Global Convergence of Average Reward Constrained MDPs with Neural Critic and General Policy Parameterization

Anirudh Satheesh[1] **Pankaj Kumar Barman**[2] **Washim Uddin Mondal**[2] **Vaneet Aggarwal**[3]

[1]Department of Computer Science, University of Maryland, College Park, MD 20740, USA
[2]Indian Institute of Technology, Kanpur, Uttar Pradesh, 208016, India
[3]Purdue University, West Lafayette, IN 47907, USA

Abstract

We study infinite-horizon Constrained Markov Decision Processes (CMDPs) with general policy parameterizations and multi-layer neural network critics. Existing theoretical analyses for constrained reinforcement learning largely rely on tabular policies or linear critics, which limits their applicability to high-dimensional and continuous control problems. We propose a primal–dual natural actor–critic algorithm that integrates neural critic estimation with natural policy gradient updates and leverages Neural Tangent Kernel (NTK) theory to control function-approximation error under Markovian sampling, without requiring access to mixing-time oracles. We establish global convergence and cumulative constraint violation rates of $\tilde{\mathcal{O}}(T^{-1/4})$ up to approximation errors induced by the policy and critic classes. Our results provide the first such guarantees for CMDPs with general policies and multi-layer neural critics, substantially extending the theoretical foundations of actor–critic methods beyond the linear-critic regime.

1 INTRODUCTION

Reinforcement Learning (RL) has achieved remarkable empirical success in solving complex, high-dimensional continuous control problems, an advancement largely driven by the representational power of deep neural networks. However, as RL is increasingly deployed in safety-critical applications such as transportation [Al-Abbasi et al., 2019], healthcare [Tamboli et al., 2024], and robotics [Gonzalez et al., 2023], it is imperative to ensure that these agents adhere to strict operational constraints. This requirement is naturally formulated as a Constrained Markov Decision Process (CMDP), wherein the objective is to maximize a primary reward signal while maintaining auxiliary cost signals below predefined thresholds. To solve continuous-space CMDPs, primal-dual actor-critic algorithms have emerged as a dominant paradigm. In these methods, the actor updates a parameterized policy via gradient ascent, a critic evaluates the policy's performance to construct the gradient, and a dual variable dynamically penalizes constraint violations.

Despite the widespread practical adoption of deep actor-critic methods, the theoretical understanding of these algorithms remains limited. Early theoretical frameworks establishing global convergence for actor-critic algorithms were predominantly restricted to tabular settings or linear function approximations [Chen et al., 2022, Bai et al., 2023, Xu et al., 2025]. While linear models provide mathematical tractability, they fail to capture the complex, non-linear feature abstractions that characterize modern deep RL. More recently, literature has begun to bridge this gap by analyzing neural network parameterizations through the lens of Neural Tangent Kernel (NTK) theory [Gaur et al., 2024, Ganesh et al., 2025]. However, these theoretical advancements have largely been confined to unconstrained Markov Decision Processes or discounted reward formulations, leaving the average reward constrained setting unexplored. Thus, we ask the following question:

Can we design a primal-dual actor-critic algorithm with general policy parameterizations and multi-layer neural network critics that achieves provable global convergence for average-reward CMDPs under Markovian sampling, without requiring a mixing-time oracle?

1.1 CHALLENGES AND CONTRIBUTIONS

Addressing this problem requires overcoming three central technical obstacles. First, handling Markovian sampling dependence is typically achieved through data dropping strategies that rely on a mixing time oracle to specify appropriate discard intervals. Such assumptions are restrictive and often impractical. Second, analyzing a multi-layer neu-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5997–6025.

Table 1: Comparison of relevant algorithms for infinite-horizon average-reward Reinforcement Learning in the model-free setting. Our work is the first to provide average reward global convergence and constraint violation results for constrained MDPs with general policy and multi-layer neural network critic parametrization. We provide a more detailed comparison to prior work in Appendix A.

References	Global Convergence	Violation	Reward Setting	Infinite State Action Space	Mixing Time Unknown?	Critic Parameterization	Policy Parameterization
[Cayci et al., 2024]	$\tilde{\mathcal{O}}(T^{-1/5})$	N/A	Discounted	✗	✓	Multi layer NN	Multi layer NN
[Gaur et al., 2024]	$\tilde{\mathcal{O}}(T^{-1/3})$	N/A	Discounted	✓	✓	Multi layer NN	General
[Ganesh et al., 2025]	$\tilde{\mathcal{O}}(T^{-1/2})$	N/A	Discounted	✓	✗	Multi layer NN	General
[Mondal and Aggarwal, 2024b]	$\tilde{\mathcal{O}}(T^{-1/2})$	$\tilde{\mathcal{O}}(T^{-1/2})$	Discounted	✓	✓	N/A	General
[Agarwal et al., 2022b,a]	$\tilde{\mathcal{O}}(T^{-1/2})$	0	Average	✗	✗	Tabular	Tabular
[Ghosh et al., 2023]	$\tilde{\mathcal{O}}(T^{-1/2})$	$\tilde{\mathcal{O}}(T^{-1/2})$	Average	✓	✗	Linear MDP	Tabular
[Bai et al., 2024]	$\tilde{\mathcal{O}}(T^{-1/5})$	$\tilde{\mathcal{O}}(T^{-1/5})$	Average	✓	✗	N/A	General
[Xu et al., 2025]	$\tilde{\mathcal{O}}(T^{-1/2})$	$\tilde{\mathcal{O}}(T^{-1/2})$	Average	✓	✓	Linear	General
This work	$\tilde{\mathcal{O}}(T^{-1/4})$	$\tilde{\mathcal{O}}(T^{-1/4})$	Average	✓	✓	Multi layer NN	General

ral critic in the Neural Tangent Kernel (NTK) regime demands precise control of its approximation bias. If this bias does not decay sufficiently fast, the resulting Natural Policy Gradient (NPG) updates become inaccurate and unstable. Third, the infinite-horizon average-reward setting presents a fundamental difficulty: unlike discounted formulations, the average-reward Bellman operator is not contractive. This lack of contraction destabilizes critic evaluation. When combined with the saddle-point structure of constrained Markov decision processes (CMDPs), the coupled estimation errors from the actor, critic, and dual variables can accumulate and potentially lead to divergence.

To address these challenges, we introduce the Primal-Dual Natural Actor-Critic with Neural Critic (PDNAC-NC) algorithm and establish its global convergence guarantees. To remove the dependence on a mixing-time oracle, we incorporate a Multi-Level Monte Carlo (MLMC) estimator that samples trajectory lengths from a geometric distribution. This construction corrects Markovian bias without discarding data. To manage the nonlinear neural critic, we restrict its parameters to remain within an NTK neighborhood around initialization and prove that its approximation bias decays proportionally to the mean-squared error, ensuring reliable NPG estimation. Finally, to handle the non-contractive Bellman operator and the primal-dual structure, we develop a refined coupled analysis that carefully tracks error propagation across actor, critic, and dual updates. This approach controls error amplification throughout the min-max dynamics and yields a convergence rate of $\tilde{\mathcal{O}}(T^{-1/4})$ in both optimality gap and constraint violation. To our knowledge, this is the first result establishing global convergence for average-reward CMDPs with multi-layer neural critics and general policy parameterizations, without assuming access to a mixing-time oracle.

2 RELATED WORKS

Neural Actor-Critic and Markovian Sampling: Recent theoretical advancements have begun to bridge the gap between deep reinforcement learning practice and theory by analyzing actor-critic methods equipped with multi-layer neural networks via Neural Tangent Kernel (NTK) theory [Fu et al., 2020, Cayci et al., 2024, Tian et al., 2023, Gaur et al., 2024, Ganesh et al., 2025]. While these contemporary works achieve state-of-the-art sample complexities for general policy parameterizations, their analyses are strictly confined to standard, unconstrained MDPs with discounted rewards. Furthermore, under Markovian sampling, these approaches rely on restrictive data-dropping techniques that discard collected transitions based on a known mixing-time oracle to mitigate statistical dependencies. However, this results in only one out of every τ_{mix} samples used for policy updates. To bypass this, recent theoretical works have adapted Multi-Level Monte Carlo (MLMC) estimation to RL [Beznosikov et al., 2023, Xu et al., 2025], providing unbiased gradient estimates that correct for Markovian bias without sample thinning. However, these MLMC techniques have previously been restricted to algorithms utilizing linear function approximations for the critic. We are the first to extend this technique to the neural critic setting, mitigating the issues of data dropping.

Constrained Average-Reward MDPs: For safe reinforcement learning, infinite-horizon average-reward Constrained MDPs (CMDPs) present a difficult challenge due to coupled primal-dual dynamics and the lack of Bellman operator contractivity. Consequently, existing theoretical guarantees for average-reward CMDPs have predominantly relied on tabular settings, exact gradients, linear MDP, or linear critic approximations [Paternain et al., 2019, Ding et al., 2020, Chen et al., 2022, Bai et al., 2024, Xu et al., 2025, Ghosh et al., 2023, Wei et al., 2022, Satheesh and Aggarwal, 2026], failing to capture the complex, non-linear feature representations characteristic of deep RL. Our work distinguishes itself by being the first to theoretically analyze a primal-dual

natural actor-critic algorithm for average-reward CMDPs utilizing multi-layer neural network critics. By uniquely integrating MLMC within a nested-loop structure to simultaneously evaluate a neural critic and estimate the Natural Policy Gradient (NPG), our approach eliminates the need for a mixing-time oracle while achieving global convergence in this setting.

3 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 stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps each state to a distribution over actions. Under policy π, the agent generates trajectories $\{(s_t, a_t)\}_{t=0}^{\infty}$ where $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$. The policy induces a state transition kernel $P^\pi : \mathcal{S} \to \Delta(\mathcal{S})$ defined by $P^\pi(s, s') = \sum_{a \in \mathcal{A}} \pi(a|s) P(s'|s, a)$ for all $s, s' \in \mathcal{S}$.

We assume the CMDP is ergodic throughout this work.

Assumption 3.1 (Ergodicity)**.** 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.

Under ergodicity, for each policy π, there exists a unique ρ-independent stationary distribution $d^\pi \in \Delta(\mathcal{S})$ satisfying $P^\pi d^\pi = d^\pi$. We define the state-action occupancy measure as $\nu^\pi(s, a) \triangleq d^\pi(s)\pi(a|s)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Definition 3.2 (Mixing Time)**.** The mixing time of the CMDP $\mathcal{M}$ with respect to a policy π is defined as

$$\tau^\pi_{\text{mix}} := \min \left\{ t \geq 1 \middle| \|(P^\pi)^t(s, \cdot) - d^\pi\|_{\text{TV}} \leq \frac{1}{4}, \forall s \right\}, \tag{1}$$

where $\|\cdot\|_{\text{TV}}$ denotes the total variation distance. The uniform mixing time is $\tau_{\text{mix}} := \sup_\pi \tau^\pi_{\text{mix}}$, which is finite under ergodicity.

The average reward and average constraint cost of a policy π are defined as

$$J^\pi_g \triangleq \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} g(s_t, a_t) \middle| s_0 \sim \rho \right], \quad g \in \{r, c\}, \tag{2}$$

where the expectation is over the distribution of π-induced trajectories. Under ergodicity, J^π_g is independent of ρ and can be written as $J^\pi_g = \mathbb{E}_{(s,a) \sim \nu^\pi}[g(s, a)]$. Our objective is to maximize the average reward while ensuring the average cost exceeds a given threshold. Without loss of generality, we formulate this as

$$\max_\pi J^\pi_r \quad \text{subject to} \quad J^\pi_c \geq 0. \tag{3}$$

When the state space $\mathcal{S}$ is large or continuous, we consider a class of parametrized policies $\{\pi_\theta \mid \theta \in \Theta\}$ indexed by a d-dimensional parameter $\theta \in \mathbb{R}^d$ where $d \ll |\mathcal{S}||\mathcal{A}|$. Using the shorthand $J_g(\theta) = J^{\pi_\theta}_g$ for $g \in \{r, c\}$, the optimization problem (3) becomes

$$\max_{\theta \in \Theta} J_r(\theta) \quad \text{subject to} \quad J_c(\theta) \geq 0. \tag{4}$$

We assume the optimization problem (4) satisfies the Slater condition, which ensures the existence of an interior point solution.

Assumption 3.3 (Slater Condition)**.** There exists $\delta_s \in (0, 1)$ and $\bar{\theta} \in \Theta$ such that $J_c(\bar{\theta}) \geq \delta_s$.

The action-value function corresponding to policy π_θ for $g \in \{r, c\}$ is defined for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ as

$$Q^{\pi_\theta}_g(s, a) = \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{\infty} (g(s_t, a_t) - J_g(\theta)) \middle| s_0 = s, a_0 = a \right]. \tag{5}$$

The state-value function is $V^{\pi_\theta}_g(s) = \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[Q^{\pi_\theta}_g(s, a)]$ for all $s \in \mathcal{S}$, and the advantage function is $A^{\pi_\theta}_g(s, a) \triangleq Q^{\pi_\theta}_g(s, a) - V^{\pi_\theta}_g(s)$. The Bellman equation for average reward MDPs takes the form

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

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $g \in \{r, c\}$. We solve (4) via a primal-dual approach based on the saddle point optimization

$$\max_{\theta \in \Theta} \min_{\lambda \geq 0} L(\theta, \lambda) \triangleq J_r(\theta) + \lambda J_c(\theta), \tag{7}$$

where $L(\cdot, \cdot)$ is the Lagrange function and $\lambda \geq 0$ is the dual variable. The policy gradient theorem Sutton et al. [1999] provides the gradient of the average reward for $g \in \{r, c\}$:

$$\nabla_\theta J_g(\theta) = \mathbb{E}_{\substack{s \sim d^{\pi_\theta} \\ a \sim \pi_\theta(\cdot|s)}} \left[A^{\pi_\theta}_g(s, a) \nabla_\theta \log \pi_\theta(a|s) \right]. \tag{8}$$

Rather than updating along the vanilla gradient, we employ the Natural Policy Gradient (NPG) direction

$$\omega^*_{g,\theta} \triangleq F(\theta)^{-1} \nabla_\theta J_g(\theta), \tag{9}$$

where $F(\theta) \in \mathbb{R}^{d \times d}$ is the Fisher information matrix defined as

$$F(\theta) = \mathbb{E} \left[\nabla_\theta \log \pi_\theta(a|s) \otimes \nabla_\theta \log \pi_\theta(a|s) \right], \tag{10}$$

with $\otimes$ denoting the outer product and the expectation taken under the state-action occupancy measure. The NPG direction $\omega^*_{g,\theta}$ can equivalently be expressed as the solution to the strongly convex optimization problem

$$\min_{\omega\in\mathbb{R}^d} f_g(\theta,\omega) := \frac{1}{2}\omega^\top F(\theta)\omega - \omega^\top \nabla_\theta J_g(\theta), \qquad (11)$$

whose gradient is $\nabla_\omega f_g(\theta,\omega) = F(\theta)\omega - \nabla_\theta J_g(\theta)$.

