde{V}_j) = \sup_x |V_j(x) - \tilde{V}_r(x)|. \tag{78}$$

Let $N_\epsilon^{V_j}$ *be the* ϵ*-covering number for the set* $\mathcal{V}_j$, *then,*

$$\log N_\epsilon^{V_j} \leq d \log\left(1 + 8H\frac{\sqrt{dk}}{\sqrt{\lambda}\epsilon'}\right) + d^2 \log\left[1 + 8d^{1/2}\beta^2/(\lambda(\epsilon')^2)\right] + \log\left(1 + \frac{\xi}{\epsilon'}\right) \tag{79}$$

where $\epsilon' = \dfrac{\epsilon}{H2\alpha(1 + \xi + H) + 1}$ *where* $\xi = 2\sqrt{K}$

Note that ϵ-covering number is dependent on ξ, the upper bound of Y_k. This is because the policy depends on the Lagrange multiplier Y which is upper bounded by ξ. Thus, we also need ϵ-covering for the Lagrange multiplier in order to obtain ϵ-close value function. In Ghosh et al. (2022), the upper bound was $2H/\gamma$ (see Appendix J). However, in our case, the upper bound is $\sqrt{K}$. Note that log of epsilon-covering number only scales as $\log(K)$ even when the upper bound of ξ is $\sqrt{K}$.

Note that the ϵ-covering does not depend on sample dependent terms. Rather it only depends on general $w_{j,h}$, Λ, and Y. Since the policy parameter is α, we also have ϵ-covering number is dependent on α.

In order to prove the above lemma, we first state and prove some additional results.

We, first, obtain the $N_\epsilon^{Q_j}$ covering number for the set $\mathcal{Q}_j$. Towards this end, we first, introduce some notations.

Definition 6. *Let $\mathcal{C}_w^\epsilon$ be an $\epsilon/2$- cover of the set $\{w \in \mathbb{R}^d | ||w|| \leq 2H\sqrt{dk/\lambda}\}$ with respect to the 2-norm. Let $\mathcal{C}_{\mathbf{A}}^\epsilon$ be an $\epsilon^2/4$-cover of the set $\{\mathbf{A} \in \mathbb{R}^{d\times d} | ||\mathbf{A}||_F \leq d^{1/2}\beta^2\lambda^{-1}\}$ with respect to the Frobenius norm.*

Lemma 19.

$$|\mathcal{C}_w^\epsilon| \leq (1 + 8H\sqrt{dk/\lambda}/\epsilon)^d, \quad |\mathcal{C}_{\mathbf{A}}^\epsilon| \leq [1 + 8d^{1/2}\beta^2/(\lambda\epsilon^2)]^{d^2} \tag{80}$$

The ϵ-covering number for the set $\mathcal{Q}_j$, for $j = r, g$, $N_\epsilon^{Q_j}$ of the set $\mathcal{Q}_j$ for $j = r, g$ satisfies the following

$$\log N_\epsilon^{Q_j} \leq d \log\left(1 + \frac{8H\sqrt{dk}}{\sqrt{\lambda}\epsilon}\right) + d^2 \log[1 + 8d^{1/2}\beta^2/(\lambda\epsilon)^2] \tag{81}$$

The distance metric is the ∞-norm, i.e., $\text{dist}(Q_1, Q_2) = \sup_{x,a}|Q_1(x,a) - Q_2(x,a)|$.

Proof. For notational simplicity, we represent $\mathbf{A} = \beta^2\Lambda^{-1}$, and reparamterized the class $\mathcal{Q}_j$ by $(w_j, \mathbf{A})$. Now,

$$\begin{aligned}
dist(Q_1, Q_2) &= \sup_{x,a} |[w_1^T\phi(x,a) + \sqrt{\phi^T(x,a)\mathbf{A}_1\phi(x,a)}] - [w_2^T\phi(x,a) + \sqrt{\phi^T(x,a)\mathbf{A}_2\phi(x,a)}]| \\
&\leq \sup_{\phi:||\phi||\leq 1} |[w_1^T\phi + \sqrt{\phi^T\mathbf{A}_1\phi}] - [w_2^T\phi + \sqrt{\phi^T\mathbf{A}_2\phi}]| \\
&\leq \sup_{\phi:||\phi||\leq 1} |(w_1 - w_2)^T\phi| + \sup_{\phi:||\phi||\leq 1} \sqrt{|\phi^T(\mathbf{A}_1 - \mathbf{A}_2)\phi|} \\
&= ||w_1 - w_2|| + \sqrt{||\mathbf{A}_1 - \mathbf{A}_2||} \leq ||w_1 - w_2|| + \sqrt{||\mathbf{A}_1 - \mathbf{A}_2||_F}
\end{aligned} \tag{82}$$

where the second-last inequality follows from the fact that $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x - y|}$. For matrices $||\cdot||$, and $||\cdot||_F$ denote matrix operator norm and the Frobenius norm respectively.

Recall that $\mathcal{C}_w$ is an $\epsilon/2$- cover of the set $\{w \in \mathbb{R}^d | ||w|| \leq 2H\sqrt{dk/\lambda}\}$ with respect to the 2-norm. Also recall that $\mathcal{C}_{\mathbf{A}}$ be an $\epsilon^2/4$-cover of the set $\{\mathbf{A} \in \mathbb{R}^{d\times d} | ||\mathbf{A}||_F \leq d^{1/2}\beta^2\lambda^{-1}\}$. Thus, from Lemma 23,

$$|\mathcal{C}_w^\epsilon| \leq (1 + 8H\sqrt{dk/\lambda}/\epsilon)^d, \quad |\mathcal{C}_{\mathbf{A}}^\epsilon| \leq [1 + 8d^{1/2}\beta^2/(\lambda\epsilon^2)]^{d^2}$$

For any $Q_j \in \mathcal{Q}_j$, there exists a $\tilde{Q}_j$ parameterized by $(w_2, \mathbf{A}_2)$ where $w_2 \in \mathcal{C}_w^\epsilon$ and $\mathbf{A}_2 \in \mathcal{C}_{\mathbf{A}}^\epsilon$ such that $\text{dist}(Q_j, \tilde{Q}_j) \leq \epsilon$. Hence, $N_\epsilon^{Q_j} \leq |\mathcal{C}_w^\epsilon||\mathcal{C}_{\mathbf{A}}^\epsilon|$, which gives the result since $\log(\cdot)$ is an increasing function. □

Since the class of Q-function is independent of the policy we do not have ξ and α in the ϵ-covering number.

