Global Convergence for Average Reward Constrained MDPs with Primal-Dual Actor Critic Algorithm

Yang Xu[*]
Purdue University, USA
xu1720@purdue.edu

Swetha Ganesh[*]
Purdue University, USA
Indian Institute of Science, Bengaluru, India
swethaganesh@iisc.ac.in

Washim Uddin Mondal
Indian Institute of Technology Kanpur
wmondal@iitk.ac.in

Qinbo Bai
Purdue University, USA
bai113@purdue.edu

Vaneet Aggarwal
Purdue University, USA
vaneet@purdue.edu

Abstract

This paper investigates infinite-horizon average reward Constrained Markov Decision Processes (CMDPs) under general parametrized policies with smooth and bounded policy gradients. We propose a Primal-Dual Natural Actor-Critic algorithm that adeptly manages constraints while ensuring a high convergence rate. In particular, our algorithm achieves global convergence and constraint violation rates of $\tilde{\mathcal{O}}(1/\sqrt{T})$ over a horizon of length T when the mixing time, τ_{mix}, is known to the learner. In absence of knowledge of τ_{mix}, the achievable rates change to $\tilde{\mathcal{O}}(1/T^{0.5-\epsilon})$ provided that $T \geq \tilde{\mathcal{O}}\left(\tau_{\mathrm{mix}}^{2/\epsilon}\right)$. Our results match the theoretical lower bound for Markov Decision Processes and establish a new benchmark in the theoretical exploration of average reward CMDPs.

1 Introduction

Reinforcement Learning (RL) is a paradigm where an agent learns to maximize its average reward in an unknown environment through repeated interactions. RL finds applications in diverse domains such as transportation, communication networks, robotics, and epidemic control [5, 28, 32, 14, 29]. Among various RL settings, the infinite horizon average reward setup is of particular significance for modeling long-term objectives in practical scenarios. This setting is critical as it aligns with real-world applications requiring persistent and consistent performance over time.

In many applications, agents must operate under certain constraints. For instance, in transportation networks, delivery times must adhere to specified windows; in communication networks, resource allocations must stay within budget limits. Constrained Markov Decision Processes (CMDPs) effectively incorporate these constraints by introducing a cost function alongside the reward function. In average reward CMDPs, agents aim to maximize the average reward while ensuring that the average cost does not exceed a predefined threshold, making them crucial for scenarios where safety, budget, or resource constraints are paramount.

Finding optimal policies for average reward CMDPs is challenging, especially if the environment is unknown. The efficiency of a solution to the CMDP is measured by the rate at which the average global optimality error and the constraint violation diminish as a function of the length of the horizon T. Although many existing works focus on the tabular policies [24, 3, 2, 15], these solutions do not apply to real-life scenarios where the state space is large or infinite.

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-3099

Algorithm	Global Convergence	Violation	Mixing Time Unknown	Model-free	Setting
Algorithm 1 in [15]	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	No	No	Tabular
Algorithm 3 in [15]	$\tilde{\mathcal{O}}\left(1/T^{1/3}\right)$	$\tilde{\mathcal{O}}\left(1/T^{1/3}\right)$	Yes	No	Tabular
UC-CURL and PS-CURL [2]	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	0	Yes	No	Tabular
Algorithm 2 in [27]	$\tilde{\mathcal{O}}\left(1/T^{1/4}\right)$	$\tilde{\mathcal{O}}\left(1/T^{1/4}\right)$	Yes	-	Linear MDP
Algorithm 3 in [27]	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	No	-	Linear MDP
Triple-QA [42]	$\tilde{\mathcal{O}}\left(1/T^{1/6}\right)$	0	Yes	Yes	Tabular
Algorithm 1 in [10]	$\tilde{\mathcal{O}}\left(1/T^{1/5}\right)$	$\tilde{\mathcal{O}}\left(1/T^{1/5}\right)$	No	Yes	General Parameterization
This paper (Theorem 4.9)	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	$\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$	No	Yes	General Parameterization
This paper (Theorem 4.10)	$\tilde{\mathcal{O}}\left(1/T^{0.5-\epsilon}\right)$	$\tilde{\mathcal{O}}\left(1/T^{0.5-\epsilon}\right)$	Yes [2]	Yes	General Parameterization
Lower bound [6]	$\Omega\left(1/\sqrt{T}\right)$	N/A	N/A	N/A	N/A

Table 1: This table summarizes the different model-based and model-free state-of-the-art algorithms available in the literature for average reward CMDPs and their results for global convergence rate and average constraint violation. The bounds of global convergence describe convergence to the best performance permitted by the chosen features and policy class; the residual approximation floor is fixed (independent of T), and the listed rates govern how fast the statistical gap decays toward that floor. General parameterization refers to parameterizations whose policy score $\nabla_\theta \log \pi_\theta(a|s)$ is uniformly bounded and Lipschitz in θ, together with a Fisher non-degeneracy condition.

General parametrization offers a useful approach for dealing with such scenarios. It indexes policies via finite-dimensional parameters, which makes them suitable for large state space CMDPs. However, as exhibited in Table 1, the current state-of-the-art algorithms for average reward CMDPs with general parametrization achieve a convergence rate of $\tilde{\mathcal{O}}(1/T^{1/5})$ [10], which is far from the theoretical lower bound of $\Omega(1/\sqrt{T})$. This significant gap highlights the need for improved algorithms that can achieve theoretical optimality.

Challenges and contributions: In this paper, we propose a primal-dual-based natural actor-critic algorithm that achieves global convergence and constraint violation rates of $\tilde{\mathcal{O}}(1/\sqrt{T})$ over a horizon of length T if the mixing time, τ_{mix}, is known. On the other hand, in the absence of knowledge of τ_{mix}, the achievable rates change to $\tilde{\mathcal{O}}(1/T^{0.5-\epsilon})$, provided that the horizon length $T \geq \tilde{\mathcal{O}}(\tau_{\mathrm{mix}}^{2/\epsilon})$. Therefore, even with unknown τ_{mix}, the achieved rates can be driven arbitrarily close to the optimal one by choosing small enough ϵ, albeit at the cost of a large T. Both results significantly improve upon the state-of-the-art convergence rate of $\tilde{\mathcal{O}}(1/T^{1/5})$ in [10] for general policy parametrization and match the theoretical lower bound.

Since CMDP does not have strong convexity in the policy parameters with general parametrization, directly applying the primal-dual approach does not lead to optimal guarantees. The lack of strong convexity prevents the convergence result for the dual problem from automatically translating to the primal problem. This issue is reflected in the current state of art convergence rate in average reward CMDPs in [10], which is $\tilde{\mathcal{O}}(1/T^{1/5})$ using a primal-dual approach, whereas the unconstrained counterpart in [11] on which the above is built upon achieves a convergence rate of $\tilde{\mathcal{O}}(1/T^{1/4})$. Thus, the previous literature for CMDP in general parametrized setups shows a gap in the results of the unconstrained and constrained setups.

It is evident that directly adding a primal-dual structure to an existing actor-critic algorithm, such as in [23], does not result in $\tilde{\mathcal{O}}(1/\sqrt{T})$ convergence rate for CMDP with general parametrization. This is specifically due to the existence of dual learning rate β, which eventually becomes one of the dominant terms in the convergence rate of the Lagrange function. If the dual learning rate is too low, the constraint violation converges very slowly. However, if β is too high, the primal updates may

[2]The convergence result holds under the condition that $T \geq \tilde{\mathcal{O}}\left(\tau_{\mathrm{mix}}^{2/\epsilon}\right)$, where τ_{mix} is the mixing time.

exhibit high variance, slowing down convergence to the optimal policy. To navigate this tradeoff, it is crucial to carefully tune the problem parameters.

Our algorithm has a nested loop structure where the outer loop of length K updates the primal-dual parameters, and the inner loops of length H run the natural policy gradient (NPG) and optimal critic parameter finding subroutines. We show that, to achieve the optimal rates, the parameter H should be nearly constant while K should be $\tilde{\Theta}(T)$. This adjustment makes our algorithm resemble a single-timescale algorithm. In contrast, for unconstrained MDPs, K and H are typically set to $\Theta(\sqrt{T})$ [23]. Achieving $\mathcal{O}(1/\sqrt{T})$ convergence in our setting requires ensuring that the bias remains sufficiently small despite a significantly low value of H. Achieving this challenging goal is made possible due to the use of Multi-level Monte Carlo (MLMC)-based estimates in the inner loop subroutines and a sharper analysis of these updates. One benefit of using MLMC-based estimates is that it allows us to design an algorithm that achieves near-optimal global convergence and constraint violation rates without knowledge of the mixing time. Related unconstrained average-reward results have also used MLMC to avoid mixing-time oracles [38, 34, 23]; here we apply this idea to CMDPs via a primal–dual natural actor–critic approach and provide global-rate and average-violation guarantees. We would like to emphasize that our work is the first in the general parameterized CMDP literature to achieve this feat.

2 Formulation

Consider an infinite-horizon average reward constrained Markov Decision Process (CMDP) denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, c, P, \rho)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space of size A, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ represents the reward function, $c : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ denotes the constraint cost function, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition function where $\Delta(\mathcal{S})$ denotes the probability simplex over $\mathcal{S}$, and $\rho \in \Delta(\mathcal{S})$ indicates the initial distribution of states. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps a state to an action distribution. The average reward and average constraint cost of a policy, π, denoted as J_r^π and J_c^π respectively, are defined as

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

The expectations computed are over the distribution of all π-induced trajectories $\{(s_t, a_t)\}_{t=0}^\infty$ where $a_t \sim \pi(s_t)$, and $s_{t+1} \sim P(s_t, a_t)$, $\forall t \in \{0, 1, \cdots\}$. For simplifying the notation, we drop the dependence on ρ when there is no confusion. We aim to maximize the average reward while ensuring that the average cost exceeds a given threshold. Without loss of generality, we can formulate this as:

$$\max_\pi J_r^\pi \text{ s.t. } J_c^\pi \geq 0 \tag{2}$$

When the underlying state space, $\mathcal{S}$ is large, the above problem becomes difficult to solve because the search space of the policy, π, increases exponentially with $|\mathcal{S} \times \mathcal{A}|$. We, therefore, consider a class of parametrized policies, $\{\pi_\theta | \theta \in \Theta\}$ that indexes the policies by a d-dimensional parameter, $\theta \in \mathbb{R}^d$ where $d \ll |\mathcal{S}||\mathcal{A}|$. The original problem in (2) can then be reformulated as follows.

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

In the remaining article, we use $J_g^{\pi_\theta} = J_g(\theta)$ for $g \in \{r, c\}$ for simplifying the notation. We assume the optimization problem in (3) obeys the Slater condition, which ensures the existence of an interior point solution. This assumption is commonly used in model-free average-reward CMDPs [42, 10].

Assumption 2.1 (Slater condition)**.** There exists a $\delta \in (0, 1)$ and $\bar{\theta} \in \Theta$ such that $J_c(\bar{\theta}) \geq \delta$.

We now make the following assumption on the CMDP.

Assumption 2.2. The CMDP $\mathcal{M}$ is ergodic, i.e., the Markov chain, $\{s_t\}_{t\geq 0}$, induced under every policy π, is irreducible and aperiodic.

Note that most works on average reward MDPs and CMDPs with general parameterizations, to the best of our knowledge, assume ergodicity [10, 22, 11]. If CMDP $\mathcal{M}$ is ergodic, then $\forall \theta \in \Theta$, there exists a ρ-independent unique stationary distribution $d^{\pi_\theta} \in \Delta(\mathcal{S})$, that obeys $P^{\pi_\theta} d^{\pi_\theta} = d^{\pi_\theta}$ where $P^{\pi_\theta}(s, s') = \sum_{a\in\mathcal{A}} \pi_\theta(a|s) P(s'|s, a)$, $\forall s, s' \in \mathcal{S}$. Since $\forall \pi$, $J_g^\pi = \sum_{s,a} g(s, a)\pi(a|s)d^\pi(s)$, the assumption of ergodicity also ensures that J_g^π is independent of ρ, $\forall g \in \{r, c\}$. Next, we define the mixing time of an ergodic CMDP.

 https://doi.org/10.52202/085713-3099

Definition 2.3. The mixing time of a CMDP $\mathcal{M}$ with respect to a policy parameter θ is defined as, $\tau^{\theta}_{\text{mix}} := \min\left\{t \geq 1 \middle| \|(P^{\pi_\theta})^t(s,\cdot) - d^{\pi_\theta}\| \leq \frac{1}{4}, \forall s \in \mathcal{S}\right\}$. We also define $\tau_{\text{mix}} := \sup_{\theta\in\Theta} \tau^{\theta}_{\text{mix}}$ as the overall mixing time. In this paper, τ_{mix} is finite due to ergodicity.

The mixing time of an MDP measures how fast the MDP reaches its stationary distribution when executing a fixed policy. We use a primal-dual actor-critic approach to solve (3). Before proceeding further, we define a few terms. The action-value function corresponding to a policy π_θ is given as

$$Q^{\pi_\theta}_g(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\left\{g(s_t,a_t) - J_g(\theta)\right\}\middle| s_0 = s, a_0 = a\right] \tag{4}$$

where $g \in \{r, c\}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$. We further write their corresponding state value function as

$$V^{\pi_\theta}_g(s) = \mathbb{E}_{a\sim\pi_\theta(s)}[Q^{\pi_\theta}_g(s,a)] \tag{5}$$

The Bellman's equation can be expressed as follows [35] $\forall(s,a) \in \mathcal{S}\times\mathcal{A}$, and $g \in \{r,c\}$.

$$Q^{\pi_\theta}_g(s,a) = g(s,a) - J_g(\theta) + \mathbb{E}_{s'\sim P(s,a)}\left[V^{\pi_\theta}_g(s')\right], \tag{6}$$

We define the advantage term for the reward and cost functions as follows $A^{\pi_\theta}_g(s,a) \triangleq Q^{\pi_\theta}_g(s,a) - V^{\pi_\theta}_g(s)$ where $g \in \{r,c\}$, $(s,a) \in \mathcal{S}\times\mathcal{A}$. With the above notations, we now present the commonly-used policy gradient theorem established by [39] for $g \in \{r,c\}$.

$$\nabla_\theta J_g(\theta) = \mathbb{E}_{\substack{s\sim d^{\pi_\theta}\\ a\sim\pi_\theta(\cdot|s)}}\left[A^{\pi_\theta}_g(s,a)\nabla_\theta \log \pi_\theta(a|s)\right] \tag{7}$$

The above term is useful in policy gradient-type algorithms where the learning direction of the parameter θ is given by $\{\nabla_\theta J_g(\theta)\}_{g\in\{r,c\}}$. In this paper, however, we will be interested in the Natural Policy Gradients $\{\omega^*_{g,\theta}\}_{g\in\{r,c\}}$ defined as follows.

$$\omega^*_{g,\theta} \triangleq F(\theta)^{-1}\nabla_\theta J_g(\theta), \tag{8}$$

where $F(\theta) \in \mathbb{R}^{d\times d}$ is the Fisher information matrix, which is formally defined as: $F(\theta) = \mathbb{E}_{(s,a)\sim\nu^{\pi_\theta}}\left[\nabla_\theta \log \pi_\theta(a|s) \otimes \nabla_\theta \log \pi_\theta(a|s)\right]$ where $\nu^{\pi_\theta}(s,a) \triangleq d^{\pi_\theta}(s)\pi_\theta(a|s)$, $\forall(s,a)$, and $\otimes$ defines the outer product. Note that NPG is similar to PG except modulated by the Fisher matrix, which accounts for the rate of change of policies with θ. The NPG direction $\omega^*_{g,\theta}$ can also be expressed as a solution to the following strongly convex optimization problem.

$$\min_{\omega\in\mathbb{R}^d} f_g(\theta,\omega) := \frac{1}{2}\omega^\top F(\theta)\omega - \omega^\top\nabla_\theta J_g(\theta) \tag{9}$$

The above formulation allows one to calculate the NPG in a gradient-based iterative procedure. In particular, the gradient of $f_g(\theta,\cdot)$ is obtained as $\nabla_\omega f_g(\theta,\omega) = F(\theta)\omega - \nabla_\theta J_g(\theta)$.

Algorithm 1 Primal-Dual Natural Actor-Critic (PDNAC)

Input: Initial parameters θ_0, $\omega_0 = 0$, $\xi_0 = 0$, $\lambda_0 = 0$, policy update stepsize α, dual update stepsize β, NPG update stepsize γ_ω, critic update stepsize γ_ξ, initial state $s_0 \sim \rho(\cdot)$, outer loop size K, inner loop size H, $T_{\max}$
for $k = 0, \cdots, K-1$ **do**
  $\omega^k_{g,0} \leftarrow \omega_0$, $\xi^k_{g,0} \leftarrow \xi_0$ $\forall g \in \{r,c\}$;
  /* Critic Subroutine */
  for $h = 0, \cdots, H-1$ **do**
    $s^0_{kh} \leftarrow s_0$, $P^k_h \sim$ Geom$(1/2)$
    $l_{kh} \leftarrow (2^{P^k_h} - 1)\mathbf{1}(2^{P^k_h} \leq T_{\max}) + 1$
    for $t = 0, \ldots, l_{kh} - 1$ **do**
      Take action $a^t_{kh} \sim \pi_{\theta_k}(\cdot|s^t_{kh})$
      Observe $s^{t+1}_{kh} \sim P(\cdot|s^t_{kh}, a^t_{kh})$
      Observe $g(s^t_{kh}, a^t_{kh})$, $g \in \{r,c\}$
    end for
    $s_0 \leftarrow s^{l_{kh}}_{kh}$
    Update $\xi^k_{g,h}$ using (18) and (20).
  end for
  /* Actor Subroutine */
  for $h = 0, \cdots, H-1$ **do**
    $s^0_{kh} \leftarrow s_0$, $Q^k_h \sim$ Geom$(1/2)$
    $l_{kh} \leftarrow (2^{Q^k_h} - 1)\mathbf{1}(2^{Q^k_h} \leq T_{\max}) + 1$
    for $t = 0, \ldots, l_{kh} - 1$ **do**
      Take action $a^t_{kh} \sim \pi_{\theta_k}(\cdot|s^t_{kh})$
      Observe $s^{t+1}_{kh} \sim P(\cdot|s^t_{kh}, a^t_{kh})$
      Observe $g(s^t_{kh}, a^t_{kh})$, $g \in \{r,c\}$
    end for
    $s_0 \leftarrow s^{l_{kh}}_{kh}$
    Update $\omega^k_{g,h}$ using (21) and (23)
  end for
  $\xi^k_g \leftarrow \xi^k_{g,H}$, $\omega^k_g \leftarrow \omega^k_{g,H}$, $g \in \{r,c\}$
  $\omega_k \leftarrow \omega^k_r + \lambda_k\omega^k_c$
  Update (θ_k, λ_k) using (12)
end for

3 Algorithm

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

$$\max_{\theta\in\Theta}\min_{\lambda\geq 0}\mathcal{L}(\theta,\lambda)\triangleq J_r(\theta)+\lambda J_c(\theta) \tag{10}$$

The term $\mathcal{L}(\cdot,\cdot)$ is defined as the Lagrange function, and λ is the Lagrange multiplier. Our algorithm aims updating (θ,λ) by the policy gradient iteration $\forall k\in\{0,\cdots,K-1\}$ as shown below from an arbitrary initial point $(\theta_0,\lambda_0=0)$.