Since the gradients in (8) and (11) depend on the unknown action-value function $Q_g^{\pi_\theta}$, we introduce a neural network critic to approximate it. Let $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^n$ be a fixed feature map satisfying $\|\phi(s,a)\| \le 1$ for all (s,a). For each $g \in \{r,c\}$, we parameterize the Q-function using an L-layer feedforward neural network defined by

$$x^{(l)} = \frac{1}{\sqrt{m}}\sigma\left(W_l x^{(l-1)}\right), \quad l \in \{1,2,\dots,L\}, \qquad (12)$$

where $x^{(0)} = \phi(s,a) \in \mathbb{R}^n$ is the input, $W_1 \in \mathbb{R}^{m\times n}$ and $W_l \in \mathbb{R}^{m\times m}$ for $2 \le l \le L$ are weight matrices, m is the network width, and $\sigma(\cdot)$ is the activation function applied element-wise. The approximate Q-function is then

$$Q_g(\phi(s,a);\zeta_g) = \frac{1}{\sqrt{m}} b_g^\top x^{(L)}, \qquad (13)$$

where $\zeta_g = (\mathrm{Vec}(W_1);\dots;\mathrm{Vec}(W_L)) \in \mathbb{R}^{m(n+(L-1)m)}$ collects all trainable weights with $\mathrm{Vec}(\cdot)$ denoting vectorization by stacking columns, and $b_g \in \mathbb{R}^m$ is a fixed random vector. The network is initialized by drawing each entry of W_l^0 i.i.d. from $\mathcal{N}(0,1)$ and each entry of b_g uniformly from $\{-1,+1\}$. We denote $\zeta_{g,0} = (\mathrm{Vec}(W_1^0);\dots;\mathrm{Vec}(W_L^0))$ as the initial parameters.

Assumption 3.4 (Smooth Activation)**.** The activation function $\sigma(\cdot)$ is L_1-Lipschitz and L_2-smooth, i.e., for all $y_1, y_2 \in \mathbb{R}$: $|\sigma(y_1) - \sigma(y_2)| \le L_1|y_1 - y_2|$ and $|\sigma'(y_1) - \sigma'(y_2)| \le L_2|y_1 - y_2|$.

Assumption 3.4 is satisfied by smooth approximations of ReLU such as Sigmoid, ELU, and GeLU, which are commonly used in practice Clevert et al. [2015], Hendrycks [2016]. This provides an $O(m^{-1/2})$-smooth property for the neural Q-function. To enable Neural Tangent Kernel (NTK) analysis, we constrain the critic parameters to a ball around initialization:

$$\mathcal{S}_R := \{\zeta : \|\zeta - \zeta_0\|_2 \le R\}, \qquad (14)$$

and denote the Euclidean projection onto $\mathcal{S}_R$ as $\Pi_{\mathcal{S}_R}$. The local linearization of the neural Q-network at initialization defines the function class

$$\mathcal{F}_{R,m} := \Big\{\hat{Q}_g(\cdot;\zeta) = Q_g(\cdot;\zeta_{g,0}) + \langle \nabla_\zeta Q_g(\cdot;\zeta_{g,0}), \zeta - \zeta_{g,0}\rangle \Big| \zeta \in \mathcal{S}_R\Big\}. \qquad (15)$$

4 ALGORITHM

We solve (4) via a primal-dual algorithm that updates the policy parameter θ and dual variable λ iteratively. Starting from an arbitrary initial point $(\theta_0, \lambda_0 = 0)$, the ideal updates are

$$\theta_{k+1} = \theta_k + \alpha F(\theta_k)^{-1}\nabla_\theta L(\theta_k,\lambda_k) \qquad (16)$$

$$\lambda_{k+1} = \mathcal{P}_{[0,2/\delta_s]}\left[\lambda_k - \beta J_c(\theta_k)\right], \qquad (17)$$

where α and β are primal and dual learning rates respectively, δ_s is the Slater parameter from Assumption 3.3, and $\mathcal{P}_\Lambda$ denotes projection onto Λ. Since $\nabla_\theta L(\theta_k,\lambda_k)$, $F(\theta_k)$, and $J_c(\theta_k)$ are not exactly computable due to the unknown transition kernel P and stationary distribution $\nu^{\pi_{\theta_k}}$, we employ approximate updates

$$\theta_{k+1} = \theta_k + \alpha\omega_k, \quad \lambda_{k+1} = \mathcal{P}_{[0,2/\delta_s]}\left[\lambda_k - \beta\eta_c^k\right], \qquad (18)$$

where ω_k estimates the NPG direction $F(\theta_k)^{-1}\nabla_\theta L(\theta_k,\lambda_k)$ and η_c^k estimates $J_c(\theta_k)$. Our algorithm, Primal-Dual Natural Actor-Critic with Neural Critic (PDNAC-NC), is presented in Algorithm 1. It employs a Multi-Level Monte Carlo (MLMC) estimation procedure within a nested loop structure where the outer loop of length K updates the primal-dual parameters, and the inner loops of length H execute the neural critic and NPG estimation subroutines.

4.1 NEURAL CRITIC ESTIMATION

At the k-th epoch, the critic estimation subroutine serves two purposes: (i) obtain an estimate η_g^k of the average reward/cost $J_g(\theta_k)$ for $g \in \{r,c\}$, and (ii) obtain the neural network parameters ζ_g^k that approximate the Q-function $Q_g^{\pi_{\theta_k}}$. For the average reward estimation, note that $J_g(\theta_k)$ solves the optimization

$$\min_{\eta\in\mathbb{R}} \mathcal{R}_g(\theta_k,\eta) = \frac{1}{2}\mathbb{E}_{\nu_g^{\pi_{\theta_k}}}\left(\eta - g(s,a)\right)^2 \qquad (19)$$

where the expectation is taken over the state-action occupancy measure. For the Q-function approximation, we minimize the mean-squared projected Bellman error using the linearized neural network class $\mathcal{F}_{R,m}$. Define the critic objective

$$\mathcal{E}_g(\theta_k,\zeta) = \frac{1}{2}\mathbb{E}_{\nu_g^{\pi_{\theta_k}}}\left(Q_g(\phi(s,a);\zeta) - Q_g^{\pi_{\theta_k}}(s,a)\right)^2 \qquad (20)$$

Thus, by the Bellman equation, the gradients are represented as

$$\nabla_\eta \mathcal{R}_g(\theta_k,\eta) = \mathbb{E}_{\nu_g^{\pi_{\theta_k}}}\left[\eta - g(s,a)\right] \qquad (21)$$

$$\nabla_\zeta \mathcal{E}_g(\theta_k,\zeta) = \mathbb{E}_{\nu_g^{\pi_{\theta_k}}}\left[\left(Q_g(\phi(s,a);\zeta) - g(s,a)\right.\right.$$

$$+ J_g(\theta_k) - \mathbb{E}\big[Q_g^{\pi_{\theta_k}}(s', a')\big]\big)\nabla_\zeta Q(\phi(s,a);\zeta)\big], \quad (22)$$

where the outer expectation for $\nabla_\zeta \mathcal{E}(\theta_k, \zeta)$ is taken over $(s,a) \sim \nu_g^{\pi_{\theta_k}}$, while the inner expectation $\mathbb{E}[Q_g^{\pi_{\theta_k}}(s', a')]$ is taken over $s' \sim P(\cdot|s,a)$, and $a' \sim \pi_{\theta_k}(\cdot|s')$. However, we cannot compute the exact gradients as we do not know the state-action occupancy measure. Thus, we perform the following gradient descent steps:

$$\eta_{g,h+1}^k = \eta_{g,h}^k - c_\gamma \gamma_\xi \hat{\nabla}_\eta \mathcal{R}(\theta_k, \eta_{g,h}^k) \quad (23)$$

Algorithm 1 Primal-Dual Natural Actor-Critic with Neural Critic (PDNAC-NC)

Require: Initial parameters θ_0, $\lambda_0 = 0$; neural critic initialization $\zeta_{g,0}$ for $g \in \{r, c\}$; step sizes $\alpha, \beta, \gamma_\xi, \gamma_\omega$; outer loops K; inner loops H; truncation $T_{\max}$; projection radius R

Initialize: $s_0 \sim \rho(\cdot)$
Critic Initialization: Draw each entry of $W_l^0 \sim \mathcal{N}(0,1)$ for $l = 1, \ldots, L$, and each entry of $b_g \sim \text{Unif}\{-1, +1\}$
for $k = 0, \ldots, K-1$ **do**
  $\xi_{g,0}^k \leftarrow [0, \zeta_{g,0}^\top]^\top$, $\omega_{g,0}^k \leftarrow \mathbf{0}$ for all $g \in \{r, c\}$
  for $h = 0, \ldots, H-1$ **do**
    $s_{kh}^0 \leftarrow s_0$, draw $P_h^k \sim \text{Geom}(1/2)$
    $\ell_{kh} \leftarrow (2^{P_h^k} - 1)\mathbf{1}(2^{P_h^k} \le T_{\max}) + 1$
    for $t = 0, \ldots, \ell_{kh} - 1$ **do**
      Take action $a_{kh}^t \sim \pi_{\theta_k}(\cdot|s_{kh}^t)$
      Observe $s_{kh}^{t+1} \sim P(\cdot|s_{kh}^t, a_{kh}^t)$
      Observe $g(s_{kh}^t, a_{kh}^t)$ for $g \in \{r, c\}$
    end for
    $s_0 \leftarrow s_{kh}^{\ell_{kh}}$
    Compute $v_g^{\text{MLMC}}(\theta_k, \xi_{g,h}^k)$ via (26) for $g \in \{r, c\}$
    Update $\xi_{g,h+1}^k \leftarrow \Pi_{\mathcal{S}_R}\left(\xi_{g,h}^k - \gamma_\xi v_g^{\text{MLMC}}(\theta_k, \xi_{g,h}^k)\right)$
  end for
  for $h = 0, \ldots, H-1$ **do**
    $s_{kh}^0 \leftarrow s_0$, draw $Q_h^k \sim \text{Geom}(1/2)$
    $\ell_{kh} \leftarrow (2^{Q_h^k} - 1)\mathbf{1}(2^{Q_h^k} \le T_{\max}) + 1$
    for $t = 0, \ldots, \ell_{kh} - 1$ **do**
      Take action $a_{kh}^t \sim \pi_{\theta_k}(\cdot|s_{kh}^t)$
      Observe $s_{kh}^{t+1} \sim P(\cdot|s_{kh}^t, a_{kh}^t)$
      Observe $g(s_{kh}^t, a_{kh}^t)$ for $g \in \{r, c\}$
    end for
    $s_0 \leftarrow s_{kh}^{\ell_{kh}}$
    Compute $u_g(\theta_k, \omega_{g,h}^k, \xi_g^k)$ via (32) for $g \in \{r, c\}$
    Update $\omega_{g,h+1}^k \leftarrow \omega_{g,h}^k - \gamma_\omega u_g(\theta_k, \omega_{g,h}^k, \xi_g^k)$
  end for
  $\xi_g^k \leftarrow \xi_{g,H}^k$, $\omega_g^k \leftarrow \omega_{g,H}^k$ for $g \in \{r, c\}$
  $\omega_k \leftarrow \omega_r^k + \lambda_k \omega_c^k$
  $\theta_{k+1} \leftarrow \theta_k + \alpha\omega_k$
  $\lambda_{k+1} \leftarrow \mathcal{P}_{[0, 2/\delta_s]}\left[\lambda_k - \beta\eta_c^k\right]$
end for

$$\zeta_{g,h+1}^k = \Pi_{\mathcal{S}_R}\left(\zeta_{g,h}^k - \gamma_\xi \hat{\nabla}_\zeta \mathcal{E}(\theta_k, \zeta_{g,h}^k)\right) \quad (24)$$

where $\Pi_{\mathcal{S}_R}$ is the projection onto NTK ball, c_γ is a scaling parameter, γ_ξ is the learning rate, and $\hat{\nabla}_\eta \mathcal{R}$ and $\hat{\nabla}_\zeta \mathcal{E}$ are estimates of the gradients. We can also represent this through the combined critic parameter $\xi_{g,h}^k = \begin{bmatrix}\eta_{g,h}^k & {\zeta_{g,h}^k}^\top\end{bmatrix}^\top$ and the combined gradient

$$v_g(\theta, \xi_{g,h}^k; z_{kh}^j) = \begin{bmatrix} c_\gamma \hat{\nabla}_\eta \mathcal{R}\left(\theta_k, \eta_{g,h}^k; z_{kh}^j\right) \\ \hat{\nabla}_\zeta \mathcal{E}\left(\theta_k, \zeta_{g,h}^k; z_{kh}^j\right) \end{bmatrix} \quad (25)$$