From the above lemma and since $Y_k \leq \xi$, we have the following,

Corollary 1. *If $\text{dist}(Q_r^k, \tilde{Q}_r) \leq \epsilon'$, $\text{dist}(Q_g^k, \tilde{Q}_g) \leq \epsilon'$, and $|\tilde{Y}_k - Y_k| \leq \epsilon'$, then, $\text{dist}(Q_r^k + Y_k Q_g^k, \tilde{Q}_r + \tilde{Y}_k\tilde{Q}_g) \leq \epsilon'(1 + \xi + H)$.*

Proof. Note that $\tilde{Q}_j \in \mathcal{Q}_j$ belongs to the ϵ' covering of the set $\mathcal{Q}$.

$$\begin{aligned}
&\text{dist}(Q_r^k + Y_k Q_g^k, \tilde{Q}_r + \tilde{Y}_k \tilde{Q}_g) = \sup_{x,a} |(Q_r^k(x,a) + Y_k Q_g^k(x,a)) - (\tilde{Q}_r(x,a) + \tilde{Y}_k \tilde{Q}_g(x,a))| \\
&\leq \sup_{x,a} |(Q_r^k(x,a) + Y_k Q_g^k(x,a)) - (\tilde{Q}_r(x,a) + Y_k \tilde{Q}_g(x,a))| + \sup_{x,a} |(\tilde{Y}_k - Y_k) Q_g^k(x,a)| \\
&\leq \sup_{x,a} |Q_r^k(x,a) - \tilde{Q}_r(x,a)| + Y_k \sup_{x,a} |Q_g^k(x,a) - \tilde{Q}_g(x,a)| + \epsilon' H \\
&\leq \epsilon'(1 + Y_k) + \epsilon' H \\
&\leq \epsilon'(1 + H + \xi)
\end{aligned} \tag{83}$$

where the first inequality follows from the property of supremum and the norm. The second inequality follows from the norm, and the fact that $|\tilde{Y}_k - Y_k| \leq \epsilon'$, and $|Q_g^k(x,a)| \leq H$. The third inequality follows from the fact that $\text{dist}(Q_j, \tilde{Q}_j) \leq \epsilon'$. □

We now show that if the there exist $\tilde{Q}_j$, and $\tilde{Y}_k$ which are close to Q_j and Y_k, then the soft-max policy is also close.

Lemma 20. *Suppose that π is the soft-max policy (temp. coefficient $1/\alpha$) corresponding to the composite Q-functions $(Q_r^k + Y_k Q_g^k)$, i.e., $\forall a \in \mathcal{A}$*

$$\pi(a|\cdot) = \textsc{Soft-Max}_\alpha^a((Q_r(\cdot,\cdot) + Y_k Q_g(\cdot,\cdot)).$$

$\tilde{\pi}$ is the soft-max policy vector with the same temp. coefficient $1/\alpha$ corresponding to the composite Q-function $(\tilde{Q}_r + \tilde{Y}_k \tilde{Q}_g)$, i.e, $\forall a \in \mathcal{A}$,

$$\tilde{\pi}(a|\cdot) = \textsc{Soft-Max}_\alpha^a((\tilde{Q}_r(\cdot,\cdot) + \tilde{Y}_k \tilde{Q}_g(\cdot,\cdot)).$$

then, for any state x,

$$||\pi(\cdot|x) - \tilde{\pi}(\cdot|x)||_1 \leq 2\alpha\epsilon'(1 + \xi + H) \tag{84}$$

where $\pi(\cdot|x) = \{\pi(a|x)\}_{a\in\mathcal{A}}$ and $\tilde{\pi}(\cdot|x) = \{\tilde{\pi}(a|x)\}_{a\in\mathcal{A}}$ when $\text{dist}(Q_{j,h}^k, \tilde{Q}_j) \leq \epsilon'$ for $j = r, g$, and $|\tilde{Y}_k - Y_k| \leq \epsilon'$.

Proof. Let $\text{Exp}^\alpha(P)$ be a soft-max corresponding to the vector P, i.e., the i-th component of $\text{Exp}^\alpha(P)$ is

$$\frac{\exp(\alpha P_i)}{\sum_i \exp(\alpha P_i)}.$$

Note from Theorem 4.4 in Epasto et al. (2020) then, we have

$$||\text{Exp}^\alpha(P_1) - \text{Exp}^\alpha(P_2)||_1 \leq 2\alpha ||P_1 - P_2||_\infty \tag{85}$$

for two vectors P_1 and P_2.

Now note that in our case for a given state x, π is equivalent to $\text{Exp}^\alpha(Q_{r,h}^k(x,\cdot) + Y_k Q_{g,h}^k(x,\cdot))$, and $\tilde{\pi}$ is equivalent to $\text{Exp}^\alpha(\tilde{Q}_r(x,\cdot) + \tilde{Y}_k \tilde{Q}_g(x,\cdot))$. Then from (85) and the fact that $\text{dist}(Q_{r,h}^k + Y_k Q_{g,h}^k, \tilde{Q}_r + \tilde{Y}_k \tilde{Q}_g) \leq \epsilon'(1 + \xi + H)$ (by Corollary 1) we have

$$||\pi(\cdot|x) - \tilde{\pi}(\cdot|x)||_1 \leq 2\alpha\epsilon'(1 + \xi + H) \tag{86}$$

Hence, the result follows. □

Based on the above two lemmas we show that when the Q-functions are close, the value functions in the class $\mathcal{V}_j$ are also close.

Lemma 21. *There exists $\tilde{V}_j \in \mathcal{V}_j$ such that*

$$\text{DIST}(V_j^k, \widetilde{V}_j) \leq H2\alpha\epsilon'(1+\xi+H) + \epsilon', \tag{87}$$

where $\text{dist}(\tilde{Q}_j, Q_j) \leq \epsilon'$, $\tilde{Q}_j \in \mathcal{Q}_j$ *for all* j*;*

$$\widetilde{V}_j(\cdot) = \sum_a [\tilde{\pi}(a|\cdot)\tilde{Q}_j(\cdot)],$$

$$\tilde{\pi}(a|\cdot) = \text{Soft-Max}_\alpha^a((\tilde{Q}_r(\cdot,\cdot) + \tilde{Y}_k\tilde{Q}_g(\cdot,\cdot)), \quad \forall a \in \mathcal{A}$$

$|\tilde{Y}_k - Y_k| \leq \epsilon'$.

Proof. For any x,

$$\begin{aligned}
&V_j^k(x) - \widetilde{V}_j(x) \\
&= |\sum_a \pi(a|x)Q_j^k(x,a) - \sum_a \tilde{\pi}(a|x)\tilde{Q}_j(x,a)| \\
&= |\sum_a \pi(a|x)Q_j^k(x,a) - \sum_a \pi(a|x)\tilde{Q}_j(x,a) + \sum_a \pi(a|x)\tilde{Q}_j(x,a) - \sum_a \tilde{\pi}(a|x)\tilde{Q}_j(x,a)| \\
&\leq |\sum_a \pi(a|x)Q_j^k(x,a) - \sum_a \pi(a|x)\tilde{Q}_j(x,a)| + |\sum_a \pi(a|x)\tilde{Q}_j(x,a) - \sum_a \tilde{\pi}(a|x)\tilde{Q}_j(x,a)| \\
&\leq \epsilon' + ||\pi(\cdot|x) - \tilde{\pi}(\cdot|x)||_1 ||\tilde{Q}_j(x)||_\infty \\
&\leq \epsilon' + H2\alpha\epsilon'(1+\xi+H)
\end{aligned} \tag{88}$$

where we use the fact that $\text{dist}(Q_j^k, \tilde{Q}_r) \leq \epsilon'$, and $\sum_a \pi(a|x) = 1$ for the first term and the Holder's inequality in the second term for the second last inequality. For the last inequality, we use Lemma 20, and the fact that $\tilde{Q}_j(x,a) \leq H$ for any (x,a). Hence, we have the result. □

Note that when $\alpha = \log(|\mathcal{A}|)\sqrt{K}/H$ as we have in Algorithm 1, the right hand side in (87) becomes

$$\epsilon' + \log(|\mathcal{A}|)\sqrt{K}\epsilon'(1+\xi+H) \tag{89}$$

We introduce one more notation which we use to prove Lemma 18.