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

where α and β denote primal and dual learning rates respectively, and δ indicates the Slater parameter introduced in Assumption 2.1. Moreover, for any $\Lambda\subset\mathbb{R}$, $\mathcal{P}_\Lambda$ indicates the projection onto Λ. Since $\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)$, $F(\theta_k)$, and $J_c(\theta_k)$ are not exactly computable due to lack of knowledge of the exact transition kernel, P, and thus, that of the occupancy measure, $\nu^{\pi_{\theta_k}}$, in most RL scenarios, we use the following approximate updates.

$$\theta_{k+1}=\theta_k+\alpha\omega_k,\lambda_{k+1}=\mathcal{P}_{[0,\frac{2}{\delta}]}[\lambda_k-\beta\eta_c^k] \tag{12}$$

where ω_k is the estimate of the NPG $F(\theta_k)^{-1}\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)$, while η_c^k estimates $J_c(\theta_k)$. Below, we discuss the detailed procedure to compute these estimates.

3.1 Estimation Procedure

We formally characterize our detailed algorithm in Algorithm 1, which utilizes a Multi-Level Monte Carlo (MLMC)-based Actor-Critic algorithm to compute the estimates stated above. For the exposition of our algorithm, it is beneficial to first discuss the critic estimation procedure, where the goal is to estimate the value functions and the average reward and cost functions. These estimates are then further used to obtain the NPG estimates, which are, in turn, used to update the policy parameter. The algorithm runs in K epochs (also called *outer loops*). At the start of kth epoch, the primal and dual parameters are denoted as θ_k,λ_k respectively.

3.1.1 Critic Estimation

At the kth epoch, one of the tasks in critic estimation is to obtain an estimate of $J_g(\theta_k)$, $g\in\{r,c\}$. Note that $J_g(\theta_k)$ can be written as a solution to the following optimization.

$$\min_{\eta\in\mathbb{R}}R_g(\theta_k,\eta)\coloneqq\frac{1}{2}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\nu_g^{\pi_{\theta_k}}(s,a)\left\{\eta-g(s,a)\right\}^2 \tag{13}$$

The second task is to estimate the value function $V_g^{\pi_{\theta_k}}$. To facilitate this objective, it is assumed that, $\forall g\in\{r,c\}$, the value function $V_g^{\pi_{\theta_k}}(\cdot)$ is well-approximated by a linear critic function $\hat{V}_g(\zeta_g^{\theta_k},\cdot)\coloneqq\langle\phi_g(\cdot),\zeta_g^{\theta_k}\rangle$ where $\phi_g:\mathcal{S}\to\mathbb{R}^m$ is a feature mapping with the property that $\|\phi_g(s)\|\leq 1,\forall s\in\mathcal{S}$, and $\zeta_g^{\theta_k}\in\mathbb{R}^m$ is a solution of the following optimization.

$$\min_{\zeta\in\mathbb{R}^m}E_g(\theta_k,\zeta)\coloneqq\frac{1}{2}\sum_{s\in\mathcal{S}}d^{\pi_{\theta_k}}(s)(V_g^{\pi_{\theta_k}}(s)-\hat{V}_g(\zeta,s))^2 \tag{14}$$

Note that the gradients of the functions $R_g(\theta_k,\cdot)$, $E_g(\theta_k,\cdot)$ can be obtained as follows.

$$\nabla_\eta R_g(\theta_k,\eta)=\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\nu_g^{\pi_{\theta_k}}(s,a)\left\{\eta-g(s,a)\right\},\;\forall\eta\in\mathbb{R}, \tag{15}$$

$$\begin{aligned}\nabla_\zeta E_g(\theta_k,\zeta)&=\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\nu_g^{\pi_{\theta_k}}(s,a)\left(\zeta^\top\phi_g(s)-Q_g^{\pi_{\theta_k}}(s,a)\right)\phi_g(s),\;\forall\zeta\in\mathbb{R}^m\\&\overset{(a)}{=}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\nu_g^{\pi_{\theta_k}}(s,a)\left(\zeta^\top\phi_g(s)-g(s,a)+J_g(\theta_k)-\mathbb{E}_{s'\sim P(s,a)}[V_g^{\pi_{\theta_k}}(s')]\right)\phi_g(s)\end{aligned} \tag{16}$$

 https://doi.org/10.52202/085713-3099

where (a) results from Bellman's equation (6). Since the above gradients cannot be exactly obtained due to lack of knowledge of the transition model P, and thus that of $\nu_g^{\pi_{\theta_k}}$, we execute the following H *inner loop* steps to compute approximations of $J_g(\theta_k)$ and $\zeta_g^{\theta_k}$ from $\eta_{g,0}^k = 0$ and $\zeta_{g,0}^k = 0$.

$$\eta_{g,h+1}^k = \eta_{g,h}^k - c_\gamma \gamma_\xi \hat{\nabla}_\eta R_g(\theta_k, \eta_{g,h}^k), \;\; \zeta_{g,h+1}^k = \zeta_{g,h}^k - \gamma_\xi \hat{\nabla}_\zeta E_g(\theta_k, \xi_{g,h}^k) \tag{17}$$

where c_γ is a constant, γ_ξ defines a learning parameter, and $h \in \{0, \cdots, H-1\}$. Moreover, the terms $\hat{\nabla}_\eta R_g(\theta_k, \eta_{g,h}^k)$ and $\hat{\nabla}_\zeta E_g(\theta_k, \xi_{g,h}^k)$ indicate estimates of $\nabla_\eta R_g(\theta_k, \eta_{g,h}^k)$ and $\nabla_\zeta E_g(\theta_k, \zeta_{g,h}^k)$ respectively where $\xi_{g,h}^k \triangleq [\eta_{g,h}^k, (\zeta_{g,h}^k)^\top]^\top$. Because the expression of $\nabla_\zeta E_g(\theta_k, \zeta_{g,h}^k)$ comprises $\zeta_{g,h}^k$ and $J_g(\theta_k)$ (refer (16)), its estimate is a function of $\eta_{g,h}^k$ and $\zeta_{g,h}^k$. At the end of inner loop iterations, we obtain $\eta_{g,H}^k$ and $\zeta_{g,H}^k$ which are estimates of $J_g(\theta_k)$ and $\zeta_g^{\theta_k}$ respectively. It remains to see how the gradient estimates used in (17) can be obtained. Note that (17) can be compactly written as

$$\xi_{g,h+1}^k = \xi_{g,h}^k - \gamma_\xi \mathbf{v}_g(\theta_k, \xi_{g,h}^k), \;\; \mathbf{v}_g(\theta_k, \xi_{g,h}^k) \triangleq [c_\gamma \hat{\nabla}_\eta R_g(\theta_k, \eta_{g,h}^k), (\hat{\nabla}_\zeta E_g(\theta_k, \xi_{g,h}^k))^\top]^\top \tag{18}$$

For a given pair (k, h), let $\mathcal{T}_{kh} = \{(s_{kh}^t, a_{kh}^t, s_{kh}^{t+1})\}_{t=0}^{l_{kh}-1}$ indicate a π_{θ_k}-induced trajectory of length $l_{kh} = 2^{Q_h^k}$ where $Q_h^k \sim \text{Geom}(1/2)$. Following (15) and (16), an estimate of $\mathbf{v}_g(\theta_k, \xi_{g,h}^k)$ based on a single state transition sample $z_{kh}^j = (s_{kh}^j, a_{kh}^j, s_{kh}^{j+1})$ can be obtained as

$$\begin{aligned}
&\mathbf{v}_g(\theta_k, \xi_{g,h}^k; z_{kh}^j) \\
&= \underbrace{\begin{pmatrix} c_\gamma & 0 \\ \phi_g(s_{kh}^j) & \phi_g(s_{kh}^j)(\phi_g(s_{kh}^j) - \phi_g(s_{kh}^{j+1}))^\top \end{pmatrix}}_{\triangleq \mathbf{A}_g(\theta_k; z_{kh}^j)} \xi_{g,h}^k - \underbrace{\begin{pmatrix} c_\gamma g(s_{kh}^j, a_{kh}^j) \\ g(s_{kh}^j, a_{kh}^j)\phi_g(s_{kh}^j) \end{pmatrix}}_{\triangleq \mathbf{b}_g(\theta_k; z_{kh}^j)}
\end{aligned} \tag{19}$$

Applying MLMC-based estimation, the term $\mathbf{v}_g(\theta_k, \xi_{g,h}^k)$ is finally computed as

$$\mathbf{v}_g(\theta_k, \xi_{g,h}^k) = \mathbf{v}_{g,kh}^0 + \begin{cases} 2^{Q_h^k}(\mathbf{v}_{g,kh}^{Q_h^k} - \mathbf{v}_{g,kh}^{Q_h^k - 1}), & \text{if } 2^{Q_h^k} \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

where $T_{\max}$ is a constant, $\mathbf{v}_{g,kh}^j = 2^{-j} \sum_{t=0}^{2^j-1} \mathbf{v}_g(\theta_k, \xi_{g,h}^k; z_{kh}^t)$, $j \in \{0, Q_h^k - 1, Q_h^k\}$. Note that the maximum number of state transition samples utilized in the MLMC estimate is $T_{\max}$. Moreover, it can be demonstrated that the samples used *on average* is $\tilde{O}(\log T_{\max})$. The advantage of the MLMC estimator is that it achieves the same bias as averaging $T_{\max}$ samples, but requires only $\tilde{\mathcal{O}}(\log T_{\max})$ samples. In addition, since drawing from a geometric distribution does not require knowledge of the mixing time, we can eliminate the mixing time knowledge assumption used in previous works [11, 22, 10]. Furthermore, these previous works utilize policy gradient methods and require saving the trajectories of length H for gradient estimations, while our approach does not. Therefore, our algorithm reduces the memory complexity by a factor of H, which is significant since the choice of H in these works scales with mixing and hitting times of the MDP.

3.1.2 Natural Policy Gradient (NPG) Estimator

Recall that the outcome of the critic estimation at the kth epoch is $\xi_{g,H}^k = [\eta_{g,H}^k, (\zeta_{g,H}^k)^\top]^\top$ where $\eta_{g,H}^k$, $\zeta_{g,H}^k$ estimate $J_g(\theta_k)$ and the critic parameter $\zeta_g^{\theta_k}$ respectively. For simplicity, we will denote $\xi_{g,H}^k$ as $\xi_g^k = [\eta_g^k, (\zeta_g^k)^\top]^\top$. We estimate the NPG ω_{g,θ_k}^* (refer to the definition (8)) using a H step inner loop as stated below $\forall h \in \{0, \cdots, H-1\}$ starting from $\omega_{g,0}^k = 0$.

$$\omega_{g,h+1}^k = \omega_{g,h}^k - \gamma_\omega \hat{\nabla}_\omega f_g(\theta_k, \omega_{g,h}^k, \xi_g^k) \tag{21}$$

where $\hat{\nabla}_\omega f_g(\theta_k, \omega_{g,h}^k, \xi_g^k)$ is an MLMC-based estimate of $\nabla_\omega f_g(\theta_k, \omega_{g,h}^k)$ where f_g is given in (9). To obtain this estimate, a π_{θ_k}-induced trajectory $\mathcal{T}_{kh} = \{(s_{kh}^t, a_{kh}^t, s_{kh}^{t+1})\}_{t=0}^{l_{kh}-1}$ of length $l_{kh} = 2^{P_h^k}$ is considered where $P_h^k \sim \text{Geom}(1/2)$. For a certain transition $z_{kh}^j = (s_{kh}^j, a_{kh}^j, s_{kh}^{j+1})$, define the

following estimate.

$$\hat{\nabla}_\omega f_g(\theta_k, \omega^k_{g,h}, \xi^k_g; z^j_{kh}) = \hat{F}(\theta_k; z^j_{kh})\omega^k_{g,h} - \hat{\nabla}_\theta J_g(\theta_k, \xi^k_g; z^j_{kh}) \text{ where} \tag{22}$$
$$\hat{F}(\theta_k; z^j_{kh}) = \nabla_\theta \log \pi_{\theta_k}(a^j_{kh}|s^j_{kh}) \otimes \nabla_\theta \log \pi_{\theta_k}(a^j_{kh}|s^j_{kh})$$
$$\hat{\nabla}_\theta J_g(\theta_k, \xi^k_g; z^j_{kh}) = \hat{A}^{\pi_{\theta_k}}_g(\xi^k_g; z^j_{kh}) \nabla_\theta \log \pi_{\theta_k}(a^j_{kh}|s^j_{kh})$$
$$\overset{(a)}{=} \left[g(s^j_{kh}, a^j_{kh}) - \eta^k_g + \zeta^k_g \left(\phi_g(s^{j+1}_{kh}) - \phi_g(s^j_{kh}) \right) \right] \nabla_\theta \log \pi_{\theta_k}(a^j_{kh}|s^j_{kh})$$

where the advantage estimate used in (a) is essentially a temporal difference (TD) error. Notice that the estimate of the policy gradient $\nabla_\theta J_g(\theta_k)$ depends on ξ^k_g obtained in the critic estimation process. The MLMC-based estimate, therefore, can be obtained as follows.

$$\hat{\nabla}_\omega f_g(\theta_k, \omega^k_{g,h}, \xi^k_g) = \mathbf{u}^0_{g,kh} + \begin{cases} 2^{P^k_h}(\mathbf{u}^{P^k_h}_{g,kh} - \mathbf{u}^{P^k_h - 1}_{g,kh}), & \text{if } 2^{P^k_h} \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where $\mathbf{u}^j_{g,kh} = 2^{-j} \sum_{t=0}^{2^j-1} \hat{\nabla}_\omega f_g(\theta_k, \omega^k_{g,h}, \xi^k_g; z^t_{kh})$, $j \in \{0, P^k_h - 1, P^k_h\}$. The above estimate is applied in (21), which finally yields the NPG estimate $\omega^k_{g,H}$. For simplicity, we denote $\omega^k_{g,H}$ as ω^k_g.

3.2 Primal and Dual Updates

The estimates ω^k_g, $g \in \{r, c\}$ obtained in section 3.1.2 can be combined to compute $\omega_k = \omega^k_r + \lambda_k \omega^k_c$. Moreover, section 3.1.1 provides η^k_c which is an estimate of $J_c(\theta_k)$. At the kth epoch, these estimates can be used to update the policy parameter θ_k and the dual parameter λ_k following (12).

4 Global Convergence Analysis

We first state some assumptions that we will be using before proceeding to the main results. Define $\mathbf{A}_g(\theta) = \mathbb{E}_\theta \mathbf{A}_g(\theta; z)$ where the expectation is over the distribution of $z = (s, a, s')$ where $(s, a) \sim \nu^{\pi_\theta}_g$, $s' \sim P(s, a)$, and the term $\mathbf{A}_g(\theta; z)$ is defined in (19). Similarly, $\mathbf{b}_g(\theta) \triangleq \mathbb{E}_\theta [\mathbf{b}_g(\theta; z)]$. Let $\xi^*_g(\theta) = [\mathbf{A}_g(\theta)]^{-1}\mathbf{b}_g(\theta) = [\eta^*_g(\theta), \zeta^*_g(\theta)]$. With these notations, we are ready to state the following assumptions regarding critic approximations.

Assumption 4.1. For $g \in \{r, c\}$, we define the worst-case critic approximation error to be $\epsilon^{\text{app}}_g = \sup_\theta \mathbb{E}_{s \sim d^{\pi_\theta}} \left[(\zeta^*_g(\theta))^\top \phi_g(s) - V^{\pi_\theta}_g(s) \right]^2$. We assume $\epsilon^{\text{app}} \triangleq \max\{\epsilon^{\text{app}}_r, \epsilon^{\text{app}}_c\}$ to be finite.

Assumption 4.2. There exists $\lambda > 0$ such that $\mathbb{E}_\theta[\phi_g(s) (\phi_g(s) - \phi_g(s'))] - \lambda I$ is positive semi-definite $\forall \theta \in \Theta$ and $g \in \{r, c\}$.

Both Assumptions 4.1 and 4.2 are frequently used in analyzing actor-critic methods [13, 47, 37, 44, 38]. Assumption 4.1 intuitively relates to the quality of the feature mapping where ϵ_{app} serves as a measure of this quality: well-crafted features result in small ϵ_{app}, whereas poorly designed features lead to a larger worst-case error. On the other hand, Assumption 4.2, is essential for ensuring the convergence of the critic updates. It can be shown that Assumption 4.2 ensures that the matrix $\mathbf{A}_g(\theta)$ is invertible for sufficiently large c_γ (details in the appendix).

Assumption 4.3. Define the following function for $\theta, \omega \in \mathbb{R}^{\text{d}}$, $\lambda \geq 0$, and $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$.

$$L_\nu(\omega, \theta, \lambda) = \mathbb{E}_{(s,a) \sim \nu} \left[\left(\nabla_\theta \log \pi_\theta(a|s) \cdot \omega - A^{\pi_\theta}_r(s, a) - \lambda A^{\pi_\theta}_c(s, a) \right)^2 \right] \tag{24}$$

Let $\omega^*_{\theta,\lambda} = \arg\min_{\omega \in \mathbb{R}^{\text{d}}} L_{\nu^{\pi_\theta}}(\omega, \theta, \lambda)$. It is assumed that $L_{\nu^{\pi^*}}(\omega^*_{\theta,\lambda}, \theta, \lambda) \leq \epsilon_{\text{bias}}$ for $\theta \in \Theta$ and $\lambda \in [0, 2/\delta]$ where π^* is the solution to the optimization (2), and ϵ_{bias} is a positive constant. The LHS of the above inequality is called the *transferred compatible function approximation error*. Note that it can be easily verified that $\omega^*_{\theta,\lambda}$ is the NPG update direction $\omega^*_{\theta,\lambda} = F(\theta)^{-1}\nabla_\theta \mathcal{L}(\theta, \lambda)$.

Assumption 4.4. For all $\theta, \theta_1, \theta_2 \in \Theta$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, the following holds for some $G_1, G_2 > 0$.