where $\hat{\nabla}_\eta \mathcal{R}\left(\theta_k, \eta_{g,h}^k; z_{kh}^j\right)$ and $\hat{\nabla}_\zeta \mathcal{E}\left(\theta_k, \zeta_{g,h}^k; z_{kh}^j\right)$ are single transition estimates of the corresponding gradients evaluated on the single transition $z_{kh}^t = (s_{kh}^t, a_{kh}^t, s_{kh}^{t+1})$ collected at inner step t (consisting of the current state, the action taken, and the resulting next state). To reduce bias from Markovian sampling while maintaining computational efficiency, we employ MLMC estimation. For each inner iteration h, we draw $P_h^k \sim \text{Geom}(1/2)$ and generate a trajectory of length $\ell_{kh} = (2^{P_h^k} - 1)\mathbf{1}(2^{P_h^k} \le T_{\max}) + 1$. The MLMC estimate is

$$v_g^{\text{MLMC}}(\theta_k, \xi_{g,h}^k) = v_{g,kh}^0 + \begin{cases} 2^{P_h^k}\left(v_{g,kh}^{P_h^k} - v_{g,kh}^{P_h^k - 1}\right) & \text{if } 2^{P_h^k} \le T_{\max} \\ 0, & \text{otherwise} \end{cases} \quad (26)$$

where $v_{g,kh}^j = 2^{-j}\sum_{t=0}^{2^j-1} v_g(\theta_k, \xi_{g,h}^k; z_{kh}^t)$ for $j \in \{0, P_h^k - 1, P_h^k\}$. The critic parameters are updated via projected gradient descent:

$$\xi_{g,h+1}^k = \Pi_{\mathcal{S}_R}\left(\xi_{g,h}^k - \gamma_\xi v_g^{\text{MLMC}}(\theta_k, \xi_{g,h}^k)\right), \quad (27)$$

where γ_ξ is the critic learning rate and $\Pi_{\mathcal{S}_R}$ projects only the ζ portion onto the set $\mathcal{S}_R$. After H inner iterations, we obtain $\xi_g^k = \xi_{g,H}^k = [\eta_g^k, (\zeta_g^k)^\top]^\top$.

The MLMC estimator achieves the same bias as averaging $T_{\max}$ samples but requires only $\mathcal{O}(\log T_{\max})$ samples on average. Moreover, since drawing from a geometric distribution does not require knowledge of the mixing time, our algorithm eliminates the mixing time oracle assumption used in prior works [Ganesh et al., 2025].

4.2 NATURAL POLICY GRADIENT ESTIMATION

Given the critic estimates $\xi_g^k = [\eta_g^k, (\zeta_g^k)^\top]^\top$ from the previous subroutine, we now estimate the NPG direction $\omega_{g,\theta_k}^* = F(\theta_k)^{-1}\nabla_\theta J_g(\theta_k)$. Recall from (11) that ω_{g,θ_k}^* minimizes $f_g(\theta_k, \omega) = \frac{1}{2}\omega^\top F(\theta_k)\omega - \omega^\top \nabla_\theta J_g(\theta_k)$.

For a transition $z_{kh}^j = (s_{kh}^j, a_{kh}^j, s_{kh}^{j+1})$, we define the single-sample estimates

$$\hat{F}(\theta_k; z_{kh}^j) = \nabla_\theta \log \pi_{\theta_k}(a_{kh}^j|s_{kh}^j) \otimes \nabla_\theta \log \pi_{\theta_k}(a_{kh}^j|s_{kh}^j), \tag{28}$$

$$\hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z_{kh}^j) = \hat{A}_g^{\pi_{\theta_k}}(\xi_g^k; z_{kh}^j) \nabla_\theta \log \pi_{\theta_k}(a_{kh}^j|s_{kh}^j), \tag{29}$$

where the advantage is estimated via the neural temporal difference error

$$\begin{aligned}\hat{A}_g^{\pi_{\theta_k}}(\xi_g^k; z_{kh}^j) = g(s_{kh}^j, a_{kh}^j) - \eta_g^k + Q_g(\phi(s_{kh}^{j+1}, a'); \zeta_g^k) \\ - Q_g(\phi(s_{kh}^j, a_{kh}^j); \zeta_g^k),\end{aligned} \tag{30}$$

with $a' \sim \pi_{\theta_k}(\cdot|s_{kh}^{j+1})$. The single-sample NPG gradient estimate is

$$\begin{aligned}&\nabla_\omega \hat{f}_g(\theta_k, \omega_{g,h}^k, \xi_g^k; z_{kh}^j) \\ &= \hat{F}(\theta_k; z_{kh}^j)\omega_{g,h}^k - \hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z_{kh}^j).\end{aligned} \tag{31}$$

Similarly to the critic estimation, we apply MLMC to obtain variance-reduced estimates. For each inner iteration h, we draw $Q_h^k \sim \text{Geom}(1/2)$ and generate a trajectory of length $\ell_{kh} = (2^{Q_h^k} - 1)\mathbf{1}(2^{Q_h^k} \le T_{\max}) + 1$. The MLMC estimate is

$$\begin{aligned}&u_g(\theta_k, \omega_{g,h}^k, \xi_g^k) \\ &= u_{g,kh}^0 + \begin{cases} 2^{Q_h^k}\left(u_{g,kh}^{Q_h^k} - u_{g,kh}^{Q_h^k - 1}\right), & \text{if } 2^{Q_h^k} \le T_{\max} \\ 0, & \text{otherwise}\end{cases}\end{aligned} \tag{32}$$

where $u_{g,kh}^j = 2^{-j}\sum_{t=0}^{2^j-1} \nabla_\omega \hat{f}_g(\theta_k, \omega_{g,h}^k, \xi_g^k; z_{kh}^t)$ for $j \in \{0, Q_h^k - 1, Q_h^k\}$. The NPG parameters are updated as

$$\omega_{g,h+1}^k = \omega_{g,h}^k - \gamma_\omega u_g(\theta_k, \omega_{g,h}^k, \xi_g^k), \tag{33}$$

where γ_ω is the NPG learning rate. After H inner iterations, we obtain $\omega_g^k = \omega_{g,H}^k$ for each $g \in \{r, c\}$.

4.3 KEY ALGORITHMIC AND TECHNICAL NOVELTIES

Several aspects of Algorithm 1 merit further discussion.

MLMC for Mixing Time Independence. The MLMC estimators (26) and (32) achieve the same bias as averaging $T_{\max}$ samples while requiring only $O(\log T_{\max})$ samples in expectation. Crucially, since the geometric distribution $\text{Geom}(1/2)$ does not depend on the mixing time, our algorithm eliminates the need for a mixing time oracle, a requirement in many prior works [Bai et al., 2024, Ganesh et al., 2025]. This allows us to achieve convergence rates without knowledge of the exact value of τ_{mix}, simply an upper bound suffices.

Neural Tangent Kernel Regime. The projection $\Pi_{\mathcal{S}_R}$ onto the ball of radius R around the initialization $\zeta_{g,0}$ ensures the critic parameters remain in the NTK regime. In this regime, the neural network behaves approximately linearly [Jacot et al., 2018], enabling us to leverage the linearized function class $\mathcal{F}_{R,m}$ for analysis. When the width m is sufficiently large, the linearization error becomes negligible.

Runtime Efficiency. Unlike recent neural actor-critic algorithms that rely on data dropping techniques to mitigate statistical dependencies [Gaur et al., 2024, Ganesh et al., 2025] by utilizing only one sample out of every τ_{mix} steps, our MLMC-based approach eliminates the need to discard data. By correcting for Markovian bias through multi-level estimation rather than sample thinning, our algorithm utilizes the entire collected trajectory.

5 ASSUMPTIONS

We now state the assumptions required for our analysis.

5.1 POLICY PARAMETERIZATION ASSUMPTIONS

Assumption 5.1 (Bounded and Lipschitz Score)**.** For all $\theta, \theta_1, \theta_2 \in \Theta$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\|\nabla_\theta \log \pi_\theta(a|s)\| \le G_1 \tag{34}$$

$$\|\nabla_\theta \log \pi_{\theta_1}(a|s) - \nabla_\theta \log \pi_{\theta_2}(a|s)\| \le G_2\|\theta_1 - \theta_2\| \tag{35}$$

Assumption 5.1 requires that the score function $\nabla_\theta \log \pi_\theta(a|s)$ is uniformly bounded and Lipschitz continuous in θ. This is a standard assumption in policy gradient analysis [Liu et al., 2020, Agarwal et al., 2021, Papini et al., 2018, Fatkhullin et al., 2023, Mondal and Aggarwal, 2024a]. It holds for many common policy parameterizations, including Gaussian policies with linearly parameterized means and bounded covariance, softmax policies with bounded features, and certain neural network policies with bounded weights.

Assumption 5.2 (Fisher Non-degeneracy)**.** There exists $\mu > 0$ such that $F(\theta) \succeq \mu I_d$ for all $\theta \in \Theta$, where I_d denotes the d-dimensional identity matrix.

Assumption 5.2 ensures that the Fisher information matrix $F(\theta)$ has eigenvalues bounded away from zero. This guarantees that the NPG direction $\omega^* = F(\theta)^{-1}\nabla_\theta J(\theta)$ is well-defined and that the optimization problem (11) is μ-strongly convex. This assumption is standard for deriving global complexity bounds in policy gradient methods [Liu et al., 2020, Zhang et al., 2021, Bai et al., 2023, Fatkhullin et al., 2023, Mondal and Aggarwal, 2024a].

Assumption 5.3 (Transferred Compatible Function Approximation)**.** Define for $\theta, \omega \in \mathbb{R}^d$, $\lambda \geq 0$, and $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$:

$$\mathcal{L}_\nu(\omega, \theta, \lambda) = \mathbb{E}_\nu \left[\left(\nabla_\theta \log \pi_\theta(a|s) \cdot \omega - A^{\pi_\theta}(s,a) \right)^2 \right]. \tag{36}$$

where $A^{\pi_\theta}(s,a) = A_r^{\pi_\theta}(s,a) + \lambda A_c^{\pi_\theta}(s,a)$ is the combined advantage function and the expectation is taken under the distribution ν. Let $\omega^*_{\theta,\lambda} \in \arg\min_\omega \mathcal{L}_{\nu^{\pi_\theta}}(\omega, \theta, \lambda)$. There exists $\epsilon_{\text{bias}} \geq 0$ such that

$$\mathcal{L}_{\nu^{\pi_{\theta^*}}}(\omega^*_{\theta,\lambda}, \theta, \lambda) \leq \epsilon_{\text{bias}} \tag{37}$$

for all $\theta \in \Theta$ and $\lambda \in [0, 2/\delta_s]$, where π_{θ^*} is optimal for the average reward CMDP (3) and $\nu^{\pi_{\theta^*}}$ is the state-action occupancy measure.

Assumption 5.3 bounds the *transferred* compatible function approximation error, measuring how well the NPG direction computed under the current policy π_θ approximates the advantage function under the optimal policy π_{θ^*}, and is a common assumption in policy gradient literature [Agarwal et al., 2021, Fatkhullin et al., 2023, Ganesh et al., 2025, Ganesh and Aggarwal, 2025, Satheesh and Aggarwal, 2026]. The term ϵ_{bias} reflects the expressivity of the policy parameterization. When the policy class is expressive enough to represent any policy, $\epsilon_{\text{bias}} = 0$. One example for where this holds is for tabular policies with softmax parameterization. If the policy class is not expressive enough, then $\epsilon_{\text{bias}} > 0$, but recent work has shown that this factor is negligble when using neural network parameterizations common in deep RL [Wang et al., 2020].

5.2 NEURAL NETWORK ASSUMPTIONS

Assumption 5.4 (Neural Critic Approximation Error)**.** The worst-case critic approximation error

$$\epsilon_{\text{app}} = \sup_{\theta \in \Theta} \max_{g \in \{r,c\}} \mathbb{E}_{\nu^{\pi_\theta}} \left[\left(Q_g^{\pi_\theta}(s,a) - \Pi_{\mathcal{F}_{R,m}} Q_g^{\pi_\theta}(s,a) \right)^2 \right] \tag{38}$$

is finite, where $\Pi_{\mathcal{F}_{R,m}}$ denotes projection onto the linearized function class (15) and the expectation is taken under the state-action occupancy measure ν^{π_θ}.

Assumption 5.4 ensures that the linearized neural network class $\mathcal{F}_{R,m}$ can approximate the true Q-function $Q_g^{\pi_\theta}$ with bounded error. This is analogous to the realizability assumption in linear function approximation but adapted to the NTK regime.

6 THEORETICAL ANALYSIS

In this section, we provide the theoretical guarantees for PDNAC-NC. Our analysis relies on the Neural Tangent Kernel (NTK) regime, where the neural network is well-approximated by its linearization around initialization. We proceed in three steps: (1) analyzing the neural critic by bounding its deviation from the linearized updates and establishing convergence, (2) bounding the error of the Natural Policy Gradient (NPG) estimator, and (3) establishing the global convergence of the primal-dual updates.

Throughout this section, let $\mathcal{F}_k = \sigma\left(\{\theta_j, \lambda_j\}_{j \leq k}\right)$ denote the σ-algebra generated by the primal-dual iterates up to the start of epoch k. We write $\mathbb{E}_{\theta_k}[\cdot] := \mathbb{E}[\cdot \mid \theta_k]$ for the expectation conditioned on the policy parameter θ_k, taken over the randomness of the inner-loop Markovian samples and the MLMC draws within epoch k. Likewise, $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$ denotes the expectation conditioned on the entire history $\mathcal{F}_k$.

