Learning General Parameterized Policies for Infinite Horizon Average Reward Constrained MDPs via Primal-Dual Policy Gradient Algorithm

Qinbo Bai
Purdue University
West Lafayette, IN 47906
bai113@purdue.edu

Washim Uddin Mondal
Indian Institute of Technology Kanpur
Kanpur, UP, India 208016
wmondal@iitk.ac.in

Vaneet Aggarwal
Purdue University
West Lafayette, IN 47906
vaneet@purdue.edu

Abstract

This paper explores the realm of infinite horizon average reward Constrained Markov Decision Processes (CMDPs). To the best of our knowledge, this work is the first to delve into the regret and constraint violation analysis of average reward CMDPs with a general policy parametrization. To address this challenge, we propose a primal dual-based policy gradient algorithm that adeptly manages the constraints while ensuring a low regret guarantee toward achieving a global optimal policy. In particular, our proposed algorithm achieves $\tilde{\mathcal{O}}(T^{4/5})$ objective regret and $\tilde{\mathcal{O}}(T^{4/5})$ constraint violation bounds.

1 Introduction

The framework of Reinforcement Learning (RL) is concerned with a class of problems where an agent learns to yield the maximum cumulative reward in an unknown environment via repeated interaction. RL finds applications in diverse areas, such as wireless communication, transportation, and epidemic control [1, 2, 3]. RL problems are mainly categorized into three setups: episodic, infinite horizon discounted reward, and infinite horizon average reward. Among them, the infinite horizon average reward setup is particularly significant for real-world applications. It aligns with most of the practical scenarios and captures their long-term goals. Some applications in real life require the learning procedure to respect the boundaries of certain constraints. In an epidemic control setup, for example, vaccination policies must take the supply shortage (budget constraint) into account. Such restrictive decision-making routines are described by constrained Markov Decision Processes (CMDP) [4, 5, 6]. Existing papers on CMDPs utilize either a tabular or a linear MDP structure. This work provides the first algorithm for an infinite horizon average reward CMDP with general parametrization and proves its sub-linear regret and constraint violation bounds.

There are two primary ways to solve a CMDP problem in the infinite horizon average reward setting. The first one, known as the model-based approach, involves constructing estimates of the transition probabilities of the underlying CMDP, which are subsequently utilized to derive policies [6, 7, 5]. The caveat of this approach is the large memory requirement to store the estimated parameters, which effectively curtails its applicability to CMDPs with large state spaces. The alternative strategy, known as the model-free approach, either directly estimates the policy function or maintains an estimate of the Q function, which is subsequently used for policy generation [8]. Model-free algorithms typically demand lower memory and computational resources than their model-based counterparts. Although the CMDP has been solved in a model-free manner in the tabular [8] and linear [9] setups, its exploration with the general parameterization is still open and is the goal of this paper.

General parameterization indexes the policies by finite-dimensional parameters (e.g., weights of neural networks) to accommodate large state spaces. The learning is manifested by updating these

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3447

Algorithm	Regret	Violation	Model-free	Setting
Algorithm 1 in [6]	$\tilde{\mathcal{O}}(\sqrt{T})$	$\tilde{\mathcal{O}}(\sqrt{T})$	No	Tabular
Algorithm 2 in [6]	$\tilde{\mathcal{O}}(T^{2/3})$	$\tilde{\mathcal{O}}(T^{2/3})$	No	Tabular
UC-CURL and PS-CURL [5]	$\tilde{\mathcal{O}}(\sqrt{T})$	0	No	Tabular
Algorithm 2 in [9]	$\tilde{\mathcal{O}}((dT)^{3/4})$	$\tilde{\mathcal{O}}((dT)^{3/4})$	No	Linear MDP
Algorithm 3 in [9]	$\tilde{\mathcal{O}}(\sqrt{T})$	$\tilde{\mathcal{O}}(\sqrt{T})$	No	Linear MDP
Triple-QA [8]	$\tilde{\mathcal{O}}(T^{5/6})$	0	Yes	Tabular
This paper	$\tilde{\mathcal{O}}(T^{\frac{4}{5}})$	$\tilde{\mathcal{O}}(T^{\frac{4}{5}})$	Yes	General Parameterization

Table 1: This table summarizes the different model-based and mode-free state-of-the-art algorithms available in the literature for average reward CMDPs. We note that our proposed algorithm is the first to analyze the regret and constraint violation for average reward CMDP with general parametrization. Here, the parameter d refers to the dimension of the feature map for linear MDPs.

parameters using policy gradient (PG)-type algorithms. Note that PG algorithms are primarily studied in discounted reward setups. For example, [10] characterizes the sample complexities of the PG and the Natural PG (NPG) algorithms with softmax and direct parameterization. Similar results for general parameterization are obtained by [11, 12]. The regret analysis of a PG algorithm with the general parameterization has been recently performed for an infinite horizon average reward MDP without constraints [13]. Similar regret and constraint violation analysis for the average reward CMDP is still missing in the literature. In this paper, we bridge this gap.

Challenges and Contribution: We propose a PG-based algorithm with general parameterized policies for the average reward CMDP and establish its sublinear regret and constraint violation bounds. In particular, assuming the underlying CMDP to be ergodic, we demonstrate that our PG algorithm achieves an average optimality rate of $\tilde{\mathcal{O}}(T^{-\frac{1}{5}})$ and average constraint violation rate of $\tilde{\mathcal{O}}(T^{-\frac{1}{5}})$. Invoking this convergence result, we establish that our algorithm achieves regret and constraint violation bounds of $\tilde{\mathcal{O}}(T^{\frac{4}{5}})$. Apart from providing the first sublinear regret guarantee for the average reward CMDP with general parameterization, our work also improves the state-of-the-art regret guarantee, $\tilde{\mathcal{O}}(T^{5/6})$ in the model-free tabular setup [8].

Despite the availability of sample complexity analysis of PG algorithms with constraints in the discounted reward setup [14, 4] and PG algorithms without constraint in average reward setup [13], obtaining sublinear regret and constraint violation bounds for their average reward counterpart is challenging.

- [14, 4] solely needs an estimate of the value function V while we additionally need the estimate of the gain function, J.
- [14, 4] assume access to a simulator to generate unbiased value estimates. In contrast, our algorithm uses a sample trajectory of length H to estimate the values and gains and does not assume the availability of a simulator.
- The first-order convergence analysis (Lemma 6) differs from that in [13]. Note that both of these papers use an ascent-like inequality. In [13], this bounds the term $J(\theta_{k+1}) - J(\theta_k)$. The final result is obtained by calculating a sum over k which cancels the intermediate terms and leaves us with $J(\theta_K) - J(\theta_1)$. We would like to emphasize that the cancellation of the intermediate terms is crucial to establishing the result. However, a similar effort in our case only leads to a bound of $J_{\mathrm{L}}(\theta_{k+1}, \lambda_k) - J_{\mathrm{L}}(\theta_k, \lambda_k)$. Note that directly performing a sum over this difference does not lead to the cancellation of intermediate terms. We had to take a different route and apply the bounds of the Lagrange multipliers and the estimate of the constraint function to achieve that goal.
- After solving the problems mentioned above, we prove $\tilde{\mathcal{O}}(T^{-\frac{1}{5}})$ convergence rate of the Lagrange function. Unfortunately, the strong duality property, which is central to proving convergence results of CMDPs for tabular and softmax policies, does not hold under the general parameterization. As a result, the convergence result for the dual problem does not automatically translate to that for the primal problem, which is a main difference from [13]. We overcome this barrier by introducing a novel constraint violation analysis and a series of intermediate results (Lemma 16-18) that help disentangle the regret and constraint violation rates from the Lagrange convergence. It is important to mention that although the techniques applied are inspired by the [14], those techniques cannot be directly adopted for average reward MDPs. This is primarily because the estimate $\hat{J}_c(\theta_k)$ is biased in the average case. To the best of our knowledge, constraint violation analysis with a biased estimate of the cost value is not available in the literature and is performed for the first time in our paper.

•Due to the presence of the Lagrange multiplier, the convergence analysis of a CMDP is much more convoluted than its unconstrained counterpart. The learning rate of the Lagrange update, β, turns out to be pivotal in determining the growth rate of regret and constraint violation. Low values of β push the regret down while simultaneously increasing the constraint violation. Finding the optimal value of β that judiciously balances these two competing goals is one of the cornerstones of our analysis.

Related work for unconstrained average reward RL: In the absence of constraints, both model-based and model-free tabular setups have been widely studied for infinite horizon average reward MDPs. For example, the model-based algorithms proposed by [15, 16] achieve the optimal regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$. Similarly, the model-free algorithm proposed by [17] for tabular MDP results in $\tilde{\mathcal{O}}(\sqrt{T})$ regret. Regret analysis for average reward MDP with general parametrization has been recently studied in [13], where a regret bound of $\tilde{\mathcal{O}}(T^{3/4})$ is derived.

Related work for constrained RL: The constrained reinforcement learning problem has been extensively studied both for infinite horizon discounted reward and episodic MDPs. For example, discounted reward CMDPs have been recently studied in the tabular setup [18], with both softmax [14, 19], and general policy parameterization [14, 19, 4, 12]. Moreover, [20, 21, 22] investigated episodic CMDPs in the tabular setting.

Recently, the infinite horizon average reward CMDPs have been investigated in model-based setups [5, 6, 7], tabular model-free setting [8] and linear CMDP setting [9]. For model-based CMDP setup, [6] proposed a model-based online mirror descent algorithm in the ergodic setting which achieves $\tilde{\mathcal{O}}(\sqrt{T})$ for regret and violation at the same time. [7] proposed algorithms based on the posterior sampling and the optimism principle that achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret with zero constraint violations in the ergodic setting. However, the above model-based algorithms cannot be extended to large state space. In the tabular model-free setup, the algorithm proposed by [8] achieves a regret of $\tilde{\mathcal{O}}(T^{5/6})$ with zero constraint violations. Finally, in the linear CMDP setting, [9] achieves $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound with zero constraint violation. Note that the linear CMDP setting assumes that the transition probability has a certain linear structure with a known feature map which is not realistic. Table 1 summarizes all relevant works. Unfortunately, none of these papers study the infinite horizon average reward CMDPs with general parametrization which is the main focus of our article.

Additionally, for the weakly communicating setting, [6] proposed a model-based algorithm achieving $\tilde{\mathcal{O}}(T^{2/3})$ for both regret and violation in tabular case. [9] further extends such result to linear MDP setting with $\tilde{\mathcal{O}}(T^{3/4})$ regret and violation. In general, it is difficult to propose a model-free algorithm with provable guarantees for Constrained MDPs (CMDPs) without considering the ergodic model. [6] pointed out several extra challenges in Weakly communicating MDP compared to the ergodic case. For example, there is no uniform bound for the span of the value function for all stationary policies. It is also unclear how to estimate a policy's bias function accurately without the estimated model, which is an important step for estimating the policy gradient.

2 Formulation

This paper analyzes an infinite-horizon average reward constrained Markov Decision Process (CMDP) denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, c, P, \rho)$ where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the action space of size A, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $c : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is the constraint cost function, $P : \mathcal{S} \times \mathcal{A} \to \Delta^{|\mathcal{S}|}$ is the state transition function where $\Delta^{|\mathcal{S}|}$ denotes a probability simplex with dimension $|\mathcal{S}|$, and $\rho \in \Delta^{|\mathcal{S}|}$ is the initial distribution of states. A policy $\pi \in \Pi : \mathcal{S} \to \Delta^A$ maps the current state to an action distribution. The average reward and cost of a policy, π, is,

$$J^{\pi}_{g,\rho} \triangleq \lim_{T\to\infty} \frac{1}{T}\mathbf{E}\left[\sum_{t=0}^{T-1} g(s_t, a_t)\middle| s_0 \sim \rho, \pi\right] \tag{1}$$

where $g = r, c$ for average reward and cost respectively. The expectation is calculated over the distribution of all sampled trajectories $\{(s_t, a_t)\}_{t=0}^{\infty}$ where $a_t \sim \pi(s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$, $\forall t \in \{0, 1, \cdots\}$. For notational convenience, we shall drop the dependence on ρ whenever there is no confusion. Our goal is to maximize the average reward function while ensuring that the average cost is above a given threshold. Without loss of generality, we can mathematically write this problem as,

$$\max_{\pi\in\Pi} J^{\pi}_r \text{ s.t. } J^{\pi}_c \geq 0 \tag{2}$$

 https://doi.org/10.52202/079017-3447

However, the above problem becomes difficult to handle when the underlying state space, $\mathcal{S}$ is large. Therefore, we consider a class of parametrized policies, $\{\pi_\theta|\theta \in \Theta\}$ whose elements are indexed by a d-dimensional parameter, $\theta \in \mathbb{R}^{\mathrm{d}}$ where $\mathrm{d} \ll |\mathcal{S}||\mathcal{A}|$. Thus, the original problem in Eq (2) can be reformulated as the following parameterized problem.

$$\max_{\theta\in\Theta} J_r^{\pi_\theta} \text{ s.t. } J_c^{\pi_\theta} \geq 0 \tag{3}$$

We denote $J_g^{\pi_\theta} = J_g(\theta)$, $g \in \{r, c\}$ for notational convenience. Let, $P^{\pi_\theta} : \mathcal{S} \to \Delta^{|\mathcal{S}|}$ be a transition function induced by π_θ and defined as, $P^{\pi_\theta}(s, s') = \sum_{a\in\mathcal{A}} P(s'|s, a)\pi_\theta(a|s)$, $\forall s, s'$. If $\mathcal{M}$ is such that for every policy π, the function, P^π is irreducible and aperiodic, then $\mathcal{M}$ is called ergodic.

Assumption 1. The CMDP $\mathcal{M}$ is ergodic.

Ergodicity is a common assumption in the literature [23, 24]. If $\mathcal{M}$ is ergodic, then $\forall\theta$, there exists a unique stationary distribution, $d^{\pi_\theta} \in \Delta^{|\mathcal{S}|}$ given as follows.

$$d^{\pi_\theta}(s) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr(s_t = s|s_0 \sim \rho, \pi_\theta) \tag{4}$$

Ergodicity implies that d^{π_θ} is independent of the initial distribution, ρ, and obeys $P^{\pi_\theta} d^{\pi_\theta} = d^{\pi_\theta}$. Hence, the average reward and cost functions can be expressed as,

$$J_g(\theta) = \mathbf{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(s)}[g(s, a)] = (d^{\pi_\theta})^T g^{\pi_\theta} \tag{5}$$

where $g^{\pi_\theta}(s) \triangleq \sum_{a\in\mathcal{A}} g(s, a)\pi_\theta(a|s),\ g \in \{r, c\}$. Note that the functions $J_g(\theta)$, $g \in \{r, c\}$ are also independent of the initial distribution, ρ. Furthermore, $\forall\theta$, there exist a function $Q_g^{\pi_\theta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that the following Bellman equation is satisfied $\forall(s, a) \in \mathcal{S} \times \mathcal{A}$.

$$Q_g^{\pi_\theta}(s, a) = g(s, a) - J_g(\theta) + \mathbf{E}_{s'\sim P(\cdot|s,a)} \left[V_g^{\pi_\theta}(s')\right] \tag{6}$$

where $g \in \{r, c\}$ and $V_g^{\pi_\theta} : \mathcal{S} \to \mathbb{R}$ is given as $V_g^{\pi_\theta}(s) = \sum_{a\in\mathcal{A}} \pi_\theta(a|s) Q_g^{\pi_\theta}(s, a),\ \forall s \in \mathcal{S}$. Note that if $Q_g^{\pi_\theta}$ satisfies (6), then it is also satisfied by $Q_g^{\pi_\theta} + c$ for any arbitrary, c. To uniquely define the value functions, we assume that $\sum_{s\in\mathcal{S}} d^{\pi_\theta}(s) V_g^{\pi_\theta}(s) = 0$. In this case, $V_g^{\pi_\theta}(s)$ is given by,

$$V_g^{\pi_\theta}(s) = \sum_{t=0}^{\infty} \sum_{s'\in\mathcal{S}} \left[(P^{\pi_\theta})^t(s, s') - d^{\pi_\theta}(s')\right] g^{\pi_\theta}(s') = \sum_{t=0}^{\infty} \mathbf{E}\left[\{g(s_t, a_t) - J_g(\theta)\}\,|s_0 = s\right] \tag{7}$$

where the expectation is computed over all π_θ-induced trajectories. In a similar way, $\forall(s, a)$, one can uniquely define $Q_g^{\pi_\theta}(s, a)$, $g \in \{r, c\}$ as follows.

$$Q_g^{\pi_\theta}(s, a) = \sum_{t=0}^{\infty} \mathbf{E}\left[\left\{g(s_t, a_t) - J_g(\theta)\right\}\big|s_0 = s, a_0 = a\right] \tag{8}$$

Moreover, the advantage function $A_g^{\pi_\theta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined such that $A_g^{\pi_\theta}(s, a) \triangleq Q_g^{\pi_\theta}(s, a) - V_g^{\pi_\theta}(s)$, $\forall(s, a)$, $\forall g \in \{r, c\}$. Assumption 1 also implies the existence of a finite mixing time. Specifically, for an ergodic MDP, $\mathcal{M}$, the mixing time is defined as follows.