$(a) \|\nabla_\theta \log \pi_\theta(a|s)\| \leq G_1$
$(b) \|\nabla_\theta \log \pi_{\theta_1}(a|s) - \nabla_\theta \log \pi_{\theta_2}(a|s)\| \leq G_2 \|\theta_1 - \theta_2\|$.

 https://doi.org/10.52202/085713-3099

Assumption 4.5 (Fisher non-degenerate policy)**.** There exists a constant $\mu > 0$ such that $F(\theta) - \mu I_d$ is positive semidefinite, where I_d denotes an identity matrix of dimension d.

Comments on Assumptions 4.3-4.5: We emphasize that these assumptions are commonly applied in the policy gradient (PG) literature [30, 1, 33, 43, 19]. The term ϵ_{bias} reflects the parametrization capacity of π_θ. For π_θ using the softmax parametrization, we directly have $\epsilon_{\text{bias}} = 0$ [1]. However, when π_θ employs a restricted parametrization that does not encompass all stochastic policies, ϵ_{bias} is greater than zero. It is known that ϵ_{bias} remains very small when utilizing rich neural parametrizations [40]. Assumption 4.4 requires that the score function is both bounded and Lipschitz continuous, which is a condition frequently assumed in the analysis of PG-based methods [30, 1, 33, 43, 19]. Assumption 4.5 ensures that the function $f_g(\theta, \cdot)$ is μ-strongly convex by mandating that the eigenvalues of the Fisher information matrix are bounded from below. This is also a standard assumption for deriving global complexity bounds for PG based algorithms [30, 46, 7, 19]. Recent studies have shown that Assumptions 4.4-4.5 hold true in various examples, including Gaussian policies with linearly parameterized means and certain neural parametrizations [30, 19].

Before proving results for the convergence rate, we first provide the following lemma to associate the global convergence rate of the Lagrange function to the convergence of the actor and critic parameters. Similar ideas have been explored in [23, 10].

Lemma 4.6. *If the policy parameters, $\{(\theta_k, \lambda_k)\}_{k=1}^K$ are updated via* (12) *and assumptions 4.3-4.5 hold, then the following inequality is satisfied*

$$\begin{aligned}\frac{1}{K}\mathbb{E}\sum_{k=0}^{K-1}\Big(\mathcal{L}(\pi^*,\lambda_k)-\mathcal{L}(\theta_k,\lambda_k)\Big) \le \sqrt{\epsilon_{\text{bias}}} + \frac{G_1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|(\mathbb{E}_k[\omega_k]-\omega_k^*)\| \\ + \frac{\alpha G_2}{2K}\sum_{k=0}^{K-1}\mathbb{E}\|\omega_k\|^2 + \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))],\end{aligned} \tag{25}$$

where $KL(\cdot\|\cdot)$ is the Kullback-Leibler divergence, π^* is the optimal policy for (2) *and $\omega_k^* \coloneqq \omega_{\theta_k,\lambda_k}^*$ is the exact NPG direction $F(\theta_k)^{-1}\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)$. Finally, $\mathbb{E}_k$ denotes conditional expectation given history up to the kth iteration.*

Observe the presence of ϵ_{bias} in (25). It dictates that due to the incompleteness of the policy class, the global convergence bound cannot be driven to zero. Note that the last term in (25) is $\mathcal{O}(1/(\alpha K))$ because the term $\mathbb{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))]$ is a constant. The term related to $\mathbb{E}\|\omega_k\|^2$ can be further decomposed as follows.

$$\begin{aligned}\frac{\alpha}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\omega_k\|^2 &\le \frac{\alpha}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\omega_k-\omega_k^*\|^2 + \frac{\alpha}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\omega_k^*\|^2 \\ &\overset{(a)}{\le} \frac{\alpha}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\omega_k-\omega_k^*\|^2 + \frac{\alpha\mu^{-2}}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)\|^2\end{aligned} \tag{26}$$

where (a) follows from Assumption 4.5 and the definition that $\omega_k^* = F(\theta_k)^{-1}\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)$. We can obtain a global convergence bound by bounding the terms $\mathbb{E}\|\omega_k-\omega_k^*\|^2$, $\mathbb{E}\|(\mathbb{E}[\omega_k|\theta_k]-\omega_k^*)\|$ and $\mathbb{E}\|\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)\|^2$. The first two terms are the variance and bias of the NPG estimator, ω_k, and the third term indicates the local convergence rate. Further, $\mathbb{E}\|\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)\|^2$ can be upper bounded by a constant (Lemma G.2 in the appendix). We now provide the convergence result for the actor and critic parameters. For brevity, we use $\lesssim$ to denote $\le \tilde{\mathcal{O}}(\cdot)$.

Theorem 4.7. *Consider Algorithm 1 and let Assumptions 2.2-4.5 hold. If J_g is L-smooth, $g \in \{r, c\}$, $\gamma_\omega = \frac{2\log T}{\mu H}$ is such that $\gamma_\omega \le \frac{\mu}{4(6G_1^4\tau_{\text{mix}}\log T_{\max}+2G_1^2\tau_{\text{mix}}^2\log T_{\max})}$, and $T_{\max}$ obeys $T_{\max} \ge \frac{8G_1^4\tau_{\text{mix}}}{\mu}$ where μ is defined in Assumption 4.5, the following inequalities hold $\forall k \in \{0,\cdots,K-1\}$.*

$$\begin{aligned}\left\|\mathbb{E}_k[\omega_g^k]-\omega_{g,k}^*\right\|^2 &\lesssim \epsilon_{\text{app}} + \frac{\tau_{\text{mix}}^2}{T^2} \\ &\quad + \frac{\tau_{\text{mix}}^2}{T_{\max}}\left\{\mathbb{E}_k\left[\left\|\xi_g^k-\xi_g^*(\theta_k)\right\|^2\right]+\tau_{\text{mix}}^2\right\} + \left\|\mathbb{E}_k[\xi_g^k]-\xi_g^*(\theta_k)\right\|^2,\end{aligned} \tag{27}$$

$$\mathbb{E}_k\left[\|\omega_g^k-\omega_{g,k}^*\|^2\right] \lesssim \epsilon_{\text{app}} + \left(\frac{1}{T^2}+\frac{\tau_{\text{mix}}}{H}+\frac{\tau_{\text{mix}}}{T_{\max}}\right)\tau_{\text{mix}}^2 + \mathbb{E}_k\left[\left\|\xi_g^k-\xi_g^*(\theta_k)\right\|^2\right] \tag{28}$$

where $\omega^*_{g,k}$ is the NPG direction $F(\theta_k)^{-1}\nabla_\theta J_g(\theta_k)$, $\xi^*_g(\theta_k)$ is defined in section 4, and $\mathbb{E}_k$ denotes conditional expectation given history up to the kth iteration.

Theorem 4.7 bounds the NPG bias $\|\mathbb{E}_k[\omega^k_g] - \omega^*_{g,k}\|$ and the second-order error $\mathbb{E}_k[\|\omega^k_g - \omega^*_{g,k}\|^2]$ in terms of the critic approximation error ϵ_{app}, and the bias $\|\mathbb{E}_k[\xi^k_g] - \xi^*_g(\theta_k)\|$ and the second-order error $\mathbb{E}_k[\|\xi^k_g - \xi^*_g(\theta_k)\|^2]$ in the critic parameter estimation. The following theorem provides bounds on these latter quantities.

Theorem 4.8. *Consider Algorithm 1 and let Assumptions 2.2-4.5 hold and $g \in \{r, c\}$. If we choose $\gamma_\xi = \frac{2\log T}{\lambda H}$ such that $\gamma_\xi \leq \frac{\lambda}{24c^2_\gamma \tau_{\text{mix}} \log T_{\max}}$ while $T_{\max}$ obeys $T_{\max} \geq \frac{8c^2_\gamma \tau_{\text{mix}}}{\lambda}$ where λ is defined in Assumption 4.2, the following inequalities hold $\forall k \in \{0, \cdots, K-1\}$.*

$$\left\|\mathbb{E}_k[\xi^k_g] - \xi^*_g(\theta_k)\right\|^2 \lesssim \frac{1}{T^2} + \frac{\tau^2_{\text{mix}}}{T_{\max}}, \tag{29}$$

$$\mathbb{E}_k[\|\xi^k_g - \xi^*_g(\theta_k)\|^2] \lesssim \frac{1}{T^2} + \frac{\tau_{\text{mix}}}{H} + \frac{\tau_{\text{mix}}}{T_{\max}} \tag{30}$$

where $\xi^*_g(\theta_k)$ is defined in section 4 and $\mathbb{E}_k$ denotes conditional expectation given history up to the kth iteration.

Invoking the bounds provided by Theorem 4.7 and 4.8 into Lemma 4.6, we can obtain a convergence rate of the Lagrange function. Our next goal is to segregate the objective convergence and constraint violation rates from this Lagrange error. This is achieved by the following theorems. Depending on whether we have access to the mixing time, two similar but slightly different results can be obtained.

Theorem 4.9. *Consider the same setup and parameters as in Theorem 4.8 and Theorem 4.7 and set $\alpha = T^{-1/2}$, $\beta = T^{-1/2}$. If τ_{mix} is known, set $H = \tilde{\Theta}(\tau^2_{\text{mix}})$ with $K = T/H$. We have:*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[J^{\pi^*}_r - J_r(\theta_k)] \lesssim \sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \frac{1}{\sqrt{T}}, \tag{31}$$

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[-J_c(\theta_k)] \lesssim \sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \frac{1}{\sqrt{T}} \tag{32}$$

Theorem 4.10. *Consider the same setup and parameters as in Theorem 4.8 and Theorem 4.7 and set $\alpha = T^{-1/2}$, $\beta = T^{-1/2}$, $H = T^\epsilon$ and $K = T^{1-\epsilon}$. With $T^\epsilon \geq \tilde{\Theta}(\tau^2_{\text{mix}})$, we have:*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[J^{\pi^*}_r - J_r(\theta_k)] \lesssim \sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \frac{1}{T^{(0.5-\epsilon)}}, \tag{33}$$

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[-J_c(\theta_k)] \lesssim \sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \frac{1}{T^{(0.5-\epsilon)}} \tag{34}$$

Theorem 4.9 dictates that one can achieve both the objective convergence and the constraint violation rates as $\tilde{\mathcal{O}}(T^{-1/2})$ up to some additive factors of ϵ_{bias} and ϵ_{app} where T is the length of the horizon. However, knowledge of the mixing time, τ_{mix}, is needed in this case to set some parameters. On the other hand, Theorem 4.10 states that, if knowledge of τ_{mix} is unavailable, one can achieve objective convergence and constraint violation rates as $\tilde{\mathcal{O}}(T^{-1/2+\epsilon})$ as long as the horizon T exceeds $\tilde{\Theta}(\tau^{2/\epsilon}_{\text{mix}})$. Note that one can get arbitrarily close to the optimal rate of $\tilde{\mathcal{O}}(T^{-1/2})$ by choosing arbitrarily small ϵ. The caveat is that the smaller the ϵ, the larger the horizon length needed to reach the desired rates.

In the setting of Theorem 4.10, for small ϵ, the horizon requirement becomes $T \gtrsim \tilde{O}\left(\tau^{2/\varepsilon}_{\text{mix}}\right)$, which can be large for slowly mixing problems. In such a scenario, one can switch to the parameter setting of Theorem 4.9, which does not impose any such restriction on T. However, since this setup requires the knowledge of τ_{mix}, one can treat τ_{mix} as one of the unknown hyperparameters of the algorithm, and fine-tune it during the training phase. This is in line with other RL algorithms in the literature that also require fine-tuning of several unknown hyperparameters, such as the Lipschitz constant or hitting time [11]. Despite these practical solutions, we acknowledge that a systematic theoretical investigation is

 https://doi.org/10.52202/085713-3099

needed to improve the requirement of the horizon length, T (in the absence of knowledge of τ_{mix}), which is left as one of the future works.

It is also worth highlighting that due to the presence of ϵ_{bias} and ϵ_{app}, the average objective error and constraint violation cannot be guaranteed to be zero, even for large T. However, for rich policy parameterization and good critic approximation, the effects of these quantities are negligibly small.

5 Conclusion

In this paper, we investigate the infinite-horizon average reward Constrained Markov Decision Processes (CMDPs) with general policy parametrization. We propose a novel algorithm, the "Primal-dual natural actor-critic," which efficiently manages constraints while achieving a global convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$, aligning with the theoretical lower bound for Markov Decision Processes (MDPs). We also extend our analysis to the setting with unknown mixing time. Future directions include narrowing the performance gap in this setting, parameterizing the critic using neural networks as in [25, 21], and relaxing the ergodicity assumption following [20].

6 Acknowledgment

The work was supported in part by the National Science Foundation under grant CCF-2149588, Office of Naval Research under grant N00014-23-1-2532, and Cisco Systems, Inc.

References

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

[2] Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Concave Utility Reinforcement Learning with Zero-Constraint Violations. *Transactions on Machine Learning Research*, 2022.

[3] Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Uncertainty in Artificial Intelligence*, pages 22–31, 2022.

[4] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[6] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[7] Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving Zero Constraint Violation for Constrained Reinforcement Learning via Primal-Dual Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:3682–3689, Jun. 2022.

[8] Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3682–3689, 2022.

[9] Qinbo Bai, Amrit Singh Bedi, and Vaneet 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*, volume 37, pages 6737–6744, 2023.

[10] Qinbo Bai, Washim Mondal, and Vaneet Aggarwal. Learning general parameterized policies for infinite horizon average reward constrained MDPs via primal-dual policy gradient algorithm. *Advances in Neural Information Processing Systems*, 37:108566–108599, 2024.

[11] Qinbo Bai, Washim Uddin Mondal, and Vaneet Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward Markov decision processes. In *AAAI Conference on Artificial Intelligence*, 2024.

[12] Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First Order Methods with Markovian Noise: from Acceleration to Variational Inequalities. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[13] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation. *CoRR*, abs/1806.02450, 2018.

[14] Jiayu Chen, Tian Lan, and Vaneet Aggarwal. Option-aware adversarial inverse reinforcement learning for robotic control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5902–5908. IEEE, 2023.

[15] Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward Markov decision process with constraints. In *International Conference on Machine Learning*, pages 3246–3270. PMLR, 2022.

[16] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

[17] Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Başar, and Mihailo R Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs. *arXiv preprint arXiv:2206.02346*, 2022.

[18] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

[19] Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869. PMLR, 2023.

[20] Swetha Ganesh and Vaneet Aggarwal. Regret analysis of average-reward unichain mdps via an actor-critic approach. In *Advances in Neural Information Processing Systems*, 2025.

[21] Swetha Ganesh, Jiayu Chen, Washim Uddin Mondal, and Vaneet Aggarwal. Order-optimal global convergence for actor-critic with general policy and neural critic parametrization. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

[22] Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. Order-Optimal Regret with Novel Policy Gradient Approaches in Infinite Horizon Average Reward MDPs. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[23] Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. A sharper global convergence analysis for average reward reinforcement learning via an actor-critic approach. *International Conference on Machine Learning*, 2025.

[24] Ather Gattami, Qinbo Bai, and Vaneet Aggarwal. Reinforcement learning for constrained markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 2656–2664. PMLR, 2021.

[25] Mudit Gaur, Amrit Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: Achieving global convergence (last iterate) of actor-critic under markovian sampling with neural network parametrization. In *International Conference on Machine Learning*, pages 15153–15179. PMLR, 2024.

[26] Jacopo Germano, Francesco Emanuele Stradi, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. A Best-of-Both-Worlds Algorithm for Constrained MDPs with Long-Term Constraints. *arXiv preprint arXiv:2304.14326*, 2023.

 https://doi.org/10.52202/085713-3099

[27] Arnob Ghosh, Xingyu Zhou, and Ness 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.

[28] Marina Haliem, Ganapathy Mani, Vaneet Aggarwal, and Bharat Bhargava. A distributed model-free ride-sharing approach for joint matching, pricing, and dispatching using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 22(12):7931–7942, 2021.

[29] Lu Ling, Washim Uddin Mondal, and Satish V Ukkusuri. Cooperating graph neural networks with deep reinforcement learning for vaccine prioritization. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[30] Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

[31] Washim Mondal and Vaneet Aggarwal. Sample-efficient constrained reinforcement learning with general parameterization. *Advances in Neural Information Processing Systems*, 37:68380–68405, 2024.

[32] Mahadesh Panju, Ramkumar Raghu, Vinod Sharma, Vaneet Aggarwal, and Rajesh Ramachandran. Queueing theoretic models for uncoded and coded multicast wireless networks with caches. *IEEE Transactions on Wireless Communications*, 21(2):1257–1271, 2021.

[33] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pages 4026–4035, 2018.

[34] Bhrij Patel, Wesley A Suttle, Alec Koppel, Vaneet Aggarwal, Brian M Sadler, Dinesh Manocha, and Amrit Bedi. Towards global optimality for practical average reward reinforcement learning without mixing time oracles. In *Forty-first International Conference on Machine Learning*, 2024.

[35] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[36] Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

[37] Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021.

[38] Wesley A Suttle, Amrit Bedi, Bhrij Patel, Brian M Sadler, Alec Koppel, and Dinesh Manocha. Beyond Exponentially Fast Mixing in Average-Reward Reinforcement Learning via Multi-Level Monte Carlo Actor-Critic. In *International Conference on Machine Learning*, pages 33240–33267, 2023.

[39] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[40] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural Policy Gradient Methods: Global Optimality and Rates of Convergence. In *International Conference on Learning Representations*, 2019.

[41] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR, 2020.

[42] Honghao Wei, Xin Liu, and Lei 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*, volume 36, pages 3868–3876, 2022.

[43] Pan Xu, Felicia Gao, and Quanquan Gu. Sample Efficient Policy Gradient Methods with Recursive Variance Reduction. In *International Conference on Learning Representations*, 2019.

[44] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551, 2020.

[45] Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

[46] Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the Convergence and Sample Efficiency of Variance-Reduced Policy Gradient Method. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[47] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. *Advances in neural information processing systems*, 32, 2019.

 https://doi.org/10.52202/085713-3099

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 claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   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 point out the limitations of our work.

   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?

   Answer: [Yes]

Justification: We provide the full set of assumptions and complete proof for each theoretical result.

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: Our 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?

 https://doi.org/10.52202/085713-3099

Answer: [NA]

Justification: Our 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: Our 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: Our 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.)
   - 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: Our 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: Our paper conforms, in every respect, with 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 potential societal impact of our work.