6.1 NEURAL CRITIC ANALYSIS

We first analyze the behavior of the neural critic parameters $\xi^k_{g,h}$ (Algorithm 1, Line 15). To facilitate the analysis, we introduce a reference sequence of *linearized updates* $\{\tilde{\xi}^k_{g,h}\}_{h \geq 0}$, defined as:

$$\tilde{\xi}^k_{g,h+1} = \Pi_{\mathcal{S}_R}\left(\tilde{\xi}^k_{g,h} - \gamma_\xi \hat{v}_g(\theta_k, \tilde{\xi}^k_{g,h}) \right), \; \tilde{\xi}^k_{g,0} = \xi^k_{g,0} = 0, \tag{39}$$

where $\hat{v}_g$ is the gradient using the linearized network approximation. The following lemma bounds the deviation between the actual neural updates and this linearized sequence.

Lemma 6.1 (Linearized Update Bounds)**.** *Suppose the assumptions in Section 5 hold. Let $\gamma_\xi = \frac{8 \log T}{\lambda H}$. Then, for any epoch k, the expected squared difference between the neural and linearized critic iterates satisfies:*

$$\begin{aligned} &\mathbb{E}_{\theta_k}\left[\|\tilde{\xi}^k_{g,h+1} - \xi^k_{g,h+1}\|^2 \right] \\ &\leq \tilde{\mathcal{O}}\left(\frac{R^2 \tau_{\text{mix}}}{\lambda T_{\max}} + \frac{R^2 \log(H/\delta) \log T_{\max}}{\lambda \sqrt{m}} \right) \end{aligned} \tag{40}$$

Proof (Sketch). This result (proven as Lemma 6.1 in the Appendix) leverages the projection $\Pi_{\mathcal{S}_R}$ to keep parameters in the NTK regime, where the Hessian of the network is small. By induction, we show that the accumulated linearization error (proportional to $m^{-1/2}$) and the difference in gradients due to MLMC noise remain bounded. The projection ensures stability, preventing the neural parameters from drifting too far from the initialization where the linear approximation is valid.

Next, we establish the convergence of the critic. Let $\xi^k_{g,*} = \left[\eta^k_{g,*}, (\zeta^k_{g,*})^\top\right]^\top$ denote the optimal critic parameter within the linearized function class $\mathcal{F}_{R,m}$, where $\eta^k_{g,*} = J_g(\theta_k)$ is the true average reward/cost and $\zeta^k_{g,*} = \arg\min_{\zeta \in \mathcal{S}_R} \mathcal{E}_g(\theta_k, \zeta)$ minimizes the critic objective (20) over the NTK ball, so that $\hat{Q}_g(\cdot\,; \zeta^k_{g,*}) = \Pi_{\mathcal{F}_{R,m}} Q_g^{\pi_{\theta_k}}$ is the best approximation of $Q_g^{\pi_{\theta_k}}$ in $\mathcal{F}_{R,m}$. We provide the second-order bound for the critic estimator.

Lemma 6.2 (Critic Second Order Bound)**.** *Under the same settings as Lemma 6.1, after H inner iterations, the critic estimator $\xi^k_{g,H}$ satisfies:*

$$\mathbb{E}_{\theta_k}\left[\|\xi^k_{g,H} - \xi^k_{g,*}\|^2\right] \leq \tilde{\mathcal{O}}\left(\frac{\|\xi^k_{g,0} - \xi^k_{g,*}\|^2}{T^2} + \frac{c_\gamma^2 \log(H/\delta)\tau_{\text{mix}}}{\lambda^2 H} + \frac{1}{\sqrt{m}}\right) \quad (41)$$

Proof (Sketch). Corresponds to Lemma C.10 in the Appendix. The proof relies on analyzing the contraction of the distance to the optimum $\xi^k_{g,*}$ under the linearized dynamics. Since the objective is strongly convex in the linearized regime (Lemma C.3 in the Appendix, which holds under the ergodicity Assumption 3.1), the error contracts geometrically. We then account for the variance of the stochastic MLMC gradients and the deviation from the true neural dynamics (bounded by Lemma 6.1). Unrolling the recursion yields the T^{-2} convergence rate for the initial error, plus steady-state error terms dominated by MLMC variance and linearization error.

6.2 NPG ESTIMATION ERROR

Using the critic bounds, we analyze the Natural Policy Gradient estimator ω^k_g. The estimator minimizes a quadratic objective whose gradient depends on the critic. We characterize its error in terms of the critic's error and MLMC sampling noise.

Theorem 6.3 (NPG Estimation Error)**.** *Suppose the assumptions in Section 5 hold. Let $\gamma_\omega = \frac{2\log T}{\mu H}$ and $\omega^k_{g,0} = 0$. Let $\omega^k_{g,*} = F(\theta_k)^{-1}\nabla_\theta J_g(\theta_k)$. Then:*

1. *The mean-squared error satisfies:*

$$\mathbb{E}_k\left[\|\omega^k_g - \omega^k_{g,*}\|^2\right] \leq \tilde{\mathcal{O}}\left(\frac{\tau^3_{\text{mix}}}{H} + \frac{1}{\sqrt{m}} + \tau^2_{\text{mix}}\epsilon_{\text{app}}\right) \quad (42)$$

2. *The squared bias satisfies:*

$$\|\mathbb{E}_k[\omega^k_g] - \omega^k_{g,*}\|^2 \leq \tilde{\mathcal{O}}\left(\frac{\log(H/\delta)}{H} + \frac{1}{\sqrt{m}} + \tau^2_{\text{mix}}\epsilon_{\text{app}}\right) \quad (43)$$

Proof (Sketch). See Theorem D.3 in the Appendix. The NPG estimator solves a quadratic problem using stochastic gradients estimated via MLMC. We first bound the bias and variance of these gradients (Lemma D.1), which depend on the quality of the critic ξ^k_g. The error in ω^k_g decomposes into optimization error (decaying with H), gradient variance (handled by MLMC), and gradient bias. The gradient bias is dominated by the critic's bias $\|\mathbb{E}[\xi^k_g] - \xi^k_{g,*}\|^2$ and the inherent function approximation error ϵ_{app}. Substituting the critic bound from Lemma 6.2 yields the final rates.

6.3 GLOBAL CONVERGENCE

Finally, we combine the NPG error bounds with the primal-dual analysis for CMDPs. The following theorem establishes the convergence rate of PDNAC-NC.

Theorem 6.4 (Global Convergence)**.** *Suppose the assumptions in Section 5 hold. Let the step sizes be $\alpha = \beta = \Theta(T^{-1/4})$, $K = \Theta(T^{1/2})$, $H = \Theta(T^{1/2})$, and $T_{\max} = T$. The sequence of policies generated by Algorithm 1 satisfies:*

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\left(J_r^{\pi^*} - J_r(\theta_k)\right)\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + T^{-1/4} + m^{-1/4}\right) \quad (44)$$

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1} -J_c(\theta_k)\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + T^{-1/4} + m^{-1/4}\right) \quad (45)$$

Proof (Sketch). See Theorem E.1 in the Appendix. The proof follows the standard primal-dual analysis for CMDPs but replaces the exact NPG update with our estimated ω^k_g. We bound the drift terms using the bias and variance of ω^k_g. The approximation errors ϵ_{bias} and ϵ_{app} appear as irreducible terms due to the limitations of the policy and critic function classes. The $m^{-1/4}$ term reflects the NTK linearization error, and taking $m = \Theta(T)$ yields an ϵ-optimal solution.

7 CONCLUSION

In this work, we provide the first global convergence guarantees for infinite-horizon average-reward Constrained MDPs (CMDPs) utilizing multi-layer neural network critics and general policy parameterizations. By leveraging Neural Tangent Kernel (NTK) theory, we successfully characterize the function approximation error of neural critics. Furthermore, by integrating a nested Multi-Level Monte Carlo (MLMC) estimation subroutine, our algorithm bypasses the need for sample-thinning techniques and mixing-time oracles, ensuring that the entirety of the collected Markovian trajectories is utilized. We establish a convergence rate of $\tilde{\mathcal{O}}(T^{-1/4})$ for both the optimality gap and constraint violation.

While our work establishes strong theoretical guarantees for neural actor-critic methods in CMDPs, the achieved rate of $\tilde{\mathcal{O}}(T^{-1/4})$ is not order-optimal compared to unconstrained natural actor-critic methods under Markovian sampling. The primary bottleneck in the constrained setting arises from the error of the critic, as prior work in the linear regime uses the sharper squared bias term [Ganesh et al., 2025]. However, applying this here is difficult due to the NTK projection operator. Overcoming this technical challenge to improve the bounds and achieve optimal rates remains an interesting open problem.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. URL http://jmlr.org/papers/v22/19-736.html.

Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Concave utility reinforcement learning with zero-constraint violations. *Transactions on Machine Learning Research*, 2022a.

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. PMLR, 2022b.

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.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

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.

Qinbo Bai, Washim U 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.

Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First order methods with markovian noise: from acceleration to variational inequalities. *Advances in Neural Information Processing Systems*, 36: 44820–44835, 2023.

Jose H Blanchet and Peter W Glynn. Unbiased monte carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE, 2015.

Semih Cayci, Niao He, and Rayadurgam Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

Semih Cayci, Niao He, and R Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*, 2024, 2024.

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.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 4(5):11, 2015.

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.

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.

Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*, 2020.

Swetha Ganesh and Vaneet Aggarwal. Regret analysis of average-reward unichain MDPs via an actor-critic approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=rl2aAyJfQz.

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 *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, pages 1358–1380, 2025.

Mudit Gaur, Amrit Singh 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 *Proceedings of the 41st International Conference on Machine Learning*, pages 15153–15179, 2024.

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.

Glebys Gonzalez, Mythra Balakuntala, Mridul Agarwal, Tomas Low, Bruce Knoth, Andrew W. Kirkpatrick, Jessica McKee, Gregory Hager, Vaneet Aggarwal, Yexiang Xue, Richard Voyles, and Juan Wachs. Asap: A semi-autonomous precise system for telesurgery during communication delays. *IEEE Transactions on Medical Robotics and Bionics*, 5(1):66–78, 2023. doi: 10.1109/TMRB.2023.3239674.

D Hendrycks. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Zhifa Ke, Zaiwen Wen, and Junyu Zhang. An improved finite-time analysis of temporal difference learning with deep neural networks. In *International Conference on Machine Learning*, pages 23407–23429. PMLR, 2024.

Sajad Khodadadian, Zaiwei Chen, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic algorithm. In *International Conference on Machine Learning*, pages 5420–5431. PMLR, 2021.

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.

Washim U Mondal and Vaneet Aggarwal. Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3097–3105. PMLR, 2024a.

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

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. PMLR, 2018.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

Anirudh Satheesh and Vaneet Aggarwal. Regret analysis of unichain average reward constrained mdps with general parameterization. *arXiv preprint arXiv:2602.08000*, 2026.

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.

Dipesh Tamboli, Jiayu Chen, Kiran Pranesh Jotheeswaran, Denny Yu, and Vaneet Aggarwal. Reinforced sequential decision-making for sepsis treatment: The posnegdm framework with mortality classifier and transformer. *IEEE Journal of Biomedical and Health Informatics*, 2024.

Haoxing Tian, Alex Olshevsky, and Yannis Paschalidis. Convergence of actor-critic with multi-layer neural networks. *Advances in neural information processing systems*, 36: 9279–9321, 2023.

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*, 2020.

Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4): 3868–3876, Jun. 2022. doi: 10.1609/aaai.v36i4.20302.

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, 2021.

Yang Xu, Swetha Ganesh, Washim Uddin Mondal, Qinbo Bai, and Vaneet Aggarwal. Global convergence for average reward constrained mdps with primal-dual actor critic algorithm. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

Global Convergence of Average Reward Constrained MDPs with Neural Critic and General Policy Parameterization (Supplementary Material)

Anirudh Satheesh[1] **Pankaj Kumar Barman**[2] **Washim Uddin Mondal**[2] **Vaneet Aggarwal**[3]

[1]Department of Computer Science, University of Maryland, College Park, MD 20740, USA
[2]Indian Institute of Technology, Kanpur, Uttar Pradesh, 208016, India
[3]Purdue University, West Lafayette, IN 47907, USA

A RELATED WORK

Constrained Markov Decision Processes (CMDPs) and Safe RL The CMDP framework is the standard mathematical formulation for safe reinforcement learning, originally formalized by Altman [2021]. Early solutions relied on linear programming, which struggled with the curse of dimensionality. Recently, primal-dual Policy Gradient (PG) and Natural Policy Gradient (NPG) methods have become the standard for high-dimensional CMDPs [Ding et al., 2020, Paternain et al., 2019]. While initial non-asymptotic analyses of these methods were limited to tabular settings or linear function approximation [Xu et al., 2021, Chen et al., 2022, Bai et al., 2023], recent works have extended them to general policy parameterizations. However, most of these guarantees are confined to the discounted reward setting.

Average-Reward Setting: While discounted reward MDPs have been extensively studied, infinite-horizon average-reward MDPs are notoriously more challenging due to the lack of contraction properties in their Bellman operators. Early theoretical advancements in average-reward CMDPs were restricted to tabular cases or required generative models. Chen et al. [2022], Agarwal et al. [2022b,a] provided convergence rates for average-reward CMDPs but relied on tabular policies. More recently, Bai et al. [2024], Xu et al. [2025] explored general parameterized policies for average-reward CMDPs via a primal-dual policy gradient algorithm, but their analysis relies heavily on access to exact gradients or limits the critic's function approximation capacity, leaving the gap for deep neural network critics unexplored.