Definition 1. The mixing time, t_{mix}^θ, of the CMDP $\mathcal{M}$ for a parameterized policy, π_θ, is defined as, $t_{\mathrm{mix}}^\theta \triangleq \min\left\{t \geq 1\big|\|(P^{\pi_\theta})^t(s, \cdot) - d^{\pi_\theta}\| \leq \frac{1}{4}, \forall s\right\}$. The overall mixing time is $t_{\mathrm{mix}} \triangleq \sup_{\theta\in\Theta} t_{\mathrm{mix}}^\theta$. In this paper, t_{mix} is finite due to ergodicity.

Mixing time characterizes how fast a CMDP converges to its stationary state distribution, d^{π_θ}, under a given policy, π_θ. We also define the hitting time as follows.

Definition 2. The hitting time of an ergodic CMDP $\mathcal{M}$ with respect to a policy, π_θ, is defined as $t_{\mathrm{hit}}^\theta \triangleq \max_{s\in\mathcal{S}}[d^{\pi_\theta}(s)]^{-1}$. The overall hitting time is defined as $t_{\mathrm{hit}} \triangleq \sup_{\theta\in\Theta} t_{\mathrm{hit}}^\theta$. In this paper, t_{hit} is finite due to ergodicity as well.

Define π^* as the optimal solution to the unparameterized problem (2). For a given CMDP $\mathcal{M}$, and a time horizon T, the regret and constraint violation of any algorithm $\mathbb{A}$ is defined as follows.

$$\mathrm{Reg}_T(\mathbb{A}, \mathcal{M}) \triangleq \sum_{t=0}^{T-1} \left(J_r^{\pi^*} - r(s_t, a_t)\right),\ \mathrm{Vio}_T(\mathbb{A}, \mathcal{M}) \triangleq -\sum_{t=0}^{T-1} c(s_t, a_t) \tag{9}$$

where the algorithm, $\mathbb{A}$, executes the actions, $\{a_t\}$, $t \in \{0, 1, \cdots\}$ based on the trajectory observed up to time, t, and the state, s_{t+1} is decided according to the state transition function, P. For simplicity, we shall denote the regret and constraint violation as Reg_T and Vio_T respectively. Our goal is to design an algorithm $\mathbb{A}$ that achieves low regret and constraint violation bounds.

3 Proposed Algorithm

We solve (3) via a primal-dual algorithm based on the following problem.

$$\max_{\theta \in \Theta} \min_{\lambda \geq 0} \; J_{\mathrm{L}}(\theta, \lambda), \tag{10}$$

where $J_{\mathrm{L}}(\theta, \lambda) \triangleq J_r(\theta) + \lambda J_c(\theta)$. The function, $J_{\mathrm{L}}(\cdot, \cdot)$, is called the Lagrange function and λ the Lagrange multiplier. Our algorithm updates the pair (θ, λ) following the policy gradient iteration as shown below $\forall k \in \{1, \cdots, K\}$ with an initial point (θ_1, λ_1), $\lambda_1 = 0$.

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J_{\mathrm{L}}(\theta_k, \lambda_k), \; \lambda_{k+1} = \mathcal{P}_{[0, \frac{2}{\delta}]}[\lambda_k - \beta J_c(\theta_k)] \tag{11}$$

where α and β are learning parameters and δ is the Slater parameter introduced in the following assumption. Finally, for any set, Λ, $\mathcal{P}_\Lambda[\cdot]$ denotes projection onto Λ. The assumption stated below ensures that we have at least one feasible interior point solution to (2).

Assumption 2 (Slater condition)**.** There exists a $\delta \in (0, 1)$ and $\bar{\theta} \in \Theta$ such that $J_c(\bar{\theta}) \geq \delta$.

Note that in (11), the dual update is projected onto the set $[0, \frac{2}{\delta}]$ because the optimal dual variable for the parameterized problem is bounded in Lemma 16. The gradient of $J_{\mathrm{L}}(\cdot, \lambda)$ can be computed by invoking a variant of the well-known policy gradient theorem [25].

Lemma 1. *The gradient of $J_{\mathrm{L}}(\cdot, \lambda)$ is computed as,*

$$\nabla_\theta J_{\mathrm{L}}(\theta, \lambda) = \mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(s)} \left[A^{\pi_\theta}_{\mathrm{L}, \lambda}(s, a) \nabla_\theta \log \pi_\theta(a|s) \right]$$

Algorithm 1 Primal-Dual Parameterized Policy Gradient

Input: Episode length H, learning rates α, β, initial parameters θ_1, λ_1, initial state $s_0 \sim \rho(\cdot)$,
$K = T/H$
for $k \in \{1, \cdots, K\}$ **do**
  $\mathcal{T}_k \leftarrow \phi$
  for $t \in \{(k-1)H, \cdots, kH - 1\}$ **do**
    Execute $a_t \sim \pi_{\theta_k}(\cdot|s_t)$
    Observe $r(s_t, a_t)$, $c(s_t, a_t)$ and s_{t+1}
    $\mathcal{T}_k \leftarrow \mathcal{T}_k \cup \{(s_t, a_t)\}$
  end for
  for $t \in \{(k-1)H, \cdots, kH - 1\}$ **do**
    Obtain $\hat{A}^{\pi_{\theta_k}}_{\mathrm{L}, \lambda_k}(s_t, a_t)$ via Algorithm 2 and $\mathcal{T}_k$
  end for
  Compute ω_k using (15)
  Update the parameters:

$$\begin{aligned} &\theta_{k+1} = \theta_k + \alpha \omega_k, \\ &\lambda_{k+1} = \mathcal{P}_{[0, \frac{2}{\delta}]} \left[\lambda_k - \beta \hat{J}_c(\theta_k) \right] \\ &\text{where } \hat{J}_c(\theta_k) = \frac{1}{H - N} \sum_{t=(k-1)H+N}^{kH-1} c(s_t, a_t) \end{aligned} \tag{12}$$

end for

where $\forall (s, a)$, $A^{\pi_\theta}_{\mathrm{L}, \lambda}(s, a) \triangleq A^{\pi_\theta}_r(s, a) + \lambda A^{\pi_\theta}_c(s, a)$, and $\{A^{\pi_\theta}_g\}_{g \in \{r, c\}}$ are the advantage functions corresponding to reward and cost. In typical RL scenarios, learners do not have access to the state transition function, P, and thereby to the functions d^{π_θ} and $A^{\pi_\theta}_{\mathrm{L}, \lambda}$. This makes computing the exact

 https://doi.org/10.52202/079017-3447

gradient a difficult task. In Algorithm 1, we demonstrate how one can still obtain good estimates of the gradient using sampled trajectories.

Algorithm 1 runs K epochs, each of duration $H = 16 t_{\rm hit} t_{\rm mix} T^{\xi} (\log T)^2$ where $\xi \in (0,1)$ defines a constant whose value is specified later. Clearly, $K = T/H$. Note that the learner is assumed to know the horizon length, T. This can be relaxed utilizing the well-known doubling trick [26]. Additionally, it is assumed that the algorithm is aware of the mixing time and the hitting time. This assumption is common in the literature [13, 17]. The first step in obtaining a gradient estimate is estimating the advantage value for a given pair (s,a). This can be accomplished via Algorithm 2. At the kth epoch, a π_{θ_k}-induced trajectory, $\mathcal{T}_k = \{(s_t, a_t)\}_{t=(k-1)H}^{kH-1}$ is obtained and subsequently passed to Algorithm 2 that searches for subtrajectories within it that start with a given state s, are of length $N = 4 t_{\rm mix} (\log T)$, and are at least N distance apart from each other. Assume that there are M such subtrajectories. Let the total reward and cost of the ith subtrajectory be $\{r_i, c_i\}$ respectively and τ_i be its starting time. The value function estimates for the kth epoch are

$$\hat{Q}_g^{\pi_{\theta_k}}(s,a) = \frac{1}{\pi_{\theta_k}(a|s)} \left[\frac{1}{M} \sum_{i=1}^{M} g_i 1(a_{\tau_i} = a) \right], \; \hat{V}_g^{\pi_{\theta_k}}(s) = \frac{1}{M} \sum_{i=1}^{M} g_i, \;\; \forall g \in \{r, c\} \tag{13}$$

This leads to the following advantage estimator.

$$\hat{A}_{{\rm L},\lambda_k}^{\pi_{\theta_k}}(s,a) = \hat{A}_r^{\pi_{\theta_k}}(s,a) + \lambda_k \hat{A}_c^{\pi_{\theta_k}}(s,a), \tag{14}$$

where $\hat{A}_g^{\pi_{\theta_k}}(s,a) = \hat{Q}_g^{\pi_{\theta_k}}(s,a) - \hat{V}_g^{\pi_{\theta_k}}(s)$, $g \in \{r, c\}$. Finally, the gradient estimator is,

$$\omega_k \triangleq \hat{\nabla}_\theta J_{\rm L}(\theta_k, \lambda_k) = \frac{1}{H} \sum_{t=t_k}^{t_{k+1}-1} \hat{A}_{{\rm L},\lambda_k}^{\pi_{\theta_k}}(s_t, a_t) \nabla_\theta \log \pi_{\theta_k}(a_t|s_t) \tag{15}$$

where $t_k = (k-1)H$ is the starting time of the kth epoch. The parameters are updated following (12). To update the Lagrange multiplier, we need an estimation of $J_c(\theta_k)$, which is obtained as the average cost of the kth epoch. It should be noted that we remove the first N samples from the kth epoch because we require the state distribution emanating from the remaining samples to be close enough to the stationary distribution $d^{\pi_{\theta_k}}$, which is the key to make $\hat{J}_c(\theta_k)$ close to $J_c(\theta_k)$. The following lemma demonstrates that $\hat{A}_{{\rm L},\lambda_k}^{\pi_{\theta_k}}(s,a)$ is a good estimator of $A_{{\rm L},\lambda_k}^{\pi_{\theta_k}}(s,a)$.

Algorithm 2 Advantage Estimation

Input: Trajectory $(s_{t_1}, a_{t_1}, \ldots, s_{t_2}, a_{t_2})$, state s, action a, Lagrange multiplier λ, and parameter θ
Initialize: $M \leftarrow 0$, $\tau \leftarrow t_1$
Define: $N = 4 t_{\rm mix} \log_2 T$.
while $\tau \leq t_2 - N$ **do**
  if $s_\tau = s$ **then**
    $M \leftarrow M + 1$, $\tau_M \leftarrow \tau$
    $g_M \leftarrow \sum_{t=\tau}^{\tau+N-1} g(s_t, a_t)$, $\forall g \in \{r, c\}$
    $\tau \leftarrow \tau + 2N$.
  else
    $\tau \leftarrow \tau + 1$.
  end if
end while
if $M > 0$ **then**
  Compute $\hat{Q}_g(s,a)$, $\hat{V}_g(s)$ via (13), $\forall g \in \{r, c\}$
else
  $\hat{V}_g(s) = 0$, $\hat{Q}_g(s,a) = 0$, $\forall g \in \{r, c\}$
end if
return $(\hat{Q}_r(s,a) - \hat{V}_r(s)) + \lambda(\hat{Q}_c(s,a) - \hat{V}_c(s))$

Lemma 2. *The following inequality holds $\forall k$, $\forall (s,a)$ and sufficiently large T.*

$$\mathbf{E}\left[\left(\hat{A}_{{\rm L},\lambda_k}^{\pi_{\theta_k}}(s,a) - A_{{\rm L},\lambda_k}^{\pi_{\theta_k}}(s,a) \right)^2 \right] \leq \mathcal{O}\left(\frac{t_{\rm hit} N^3 \log T}{\delta^2 H \pi_{\theta_k}(a|s)} \right) = \mathcal{O}\left(\frac{t_{\rm mix}^2 (\log T)^2}{\delta^2 T^{\xi} \pi_{\theta_k}(a|s)} \right) \tag{16}$$

Lemma 2 shows that the L_2 error of our proposed advantage estimator can be bounded above as $\tilde{\mathcal{O}}(T^{-\xi})$. We later utilize the above result to prove the goodness of the gradient estimator. It is to be clarified that our Algorithm 2 is inspired by Algorithm 2 of [17]. However, while the authors of [17] choose $H = \tilde{\mathcal{O}}(1)$, we adapt $H = \tilde{\mathcal{O}}(T^{\xi})$. This subtle change is important in proving a sublinear regret for general parametrization.

4 Global Convergence Analysis

This section first shows that the sequence $\{\theta_k, \lambda_k\}_{k=1}^{K}$ produced by Algorithm 1 is such that their associated Lagrange sequence $\{J_{\mathrm{L}}(\theta_k, \lambda_k)\}_{k=1}^{\infty}$ converges globally. By expanding the Lagrange function, we then exhibit convergence of each of its components $\{J_g(\theta_k, \lambda_k)\}_{k=1}^{K}$, $g \in \{r, c\}$. This is later used for regret and constraint violation analysis. Before delving into the details, we would like to state a few necessary assumptions.

Assumption 3. The score function (stated below) is G-Lipschitz and B-smooth. Specifically, $\forall \theta, \theta_1, \theta_2 \in \Theta$, and $\forall (s, a)$, the following inequalities hold.

$$\|\nabla_\theta \log \pi_\theta(a|s)\| \leq G,\ \|\nabla_\theta \log \pi_{\theta_1}(a|s) - \nabla_\theta \log \pi_{\theta_2}(a|s)\| \leq B\|\theta_1 - \theta_2\|$$

Remark 1. The Lipschitz and smoothness properties of the score function are commonly assumed for policy gradient analyses [27, 28, 29]. These assumptions hold for simple parameterization classes such as Gaussian policies.

Note that by combining Assumption 3 with Lemma 2 and using the gradient estimator as given in (15), one can deduce the following result.

Lemma 3. *The following inequality holds $\forall k$ provided that assumptions 1 and 3 are true.*

$$\mathbf{E}\left[\|\omega_k - \nabla_\theta J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2\right] \leq \tilde{\mathcal{O}}\left(\delta^{-2} A G^2 t_{\mathrm{mix}}^2 T^{-\xi}\right) \tag{17}$$

Lemma 3 claims that the gradient estimation error can be bounded as $\tilde{\mathcal{O}}(T^{-\xi})$. We will use this result later to prove the global convergence of our algorithm.

Assumption 4. Let the transferred compatible function approximation error be defined as follows.

$$L_{d^{\pi^*}, \pi^*}(\omega^*_{\theta,\lambda}, \theta, \lambda) = \mathbf{E}_{s\sim d^{\pi^*}} \mathbf{E}_{a\sim\pi^*(s)}\left[\left(\nabla_\theta \log \pi_\theta(a|s) \cdot \omega^*_{\theta,\lambda} - A^{\pi_\theta}_{\mathrm{L},\lambda}(s,a)\right)^2\right] \tag{18}$$

where π^* is the optimal solution of unparameterized problem in (2) and

$$\omega^*_{\theta,\lambda} = \arg\min_{\omega\in\mathbb{R}^{\mathrm{d}}} \mathbf{E}_{s\sim d^{\pi_\theta}} \mathbf{E}_{a\sim\pi_\theta(s)}\left[\left(\nabla_\theta \log \pi_\theta(a|s) \cdot \omega - A^{\pi_\theta}_{\mathrm{L},\lambda}(s,a)\right)^2\right] \tag{19}$$

We assume that $L_{d^{\pi^*}, \pi^*}(\omega^*_{\theta,\lambda}, \theta, \lambda) \leq \epsilon_{\mathrm{bias}}$, $\lambda \in [0, \frac{2}{\delta}]$ and $\theta \in \Theta$ where ϵ_{bias} is a positive constant.

Remark 2. The transferred compatible function approximation error quantifies the expressivity of the parameterized policy class. We can show that $\epsilon_{\mathrm{bias}} = 0$ for softmax parameterization [10] and linear MDPs [30]. If the policy class is restricted, i.e., it does not contain all stochastic policies, ϵ_{bias} turns out to be strictly positive. However, if the policy class is parameterized by a rich neural network, then ϵ_{bias} can be assumed to be negligibly small [31]. Such assumptions are common [29, 10].

Remark 3. Note that $\omega^*_{\theta,\lambda}$ defined in (19) can be written as,

$$\omega^*_{\theta,\lambda} = F_\rho(\theta)^\dagger \mathbf{E}_{s\sim d_\rho^{\pi_\theta}} \mathbf{E}_{a\sim\pi_\theta(s)}\left[\nabla_\theta \log \pi_\theta(a|s) A^{\pi_\theta}_{\mathrm{L},\lambda}(s,a)\right]$$

where $\dagger$ is the Moore-Penrose pseudoinverse and $F_\rho(\theta)$ is the Fisher information matrix defined as,

$$F_\rho(\theta) = \mathbf{E}_{s\sim d_\rho^{\pi_\theta}} \mathbf{E}_{a\sim\pi_\theta(\cdot|s)}[\nabla_\theta \log \pi_\theta(a|s)(\nabla_\theta \log \pi_\theta(a|s))^T]$$

Assumption 5. There exists a constant $\mu_F > 0$ such that $F_\rho(\theta) - \mu_F I_{\mathrm{d}}$ is positive semidefinite where I_{d} is an identity matrix of dimension, d.