    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

 https://doi.org/10.52202/085713-3099

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: Our 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: Our 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.
    - 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: Our paper does not 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: Our paper involves neither crowdsourcing nor 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: Our paper neither involves crowdsourcing nor 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.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

 https://doi.org/10.52202/085713-3099

Justification: We do not use LLM in our paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Related Works

The constrained reinforcement learning problem has been widely explored for both infinite horizon discounted reward and episodic MDPs. Recent studies have examined discounted reward CMDPs in various contexts, including the tabular setting [8], with softmax parameterization [16, 45], and with general policy parameterization setting [16, 45, 9, 31]. Additionally, episodic CMDPs have been explored in the tabular setting by [18, 36, 26]. Recent research has also focused on infinite horizon average reward CMDPs, examining various approaches including model-based setups [15, 3, 2], tabular model-free settings [42], linear CMDP setting [27] and general policy parametrization setting [10]. In model-based CMDP setups, [2] introduced algorithms leveraging posterior sampling and the optimism principle, achieving a convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$ with zero constraint violations. For tabular model-free approach, [42] attains a convergence rate of $\tilde{\mathcal{O}}(T^{-1/6})$ with zero constraint violations. In linear CMDP, [27] obtains $\tilde{\mathcal{O}}(T^{-1/2})$ convergence rate with zero constraint violation. Note that linear CMDP assumes a linear structure in the transition probability based on a known feature map, which is not realistic. Finally, [10] studied the infinite horizon average reward CMDPs with general parametrization and achieved a global convergence rate of $\tilde{\mathcal{O}}(1/T^{1/5})$. Table 1 summarizes all relevant works on average reward CMDPs.

In unconstrained average reward MDPs, both model-based and model-free tabular setups have been widely studied. For instance, the model-based algorithms by [4, 6] obtain the optimal convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$. Similarly, the model-free algorithm proposed by [41] achieves $\tilde{\mathcal{O}}(1/\sqrt{T})$ convergence rate for tabular MDP. Average reward MDP with general parametrization has been recently studied in [23, 22], where global convergence rates of $\tilde{\mathcal{O}}(1/\sqrt{T})$ are achieved. In particular, [23] leverages Actor-Critic methods and achieves global convergence without knowledge of mixing time.

B Preliminary Results for Global Convergence Analysis

To prove Theorem 4.8 and 4.7, we first discuss a more general case of linear recursion with biased estimators.

Theorem B.1. *Consider the stochastic linear recursion that aims to approximate $x^* = P^{-1}q$.*

$$x_{h+1} = x_h - \bar{\beta}(\hat{P}_h x_h - \hat{q}_h) \tag{35}$$

where $\hat{P}_h$, $\hat{q}_h$ are estimates of the matrices $P \in \mathbb{R}^{n\times n}$, $q \in \mathbb{R}^n$ respectively, and $h \in \{0, \cdots, H-1\}$. Assume that the following bounds hold $\forall h$.

$$\begin{aligned} &\mathbb{E}_h\left[\left\|\hat{P}_h - P\right\|^2\right] \le \sigma_P^2,\ \left\|\mathbb{E}_h\left[\hat{P}_h\right] - P\right\|^2 \le \delta_P^2, \\ &\mathbb{E}_h\left[\|\hat{q}_h - q\|^2\right] \le \sigma_q^2,\ \|\mathbb{E}_h\left[\hat{q}_h\right] - q\|^2 \le \delta_q^2,\ \textit{and}\ \|\mathbb{E}\left[\hat{q}_h\right] - q\|^2 \le \bar{\delta}_q^2 \end{aligned} \tag{36}$$

where $\mathbb{E}_h$ denotes conditional expectation given history up to step h. Since $\mathbb{E}[\hat{q}_h] = \mathbb{E}[\mathbb{E}_h[\hat{q}_h]]$, we have $\bar{\delta}_q^2 \le \delta_q^2$. Additionally, assume that

$$0 < \lambda_P \le \|P\| \le \Lambda_P \ \textit{and}\ \|q\| \le \Lambda_q \tag{37}$$

The condition that $\lambda_P > 0$ implies that P is invertible. The goal of recursion (35) *is to approximate the term $x^* = P^{-1}q$. If $\delta_P \le \lambda_P/8$, and $\bar{\beta} \le \lambda_P/[4(6\sigma_P^2 + 2\Lambda_P^2)]$, the following relation holds.*

$$\mathbb{E}\left[\|x_H - x^*\|^2\right] \le \exp\left(-H\bar{\beta}\lambda_P\right)\mathbb{E}\|x_0 - x^*\|^2 + \tilde{\mathcal{O}}\left(\bar{\beta}R_0 + R_1\right) \tag{38}$$

where $R_0 = \lambda_P^{-3}\Lambda_q^2\sigma_P^2 + \lambda_P^{-1}\sigma_q^2$, $R_1 = \lambda_P^{-2}\left[\delta_P^2\lambda_P^{-2}\Lambda_q^2 + \delta_q^2\right]$, and $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors of H. Moreover,

$$\begin{aligned} \|\mathbb{E}[x_H] - x^*\|^2 &\le \exp\left(-H\bar{\beta}\lambda_P\right)\|\mathbb{E}[x_0] - x^*\|^2 \\ &\quad + \frac{6}{\lambda_P}\left(\bar{\beta} + \frac{1}{\lambda_P}\right)\left[\delta_P^2\left\{\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \mathcal{O}\left(\bar{\beta}R_0 + R_1\right)\right\} + \bar{R}_1\right] \end{aligned} \tag{39}$$

where $\bar{R}_1 = \delta_P^2\lambda_P^{-2}\Lambda_q^2 + \bar{\delta}_q^2$.

https://doi.org/10.52202/085713-3099

Proof. Let $g_h = \hat{P}_h x_h - \hat{q}_h$. To prove the first statement, observe the following relations.

$$\begin{aligned}
&\|x_{h+1} - x^*\|^2 = \|x_h - \bar{\beta} g_h - x^*\|^2 \\
&= \|x_h - x^*\|^2 - 2\bar{\beta}\langle x_h - x^*, g_h\rangle + \bar{\beta}^2\|g_h\|^2 \\
&= \|x_h - x^*\|^2 - 2\bar{\beta}\langle x_h - x^*, P(x_h - x^*)\rangle - 2\bar{\beta}\langle x_h - x^*, g_h - P(x_h - x^*)\rangle + \bar{\beta}^2\|g_h\|^2 \\
&\overset{(a)}{\leq} \|x_h - x^*\|^2 - 2\bar{\beta}\lambda_P\|x_h - x^*\|^2 - 2\bar{\beta}\langle x_h - x^*, g_h - P(x_h - x^*)\rangle \\
&\qquad + 2\bar{\beta}^2\|g_h - P(x_h - x^*)\|^2 + 2\bar{\beta}^2\|P(x_h - x^*)\|^2 \\
&\overset{(b)}{\leq} \|x_h - x^*\|^2 - 2\bar{\beta}\lambda_P\|x_h - x^*\|^2 - 2\bar{\beta}\langle x_h - x^*, g_h - P(x_h - x^*)\rangle \\
&\qquad + 2\bar{\beta}^2\|g_h - P(x_h - x^*)\|^2 + 2\Lambda_P^2\bar{\beta}^2\|x_h - x^*\|^2
\end{aligned}$$

where (a), (b) follow from $\lambda_P \leq \|P\| \leq \Lambda_P$. Taking the conditional expectation $\mathbb{E}_h$ on both sides,

$$\begin{aligned}
\mathbb{E}_h\left[\|x_{h+1} - x^*\|^2\right] \leq (1 - 2\bar{\beta}\lambda_P + 2\Lambda_P^2\bar{\beta}^2)\|x_h - x^*\|^2 \\
- 2\bar{\beta}\langle x_h - x^*, \mathbb{E}_h\left[g_h - P(x_h - x^*)\right]\rangle + 2\bar{\beta}^2\mathbb{E}_h\left\|g_h - P(x_h - x^*)\right\|^2
\end{aligned} \tag{40}$$

Observe that the third term in (40) can be bounded as follows.

$$\begin{aligned}
\|g_h - P(x_h - x^*)\|^2 &= \|(\hat{P}_h - P)(x_h - x^*) + (\hat{P}_h - P)x^* + (q - \hat{q}_h)\|^2 \\
&\leq 3\|\hat{P}_h - P\|^2\|x_h - x^*\|^2 + 3\|\hat{P}_h - P\|^2\|x^*\|^2 + 3\|q - \hat{q}_h\|^2 \\
&\leq 3\|\hat{P}_h - P\|^2\|x_h - x^*\|^2 + 3\lambda_P^{-2}\Lambda_q^2\|\hat{P}_h - P\|^2 + 3\|q - \hat{q}_h\|^2
\end{aligned}$$

where the last inequality follows from $\|x^*\|^2 = \left\|P^{-1}q\right\|^2 \leq \lambda_P^{-2}\Lambda_q^2$. Taking the expectation yields

$$\begin{aligned}
&\mathbb{E}_h\|g_h - P(x_h - x^*)\|^2 \\
&\quad \leq 3\mathbb{E}_h\|\hat{P}_h - P\|^2\|x_h - x^*\|^2 + 3\lambda_P^{-2}\Lambda_q^2\mathbb{E}_h\|\hat{P}_h - P\|^2 + 3\mathbb{E}_h\|\hat{q}_h - q\|^2 \\
&\quad \leq 3\sigma_P^2\left\|x_h - x^*\right\|^2 + 3\lambda_P^{-2}\Lambda_q^2\sigma_P^2 + 3\sigma_q^2
\end{aligned} \tag{41}$$

The second term in (40) can be bounded as

$$\begin{aligned}
-\langle x_h - x^*, \mathbb{E}_h\left[g_h - P(x_h - x^*)\right]\rangle &\leq \frac{\lambda_P}{4}\|x_h - x^*\|^2 + \frac{1}{\lambda_P}\left\|\mathbb{E}_h[g_h - P(x_h - x^*)]\right\|^2 \\
&\leq \frac{\lambda_P}{4}\|x_h - x^*\|^2 + \frac{1}{\lambda_P}\left\|\left\{\mathbb{E}_h[\hat{P}_h] - P\right\}x_h + \left\{q - \mathbb{E}_h\left[\hat{q}_h\right]\right\}\right\|^2 \\
&\leq \frac{\lambda_P}{4}\|x_h - x^*\|^2 + \frac{2\delta_P^2\|x_h\|^2 + 2\delta_q^2}{\lambda_P} \\
&\leq \frac{\lambda_P}{4}\|x_h - x^*\|^2 + \frac{4\delta_P^2\|x_h - x^*\|^2 + 4\delta_P^2\lambda_P^{-2}\Lambda_q^2 + 2\delta_q^2}{\lambda_P}
\end{aligned} \tag{42}$$

where the last inequality follows from $\|x^*\|^2 = \left\|P^{-1}q\right\|^2 \leq \lambda_P^{-2}\Lambda_q^2$. Substituting the above bounds in (40), we arrive at the following.

$$\begin{aligned}
\mathbb{E}_h\left[\|x_{h+1} - x^*\|^2\right] \leq \left(1 - \frac{3\bar{\beta}\lambda_P}{2} + \frac{8\bar{\beta}\delta_P^2}{\lambda_P} + 6\bar{\beta}^2\sigma_P^2 + 2\bar{\beta}^2\Lambda_P^2\right)\|x_h - x^*\|^2 \\
+ \frac{4\bar{\beta}}{\lambda_P}\left[2\delta_P^2\lambda_P^{-2}\Lambda_q^2 + \delta_q^2\right] + 6\bar{\beta}^2\left[\lambda_P^{-2}\Lambda_q^2\sigma_P^2 + \sigma_q^2\right]
\end{aligned}$$

For $\delta_P \leq \lambda_P/8$, and $\bar{\beta} \leq \lambda_P/[4(6\sigma_P^2 + 2\Lambda_P^2)]$, we can modify the above inequality to the following.

$$\mathbb{E}_h[\|x_{h+1} - x^*\|^2] \leq \left(1 - \bar{\beta}\lambda_P\right)\|x_h - x^*\|^2 + \frac{4\bar{\beta}}{\lambda_P}\left[2\delta_P^2\lambda_P^{-2}\Lambda_q^2 + \delta_q^2\right] + 6\bar{\beta}^2\left[\lambda_P^{-2}\Lambda_q^2\sigma_P^2 + \sigma_q^2\right]$$

Taking the expectation on both sides and unrolling the recursion yields

$$\begin{aligned}
\mathbb{E}[\|x_H - x^*\|^2] &\leq \left(1-\bar{\beta}\lambda_P\right)^H \mathbb{E}\|x_0 - x^*\|^2 \\
&\quad + \sum_{h=0}^{H-1} \left(1-\bar{\beta}\lambda_P\right)^h \left\{ \frac{4\bar{\beta}}{\lambda_P}\left[2\delta_P^2\lambda_P^{-2}\Lambda_q^2 + \delta_q^2\right] + 6\bar{\beta}^2\left[\lambda_P^{-2}\Lambda_q^2\sigma_P^2 + \sigma_q^2\right]\right\} \\
&\leq \exp\left(-H\bar{\beta}\lambda_P\right)\mathbb{E}\|x_0 - x^*\|^2 + \frac{1}{\bar{\beta}\lambda_P}\left\{ \frac{4\bar{\beta}}{\lambda_P}\left[2\delta_P^2\lambda_P^{-2}\Lambda_q^2 + \delta_q^2\right] + 6\bar{\beta}^2\left[\lambda_P^{-2}\Lambda_q^2\sigma_P^2 + \sigma_q^2\right]\right\} \\
&= \exp\left(-H\bar{\beta}\lambda_P\right)\mathbb{E}\|x_0 - x^*\|^2 + \left\{ 4\lambda_P^{-2}\left[2\delta_P^2\lambda_P^{-2}\Lambda_q^2 + \delta_q^2\right] + 6\bar{\beta}\lambda_P^{-1}\left[\lambda_P^{-2}\Lambda_q^2\sigma_P^2 + \sigma_q^2\right]\right\}
\end{aligned}$$

To prove the second statement, observe that we have the following recursion.

$$\begin{aligned}
\|\mathbb{E}[x_{h+1}] - x^*\|^2 &= \|\mathbb{E}[x_h] - \bar{\beta}\mathbb{E}[g_h] - x^*\|^2 \\
&= \|\mathbb{E}[x_h] - x^*\|^2 - 2\bar{\beta}\langle \mathbb{E}[x_h] - x^*, \mathbb{E}[g_h]\rangle + \bar{\beta}^2\|\mathbb{E}[g_h]\|^2 \\
&= \|\mathbb{E}[x_h] - x^*\|^2 - 2\bar{\beta}\langle \mathbb{E}[x_h] - x^*, P(\mathbb{E}[x_h] - x^*)\rangle \\
&\qquad - 2\bar{\beta}\langle \mathbb{E}[x_h] - x^*, \mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\rangle + \bar{\beta}^2\|\mathbb{E}[g_h]\|^2 \\
&\overset{(a)}{\leq} \|\mathbb{E}[x_h] - x^*\|^2 - 2\bar{\beta}\lambda_P\|\mathbb{E}[x_h] - x^*\|^2 - 2\bar{\beta}\langle \mathbb{E}[x_h] - x^*, \mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\rangle \\
&\qquad + 2\bar{\beta}^2\|\mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\|^2 + 2\bar{\beta}^2\|P(\mathbb{E}[x_h] - x^*)\|^2 \\
&\overset{(b)}{\leq} \|\mathbb{E}[x_h] - x^*\|^2 - 2\bar{\beta}\lambda_P\|\mathbb{E}[x_h] - x^*\|^2 - 2\bar{\beta}\langle \mathbb{E}[x_h] - x^*, \mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\rangle \\
&\qquad + 2\bar{\beta}^2\|\mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\|^2 + 2\Lambda_P^2\bar{\beta}^2\|\mathbb{E}[x_h] - x^*\|^2 \\
&\leq (1 - 2\bar{\beta}\lambda_P + 2\Lambda_P^2\bar{\beta}^2)\|\mathbb{E}[x_h] - x^*\|^2 - 2\bar{\beta}\langle \mathbb{E}[x_h] - x^*, \mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\rangle \\
&\qquad + 2\bar{\beta}^2\left\|\mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\right\|^2 \qquad (43)
\end{aligned}$$

where (a) and (b) are consequences of $\lambda_P \leq \|P\| \leq \Lambda_P$. The third term in the last line of (43) can be bounded as follows.

$$\begin{aligned}
\|\mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\|^2 &= \left\|\mathbb{E}\left[(\hat{P}_h - P)(x_h - x^*)\right] + (\mathbb{E}[\hat{P}_h] - P)x^* + (q - \mathbb{E}[\hat{q}_h])\right\|^2 \\
&\leq 3\mathbb{E}\left[\|\mathbb{E}_h[\hat{P}_h] - P\|^2\|x_h - x^*\|^2\right] + 3\mathbb{E}\left[\|\mathbb{E}_h[\hat{P}_h] - P\|^2\right]\|x^*\|^2 + 3\left\|q - \mathbb{E}[\hat{q}_h]\right\|^2 \\
&\leq 3\delta_P^2\mathbb{E}\left[\|x_h - x^*\|^2\right] + 3\lambda_P^{-2}\Lambda_q^2\delta_P^2 + 3\bar{\delta}_q^2 \\
&\overset{(a)}{\leq} 3\delta_P^2\left\{\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \mathcal{O}\left(\bar{\beta}R_0 + R_1\right)\right\} + 3\lambda_P^{-2}\Lambda_q^2\delta_P^2 + 3\bar{\delta}_q^2
\end{aligned}$$

where (a) follows from (38). The second term in the last line of (43) can be bounded as follows.

$$\begin{aligned}
&-\langle \mathbb{E}[x_h] - x^*, \mathbb{E}_h\left[\mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\right]\rangle \\
&\leq \frac{\lambda_P}{4}\|\mathbb{E}[x_h] - x^*\|^2 + \frac{1}{\lambda_P}\left\|\mathbb{E}[g_h] - P(\mathbb{E}[x_h] - x^*)\right\|^2 \\
&\leq \frac{\lambda_P}{4}\|\mathbb{E}[x_h] - x^*\|^2 + \frac{3}{\lambda_P}\left[\delta_P^2\left\{\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \mathcal{O}\left(\bar{\beta}R_0 + R_1\right)\right\} + \lambda_P^{-2}\Lambda_q^2\delta_P^2 + \bar{\delta}_q^2\right]
\end{aligned}$$

Substituting the above bounds in (43), we obtain the following recursion.

$$\begin{aligned}
\|\mathbb{E}[x_{h+1}] - x^*\|^2 &\leq \left(1 - \frac{3\bar{\beta}\lambda_P}{2} + 2\Lambda_P^2\bar{\beta}^2\right)\|\mathbb{E}[x_h] - x^*\|^2 \\
&\qquad + 6\bar{\beta}\left(\bar{\beta} + \frac{1}{\lambda_P}\right)\left[\delta_P^2\left\{\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \mathcal{O}\left(\bar{\beta}R_0 + R_1\right)\right\} + \lambda_P^{-2}\Lambda_q^2\delta_P^2 + \bar{\delta}_q^2\right]
\end{aligned}$$

If $\bar{\beta} < \lambda_P/(4\Lambda_P^2)$, the above bound implies the following.

$$\begin{aligned}
\|\mathbb{E}[x_{h+1}] - x^*\|^2 &\leq \left(1 - \bar{\beta}\lambda_P\right)\|\mathbb{E}[x_h] - x^*\|^2 \\
&\qquad + 6\bar{\beta}\left(\bar{\beta} + \frac{1}{\lambda_P}\right)\left[\delta_P^2\left\{\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \mathcal{O}\left(\bar{\beta}R_0 + R_1\right)\right\} + \lambda_P^{-2}\Lambda_q^2\delta_P^2 + \bar{\delta}_q^2\right]
\end{aligned}$$

https://doi.org/10.52202/085713-3099

Unrolling the above recursion, we obtain

$$\begin{aligned}
&\|\mathbb{E}[x_H] - x^*\|^2 \leq \left(1 - \bar{\beta}\lambda_P\right)^H \|\mathbb{E}[x_0] - x^*\|^2 \\
&+ \sum_{h=0}^{H-1} 6\left(1 - \bar{\beta}\lambda_P\right)^h \bar{\beta}\left(\bar{\beta} + \frac{1}{\lambda_P}\right)\left[\delta_P^2\left\{\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \mathcal{O}\left(\bar{\beta}R_0 + R_1\right)\right\} + \lambda_P^{-2}\Lambda_q^2\delta_P^2 + \bar{\delta}_q^2\right] \\
&\leq \exp\left(-H\bar{\beta}\lambda_P\right)\|\mathbb{E}[x_0] - x^*\|^2 \\
&\qquad + \frac{6}{\lambda_P}\left(\bar{\beta} + \frac{1}{\lambda_P}\right)\left[\delta_P^2\left\{\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \mathcal{O}\left(\bar{\beta}R_0 + R_1\right)\right\} + \bar{R}_1\right]
\end{aligned}$$

where $\bar{R}_1 = \delta_P^2\lambda_P^{-2}\Lambda_q^2 + \bar{\delta}_q^2$. This concludes the result. □

We now provide some bounds on MLMC estimates.