Definition 7. *Let $\mathcal{C}_\xi^\epsilon$ be an ϵ cover for $Y \in [0,\xi]$. Hence,* $|\mathcal{C}_\xi^\epsilon| \leq \left(1 + \frac{\xi}{\epsilon}\right)$

Note that $\mathcal{C}_\xi^\epsilon$ consists of points which is ϵ-close to any point within the interval $[0,\xi]$. Since we have defined ϵ-cover for all the parameters, we are now ready to prove Lemma 18.

Proof. Fix an ϵ. Let $\epsilon' = \dfrac{\epsilon}{H2\alpha(1+\xi+H)+1}$, then from Lemma 21, we have $\text{DIST}(V_j^k, \widetilde{V}_j) \leq \epsilon$. Thus, we only need to find parameters in the ϵ'-covering of the Q-functions as described in Lemma 19 in order to obtain ϵ-close value function.

Recall the Definition 6. Then, there exists $\tilde{w}_r, \tilde{w}_g \in \mathcal{C}_w^{\epsilon'}$ such that $||\tilde{w}_r - w_r|| \leq \dfrac{\epsilon'}{2}$, $||\tilde{w}_g - w_g|| \leq \dfrac{\epsilon'}{2}$. Further, there exists $\mathbf{A}_2 \in \mathcal{C}_{\mathbf{A}}^{\epsilon'}$ such that $||\mathbf{A} - \tilde{\mathbf{A}}||_F \leq \dfrac{\epsilon'^2}{4}$, $\mathbf{A} = \beta^2(\Lambda^k)^{-1}$, $\tilde{\mathbf{A}} = \beta^2(\tilde{\Lambda})^{-1}$, for some $\tilde{\Lambda}$, and $Y_k, \tilde{Y}_k$ such that $|Y_k - \tilde{Y}_k| \leq \epsilon'$. Then we obtain $\tilde{Q}_j$ parameterized by $(\tilde{w}_j, \beta, \tilde{\Lambda})$ for $j = r, g$, such that $\text{dist}(Q_j, \tilde{Q}_j) \leq \epsilon'$ (by Lemma 19).

Now define $\widetilde{V}_j = \sum_a \tilde{\pi}(a|\cdot)\tilde{Q}_j$, where

$$\tilde{\pi}(a|\cdot) = \text{Soft-Max}_\alpha^a((\tilde{Q}_r(\cdot,\cdot) + \tilde{Y}_k\tilde{Q}_g(\cdot,\cdot)).$$

Thus, from Lemma 21, we have $\text{DIST}(V_j^k, \tilde{V}_j) \leq \epsilon$. Hence, there exists $\tilde{V}_j$ parameterized by $\tilde{w}_r, \tilde{w}_g, \tilde{Y}_k, \tilde{\mathbf{A}}$, such that $\text{Dist}(\tilde{V}_j, V_j^k) \leq \epsilon$. Hence, $N_\epsilon^V \leq |\mathcal{C}_w^{\epsilon'}||\mathcal{C}_{\mathbf{A}}^{\epsilon'}||\mathcal{C}_\xi^{\epsilon'}|$. Thus, from Lemma 19 and Definition 7, the ϵ-covering number $N_\epsilon^{V_j}$ for the set $\mathcal{V}_j$ satisfies the following

$$\log N_\epsilon^{V_j} \leq d \log\left(1 + 8H\frac{\sqrt{dk}}{\sqrt{\lambda}\epsilon'}\right) + d^2 \log\left[1 + 8d^{1/2}\beta^2/(\lambda(\epsilon')^2)\right] + \log\left(1 + \frac{\xi}{\epsilon'}\right).$$

Hence, the result follows. □

From Lemma 18, note that we need ϵ' covering for the Q-functions where $\epsilon' = \dfrac{\epsilon}{(H2\alpha(1+\xi)+1)}$ if we need to bound DIST $(V_j, \tilde{V}_j)$ by ϵ.

Now, we are ready to prove Lemma 10.

Proof. By Lemma 18, we know that there exists $\tilde{V}_j$ in the ϵ-covering for $\mathcal{V}_j$ such that for every x,

$$V_j(x) = \tilde{V}_j(x) + \Delta V(x) \tag{90}$$

where $\sup_x \Delta V(x) \leq \epsilon$.

Hence,

$$\begin{aligned}\left\|\sum_{\tau=1}^k \phi^\tau (V_j(x_\tau) - \mathbb{E}[V_j(x_\tau)|\mathcal{F}_{\tau-1}])\right\|_{(\Lambda^k)^{-1}}^2 &\leq 2\left\|\sum_{\tau=1}^k \phi^\tau (\tilde{V}_j(x_\tau) - \mathbb{E}[\tilde{V}_j(x_\tau)|\mathcal{F}_{\tau-1}])\right\|_{(\Lambda^k)^{-1}}^2 \\ &+ 2\left\|\sum_{\tau=1}^k \phi^\tau (\Delta V(x_\tau) - \mathbb{E}[\Delta V(x_\tau)|\mathcal{F}_{\tau-1}])\right\|_{(\Lambda^k)^{-1}}^2 \end{aligned} \tag{91}$$

The last expression is bounded by $\dfrac{8k^2\epsilon^2}{\lambda}$.

Now, we bound the first term. Note from Lemma 18 that in order to obtain $\tilde{V}_j$ which satisfies (90), we need to obtain we need N_ϵ^V number of elements to obtain such $(\tilde{w}_r, \tilde{w}_g, \beta, \tilde{\Lambda}, \tilde{Y})$. Such $\tilde{V}_j$ is independent of samples. Hence, we can use the Elliptical lemma for self-normalization (Theorem 2). From Theorem 2 and the union bound we obtain

$$\left\|\sum_{\tau=1}^k \phi^\tau (\tilde{V}_j(x_\tau) - \mathbb{E}[\tilde{V}_j(x_\tau)|\mathcal{F}_{\tau-1}])\right\|_{(\Lambda^k)^{-1}}^2 \leq 2H^2\left[d\log\left(\frac{k+\lambda}{\lambda}\right) + \log\left(\frac{N_\epsilon^V}{\delta}\right)\right] \tag{92}$$

where N_ϵ^V is upper bounded in (79). β is equal to $C_1 dH\sqrt{\iota}$ for some constant C_1, and $\iota = \log(\log(|\mathcal{A}|)4dT/p)$. Further, $\xi = \sqrt{K}$. We obtain from (92)

$$\begin{aligned}&\left\|\sum_{\tau=1}^k \phi^\tau (\tilde{V}_j(x_\tau) - \mathbb{E}[\tilde{V}_j(x_\tau)|\mathcal{F}_{\tau-1}])\right\|_{(\Lambda^k)^{-1}}^2 \leq \\ &4H^2\left[\frac{d}{2}\log\left(\frac{k+\lambda}{\lambda}\right) + d\log\left(1 + \frac{8H\sqrt{dk}}{\epsilon'\sqrt{\lambda}}\right) + d^2\log\left(1 + \frac{8d^{1/2}\beta^2}{\epsilon'^2\lambda}\right) + \log\left(1 + \frac{\sqrt{K}}{\epsilon'}\right) + \log\left(\frac{4}{p}\right)\right]\end{aligned} \tag{93}$$

where $\epsilon' = \dfrac{\epsilon}{(H2\alpha(1+\xi+H)+1)}$. Set $\epsilon = \dfrac{dH}{k}$, $\lambda = 1$. Thus, $\epsilon' = \dfrac{dH}{(2H\alpha(1+\xi+H)+1)k}$. Plugging in the above, and putting $\alpha = \log(|\mathcal{A}|)\sqrt{K}/H$, we obtain from (93)

$$||\sum_{\tau=1}^k \phi^\tau (\tilde{V}_j(x_\tau) - \mathbb{E}[\tilde{V}_j(x_\tau)|\mathcal{F}_{\tau-1}])||_{\Lambda_k^{-1}}^2 \leq C_2 H^2 d^2 \log\left(\frac{4(C_1+1)\log(|\mathcal{A}|)dT}{p}\right) \tag{94}$$

for some constant C_2. Hence, the result follows.