Assumption 5 is also called Fisher-non-degenerate policy assumption and is quite common in the literature [29, 32, 33] in the policy gradient analysis. [29][Assumption 2.1] provided a detailed discussion on the requirement of policy class to satisfy the assumption 5. Moreover, [34] describes a class of policies that obeys assumptions $3 - 5$ simultaneously. The Lagrange difference lemma stated below is important in establishing global convergence.

https://doi.org/10.52202/079017-3447

Lemma 4. *With a slight abuse of notation, let $J_{\mathrm{L}}(\pi,\lambda) = J_r^{\pi} + \lambda J_c^{\pi}$. For any two policies π, π', the following result holds $\forall \lambda > 0$.*

$$J_{\mathrm{L}}(\pi,\lambda) - J_{\mathrm{L}}(\pi',\lambda) = \mathbf{E}_{s\sim d^{\pi}}\mathbf{E}_{a\sim\pi(s)}\left[A_{\mathrm{L},\lambda}^{\pi'}(s,a)\right]$$

We now present a general framework for the convergence analysis of Algorithm 1.

Lemma 5. *If the policy parameters, $\{\theta_k,\lambda_k\}_{k=1}^{K}$ are updated via* (12) *and assumptions 3, 4,and 5 hold, then we have the following inequality for any K,*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big(J_{\mathrm{L}}(\pi^*,\lambda_k) - J_{\mathrm{L}}(\theta_k,\lambda_k)\Big) \leq \sqrt{\epsilon_{\mathrm{bias}}} + \frac{G}{K}\sum_{k}^{K}\mathbf{E}\|(\omega_k-\omega_k^*)\| + \frac{B\alpha}{2K}\sum_{k=1}^{K}\mathbf{E}\|\omega_k\|^2 + \frac{1}{\alpha K}\mathbf{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))\right]$$

where $\omega_k^* := \omega_{\theta_k,\lambda_k}^*$, $\omega_{\theta_k,\lambda_k}^*$ is defined in (19)*, and π^* is the optimal solution to the problem* (2)*.*

Lemma 5 proves that the optimality error of the Lagrange sequence can be bounded by the average first-order and second-order norms of the intermediate gradients. Note the presence of ϵ_{bias} in the result. If the policy class is severely restricted, the optimality bound loses its importance. Consider the expectation of the second term in (20). Note that,

$$\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\|\omega_k-\omega_k^*\|\right)^2 \leq \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big[\|\omega_k-\omega_k^*\|^2\Big] = \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big[\|\omega_k - F_{\rho}(\theta_k)^{\dagger}\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2\Big]$$

$$\leq \frac{2}{K}\sum_{k=1}^{K}\left\{\mathbf{E}\Big[\|\omega_k - \nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2\Big] + \mathbf{E}\Big[\|\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k) - F_{\rho}(\theta_k)^{\dagger}\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2\Big]\right\}$$

$$\overset{(a)}{\leq} \frac{2}{K}\sum_{k=1}^{K}\mathbf{E}\Big[\|\omega_k - \nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2\Big] + \frac{2}{K}\sum_{k=1}^{K}\left(1+\frac{1}{\mu_F^2}\right)\mathbf{E}\Big[\|\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2\Big]$$

where (a) follows from Assumption 5. The expectation of the third term in (20) can be bounded as

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big[\|\omega_k\|^2\Big] \leq \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left[\|\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2\right] + \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big[\|\omega_k - \nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2\Big]$$

In both (4) and (20), $\mathbf{E}\|\omega_k - \nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2$ is bounded above by Lemma 3. To bound the term, $\mathbf{E}\|\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2$, the following lemma is applied.

Lemma 6. *Let $J_g(\cdot)$ be L-smooth, $\forall g \in \{r,c\}$ and $\alpha = \frac{1}{4L(1+\frac{2}{\delta})}$. Then the following holds.*

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2 \leq \frac{288L}{\delta^2 K} + \frac{3}{K}\sum_{k=1}^{K}\|\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k) - \omega_k\|^2 + \beta \tag{20}$$

Note the presence of β in (20). To ensure convergence, β must be a function of T. Invoking Lemma 3, we get the following relation under the same set of assumptions and the choice of parameters as in Lemma 6.

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\|\nabla_{\theta}J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2 \leq \tilde{\mathcal{O}}\left(\frac{AG^2t_{\mathrm{mix}}^2}{\delta^2 T^{\xi}}\right) + \tilde{\mathcal{O}}\left(\frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}{\delta^2 T^{1-\xi}}\right) + \beta \tag{21}$$

Applying Lemma 3 and (21) in (20), we arrive at,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big[\|\omega_k\|^2\Big] \leq \tilde{\mathcal{O}}\left(\frac{AG^2t_{\mathrm{mix}}^2}{\delta^2 T^{\xi}} + \frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}{\delta^2 T^{1-\xi}}\right) + \beta \tag{22}$$

Similarly, using (4), we deduce the following.

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\|\omega_k-\omega_k^*\| \leq \left(1+\frac{1}{\mu_F}\right)\sqrt{\beta} + \left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\frac{\sqrt{A}Gt_{\mathrm{mix}}}{\delta T^{\xi/2}} + \frac{\sqrt{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}}{\delta T^{(1-\xi)/2}}\right) \tag{23}$$

Inequalities (22) and (23) lead to the following global convergence of the Lagrange function.

Lemma 7. *Let $\{\theta_k\}_{k=1}^K$ be as described in Lemma 5. If assumptions 1−5 hold, $\{J_g(\cdot)\}_{g\in\{r,c\}}$ are L-smooth functions, $\alpha = \frac{1}{4L(1+\frac{2}{\delta})}$, $K = \frac{T}{H}$, and $H = 16t_{\rm mix}t_{\rm hit}T^{\xi}(\log_2 T)^2$, then*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big(J_{\rm L}(\pi^*,\lambda_k) - J_{\rm L}(\theta_k,\lambda_k)\Big) \le G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\sqrt{\beta} + \frac{\sqrt{A}Gt_{\rm mix}}{\delta T^{\xi/2}} + \frac{\sqrt{Lt_{\rm mix}t_{\rm hit}}}{\delta T^{(1-\xi)/2}}\right)$$
$$+ \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{AG^2t_{\rm mix}^2}{\delta^2T^{\xi}} + \frac{Lt_{\rm mix}t_{\rm hit}}{\delta^2T^{1-\xi}} + \beta\right) + \tilde{\mathcal{O}}\left(\frac{Lt_{\rm mix}t_{\rm hit}\mathbf{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))]}{T^{1-\xi}\delta}\right) + \sqrt{\epsilon_{\rm bias}}$$

Lemma 7 establishes that the average difference between $J_{\rm L}(\pi^*,\lambda_k)$ and $J_{\rm L}(\theta_k,\lambda_k)$ is $\tilde{\mathcal{O}}(\sqrt{\beta} + T^{-\xi/2} + T^{-(1-\xi)/2})$. Expanding the function, $J_{\rm L}$, and utilizing the update rule of the Lagrange multiplier, we achieve the global convergence for the objective and the constraint in Theorem 1 (stated below). In its proof, Lemma 18 (stated in the appendix) serves as an important tool in disentangling the convergence rates of regret and constraint violation. Interestingly, Lemma 18 is built upon the strong duality property of the unparameterized optimization (2) and has no apparent direct connection with the parameterized setup.

Theorem 1. *Consider the same parameters as in Lemma 7 and set $\beta = T^{-2/5}$, $\xi = 2/5$. We have,*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big(J_r^{\pi^*} - J_r(\theta_k)\Big) \le \sqrt{\epsilon_{\rm bias}} + \frac{\sqrt{A}G^2t_{\rm mix}}{\delta}\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(T^{-1/5}\right)$$
$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big(-J_c(\theta_k)\Big) \le \delta\sqrt{\epsilon_{\rm bias}} + \tilde{\mathcal{O}}\left(\frac{t_{\rm mix}t_{\rm hit}}{\delta T^{1/5}}\right) + \sqrt{A}G^2t_{\rm mix}\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(T^{-1/5}\right)$$

where π^* is a solution to (2)*. In the above bounds, we write only the dominating terms of T.*

Theorem 1 establishes $\tilde{\mathcal{O}}(T^{-1/5})$ convergence rates for both the objective and the constraint violation.

5 Regret and Violation Analysis

In this section, we use the convergence result of the previous section to bound the expected regret and constraint violation of Algorithm 1. Note that the regret and constraint violation decompose as,

$$\mathrm{Reg}_T = \sum_{t=0}^{T-1}\Big(J_r^{\pi^*} - r(s_t,a_t)\Big) = H\sum_{k=1}^{K}\Big(J_r^{\pi^*} - J(\theta_k)\Big) + \sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}(J(\theta_k) - r(s_t,a_t))$$
$$\mathrm{Vio}_T = \sum_{t=0}^{T-1}(-c(s_t,a_t)) = H\sum_{k=1}^{K}(-J_c(\theta_k)) + \sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}(J_c(\theta_k) - c(s_t,a_t))$$

where $\mathcal{I}_k \triangleq \{(k-1)H,\cdots,kH-1\}$. Observe that the expectation of the first terms in regret and violation can be bounded by Theorem 1. The expectation of the second term in regret and violation can be expanded as follows,

$$\begin{aligned}
&\mathbf{E}\left[\sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}(J_g(\theta_k) - g(s_t,a_t))\right] \overset{(a)}{=} \mathbf{E}\left[\sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}\mathbf{E}_{s'\sim P(\cdot|s_t,a_t)}[V_g^{\pi_{\theta_k}}(s')] - Q_g^{\pi_{\theta_k}}(s_t,a_t)\right]\\
&\overset{(b)}{=} \mathbf{E}\left[\sum_{k=}^{K}\sum_{t\in\mathcal{I}_k}V_g^{\pi_{\theta_k}}(s_{t+1}) - V_g^{\pi_{\theta_k}}(s_t)\right] = \mathbf{E}\left[\sum_{k=1}^{K}V_g^{\pi_{\theta_k}}(s_{kH}) - V_g^{\pi_{\theta_k}}(s_{(k-1)H})\right] \qquad (24)\\
&= \mathbf{E}\left[\sum_{k=1}^{K-1}V_g^{\pi_{\theta_{k+1}}}(s_{kH}) - V_g^{\pi_{\theta_k}}(s_{kH})\right] + \mathbf{E}\left[V_g^{\pi_{\theta_K}}(s_T) - V_g^{\pi_{\theta_0}}(s_0)\right]
\end{aligned}$$

where $g\in\{r,c\}$. Equality (a) uses the Bellman equation and (b) follows from the definition of Q_g. The first term in the last line of Eq. (24) can be upper bounded by Lemma 8 (stated below). On the other hand, the second term can be upper bounded as $\mathcal{O}(t_{\rm mix})$ using Lemma 9.

https://doi.org/10.52202/079017-3447

Lemma 8. *If assumptions 1 and 3 hold, then for $K = \frac{T}{H}$ where $H = 16 t_{\text{mix}} t_{\text{hit}} T^{\frac{2}{5}} (\log_2 T)^2$, the following inequalities hold $\forall k$, $\forall (s, a)$ and sufficiently large T:*

$$(a)\ |\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)| \leq G \|\theta_{k+1} - \theta_k\|$$

$$(b)\ \sum_{k=1}^{K} \mathbf{E}|J_g(\theta_{k+1}) - J_g(\theta_k)| \leq \tilde{\mathcal{O}}\left(\frac{\alpha A G}{\delta t_{\text{hit}}}\left[\left(\sqrt{A} G t_{\text{mix}} + \delta\right) T^{\frac{2}{5}} + \sqrt{L t_{\text{mix}} t_{\text{hit}}} T^{\frac{3}{10}}\right]\right)$$

$$(c)\ \sum_{k=1}^{K} \mathbf{E}|V_g^{\pi_{\theta_{k+1}}}(s_k) - V_g^{\pi_{\theta_k}}(s_k)| \leq \tilde{\mathcal{O}}\left(\frac{\alpha A G t_{\text{mix}}}{\delta t_{\text{hit}}}\left[\left(\sqrt{A} G t_{\text{mix}} + \delta\right) T^{\frac{2}{5}} + \sqrt{L t_{\text{mix}} t_{\text{hit}}} T^{\frac{3}{10}}\right]\right)$$

where $g \in \{r, c\}$, and $\{s_k\}_{k=1}^{K}$ is an arbitrary sequence of states.

Lemma 8 states that the obtained policy parameters are such that the average consecutive difference in the sequence $\{J_g(\theta_k)\}_{k=1}^{K}$, $g \in \{r, c\}$ decreases with time horizon, T. We would like to emphasize that Lemma 8 works for both reward and constraint functions. Hence, we can prove our regret guarantee and constraint violation as shown below.

Theorem 2. *If assumptions 1−5 hold, $J_g(\cdot)$'s are L-smooth, $\forall g \in \{r, c\}$ and T are sufficiently large, then our proposed Algorithm 1 achieves the following expected regret and constraint violation bounds with learning rates $\alpha = \frac{1}{4L(1+\frac{2}{\delta})}$ and $\beta = T^{-2/5}$.*

$$\mathbf{E}\left[\text{Reg}_T\right] \leq T\sqrt{\epsilon_{\text{bias}}} + \tilde{\mathcal{O}}(T^{4/5}) + \mathcal{O}(t_{\text{mix}}) \tag{25}$$

$$\mathbf{E}\left[\text{Vio}_T\right] \leq T\delta\sqrt{\epsilon_{\text{bias}}} + \tilde{\mathcal{O}}(T^{4/5}) + \mathcal{O}(t_{\text{mix}}) \tag{26}$$

The detailed expressions of these bounds are provided in the Appendix. Here, we keep only those terms that emphasize the order of T. Note that our result outperforms the state-of-the-art model-free tabular result in average-reward CMDP [8]. However, our regret bound is worse than that achievable in average reward unconstrained MDP with general parameterization [13]. Interestingly, the gap between the convergence results of constrained and unconstrained setups is a common observation across the literature. For example, in the tabular model-free average reward MDP, the state-of-the-art regret bound for unconstrained setup, $\tilde{\mathcal{O}}(T^{1/2})$ [17], is better than that in the constrained setup, $\tilde{\mathcal{O}}(T^{5/6})$ [8].

6 Conclusions

This paper establishes the first sublinear regret and constraint violation bounds in the average reward CMDP setup with general parametrization (and do not assume the underlying constrained Markov Decision Process (CMDP) to be tabular or linear). We show that our proposed algorithm achieves $\tilde{\mathcal{O}}(T^{4/5})$ regret and constraint violation bounds where T is the time horizon. Note that the state of the art in unconstrained counterpart is $\tilde{\mathcal{O}}(T^{3/4})$. Closing this gap by designing more efficient algorithms is an open question in the average reward CMDP literature with the general parametrization. Moreover, our current algorithm requires the knowledge of mixing time. Relaxing such assumptions is another important future direction in realistic settings. For further discussions on future work directions, the readers are referred to [35].

7 Acknowledgement

This research was supported in part by the National Science Foundation under grant CCF-2149588 and Cisco, Inc.

References

[1] Liu, C., N. Geng, et al. Cmix: Deep multi-agent reinforcement learning with peak and average constraints. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, pages 157–173. Springer, 2021.

[2] Al-Abbasi, A. O., A. Ghosh, V. Aggarwal. Deeppool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4714–4727, 2019.

[3] Ling, L., W. U. Mondal, S. V. Ukkusuri. Cooperating graph neural networks with deep reinforcement learning for vaccine prioritization. *arXiv preprint arXiv:2305.05163*, 2023.

[4] Bai, Q., A. S. Bedi, V. Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6737–6744. 2023.

[5] Agarwal, M., Q. Bai, V. Aggarwal. Concave utility reinforcement learning with zero-constraint violations. *Transactions on Machine Learning Research*, 2022.

[6] Chen, L., R. Jain, H. Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pages 3246–3270. PMLR, 2022.

[7] Agarwal, M., Q. Bai, V. Aggarwal. Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Uncertainty in Artificial Intelligence*, pages 22–31. PMLR, 2022.

[8] Wei, H., X. Liu, L. Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3868–3876. 2022.

[9] Ghosh, A., X. Zhou, N. Shroff. Achieving sub-linear regret in infinite horizon average reward constrained mdp with linear function approximation. In *The Eleventh International Conference on Learning Representations*. 2023.

[10] Agarwal, A., S. M. Kakade, J. D. Lee, G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

[11] Mondal, W. U., V. Aggarwal. Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward markov decision processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2024.

[12] —. Sample-efficient constrained reinforcement learning with general parameterization. *arXiv preprint arXiv:2405.10624*, 2024.

[13] Bai, Q., W. U. Mondal, V. Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024.

[14] Ding, D., K. Zhang, T. Basar, M. Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

[15] Agrawal, S., R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

[16] Auer, P., T. Jaksch, R. Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[17] Wei, C.-Y., M. J. Jahromi, H. Luo, H. Sharma, R. Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR, 2020.

[18] Bai, Q., A. S. Bedi, M. Agarwal, A. Koppel, V. Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3682–3689. 2022.

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

[20] Efroni, Y., S. Mannor, M. Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

 https://doi.org/10.52202/079017-3447

[21] Qiu, S., X. Wei, Z. Yang, J. Ye, Z. Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

[22] Germano, J., F. E. Stradi, G. Genalti, M. Castiglioni, A. Marchesi, N. Gatti. A best-of-both-worlds algorithm for constrained mdps with long-term constraints. *arXiv preprint arXiv:2304.14326*, 2023.