Lemma B.2. *Consider a time-homogeneous, ergodic Markov chain $(Z_t)_{t\geq 0}$ with a unique invariant distribution d_Z and a mixing time τ_{mix}. Assume that $\nabla F(x, Z)$ is an estimate of the gradient $\nabla F(x)$. Let $\|\mathbb{E}_{d_Z}[\nabla F(x, Z)] - \nabla F(x)\|^2 \leq \delta^2$ and $\|\nabla F(x, Z_t) - \mathbb{E}_{d_Z}\left[\nabla F(x, Z)\right]\|^2 \leq \sigma^2$ for all $t \geq 0$. If $Q \sim \mathrm{Geom}(1/2)$, then the following MLMC estimator*

$$g_{\mathrm{MLMC}} = g^0 + \begin{cases} 2^Q\left(g^Q - g^{Q-1}\right), & \text{if } 2^Q \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \quad \text{where } g^j = 2^{-j}\sum\nolimits_{t=0}^{2^j-1}\nabla F(x, Z_t) \tag{44}$$

satisfies the inequalities stated below.

(a) $\mathbb{E}[g_{\mathrm{MLMC}}] = \mathbb{E}[g^{\lfloor \log T_{\max}\rfloor}]$

(b) $\mathbb{E}[\|\nabla F(x) - g_{\mathrm{MLMC}}\|^2] \leq \mathcal{O}\left(\sigma^2\tau_{\mathrm{mix}}\log_2 T_{\max} + \delta^2\right)$

(c) $\|\nabla F(x) - \mathbb{E}[g_{\mathrm{MLMC}}]\|^2 \leq \mathcal{O}\left(\sigma^2\tau_{\mathrm{mix}}T_{\max}^{-1} + \delta^2\right)$

Before proceeding to the proof, we state a useful lemma.

Lemma B.3 (Lemma 1, [12])**.** *Consider the same setup as in Lemma B.2. The following holds.*

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{t=0}^{N-1}\nabla F(x, Z_t) - \mathbb{E}_{d_Z}\left[\nabla F(x, Z)\right]\right\|^2\right] \leq \frac{C_1 t_{\mathrm{mix}}}{N}\sigma^2, \tag{45}$$

where N is a constant, $C_1 = 16(1 + \frac{1}{\ln^2 4})$, and the expectation is over the distribution of $\{Z_t\}_{t=0}^{N-1}$ emanating from any arbitrary initial distribution.

Proof of Lemma B.2. The statement (a) can be proven as follows.

$$\begin{aligned}
\mathbb{E}[g_{\mathrm{MLMC}}] &= \mathbb{E}[g^0] + \sum_{j=1}^{\lfloor \log_2 T_{\max}\rfloor}\Pr\{Q = j\}\cdot 2^j\mathbb{E}[g^j - g^{j-1}] \\
&= \mathbb{E}[g^0] + \sum_{j=1}^{\lfloor \log_2 T_{\max}\rfloor}\mathbb{E}[g^j - g^{j-1}] = \mathbb{E}[g^{\lfloor \log_2 T_{\max}\rfloor}],
\end{aligned}$$

For the proof of (b), notice that

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbb{E}_{d_Z}\left[\nabla F(x, Z)\right] - g_{\mathrm{MLMC}}\|^2\right] \leq 2\mathbb{E}\left[\left\|\mathbb{E}_{d_Z}\left[\nabla F(x_t)\right] - g^0\right\|^2\right] + 2\mathbb{E}\left[\left\|g_{\mathrm{MLMC}} - g^0\right\|^2\right] \\
&= 2\mathbb{E}\left[\left\|\mathbb{E}_{d_Z}\left[\nabla F(x_t)\right] - g^0\right\|^2\right] + 2\sum\nolimits_{j=1}^{\lfloor \log_2 T_{\max}\rfloor}\Pr\{Q = j\}\cdot 4^j\mathbb{E}\left[\left\|g^j - g^{j-1}\right\|^2\right] \\
&= 2\mathbb{E}\left[\left\|\mathbb{E}_{d_Z}\left[\nabla F(x_t)\right] - g^0\right\|^2\right] + 2\sum\nolimits_{j=1}^{\lfloor \log_2 T_{\max}\rfloor}2^j\mathbb{E}\left[\left\|g^j - g^{j-1}\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\mathbb{E}_{d_Z}\left[\nabla F(x_t)\right] - g^0\right\|^2\right] \\
&\qquad + 4\sum\nolimits_{j=1}^{\lfloor \log_2 T_{\max}\rfloor}2^j\left(\mathbb{E}\left[\left\|\mathbb{E}_{d_z}\left[\nabla F(x, Z)\right] - g^{j-1}\right\|^2\right] + \mathbb{E}\left[\left\|g^j - \mathbb{E}_{d_Z}\left[\nabla F(x, Z)\right]\right\|^2\right]\right) \\
&\overset{(a)}{\leq} C_1 t_{\mathrm{mix}}\sigma^2\left[2 + 4\sum\nolimits_{j=1}^{\lfloor \log_2 T_{\max}\rfloor}2^j\left(\frac{1}{2^{j-1}} + \frac{1}{2^j}\right)\right] = \mathcal{O}\left(\sigma^2 t_{\mathrm{mix}}\log_2 T_{\max}\right)
\end{aligned}$$

where (a) follows from Lemma B.3. Using this result, we obtain the following.

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla F(x)-g_{\mathrm{MLMC}}\right\|^2\right] &\leq 2\mathbb{E}\left[\left\|\nabla F(x)-\mathbb{E}_{d_z}\left[\nabla F(x,Z)\right]\right\|^2\right]+2\mathbb{E}\left[\left\|\mathbb{E}_{d_z}\left[\nabla F(x,Z)\right]-g_{\mathrm{MLMC}}\right\|^2\right]\\ &\leq \mathcal{O}\left(\sigma^2 t_{\mathrm{mix}}\log_2 T_{\max}+\delta^2\right)\end{aligned}$$

This completes the proof of statement (b). For part (c), we have

$$\begin{aligned}&\left\|\nabla F(x)-\mathbb{E}\left[g_{\mathrm{MLMC}}\right]\right\|^2\\ &\leq 2\left\|\nabla F(x)-\mathbb{E}_{d_Z}\left[\nabla F(x,Z)\right]\right\|^2+2\left\|\mathbb{E}_{d_Z}\left[\nabla F(x,Z)\right]-\mathbb{E}\left[g_{\mathrm{MLMC}}\right]\right\|^2\\ &\leq 2\delta^2+2\left\|\mathbb{E}_{d_Z}\left[\nabla F(x,Z)\right]-\mathbb{E}[g^{\lfloor\log_2 T_{\max}\rfloor}]\right\|^2 \overset{(a)}{\leq} 2\delta^2+\frac{2C_1 t_{\mathrm{mix}}}{T_{\max}}\sigma^2\end{aligned} \tag{46}$$

where (a) follows from Lemma B.3. This concludes the proof of Lemma B.2. □

C Proof of Lemma 4.6

We begin by stating a useful lemma:

Lemma C.1 (Lemma 4, [11])**.** *The performance difference between two policies π_θ, $\pi_{\theta'}$ is bounded as follows where $g \in \{r, c\}$.*

$$J(\theta)-J(\theta')=\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_{\theta'}}(s,a)\right]. \tag{47}$$

Continuing with the proof of Lemma 4.6, we start with the definition of KL divergence. For notational simplicity, we shall use $A^{\pi_{\theta_k}}_{r+\lambda_k c}$ to denote $A^{\pi_{\theta_k}}_r+\lambda_k A^{\pi_{\theta_k}}_c$.

$$\begin{aligned}&\mathbb{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_k}(\cdot|s))-KL(\pi^*(\cdot|s)\|\pi_{\theta_{k+1}}(\cdot|s))]=\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}\left[\log\frac{\pi_{\theta_{k+1}}(a|s)}{\pi_{\theta_k}(a|s)}\right]\\
&\overset{(a)}{\geq}\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot(\theta_{k+1}-\theta_k)]-\frac{G_2}{2}\|\theta_{k+1}-\theta_k\|^2\\
&=\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot\omega_k]-\frac{G_2\alpha^2}{2}\|\omega_k\|^2\\
&=\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot\omega_k^*]+\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot(\omega_k-\omega_k^*)]-\frac{G_2\alpha^2}{2}\|\omega_k\|^2\\
&=\alpha[\mathcal{L}(\pi^*,\lambda_k)-\mathcal{L}(\theta_k,\lambda_k)]+\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot\omega_k^*]\\
&-\alpha[\mathcal{L}(\pi^*,\lambda_k)-\mathcal{L}(\theta_k,\lambda_k)]+\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot(\omega_k-\omega_k^*)]-\frac{G_2\alpha^2}{2}\|\omega_k\|^2\\
&\overset{(b)}{=}\alpha[\mathcal{L}(\pi^*,\lambda_k)-\mathcal{L}(\theta_k,\lambda_k)]+\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}\left[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot\omega_k^*-A^{\pi_{\theta_k}}_{r+\lambda_k c}(s,a)\right]\\
&+\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot(\omega_k-\omega_k^*)]-\frac{G_2\alpha^2}{2}\|\omega_k\|^2\\
&\overset{(c)}{\geq}\alpha[\mathcal{L}(\pi^*,\lambda_k)-\mathcal{L}(\theta_k,\lambda_k)]-\alpha\sqrt{\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}\left[\left(\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot\omega_k^*-A^{\pi_{\theta_k}}_{r+\lambda_k c}(s,a)\right)^2\right]}\\
&+\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot(\omega_k-\omega_k^*)]-\frac{G_2\alpha^2}{2}\|\omega_k\|^2\\
&\overset{(d)}{\geq}\alpha[\mathcal{L}(\pi^*,\lambda_k)-\mathcal{L}(\theta_k,\lambda_k)]-\alpha\sqrt{\epsilon_{\mathrm{bias}}}+\alpha\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\nabla_\theta\log\pi_{\theta_k}(a|s)\cdot(\omega_k-\omega_k^*)]-\frac{G_2\alpha^2}{2}\|\omega_k\|^2\end{aligned}$$

where $\mathcal{L}(\pi^*,\lambda_k)\coloneqq J_r^{\pi^*}+\lambda_k J_c^{\pi^*}$, the relations (a), (b) result from Assumption 4.4(b) and Lemma C.1, respectively. Inequality (c) arises from the convexity of the function $f(x)=x^2$. Lastly, (d) is a

 https://doi.org/10.52202/085713-3099

consequence of Assumption 4.3. By taking expectations on both sides, we derive:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_k}(\cdot|s)) - KL(\pi^*(\cdot|s)\|\pi_{\theta_{k+1}}(\cdot|s))\right]\right] \\
&\geq \alpha[\mathcal{L}(\pi^*,\lambda_k) - \mathbb{E}[\mathcal{L}(\theta_k,\lambda_k)]] - \alpha\sqrt{\epsilon_{\text{bias}}} \\
&\qquad + \alpha\mathbb{E}\left[\mathbb{E}_{s\sim d^{\pi^*}}\mathbb{E}_{a\sim\pi^*(\cdot|s)}[\nabla_\theta \log \pi_{\theta_k}(a|s)\cdot(\mathbb{E}[\omega_k|\theta_k] - \omega_k^*)]\right] - \frac{G_2\alpha^2}{2}\mathbb{E}\left[\|\omega_k\|^2\right] \\
&\geq \alpha[\mathcal{L}(\pi^*,\lambda_k) - \mathbb{E}[\mathcal{L}(\theta_k,\lambda_k)]] - \alpha\sqrt{\epsilon_{\text{bias}}} \\
&\qquad - \alpha\mathbb{E}\left[\mathbb{E}_{(s,a)\sim\nu^{\pi^*}}[\|\nabla_\theta \log \pi_{\theta_k}(a|s)\|\|\mathbb{E}[\omega_k|\theta_k] - \omega_k^*\|]\right] - \frac{G_2\alpha^2}{2}\mathbb{E}\left[\|\omega_k\|^2\right] \\
&\overset{(a)}{\geq} \alpha[\mathcal{L}(\pi^*,\lambda_k) - \mathbb{E}[\mathcal{L}(\theta_k,\lambda_k)]] - \alpha\sqrt{\epsilon_{\text{bias}}} \\
&\qquad - \alpha G_1\mathbb{E}\|(\mathbb{E}[\omega_k|\theta_k] - \omega_k^*)\| - \frac{G_2\alpha^2}{2}\mathbb{E}\left[\|\omega_k\|^2\right]
\end{aligned} \tag{48}$$

where (a) follows from Assumption 4.4(a). Rearranging the terms, we get,

$$\begin{aligned}
\mathcal{L}(\pi^*,\lambda_k) - \mathbb{E}[\mathcal{L}(\theta_k,\lambda_k)] &\leq \sqrt{\epsilon_{\text{bias}}} + G_1\mathbb{E}\|(\mathbb{E}[\omega_k|\theta_k] - \omega_k^*)\| + \frac{G_2\alpha}{2}\mathbb{E}\|\omega_k\|^2 \\
&\quad + \frac{1}{\alpha}\mathbb{E}\left[\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_k}(\cdot|s)) - KL(\pi^*(\cdot|s)\|\pi_{\theta_{k+1}}(\cdot|s))\right]\right]
\end{aligned} \tag{49}$$

Adding the above inequality from $k = 0$ to $K-1$, using the non-negativity of KL divergence and dividing the resulting expression by K, we obtain the final result.

D Proof of Theorem 4.8

Recall that $\mathbf{A}_g(\theta) = \mathbb{E}_\theta[\mathbf{A}_g(\theta;z)]$, $\mathbf{b}_g(\theta) = \mathbb{E}_\theta[\mathbf{b}_g(\theta;z)]$, and $\xi_g^*(\theta) = (\mathbf{A}_g(\theta))^{-1}\mathbf{b}_g(\theta)$ where $\mathbb{E}_\theta$ denotes the expectation over the distribution of $z = (s,a,s')$ where $(s,a)\sim\nu^{\pi_\theta}$, $s'\sim P(s,a)$ (refer to section 4).

Lemma D.1. *For large c_γ, Assumption 4.2 implies that $\mathbf{A}_g(\theta_k) - (\lambda/2)I$ is positive semi-definite for both $g\in\{r,c\}$ and for all k where I is an identity matrix, i.e., $\xi^\top\mathbf{A}_g(\theta_k)\xi \geq \lambda/2\cdot\|\xi\|^2$, for all ξ.*

Proof of Lemma D.1. Note that for any $\xi = [\eta,\zeta]$, we have

$$\begin{aligned}
\xi^\top\mathbf{A}_g(\theta_k)\xi &= c_\gamma\eta^2 + \eta\zeta^\top\mathbb{E}_{\theta_k}\left[\phi_g(s)\right] + \zeta^\top\mathbb{E}_{\theta_k}\left[\phi_g(s)\left[\phi_g(s) - \phi_g(s')\right]^\top\right]\zeta \\
&\overset{(a)}{\geq} c_\gamma\eta^2 - |\eta|\,\|\zeta\| + \lambda\|\zeta\|^2 \\
&\geq \|\xi\|^2\left\{\min_{u\in[0,1]} c_\gamma u - \sqrt{u(1-u)} + \lambda(1-u)\right\} \overset{(b)}{\geq} (\lambda/2)\|\xi\|^2
\end{aligned} \tag{50}$$

where (a) is a consequence of Assumption 4.2 and the fact that $\|\phi_g(s)\|\leq 1, \forall s\in\mathcal{S}$. Finally, (b) holds when $c_\gamma \geq \lambda + \sqrt{\frac{1}{\lambda^2}-1}$. This completes the proof. □

Let $\mathbf{A}_{g,kh}^{\text{MLMC}}$ and $\mathbf{b}_{g,kh}^{\text{MLMC}}$ be the MLMC estimators of $\{\mathbf{A}_g(\theta_k;z_{kh}^t)\}_{t=0}^{l_{kh}-1}$ and $\{\mathbf{b}_g(\theta_k;z_{kh}^t)\}_{t=0}^{l_{kh}-1}$ respectively (see (19)) i.e.

$$\mathbf{A}_{g,kh}^{\text{MLMC}} = \mathbf{A}_{g,kh}^0 + \begin{cases} 2^{P_h^k}(\mathbf{A}_{g,kh}^{Q_h^k} - \mathbf{A}_{g,kh}^{Q_h^k-1}), & \text{if } 2^{Q_h^k}\leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{51}$$

where $\mathbf{A}_{g,kh}^j = \frac{1}{2^j}\sum_{t=0}^{2^j-1}\mathbf{A}_g(\theta_k;z_{kh}^t)$ and

$$\mathbf{b}_{g,kh}^{\text{MLMC}} = \mathbf{b}_{g,kh}^0 + \begin{cases} 2^{P_h^k}(\mathbf{b}_{g,kh}^{Q_h^k} - \mathbf{b}_{g,kh}^{Q_h^k-1}), & \text{if } 2^{Q_h^k}\leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{52}$$

where $\mathbf{b}_{g,kh}^j = \frac{1}{2^j}\sum_{t=0}^{2^j-1}\mathbf{b}_g(\theta_k;z_{kh}^t)$.