□

G Supporting Results

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

Lemma 22. *Let $\{\phi_t\}_{t\geq 0}$ be a sequence in $\Re^d$ satisfying $\sup_{t\geq 0}||\phi_t||\leq 1$. For any $t \geq 0$, we define $\Lambda_t = \Lambda_0 + \sum_{j=0}^{t} \phi_j \phi_j^T \phi_j$. Then if the smallest eigen value of Λ_0 be at least 1, we have*

$$\log\left[\frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^1)}\right] \leq \sum_{k=1}^{K} (\phi_h^k)^T (\Lambda_h^k)^{-1} \phi_h^k \leq 2\log\left[\frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^1)}\right] \tag{95}$$

Theorem 2. *[Concentration of Self-Normalized Process Abbasi-Yadkori et al. (2011)] Let $\{\epsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let $\epsilon_t|\mathcal{F}_{t-1}$ be a zero mean and σ sub-Gaussian, i.e., $\mathbb{E}[\epsilon_t|\mathcal{F}_{t-1}] = 0$, and*

$$\forall \zeta \in \Re, \quad \mathbb{E}[e^{\zeta \epsilon_t}|\mathcal{F}_{t-1}] \leq e^{\zeta^2\sigma^2/2}. \tag{96}$$

Let $\{\phi_t\}_{t=1}^{\infty}$ be a $\Re^d$-valued Stochastic process where $\phi_t \in \mathcal{F}_{t-1}$. Assume $\Lambda_0 \in \Re^{d\times d}$ is a positive-define matrix, let, $\Lambda_t = \Lambda_0 + \sum_{j=0}^{t} \phi_j \phi_j^T$. Then for any $\delta > 0$ with probability at least $1-\delta$, we have

$$||\sum_{s=1}^{t} \phi_s \epsilon_s||^2_{\Lambda_t^{-1}} \leq 2\sigma^2 \log\left[\frac{\det(\Lambda_t)^{1/2}\det(\Lambda_0)^{-1/2}}{\delta}\right] \tag{97}$$

The next result characterizes the covering number of an Euclidean ball (Lemma 5.2 in Vershynin (2010)).

Lemma 23. *[Covering Number of Euclidean Ball] For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius R is upper bounded by $(1 + 2R/\epsilon)^d$.*

The following lemma is similar to Lemma C.4 in Jin et al. (2020).

Lemma 24. *Let $\{\epsilon_\tau\}$ be any sequence so that $|\epsilon_\tau| \leq B$ for any τ. Then, we have for any $(h,k) \in [H]\times[K]$ and any $\phi \in \Re^d$:*

$$|\phi^T(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi_h^\tau \epsilon_\tau| \leq B\sqrt{k}||\phi||_{(\Lambda_h^k)^{-1}}$$

H Tabular Setup

In this section, we describe the tabular setup in detail. First, we describe a structure for the tabular setup where one does not need to take the inverse of Λ_h^k. Further, we note that $\eta = \mathcal{O}(K^{-H})$ is enough rater $\eta = \mathcal{O}(K^{-1.5H})$.

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

$$w_{r,h}^{k,Y}(x,a) = \frac{1}{(n_h^k(x,a)+\lambda)} \sum_{\tau=1}^{n_h^k(x,a)} (r_h(x_h^\tau, a_h^\tau) + V_{r,h+1}^{k,Y}(x_{h+1}^\tau))$$

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

$$Q_{r,h}^{k,Y}(x,a) = \min\{\langle w_{r,h}^{k,Y}(x,a), \phi(x,a)\rangle + \beta\sqrt{1/(n_h^k(x,a)+\lambda)}, H\}.$$

In a similar manner, we can update $Q^k_{g,h}$.

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

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

In this case, we do not need to go through all samples at each iteration and do not even need to store the old samples. The memory complexity of maintaining the counts $\{n_h(x,a,\tilde{x})\}$ is $O\left(H|\mathcal{S}|^2|\mathcal{A}|\right)$, which matches model-based algorithms for tabular settings such as Efroni et al. (2020).

It is clear that if $|V^{k,Y_k}_{r,h+1}(x_{h+1}) - V^{k,Y'_k}_{r,h+1}(x_{h+1})| \leq \epsilon$ then $|w^{k,Y_k}_{r,h}(x,a) - w^{k,Y'_k}_{r,h}(x,a)| \leq \epsilon$ for any (x,a).

Lemma 25. *$|V^{k,Y_k}_{g,1}(x_1) - V^{k,Y'_k}_{g,1}(x_1)| \leq \mathcal{O}(K^{-1})$, if $|Y_k - Y'_k| \leq \frac{1}{\log(|\mathcal{A}|)^H K^{H+1}}$.*

Proof. We prove the result via induction. For step $h = H$, $V^k_{j,H+1}(x) = 0$, hence, $Q^{k,Y_k}_{j,H}(x,a) = Q^{k,Y'_k}_{j,H}(x,a)$, thus, from Lemma 15,

$$|V^{k,Y_k}_{j,H}(x) - V^{k,Y'_k}_{j,H}(x)| \leq 2H^2\alpha \frac{1}{H\log(|\mathcal{A}|)K^{H+1}} = \frac{1}{\log(|\mathcal{A}|)^{H-1}K^{H+0.5}} \tag{98}$$

where we have used the fact that $\epsilon'' = \frac{1}{\log(|\mathcal{A}|)^H K^{H+1}}$, and $\alpha = \log(|\mathcal{A}|)\sqrt{K}/(4H)$. Now, let us assume that the result is true for $h+1$. We have $|V^{k,Y_k}_{j,h+1}(x) - V^{k,Y'_k}_{j,h+1}(x)| \leq \frac{1}{K^{h+1}}$. Thus, $|Q^{k,Y_k}_{j,h}(x,a) - Q^{k,Y'_k}_{j,h}(x,a)| \leq \frac{1}{\log(|\mathcal{A}|)^h K^{h+1}}$. Hence, we have from Lemma 15

$$\begin{aligned} |V^k_{j,h}(x) - V^k_{j,h}(x)| &\leq 2H\alpha \left(H\epsilon'' + \sqrt{K}\frac{1}{K^{h+1}} \right) \\ &\leq \mathcal{O}\left(\frac{1}{\log(|\mathcal{A}|)^{h-1}K^h} \right) \end{aligned} \tag{99}$$

where $\epsilon'' = \frac{1}{\log(|\mathcal{A}|)^H K^{H+1}}$, and $\alpha = \sqrt{K}/(4H)$ Hence, the result follows after putting $h = 1$. □

Hence, for tabular case, $\eta = \mathcal{O}(1/K^{H+1})$ is enough. Thus, the maximum number of times the while loop may continue is $\mathcal{O}(K^{H+1.5})$ since $Y_k \leq \sqrt{K}$.