[23] Pesquerel, F., O.-A. Maillard. Imed-rl: Regret optimal learning of ergodic markov decision processes. In *NeurIPS 2022-Thirty-sixth Conference on Neural Information Processing Systems*. 2022.

[24] Gong, H., M. Wang. A duality approach for regret minimization in average-award ergodic markov decision processes. In *Learning for Dynamics and Control*, pages 862–883. PMLR, 2020.

[25] Sutton, R. S., D. McAllester, S. Singh, Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[26] Lattimore, T., C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[27] Agarwal, A., S. M. Kakade, J. D. Lee, G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

[28] Zhang, J., C. Ni, C. Szepesvari, M. Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

[29] Liu, Y., K. Zhang, T. Basar, W. Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

[30] Jin, C., Z. Yang, Z. Wang, M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In J. Abernethy, S. Agarwal, eds., *Proceedings of Thirty Third Conference on Learning Theory*, vol. 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 2020.

[31] Wang, L., Q. Cai, Z. Yang, Z. Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*. 2019.

[32] Yuan, R., R. M. Gower, A. Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR, 2022.

[33] Fatkhullin, I., A. Barakat, A. Kireeva, N. He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869. PMLR, 2023.

[34] Mondal, W. U., V. Aggarwal, S. V. Ukkusuri. Mean-field control based approximation of multi-agent reinforcement learning in presence of a non-decomposable shared global state. *Transactions on Machine Learning Research*, 2023.

[35] Aggarwal, V., W. U. Mondal, Q. Bai. Constrained reinforcement learning with average reward objective: Model-based and model-free algorithms. *Found. Trends Optim.*, 6(4):193–298, 2024.

[36] Dorfman, R., K. Y. Levy. Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.

[37] Ding, D., K. Zhang, J. Duan, T. Başar, M. R. Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. *arXiv preprint arXiv:2206.02346*, 2022.

[38] Bai, Q., V. Aggarwal, A. Gattami. Provably sample-efficient model-free algorithm for mdps with peak constraints. *Journal of Machine Learning Research*, 24(60):1–25, 2023.

A Proofs for Lemmas in Section 3

A.1 Proof of Lemma 1

Since the first step of the proof works in the same way for functions J_r and J_c, we use the generic notations J_g, V_g, Q_g where $g = r, c$ and derive the following.

$$\begin{aligned}
&\nabla_\theta V_g^{\pi_{\theta_k}}(s) = \nabla_\theta \Bigg(\sum_a \pi_\theta(a|s) Q_g^{\pi_\theta}(s,a) \Bigg) \\
&= \sum_a \Bigg(\nabla_\theta \pi_\theta(a|s) \Bigg) Q_g^{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s) \nabla_\theta Q_g^{\pi_\theta}(s,a) \\
&\overset{(a)}{=} \sum_a \pi_\theta(a|s) \Bigg(\nabla_\theta \log \pi_\theta(a|s) \Bigg) Q_g^{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s) \nabla_\theta \Bigg(g(s,a) - J_g(\theta) + \sum_{s'} P(s'|s,a) V_g^{\pi_\theta}(s') \Bigg) \\
&= \sum_a \pi_\theta(a|s) \Bigg(\nabla_\theta \log \pi_\theta(a|s) \Bigg) Q_g^{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s) \Bigg(\sum_{s'} P(s'|s,a) \nabla_\theta V_g^{\pi_\theta}(s') \Bigg) - \nabla_\theta J_g(\theta)
\end{aligned} \tag{27}$$

where the step (a) is a consequence of $\nabla_\theta \log \pi_\theta = \frac{\nabla \pi_\theta}{\pi_\theta}$ and the Bellman equation. Multiplying both sides by $d^{\pi_\theta}(s)$, taking a sum over $s \in \mathcal{S}$, and rearranging the terms, we obtain the following.

$$\begin{aligned}
&\nabla_\theta J_g(\theta) = \sum_s d^{\pi_\theta}(s) \nabla_\theta J_g(\theta) \\
&= \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \Bigg(\nabla_\theta \log \pi_\theta(a|s) \Bigg) Q_g^{\pi_\theta}(s,a) + \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \Bigg(\sum_{s'} P(s'|s,a) \nabla_\theta V_g^{\pi_\theta}(s') \Bigg) \\
&- \sum_s d^{\pi_\theta}(s) \nabla_\theta V_g^{\pi_\theta}(s) \\
&= \mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \Big[Q_g^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \Big] + \sum_s d^{\pi_\theta}(s) \sum_{s'} P^{\pi_\theta}(s'|s) \nabla_\theta V_g^{\pi_\theta}(s') - \sum_s d^{\pi_\theta}(s) \nabla_\theta V_g^{\pi_\theta}(s) \\
&\overset{(a)}{=} \mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \Big[Q_g^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \Big] + \sum_{s'} d^{\pi_\theta}(s') \nabla_\theta V_g^{\pi_\theta}(s') - \sum_s d^{\pi_\theta}(s) \nabla_\theta V_g^{\pi_\theta}(s) \\
&= \mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \Big[Q_g^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \Big]
\end{aligned} \tag{28}$$

where (a) uses the fact that d^{π_θ} is a stationary distribution. Note that,

$$\begin{aligned}
&\mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \Big[V_g^{\pi_\theta}(s) \nabla \log \pi_\theta(a|s) \Big] \\
&= \mathbf{E}_{s \sim d^{\pi_\theta}} \left[\sum_{a \in \mathcal{A}} V_g^{\pi_\theta}(s) \nabla_\theta \pi_\theta(a|s) \right] \\
&= \mathbf{E}_{s \sim d^{\pi_\theta}} \left[V_g^{\pi_\theta}(s) \nabla_\theta \left(\sum_{a \in \mathcal{A}} \pi_\theta(a|s) \right) \right] = \mathbf{E}_{s \sim d^{\pi_\theta}} \Big[V_g^{\pi_\theta}(s) \nabla_\theta(1) \Big] = 0
\end{aligned} \tag{29}$$

We can, therefore, replace the function $Q_g^{\pi_\theta}$ in the policy gradient with the advantage function $A_g^{\pi_\theta}(s,a) = Q_g^{\pi_\theta}(s,a) - V_g^{\pi_\theta}(s), \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. Thus,

$$\nabla_\theta J_g(\theta) = \mathbf{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \Big[A_g^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \Big] \tag{30}$$

The proof is completed using the definitions of $J_{\mathrm{L},\lambda}$ and $A_{\mathrm{L},\lambda}$.

A.2 Proof of Lemma 2

Proof. The proof is similar to the proof of [17, Lemma 6]. Consider the kth epoch and assume that π_{θ_k} is denoted as π for notational convenience. Let, M be the number of disjoint sub-trajectories of

https://doi.org/10.52202/079017-3447

length N that start with the state s and are at least N distance apart (found by Algorithm 2). Let, $g_{k,i}$ be the sum of rewards or constraint ($g = r, c$ accordingly) observed in the ith sub-trajectory and τ_i denote its starting time. The advantage function estimate is,

$$\hat{A}_g^\pi(s,a) = \begin{cases} \frac{1}{\pi(a|s)} \left[\frac{1}{M} \sum_{i=1}^M g_{k,i} 1(a_{\tau_i} = a) \right] - \frac{1}{M} \sum_{i=1}^M g_{k,i} & \text{if } M > 0 \\ 0 & \text{if } M = 0 \end{cases} \tag{31}$$

Note the following,

$$\begin{aligned}
&\mathbf{E}\left[g_{k,i} \middle| s_{\tau_i} = s, a_{\tau_i} = a \right] = g(s,a) + \mathbf{E}\left[\sum_{t=\tau_i+1}^{\tau_i+N} g(s_t, a_t) \middle| s_{\tau_i} = s, a_{\tau_i} = a \right] \\
&= g(s,a) + \sum_{s'} P(s'|s,a) \mathbf{E}\left[\sum_{t=\tau_i+1}^{\tau_i+N} g(s_t, a_t) \middle| s_{\tau_i+1} = s' \right] \\
&= g(s,a) + \sum_{s'} P(s'|s,a) \left[\sum_{j=0}^{N-1} (P^\pi)^j (s', \cdot) \right]^T g^\pi \\
&= g(s,a) + \sum_{s'} P(s'|s,a) \left[\sum_{j=0}^{N-1} (P^\pi)^j (s', \cdot) - d^\pi \right]^T g^\pi + N (d^\pi)^T g^\pi \\
&\overset{(a)}{=} g(s,a) + \sum_{s'} P(s'|s,a) \left[\sum_{j=0}^{\infty} (P^\pi)^j (s', \cdot) - d^\pi \right]^T g^\pi + N J_g^\pi - \underbrace{\sum_{s'} P(s'|s,a) \left[\sum_{j=N}^{\infty} (P^\pi)^j (s', \cdot) - d^\pi \right]^T g^\pi}_{\triangleq \mathrm{E}_T^\pi(s,a)} \\
&\overset{(b)}{=} g(s,a) + \sum_{s'} P(s'|s,a) V_g^\pi(s') + N J_g^\pi - \mathrm{E}_T^\pi(s,a) \overset{(c)}{=} Q_g^\pi(s,a) + (N+1) J_g^\pi - \mathrm{E}_T^\pi(s,a)
\end{aligned} \tag{32}$$

where (a) follows from the definition of J_g^π as given in (5), (b) is an application of the definition of V_g^π given in (7), and (c) follows from the Bellman equation. Define the following quantity.

$$\delta^\pi(s,T) \triangleq \sum_{t=N}^{\infty} \left\| (P^\pi)^t (s, \cdot) - d^\pi \right\|_1 \text{ where } N = 4 t_{\mathrm{mix}} (\log_2 T) \tag{33}$$

Using Lemma 10, we get $\delta^\pi(s,T) \leq \frac{1}{T^3}$ which implies, $|\mathrm{E}_T^\pi(s,a)| \leq \frac{1}{T^3}$. Observe that,

$$\begin{aligned}
&\mathbf{E}\left[\left(\frac{1}{\pi(a|s)} g_{k,i} 1(a_{\tau_i} = a) - g_{k,i} \right) \middle| s_{\tau_i} = s \right] \\
&= \mathbf{E}\left[g_{k,i} \middle| s_{\tau_i} = s, a_{\tau_i} = a \right] - \sum_{a'} \pi(a'|s) \mathbf{E}\left[g_{k,i} \middle| s_{\tau_i} = s, a_{\tau_i} = a' \right] \\
&= Q_g^\pi(s,a) + (N+1) J_g^\pi - \mathrm{E}_T^\pi(s,a) - \sum_{a'} \pi(a'|s) [Q^\pi(s,a) + (N+1) J_g^\pi - \mathrm{E}_T^\pi(s,a)] \\
&= Q_g^\pi(s,a) - V_g^\pi(s) - \left[\mathrm{E}_T(s,a) - \sum_{a'} \pi(a'|s) \mathrm{E}_T^\pi(s,a') \right] \\
&= A_g^\pi(s,a) - \Delta_T^\pi(s,a)
\end{aligned} \tag{34}$$

where $\Delta_T^\pi(s,a) \triangleq \mathrm{E}_T(s,a) - \sum_{a'} \pi(a'|s) \mathrm{E}_T^\pi(s,a')$. Using the bound on $\mathrm{E}_T^\pi(s,a)$, we derive, $|\Delta_T^\pi(s,a)| \leq \frac{2}{T^3}$, which implies,

$$\left| \mathbf{E}\left[\left(\frac{1}{\pi(a|s)} g_{k,i} 1(a_{\tau_i} = a) - g_{k,i} \right) \middle| s_{\tau_i} = s \right] - A_g^\pi(s,a) \right| \leq |\Delta_T^\pi(s,a)| \leq \frac{2}{T^3} \tag{35}$$

Note that (35) cannot be directly used to bound the bias of $\hat{A}_g^\pi(s,a)$. This is because the random variable M is correlated with the variables $\{g_{k,i}\}_{i=1}^M$. To decorrelate them, imagine a CMDP where the state distribution resets to the stationary distribution, d^π after exactly N time steps since the completion of a sub-trajectory. In other words, if a sub-trajectory starts at τ_i, and ends at $\tau_i + N$, then the system 'rests' for additional N steps before rejuvenating with the state distribution, d^π at $\tau_i + 2N$. Clearly, the wait time between the reset after the $(i-1)$th sub-trajectory and the start of the ith sub-trajectory is, $w_i = \tau_i - (\tau_{i-1} + 2N)$, $i > 1$. Let w_1 be the difference between the start time of the kth epoch and the start time of the first sub-trajectory. Note that,

(a) w_1 only depends on the initial state, $s_{(k-1)H}$ and the induced transition function, P^π,

(b) w_i, where $i > 1$, depends on the stationary distribution, d^π, and the induced transition function, P^π,

(c) M only depends on $\{w_1, w_2, \cdots\}$ as other segments of the epoch have fixed length, $2N$.

Clearly, in this imaginary CMDP, the sequence, $\{w_1, w_2, \cdots\}$, and hence, M is independent of $\{g_{k,1}, g_{k,2}, \cdots\}$. Let, $\mathbf{E}'$ denote the expectation operation and $\Pr'$ denote the probability of events in this imaginary system. Define the following.

$$\Delta_i \triangleq \frac{g_{k,i} 1(a_{\tau_i} = a)}{\pi(a|s)} - g_{k,i} - A_g^\pi(s,a) + \Delta_T^\pi(s,a) \tag{36}$$

where $\Delta_T^\pi(s,a)$ is given in (34). Note that we have suppressed the dependence on T, s, a, and π while defining Δ_i to remove clutter. Using (34), one can write $\mathbf{E}'\left[\Delta_i(s,a)|\{w_i\}\right] = 0$. Moreover,

$$\begin{aligned}
&\mathbf{E}'\left[\left(\hat{A}_g^\pi(s,a) - A_g^\pi(s,a)\right)^2\right] \\
&= \mathbf{E}'\left[\left(\hat{A}_g^\pi(s,a) - A_g^\pi(s,a)\right)^2 \Big| M > 0\right] \times \Pr'(M > 0) + \left(A_g^\pi(s,a)\right)^2 \times \Pr'(M = 0) \\
&= \mathbf{E}'\left[\left(\frac{1}{M}\sum_{i=1}^M \Delta_i - \Delta_T^\pi(s,a)\right)^2 \Big| M > 0\right] \times \Pr'(M > 0) + \left(A_g^\pi(s,a)\right)^2 \times \Pr'(M = 0) \\
&\leq 2\mathbf{E}'_{\{w_i\}}\left[\mathbf{E}'\left[\left(\frac{1}{M}\sum_{i=1}^M \Delta_i\right)^2 \Big| \{w_i\}\right] \Big| w_1 \leq H - N\right] \times \Pr'(w_1 \leq H - N) \\
&+ 2\left(\Delta_T^\pi(s,a)\right)^2 + \left(A_g^\pi(s,a)\right)^2 \times \Pr'(M = 0) \\
&\overset{(a)}{\leq} 2\mathbf{E}'_{\{w_i\}}\left[\frac{1}{M^2}\sum_{i=1}^M \mathbf{E}'\left[\Delta_i^2 \big| \{w_i\}\right] \Big| w_1 \leq H - N\right] \times \Pr'(w_1 \leq H - N) \\
&+ \frac{8}{T^6} + \left(A_g^\pi(s,a)\right)^2 \times \Pr'(M = 0)
\end{aligned} \tag{37}$$

where (a) uses the bound $|\Delta_T^\pi(s,a)| \leq \frac{2}{T^3}$ derived in (35), and the fact that $\{\Delta_i\}$ are zero mean independent random variables conditioned on $\{w_i\}$. Note that $|g_{k,i}| \leq N$ almost surely, $|A_g^\pi(s,a)| \leq \mathcal{O}(t_{\text{mix}})$ via Lemma 9, and $|\Delta_T^\pi(s,a)| \leq \frac{2}{T^3}$ as shown in (35). Combining, we get, $\mathbf{E}'[|\Delta_i|^2 \big| \{w_i\}] \leq \mathcal{O}(N^2/\pi(a|s))$ (see the definition of Δ_i in (36)). Invoking this bound into (37), we get the following result.

$$\begin{aligned}
\mathbf{E}'\left[\left(\hat{A}_g^\pi(s,a) - A_g^\pi(s,a)\right)^2\right] &\leq 2\mathbf{E}'\left[\frac{1}{M}\Big| w_1 \leq H - N\right] \mathcal{O}\left(\frac{N^2}{\pi(a|s)}\right) + \frac{8}{T^6} \\
&+ \mathcal{O}(t_{\text{mix}}^2) \times \Pr'(w_1 > H - N)
\end{aligned} \tag{38}$$

Note that, one can use Lemma 11 to bound the following violation probability.

$$\Pr'(w_1 > H - N) \leq \left(1 - \frac{3d^\pi(s)}{4}\right)^{4t_{\text{hit}} T^\xi (\log T) - 1} \overset{(a)}{\leq} \left(1 - \frac{3d^\pi(s)}{4}\right)^{\frac{4}{d^\pi(s)}(\log T)} \leq \frac{1}{T^3} \tag{39}$$

https://doi.org/10.52202/079017-3447

where (a) is a consequence of the fact that $4t_{\rm hit}T^{\xi}(\log_2 T)-1 \geq \frac{4}{d^{\pi}(s)}\log_2 T$ for sufficiently large T. Finally, note that, if $M < M_0$, where M_0 is defined as,

$$M_0 \triangleq \frac{H-N}{2N+\dfrac{4N\log T}{d^{\pi}(s)}} \tag{40}$$

then there exists at least one w_i that exceeds $4N\log_2 T/d^{\pi}(s)$ which can happen with the following maximum probability according to Lemma 11.