We can bound the bias and variance of $\mathbf{A}_{g,kh}^{\text{MLMC}}$ and $\mathbf{b}_{g,kh}^{\text{MLMC}}$ as follows.

Lemma D.2. *Consider Algorithm 1 with a policy parameter θ_k and assume assumptions 2.2-4.5 hold. The MLMC estimators $\mathbf{A}_{g,kh}^{\mathrm{MLMC}}$ and $\mathbf{b}_{g,kh}^{\mathrm{MLMC}}$ obey the following bounds.*

(a) $\|\mathbb{E}_{k,h}[\mathbf{A}_{g,kh}^{\mathrm{MLMC}}] - \mathbf{A}_g(\theta_k)\|^2 \leq \mathcal{O}\left(c_\gamma^2 \tau_{\mathrm{mix}} T_{\max}^{-1}\right)$

(b) $\|\mathbb{E}_{k,h}[\mathbf{b}_{g,kh}^{\mathrm{MLMC}}] - \mathbf{b}_g(\theta_k)\|^2 \leq \mathcal{O}\left(c_\gamma^2 \tau_{\mathrm{mix}} T_{\max}^{-1}\right)$

(c) $\mathbb{E}_{k,h}\left[\|\mathbf{A}_{g,kh}^{\mathrm{MLMC}} - \mathbf{A}_g(\theta_k)\|^2\right] \leq \mathcal{O}\left(c_\gamma^2 \tau_{\mathrm{mix}} \log T_{\max}\right)$

(d) $\mathbb{E}_{k,h}\left[\|\mathbf{b}_{g,kh}^{\mathrm{MLMC}} - \mathbf{b}_g(\theta_k)\|^2\right] \leq \mathcal{O}\left(c_\gamma^2 \tau_{\mathrm{mix}} \log T_{\max}\right)$

where $h \in \{0, \cdots, H-1\}$ and $\mathbb{E}_{k,h}$ defines the conditional expectation given the history up to the inner loop step h of the critic within the kth outer loop instant.

Proof. Recall from (19) the definitions of $\mathbf{A}_g(\theta; z)$ and $\mathbf{b}(\theta_k; z)$ for any $z = (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. We have the following inequalities.

$$\|\mathbf{A}_g(\theta_k; z)\| \leq |c_\gamma| + \|\phi_g(s)\| + \left\|\phi_g(s)(\phi_g(s) - \phi_g(s'))^\top\right\| \overset{(a)}{\leq} c_\gamma + 3 = \mathcal{O}(c_\gamma), \tag{53}$$

$$\|\mathbf{b}_g(\theta_k; z)\| \leq |c_\gamma g(s, a)| + \|g(s, a)\phi_g(s)\| \overset{(b)}{\leq} c_\gamma + 1 = \mathcal{O}(c_\gamma) \tag{54}$$

where (a), (b) hold since $|g(s, a)| \leq 1$ and $\|\phi_g(s)\| \leq 1$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, $\forall g \in \{r, c\}$. Hence, for any $z_{kh}^j \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have the inequalities stated below.

$$\left\|\mathbf{A}_g(\theta_k; z_{kh}^j) - \mathbf{A}_g(\theta_k)\right\|^2 \leq \mathcal{O}(c_\gamma^2), \text{ and } \left\|\mathbf{b}_g(\theta_k; z_{kh}^j) - \mathbf{b}_g(\theta_k)\right\|^2 \leq \mathcal{O}(c_\gamma^2)$$

Combining the above results with Lemma B.2, and using the definitions that $\mathbf{A}_g(\theta) = \mathbb{E}_\theta[\mathbf{A}_g(\theta; z)]$, $\mathbf{b}_g(\theta) = \mathbb{E}_\theta[\mathbf{b}_g(\theta; z)]$, we establish the result. □

Combining Lemma D.1 with (53), (54), we obtain the following for any θ_k with $c_\gamma \geq \lambda + \sqrt{\frac{1}{\lambda^2} - 1}$.

$$\frac{\lambda}{2} \leq \|\mathbf{A}_g(\theta_k)\| \leq \mathcal{O}(c_\gamma), \text{ and } \|\mathbf{b}_g(\theta_k)\| \leq \mathcal{O}(c_\gamma) \tag{55}$$

Combining the above results with Lemma D.1 along with Theorem B.1, we then obtain the following inequalities given that $T_{\max} \geq \frac{8c_\gamma^2 \tau_{\mathrm{mix}}}{\lambda}$ and $\gamma_\xi \leq \frac{\lambda}{24 c_\gamma^2 \tau_{\mathrm{mix}} \log T_{\max}}$.

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|\xi_{g,H}^k - \xi_g^*(\theta_k)\right\|^2\right] \leq \exp\left(-H\gamma_\xi \lambda\right)\left\|\xi_0 - \xi_g^*(\theta_k)\right\|^2 + \tilde{\mathcal{O}}\left(\gamma_\xi R_0 + R_1\right), \\
&\mathbb{E}_k\left[\left\|\mathbb{E}_k[\xi_{g,H}^k] - \xi_g^*(\theta_k)\right\|^2\right] \leq \exp\left(-H\gamma_\xi \lambda\right)\left\|\xi_0 - \xi_g^*(\theta_k)\right\|^2 \\
&\qquad + \frac{\lambda\gamma_\xi + 1}{\lambda^2}\left[\mathcal{O}(T_{\max}^{-1} c_\gamma^2 \tau_{\mathrm{mix}})\left\{\left\|\xi_0 - \xi_g^*(\theta_k)\right\|^2 + \mathcal{O}\left(\gamma_\xi R_0 + R_1\right)\right\} + \bar{R}_1\right]
\end{aligned}$$

where the terms $R_0, R_1, \bar{R}_1$ are defined as follows.

$$\begin{aligned}
R_0 &= \tilde{\mathcal{O}}\left(\lambda^{-3} c_\gamma^4 \tau_{\mathrm{mix}} + \lambda^{-1} c_\gamma^2 \tau_{\mathrm{mix}}\right) = \tilde{\mathcal{O}}\left(\lambda^{-3} c_\gamma^4 \tau_{\mathrm{mix}}\right), \\
R_1 &= \mathcal{O}\left(T_{\max}^{-1} \lambda^{-4} c_\gamma^4 \tau_{\mathrm{mix}} + T_{\max}^{-1} \lambda^{-2} c_\gamma^2 \tau_{\mathrm{mix}}\right) = \mathcal{O}\left(T_{\max}^{-1} \lambda^{-4} c_\gamma^4 \tau_{\mathrm{mix}}\right) \\
\bar{R}_1 &= \mathcal{O}\left(T_{\max}^{-1} \lambda^{-2} c_\gamma^4 t_{\mathrm{mix}} + T_{\max}^{-1} c_\gamma^2 t_{\mathrm{mix}}\right) = \mathcal{O}\left(T_{\max}^{-1} \lambda^{-2} c_\gamma^4 \tau_{\mathrm{mix}}\right)
\end{aligned}$$

If we further set $\gamma_\xi = \frac{2 \log T}{\lambda H}$ while ensuring $\frac{2 \log T}{\lambda H} \leq \frac{\lambda}{24 c_\gamma^2 \tau_{\mathrm{mix}} \log T_{\max}}$, we have the following results.

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|\xi_{g,H}^k - \xi_g^*(\theta_k)\right\|^2\right] \leq \frac{1}{T^2}\left\|\xi_0 - \xi_g^*(\theta_k)\right\|^2 + \tilde{\mathcal{O}}\left(\lambda^{-4} c_\gamma^4 \tau_{\mathrm{mix}}\left(\frac{1}{H} + \frac{1}{T_{\max}}\right)\right), \\
&\left\|\mathbb{E}_k[\xi_{g,H}^k] - \xi_g^*(\theta_k)\right\|^2 \leq \tilde{\mathcal{O}}\left(\left(\frac{1}{T^2} + \frac{c_\gamma^2 \tau_{\mathrm{mix}}}{\lambda^2 T_{\max}}\right)\left\|\xi_0 - \xi_g^*(\theta_k)\right\|^2\right) + \tilde{\mathcal{O}}\left(\frac{c_\gamma^6 \tau_{\mathrm{mix}}^2}{\lambda^6 T_{\max}}\right)
\end{aligned}$$

https://doi.org/10.52202/085713-3099

We used the fact that $H^{-1} + T_{\max}^{-1} = \mathcal{O}(1)$ in the last inequality. Utilizing $\xi_{g,H}^k = \xi_g^k$, $\xi_0 = 0$, and $\|\xi_g^*(\theta_k)\| = \|[\mathbf{A}_g(\theta_k)]^{-1}\mathbf{b}_g(\theta_k)\| = \mathcal{O}(\lambda^{-2}c_\gamma^2)$ (via (55)), we conclude:

$$\mathbb{E}_k\left[\left\|\xi_g^k - \xi_g^*(\theta_k)\right\|^2\right] \leq \frac{c_\gamma^2}{\lambda^2 T^2} + \tilde{\mathcal{O}}\left(\lambda^{-4}c_\gamma^4\tau_{\text{mix}}\left(\frac{1}{H} + \frac{1}{T_{\max}}\right)\right), \tag{56}$$

$$\left\|\mathbb{E}_k[\xi_g^k] - \xi_g^*(\theta_k)\right\|^2 \leq \tilde{\mathcal{O}}\left(\left(\frac{1}{T^2} + \frac{c_\gamma^2\tau_{\text{mix}}}{\lambda^2 T_{\max}}\right)\frac{c_\gamma^2}{\lambda^2} + \frac{c_\gamma^6\tau_{\text{mix}}^2}{\lambda^6 T_{\max}}\right) = \tilde{\mathcal{O}}\left(\frac{c_\gamma^2}{\lambda^2 T^2} + \frac{c_\gamma^6\tau_{\text{mix}}^2}{\lambda^6 T_{\max}}\right) \tag{57}$$

This completes the proof.

E Proof of Theorem 4.7

To prove Theorem 4.7, we follow a proof structure similar to that of Theorem 4.8. Let F_{kh}^{MLMC} and $\nabla_\theta^{\text{MLMC}} J_{g,kh}$ denote the MLMC estimators defined as follows.

$$F_{kh}^{\text{MLMC}} = F_{kh}^0 + \begin{cases} 2^{Q_h^k}(F_{kh}^{Q_h^k} - F_{kh}^{Q_h^k - 1}), & \text{if } 2^{Q_h^k} \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{58}$$

with $F_{kh}^j = \frac{1}{2^j}\sum_{t=0}^{2^j-1} F(\theta_k; z_{kh}^t)$ (see (22)) and

$$\nabla_\theta^{\text{MLMC}} J_{g,kh} = \nabla_\theta^0 J_{g,kh}^0 + \begin{cases} 2^{Q_h^k}(\nabla_\theta^{Q_h^k} J_{g,kh} - \nabla_\theta^{Q_h^k - 1} J_{g,kh}), & \text{if } 2^{Q_h^k} \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{59}$$

where $\nabla_\theta^j J_{g,kh} = 2^{-j}\sum_{t=0}^{2^j-1} \hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z_{kh}^t)$ (see (22)).

Using the above inequalities, we can bound the bias and variance of the MLMC estimators as follows.

Lemma E.1. *Consider Algorithm 1 with given policy parameter θ_k. Under the assumptions 2.2-4.5, the following statements hold.*

(a) $\|\mathbb{E}_{k,h}[F_{kh}^{\text{MLMC}}] - F(\theta_k)\|^2 \leq \mathcal{O}\left(G_1^4\tau_{\text{mix}}T_{\max}^{-1}\right)$

(b) $\|\mathbb{E}_{k,h}[\nabla_\theta^{\text{MLMC}} J_{g,kh}] - \nabla_\theta J_g(\theta_k)\|^2 \leq \mathcal{O}\left(\sigma_{k,g}^2\tau_{\text{mix}}T_{\max}^{-1} + \delta_{k,g}^2\right)$

(c) $\mathbb{E}_{k,h}\left[\|F_{kh}^{\text{MLMC}} - F(\theta_k)\|^2\right] \leq \mathcal{O}\left(G_1^4\tau_{\text{mix}}\log T_{\max}\right)$

(d) $\mathbb{E}_{k,h}\left[\|\nabla_\theta^{\text{MLMC}} J_{g,kh} - \nabla_\theta J_g(\theta_k)\|^2\right] \leq \mathcal{O}\left(\sigma_{k,g}^2\tau_{\text{mix}}\log T_{\max} + \delta_{k,g}^2\right)$.

(e) $\|\mathbb{E}_k[\nabla_\theta^{\text{MLMC}} J_{g,kh}] - \nabla_\theta J_g(\theta_k)\|^2 \leq \mathcal{O}\left(\bar{\sigma}_{k,g}^2\tau_{\text{mix}}T_{\max}^{-1} + \bar{\delta}_{k,g}^2\right)$

where $\mathbb{E}_{k,h}$ defines the conditional expectation given the history up to the inner loop step h of the actor within the kth outer loop instant, while $\mathbb{E}_k$ is the conditional expectation given the history up to the kth outer loop instant. Moreover,

$$\begin{aligned}
\sigma_{k,g}^2 &= \mathcal{O}\left(G_1^2\left\|\xi_g^k\right\|^2\right), \\
\bar{\sigma}_{k,g}^2 &= \mathcal{O}\left(G_1^2\mathbb{E}_k\left[\left\|\xi_g^k\right\|^2\right]\right), \\
\delta_{k,g}^2 &= \mathcal{O}\left(G_1^2\left\|\xi_g^k - \xi_g^*(\theta_k)\right\|^2 + G_1^2\epsilon_{\text{app}}\right), \\
\bar{\delta}_{k,g}^2 &= \mathcal{O}\left(G_1^2\left\|\mathbb{E}_k[\xi_g^k] - \xi_g^*(\theta_k)\right\|^2 + G_1^2\epsilon_{\text{app}}\right)
\end{aligned}$$

for $h \in \{0, 1, \cdots, H-1\}$.

Proof of Lemma E.1. Fix an outer loop instant k and an inner loop instant $h \in \{0, \cdots, H-1\}$ of the actor subroutine. Recall the definition of $F(\theta_k; \cdot)$ from (22). The following inequalities hold for any θ_k and $z_{kh}^j \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

$$\mathbb{E}_{\theta_k}\left[F(\theta_k; z)\right] \overset{(a)}{=} F(\theta_k), \text{ and } \left\|F(\theta_k; z_{kh}^j) - \mathbb{E}_{\theta_k}\left[F(\theta_k; z)\right]\right\|^2 \overset{(b)}{\leq} 2G_1^4$$

where $\mathbb{E}_{\theta_k}$ denotes the expectation over the distribution $z = (s, a, s')$ where $(s, a) \sim \nu^{\pi_{\theta_k}}$, $s' \sim P(\cdot|s, a)$. The equality (a) follows from the definition of the Fisher matrix, and (b) is a consequence of assumption 4.4. Statements (a) and (c), therefore, directly follow from Lemma B.2.

To prove the other statements, recall the definition of $\hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; \cdot)$ from (22). Note the following relations for arbitrary $\theta_k, \xi_g^k = [\eta_g^k, \zeta_g^k]$.

$$
\begin{aligned}
&\mathbb{E}_{\theta_k}\left[\hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z)\right] - \nabla_\theta J_g(\theta_k) \\
&\quad= \mathbb{E}_{\theta_k}\left[\left\{g(s,a) - \eta_g^k + \langle \phi_g(s') - \phi_g(s), \zeta_g^k\rangle\right\}\nabla_\theta \log \pi_{\theta_k}(a|s)\right] - \nabla_\theta J_g(\theta_k) \\
&\quad\overset{(a)}{=} \underbrace{\mathbb{E}_{\theta_k}\left[\left\{\eta_g^*(\theta_k) - \eta_g^k + \langle \phi_g(s') - \phi_g(s), \zeta_g^k - \zeta_g^*(\theta_k)\rangle\right\}\nabla_\theta \log \pi_{\theta_k}(a|s)\right]}_{T_0} \\
&\quad+ \underbrace{\mathbb{E}_{\theta_k}\left[\left\{V_g^{\pi_{\theta_k}}(s) - \langle \phi_g(s), \zeta_g^*(\theta_k)\rangle + \langle \phi_g(s'), \zeta_g^*(\theta_k)\rangle - V_g^{\pi_{\theta_k}}(s')\right\}\nabla_\theta \log \pi_{\theta_k}(a|s)\right]}_{T_1} \\
&\quad+ \underbrace{\mathbb{E}_{\theta_k}\left[\left\{V_g^{\pi_{\theta_k}}(s') - \eta_g^*(\theta_k) + g(s,a) - V_g^{\pi_{\theta_k}}(s)\right\}\nabla_\theta \log \pi_{\theta_k}(a|s)\right] - \nabla_\theta J_g(\theta_k)}_{T_2}
\end{aligned}
$$

We will use the notation that $\xi_g^*(\theta_k) = [\eta_g^*(\theta_k), \zeta_g^*(\theta_k)]^\top$. Observe that

$$
\|T_0\|^2 \overset{(a)}{=} \mathcal{O}\left(G_1^2 \left\|\xi_g^k - \xi_g^*(\theta_k)\right\|^2\right),\ \|T_1\|^2 \overset{(b)}{=} \mathcal{O}\left(G_1^2 \epsilon_{\text{app}}\right),\ \text{and } T_2 \overset{(c)}{=} 0 \tag{60}
$$

where (a) follows from Assumption 4.4 and the boundedness of the feature map, ϕ while (b) is a consequence of Assumption 4.4 and 4.1. Finally, (c) is an application of Bellman's equation and the fact that $\eta_g^*(\theta_k) = J_g^{\pi_{\theta_k}}$, which can be easily verified. We get,

$$
\left\|\mathbb{E}_{\theta_k}\left[\hat{\nabla}_\theta J_g(\theta_k; z)\right] - \nabla_\theta J_g(\theta_k)\right\|^2 \leq \delta_{k,g}^2 = \mathcal{O}\left(G_1^2 \left\|\xi_g^k - \xi_g^*(\theta_k)\right\|^2 + G_1^2 \epsilon_{\text{app}}\right) \tag{61}
$$