Improved Bounds for Tabular Case Using the finite state-space, we obtain a better bound for the regret and hard constraint violation.

Theorem 3. *Fix any $p \in (0,1)$. If we set $\lambda = 1$, $\beta = C_4\sqrt{|S|\log(4|S||A|\log(|A|)T/p)}$ for tabular case in Algorithm 1 for some absolute constant C_4. With probability $1-2p$, we have*

$$\begin{aligned} &\text{Regret}(K) \leq C(\sqrt{|S|^2|A|H^3T}\log(4|S||A|\log(|A|)T/p)), \\ &\text{Violation}_H(K) \leq C'\sqrt{|S|^2|A|H^3T}\log(4|S||A|\log(|A|)T/p) \end{aligned}$$

for some absolute constants C, and C'.

Proof. In order to prove the above result, we show that new bonus term is enough for optimisim.

Let us recall the difference $Q^k_{r,h}(x,a) - Q^\pi_{r,h}(x,a) = \phi(x,a)^T[w^k_{r,h} - w^\pi_{r,h}]$ for any policy π. Recall from (22) that

$$\begin{aligned} w^k_{r,h} - w^\pi_{r,h} &= -\lambda(\Lambda^k_h)^{-1}(w^\pi_{r,h}) + (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[V^k_{r,h+1}(x^\tau_{h+1}) - \mathbb{P}_h V^k_{r,h+1}(x^\tau_h,a^\tau_h)] \\ &+ (\Lambda^k_h)^{-1}\sum_{\tau=1}^{k-1}\phi^\tau_h[\mathbb{P}_h V^k_{r,h+1}(x^\tau_h,a^\tau_h) - \mathbb{P}_h V^\pi_{r,h+1}(x^\tau_h,a^\tau_h)] \end{aligned} \tag{100}$$

Now, note that $w^{\pi}_{r,h} = \theta_{j,h} + \sum_{s'} \mathbb{P}_h V^{\pi}_{r,h+1}(s')$ for tabular case. Hence $||w^{\pi}_{r,h}|| \leq H\sqrt{|S|}$. For the third term, note that

$$\begin{aligned}
&\langle \phi(x,a), (\Lambda^k_h)^{-1} \sum_{\tau=1}^{k-1} \phi^{\tau}_h [\mathbb{P}_h (V^k_{r,h+1} - V^{\pi}_{r,h+1})(x^{\tau}_h, a^{\tau}_h)] \rangle \\
&= \langle \phi(x,a), (\Lambda^k_h)^{-1} \sum_{\tau=1}^{k-1} \phi^{\tau}_h (\phi^{\tau}_h)^T \int (V^k_{r,h+1} - V^{\pi}_{r,h+1})(x') d\mu_h(x') \rangle \\
&= \langle \phi(x,a), \int (V^k_{r,h+1} - V^{\pi}_{r,h+1})(x') d\mu_h(x') \rangle \\
&- \langle \phi(x,a), \lambda (\Lambda^k_h)^{-1} \int (V^k_{r,h+1} - V^{\pi}_{r,h+1})(x') d\mu_h(x') \rangle
\end{aligned} \tag{101}$$

The last term in (101) can be bounded as the following

$$|\langle \phi(x,a), \lambda (\Lambda^k_h)^{-1} \sum_{x'} \mathbb{P}_h(x'|x,a)(V^k_{r,h+1} - V^{\pi}_{r,h+1})(x') \rangle| \leq 2H\sqrt{|S|} \sqrt{\phi(x,a)^T (\Lambda^k_h)^{-1} \phi(x,a)} \tag{102}$$

The first term in (101) is equal to

$$\mathbb{P}_h (V^k_{r,h+1} - V^{\pi}_{r,h+1})(x,a)$$

We have to bound the second term in (100). Note that

$$\begin{aligned}
&\phi(x,a)^T (\Lambda^k_h)^{-1} \sum_{\tau=1}^{k-1} \phi^{\tau}_h [V^k_{r,h+1}(x^{\tau}_{h+1}) - \mathbb{P}_h V^k_{r,h+1}(x^{\tau}_h, a^{\tau}_h)] = \\
&(n^k_h(x,a) + \lambda)^{-1} \sum_{\tau} \mathbf{1}\{(x,a) = (x^{\tau}_h, a^{\tau}_h)\} (V^k_{r,h+1}(x^{\tau}_{h+1}) - \mathbb{P}_h V^k_{r,h+1}(x,a))
\end{aligned} \tag{103}$$

We now bound the above. First, we express $V_{r,h+1} = \tilde{V}_{h+1} + \Delta V_j$ where $|\Delta V_j| \leq \epsilon$ for $\tilde{V}_j$ in the ϵ-covering set of V_j. Since the upper bound for the value function is H, one can trivially obtain the ϵ-covering number for the value function as $N^V_{\epsilon} = (1 + 2|S|H/\epsilon)^{|S|}$. Now, we have

$$\begin{aligned}
&(n^k_h(x,a) + \lambda)^{-1} \sum_{\tau} \mathbf{1}\{(x,a) = (x^{\tau}_h, a^{\tau}_h)\} (V^k_{r,h+1}(x^{\tau}_{h+1}) - \mathbb{P}_h V^k_{r,h+1}(x,a)) = \\
&(n^k_h(x,a) + \lambda)^{-1} \sum_{\tau} \mathbf{1}\{(x,a) = (x^{\tau}_h, a^{\tau}_h)\} (\tilde{V}^k_{h+1}(x^{\tau}_{h+1}) - \mathbb{P}_h \tilde{V}^k_{h+1}(x,a)) + 2(n^k_h(x,a) + \lambda)^{-1} \epsilon
\end{aligned} \tag{104}$$

From Theorem 2, we obtain with probability $1 - \delta$ for a specific $(x,a) \in [S] \times [A]$

$$\begin{aligned}
(n^k_h(x,a) + \lambda)^{-1/2} \sum_{\tau=1}^{k} \mathbf{1}\{(x,a) = (x^{\tau}_h, a^{\tau}_h)\} (\tilde{V}_j(x_{\tau}) - \mathbb{E}[\tilde{V}_j(x_{\tau})|x,a]) &\leq \sqrt{2H^2 \left[\log \left(\frac{(n^k_h(x,a) + \lambda)^{1/2} \lambda^{-1/2}}{\delta} \right) \right]} \\
&\leq \sqrt{2H^2 \log(T/\delta)}
\end{aligned}$$

Hence, using the union bound where $\delta = p/(N^{\epsilon}_V |S||A|)$, we obtain with probability $1 - p/2$ for any (x,a)

$$\begin{aligned}
&(n^k_h(x,a) + \lambda)^{-1/2} \sum_{\tau=1}^{k} \mathbf{1}\{(x,a) = (x^{\tau}_h, a^{\tau}_h)\} (\tilde{V}_j(x_{\tau}) - \mathbb{E}[\tilde{V}_j(x_{\tau})|x,a]) \\
&\leq \sqrt{2H^2 \left[\log(2T|S||A|/p) + |S| \log(1 + 2H|S|/\epsilon) \right]}
\end{aligned} \tag{105}$$