$$\Pr{}'\left(M<M_0\right) \leq \left(1-\frac{3d^{\pi}(s)}{4}\right)^{\frac{4\log T}{d^{\pi}(s)}} \leq \frac{1}{T^3} \tag{41}$$

The above probability bound can be used to obtain the following result,

$$\begin{aligned}\mathbf{E}'\left[\frac{1}{M}\middle| M>0\right] = \frac{\sum_{m=1}^{\infty}\frac{1}{m}\Pr{}'(M=m)}{\Pr{}'(M>0)} &\leq \frac{1\times\Pr{}'(M\leq M_0)+\frac{1}{M_0}\Pr{}'(M>M_0)}{\Pr{}'(M>0)} \\ &\leq \frac{\dfrac{1}{T^3}+\dfrac{2N+\dfrac{4N\log T}{d^{\pi}(s)}}{H-N}}{1-\dfrac{1}{T^3}} \leq \mathcal{O}\left(\frac{N\log T}{Hd^{\pi}(s)}\right)\end{aligned} \tag{42}$$

Injecting (39) and (42) into (38), we finally obtain the following.

$$\begin{aligned}\mathbf{E}'\left[\left(\hat{A}_g^{\pi}(s,a)-A_g^{\pi}(s,a)\right)^2\right] &\leq \mathcal{O}\left(\frac{N^3\log T}{Hd^{\pi}(s)\pi(a|s)}\right) \\ &= \mathcal{O}\left(\frac{N^3 t_{\rm hit}\log T}{H\pi(a|s)}\right) = \mathcal{O}\left(\frac{t_{\rm mix}^2(\log T)^2}{T^{\xi}\pi(a|s)}\right)\end{aligned} \tag{43}$$

Eq. (43) demonstrates that our desired inequality is obeyed in the imaginary system. We now need a mechanism to translate this result to our actual CMDP. Note that $(\hat{A}_g^{\pi}(s,a)-A_g^{\pi}(s,a))^2 = f(X)$ where $X=(M,\tau_1,\mathcal{T}_1,\cdots,\tau_M,\mathcal{T}_M)$, and $\mathcal{T}_i=(a_{\tau_i},s_{\tau_i+1},a_{\tau_i+1},\cdots,s_{\tau_i+N},a_{\tau_i+N})$. We have,

$$\frac{\mathbf{E}[f(X)]}{\mathbf{E}'[f(X)]} = \frac{\sum_X f(X)\Pr(X)}{\sum_X f(X)\Pr{}'(X)} \leq \max_X \frac{\Pr(X)}{\Pr{}'(X)} \tag{44}$$

The last inequality uses the non-negativity of $f(\cdot)$. Observe that, for a fixed sequence, X, we have,

$$\begin{aligned}\Pr(X) = \Pr(\tau_1)\times\Pr(\mathcal{T}_1|\tau_1)\times\Pr(\tau_2|\tau_1,\mathcal{T}_1)\times\Pr(\mathcal{T}_2|\tau_2)\times\cdots \\ \times\Pr(\tau_M|\tau_{M-1},\mathcal{T}_{M-1})\times\Pr(\mathcal{T}_M|\tau_M)\times\Pr(s_t\neq s,\forall t\in[\tau_M+2N,kH-N]|\tau_M,\mathcal{T}_M),\end{aligned} \tag{45}$$

$$\begin{aligned}\Pr{}'(X) = \Pr(\tau_1)\times\Pr(\mathcal{T}_1|\tau_1)\times\Pr{}'(\tau_2|\tau_1,\mathcal{T}_1)\times\Pr(\mathcal{T}_2|\tau_2)\times\cdots \\ \times\Pr{}'(\tau_M|\tau_{M-1},\mathcal{T}_{M-1})\times\Pr(\mathcal{T}_M|\tau_M)\times\Pr(s_t\neq s,\forall t\in[\tau_M+2N,kH-N]|\tau_M,\mathcal{T}_M),\end{aligned} \tag{46}$$

The difference between $\Pr(X)$ and $\Pr{}'(X)$ arises because $\Pr(\tau_{i+1}|\tau_i,\mathcal{T}_i)\neq\Pr{}'(\tau_{i+1}|\tau_i,\mathcal{T}_i), \forall i\in\{1,\cdots,M-1\}$. Note that the ratio of these two terms can be bounded as follows,

$$\begin{aligned}&\frac{\Pr(\tau_{i+1}|\tau_i,\mathcal{T}_i)}{\Pr{}'(\tau_{i+1}|\tau_i,\mathcal{T}_i)} \\ &= \frac{\sum_{s'\neq s}\Pr(s_{\tau_i+2N}=s'|\tau_i,\mathcal{T}_i)\times\Pr(s_t\neq s,\forall t\in[\tau_i+2N,\tau_{i+1}-1],s_{\tau_{i+1}}=s|s_{\tau_i+2N}=s')}{\sum_{s'\neq s}\Pr{}'(s_{\tau_i+2N}=s'|\tau_i,\mathcal{T}_i)\times\Pr(s_t\neq s,\forall t\in[\tau_i+2N,\tau_{i+1}-1],s_{\tau_{i+1}}=s|s_{\tau_i+2N}=s')} \\ &\leq \max_{s'}\frac{\Pr(s_{\tau_i+2N}=s'|\tau_i,\mathcal{T}_i)}{\Pr{}'(s_{\tau_i+2N}=s'|\tau_i,\mathcal{T}_i)} \\ &= \max_{s'} 1+\frac{\Pr(s_{\tau_i+2N}=s'|\tau_i,\mathcal{T}_i)-d^{\pi}(s')}{d^{\pi}(s')} \overset{(a)}{\leq} \max_{s'} 1+\frac{1}{T^3 d^{\pi}(s')} \leq 1+\frac{t_{\rm hit}}{T^3} \leq 1+\frac{1}{T^2}\end{aligned} \tag{47}$$

where (a) is a consequence of Lemma 10. We have,

$$\frac{\Pr(X)}{\Pr'(X)} \leq \left(1+\frac{1}{T^2}\right)^M \leq e^{\frac{M}{T^2}} \overset{(a)}{\leq} e^{\frac{1}{T}} \leq \mathcal{O}\left(1+\frac{1}{T}\right) \tag{48}$$

where (a) uses the fact that $M \leq T$. Combining (44) and (48), we get,

$$\begin{aligned}\mathbf{E}\left[\left(\hat{A}_g^\pi(s,a)-A_g^\pi(s,a)\right)^2\right] &\leq \mathcal{O}\left(1+\frac{1}{T}\right)\mathbf{E}'\left[\left(\hat{A}_g^\pi(s,a)-A_g^\pi(s,a)\right)^2\right] \\ &\overset{(a)}{\leq} \mathcal{O}\left(\frac{t_{\text{mix}}^2(\log T)^2}{T^\xi \pi(a|s)}\right)\end{aligned} \tag{49}$$

where (a) follows from (43). Using the definition of $A_{\text{L},\lambda}$, we get,

$$\begin{aligned}&\mathbf{E}\left[\left(\hat{A}_{\text{L},\lambda}^\pi(s,a)-A_{\text{L},\lambda}^\pi(s,a)\right)^2\right] \\ &= \mathbf{E}\left[\left((\hat{A}_r^\pi(s,a)-A_r^\pi(s,a))+\lambda(\hat{A}_c^\pi(s,a)-A_c^\pi(s,a))\right)^2\right] \\ &\leq 2\mathbf{E}\left[\left(\hat{A}_r^\pi(s,a)-A_r^\pi(s,a)\right)^2\right]+2\lambda^2\mathbf{E}\left[\left(\hat{A}_c^\pi(s,a)-A_c^\pi(s,a)\right)^2\right] \leq \mathcal{O}\left(\frac{t_{\text{mix}}^2(\log T)^2}{\delta^2 T^\xi \pi(a|s)}\right)\end{aligned}$$

This concludes the proof. □

B Proofs for the Section of Global Convergence Analysis

B.1 Proof of Lemma 3

Proof. Recall from Eq. (15) that,

$$\omega_k = \frac{1}{H}\sum_{t=t_k}^{t_{k+1}-1} \hat{A}_{\text{L},\lambda}^{\pi_{\theta_k}}(s_t,a_t)\nabla_\theta \log \pi_{\theta_k}(a_t|s_t), \tag{50}$$

Define the following quantity,

$$\bar{\omega}_k = \frac{1}{H}\sum_{t=t_k}^{t_{k+1}-1} A_{\text{L},\lambda}^{\pi_{\theta_k}}(s_t,a_t)\nabla_\theta \log \pi_{\theta_k}(a_t|s_t) \tag{51}$$

where $t_k = (k-1)H$ is the starting time of the kth epoch. Note that the true gradient is given by,

$$\nabla_\theta J_{\text{L},\lambda}(\theta_k) = \mathbf{E}_{s\sim d^{\pi_{\theta_k}}, a\sim\pi_{\theta_k}(\cdot|s)}\left[A_{\text{L},\lambda}^{\pi_{\theta_k}}(s,a)\nabla_\theta \log \pi_\theta(a|s)\right] \tag{52}$$

Using Assumption 3, Lemma 9, and $\lambda \in [0, \frac{2}{\delta}]$, one can exhibit that $|A_{\text{L},\lambda}^{\pi_{\theta_k}}(s,a)\nabla_\theta \log \pi_\theta(a|s)| \leq \mathcal{O}(\frac{t_{\text{mix}}G}{\delta})$, $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$ which implies $|\nabla_\theta J_{\text{L},\lambda}(\theta_k)| \leq \mathcal{O}(\frac{t_{\text{mix}}G}{\delta})$. Applying Lemma 14, one, therefore, arrives at

$$\mathbf{E}\left[\|\bar{\omega}_k - \nabla_\theta J_{L,\lambda}(\theta_k)\|^2\right] \leq \mathcal{O}\left(\frac{1}{\delta^2}G^2 t_{\text{mix}}^2 \log T\right) \times \mathcal{O}\left(\frac{t_{\text{mix}}\log T}{H}\right) = \mathcal{O}\left(\frac{G^2 t_{\text{mix}}^2}{\delta^2 t_{\text{hit}} T^\xi}\right) \tag{53}$$

Finally, the difference, $\mathbf{E}\|\omega_k - \bar{\omega}_k\|^2$ can be bounded as follows.

$$\begin{aligned}&\mathbf{E}\|\omega_k - \bar{\omega}_k\|^2 \\ &= \mathbf{E}\left[\left\|\frac{1}{H}\sum_{t=t_k}^{t_{k+1}-1}\hat{A}_{\text{L},\lambda}^{\pi_{\theta_k}}(s_t,a_t)\nabla_\theta \log \pi_{\theta_k}(a_t|s_t) - \frac{1}{H}\sum_{t=t_k}^{t_{k+1}-1}\hat{A}_{\text{L},\lambda}^{\pi_{\theta_k}}(s_t,a_t)\nabla_\theta \log \pi_{\theta_k}(a_t|s_t)\right\|^2\right] \\ &\overset{(a)}{\leq} \frac{G^2}{H}\sum_{t=t_k}^{t_{k+1}-1}\mathbf{E}\left[\left(\hat{A}_{\text{L},\lambda}^{\pi_{\theta_k}}(s_t,a_t) - A_{\text{L},\lambda}^{\pi_{\theta_k}}(s_t,a_t)\right)^2\right] \\ &\leq \frac{G^2}{H}\sum_{t=t_k}^{t_{k+1}-1}\mathbf{E}\left[\sum_a \pi_{\theta_k}(a|s_t)\mathbf{E}\left[\left(\hat{A}_{\text{L},\lambda}^{\pi_{\theta_k}}(s_t,a) - A_{\text{L},\lambda}^{\pi_{\theta_k}}(s_t,a)\right)^2 \middle| s_t\right]\right] \overset{(b)}{\leq} \mathcal{O}\left(\frac{AG^2 t_{\text{mix}}^2(\log T)^2}{\delta^2 T^\xi}\right)\end{aligned} \tag{54}$$

https://doi.org/10.52202/079017-3447

where (a) follows from Assumption 3 and Jensen's inequality whereas (b) follows from Lemma 2. Combining, (53) and (54), we conclude the result. □

B.2 Proof of Lemma 4

Proof. Using the Lemma 12, it is obvious to see that

$$\begin{aligned} J_g^{\pi} - J_g^{\pi'} &= \sum_s \sum_a d^{\pi}(s)(\pi(a|s) - \pi'(a|s))Q_g^{\pi'}(s,a) \\ &= \sum_s \sum_a d^{\pi}(s)\pi(a|s)Q_g^{\pi'}(s,a) - \sum_s d^{\pi}(s)V_g^{\pi'}(s) \\ &= \sum_s \sum_a d^{\pi}(s)\pi(a|s)Q_g^{\pi'}(s,a) - \sum_s \sum_a d^{\pi}(s)\pi(a|s)V_g^{\pi'}(s) \\ &= \sum_s \sum_a d^{\pi}(s)\pi(a|s)[Q_g^{\pi'}(s,a) - V_g^{\pi'}(s)] = \mathbf{E}_{s\sim d^{\pi}}\mathbf{E}_{a\sim\pi(\cdot|s)}\left[A_g^{\pi'}(s,a)\right] \end{aligned} \tag{55}$$

We conclude the lemma using the definition of $J_{\mathrm{L},\lambda}$ and $A_{\mathrm{L},\lambda}$. □

B.3 Proof of Lemma 5

Proof. We start with the definition of KL divergence.

$$\begin{aligned} &\mathbf{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_k}(\cdot|s)) - KL(\pi^*(\cdot|s)\|\pi_{\theta_{k+1}}(\cdot|s))\right] \\ &= \mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}\left[\log\frac{\pi_{\theta_{k+1}}(a|s)}{\pi_{\theta_k}(a|s)}\right] \\ &\overset{(a)}{\geq} \mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}[\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot(\theta_{k+1}-\theta_k)] - \frac{B}{2}\|\theta_{k+1}-\theta_k\|^2 \\ &= \alpha\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}[\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot\omega_k] - \frac{B\alpha^2}{2}\|\omega_k\|^2 \\ &= \alpha\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}[\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot\omega_k^*] + \alpha\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}[\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot(\omega_k-\omega_k^*)] - \frac{B\alpha^2}{2}\|\omega_k\|^2 \\ &= \alpha[J_{\mathrm{L}}(\pi^*,\lambda_k) - J_{\mathrm{L}}(\theta_k,\lambda_k)] + \alpha\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}[\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot\omega_k^*] - \alpha[J_{\mathrm{L}}(\pi^*,\lambda_k) - J_{\mathrm{L}}(\theta_k,\lambda_k)] \\ &+ \alpha\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}[\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot(\omega_k-\omega_k^*)] - \frac{B\alpha^2}{2}\|\omega_k\|^2 \\ &\overset{(b)}{=} \alpha[J_{\mathrm{L}}(\pi^*,\lambda_k) - J_{\mathrm{L}}(\theta_k,\lambda_k)] + \alpha\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}\left[\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot\omega_k^* - A_{\mathrm{L},\lambda_k}^{\pi_{\theta_k}}(s,a)\right] \\ &+ \alpha\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}[\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot(\omega_k-\omega_k^*)] - \frac{B\alpha^2}{2}\|\omega_k\|^2 \\ &\overset{(c)}{\geq} \alpha[J_{\mathrm{L}}(\pi^*,\lambda_k) - J_{\mathrm{L}}(\theta_k,\lambda_k)] - \alpha\sqrt{\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}\left[\left(\nabla_\theta \log\pi_{\theta_k}(a|s)\cdot\omega_k^* - A_{\mathrm{L},\lambda_k}^{\pi_{\theta_k}}(s,a)\right)^2\right]} \\ &- \alpha\mathbf{E}_{s\sim d^{\pi^*}}\mathbf{E}_{a\sim\pi^*(\cdot|s)}\|\nabla_\theta \log\pi_{\theta_k}(a|s)\|_2\|(\omega_k-\omega_k^*)\| - \frac{B\alpha^2}{2}\|\omega_k\|^2 \\ &\overset{(d)}{\geq} \alpha[J_{\mathrm{L}}(\pi^*,\lambda_k) - J_{\mathrm{L}}(\theta_k,\lambda_k)] - \alpha\sqrt{\epsilon_{\mathrm{bias}}} - \alpha G\|(\omega_k-\omega_k^*)\| - \frac{B\alpha^2}{2}\|\omega_k\|^2 \end{aligned} \tag{56}$$

where the step (a) holds by Assumption 3 and step (b) holds by Lemma 4. Step (c) uses the convexity of the function $f(x) = x^2$. Finally, step (d) comes from the Assumption 4. Rearranging items, we have

$$\begin{aligned} J_{\mathrm{L}}(\pi^*,\lambda_k) - J_{\mathrm{L}}(\theta_k,\lambda_k) \leq \sqrt{\epsilon_{\mathrm{bias}}} + G\|(\omega_k-\omega_k^*)\| + \frac{B\alpha}{2}\|\omega_k\|^2 \\ + \frac{1}{\alpha}\mathbf{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_k}(\cdot|s)) - KL(\pi^*(\cdot|s)\|\pi_{\theta_{k+1}}(\cdot|s))\right] \end{aligned} \tag{57}$$