Actor-Critic Methods and Neural Function Approximation: Actor-Critic (AC) algorithms are ubiquitous in practical RL, but their theoretical analysis has historically lagged behind their empirical success. Foundational analyses of AC algorithms established finite-time global convergence but primarily assumed linear critic approximations [Khodadadian et al., 2021, Cayci et al., 2021]. To bridge the gap between theory and deep RL practice, recent literature has adopted Neural Tangent Kernel (NTK) theory to analyze multi-layer neural network critics [Jacot et al., 2018]. Works such as Fu et al. [2020], Cayci et al. [2024] and Tian et al. [2023] integrated neural critics into AC methods but analyzed them strictly within standard, unconstrained MDPs. Even the most recent state-of-the-art analyses of neural actor-critic methods such as Gaur et al. [2024] and Ganesh et al. [2025] achieve sample complexities of $\tilde{\mathcal{O}}(\epsilon^{-3})$ and $\tilde{\mathcal{O}}(\epsilon^{-2})$ respectively, but are limited to the discounted, unconstrained setting. Our work fundamentally extends this line of inquiry by applying NTK analysis to the coupled primal-dual dynamics required for the constrained average-reward setting.

Markovian Sampling and Multi-Level Monte Carlo (MLMC): A primary bottleneck in the analysis of RL algorithms under Markovian sampling is the statistical dependence across sequential transitions. Traditional analyses mitigate this by utilizing data-dropping or sample-thinning techniques, where only one out of every τ_{mix} samples is used to construct gradient estimates [Gaur et al., 2024, Ganesh et al., 2025]. This approach not only wastes collected data but also necessitates the restrictive assumption of possessing a mixing-time oracle to know the exact value of the mixing time. To circumvent this, recent theoretical works have adapted Multi-Level Monte Carlo (MLMC) estimation techniques [Blanchet and Glynn, 2015] to RL [Beznosikov et al., 2023]. By drawing trajectory lengths from a geometric distribution, MLMC provides unbiased gradient estimates that systematically correct for Markovian bias without requiring sample thinning. While Xu et al. [2025] successfully applied MLMC to primal-dual natural policy gradients, their work was restricted to linear critics. Our algorithm is the first to integrate MLMC to simultaneously evaluate a neural network critic and estimate the NPG direction, thereby entirely removing the need for a mixing-time oracle in neural actor-critic methods.

B PRELIMINARY RESULTS FOR MLMC ESTIMATION

We begin by establishing fundamental properties of the Multi-Level Monte Carlo (MLMC) estimator that will be used throughout the analysis.

B.1 MARKOVIAN SAMPLING BOUNDS

The following lemma bounds the variance of sample averages under Markovian sampling.

Lemma B.1 (Lemma 1, Beznosikov et al. [2023])**.** *Consider a time-homogeneous, ergodic Markov chain $\{Z_t\}_{t\geq 0}$ with unique stationary distribution d_Z and mixing time τ_{mix}. Let $h(Z)$ be a function satisfying $\|h(Z_t) - \mathbb{E}_{d_Z}[h(Z)]\|^2 \leq \sigma^2$ for all $t \geq 0$. Then*

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{t=0}^{N-1} h(Z_t) - \mathbb{E}_{d_Z}[h(Z)]\right\|^2\right] \leq \frac{C_1\tau_{\mathrm{mix}}}{N}\sigma^2, \tag{46}$$

where $C_1 = 16(1 + 1/\ln^2 4)$ and the expectation is over trajectories from any initial distribution.

B.2 MLMC ESTIMATOR PROPERTIES

The MLMC estimator combines samples at multiple resolutions to achieve low bias with few samples.

Lemma B.2 (Lemma B.2, Xu et al. [2025])**.** *Consider a time-homogeneous, ergodic Markov chain $\{Z_t\}_{t\geq 0}$ with stationary distribution d_Z and mixing time τ_{mix}. Let $g(\cdot)$ satisfy:*

- $\|\mathbb{E}_{d_Z}[g(Z)] - \nabla F(x)\|^2 \leq \delta^2$ *(bias at stationarity)*
- $\|g(Z_t) - \mathbb{E}_{d_Z}[g(Z)]\|^2 \leq \sigma^2$ *for all $t \geq 0$ (bounded variance)*

Let $Q \sim \mathrm{Geom}(1/2)$ and define the MLMC estimator

$$g^{\mathrm{MLMC}} = g^0 + \begin{cases} 2^Q(g^Q - g^{Q-1}), & \text{if } 2^Q \leq T_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{47}$$

where $g^j = 2^{-j}\sum_{t=0}^{2^j-1} g(Z_t)$ for $j \in \{0, Q-1, Q\}$. Then:

$$\mathbb{E}[g^{\mathrm{MLMC}}] = \mathbb{E}[g^{\lfloor \log_2 T_{\max}\rfloor}] \tag{48}$$

$$\mathbb{E}[\|g^{\mathrm{MLMC}} - \nabla F(x)\|^2] \leq O(\sigma^2\tau_{\mathrm{mix}}\log^2 T_{\max} + \delta^2) \tag{49}$$

$$\|\mathbb{E}[g^{\mathrm{MLMC}}] - \nabla F(x)\|^2 \leq O(\sigma^2\tau_{\mathrm{mix}}T_{\max}^{-1} + \delta^2) \tag{50}$$

C NEURAL CRITIC ANALYSIS

We analyze the convergence of the neural critic estimator. Recall from Section 4.1 that the combined critic parameter is ξ_g^k. We define the linearized single-sample gradient of the critic objective at $\xi_{g,0}$ as

$$\hat{v}_g(\theta_k, \xi_{g,h}^k; z_{kh}^j) = \mathbf{A}_g(\theta_k; z_{kh}^j)\xi_{g,h}^k - \mathbf{b}_g(\theta_k; z_{kh}^j), \tag{51}$$

where the single-sample matrix and vector are defined as:

$$\mathbf{A}_g(\theta_k; z_{kh}^j) = \begin{bmatrix} c_\gamma & 0 \\ \psi_g^j & \psi_g^j\left(\psi_g^j - \psi_g^{j+1}\right)^\top \end{bmatrix}, \tag{52}$$

$$\mathbf{b}_g(\theta_k; z_{kh}^j) = \begin{bmatrix} c_\gamma g(s_{kh}^j, a_{kh}^j) \\ \left(g(s_{kh}^j, a_{kh}^j) - Q_{g,0}^j + Q_{g,0}^{j+1}\right)\psi_g^j \end{bmatrix}, \tag{53}$$

where $\psi_g^j := \nabla_\zeta Q_g(\phi(s_{kh}^j, a_{kh}^j); \zeta_{g,0})$, $\psi_g^{j+1} := \nabla_\zeta Q_g(\phi(s_{kh}^{j+1}, a'); \zeta_{g,0})$, $Q_{g,0}^j = Q(\phi(s_{kh}^j, a_{kh}^j); \zeta_{g,0})$, $Q_{g,0}^{j+1} = Q(\phi(s_{kh}^{j+1}, a'); \zeta_{g,0})$, and we use the zero-initialization for $\zeta_{g,0}$. We define the expected population matrices as:

$$\mathbf{A}_g(\theta_k) = \mathbb{E}_{(s,a)\sim\nu_g^{\pi_{\theta_k}}, s'\sim P(s,a), a'\sim\pi_{\theta_k}(\cdot|s')}\left[\mathbf{A}_g(\theta_k; z)\right], \tag{54}$$

$$\mathbf{b}_g(\theta_k) = \mathbb{E}_{(s,a)\sim\nu_g^{\pi_{\theta_k}}, s'\sim P(s,a), a'\sim\pi_{\theta_k}(\cdot|s')}\left[\mathbf{b}_g(\theta_k; z)\right]. \tag{55}$$

where z is the transition. For the following sections, we denote the above expectation as $\mathbb{E}_{\theta_k}$.

Lemma C.1 (NTK Critic Bounds)**.** *Fix an outer iteration index k. Let $\zeta_{g,h}^k \in \mathcal{S}_R \quad \forall h \in \{0, 1, \ldots, H\}$ and $R = O(\log T)$. Then, for all inner steps $h \in \{1, \ldots, H\}$, there exist positive constants $C_1, C_1', \{C_i\}_{i=2}^5$ such that with probability at least $1 - \delta - 2L\exp(-Cm)$, the following statements hold:*

$$\|\nabla_\zeta Q(\phi(s,a); \zeta_{g,h}^k)\| \le C_1, \quad |Q(\phi(s,a); \zeta_{g,h}^k)| \le C_1'\sqrt{\log(H/\delta)} \tag{56}$$

$$\|v_g(\theta_k, \zeta_{g,h}^k; z_{kh}^j) - \hat{v}_g(\theta_k, \zeta_{g,h}^k; z_{kh}^j)\| \le C_2 R m^{-1/2}\sqrt{\log(H/\delta)} \tag{57}$$

$$\langle v_g(\theta_k, \zeta_{g,h}^k; z_{kh}^j) - \hat{v}_g(\theta_k, \zeta_{g,h}^k; z_{kh}^j), \zeta_{g,h}^k - \zeta_g^*\rangle \le C_3 R m^{-1/2}\sqrt{\log(H/\delta)} \tag{58}$$

$$|\hat{Q}(\phi(s,a); \zeta_{g,h}^k) - Q(\phi(s,a); \zeta_{g,h}^k)| \le C_4 m^{-1/2}\sqrt{\log(H/\delta)} \tag{59}$$

$$\|\nabla_\zeta Q(\phi(s,a); \zeta_0) - \nabla_\zeta Q(\phi(s,a); \zeta_{g,h}^k)\| \le C_5 m^{-1/2}\sqrt{\log(H/\delta)} \tag{60}$$

Proof. Equation (56) follows from Lemma D.2 and D.3 in Ke et al. [2024], equations (57) and (58) follow from Lemma D.5, and equations (59) and (60) follow from Lemma D.4. □

Lemma C.2 (Bounds on Single-Sample Linearized Critic Matrices)**.** *Fix an outer iteration index k and an inner step $h \in \{0, \ldots, H\}$. Let $z = (s, a, s')$ denote a single transition with $a' \sim \pi_{\theta_k}(\cdot \mid s')$. Then, under the NTK bounds in Lemma C.1, we have:*

$$\|\mathbf{A}_g(\theta_k; z)\| \le c_\gamma + 3C_1^2, \tag{61}$$

$$\|\mathbf{b}_g(\theta_k; z)\| \le c_\gamma + 3C_1 C_1'\sqrt{\log(H/\delta)}, \tag{62}$$

$$\|\mathbf{A}_g(\theta_k; z) - \mathbf{A}_g(\theta_k)\|^2 \le \mathcal{O}\left(c_\gamma^2 + C_1^4\right), \tag{63}$$

$$\|\mathbf{b}_g(\theta_k; z) - \mathbf{b}_g(\theta_k)\|^2 \le \mathcal{O}\left(c_\gamma^2 + C_1^2 C_1'^2\log(H/\delta)\right) \tag{64}$$

Proof. The bound for $\mathbf{A}_g$ follows from the triangle inequality:

$$\begin{aligned}
\|\mathbf{A}_g(\theta_k; z)\| &\le c_\gamma + \|\nabla_\zeta Q_g(\phi(s,a); \zeta_{g,0})\| + \|\nabla_\zeta Q_g(\phi(s,a); \zeta_{g,0})\|\|\nabla_\zeta Q_g(\phi(s,a); \zeta_{g,0}) - \nabla_\zeta Q_g(\phi(s',a'); \zeta_{g,0})\| && (65)\\
&\le c_\gamma + C_1 + C_1(2C_1) && (66)\\
&\le c_\gamma + 3C_1^2 && (67)
\end{aligned}$$

Similarly for $\mathbf{b}_g$, utilizing the fact that $|g(s,a)| \le 1$:

$$\begin{aligned}
\|\mathbf{b}_g(\theta_k; z)\| &\le c_\gamma|g(s,a)| + |g(s,a) - Q(\phi(s,a); \zeta_{g,0}) + Q(\phi(s',a'); \zeta_{g,0})|\|\nabla_\zeta Q(\phi(s,a); \zeta_{g,0})\| && (68)\\
&\le c_\gamma + \left(1 + 2C_1'\sqrt{\log(H/\delta)}\right)C_1 && (69)\\
&\le c_\gamma + 3C_1 C_1'\sqrt{\log(H/\delta)} && (70)
\end{aligned}$$

For the squared bias bounds, we have

$$\begin{aligned}
\|\mathbf{A}_g(\theta_k; z) - \mathbf{A}_g(\theta_k)\|^2 &\le 2\|\mathbf{A}_g(\theta_k; z)\|^2 + 2\|\mathbf{A}_g(\theta_k)\|^2 && (71)\\
&\le 2\|\mathbf{A}_g(\theta_k; z)\|^2 + \mathbb{E}_{\theta_k}\left[2\|\mathbf{A}_g(\theta_k; z)\|^2\right] && (72)\\
&\le \mathcal{O}\left(c_\gamma^2 + C_1^4\right) && (73)
\end{aligned}$$

and

$$\|\mathbf{b}_g(\theta_k; z) - \mathbf{b}_g(\theta_k)\|^2 \le 2\|\mathbf{b}_g(\theta_k; z)\|^2 + 2\|\mathbf{b}_g(\theta_k)\|^2 \tag{74}$$

$$\leq 2\|\mathbf{b}_g(\theta_k; z)\|^2 + 2\mathbb{E}_{\theta_k}[\|\mathbf{b}_g(\theta_k; z)\|^2] \tag{75}$$

$$\leq \mathcal{O}\left(c_\gamma^2 + C_1^2 C_1^{'2} \log(H/\delta)\right) \tag{76}$$

□

Lemma C.3 (PSD of $\mathbf{A}_g(\theta_k)$)**.** *Fix k and let Assumption 3.1 hold. Then if $\xi_g \in \ker(\mathbf{A}_g(\theta_k))^\perp$, there exists a $\lambda > 0$ such that the following holds:*

$$\xi_g^\top \mathbf{A}_g(\theta_k)\xi_g \geq \frac{\lambda}{2}\|\xi_g\|_2^2 \tag{77}$$