Using $\epsilon = H/K$, we obtain from (104) and (105) as

$$\begin{aligned}
&(n^k_h(x,a) + \lambda)^{-1} \sum_{\tau=1}^{k} \mathbf{1}\{(x,a) = (x^{\tau}_h, a^{\tau}_h)\} (V_{r,h+1}(x_{\tau}) - \mathbb{E}[\tilde{V}_{r,h+1}(x_{\tau})|x,a]) \leq \\
&\sqrt{2H^2 |S| \log(4T|S||A|/p)} (n^k_h(x,a) + \lambda)^{-1/2}
\end{aligned} \tag{106}$$

Hence, we can write

$$Q^k_{r,h}(x,a) - Q^{\pi}_{r,h}(x,a) \leq C_4 H \sqrt{|S|\log(4(C_5+1)|S||A|T)/p)} \sqrt{(\phi(x,a)^T(\Lambda^k_h)^{-1}\phi(x,a)} \tag{107}$$

for some constant C_4 and C_5.

Hence, in Lemma 11, we can use the β value as $C_4 H\sqrt{|S|\log(|S||A|4T/p)}$. Recall equation (67)

$$\sum_{k=1}^{K} V^k_{j,1}(x_1) - V^{\pi_k}_{j,1}(x_1) \leq \sum_{k=1}^{K}\sum_{h=1}^{H}(D^k_{j,h,1} + D^k_{j,h,2}) + \sum_{k=1}^{K}\sum_{h=1}^{H} 2\beta\sqrt{\phi(x^k_h,a^k_h)^T(\Lambda^k_h)^{-1}\phi(x^k_h,a^k_h)}$$

Hence, following the same arguments as in Lemmas 3, and 6, we obtain

$$\sum_{k=1}^{K} V^k_{j,1}(x_1) - V^{\pi_k}_{j,1}(x_1) \leq \mathcal{O}(\sqrt{|S|^2|A|H^3T((\log(|S||A|\log(|A|)T/p))^2)}$$

where we use (76) to bound the above. The above is enough to obtain the improved regret and hard constraint violation bound as the rest of the argument will follow the same logic. □

I Difference from the Bandit-setup

Recently, Guo et al. (2022) proposed an algorithm which achieves $\tilde{\mathcal{O}}(\sqrt{T})$ regret and hard-constraint violation in various bandit setups using primal-dual approach. The episodic RL-setup with $H = 1$ is equivalent to the bandit setup. Thus, our approach is applicable to the bandit setup as well. However, unfortunately, the approach in Guo et al. (2022) can not be extended to the episodic CMDP setup. *We will describe the main issue next.*

For the bandit set-up, Guo et al. (2022) considered the following problem

$$\text{maximize}_{\ \pi} \sum_a \pi(a) f(a) \quad \text{subject to } \sum_a \pi(a) g(a) \leq 0 \tag{108}$$

In Guo et al. (2022), a dual variable $\hat{Y}_k \geq \sqrt{K}$ is used (hence, the dual variable is always greater than or equal to $\sqrt{K}$). Since there is no need of multiple steps in bandit setup, there is no need of value function. Rather, one only needs to estimate the reward and utility function f and g respectively. Guo et al. (2022) estimated an optimistic reward function $\hat{f}(a)$ and utility function $\hat{g}(a)$ for each a. Then, Guo et al. (2022) proposed an algorithm according to the greedy policy

$$a = \arg\max_{a'}(\hat{f}(a') - \hat{Y}_k(\hat{g}(a'))_+) \tag{109}$$

Note that if $\hat{g}(a')$ is positive, then it negates the reward which means that such an action would be avoided. From the optimism, one can show that $\hat{f}(a') - Y_k(\hat{g}(a'))_+ \geq f(a^*) - Y_k(g(a^*))_+$ where a^* is the optimal solution. Using the above, Guo et al. (2022) obtained the regret and hard constraint violation bound.

The regret and violation bound obtained by Guo et al. (2022) is $\tilde{\mathcal{O}}(\sqrt{T})$ *which is the same as ours.* However, the computation complexity is much less. In particular, there is no need to obtain the dual-variable Y_k at every episode to balance between the reward and the utility maximization which we proposed. Rather, any $\hat{Y}_k \geq \sqrt{K}$ would be sufficient.

It is natural to ask whether we can extend the above approach for the RL setup. Readers would note that in the RL one would replace the f and g with $Q_{r,h}$ and $Q_{g,h}$ respectively. In particular, one would be tempted to take action according to the greedy policy

$$a = \arg\max_{a'}(Q^k_{r,h}(x,a') - Y_k(b - Q^k_{g,h})_+) \tag{110}$$

However, the above would not work.

First, as mentioned in Ghosh et al. (2022) greedy policy can not provide required ϵ-covering number for individual value function which is needed to show uniform concentration bound for linear CMDP. Ghosh et al. (2022) showed that using the soft-max policy instead of greedy policy would solve the above issue. Thus, it is natural to ask whether using the soft-max policy instead of greedy-policy would solve the above problem.

However, even if we use the soft-max policy on the composite state-action value function $Q^k_{r,h}(x,a) - Y_k(b - Q^k_{g,h}(x,a))_+$ we can not guarantee optimism result. The reason is that in (5) the constraint only entails that the expected cumulative utility must be greater than or equal to b (i.e., $V^{\pi_k}_{g,1}(x_1) \geq b$). However, it does not imply for what value function at h-th step (i.e.,$V^{\pi_k}_{g,h}(x)$) for $h > 1$ would lead to a feasible policy. Certainly, requiring that $V^{\pi_k}_{g,h} \geq b$ would be too pessimistic. Thus, such a policy would most likely reject even the optimal policy π^*. Hence, even if one correctly estimates $V^{\pi_k}_{g,h}$ it does not know whether it is feasible or not unless $H = 1$. Hence, the optimism result no longer holds, unlike the bandit scenario. Instead, our algorithm tries to find the dual-variable which would perfectly balance between the reward maximization and utility maximization.

Further, Guo et al. (2022) inherently assumed that the optimal policy is deterministic as the self-bounding property in the proof of Theorem 1 (Appendix A in Guo et al. (2022)) only holds when the optimal policy is deterministic. It is already known that CMDP admits stochastic policy Altman (1999), thus, the above approach can not be applied to the RL-setup.

J Strictly feasible Policy

In order to obtain the soft-violation bound, all the existing works assume that there exists a strictly feasible policy $\bar{\pi}$ such that $V^{\bar{\pi}}_{g,1}(x_1) \geq b + \gamma$ for some $\gamma > 0$. Then, from the theory of strong duality, one obtains the optimal dual variable is upper bounded by $2H/\gamma$ Ghosh et al. (2022); Ding et al. (2021). As we mentioned before, in our approach, we do not need such strict feasibility assumption. However, if we assume strict feasibility, we only need to search for dual variable till H/γ rather than $\sqrt{K}$. In fact, it is guaranteed that while loop will return a Y_k such that $V^{k,Y_k}_{g,1}(x_1) \geq b - (\gamma/H)K^{-1/2}$ which would help us to prove the violation bound. We now formalize the result in the following

Lemma 26. *Let there be a strictly feasible policy such that $V^{\bar{\pi}}_{g,1} \geq b + \gamma$ where $\gamma > 0$, Algorithm 1 returns $V^{k,Y_k}_{g,1}(x_1) \geq b - \gamma/H(K^{-1/2})$ for $Y_k \leq H/\gamma$.*

Proof. If while loop is terminated before H/γ, then we are certain that $V^{k,Y}_{g,1} \geq b$, hence, the statement of the lemma is trivially true. Thus, we consider the case when Y_k reaches H/γ..