Summing from $k = 1$ to K, using the non-negativity of KL divergence and dividing the resulting expression by K, we get the desired result. □

B.4 Proof of Lemma 6

Proof. By the L-smooth property of the objective function and constraint function, we know that $J_{\mathrm{L}}(\cdot, \lambda)$ is a $L(1+\lambda)$-smooth function. Thus,

$$
\begin{aligned}
&J_{\mathrm{L}}(\theta_{k+1}, \lambda_k) \geq J_{\mathrm{L}}(\theta_k, \lambda_k) + \langle \nabla J_{\mathrm{L}}(\theta_k, \lambda_k), \theta_{k+1} - \theta_k \rangle - \frac{L(1+\lambda_k)}{2} \|\theta_{k+1} - \theta_k\|^2 \\
&\overset{(a)}{=} J_{\mathrm{L}}(\theta_k, \lambda_k) + \alpha \nabla J_{\mathrm{L}}(\theta_k, \lambda_k)^T \omega_k - \frac{L(1+\lambda_k)\alpha^2}{2} \|\omega_k\|^2 \\
&= J_{\mathrm{L}}(\theta_k, \lambda_k) + \alpha \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 - \alpha \langle \nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k, \nabla J_{\mathrm{L}}(\theta_k, \lambda_k) \rangle \\
&\quad - \frac{L(1+\lambda_k)\alpha^2}{2} \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k - \nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \\
&\overset{(b)}{\geq} J_{\mathrm{L}}(\theta_k, \lambda_k) + \alpha \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 - \frac{\alpha}{2} \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2 - \frac{\alpha}{2} \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \\
&\quad - L(1+\lambda_k)\alpha^2 \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2 - L(1+\lambda_k)\alpha^2 \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \\
&= J_{\mathrm{L}}(\theta_k, \lambda_k) + \left(\frac{\alpha}{2} - L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 - \left(\frac{\alpha}{2} + L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2
\end{aligned}
\tag{58}
$$

where step (a) follows from the fact that $\theta_{k+1} = \theta_k + \alpha\omega_k$ and inequality (b) holds due to the Cauchy-Schwarz inequality. Now, adding $J_{\mathrm{L}}(\theta_{k+1}, \lambda_{k+1})$ on both sides, we have

$$
\begin{aligned}
J_{\mathrm{L}}(\theta_{k+1}, \lambda_{k+1}) &\geq J_{\mathrm{L}}(\theta_{k+1}, \lambda_{k+1}) - J_{\mathrm{L}}(\theta_{k+1}, \lambda_k) + J_{\mathrm{L}}(\theta_k, \lambda_k) + \left(\frac{\alpha}{2} - L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \\
&\quad - \left(\frac{\alpha}{2} + L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2 \\
&\overset{(a)}{=} (\lambda_{k+1} - \lambda_k) J_c(\theta_{k+1}) + J_{\mathrm{L}}(\theta_k, \lambda_k) + \left(\frac{\alpha}{2} - L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \\
&\quad - \left(\frac{\alpha}{2} + L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2 \\
&\overset{(b)}{\geq} -\beta + J_{\mathrm{L}}(\theta_k, \lambda_k) + \left(\frac{\alpha}{2} - L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \\
&\quad - \left(\frac{\alpha}{2} + L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2
\end{aligned}
\tag{59}
$$

where (a) holds by the definition of $J_{\mathrm{L}}(\theta, \lambda)$ and step (b) is true because $|J_c(\theta)| \leq 1, \forall \theta$ and $|\lambda_{k+1} - \lambda_k| \leq \beta |\hat{J}_c(\theta_k)| \leq \beta$ where the last inequality uses the fact that $|\hat{J}_c(\theta_k)| \leq 1$. Summing over $k \in \{1, \cdots, K\}$, we have,

$$
\begin{aligned}
\sum_{k=1}^{K} \Big[J_{\mathrm{L}}(\theta_{k+1}, \lambda_{k+1}) - J_{\mathrm{L}}(\theta_k, \lambda_k) \Big] &\geq -\beta K + \sum_{k=1}^{K} \left(\frac{\alpha}{2} - L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \\
&\quad - \sum_{k=1}^{K} \left(\frac{\alpha}{2} + L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2
\end{aligned}
\tag{60}
$$

which leads to the following.

$$
\begin{aligned}
J_{\mathrm{L}}(\theta_{K+1}, \lambda_{K+1}) - J_{\mathrm{L}}(\theta_1, \lambda_1) &\geq -\beta K + \sum_{k=1}^{K} \left(\frac{\alpha}{2} - L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \\
&\quad - \sum_{k=1}^{K} \left(\frac{\alpha}{2} + L(1+\lambda_k)\alpha^2\right) \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2
\end{aligned}
\tag{61}
$$

Rearranging the terms and using $0 \leq \lambda_k \leq \frac{2}{\delta}$ due to the dual update, we arrive at the following.

$$
\sum_{k=1}^{K} \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k)\|^2 \leq \frac{J_{\mathrm{L}}(\theta_{K+1}, \lambda_{K+1}) - J_{\mathrm{L}}(\theta_1, \lambda_1) + \beta K + (\frac{\alpha}{2} + L(1+\frac{2}{\delta})\alpha^2) \sum_{k=1}^{K} \|\nabla J_{\mathrm{L}}(\theta_k, \lambda_k) - \omega_k\|^2}{\frac{\alpha}{2} - L(1+\frac{2}{\delta})\alpha^2}
\tag{62}
$$

https://doi.org/10.52202/079017-3447

Choosing $\alpha = \frac{1}{4L(1+\frac{2}{\delta})}$ and dividing both sides by K, we conclude the result.

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\|\nabla J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2 \leq \frac{16L(1+\frac{2}{\delta})}{K}\left[J_{\mathrm{L}}(\theta_{K+1},\lambda_{K+1}) - J_{\mathrm{L}}(\theta_1,\lambda_1)\right] \\ + \frac{3}{K}\sum_{k=1}^{K}\|\nabla J_{\mathrm{L}}(\theta_k,\lambda_k) - \omega_k\|^2 + \beta\end{aligned} \tag{63}$$

Recall that $|J_{\mathrm{L}}(\theta,\lambda)| \leq 1+\lambda \leq 1+\frac{2}{\delta} \leq \frac{3}{\delta}, \forall\theta\in\Theta, \forall\lambda\geq 0$. Thus,

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla J_{\mathrm{L}}(\theta_k,\lambda_k)\|^2 \leq \frac{288L}{\delta^2 K} + \frac{3}{K}\sum_{k=1}^{K}\|\nabla J_{\mathrm{L}}(\theta_k,\lambda_k) - \omega_k\|^2 + \beta \tag{64}$$

This completes the proof. □

B.5 Proof of Theorem 1

B.5.1 Rate of Convergence of the Objective

Recall the definition of $J_{\mathrm{L}}(\theta,\lambda) = J_r(\theta) + \lambda J_c(\theta)$. Using Lemma 7, we arrive at the following.

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left(J_r^{\pi^*} - J_r(\theta_k)\right) \leq G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\sqrt{\beta} + \frac{\sqrt{A}Gt_{\mathrm{mix}}}{\delta T^{\xi/2}} + \frac{\sqrt{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}}{\delta T^{(1-\xi)/2}}\right) \\ &+ \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{AG^2t_{\mathrm{mix}}^2}{\delta^2 T^{\xi}} + \frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}{\delta^2 T^{1-\xi}} + \beta\right) + \tilde{\mathcal{O}}\left(\frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}\mathbf{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))\right]}{T^{1-\xi}\delta}\right) \\ &- \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left[\lambda_k\left(J_c^{\pi^*} - J_c(\theta_k)\right)\right] + \sqrt{\epsilon_{\mathrm{bias}}}\end{aligned} \tag{65}$$

Thus, we need to find a bound for the last term in the above equation.

$$\begin{aligned}0 &\leq (\lambda_{K+1})^2 \\ &\overset{(a)}{=} \sum_{k=1}^{K}\left((\lambda_{k+1})^2 - (\lambda_k)^2\right) \\ &= \sum_{k=1}^{K}\left(\mathcal{P}_{[0,\frac{2}{\delta}]}\left[\lambda_k - \beta\hat{J}_c(\theta_k)\right]^2 - (\lambda_k)^2\right) \\ &\leq \sum_{k=1}^{K}\left(\left[\lambda_k - \beta\hat{J}_c(\theta_k)\right]^2 - (\lambda_k)^2\right) \\ &= -2\beta\sum_{k=1}^{K}\lambda_k\hat{J}_c(\theta_k) + \beta^2\sum_{k=1}^{K}\hat{J}_c(\theta_k)^2 \\ &\overset{(b)}{\leq} 2\beta\sum_{k=1}^{K}\lambda_k(J_c^{\pi^*} - \hat{J}_c(\theta_k)) + \beta^2\sum_{k=1}^{K}\hat{J}_c(\theta_k)^2 \\ &\leq 2\beta\sum_{k=1}^{K}\lambda_k(J_c^{\pi^*} - \hat{J}_c(\theta_k)) + 2\beta^2\sum_{k=1}^{K}\hat{J}_c(\theta_k)^2 \\ &= 2\beta\sum_{k=1}^{K}\lambda_k(J_c^{\pi^*} - J_c(\theta_k)) + 2\beta\sum_{k=1}^{K}\lambda_k(J_c(\theta_k) - \hat{J}_c(\theta_k)) + 2\beta^2\sum_{k=1}^{K}\hat{J}_c(\theta_k)^2\end{aligned} \tag{66}$$

where (a) uses $\lambda_1 = 0$ and inequality (b) holds because θ^* is a feasible solution to the constrained optimization problem. Rearranging items and taking the expectation, we have,

$$
\begin{aligned}
-\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big[\lambda_k(J_c^{\pi^*}-J_c(\theta_k))\Big] &\le \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big[\lambda_k(J_c(\theta_k)-\hat{J}_c(\theta_k))\Big]+\frac{\beta}{K}\sum_{k=1}^{K}\mathbf{E}[\hat{J}_c(\theta_k)]^2 \\
&\overset{(a)}{\le} \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left[\lambda_k\left(J_c(\theta_k)-\hat{J}_c(\theta_k)\right)\right]+\beta \\
&\overset{(b)}{=} \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left[\lambda_k\left(J_c(\theta_k)-\mathbf{E}\left[\hat{J}_c(\theta_k)\middle|\theta_k\right]\right)\right]+\beta \\
&\le \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left[\lambda_k\left|J_c(\theta_k)-\mathbf{E}\left[\hat{J}_c(\theta_k)\middle|\theta_k\right]\right|\right]+\beta \overset{(c)}{\le} \frac{2}{\delta T^2}+\beta
\end{aligned}
\tag{67}
$$

where (a) results from $|\hat{J}_{c,\rho}(\theta)|^2 \le 1$, $\forall\theta\in\Theta$ and (b) uses the fact that $\hat{J}_{c,\rho}(\theta_k)$ and λ_k are conditionally independent given θ_k. Finally, (c) is a consequence of Lemma 13. Combining (67) with (65), we deduce,

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big(J_r^{\pi^*}-J_r(\theta_k)\Big) \\
&\le \sqrt{\epsilon_{\text{bias}}}+G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\sqrt{\beta}+\frac{\sqrt{A}Gt_{\text{mix}}}{\delta T^{\xi/2}}+\frac{\sqrt{Lt_{\text{mix}}t_{\text{hit}}}}{\delta T^{(1-\xi)/2}}\right)+\mathcal{O}\left(\frac{1}{\delta T^2}+\beta\right) \\
&+\frac{B}{L}\tilde{\mathcal{O}}\left(\frac{AG^2t_{\text{mix}}^2}{\delta^2T^{\xi}}+\frac{Lt_{\text{mix}}t_{\text{hit}}}{\delta^2T^{1-\xi}}+\beta\right)+\tilde{\mathcal{O}}\left(\frac{Lt_{\text{mix}}t_{\text{hit}}\mathbf{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))]}{T^{1-\xi}\delta}\right) \\
&\le \sqrt{\epsilon_{\text{bias}}}+G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\sqrt{\beta}+\frac{\sqrt{A}Gt_{\text{mix}}}{\delta T^{\xi/2}}+\frac{\sqrt{Lt_{\text{mix}}t_{\text{hit}}}}{\delta T^{(1-\xi)/2}}\right)
\end{aligned}
\tag{68}
$$

The last inequality presents only the dominant terms of β and T.

B.5.2 Rate of Constraint Violation

Since $\{\lambda_k\}_{k=1}^K$ are derived by applying the dual update in Algorithm 1, we have,

$$
\begin{aligned}
&\mathbf{E}\left|\lambda_{k+1}-\frac{2}{\delta}\right|^2 \overset{(a)}{\le} \mathbf{E}\left|\lambda_k-\beta\hat{J}_c(\theta_k)-\frac{2}{\delta}\right|^2 \\
&=\mathbf{E}\left|\lambda_k-\frac{2}{\delta}\right|^2-2\beta\mathbf{E}\left[\hat{J}_c(\theta_k)\left(\lambda_k-\frac{2}{\delta}\right)\right]+\beta^2\mathbf{E}\left[\hat{J}_c^2(\theta_k)\right] \\
&\overset{(b)}{\le} \mathbf{E}\left|\lambda_k-\frac{2}{\delta}\right|^2-2\beta\mathbf{E}\left[J_c(\theta_k)\left(\lambda_k-\frac{2}{\delta}\right)\right]-2\beta\mathbf{E}\left[\left(\hat{J}_c(\theta_k)-J_c(\theta_k)\right)\left(\lambda_k-\frac{2}{\delta}\right)\right]+\beta^2 \\
&\overset{(c)}{=} \mathbf{E}\left|\lambda_k-\frac{2}{\delta}\right|^2-2\beta\mathbf{E}\left[J_c(\theta_k)\left(\lambda_k-\frac{2}{\delta}\right)\right]-2\beta\mathbf{E}\left[\left(\mathbf{E}\left[\hat{J}_c(\theta_k)\middle|\theta_k\right]-J_c(\theta_k)\right)\left(\lambda_k-\frac{2}{\delta}\right)\right]+\beta^2 \\
&\le \mathbf{E}\left|\lambda_k-\frac{2}{\delta}\right|^2-2\beta\mathbf{E}\left[J_c(\theta_k)\left(\lambda_k-\frac{2}{\delta}\right)\right]+2\beta\mathbf{E}\left[\left|\mathbf{E}\left[\hat{J}_c(\theta_k)\middle|\theta_k\right]-J_c(\theta_k)\right|\left|\lambda_k-\frac{2}{\delta}\right|\right]+\beta^2 \\
&\overset{(d)}{\le} \mathbf{E}\left|\lambda_k-\frac{2}{\delta}\right|^2-2\beta\mathbf{E}\left[J_c(\theta_k)\left(\lambda_k-\frac{2}{\delta}\right)\right]+\frac{4\beta}{\delta T^2}+\beta^2
\end{aligned}
\tag{69}
$$

where (a) is due to the non-expansiveness of the projection $\mathcal{P}_{[0,\frac{2}{\delta}]}$ and (b) holds because $\hat{J}_c(\theta)\in[0,1]$, $\forall\theta\in\Theta$ according to its definition in Algorithm 1. Finally, (c) is a consequence of the fact that $\hat{J}_c(\theta_k)$

https://doi.org/10.52202/079017-3447

and λ_k are conditionally independent given θ_k whereas (d) applies $|\lambda_k - \frac{2}{\delta}| \leq \frac{2}{\delta}$ and Lemma 13. Averaging (69) over $k \in \{1, \dots, K\}$, we get,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left[J_c(\theta_k)\left(\lambda_k - \frac{2}{\delta}\right)\right] \leq \frac{\left|\lambda_1 - \frac{2}{\delta}\right|^2 - \left|\lambda_{K+1} - \frac{2}{\delta}\right|^2}{2\beta K} + \frac{2}{\delta T^2} + \frac{\beta}{2} \overset{(a)}{\leq} \frac{2}{\delta^2\beta K} + \frac{2}{\delta T^2} + \frac{\beta}{2} \tag{70}$$

where (a) uses $\lambda_1 = 0$. Note that $\lambda_k J_c^{\pi^*} \geq 0, \forall k$. Adding the above inequality to (65) at both sides, we, therefore, have,

$$\begin{aligned}
&\mathbf{E}\left[J_r^{\pi^*} - \frac{1}{K}\sum_{k=1}^{K}J_r(\theta_k)\right] + \frac{2}{\delta}\mathbf{E}\left[\frac{1}{K}\sum_{k=1}^{K} -J_c(\theta_k)\right] \leq \sqrt{\epsilon_{\text{bias}}} + \frac{2}{\delta^2\beta K} + \frac{2}{T^2\delta} + \frac{\beta}{2} \\
&+ G\left(1 + \frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\sqrt{\beta} + \frac{\sqrt{A}Gt_{\text{mix}}}{\delta T^{\xi/2}} + \frac{\sqrt{Lt_{\text{mix}}t_{\text{hit}}}}{\delta T^{(1-\xi)/2}}\right) + \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{AG^2t_{\text{mix}}^2}{\delta^2T^{\xi}} + \frac{Lt_{\text{mix}}t_{\text{hit}}}{\delta^2T^{1-\xi}} + \beta\right) \\
&+ \tilde{\mathcal{O}}\left(\frac{Lt_{\text{mix}}t_{\text{hit}}\mathbf{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))\right]}{T^{1-\xi}\delta}\right)
\end{aligned} \tag{71}$$