Proof. We first observe $\ker(\mathbf{A}_g(\theta_k))$. By definition, for any $\xi_g \in \ker(\mathbf{A}_g(\theta_k))$, we have

$$\mathbf{A}_g(\theta_k)\xi_g = \mathbb{E}_{\theta_k}\begin{bmatrix} c_\gamma & 0 \\ \psi_g & \psi_g\left(\psi_g - \psi_g^{'}\right)^\top \end{bmatrix}\xi_g = \mathbf{0} \implies c_\gamma \eta_g = 0 \implies \eta_g = 0 \tag{78}$$

This yields

$$\mathbf{A}_g^\zeta(\theta_k)\zeta_g := \mathbb{E}_{\theta_k}\left[\psi_g\left(\psi_g - \psi_g^{'}\right)^\top \zeta_g\right] = \mathbf{0} \tag{79}$$

We then define

$$\mathcal{S}_e = \left\{\zeta_g \in \mathbb{R}^{m(n+(L-1)m)} \mid \psi_g(s,a)^\top \zeta_g = c,\ \forall (s,a) \in \mathrm{supp}(\nu_g^{\pi_{\theta_k}}), c \in \mathbb{R}\right\} \tag{80}$$

We wish to show that $\ker\left(\mathbf{A}_g(\theta_k)\right) = \mathcal{S}_e$. First, we choose an arbitrary $\zeta_g \in \mathcal{S}_e$. Then

$$\mathbf{A}_g^\zeta(\theta_k)\zeta_g = \mathbb{E}_{\theta_k}\left[\psi_g\left(\psi_g^\top \zeta_g - \psi_g^{'\top}\zeta_g\right)\right] \tag{81}$$

$$\overset{(a)}{=} \mathbb{E}_{\theta_k}\left[\psi_g(0)\right] \tag{82}$$

$$= 0 \tag{83}$$

where (a) holds by the definition of $\mathcal{S}_e$. Thus, $\zeta_g \in \ker(\mathbf{A}_g^\zeta(\theta_k))$ and $\mathcal{S}_e \subseteq \ker(\mathbf{A}_g^\zeta(\theta_k))$. Similarly, we choose an arbitrary $\zeta_g' \in \ker(\mathbf{A}_g^\zeta(\theta_k))$. By definition, this yields

$$\mathbf{A}_g^\zeta(\theta_k)\zeta_g' = \mathbf{0} \implies \zeta_g'^\top \mathbf{A}_g^\zeta(\theta_k)\zeta_g' = 0 \tag{84}$$

Expanding out this quadratic form yields

$$\zeta_g'^\top \mathbf{A}_g^\zeta(\theta_k)\zeta_g' = \zeta_g'^\top \mathbb{E}_{\theta_k}\left[\psi_g\left(\psi_g - \psi_g^{'}\right)^\top\right]\zeta_g' = \mathbb{E}_{\theta_k}\left[\left(\zeta_g'^\top \psi_g\right)\left(\psi_g^\top \zeta_g' - \psi_g^{'\top}\zeta_g'\right)\right] = \mathcal{E}(\psi_g^\top \zeta_g', \psi_g^\top \zeta_g') = 0 \tag{85}$$

where $\mathcal{E}(\psi_g^\top \zeta_g', \psi_g^\top \zeta_g')$ is the Dirichlet form. Since the Dirichlet form is 0 and our MDP is irreducible and aperiodic by ergodicity, $\psi_g^\top \zeta_g'$ is constant on the support of $\nu_g^{\pi_{\theta_k}}$. Thus, $\zeta_g' \in \mathcal{S}_e$ and $\ker(\mathbf{A}_g^\zeta(\theta_k)) \subseteq \mathcal{S}_e$. Since both $\mathcal{S}_e \subseteq \ker(\mathbf{A}_g^\zeta(\theta_k))$ and $\ker(\mathbf{A}_g^\zeta(\theta_k)) \subseteq \mathcal{S}_e$, $\ker(\mathbf{A}_g^\zeta(\theta_k)) = \mathcal{S}_e$ and $\ker(\mathbf{A}_g^\zeta(\theta_k))^\perp = \mathcal{S}_e^\perp$.

We have that the Dirichlet form is related to the variance through the spectral gap of the MDP λ^*:

$$\mathcal{E}\left(\psi_g^\top \zeta_g, \psi_g^\top \zeta_g\right) \geq \lambda^* \mathrm{Var}\left(\psi_g^\top \zeta_g\right) \tag{86}$$

Since we are given $\xi_g \in \ker\left(\mathbf{A}_g(\theta_k)\right)^\perp$, $\zeta_g \in \mathcal{S}_e^\perp$, and $\mathrm{Var}(\psi_g^\top \zeta_g) > 0$ on the support of $\nu_g^{\pi_{\theta_k}}$. We define

$$\lambda = \min_{\|\zeta_g\|_2^2 = 1, \zeta_g \in \mathcal{S}_e^\perp} \zeta_g^\top \mathbf{A}_g^\zeta(\theta_k)\zeta_g > 0 \tag{87}$$

which implies that $\zeta_g^\top \mathbf{A}_g^\zeta(\theta_k)\zeta_g > \lambda\|\zeta_g\|_2^2$. Finally, we have

$$\xi_g^\top \mathbf{A}_g(\theta_k)\xi_g = c_\gamma \eta_g^2 + \eta_g \zeta_g^\top \mathbb{E}_{\theta_k}\left[\psi_g\right] + \zeta_g^\top \mathbf{A}_g^\zeta(\theta_k)\zeta_g \tag{88}$$

$$\overset{(a)}{\geq} c_\gamma \eta_g^2 - C_1|\eta_g||\zeta_g| + \lambda\|\zeta_g\|_2^2 \tag{89}$$

$$\overset{(b)}{\geq} \frac{\lambda}{2}\|\xi_g\|_2^2 \tag{90}$$

where (a) uses Lemma C.1a, and (b) holds when $c_\gamma \geq \frac{\lambda}{2} + \frac{C_1^2}{2\lambda}$. □

Lemma C.4 (MLMC Linearized Critic Matrix Error Bounds)**.** *Fix an outer iteration index k and inner iteration $h \in \{0, \ldots, H\}$. Let $\mathbf{A}_{g,kh}^{\mathrm{MLMC}}$ and $\mathbf{b}_{g,kh}^{\mathrm{MLMC}}$ denote the MLMC estimates of the linearized critic matrices at step h. Under the NTK bounds in Lemma C.1 and the single-sample bounds in Lemma C.2, we have*

$$\left\|\mathbb{E}_{k,h}\left[\mathbf{A}_{g,kh}^{\mathrm{MLMC}}\right] - \mathbf{A}_g(\theta_k)\right\|^2 = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^4\right)\tau_{\mathrm{mix}} T_{\max}^{-1}\Big), \tag{91}$$

$$\mathbb{E}_{k,h}\left[\left\|\mathbf{A}_{g,kh}^{\mathrm{MLMC}} - \mathbf{A}_g(\theta_k)\right\|^2\right] = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^4\right)\tau_{\mathrm{mix}} \log T_{\max}\Big), \tag{92}$$

$$\left\|\mathbb{E}_{k,h}\left[\mathbf{b}_{g,kh}^{\mathrm{MLMC}}\right] - \mathbf{b}_g(\theta_k)\right\|^2 = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right)\tau_{\mathrm{mix}} T_{\max}^{-1}\Big), \tag{93}$$

$$\mathbb{E}_{k,h}\left[\left\|\mathbf{b}_{g,kh}^{\mathrm{MLMC}} - \mathbf{b}_g(\theta_k)\right\|^2\right] = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right)\tau_{\mathrm{mix}} \log T_{\max}\Big), \tag{94}$$

where $T_{\max}$ is the maximum rollout length and $\mathbb{E}_{k,h}$ denotes the conditional expectation at timestep k given the history of the current inner loop.

Proof. To find the MLMC error bounds, we first find the values of $\sigma_{\mathbf{A}}^2$, $\sigma_{\mathbf{b}}^2$, $\delta_{\mathbf{A}}^2$, $\delta_{\mathbf{b}}^2$, which denote the squared bias and variance terms for the single sample estimates. Since by definition $\mathbb{E}_{\theta_k}\left[\mathbf{A}_g(\theta_k; z)\right] = \mathbf{A}_g(\theta_k)$ and $\mathbb{E}_{\theta_k}\left[\mathbf{b}_g(\theta_k; z)\right] = \mathbf{b}_g(\theta_k)$, $\delta_{\mathbf{A}}^2 = \delta_{\mathbf{b}}^2 = 0$. Additionally, from Lemma C.2, we have

$$\mathbb{E}_{\theta_k}\left[\|\mathbf{A}_g(\theta_k; z) - \mathbf{A}_g(\theta_k)\|^2\right] \leq \mathcal{O}\left(c_\gamma^2 + C_1^4\right) \tag{95}$$

$$\mathbb{E}_{\theta_k}\left[\|\mathbf{b}_g(\theta_k; z) - \mathbf{b}_g(\theta_k)\|^2\right] \leq \mathcal{O}\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right) \tag{96}$$

Thus, by Lemma B.2, we have

$$\left\|\mathbb{E}_{k,h}\left[\mathbf{A}_{g,kh}^{\mathrm{MLMC}}\right] - \mathbf{A}_g(\theta_k)\right\|^2 = \mathcal{O}\left(\sigma_{\mathbf{A}}^2 \tau_{\mathrm{mix}} T_{\max}^{-1} + \delta_{\mathbf{A}}^2\right) = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^4\right)\tau_{\mathrm{mix}} T_{\max}^{-1}\Big), \tag{97}$$

$$\mathbb{E}_{k,h}\left[\left\|\mathbf{A}_{g,kh}^{\mathrm{MLMC}} - \mathbf{A}_g(\theta_k)\right\|^2\right] = \mathcal{O}\left(\sigma_{\mathbf{A}}^2 \tau_{\mathrm{mix}} \log T_{\max} + \delta_{\mathbf{A}}^2\right) = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^4\right)\tau_{\mathrm{mix}} \log T_{\max}\Big), \tag{98}$$

$$\left\|\mathbb{E}_{k,h}\left[\mathbf{b}_{g,kh}^{\mathrm{MLMC}}\right] - \mathbf{b}_g(\theta_k)\right\|^2 = \mathcal{O}\left(\sigma_{\mathbf{b}}^2 \tau_{\mathrm{mix}} T_{\max}^{-1} + \delta_{\mathbf{b}}^2\right) = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right)\tau_{\mathrm{mix}} T_{\max}^{-1}\Big), \tag{99}$$

$$\mathbb{E}_{k,h}\left[\left\|\mathbf{b}_{g,kh}^{\mathrm{MLMC}} - \mathbf{b}_g(\theta_k)\right\|^2\right] = \mathcal{O}\left(\sigma_{\mathbf{b}}^2 \tau_{\mathrm{mix}} \log T_{\max} + \delta_{\mathbf{b}}^2\right) = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right)\tau_{\mathrm{mix}} \log T_{\max}\Big), \tag{100}$$

□

We also provide the following bounds on the original MLMC gradient estimate.

Lemma C.5 (MLMC Critic Parameter Error Bounds)**.** *Assume the same setting as Lemma C.4. Let $v_g(\theta_k, \xi_{g,h}^k) = \mathbb{E}_{\theta_k}\Big[v_g(\theta_k, \xi_{g,h}^k; z)\Big]$. Then we have the following bounds:*

$$\|\mathbb{E}_{k,h}\left[v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\right] - v_g(\theta_k, \xi_{g,h}^k)\|^2 \leq \mathcal{O}\Big(\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right)\tau_{\mathrm{mix}} T_{\max}^{-1}\Big) \tag{101}$$

$$\mathbb{E}_{k,h}\left\|v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - v_g(\theta_k, \xi_{g,h}^k)\right\|^2 \leq \mathcal{O}\Big(\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right)\tau_{\mathrm{mix}} \log T_{\max}\Big) \tag{102}$$

Proof. By the definition of the single sample combined gradient estimate, we have

$$\|v_g(\theta_k, \xi_{g,h}^k; z)\| \leq c_\gamma \|\hat{\nabla}_\eta \mathcal{R}(\theta_k, \eta_{g,h}^k; z)\| + \|\hat{\nabla}_\zeta \mathcal{E}(\theta_k, \zeta_{g,h}^k; z)\| \tag{103}$$

$$\begin{aligned} &= c_\gamma \|\eta_{g,h}^k - g(s_{kh}^j, a_{kh}^j)\| \\ &\quad + \|Q_g(\phi(s,a); \zeta_{g,h}^k) + \eta_{g,h}^k - g(s,a) - Q_g(\phi(s',a'); \zeta_{g,h}^k)\|\|\nabla_\zeta Q_g(\phi(s,a); \zeta_{g,h}^k)\| \end{aligned} \tag{104}$$

$$\leq \mathcal{O}\left(c_\gamma + C_1 C_1' \sqrt{\log(H/\delta)}\right) \tag{105}$$

Additionally, we have

$$\sigma_v^2 = \|v_g(\theta_k, \xi_{g,h}^k; z) - v_g(\theta_k, \xi_{g,h}^k)\|^2 \leq 2\|v_g(\theta_k, \xi_{g,h}^k; z)\|^2 + 2\|v_g(\theta_k, \xi_{g,h}^k)\|^2 \tag{106}$$

$$\leq 2\|v_g(\theta_k, \xi_{g,h}^k; z)\|^2 + \mathbb{E}_{\theta_k}\left[2\|v_g(\theta_k, \xi_{g,h}^k)\|^2\right] \tag{107}$$

$$\leq \mathcal{O}\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right) \tag{108}$$

and

$$\delta_v^2 = \|\mathbb{E}_{\theta_k}\left[v_g(\theta_k, \xi_{g,h}^k; z)\right] - v_g(\theta_k, \xi_{g,h}^k)\| = 0 \tag{109}$$