We prove by contradiction. Suppose that the above does not hold. Then, for all $Y_k \leq H/\gamma$, $V^{k,Y_k}_{g,1}(x_1) < b - \gamma/H(K^{-1/2})$. However, note that $V^{\bar{\pi}}_{g,1}(x_1) = b + \gamma$. Hence,

$$V^{\bar{\pi}}_{r,1}(x_1) + H/\gamma(V^{\bar{\pi}}_{g,1}(x_1)) \geq Hb/\gamma + H \tag{111}$$

where we used the fact that $V^{\bar{\pi}}_{r,1}(x_1) \geq 0$. Now, note that

$$V^{k,H/\gamma}_{r,1}(x_1) + H/\gamma V^{k,H/\gamma}_{g,1}(x_1) < H + Hb/\gamma - K^{-1/2} \tag{112}$$

From Lemma 9 for any Y,$V^{k,Y}_{r,1}(x_1) + YV^{k,Y}_{g,1}(x_1) + K^{-1/2} \geq V^{\bar{\pi}}_{r,1}(x_1) + YV^{\bar{\pi}}_{g,1}(x_1)$. Thus,

$$V^{k,H/\gamma}_{r,1}(x_1) + Y_kV^k_{g,1}(x_1) \geq V^{\bar{\pi}}_{r,1}(x_1) + Y_kV^{\bar{\pi}}_{g,1}(x_1) - K^{-1/2} \geq H + Hb\gamma - K^{-1/2} \tag{113}$$

However, it contradicts (112). Hence, the result follows. □

Note that we use the fact that $Y_k \geq \sqrt{K}$ to show the violation bound. Hence, the regret bound can be proved in a similar way as we have proved. In the following, we prove the violation bound.

We know that at least when the loop ends we have $V^{k,Y_k}_{g,1}(x_1) \geq b - K^{-1/2}\gamma/H$. Equating $V^{k,Y_k}_{g,1} = V^k_{g,1}$ for the chosen dual value, we have

$$\sum_k (b - V^k_{g,1}(x_1))_+ \leq K^{1/2}\gamma/H. \tag{114}$$

which gives the bound on term $\mathcal{T}_5$. This shows the bound on violation in (12). Hence, the result follows.

Since we only need to search for dual variable till H/γ instead of $\sqrt{K}$, we can reduce the maximum no. of steps required to find Y_k. Note that similar to Ghosh et al. (2022); Ding et al. (2021) we do not need to know the strictly feasible policy, rather, we only need to know (estimate) γ.

K Numerical Evaluations

Hyper-parameter Selection: Throughout this section, we use $\alpha = \sqrt{K}/H$, $p = 0.05$,and $\eta = \frac{1}{\sqrt{KH}}$. Thus, we are using a larger η compared to the one described in Algorithm 1. However, such a higher η decreases the computation time significantly, and yet, we observe good empirical behavior. For algorithm proposed in Ghosh et al. (2022) we use $\alpha = K/H$, and the dual variable learning rate $\eta = 10/\sqrt{KH^2}$ as suggested by the paper. For OptPess-PrimalDual (Liu et al., 2021a) we use the following set of hyper parameters: $\eta^k = 10H\sqrt{k}$, $\epsilon^k = H^2\sqrt{|S|^3|A|}\log(k/\delta'+1)/\sqrt{k\log(k/\delta')}$, $\delta' = p/(|S|^2|A|H)$ for episode k since we get the best result for this set of hyper-parameters.

K.1 Setup in Ghosh et al. (2022)

Similar to Ghosh et al. (2022) we consider that even if the scheduler schedules a job, the machine might not be able to complete the 2 jobs. We consider $\mathbb{P}(x_{h+1} = (x_h - 2a)_+|x_h, a) = 0.8$, $\mathbb{P}(x_{h+1} = (x_h - a)_+|x_h, a) = 0.1$, and $\mathbb{P}(x_{h+1} = x_h|x_h, a) = 0.1$.

$$x_{h+1} = \begin{cases} \max\{x_h - 2a, 0\} & \text{w.p. } 0.8 \\ \max\{x_h - a, 0\} & \text{w.p. } 0.1 \\ x_h, & \text{otherwise} \end{cases}$$

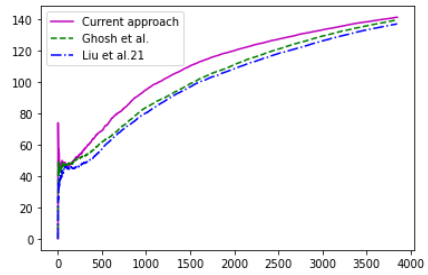
Thus, if $a = 0$, the state $x_{h+1} = x_h$. We want that utility to be less than or equal to 4 at the end of every episode.

We evaluate Algorithm 1 on a simulated model (same as in Ghosh et al. (2022)) to validate our theoretical results. We consider that the number of jobs belongs to the discrete state $\{0, 1, \ldots, 9\}$ where 0 means that there is no job. The length of the episode (H)is 10. At the start of each episode, the state of the job is 9, i.e., the job stack is full. The agent needs to decide whether to send job ($a = 1$) or not ($a = 0$) to a machine. The environment is similar to Ghosh et al. (2022). In particular, we assume that at time steps from 3 to 6, the reward is $1 - 0.9a$, In other time steps, the reward is $1 - 0.2a$. When $a = 1$, the job state decreases with the same probability as in Ghosh et al. (2022). This mimics the setup where at a certain time, it might be more costly to process a job (for example, electricity cost might be higher, or the machine needs to abandon an important job). The agent gets an utility of $g(x_h, a_h, x_{h+1}) = (x_h - x_{h+1})/2$. b is set at 3.5. This will ensure that at most 2 job can remain at the end of each episode.

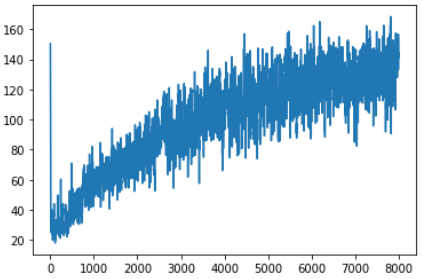
We run Algorithm 1 for 3×10^5 episodes (K). Note that the setup can be represented in a tabular form (Appendix H). We plot the reward achieved in an episode (averaged over the no. of episodes) and cumulative hard constraint violations in Figure 1. As predicted by our theory, the hard constraint violation scales smaller than $\sqrt{K}$. In fact, our approach employs policies that are close to satisfying the constraint after 1.5×10^5 episodes. Further, we observe that the hard constraint violations achieved by the algorithm proposed in Ghosh et al. (2022) and OptPess-PrimalDual (Liu et al., 2021a) are much higher and grow at a faster scale compared to ours. The average reward achieved by our approach is close to the optimal one. The above shows the efficacy of our approach in achieving sub-linear hard constraint violation compared to other primal-dual based approaches which mostly focus on reducing the soft constraint violation.