Since the functions $\{J_g(\theta_k)\}, k \in \{0, \cdots, K-1\}, g \in \{r, c\}$ are linear in occupancy measure, there exists a policy $\bar{\pi}$ such that the following holds $\forall g \in \{r, c\}$.

$$\frac{1}{K}\sum_{k=1}^{K}J_g(\theta_k) = J_g^{\bar{\pi}} \tag{72}$$

Injecting the above relation to (71), we have

$$\begin{aligned}
&\mathbf{E}\left[J_r^{\pi^*} - J_r^{\bar{\pi}}\right] + \frac{2}{\delta}\mathbf{E}\left[-J_c^{\bar{\pi}}\right] \leq \sqrt{\epsilon_{\text{bias}}} + \frac{2}{\delta^2\beta K} + \frac{2}{T^2\delta} + \frac{\beta}{2} \\
&+ G\left(1 + \frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\sqrt{\beta} + \frac{\sqrt{A}Gt_{\text{mix}}}{\delta T^{\xi/2}} + \frac{\sqrt{Lt_{\text{mix}}t_{\text{hit}}}}{\delta T^{(1-\xi)/2}}\right) \\
&+ \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{AG^2t_{\text{mix}}^2}{\delta^2T^{\xi}} + \frac{Lt_{\text{mix}}t_{\text{hit}}}{\delta^2T^{1-\xi}} + \beta\right) + \tilde{\mathcal{O}}\left(\frac{Lt_{\text{mix}}t_{\text{hit}}\mathbf{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))\right]}{T^{1-\xi}\delta}\right)
\end{aligned} \tag{73}$$

By Lemma 18, we arrive at,

$$\begin{aligned}
&\mathbf{E}\left[-J_c^{\bar{\pi}}\right] \\
&\leq \delta\sqrt{\epsilon_{\text{bias}}} + \frac{2}{\delta\beta K} + \frac{2}{T^2} + \frac{\delta\beta}{2} + G\left(1 + \frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\delta\sqrt{\beta} + \frac{\sqrt{A}Gt_{\text{mix}}}{T^{\xi/2}} + \frac{\sqrt{Lt_{\text{mix}}t_{\text{hit}}}}{T^{(1-\xi)/2}}\right) \\
&+ \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{AG^2t_{\text{mix}}^2}{\delta T^{\xi}} + \frac{Lt_{\text{mix}}t_{\text{hit}}}{\delta T^{1-\xi}} + \delta\beta\right) + \tilde{\mathcal{O}}\left(\frac{Lt_{\text{mix}}t_{\text{hit}}\mathbf{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))\right]}{T^{1-\xi}}\right) \\
&\leq \delta\sqrt{\epsilon_{\text{bias}}} + \tilde{\mathcal{O}}\left(\frac{2t_{\text{mix}}t_{\text{hit}}}{\delta\beta T^{1-\xi}}\right) + G\left(1 + \frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\delta\sqrt{\beta} + \frac{\sqrt{A}Gt_{\text{mix}}}{T^{\xi/2}} + \frac{\sqrt{Lt_{\text{mix}}t_{\text{hit}}}}{T^{(1-\xi)/2}}\right)
\end{aligned} \tag{74}$$

The last inequality presents only the dominant terms of β and T.

B.5.3 Optimal Choice of β and ξ

If we choose $\beta = T^{-\eta}$ for some $\eta \in (0, 1)$, then following (68) and (74), we can write,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left(J_r^{\pi^*} - J_r(\theta_k)\right) \leq \sqrt{\epsilon_{\text{bias}}} + \tilde{\mathcal{O}}\left(T^{-\eta/2} + T^{-\xi/2} + T^{-(1-\xi)/2}\right), \tag{75}$$

$$\mathbf{E}\left[\frac{1}{K}\sum_{k=1}^{K} -J_c(\theta_k)\right] \leq \delta\sqrt{\epsilon_{\text{bias}}} + \tilde{\mathcal{O}}\left(T^{-(1-\xi-\eta)} + T^{-\eta/2} + T^{-\xi/2} + T^{-(1-\xi)/2}\right) \tag{76}$$

Clearly, the optimal values of η and ξ can be obtained by solving the following optimization.

$$\max_{(\eta,\xi)\in(0,1)^2} \min\left\{1-\xi-\eta, \frac{\eta}{2}, \frac{\xi}{2}, \frac{1-\xi}{2}\right\} \tag{77}$$

One can easily verify that $(\xi, \eta) = (2/5, 2/5)$ is the solution of the above optimization. Therefore, the convergence rate of the objective function can be written as follows.

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big(J_r^{\pi^*} - J_r(\theta_k)\Big) \\
&\leq \sqrt{\epsilon_{\text{bias}}} + G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\frac{1}{T^{1/5}} + \frac{\sqrt{A}Gt_{\text{mix}}}{\delta T^{1/5}} + \frac{\sqrt{Lt_{\text{mix}}t_{\text{hit}}}}{\delta T^{3/10}}\right) + \mathcal{O}\left(\frac{1}{\delta T^2} + \frac{1}{T^{2/5}}\right) \\
&+ \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{\delta^2 + AG^2t_{\text{mix}}^2}{\delta^2 T^{2/5}} + \frac{Lt_{\text{mix}}t_{\text{hit}}}{\delta^2 T^{3/5}}\right) + \tilde{\mathcal{O}}\Big(\frac{Lt_{\text{mix}}t_{\text{hit}}\mathbf{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))]}{T^{3/5}\delta}\Big) \\
&\leq \sqrt{\epsilon_{\text{bias}}} + \frac{\sqrt{A}G^2t_{\text{mix}}}{\delta}\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(T^{-1/5}\right)
\end{aligned} \tag{78}$$

The last expression only considers the dominant terms of T. Similarly, the constraint violation rate can be computed as,

$$\begin{aligned}
&\mathbf{E}\Big[\frac{1}{K}\sum_{k=1}^{K} -J_c(\theta_k)\Big] \\
&\leq \delta\sqrt{\epsilon_{\text{bias}}} + \tilde{\mathcal{O}}\left(\frac{t_{\text{mix}}t_{\text{hit}}}{\delta T^{1/5}} + \frac{1}{T^2} + \frac{\delta}{T^{2/5}}\right) + G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\frac{\delta + \sqrt{A}Gt_{\text{mix}}}{T^{1/5}} + \frac{\sqrt{Lt_{\text{mix}}t_{\text{hit}}}}{T^{3/10}}\right) \\
&+ \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{\delta^2 + AG^2t_{\text{mix}}^2}{\delta T^{2/5}} + \frac{Lt_{\text{mix}}t_{\text{hit}}}{\delta T^{3/5}}\right) + \tilde{\mathcal{O}}\Big(\frac{Lt_{\text{mix}}t_{\text{hit}}\mathbf{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))]}{T^{3/5}}\Big) \\
&\leq \delta\sqrt{\epsilon_{\text{bias}}} + \tilde{\mathcal{O}}\left(\frac{t_{\text{mix}}t_{\text{hit}}}{\delta T^{1/5}}\right) + \sqrt{A}G^2t_{\text{mix}}\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(T^{-1/5}\right)
\end{aligned} \tag{79}$$

where the last expression contains only the dominant terms of T. This concludes the theorem.

C Proofs for the Regret and Violation Analysis

C.1 Proof of Lemma 8

Proof. Using Taylor's expansion, we can write the following $\forall(s,a) \in \mathcal{S}\times\mathcal{A}, \forall k$.

$$\begin{aligned}
|\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)| &= \left|(\theta_{k+1}-\theta_k)^T\nabla_\theta\pi_{\bar{\theta}}(a|s)\right| \\
&= \pi_{\bar{\theta}_k}(a|s)\left|(\theta_{k+1}-\theta_k)^T\nabla_\theta\log\pi_{\bar{\theta}_k}(a|s)\right| \\
&\leq \pi_{\bar{\theta}_k}(a|s)\|\theta_{k+1}-\theta_k\|\|\nabla_\theta\log\pi_{\bar{\theta}_k}(a|s)\| \overset{(a)}{\leq} G\|\theta_{k+1}-\theta_k\|
\end{aligned} \tag{80}$$

https://doi.org/10.52202/079017-3447

where $\bar{\theta}_k$ is some convex combination[1] of θ_k and θ_{k+1} and (a) results from Assumption 3. This concludes the first statement. Applying (80) and Lemma 12, we obtain the following for $g \in \{r, c\}$.

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbf{E}\big|J_g(\theta_{k+1}) - J_g(\theta_k)\big| = \sum_{k=1}^{K} \mathbf{E}\left|\sum_{s,a} d^{\pi_{\theta_{k+1}}}(s)(\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s))Q_g^{\pi_{\theta_k}}(s,a)\right| \\
&\leq \sum_{k=1}^{K} \mathbf{E}\left[\sum_{s,a} d^{\pi_{\theta_{k+1}}}(s)\left|\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s)\right|\left|Q_g^{\pi_{\theta_k}}(s,a)\right|\right] \\
&\leq G\sum_{k=1}^{K} \mathbf{E}\left[\sum_{s,a} d^{\pi_{\theta_{k+1}}}(s)\|\theta_{k+1} - \theta_k\||Q_g^{\pi_{\theta_k}}(s,a)|\right] \\
&\overset{(a)}{\leq} G\alpha\sum_{k=1}^{K} \mathbf{E}\left[\sum_{a}\underbrace{\sum_{s} d^{\pi_{\theta_{k+1}}}(s)}_{=1}\|\omega_k\| \cdot 6t_{\text{mix}}\right] = 6AG\alpha t_{\text{mix}}\sum_{k=1}^{K}\mathbf{E}\|\omega_k\| \\
&\overset{(b)}{\leq} 6AG\alpha t_{\text{mix}}\sqrt{K}\left(\sum_{k=1}^{K}\mathbf{E}\|\omega_k\|^2\right)^{\frac{1}{2}} \\
&\overset{(c)}{\leq} \tilde{\mathcal{O}}\left(\frac{\alpha AG}{\delta t_{\text{hit}}}\left[\left(\sqrt{A}Gt_{\text{mix}} + \delta\right)T^{\frac{2}{5}} + \sqrt{Lt_{\text{mix}}t_{\text{hit}}}T^{\frac{3}{10}}\right]\right)
\end{aligned} \tag{81}$$

Inequality (a) uses Lemma 9 and the update rule $\theta_{k+1} = \theta_k + \alpha\omega_k$. Step (b) holds by the Cauchy inequality and Jensen inequality whereas (c) can be derived using (22) and substituting $K = T/H$. This establishes the second statement. Next, recall from (5) that for any policy π_θ, $g^{\pi_\theta}(s) \triangleq \sum_a \pi_\theta(a|s)g(s,a)$. Note that, for any policy parameter θ, and any state $s \in \mathcal{S}$, the following holds.

$$V_g^{\pi_\theta}(s) = \sum_{t=0}^{\infty}\left\langle (P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}, g^{\pi_\theta}\right\rangle = \sum_{t=0}^{N-1}\left\langle (P^{\pi_\theta})^t(s,\cdot), g^{\pi_\theta}\right\rangle - NJ(\theta) + \sum_{t=N}^{\infty}\left\langle (P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}, g^{\pi_\theta}\right\rangle. \tag{82}$$

Define the following quantity.

$$\delta^{\pi_\theta}(s,T) \triangleq \sum_{t=N}^{\infty}\|(P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}\|_1 \text{ where } N = 4t_{\text{mix}}(\log_2 T) \tag{83}$$

Lemma 10 states that for sufficiently large T, we have $\delta^{\pi_\theta}(s,T) \leq \frac{1}{T^3}$ for any policy π_θ and state s. Combining this result with the fact that the g^{π_θ} function is absolutely bounded in $[0,1]$, we obtain,

$$\begin{aligned}
&\sum_{k=1}^{K}\mathbf{E}|V_g^{\pi_{\theta_{k+1}}}(s_k) - V_g^{\pi_{\theta_k}}(s_k)| \\
&\leq \sum_{k=1}^{K}\mathbf{E}\left|\sum_{t=0}^{N-1}\left\langle (P^{\pi_{\theta_{k+1}}})^t(s_k,\cdot) - (P^{\pi_{\theta_k}})^t(s_k,\cdot), g^{\pi_{\theta_{k+1}}}\right\rangle\right| + \sum_{k=1}^{K}\mathbf{E}\left|\sum_{t=0}^{N-1}\left\langle (P^{\pi_{\theta_k}})^t(s_k,\cdot), g^{\pi_{\theta_{k+1}}} - g^{\pi_{\theta_k}}\right\rangle\right| \\
&+ N\sum_{k=1}^{K}\mathbf{E}|J_g(\theta_{k+1}) - J_g(\theta_k)| + \frac{2K}{T^3} \\
&\overset{(a)}{\leq} \sum_{k=1}^{K}\sum_{t=0}^{N-1}\mathbf{E}\|(P^{\pi_{\theta_{k+1}}})^t - (P^{\pi_{\theta_k}})^t)g^{\pi_{\theta_{k+1}}}\|_\infty + \sum_{k=1}^{K}\sum_{t=0}^{N-1}\mathbf{E}\|g^{\pi_{\theta_{k+1}}} - g^{\pi_{\theta_k}}\|_\infty \\
&+ \tilde{\mathcal{O}}\left(\frac{\alpha AGt_{\text{mix}}}{\delta t_{\text{hit}}}\left[\left(\sqrt{A}Gt_{\text{mix}} + \delta\right)T^{\frac{2}{5}} + \sqrt{Lt_{\text{mix}}t_{\text{hit}}}T^{\frac{3}{10}}\right]\right)
\end{aligned} \tag{84}$$

[1]Note that, in general, $\bar{\theta}_k$ is dependent on (s,a).

where (a) follows from (81) and substituting $N = 4t_{\text{mix}}(\log_2 T)$. For the first term, note that,

$$\begin{aligned}
&\|((P^{\pi_{\theta_{k+1}}})^t - (P^{\pi_{\theta_k}})^t)g^{\pi_{\theta_{k+1}}}\|_\infty \\
&\leq \|P^{\pi_{\theta_{k+1}}}((P^{\pi_{\theta_{k+1}}})^{t-1} - (P^{\pi_{\theta_k}})^{t-1})g^{\pi_{\theta_{k+1}}}\|_\infty + \|(P^{\pi_{\theta_{k+1}}} - P^{\pi_{\theta_k}})(P^{\pi_{\theta_k}})^{t-1}g^{\pi_{\theta_{k+1}}}\|_\infty \\
&\overset{(a)}{\leq} \|((P^{\pi_{\theta_{k+1}}})^{t-1} - (P^{\pi_{\theta_k}})^{t-1})g^{\pi_{\theta_{k+1}}}\|_\infty + \max_s \|P^{\pi_{\theta_{k+1}}}(s,\cdot) - P^{\pi_{\theta_k}}(s,\cdot)\|_1
\end{aligned} \tag{85}$$

Inequality (a) holds since every row of $P^{\pi_{\theta_k}}$ sums to 1 and $\|(P^{\pi_{\theta_k}})^{t-1}g^{\pi_{\theta_{k+1}}}\|_\infty \leq 1$. Moreover, invoking (80), and the parameter update rule $\theta_{k+1} = \theta_k + \alpha\omega_k$, we get,

$$\begin{aligned}
\max_s \|P^{\pi_{\theta_{k+1}}}(s,\cdot) - P^{\pi_{\theta_k}}(s,\cdot)\|_1 &= \max_s \left|\sum_{s'}\sum_a (\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s))P(s'|s,a)\right| \\
&\leq G\|\theta_{k+1} - \theta_k\| \max_s \left|\sum_{s'}\sum_a P(s'|s,a)\right| \\
&\leq \alpha AG\|\omega_k\|
\end{aligned}$$