Then by Lemma B.2, we have

$$\|\mathbb{E}_{k,h}\left[v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\right] - v_g(\theta_k, \xi_{g,h}^k)\|^2 \leq \mathcal{O}\left(\sigma_v^2 \tau_{\mathrm{mix}} T_{\max}^{-1} + \delta_v^2\right) = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right) \tau_{\mathrm{mix}} T_{\max}^{-1}\Big)$$

$$\mathbb{E}_{k,h}\left\|v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - v_g(\theta_k, \xi_{g,h}^k)\right\|^2 \leq \mathcal{O}\left(\sigma_v^2 \tau_{\mathrm{mix}} \log T_{\max} + \delta_v^2\right) = \mathcal{O}\Big(\left(c_\gamma^2 + C_1^2 C_1'^2 \log(H/\delta)\right) \tau_{\mathrm{mix}} \log T_{\max}\Big)$$

□

Lemma C.6 (First Order Linearization Error of the MLMC Critic Estimator)**.** *Let the MLMC estimator use truncation level* $T_{\max}$*. Then*

$$\mathbb{E}_{k,h}\left[\left\|v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - \hat{v}_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\right\|\right] = \mathcal{O}\left(C_2 R m^{-1/2} \sqrt{\log(H/\delta)} \log T_{\max}\right).$$

Proof. We already have from Lemma C.1 that $\|v_g(\theta_k, \xi_{g,h}^k; z_{kh}^j) - \hat{v}_g(\theta_k, \xi_{g,h}^k; z_{kh}^j)\| \leq C_2 R m^{-1/2} \sqrt{\log(H/\delta)}$ from equation (57). Then from the definition of the MLMC estimator, we have

$$\mathbb{E}_{k,h}\left[\|v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - \hat{v}_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\|\right] \tag{110}$$

$$= \mathbb{E}_{k,h}\left[\mathbb{E}_{P_h^k}\left[\left\|v_{g,kh}^0 - \hat{v}_{g,kh}^0 + \begin{cases} 2^{P_h^k}\left(v_{g,kh}^{P_h^k} - v_{g,kh}^{P_h^k - 1} - \hat{v}_{g,kh}^{P_h^k} + \hat{v}_{g,kh}^{P_h^k - 1}\right) & \text{if } 2^{P_h^k} \leq T_{\max} \\ 0 & \text{otherwise} \end{cases}\right\| \middle| k, h\right]\right] \tag{111}$$

By definition, we have

$$\left\|v_{g,kh}^j - \hat{v}_{g,kh}^j\right\| = \left\|\frac{1}{2^j}\sum_{t=0}^{2^j-1}\left(v_g(\theta_k, \xi_{g,h}^k; z_{kh}^t) - \hat{v}_g(\theta_k, \xi_{g,h}^k; z_{kh}^t)\right)\right\| \tag{112}$$

$$\leq \frac{1}{2^j}\sum_{t=0}^{2^j-1}\left\|v_g(\theta_k, \xi_{g,h}^k; z_{kh}^t) - \hat{v}_g(\theta_k, \xi_{g,h}^k; z_{kh}^t)\right\| \tag{113}$$

$$\overset{(a)}{\leq} \frac{1}{2^j}\sum_{t=0}^{2^j-1} C_2 R m^{-1/2}\sqrt{\log(H/\delta)} \tag{114}$$

$$= C_2 R m^{-1/2}\sqrt{\log(H/\delta)} \tag{115}$$

where (a) uses equation (57). Thus, we have

$$\mathbb{E}_{k,h}\left[\|v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - \hat{v}_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\|\right] \tag{116}$$

$$\leq \mathbb{E}_{k,h}\left[\mathbb{E}_{P_h^k}\left[C_2 R m^{-1/2}\sqrt{\log(H/\delta)} + 2^{P_h^k+1}\left(C_2 R m^{-1/2}\sqrt{\log(H/\delta)}\right)\mathbb{I}\left(2^{P_h^k} \leq T_{\max}\right)\middle| k, h\right]\right] \tag{117}$$

$$\overset{(a)}{\leq} \mathbb{E}_{k,h}\left[C_2 R m^{-1/2}\sqrt{\log(H/\delta)} + \sum_{p=0}^{\lfloor \log_2 T_{\max}\rfloor} 2^p\left(2C_2 R m^{-1/2}\sqrt{\log(H/\delta)}\right)\frac{1}{2}\frac{1}{2^{p-1}}\right] \tag{118}$$

$$= \mathbb{E}_{k,h}\left[C_2 R m^{-1/2}\sqrt{\log(H/\delta)} + \sum_{p=0}^{\lfloor \log_2 T_{\max}\rfloor}\left(2C_2 R m^{-1/2}\sqrt{\log(H/\delta)}\right)\right] \tag{119}$$

$$= \mathcal{O}\left(\log_2(T_{\max}) C_2 R m^{-1/2}\sqrt{\log(H/\delta)}\right) \tag{120}$$

where (a) uses the definition of the expectation for the truncated geometric distribution. □

Lemma C.7 (Second Order Linearization Error of the MLMC Critic Estimator)**.** *Assume the same setting as Lemma C.6. Then*

$$\mathbb{E}_{k,h}\left[\|v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - \hat{v}_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\|^2\right] \leq \mathcal{O}\left(C_2^2 R^2 m^{-1}\log(H/\delta) T_{\max}\right) \tag{121}$$

Proof. We have

$$\mathbb{E}_{k,h}\left[\|v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - \hat{v}_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\|^2\right] \tag{122}$$

$$= \mathbb{E}_{k,h}\left[\mathbb{E}_{P_h^k}\left[\left\|v_{g,kh}^0 - \hat{v}_{g,kh}^0 + 2^{P_h^k}\left(v_{g,kh}^{P_h^k} - v_{g,kh}^{P_h^k-1} - \hat{v}_{g,kh}^{P_h^k} + \hat{v}_{g,kh}^{P_h^k-1}\right)\mathbb{I}\left(2^{P_h^k} \leq T_{\max}\right)\right\|^2 \middle| k, h\right]\right] \tag{123}$$

$$\leq 2\mathbb{E}_{k,h}\left[\|v_{g,kh}^0 - \hat{v}_{g,kh}^0\|^2\right] \tag{124}$$

$$+ 2\mathbb{E}_{k,h}\left[\mathbb{E}_{P_h^k}\left[\left\|2^{P_h^k}\left(v_{g,kh}^{P_h^k} - v_{g,kh}^{P_h^k-1} - \hat{v}_{g,kh}^{P_h^k} + \hat{v}_{g,kh}^{P_h^k-1}\right)\mathbb{I}\left(2^{P_h^k} \leq T_{\max}\right)\right\|^2 \middle| k, h\right]\right] \tag{125}$$

$$\overset{(a)}{\leq} 2C_2^2 R^2 m^{-1}\log(H/\delta) + 4C_2^2 R^2 m^{-1}\log(H/\delta)\mathbb{E}_{k,h}\left[\mathbb{E}_{P_h^k}\left[2^{2P_h^k}\mathbb{I}\left(2^{P_h^k} \leq T_{\max}\right) \middle| k, h\right]\right] \tag{126}$$

$$= 2C_2^2 R^2 m^{-1}\log(H/\delta) + 4C_2^2 R^2 m^{-1}\log(H/\delta)\sum_{p=0}^{\lfloor \log_2 T_{\max}\rfloor} 2^{2p}\frac{1}{2^{p+1}} \tag{127}$$

$$= \mathcal{O}\left(C_2^2 R^2 m^{-1}\log(H/\delta) T_{\max}\right) \tag{128}$$

where (a) follows from Lemma C.6 □

Now that we have bounds for the estimation error for the MLMC gradient and its parameters, we can find the bound between the linearized and original updates.

Lemma C.8 (Linearized Update Bounds)**.** *We define the linearized sequence* $\{\tilde{\xi}_{g,h}^k\}_{h\geq 0}$ *that represents the linearized iterates of the neural update* $\hat{v}_g$*:*

$$\tilde{\xi}_{g,h+1}^k = \Pi_{\mathcal{S}_R}\left(\tilde{\xi}_{g,h}^k - \gamma_\xi \hat{v}_g(\theta_k, \tilde{\xi}_{g,h}^k)\right) \quad \tilde{\xi}_{g,0}^k = \xi_{g,0}^k = 0 \tag{129}$$

Let $\Pi_\perp$ *be the orthogonal projection onto* $\ker\left(\mathbf{A}_g(\theta_k)\right)$*. Then the following result holds:*

$$\mathbb{E}_{\theta_k}\left\|\tilde{\xi}_{g,h+1}^k - \xi_{g,h+1}^k\right\|^2 \leq \mathcal{O}\left(\frac{R^2(c_\gamma^2 + C_1^4)\tau_{\mathrm{mix}}\log T}{\lambda T_{\max}} + \frac{R^2\sqrt{\log(H/\delta)}\log T_{\max}\log T}{\lambda m^{1/2}} + \frac{R^2 T_{\max}\log(H/\delta)\log T}{H\lambda m}\right) \tag{130}$$

Proof. Using the non-expansiveness of $\Pi_{\mathcal{S}_R}$ and Lemma C.1, we have

$$\mathbb{E}_{k,h}\left\|\tilde{\xi}_{g,h+1}^k - \xi_{g,h+1}^k\right\|^2 = \mathbb{E}_{k,h}\left\|\Pi_{\mathcal{S}_R}\left(\xi_{g,h}^k - \gamma_\xi v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\right) - \Pi_{\mathcal{S}_R}\left(\tilde{\xi}_{g,h}^k - \gamma_\xi \hat{v}_g^{\mathrm{MLMC}}\left(\theta_k, \tilde{\xi}_{g,h}^k\right)\right)\right\|^2$$

$$\leq \mathbb{E}_{k,h}\left\|\xi_{g,h}^k - \gamma_\xi v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - \left(\tilde{\xi}_{g,h}^k - \gamma_\xi \hat{v}_g^{\mathrm{MLMC}}\left(\theta_k, \tilde{\xi}_{g,h}^k\right)\right)\right\|^2$$

$$\leq \mathbb{E}_{k,h}\left\|\left(\xi_{g,h}^k - \tilde{\xi}_{g,h}^k\right) - \gamma_\xi\left(\mathbf{A}_{g,kh}^{\mathrm{MLMC}}(\theta_k)\left(\xi_{g,h}^k - \tilde{\xi}_{g,h}^k\right) + v_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k) - \hat{v}_g^{\mathrm{MLMC}}(\theta_k, \xi_{g,h}^k)\right)\right\|^2$$

$$
\begin{aligned}
&\leq \mathbb{E}_{k,h}\left\|\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right\|^2-2\gamma_\xi\mathbb{E}_{k,h}\left\langle\xi_{g,h}^k-\tilde{\xi}_{g,h}^k,\mathbf{A}_{g,kh}^{\mathrm{MLMC}}(\theta_k)\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)\right\rangle\\
&\quad-2\gamma_\xi\mathbb{E}_{k,h}\left\langle\xi_{g,h}^k-\tilde{\xi}_{g,h}^k,v_g^{\mathrm{MLMC}}(\theta_k,\xi_{g,h}^k)-\hat{v}_g^{\mathrm{MLMC}}(\theta_k,\xi_{g,h}^k)\right\rangle+\mathcal{O}\left(\gamma_\xi^2C_2^2R^2m^{-1}\log(H/\delta)T_{\max}\right)\\
&\quad+2\gamma_\xi^2\mathbb{E}_{k,h}\left\|\mathbf{A}_{g,kh}^{\mathrm{MLMC}}(\theta_k)\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)\right\|^2\\
&\leq \mathbb{E}_{k,h}\left\|\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right\|^2-2\gamma_\xi\mathbb{E}_{k,h}\left\langle\xi_{g,h}^k-\tilde{\xi}_{g,h}^k,\mathbf{A}_{g,kh}^{\mathrm{MLMC}}(\theta_k)\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)\right\rangle\\
&\quad+\mathcal{O}\left(\gamma_\xi R^2\log T_{\max}m^{-1/2}\sqrt{\log(H/\delta)}\right)+\mathcal{O}\left(\gamma_\xi^2C_2^2R^2m^{-1}\log(H/\delta)T_{\max}\right)\\
&\quad+4\gamma_\xi^2\mathbb{E}_{k,h}\left\|\mathbf{A}_g(\theta_k)\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)\right\|^2+4\gamma_\xi^2\mathbb{E}_{k,h}\left\|\mathbf{A}_{g,kh}^{\mathrm{MLMC}}(\theta_k)\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)-\mathbf{A}_g(\theta_k)\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)\right\|^2\\
&\leq \mathbb{E}_{k,h}\left\|\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right\|^2-2\gamma_\xi\mathbb{E}_{k,h}\left\langle\Pi_\perp\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right),\mathbf{A}_g(\theta_k)\Pi_\perp\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)\right\rangle+\mathcal{O}\left(\frac{\gamma_\xi R^2(c_\gamma^2+C_1^4)\tau_{\mathrm{mix}}}{T_{\max}}\right)\\
&\quad+\mathcal{O}\left(\gamma_\xi R^2\log T_{\max}m^{-1/2}\sqrt{\log(H/\delta)}\right)+\mathcal{O}\left(\gamma_\xi^2C_2^2R^2m^{-1}\log(H/\delta)T_{\max}\right)\\
&\quad+\mathcal{O}\left(\gamma_\xi^2\left(c_\gamma^2+C_1^4\right)\right)\mathbb{E}_{k,h}\left\|\Pi_\perp\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)\right\|^2+\mathcal{O}\left(\gamma_\xi^2\left(c_\gamma^2+C_1^4\right)\tau_{\mathrm{mix}}\log T_{\max}\right)\left\|\Pi_\perp\left(\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right)\right\|^2\\
&\leq \mathbb{E}_{k,h}\left\|\xi_{g,h}^k-\tilde{\xi}_{g,h}^k\right\|^2+\mathcal{O}\left(\frac{\gamma_\xi R^2(c_\gamma^2+C_1^4)\tau_{\mathrm{mix}}}{T_{\max}}\right)+\mathcal{O}\left(\gamma_\xi R^2\log T_{\max}m^{-1/2}\sqrt{\log(H/\delta)}\right)\\
&\quad+\mathcal{O}\left(\gamma_\xi^2C_2^2R^2m^{-1}\log(H/\delta)T_{\max}\right)\\
&\leq \mathcal{O}\left(\frac{(h+1)\gamma_\xi R^2(c_\gamma^2+C_1^4)\tau_{\mathrm{mix}}}{T_{\max}}\right)+\mathcal{O}\left((h+1)\gamma_\xi R^2m^{-1/2}\sqrt{\log(H/\delta)}\log T_{\max}\right)\\
&\quad+\mathcal{O}\left((h+1)\gamma_\xi^2C_2^2R^2m^{-1}\log(H/\delta)T_{\max}\right)
\end{aligned}
$$

If we substitute $\gamma_\xi=\frac{8\log T}{\lambda H}$ and take expectations on both sides, we have

$$
\mathbb{E}_{\theta_k}\left\|\tilde{\xi}_{g,h+1}^k-\xi_{g,h+1}^k\right\|^2\leq\mathcal{O}\left(\frac{R^2(c_\gamma^2+C_1^4)\tau_{\mathrm{mix}}\log T}{\lambda T_{\max}}+\frac{R^2\sqrt{\log(H/\delta)}\log T_{\max}\log T}{\lambda m^{1/2}}+\frac{R^2T_{\max}\log(H/\delta)\log T}{H\lambda m}\right)
$$

□

Before proceeding further with the neural critic analysis, we provide the following Lemma from Ganesh et al. [2025], adopted to the CMDP setting.