K.2 Cartpole

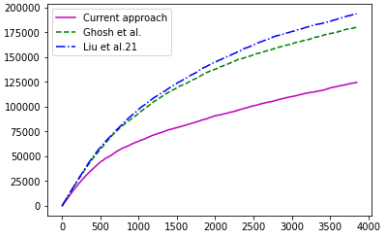
We also consider the traditional cartpole environment of OpenAIGym Brockman et al. (2016). Similar to Xu et al. (2021), we consider that the agent gets a reward of 1 if the cartpole is kept upright, and gets 1 utility unless the absolute value of the angle (θ) exceeds 6°. In which case the utility is 0. Each episode is of length 200 steps. We set $b = 140$.



(a) Moving Average (of Last 10 episodes) of Reward for our approach, Ghosh et al. (2022), and Liu et al. (2021a)



(b) Moving average of utility for our approach. The value of b is 140.



(c) Hard-violation of our approach, Ghosh et al. (2022), and Liu et al. (2021a)

Figure 2: Empirical evaluations for Cartpole environment of OpenAIGym Brockman et al. (2016).

The state space consists of location, speed, angle, and angular velocity. We discrete each in evenly-spaced 15 states, thus, the total state space is 15^4. Even though the state space is large, we observe that our approach learns to achieve the optimal policy. It shows that our approach can be applied to continuous state space as well. Further, our approach also learns to satisfy constraints.

From Figure 2, we observe that the hard constraint violation in our approach grows on a much smaller scale compared to both Ghosh et al. (2022) and OptPess-PrimalDual. In fact, even the reward achieved by our approach is higher. The highest achievable reward is 200 and as we can see our approach indeed approaches the optimal reward as K increases.

K.3 A paradoxical CMDP

This MDP is proposed by Moskovitz et al. (2023). If one takes action $a = 0$, one would stay in state s_0 and gets a reward 1. However, taking action $a = 1$ would take the agent to state s_1 and reward 0. At state s_1, the action $a = 1$ would get the agent to state s_0 and reward 1. On the other hand, the agent will remain in state s_1 if the action $a = 0$ and the reward will be 0 (see Figure 3a). The length of the horizon $H = 10$. The utility $g = r$. The CMDP problem is

$$\text{maximize } V_{r,1}(s_0) \quad \text{s.t } V_{g,1}(s_0) \leq 5 \tag{115}$$

Hence, the agent should be in state s_0 half of the time and state s_1 for the rest. Though the maximum cumulative reward can be 10 where the agent can remain in state s_0. However, such a strategy is not feasible.

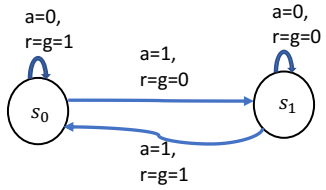
From Figure 3 it is evident that the hard-constraint violation in our approach scales at most $\mathcal{O}(\sqrt{K})$ and scales at a much smaller scale compared to Liu et al. (2021a) and Ghosh et al. (2022). Hence, it shows that our algorithm is able to achieve feasible policy at a faster scale compared to the existing state-of-the-art approaches who have only focused on reducing the soft constaint violation. Figure 3 also shows that initially, the reward is higher than 5 as our algorithm still explores. Hence, the algorithm chooses infeasible policies more frequently. Finally, the reward decreases the reward converges to the optimal value of 5 as our algorithm chooses feasible optimal policies more frequently.

K.4 Experiment on Frozen Lake

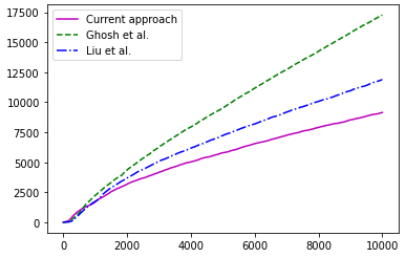
We also simulated our method on the frozen lake environment of OpenAIgym Brockman et al. (2016). We consider 4×4 grid. The agent gets a reward 1 when it reaches the goal state. We consider an episode length of $H = 9$. The agent stays in the goal state once it reaches there. The agent is also permanently in the hole if it reaches a hole. In this case, the reward will be 0.

In the original frozen-lake experiment, there are two optimal ways to get to the goal. We add a constraint to ensure that one of the paths is infeasible. In particular, we add a utility function where the agent gets a utility of 1 except when the agent falls into a hole or goes to any of the blocks on the extreme left-hand column (Figure 4a). This will ensure that the optimal path of the left-hand side is infeasible. We set b to 8 which ensures that only one path is feasible and optimal.

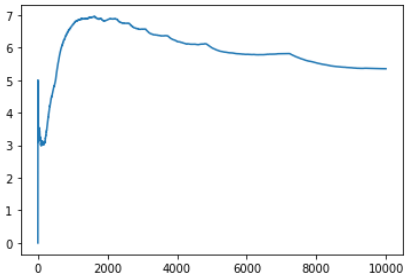
Our simulation result shows that our approach indeed identifies the optimal path (Figure 4). The violation



(a) Paradoxical CMDP



(b) The plot of hard-violation of our approach,that of Ghosh et al. (2022), and Liu et al. (2021b)



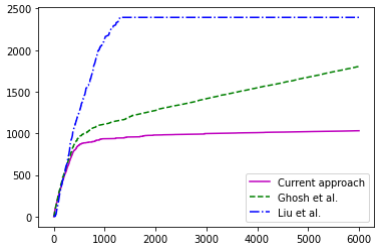
(c) Avg. reward of our approach.

Figure 3: Empirical evaluations for Paradoxical CMDP inspired from Moskovitz et al. (2023).

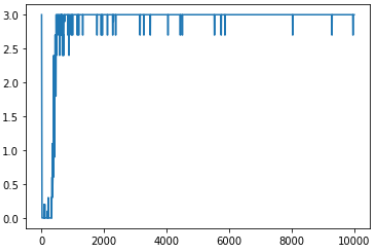
S, g=1	F, g=1	F, g=1	F,g=1
F, g=0,	H, g=0	F, g=1	H,g=0
F, g=0	F, g=1	F, g=1	H, g=0
H, g=0	F, g=1	F, g=1	G, g=1

The optimal solution– Move rightward for 2 steps, and downward for 3 steps, and then rightward for 1 step to reach the goal state.

(a) Frozen-Lake CMDP



(b) The plot of hard-violation of our approach, that of Ghosh et al. (2022), and Liu et al. (2021a).



(c) Moving Average reward of our approach for Frozen-lake.

Figure 4: Empirical evaluations for Frozen-Lake Environment of OpenAIGym.

becomes almost 0 after 1000 episodes (Figure 4b). Hence, our algorithm indeed achieves the feasible optimal policy. The hard constraint violation achieved by Ghosh et al. (2022) is much higher compared to ours and is unable to obtain feasible policy. The hard constraint violation achieved by Liu et al. (2021a) is higher compared to even Ghosh et al. (2022).Further, the reward achieved by our approach converges to the optimal one (3) after only 1000 episodes (Figure 4c). Again, it shows that our algorithm is able to identify optimal policy even within smaller number of episode while the other algorithms are unable to find feasible policy.