Moreover, observe that, for arbitrary $z_{kh}^j \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$

$$
\left\|\hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z_{kh}^j) - \mathbb{E}_{\theta_k}\left[\hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z)\right]\right\|^2 \overset{(a)}{\leq} \sigma_{k,g}^2 = \mathcal{O}\left(G_1^2 \left\|\xi_g^k\right\|^2\right) \tag{62}
$$

where (a) results from Assumption 4.4 and the boundedness of ϕ. We can, thus, conclude statements (c) and (d) by applying (61) and (62) in Lemma B.2. To prove the statement (e), note that

$$
\begin{aligned}
&\mathbb{E}_{\theta_k}\left[\mathbb{E}_k\left[\hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z)\right]\right] - \nabla_\theta J_g(\theta_k) \\
&\quad= \mathbb{E}_{\theta_k}\left[\left\{g(s,a) - \mathbb{E}_k[\eta_g^k] + \langle \phi_g(s') - \phi_g(s), \mathbb{E}_k[\zeta_g^k]\rangle\right\}\nabla_\theta \log \pi_{\theta_k}(a|s)\right] - \nabla_\theta J_g(\theta_k) \\
&\quad\overset{(a)}{=} \underbrace{\mathbb{E}_{\theta_k}\left[\left\{\eta_g^*(\theta_k) - \mathbb{E}_k[\eta_g^k] + \langle \phi_g(s') - \phi_g(s), \mathbb{E}_k[\zeta_g^k] - \zeta_g^*(\theta_k)\rangle\right\}\nabla_\theta \log \pi_{\theta_k}(a|s)\right]}_{T_0} + \\
&\quad+ \underbrace{\mathbb{E}_{\theta_k}\left[\left\{V_g^{\pi_{\theta_k}}(s) - \langle \phi_g(s), \zeta_g^*(\theta_k)\rangle + \langle \phi_g(s'), \zeta_g^*(\theta_k)\rangle - V_g^{\pi_{\theta_k}}(s')\right\}\nabla_\theta \log \pi_{\theta_k}(a|s)\right]}_{T_1} \\
&\quad+ \underbrace{\mathbb{E}_{\theta_k}\left[\left\{V_g^{\pi_{\theta_k}}(s') - \eta_g^*(\theta_k) + g(s,a) - V_g^{\pi_{\theta_k}}(s)\right\}\nabla_\theta \log \pi_{\theta_k}(a|s)\right] - \nabla_\theta J_g(\theta_k)}_{T_2}
\end{aligned}
$$

Observe the following bounds.

$$
\|T_0\|^2 \overset{(a)}{=} \mathcal{O}\left(G_1^2 \left\|\mathbb{E}_k[\xi_g^k] - \xi_g^*(\theta_k)\right\|^2\right),\ \|T_1\|^2 \overset{(b)}{=} \mathcal{O}\left(G_1^2 \epsilon_{\text{app}}\right),\ \text{and } T_2 \overset{(c)}{=} 0 \tag{63}
$$

https://doi.org/10.52202/085713-3099

where (a) follows from Assumption 4.4 and the boundedness of the feature map, ϕ while (b) is a consequence of Assumption 4.4 and 4.1. Finally, (c) is an application of Bellman's equation. We get,

$$\left\|\mathbb{E}_{\theta_k}\left[\mathbb{E}_k\left[\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)\right]\right]-\nabla_\theta J_g(\theta_k)\right\|^2 \leq \bar{\delta}_{k,g}^2 = \mathcal{O}\left(G_1^2\left\|\mathbb{E}_k[\xi_g^k]-\xi_g^*(\theta_k)\right\|^2+G_1^2\epsilon_{\text{app}}\right) \tag{64}$$

Using the above bound, we deduce the following.

$$\begin{aligned}
&\left\|\mathbb{E}_k\left[\nabla_\theta J_{g,kh}^{\text{MLMC}}\right]-\nabla_\theta J_g(\theta_k)\right\|^2 \\
&\leq 2\left\|\mathbb{E}_k\left[\nabla_\theta^{\text{MLMC}} J_{g,kh}\right]-\mathbb{E}_{\theta_k}\left[\mathbb{E}_k\left[\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)\right]\right]\right\|^2 \\
&\qquad +2\left\|\mathbb{E}_{\theta_k}\left[\mathbb{E}_k\left[\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)\right]\right]-\nabla_\theta J_g(\theta_k)\right\|^2 \\
&\overset{(a)}{\leq} 2\mathbb{E}_k\left\|\mathbb{E}_{k,h}\left[\nabla_\theta^{\text{MLMC}} J_{g,kh}\right]-\mathbb{E}_{\theta_k}\left[\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)\right]\right\|^2+\mathcal{O}\left(\bar{\delta}_{k,g}^2\right) \\
&\overset{(b)}{\leq} \mathcal{O}\left(\tau_{\text{mix}}T_{\max}^{-1}\bar{\sigma}_{k,g}^2+\bar{\delta}_{k,g}^2\right)
\end{aligned}$$

where (a) follows from (64), (b) follows from Lemma B.2(a), B.3, and the definition of $\bar{\sigma}_{k,g}^2$. □

Combining Lemma G.1 with Assumptions 4.4 and 4.5, we obtain the following for a given policy parameter θ and $\forall g \in \{r,c\}$,

$$\mu \leq \|F(\theta)\| \leq G_1^2, \text{ and } \|\nabla_\theta J_g(\theta)\| \leq \mathcal{O}\left(G_1\tau_{\text{mix}}\right) \tag{65}$$

Combining the above results with Lemma E.1 and invoking Theorem B.1, we then obtain given that $T_{\max} \geq \frac{8G_1^4\tau_{\text{mix}}}{\mu}$ and $\gamma_\omega \leq \frac{\mu}{4(6G_1^4\tau_{\text{mix}}\log T_{\max}+2G_1^2\tau_{\text{mix}}^2\log T_{\max})}$.

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\omega_g^k-\omega_{g,k}^*\right\|^2\right] &\leq \exp\left(-H\gamma_\omega\mu\right)\left\|\omega_0-\omega_{g,k}^*\right\|^2+\tilde{\mathcal{O}}\left(\gamma_\omega R_0+R_1\right), \\
\left\|\mathbb{E}_k[\omega_g^k]-\omega_{g,k}^*\right\|^2 &\leq \exp\left(-H\gamma_\omega\mu\right)\left\|\omega_0-\omega_{g,k}^*\right\|^2 \\
&\quad+\frac{\mu\gamma_\omega+1}{\mu^2}\left[\mathcal{O}(T_{\max}^{-1}G_1^4\tau_{\text{mix}})\left\{\left\|\omega_0-\omega_{g,k}^*\right\|^2+\mathcal{O}\left(\gamma_\omega R_0+R_1\right)\right\}+\bar{R}_1\right]
\end{aligned}$$

where the terms $R_0, R_1, \bar{R}_1$ are defined as follows.

$$\begin{aligned}
R_0 &= \tilde{\mathcal{O}}\left(\mu^{-3}G_1^6\tau_{\text{mix}}^3+\mu^{-1}G_1^2\tau_{\text{mix}}\mathbb{E}_k\left[\left\|\xi_g^k\right\|^2\right]+\mu^{-1}G_1^2\mathbb{E}_k\left[\left\|\xi_g^k-\xi_g^*(\theta_k)\right\|^2\right]+\mu^{-1}G_1^2\epsilon_{\text{app}}\right), \\
R_1 &= \mu^{-2}G_1^2\,\mathcal{O}\left(T_{\max}^{-1}\mu^{-2}G_1^4\tau_{\text{mix}}^3+T_{\max}^{-1}\tau_{\text{mix}}\mathbb{E}_k\left[\left\|\xi_g^k\right\|^2\right]+\mathbb{E}_k\left[\left\|\xi_g^k-\xi_g^*(\theta_k)\right\|^2\right]+\epsilon_{\text{app}}\right), \\
\bar{R}_1 &= \mathcal{O}\left(T_{\max}^{-1}\mu^{-2}G_1^6\tau_{\text{mix}}^3+T_{\max}^{-1}G_1^2\tau_{\text{mix}}\mathbb{E}_k\left[\left\|\xi_g^k\right\|^2\right]+G_1^2\left\|\mathbb{E}_k[\xi_g^k]-\xi_g^*(\theta_k)\right\|^2+G_1^2\epsilon_{\text{app}}\right)
\end{aligned}$$

Moreover, note that

$$\mathbb{E}_k\left[\left\|\xi_g^k\right\|^2\right] \leq 2\mathbb{E}_k\left[\left\|\xi_g^k-\xi_g^*(\theta_k)\right\|^2\right]+2\mathbb{E}_k\left[\left\|\xi_g^*(\theta_k)\right\|^2\right] \overset{(a)}{\leq} \mathcal{O}\left(\mathbb{E}_k\left[\left\|\xi_g^k-\xi_g^*(\theta_k)\right\|^2\right]+\lambda^{-2}c_\gamma^2\right)$$

where (a) uses (55) for sufficiently large c_γ and the definition that $\xi_g^*(\theta_k)=[\mathbf{A}_g(\theta_k)]^{-1}\mathbf{b}_g(\theta_k)$. If we set $\gamma_\omega = \frac{2\log T}{\mu H} \leq \frac{\mu}{4(6G_1^4\tau_{\text{mix}}\log T_{\max}+2G_1^2\tau_{\text{mix}}^2\log T_{\max})}$, we get

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\omega_g^k-\omega_{g,k}^*\right\|^2\right] &\leq \tilde{\mathcal{O}}\left(\left(\frac{1}{H}+\frac{1}{T_{\max}}\right)\left\{\mu^{-4}G_1^6\tau_{\text{mix}}^3+\mu^{-2}\lambda^{-2}G_1^2c_\gamma^2\tau_{\text{mix}}\right\}\right) \\
&\quad+\frac{1}{T^2}\left\|\omega_0-\omega_{g,k}^*\right\|^2+\tilde{\mathcal{O}}\left(\mu^{-2}G_1^2\left\{\mathbb{E}_k\left[\left\|\xi_g^k-\xi_g^*(\theta_k)\right\|^2\right]+\epsilon_{\text{app}}\right\}\right), \\
\left\|\mathbb{E}_k[\omega_g^k]-\omega_{g,k}^*\right\|^2 &\leq \mu^{-2}G_1^2\tilde{\mathcal{O}}\left(\left\|\mathbb{E}_k[\xi_g^k]-\xi_g^*(\theta_k)\right\|^2+\epsilon_{\text{app}}\right)+\left(\frac{1}{T^2}+\frac{G_1^4\tau_{\text{mix}}}{\mu^2T_{\max}}\right)\left\|\omega_0-\omega_{g,k}^*\right\|^2 \\
&\quad+\tilde{\mathcal{O}}\left(\frac{G_1^4\tau_{\text{mix}}}{\mu^2T_{\max}}\left\{\mu^{-2}G_1^2\tau_{\text{mix}}\mathbb{E}_k\left[\left\|\xi_g^k-\xi_g^*(\theta_k)\right\|^2\right]+\mu^{-4}G_1^6\tau_{\text{mix}}^3+\mu^{-2}\lambda^{-2}G_1^2c_\gamma^2\tau_{\text{mix}}\right\}\right)
\end{aligned}$$

Substituting $\omega_0 = 0$, $\|\omega_{g,k}^*\| = \|[F(\theta_k)]^{-1}\nabla_\theta J_g(\theta_k)\| = \mathcal{O}(\mu^{-1}G_1\tau_{\text{mix}})$ (follows from Lemma G.1, and assumptions 4.4and 4.5), we can simplify the above bounds as follows.

$$\mathbb{E}_k\left[\left\|\omega_g^k - \omega_{g,k}^*\right\|^2\right] \leq \tilde{\mathcal{O}}\left(\left(\frac{1}{H} + \frac{1}{T_{\max}}\right)\mu^{-4}\lambda^{-2}G_1^6 c_\gamma^2 \tau_{\text{mix}}^3\right) + \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2} + \tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\left\{\mathbb{E}_k\left[\left\|\xi_g^k - \xi_g^*(\theta_k)\right\|^2\right] + \epsilon_{\text{app}}\right\}\right), \tag{66}$$

$$\left\|\mathbb{E}_k[\omega_g^k] - \omega_{g,k}^*\right\|^2 \leq \frac{G_1^2}{\mu^2}\tilde{\mathcal{O}}\left(\left\|\mathbb{E}_k[\xi_g^k] - \xi_g^*(\theta_k)\right\|^2 + \epsilon_{\text{app}}\right) + \left(\frac{1}{T^2} + \frac{G_1^4\tau_{\text{mix}}}{\mu^2 T_{\max}}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2} + \tilde{\mathcal{O}}\left(\frac{G_1^6\tau_{\text{mix}}^2}{\mu^4 T_{\max}}\left\{\mathbb{E}_k\left[\left\|\xi_g^k - \xi_g^*(\theta_k)\right\|^2\right] + \mu^{-2}\lambda^{-2}G_1^4 c_\gamma^2\tau_{\text{mix}}^2\right\}\right) \tag{67}$$

This completes the proof of Theorem 4.7.

F Proof of Main Theorems (Theorem 4.9 and Theorem 4.10)

F.0.1 Rate of Convergence of the Objective

Combining (56), (57) and (66) and (67), we obtain the following results.

$$\mathbb{E}_k\left[\left\|\omega_g^k - \omega_{g,k}^*\right\|^2\right] \leq \tilde{\mathcal{O}}\left(\frac{G_1^6 c_\gamma^2\tau_{\text{mix}}^3}{\mu^4\lambda^2} + \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2}\right) + \tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\left\{\frac{c_\gamma^2}{\lambda^2 T^2} + \frac{c_\gamma^4\tau_{\text{mix}}}{\lambda^4} + \epsilon_{\text{app}}\right\}\right)$$
$$= \tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}} + \frac{G_1^6 c_\gamma^4\tau_{\text{mix}}^3}{\mu^4\lambda^4}\right), \tag{68}$$

$$\left\|\mathbb{E}_k[\omega_g^k] - \omega_{g,k}^*\right\|^2 \leq \frac{G_1^2}{\mu^2}\tilde{\mathcal{O}}\left(\frac{c_\gamma^2}{\lambda^2 T^2} + \frac{c_\gamma^6\tau_{\text{mix}}^2}{\lambda^6 T_{\max}} + \epsilon_{\text{app}}\right) + \left(\frac{1}{T^2} + \frac{G_1^4\tau_{\text{mix}}}{\mu^2 T_{\max}}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2} + \tilde{\mathcal{O}}\left(\frac{G_1^6\tau_{\text{mix}}^2}{\mu^4 T_{\max}}\left\{\frac{c_\gamma^2}{\lambda^2 T^2} + \frac{c_\gamma^4\tau_{\text{mix}}}{\lambda^4} + \frac{G_1^4 c_\gamma^2\tau_{\text{mix}}^2}{\mu^2\lambda^2}\right\}\right)$$
$$= \tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}} + \frac{G_1^2 c_\gamma^2\tau_{\text{mix}}^2}{\mu^2\lambda^2 T^2} + \frac{G_1^{10} c_\gamma^6\tau_{\text{mix}}^4}{\mu^6\lambda^6 T_{\max}}\right) \tag{69}$$

We can decompose $\mathbb{E}\|(\mathbb{E}_k[\omega_k] - \omega_k^*)\|$ and $\mathbb{E}\|\omega_k - \omega_k^*\|^2$ using the definition that $\omega_k = \omega_r^k + \lambda_k\omega_c^k$. Using (68) and (69), and the fact that $\lambda_k \in [0, 2/\delta]$, we obtain

$$\mathbb{E}\|(\mathbb{E}_k[\omega_k] - \omega_k^*)\| \leq \left(1 + \frac{2}{\delta}\right)\tilde{\mathcal{O}}\left(\frac{G_1}{\mu}\sqrt{\epsilon_{\text{app}}} + \frac{G_1 c_\gamma\tau_{\text{mix}}}{\mu\lambda T} + \frac{G_1^5 c_\gamma^3\tau_{\text{mix}}^2}{\mu^3\lambda^3\sqrt{T_{\max}}}\right) \tag{70}$$

$$\mathbb{E}\|\omega_k - \omega_k^*\|^2 \leq \left(1 + \frac{4}{\delta^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}} + \frac{G_1^6 c_\gamma^4\tau_{\text{mix}}^3}{\mu^4\lambda^4}\right) \tag{71}$$

Setting $T_{\max} = T$ in (70) and (71), and using Lemma G.2 and (26) in (25) would obtain

$$\frac{1}{K}\mathbb{E}\sum_{k=0}^{K-1}\left(\mathcal{L}(\pi^*, \lambda_k) - \mathcal{L}(\theta_k, \lambda_k)\right) \leq \sqrt{\epsilon_{\text{bias}}} + \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))]$$
$$+ \left(1 + \frac{2}{\delta}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu}\sqrt{\epsilon_{\text{app}}} + \frac{G_1^6 c_\gamma^3\tau_{\text{mix}}^2}{\mu^3\lambda^3\sqrt{T}}\right) + \alpha G_2\left(1 + \frac{4}{\delta^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}} + \frac{G_1^6 c_\gamma^4\tau_{\text{mix}}^3}{\mu^4\lambda^4}\right)$$
$$+ \alpha\,\mathcal{O}\left(\left(1 + \frac{4}{\delta^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right)$$

https://doi.org/10.52202/085713-3099

Using the definition of the Lagrange function, the above inequality can be equivalently written as

$$\frac{1}{K}\mathbb{E}\sum_{k=0}^{K-1}\left(J_r^{\pi^*}-J_r(\theta_k)\right)\leq\sqrt{\epsilon_{\text{bias}}}+\frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right]$$
$$+\left(1+\frac{2}{\delta}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu}\sqrt{\epsilon_{\text{app}}}+\frac{G_1^6c_\gamma^3\tau_{\text{mix}}^2}{\mu^3\lambda^3\sqrt{T}}\right)+\alpha G_2\left(1+\frac{4}{\delta^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}}+\frac{G_1^6c_\gamma^4\tau_{\text{mix}}^3}{\mu^4\lambda^4}\right)$$
$$+\alpha\,\mathcal{O}\left(\left(1+\frac{4}{\delta^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right)-\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\lambda_k\left(J_c^{\pi^*}-J_c(\theta_k)\right)\right] \tag{72}$$

We will now extract the objective convergence rate from the convergence rate of the Lagrange function stated above. Note that,

$$\begin{aligned}
0\leq(\lambda_K)^2&\leq\sum_{k=0}^{K-1}\left((\lambda_{k+1})^2-(\lambda_k)^2\right)\\
&=\sum_{k=0}^{K-1}\left(\mathcal{P}_{[0,\frac{2}{\delta}]}\left[\lambda_k-\beta\eta_c^k\right]^2-(\lambda_k)^2\right)\\
&\leq\sum_{k=0}^{K-1}\left(\left[\lambda_k-\beta\eta_c^k\right]^2-(\lambda_k)^2\right)\\
&=-2\beta\sum_{k=0}^{K-1}\lambda_k\eta_c^k+\beta^2\sum_{k=0}^{K-1}(\eta_c^k)^2\\
&\overset{(a)}{\leq}2\beta\sum_{k=0}^{K-1}\lambda_k(J_c^{\pi^*}-\eta_c^k)+2\beta^2\sum_{k=0}^{K-1}(\eta_c^k)^2\\
&=2\beta\sum_{k=0}^{K-1}\lambda_k(J_c^{\pi^*}-J_c(\theta_k))+2\beta\sum_{k=0}^{K-1}\lambda_k(J_c(\theta_k)-\eta_c^k)+2\beta^2\sum_{k=0}^{K-1}(\eta_c^k)^2
\end{aligned} \tag{73}$$