Plugging the above result into (85) and using a recursive argument, we get,

$$\begin{aligned}
\|((P^{\pi_{\theta_{k+1}}})^t - (P^{\pi_{\theta_k}})^t)g^{\pi_{\theta_{k+1}}}\|_\infty &\leq \sum_{t'=1}^{t} \max_s \|P^{\pi_{\theta_{k+1}}}(s,\cdot) - P^{\pi_{\theta_k}}(s,\cdot)\|_1 \\
&\leq \sum_{t'=1}^{t} \alpha AG\|\omega_k\| \leq \alpha tAG\|\omega_k\|
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{t=0}^{N-1} \mathbf{E}\|((P^{\pi_{\theta_{k+1}}})^t - (P^{\pi_{\theta_k}})^t)g^{\pi_{\theta_{k+1}}}\|_\infty \\
&\qquad \leq \sum_{k=1}^{K}\sum_{t=0}^{N-1} \alpha tAG\|\omega_k\| \\
&\qquad \leq \mathcal{O}(\alpha AGN^2)\sum_{k=1}^{K}\mathbf{E}\|\omega_k\| \\
&\qquad \leq \mathcal{O}(\alpha AGN^2\sqrt{K})\left(\sum_{k=1}^{K}\mathbf{E}\|\omega_k\|^2\right)^{\frac{1}{2}} \\
&\qquad \overset{(a)}{=} \tilde{\mathcal{O}}\left(\frac{\alpha AGt_{\text{mix}}}{\delta t_{\text{hit}}}\left[\left(\sqrt{A}Gt_{\text{mix}} + \delta\right)T^{\frac{2}{5}} + \sqrt{Lt_{\text{mix}}t_{\text{hit}}}T^{\frac{3}{10}}\right]\right)
\end{aligned} \tag{86}$$

where (a) follows from (22). Moreover, notice that,

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{t=0}^{N-1} \mathbf{E}\|g^{\pi_{\theta_{k+1}}} - g^{\pi_{\theta_k}}\|_\infty &\leq \sum_{k=1}^{K}\sum_{t=0}^{N-1} \mathbf{E}\left[\max_s \left|\sum_a (\pi_{\theta_{k+1}}(a|s) - \pi_{\theta_k}(a|s))g(s,a)\right|\right] \\
&\overset{(a)}{\leq} \alpha AGN\sum_{k=1}^{K}\mathbf{E}\|\omega_k\| \\
&\leq \alpha AGN\sqrt{K}\left(\sum_{k=1}^{K}\mathbf{E}\|\omega_k\|^2\right)^{\frac{1}{2}} \\
&\overset{(b)}{\leq} \tilde{\mathcal{O}}\left(\frac{\alpha AG}{\delta t_{\text{hit}}}\left[\left(\sqrt{A}Gt_{\text{mix}} + \delta\right)T^{\frac{2}{5}} + \sqrt{Lt_{\text{mix}}t_{\text{hit}}}T^{\frac{3}{10}}\right]\right)
\end{aligned} \tag{87}$$

where (a) follows from (80) and the update rule $\theta_{k+1} = \theta_k + \alpha\omega_k$ whereas (b) is a consequence of (22). Combining (84), (86), and (87), we establish the third statement. □

https://doi.org/10.52202/079017-3447

C.2 Proof of Theorem 2

Proof. Recall the decomposition of the regret in section 5 and take the expectation.

$$\begin{aligned}\mathbf{E}[\mathrm{Reg}_T] &= \sum_{t=0}^{T-1}\left(J_r^{\pi^*} - r(s_t,a_t)\right) = H\sum_{k=1}^{K}\left(J_r^{\pi^*} - J_r(\theta_k)\right) + \sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}\left(J_r(\theta_k) - r(s_t,a_t)\right)\\ &= H\sum_{k=1}^{K}\left(J_r^{\pi^*} - J_r(\theta_k)\right) + \mathbf{E}\left[\sum_{k=1}^{K-1} V_r^{\pi_{\theta_{k+1}}}(s_{kH}) - V_r^{\pi_{\theta_k}}(s_{kH})\right] + \mathbf{E}\left[V_r^{\pi_{\theta_K}}(s_T) - V_r^{\pi_{\theta_0}}(s_0)\right]\end{aligned} \tag{88}$$

Using the result in (78), Lemma 8 and Lemma 9, we get,

$$\begin{aligned}\mathbf{E}[\mathrm{Reg}_T] &\le T\sqrt{\epsilon_{\mathrm{bias}}} + G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(T^{\frac{4}{5}} + \frac{\sqrt{A}Gt_{\mathrm{mix}}}{\delta}T^{\frac{4}{5}} + \frac{\sqrt{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}}{\delta}T^{\frac{7}{10}}\right) + \mathcal{O}\left(\frac{1}{T}+T^{\frac{3}{5}}\right)\\ &+ \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{\delta^2 + AG^2t_{\mathrm{mix}}^2}{\delta^2}T^{\frac{3}{5}} + \frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}{\delta^2}T^{\frac{2}{5}}\right) + \tilde{\mathcal{O}}\left(\frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}\mathbf{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))]}{\delta}T^{\frac{2}{5}}\right)\\ &+ \tilde{\mathcal{O}}\left(\frac{\alpha AGt_{\mathrm{mix}}}{\delta t_{\mathrm{hit}}}\left[\left(\sqrt{A}Gt_{\mathrm{mix}}+\delta\right)T^{\frac{2}{5}} + \sqrt{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}T^{\frac{3}{10}}\right]\right) + \mathcal{O}(t_{\mathrm{mix}})\end{aligned} \tag{89}$$

Similarly, for the constraint violation, we have

$$\begin{aligned}\mathbf{E}[\mathrm{Vio}_T] &= \sum_{t=0}^{T-1}\left(-c(s_t,a_t)\right) = H\sum_{k=1}^{K} -J_c(\theta_k) + \sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}\left(J_c(\theta_k) - c(s_t,a_t)\right)\\ &= -H\sum_{k=1}^{K} J_c(\theta_k) + \mathbf{E}\left[\sum_{k=1}^{K-1} V_c^{\pi_{\theta_{k+1}}}(s_{kH}) - V_c^{\pi_{\theta_k}}(s_{kH})\right] + \mathbf{E}\left[V_c^{\pi_{\theta_K}}(s_T) - V_c^{\pi_{\theta_0}}(s_0)\right]\end{aligned} \tag{90}$$

Using the result in (79), Lemma 8 and Lemma 9, we get,

$$\begin{aligned}\mathbf{E}[\mathrm{Vio}_T] &\le T\delta\sqrt{\epsilon_{\mathrm{bias}}} + G\left(1+\frac{1}{\mu_F}\right)\tilde{\mathcal{O}}\left(\left[\delta + \sqrt{A}Gt_{\mathrm{mix}}\right]T^{\frac{4}{5}} + \sqrt{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}T^{\frac{7}{10}}\right)\\ &+ \mathcal{O}\left(\frac{t_{\mathrm{mix}}t_{\mathrm{hit}}}{\delta}T^{\frac{4}{5}} + \frac{1}{\delta T} + \delta T^{\frac{3}{5}}\right) + \frac{B}{L}\tilde{\mathcal{O}}\left(\frac{\delta^2 + AG^2t_{\mathrm{mix}}^2}{\delta}T^{\frac{3}{5}} + \frac{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}{\delta}T^{\frac{2}{5}}\right)\\ &+ \tilde{\mathcal{O}}\left(Lt_{\mathrm{mix}}t_{\mathrm{hit}}\mathbf{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_1}(\cdot|s))]T^{\frac{2}{5}}\right)\\ &+ \tilde{\mathcal{O}}\left(\frac{\alpha AGt_{\mathrm{mix}}}{\delta t_{\mathrm{hit}}}\left[\left(\sqrt{A}Gt_{\mathrm{mix}}+\delta\right)T^{\frac{2}{5}} + \sqrt{Lt_{\mathrm{mix}}t_{\mathrm{hit}}}T^{\frac{3}{10}}\right]\right) + \mathcal{O}(t_{\mathrm{mix}})\end{aligned} \tag{91}$$

This concludes the theorem. □

D Some Auxiliary Lemmas for the Proofs

Lemma 9. *[17, Lemma 14] For any ergodic MDP with mixing time t_{mix}, the following holds $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$, any policy π and $\forall g\in\{r,c\}$.*

$$(a)|V_g^\pi(s)| \le 5t_{\mathrm{mix}},\quad (b)|Q_g^\pi(s,a)| \le 6t_{\mathrm{mix}}$$

Lemma 10. *[17, Corollary 13.2] Let $\delta^\pi(\cdot,T)$ be defined as written below for an arbitrary policy π.*

$$\delta^\pi(s,T) \triangleq \sum_{t=N}^{\infty}\left\|(P^\pi)^t(s,\cdot) - d^\pi\right\|_1,\ \forall s\in\mathcal{S}\ \text{ where } N = 4t_{\mathrm{mix}}(\log_2 T) \tag{92}$$

If $t_{\mathrm{mix}} < T/4$, we have the following inequality $\forall s\in\mathcal{S}$: $\delta^\pi(s,T) \le \frac{1}{T^3}$.

Lemma 11. *[17, Lemma 16] Let $\mathcal{I} = \{t_1 + 1, t_1 + 2, \cdots, t_2\}$ be a certain period of an epoch k of Algorithm 2 with length N. Then for any s, the probability that the algorithm never visits s in $\mathcal{I}$ is upper bounded by*

$$\left(1 - \frac{3d^{\pi_{\theta_k}}(s)}{4}\right)^{\left\lfloor \frac{\lfloor \mathcal{I} \rfloor}{N} \right\rfloor} \tag{93}$$

Lemma 12. *[17, Lemma 15] The difference of the values of the function J_g, $g \in \{r, c\}$ at policies π and π', is*

$$J_g^{\pi} - J_g^{\pi'} = \sum_s \sum_a d^{\pi}(s)(\pi(a|s) - \pi'(a|s))Q_g^{\pi'}(s, a) \tag{94}$$

Lemma 13. *[6, Lemma 7] The term $\hat{J}_c(\theta)$ for any $\theta \in \Theta$ is a good estimator of $J_c(\theta)$, which means*

$$\left|\mathbf{E}[\hat{J}_c(\theta)] - J_c(\theta)\right| \leq \frac{1}{T^2} \tag{95}$$

Lemma 14. *[36, Lemma A.6] Let $\theta \in \Theta$ be a policy parameter. Fix a trajectory $z = \{(s_t, a_t, r_t, s_{t+1})\}_{t \in \mathbb{N}}$ generated by following the policy π_θ starting from some initial state $s_0 \sim \rho$. Let, $\nabla L(\theta)$ be the gradient that we wish to estimate over z, and $l(\theta, \cdot)$ is a function such that $\mathbf{E}_{z \sim d^{\pi_\theta}, \pi_\theta} l(\theta, z) = \nabla L(\theta)$. Assume that $\|l(\theta, z)\|, \|\nabla L(\theta)\| \leq G_L$, $\forall \theta \in \Theta$, $\forall z \in \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$. Define $l^Q = \frac{1}{Q} \sum_{i=1}^{Q} l(\theta, z_i)$. If $P = 2t_{\text{mix}} \log T$, then the following holds as long as $Q \leq T$,*

$$\mathbf{E}\left[\left\|l^Q - \nabla L(\theta)\right\|^2\right] \leq \mathcal{O}\left(G_L^2 \log(PQ) \frac{P}{Q}\right) \tag{96}$$

Lemma 15 (Strong duality)**.** *[37, Lemma 3] For convenience, we rewrite the unparameterized problem* (2)*.*

$$\begin{aligned} &\max_{\pi \in \Pi} \; J_r^{\pi} \\ &\text{s.t. } J_c^{\pi} \geq 0 \end{aligned} \tag{97}$$

Define π^* as the optimal solution to the above problem. Define the associated dual function as

$$J_D^{\lambda} \triangleq \max_{\pi \in \Pi} J_r^{\pi} + \lambda J_c^{\pi} \tag{98}$$

and denote $\lambda^* = \arg\min_{\lambda \geq 0} J_D^{\lambda}$. We have the following strong duality property for the unparameterized problem whenever Assumption 2 holds.

$$J_r^{\pi^*} = J_D^{\lambda^*} \tag{99}$$

Although the strong duality holds for the unparameterized problem, the same is not true for parameterized class $\{\pi_\theta | \theta \in \Theta\}$. To formalize this statement, define the dual function associated with the parameterized problem as follows.

$$J_{D,\Theta}^{\lambda} \triangleq \max_{\theta \in \Theta} J_r(\theta) + \lambda J_c(\theta) \tag{100}$$

and denote $\lambda_\Theta^* = \arg\min_{\lambda \geq 0} J_{D,\Theta}^{\lambda}$. The lack of strong duality states that, in general, $J_{D,\Theta}^{\lambda_\Theta^*} \neq J_r(\theta^*)$ where θ^* is a solution of the parameterized constrained optimization (3). However, the parameter λ_Θ^*, as we demonstrate below, must obey some restrictions.

Lemma 16. *Under Assumption 2, the optimal dual variable for the parameterized problem is bounded as*

$$0 \leq \lambda_\Theta^* \leq \frac{J_r^{\pi^*} - J_r(\bar{\theta})}{\delta} \leq \frac{1}{\delta} \tag{101}$$

Proof. The proof follows the approach in [37, Lemma 3], but is revised to the general parameterization setup. Let $\Lambda_a \triangleq \{\lambda \geq 0 \,|\, J_{D,\Theta}^{\lambda} \leq a\}$ be a sublevel set of the dual function for $a \in \mathbb{R}$. If Λ_a is non-empty, then for any $\lambda \in \Lambda_a$,

$$a \geq J_{D,\Theta}^{\lambda} \geq J_r(\bar{\theta}) + \lambda J_c(\bar{\theta}) \geq J_r(\bar{\theta}) + \lambda\delta \tag{102}$$

where $\bar{\theta}$ is a Slater point in Assumption 2. Thus, $\lambda \leq (a - J_r(\bar{\theta}))/\delta$. If we take $a = J_{D,\Theta}^{\lambda_\Theta^*} \leq J_{D,\Theta}^{\lambda^*} \leq J_D^{\lambda^*} = J_r^{\pi^*}$, then we have $\lambda_\Theta^* \in \Lambda_a$, which proves the Lemma. The last inequality holds since $J_r^{\pi} \in [0, 1]$ for any policy, π. □

 https://doi.org/10.52202/079017-3447

Since the above inequality holds for arbitrary Θ, we also have, $0 \leq \lambda^* \leq \frac{1}{\delta}$. Define $v(\tau) \triangleq \max_{\pi \in \Pi}\{J_r^\pi | J_c^\pi \geq \tau\}$. Using the strong duality property of the unparameterized problem (97), we establish the following property of the function, $v(\cdot)$.

Lemma 17. *Assume that the Assumption 2 holds, we have for any $\tau \in \mathbb{R}$,*

$$v(0) - \tau\lambda^* \geq v(\tau) \tag{103}$$

Proof. By the definition of $v(\tau)$, we have $v(0) = J_r^{\pi^*}$. With a slight abuse of notation, denote $J_{\mathrm{L}}(\pi, \lambda) = J_r^\pi + \lambda J_c^\pi$. By the strong duality stated in Lemma 15, we have the following for any $\pi \in \Pi$.

$$J_{\mathrm{L}}(\pi, \lambda^*) \leq \max_{\pi \in \Pi} J_{\mathrm{L}}(\pi, \lambda^*) \overset{Def}{=} J_D^{\lambda^*} \overset{(99)}{=} J_r^{\pi^*} = v(0) \tag{104}$$

Thus, for any $\pi \in \{\pi \in \Pi \,|\, J_c^\pi \geq \tau\}$,

$$\begin{aligned} v(0) - \tau\lambda^* &\geq J_{\mathrm{L}}(\pi, \lambda^*) - \tau\lambda^* \\ &= J_r^\pi + \lambda^*(J_c^\pi - \tau) \geq J_r^\pi \end{aligned} \tag{105}$$

Maximizing the right-hand side of this inequality over $\{\pi \in \Pi | J_c^\pi \geq \tau\}$ yields

$$v(0) - \tau\lambda^* \geq v(\tau) \tag{106}$$

This completes the proof of the lemma. □

We note that a similar result was shown in [38, Lemma 15]. However, the setup of the stated paper is different from that of ours. Specifically, [38] considers a tabular setup with peak constraints. Note that Lemma 17 has no direct connection with the parameterized setup since its proof uses strong duality and the function, $v(\cdot)$, is defined via a constrained optimization over the entire policy set, Π, rather than the parameterized policy set. Interestingly, however, the relationship between $v(\tau)$ and $v(0)$ leads to the lemma stated below which turns out to be pivotal in establishing regret and constraint violation bounds in the parameterized setup.

Lemma 18. *Let Assumption 2 hold. For any constant $C \geq 2\lambda^*$, if there exists a $\pi \in \Pi$ and $\zeta > 0$ such that $J_r^{\pi^*} - J_r^\pi + C[-J_c^\pi] \leq \zeta$, then*

$$-J_c^\pi \leq 2\zeta/C \tag{107}$$

Proof. Let $\tau = J_c^\pi$. Using the definition of $v(\tau)$, one can write,

$$J_r^\pi \leq v(\tau) \tag{108}$$

Combining Eq. (106) and (108), we obtain the following.

$$J_r^\pi - J_r^{\pi^*} \leq v(\tau) - v(0) \leq -\tau\lambda^* \tag{109}$$

The condition in the Lemma leads to,

$$(C - \lambda^*)(-\tau) = \tau\lambda^* + C(-\tau) \leq J_r^{\pi^*} - J_r^\pi + C[-J_c^\pi] \leq \zeta \tag{110}$$

Finally, we have,

$$-\tau \leq \frac{\zeta}{C - \lambda^*} \leq \frac{2\zeta}{C} \tag{111}$$

which completes the proof. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contribution and challenges are clearly described at the end of the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We add all the assumptions in the work, list the gap with lower bound in Table 1, and give future work direction.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-3447

Answer: [Yes]

Justification: All the assumptions are clearly stated with a remark to discuss. All the proof are given in the appendix in details.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: the paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: the paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: the paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: the paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

 https://doi.org/10.52202/079017-3447

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: the paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the research conducted in the paper satisfies the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: the paper poses no such risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: the paper does not use existing assets

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

 https://doi.org/10.52202/079017-3447

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not use release new assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.