Inequality (a) holds because θ^* is a feasible solution to the constrained optimization problem. Rearranging items and taking the expectation, we have

$$-\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\lambda_k(J_c^{\pi^*}-J_c(\theta_k))\Big]\leq\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\lambda_k(J_c(\theta_k)-\eta_c^k)\Big]+\frac{\beta}{K}\sum_{k=0}^{K-1}\mathbb{E}[\eta_c^k]^2 \tag{74}$$

Note that, unlike the discounted reward case, the average reward estimate η_c^k is no longer unbiased. However, by Theorem 4.8 and the facts that $|c(s,a)|\leq 1,\forall(s,a)\in\mathcal{S}\times\mathcal{A}$, and $\eta_c^*(\theta_k)=J_c(\theta_k)$, we get the following inequality.

$$\begin{aligned}
-\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\lambda_k(J_c^{\pi^*}-J_c(\theta_k))\Big]&\leq\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\lambda_k\left(\eta_c^*(\theta_k)-\mathbb{E}_k[\eta_c^k]\right)\right]+\beta\\
&\leq\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[|\lambda_k|\left|\eta_c^*(\theta_k)-\mathbb{E}_k[\eta_c^k]\right|\right]+\beta\\
&\leq\tilde{\mathcal{O}}\left(\frac{c_\gamma^3\tau_{\text{mix}}}{\lambda^3\delta\sqrt{T}}+\beta\right)
\end{aligned} \tag{75}$$

where the last inequality utilizes (57), the fact that $\lambda_k \in [0, 2/\delta]$, and $T_{\max} = T$. Combining with (72), we arrive at the following result.

$$\frac{1}{K}\mathbb{E}\sum_{k=0}^{K-1}\left(J_r^{\pi^*} - J_r(\theta_k)\right) \leq \sqrt{\epsilon_{\text{bias}}} + \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right]$$
$$+ \left(1+\frac{2}{\delta}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu}\sqrt{\epsilon_{\text{app}}} + \frac{G_1^6 c_\gamma^3 \tau_{\text{mix}}^2}{\mu^3\lambda^3\sqrt{T}}\right) + \alpha G_2\left(1+\frac{4}{\delta^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}} + \frac{G_1^6 c_\gamma^4 \tau_{\text{mix}}^3}{\mu^4\lambda^4}\right)$$
$$+ \alpha\,\mathcal{O}\left(\left(1+\frac{4}{\delta^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right) + \tilde{\mathcal{O}}\left(\frac{c_\gamma^3\tau_{\text{mix}}}{\lambda^3\delta\sqrt{T}} + \beta\right)$$
$$\leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \alpha + \beta + \frac{\tau_{\text{mix}}^2}{\sqrt{T}} + \frac{1}{\alpha K}\right) \tag{76}$$

F.0.2 Rate of Constraint Violation

Given the dual update in algorithm 1, we have

$$\begin{aligned}\left|\lambda_{k+1} - \frac{2}{\delta}\right|^2 &\overset{(a)}{\leq} \left|\lambda_k - \beta J_c(\theta_k) - \frac{2}{\delta}\right|^2 \\ &= \left|\lambda_k - \frac{2}{\delta}\right|^2 - 2\beta J_c(\theta_k)\left(\lambda_k - \frac{2}{\delta}\right) + \beta^2 J_c(\theta_k)^2 \\ &\leq \left|\lambda_k - \frac{2}{\delta}\right|^2 - 2\beta J_c(\theta_k)\left(\lambda_k - \frac{2}{\delta}\right) + \beta^2\end{aligned} \tag{77}$$

where (a) is because of the non-expansiveness of projection $\mathcal{P}_\Lambda$. Averaging the above inequality over $k = 0, \ldots, K-1$ yields

$$0 \leq \frac{1}{K}\left|\lambda_K - \frac{2}{\delta}\right|^2 \leq \frac{1}{K}\left|\lambda_0 - \frac{2}{\delta}\right|^2 - \frac{2\beta}{K}\sum_{k=0}^{K-1} J_c(\theta_k)\left(\lambda_k - \frac{2}{\delta}\right) + \beta^2 \tag{78}$$

Taking expectations on both sides,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[J_c(\theta_k)\left(\lambda_k - \frac{2}{\delta}\right)\right] \leq \frac{1}{2\beta K}\left|\lambda_0 - \frac{2}{\delta}\right|^2 + \frac{\beta}{2} \leq \frac{2}{\delta^2\beta K} + \frac{\beta}{2} \tag{79}$$

where the last inequality utilizes $\lambda_0 = 0$. Notice that $\lambda_k J_c^{\pi^*} \geq 0, \forall k$. Using the above inequality to (72), we therefore have,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big(J_r^{\pi^*} - J_r(\theta_k)\Big) + \frac{2}{\delta}\sum_{k=0}^{K-1}\frac{1}{K}\mathbb{E}\Big[-J_c(\theta_k)\Big] \leq \sqrt{\epsilon_{\text{bias}}} + \frac{2}{\delta^2\beta K} + \frac{\beta}{2}$$
$$+ \left(1+\frac{2}{\delta}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu}\sqrt{\epsilon_{\text{app}}} + \frac{G_1^6 c_\gamma^3 \tau_{\text{mix}}^2}{\mu^3\lambda^3\sqrt{T}}\right) + \alpha G_2\left(1+\frac{4}{\delta^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}} + \frac{G_1^6 c_\gamma^4 \tau_{\text{mix}}^3}{\mu^4\lambda^4}\right)$$
$$+ \alpha\,\mathcal{O}\left(\left(1+\frac{4}{\delta^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right) + \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right] \tag{80}$$

We define a new policy $\bar{\pi}$ which uniformly chooses the policy π_{θ_k} for $k \in [K]$. By the occupancy measure method, $J_g(\theta_k)$ is linear in terms of an occupancy measure induced by policy π_{θ_k}. Thus,

$$\frac{1}{K}\sum_{k=0}^{K-1} J_g(\theta_k) = J_g^{\bar{\pi}},\ g \in \{r, c\} \tag{81}$$

https://doi.org/10.52202/085713-3099

Injecting the above relation to (80), we have

$$\mathbb{E}\Big[J_r^{\pi^*} - J_r^{\bar{\pi}}\Big] + \frac{2}{\delta}\mathbb{E}\Big[-J_c^{\bar{\pi}}\Big] \leq \sqrt{\epsilon_{\text{bias}}} + \frac{2}{\delta^2\beta K} + \frac{\beta}{2}$$
$$+ \left(1+\frac{2}{\delta}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu}\sqrt{\epsilon_{\text{app}}} + \frac{G_1^6 c_\gamma^3 \tau_{\text{mix}}^2}{\mu^3\lambda^3\sqrt{T}}\right) + \alpha G_2\left(1+\frac{4}{\delta^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}} + \frac{G_1^6 c_\gamma^4 \tau_{\text{mix}}^3}{\mu^4\lambda^4}\right)$$
$$+ \alpha\,\mathcal{O}\left(\left(1+\frac{4}{\delta^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right) + \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))]$$

By Lemma G.6, we arrive at,

$$\mathbb{E}\Big[-J_c^{\bar{\pi}}\Big] \leq \delta\sqrt{\epsilon_{\text{bias}}} + \frac{1}{\delta\beta K} + \frac{\delta\beta}{2}$$
$$+ \left(1+\frac{\delta}{2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu}\sqrt{\epsilon_{\text{app}}} + \frac{G_1^6 c_\gamma^3 \tau_{\text{mix}}^2}{\mu^3\lambda^3\sqrt{T}}\right) + \alpha G_2\left(\frac{\delta}{2}+\frac{2}{\delta}\right)\tilde{\mathcal{O}}\left(\frac{G_1^2}{\mu^2}\epsilon_{\text{app}} + \frac{G_1^6 c_\gamma^4 \tau_{\text{mix}}^3}{\mu^4\lambda^4}\right)$$
$$+ \alpha\,\mathcal{O}\left(\left(\frac{\delta}{2}+\frac{2}{\delta}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right) + \frac{\delta}{2\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))]$$
$$= \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \alpha + \beta + \frac{1}{\alpha K} + \frac{1}{\beta K} + \frac{\tau_{\text{mix}}^2}{\sqrt{T}}\right) \tag{82}$$

F.0.3 Optimal Choice of α, β, K, and H

If τ_{mix} is unknown, we can take $\alpha = T^{-a}$, $\beta = T^{-b}$ for some $a, b \in (0,1)$, and $H = T^\epsilon$, $K = T^{1-\epsilon}$ for $\epsilon \in (0,1)$ then following (76) and (82), we can write,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big(J_r^{\pi^*} - J_r(\theta_k)\Big) \leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + T^{-a} + T^{-b} + T^{-0.5} + T^{-1+\epsilon+a}\right),$$
$$\mathbb{E}\Big[\frac{1}{K}\sum_{k=0}^{K-1} -J_c(\theta_k)\Big] \leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + T^{-a} + T^{-b} + T^{-0.5} + T^{-1+\epsilon+a} + T^{-1+\epsilon+b}\right)$$

Clearly, the optimal values of a and b can be obtained by solving the following optimization.

$$\max_{(a,b)\in(0,1)^2} \min\{a, b, 1-\epsilon-a, 1-\epsilon-b\} \tag{83}$$

By choosing $(a,b) = (1/2, 1/2)$, this would obtain the solution of the above optimization problem. Thus, the convergence rate and constraint violation would both become:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big(J_r^{\pi^*} - J_r(\theta_k)\Big) \leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \frac{1}{T^{(0.5-\epsilon)}}\right) \tag{84}$$
$$\mathbb{E}\Big[\frac{1}{K}\sum_{k=0}^{K-1} -J_c(\theta_k)\Big] \leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \frac{1}{T^{(0.5-\epsilon)}}\right) \tag{85}$$

Recall that the above result holds when the conditions for Theorem 4.8 and Theorem 4.7 are satisfied.

$$\frac{2\log T}{H} \leq \frac{\lambda^2}{24c_\gamma^2\tau_{\text{mix}}\log T_{\max}} = \Theta\left(\frac{1}{\tau_{\text{mix}}\log T}\right) \tag{86}$$
$$\frac{2\log T}{H} \leq \frac{\mu^2}{4(6G_1^4\tau_{\text{mix}}\log T_{\max} + 2G_1^2\tau_{\text{mix}}^2\log T_{\max})} = \Theta\left(\frac{1}{\tau_{\text{mix}}^2\log T}\right) \tag{87}$$

In light of the above conditions, (84) and (85) hold if $H = T^\epsilon \geq \tilde{\Theta}(\tau_{\text{mix}}^2)$. If τ_{mix} is known, we can set $H = \tilde{\Theta}(\tau_{\text{mix}}^2)$ that satisfies (86) and (87). Moreover, we can take $K = \mathcal{O}(T)$. Thus, by following

the same analysis as above, we can obtain:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\bigg(J_r^{\pi^*}-J_r(\theta_k)\bigg)\leq\tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}}+\sqrt{\epsilon_{\text{app}}}+\frac{1}{\sqrt{T}}\right) \tag{88}$$

$$\mathbb{E}\bigg[\frac{1}{K}\sum_{k=0}^{K-1}-J_c(\theta_k)\bigg]\leq\tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}}+\sqrt{\epsilon_{\text{app}}}+\frac{1}{\sqrt{T}}\right) \tag{89}$$

This concludes the proof of Theorem 4.9 and Theorem 4.10.

G Some Auxiliary Lemmas for the Proofs

Lemma G.1. *[41, Lemma 14] For any ergodic MDP with mixing time τ_{mix}, the following holds $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$, any policy π and $\forall g\in\{r,c\}$.*

$$(a)|V_g^\pi(s)|\leq 5\tau_{\text{mix}},\ \ (b)|Q_g^\pi(s,a)|\leq 6\tau_{\text{mix}},\ \ (c)|A_g^\pi(s,a)|=\mathcal{O}(\tau_{\text{mix}})$$

Lemma G.2. *For any ergodic MDP with mixing time τ_{mix}, the following holds $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$ for any policy π_{θ_k} with θ_k satisfying Assumption 4.4, the dual parameter $\lambda_k\in[0,2/\delta]$, and $g\in\{r,c\}$.*

$$(a)\|\nabla_\theta J_g(\theta_k)\|\leq 6\tau_{\text{mix}}G_1,\ \ (b)\|\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)\|\leq(6+\frac{12}{\delta})\tau_{\text{mix}}G_1$$

Proof. Using the policy gradient theorem (7), we have the following relation.

$$\nabla_\theta J_g(\theta_k)=\mathbb{E}_{(s,a)\sim\nu^{\pi_{\theta_k}}}\left[Q_g^{\pi_{\theta_k}}(s,a)\nabla_\theta\log\pi_{\theta_k}(a|s)\right] \tag{90}$$

Applying Lemma G.1(b) and Assumption 4.4(a), we get

$$\|\nabla_\theta J_g(\theta_k)\|=\left\|\mathbb{E}_{(s,a)\sim\nu^{\pi_{\theta_k}}}\left[Q_g^{\pi_{\theta_k}}(s,a)\nabla_\theta\log\pi_{\theta_k}(a|s)\right]\right\|\leq 6\tau_{\text{mix}}G_1 \tag{91}$$

Combining the above result with the definition of the Lagrange function and the bound that $\lambda_k\leq\frac{2}{\delta}$, we arrive at the following result.

$$\begin{aligned}\|\nabla_\theta\mathcal{L}(\theta_k,\lambda_k)\|&=\|\nabla_\theta J_r(\theta_k)+\lambda_k\nabla_\theta J_c(\theta_k)\|\\&\leq\|\nabla_\theta J_r(\theta_k)\|+\lambda_k\|\nabla_\theta J_c(\theta_k)\|\leq(6+\frac{12}{\delta})\tau_{\text{mix}}G_1\end{aligned} \tag{92}$$

This concludes the proof of Lemma G.2. □

Lemma G.3 (Strong duality)**.** *[17, Lemma 3] We restate the problem* (2) *below for convenience.*

$$\max_{\pi\in\Pi}J_r^\pi\ \ \textit{s.t. }J_c^\pi\geq 0 \tag{93}$$

where Π is the set of all policies. Define π^* as an optimal solution to the above optimization problem. Define the associated dual function as

$$J_D^\lambda\triangleq\max_{\pi\in\Pi}J_r^\pi+\lambda J_c^\pi \tag{94}$$

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

$$J_r^{\pi^*}=J_D^{\lambda^*} \tag{95}$$

Lemma G.4 (Lemma 16, [10])**.** *Consider the parameterized problem* (3) *where Θ is the collection of all parameters. Under Assumption 2.1, the optimal dual variable, $\lambda_\Theta^*=\arg\min_{\lambda\geq 0}\max_{\theta\in\Theta}J_r^{\pi_\theta}+\lambda J_c^{\pi_\theta}$, for the parameterized problem can be bounded as follows.*

$$0\leq\lambda_\Theta^*\leq\frac{J_r^{\pi^*}-J_r(\bar{\theta})}{\delta}\leq\frac{1}{\delta} \tag{96}$$

where π^* is an optimal solution to the unparameterized problem (93) *and $\bar{\theta}$ is a feasible parameter mentioned in Assumption 2.1.*

https://doi.org/10.52202/085713-3099

Notice that in Eq. (12), 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 G.4. We note that our dual updating technique remains the same as [10], however we were able to achieve an improvement of global convergence rate due to the use of natural policy gradient and a more prudent choice of stepsizes (α, β).

Lemma G.5. *Assume that the Assumption 2.1 holds. Define* $v(\tau) = \max_{\pi \in \Pi} \{J_r^\pi | J_c^\pi \geq \tau\}$ *where* Π *is the set of all policies. The following holds for any* $\tau \in \mathbb{R}$ *where* λ^* *is the optimal dual parameter for the unparameterized problem as stated in Lemma G.3.*

$$v(0) - \tau\lambda^* \geq v(\tau) \tag{97}$$

Proof. Using the definition of $v(\tau)$, we get $v(0) = J_r^{\pi^*}$ where π^* is a solution to the unparameterized problem (93). Denote $\mathcal{L}(\pi, \lambda) = J_r^\pi + \lambda J_c^\pi$. By the strong duality stated in Lemma G.3, we have the following for any $\pi \in \Pi$.

$$\mathcal{L}(\pi, \lambda^*) \leq \max_{\pi \in \Pi} \mathcal{L}(\pi, \lambda^*) \overset{Def}{=} J_D^{\lambda^*} \overset{(95)}{=} J_r^{\pi^*} = v(0) \tag{98}$$

Thus, for any $\pi \in \{\pi \in \Pi \,|\, J_c^\pi \geq \tau\}$, we can deduce the following.

$$v(0) - \tau\lambda^* \geq \mathcal{L}(\pi, \lambda^*) - \tau\lambda^* = J_r^\pi + \lambda^*(J_c^\pi - \tau) \geq J_r^\pi \tag{99}$$

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

$$v(0) - \tau\lambda^* \geq v(\tau) \tag{100}$$

This completes the proof of Lemma G.5. □

Lemma G.6. *Let Assumption 2.1 hold and* (π^*, λ^*) *be the optimal primal and dual parameters for the unparameterized problem* (93)*. For any constant* $C \geq 2\lambda^*$*, if there exist a* $\pi \in \Pi$ *and* $\zeta > 0$ *such that* $J_r^{\pi^*} - J_r^\pi + C[-J_c^\pi] \leq \zeta$*, then*

$$-J_c^\pi \leq 2\zeta/C \tag{101}$$

Proof. Let $\tau = J_c^\pi$. We have the following inequality following the definition of $v(\tau)$ provided in Lemma G.5.

$$J_r^\pi \leq v(\tau) \tag{102}$$

Combining Eq. (100) and (102), and using the fact that $v(0) = J_r^{\pi^*}$, we have the following inequality.

$$J_r^\pi - J_r^{\pi^*} \leq v(\tau) - v(0) \leq -\tau\lambda^* \tag{103}$$

Using the condition in the Lemma, we have

$$(C - \lambda^*)(-\tau) = \tau\lambda^* + C(-\tau) \leq J_r^{\pi^*} - J_r^\pi + C(-\tau) \leq \zeta \tag{104}$$

Since $C - \lambda^* \geq C/2 > 0$, we finally have the following inequality.

$$(-\tau) \leq \frac{\zeta}{C - \lambda^*} \leq \frac{2\zeta}{C} \tag{105}$$

This completes the proof of Lemma G.6. □