Lemma C.9 (Lemma 9, Ganesh et al. [2025]). *Let $Z_k:=\{z\in\mathbb{R}^d:z=\mathbf{A}_g(\theta_k)^\dagger\mathbf{b}_g(\theta_k)+v,\,v\in\ker(\mathbf{A}_g(\theta_k))\}$ denote the set of minimum-norm least-squares solutions to $\mathbf{A}_g(\theta_k)z\approx\mathbf{b}_g(\theta_k)$, and let $\xi_{g,*}^k$ be the projection of a fixed point ξ_0 onto Z_k. Then, under the update rule*

$$
\tilde{\xi}_{g,h}^k=\Pi_{\mathcal{S}_R}\left(\tilde{\xi}_{g,h-1}^k-\gamma_\xi\hat{v}_g(\theta_k;\tilde{\xi}_{g,h-1}^k)\right),
$$

where $\hat{v}_g(\theta_k;\tilde{\xi}_{g,h}^k)\in\ker(\mathbf{A}_g(\theta_k))^\perp$, it holds that

$$
\tilde{\xi}_{g,h}^k-\xi_{g,*}^k\in\ker(\mathbf{A}_g(\theta_k))^\perp,\quad\textit{for all } h\geq 0.
$$

Next, we derive the second order bound.

Lemma C.10 (Critic Second Order Bound). *Assume the same setting as Lemma C.8. Then, after H inner critic updates, the critic iterate $\xi_{g,H}^k$ satisfies*

$$
\begin{aligned}
\mathbb{E}_{\theta_k}\left[\|\xi_{g,H}^k-\xi_{g,*}^k\|^2\right]\leq\mathcal{O}\Bigg(&\frac{\mathbb{E}_{\theta_k}\|\tilde{\xi}_{g,0}^k-\xi_{g,*}^k\|^2}{T^2}+\frac{c_\gamma^4\tau_{\mathrm{mix}}\log(H/\delta)R^2\log T}{\lambda^4T_{\max}}+\frac{R^2\sqrt{\log(H/\delta)}\log T\log T_{\max}}{\lambda m^{-1/2}}\\
&+\frac{R^2T_{\max}\log(H/\delta)\log T}{H\lambda m}+\frac{c_\gamma^2\log(H/\delta)\tau_{\mathrm{mix}}\log T_{\max}\log T}{\lambda^2H}\Bigg)
\end{aligned}
\tag{131}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_{k,h}\left\|\tilde{\xi}^k_{g,h+1}-\xi^k_{g,*}\right\|^2=\mathbb{E}_{k,h}\|\Pi_{\mathcal{S}_R}\left(\tilde{\xi}^k_{g,h}-\gamma_\xi v_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})\right)-\Pi_{\mathcal{S}_R}(\xi^k_{g,*})\|^2\\
&\le \mathbb{E}_{k,h}\|\tilde{\xi}^k_{g,h}-\gamma_\xi v_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})-\xi^k_{g,*}\|^2\\
&=\mathbb{E}_{k,h}\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2-2\gamma_\xi\mathbb{E}_{k,h}\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},v_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})\right\rangle+\gamma_\xi^2\mathbb{E}_{k,h}\|v_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})\|^2\\
&=\mathbb{E}_{k,h}\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2-2\gamma_\xi\mathbb{E}_{k,h}\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},\hat{v}_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})\right\rangle+\gamma_\xi^2\mathbb{E}_{k,h}\|v_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})\|^2\\
&\quad-2\gamma_\xi\mathbb{E}_{k,h}\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},v_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})-\hat{v}_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})\right\rangle\\
&\le\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2-2\gamma_\xi\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},\hat{v}_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})\right\rangle-2\gamma_\xi\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},v_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})-\hat{v}_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})\right\rangle\\
&\quad+2\gamma_\xi^2\mathbb{E}_{k,h}\|v_g(\theta_k;\tilde{\xi}^k_{g,h})\|^2+2\gamma_\xi^2\mathbb{E}_{k,h}\left\|v_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})-v_g(\theta_k;\tilde{\xi}^k_{g,h})\right\|^2\\
&\le\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2-2\gamma_\xi\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},\mathbf{A}_g(\theta_k)(\tilde{\xi}^k_{g,h}-\xi^k_{g,*})\right\rangle+\mathcal{O}\left(\gamma_\xi^2\left(c_\gamma^2+C_1'^2C_1^2\right)\log(H/\delta)\tau_{\mathrm{mix}}\log T_{\max}\right)\\
&\quad-2\gamma_\xi\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},\hat{v}_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})-\mathbf{A}_g(\theta_k)(\tilde{\xi}^k_{g,h}-\xi^k_{g,*})\right\rangle+\mathcal{O}\left(\gamma_\xi\log(T_{\max})C_2R^2m^{-1/2}\sqrt{\log(H/\delta)}\right)\\
&\overset{(a)}{\le}\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2-\gamma_\xi\lambda\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|-2\gamma_\xi\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},v_g(\theta_k;\tilde{\xi}^k_{g,h})-\mathbf{A}_g(\theta_k)(\tilde{\xi}^k_{g,h}-\xi^k_{g,*})\right\rangle\\
&\quad+\mathcal{O}\left(\gamma_\xi\log(T_{\max})C_2R^2m^{-1/2}\sqrt{\log(H/\delta)}\right)+\mathcal{O}\left(\gamma_\xi^2\left(c_\gamma^2+C_1'^2C_1^2\right)\log(H/\delta)\tau_{\mathrm{mix}}\log T_{\max}\right)
\end{aligned}
$$

where we use Lemma C.1 to bound $v_g(\theta_k,\xi^k_{g,h})$ and (a) follows from Lemma C.3. We take the conditional expectation $\mathbb{E}_{k,h}$

$$
\begin{aligned}
\mathbb{E}_{k,h}\left\|\tilde{\xi}^k_{g,h+1}-\xi^k_{g,*}\right\|^2&\le(1-\gamma_\xi\lambda)\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2-2\gamma_\xi\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},\mathbb{E}_{k,h}\left[\hat{v}_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})-\mathbf{A}_g(\theta_k)(\tilde{\xi}^k_{g,h}-\xi^k_{g,*})\right]\right\rangle\\
&\quad+\mathcal{O}\left(\gamma_\xi\log(T_{\max})C_2R^2m^{-1/2}\sqrt{\log(H/\delta)}\right)+\mathcal{O}\left(\gamma_\xi^2\left(c_\gamma^2+C_1'^2C_1^2\right)\log(H/\delta)\tau_{\mathrm{mix}}\log T_{\max}\right)
\end{aligned}\tag{132}
$$

We bound the second term as

$$
\begin{aligned}
&-2\gamma_\xi\left\langle\tilde{\xi}^k_{g,h}-\xi^k_{g,*},\mathbb{E}_{k,h}\left[\hat{v}_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})-\mathbf{A}_g(\theta_k)(\tilde{\xi}^k_{g,h}-\xi^k_{g,*})\right]\right\rangle && (133)\\
&\le\frac{\gamma_\xi\lambda}{2}\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2+\frac{2\gamma_\xi}{\lambda}\left\|\mathbb{E}_{k,h}\left[\hat{v}_g^{\mathrm{MLMC}}(\theta_k;\tilde{\xi}^k_{g,h})-\mathbf{A}_g(\theta_k)\left(\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\right)\right]\right\|^2 && (134)\\
&\le\frac{\gamma_\xi\lambda}{2}\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2+\frac{2\gamma_\xi}{\lambda}\left\|\left(\mathbb{E}_{k,h}\left[\mathbf{A}_{g,kh}^{\mathrm{MLMC}}\right]-\mathbf{A}_g(\theta_k)\right)\tilde{\xi}^k_{g,h}-\left(\mathbf{b}_g(\theta_k)-\mathbb{E}_{k,h}\left[\mathbf{b}_{g,kh}^{\mathrm{MLMC}}\right]\right)\right\|^2 && (135)\\
&\le\frac{\gamma_\xi\lambda}{2}\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2+\frac{4\gamma_\xi\Delta_{\mathbf{A}}^2\|\tilde{\xi}^k_{g,h}\|^2+4\gamma_\xi\Delta_{\mathbf{b}}^2}{\lambda} && (136)\\
&\le\frac{\gamma_\xi\lambda}{2}\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2+\frac{8\gamma_\xi\Delta_{\mathbf{A}}^2\left(\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|+4\lambda^{-2}\Lambda_{\mathbf{b}}^2\right)+4\gamma_\xi\Delta_b^2}{\lambda} && (137)
\end{aligned}
$$

where $\Delta_{\mathbf{A}}^2=\left\|\mathbb{E}_{k,h}\left[\mathbf{A}_{g,kh}^{\mathrm{MLMC}}\right]-\mathbf{A}_g(\theta_k)\right\|^2$, $\Delta_{\mathbf{b}}^2=\left\|\mathbb{E}_{k,h}\left[\mathbf{b}_{g,kh}^{\mathrm{MLMC}}\right]-\mathbf{b}_g(\theta_k)\right\|^2$, and $\Lambda_{\mathbf{b}}^2=\|\mathbf{b}_{g,kh}^{\mathrm{MLMC}}\|^2$. The last inequality is due to $\|\xi^k_*\|^2=\|\mathbf{A}_g(\theta_k)^{-1}\mathbf{b}_g(\theta_k)\|^2\le\frac{4\Lambda_{\mathbf{b}}^2}{\lambda^2}$. We substitute this back to yield

$$
\begin{aligned}
\mathbb{E}_{k,h}\left\|\tilde{\xi}^k_{g,h+1}-\xi^k_{g,*}\right\|^2&\le\left(1-\frac{\gamma_\xi\lambda}{2}+\frac{8\gamma_\xi\Delta_{\mathbf{A}}^2}{\lambda}\right)\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2+\frac{4\gamma_\xi}{\lambda}\left(8\lambda^{-2}\Lambda_{\mathbf{b}}^2\Delta_{\mathbf{A}}^2+\Delta_{\mathbf{b}}^2\right) && (138)\\
&\quad+\mathcal{O}\left(\gamma_\xi\log(T_{\max})C_2R^2m^{-1/2}\sqrt{\log(H/\delta)}\right)+\mathcal{O}\left(\gamma_\xi^2\left(c_\gamma^2+C_1'^2C_1^2\right)\log(H/\delta)\tau_{\mathrm{mix}}\log T_{\max}\right) && (139)
\end{aligned}
$$

For $\Delta_{\mathbf{A}}^2\le\frac{\lambda^2}{32}$, we have

$$
\begin{aligned}
\mathbb{E}_{k,h}\left\|\tilde{\xi}^k_{g,h+1}-\xi^k_{g,*}\right\|^2&\le\left(1-\frac{\gamma_\xi\lambda}{4}\right)\|\tilde{\xi}^k_{g,h}-\xi^k_{g,*}\|^2+\frac{4\gamma_\xi}{\lambda}\left(8\lambda^{-2}\Lambda_{\mathbf{b}}^2\Delta_{\mathbf{A}}^2+\Delta_{\mathbf{b}}^2\right) && (140)\\
&\quad+\mathcal{O}\left(\gamma_\xi\log(T_{\max})C_2R^2m^{-1/2}\sqrt{\log(H/\delta)}\right)+\mathcal{O}\left(\gamma_\xi^2\left(c_\gamma^2+C_1'^2C_1^2\right)\log(H/\delta)\tau_{\mathrm{mix}}\log T_{\max}\right) && (141)
\end{aligned}
$$

Taking expectations on both sides with respect to θ_k and substituting $\gamma_\xi = \frac{8\log T}{\lambda H}$ yields

$$\begin{aligned}
\mathbb{E}_{\theta_k}\left\|\tilde{\xi}^k_{g,H} - \xi^k_{g,*}\right\|^2 &\leq \left(1 - \frac{\gamma_\xi \lambda}{4}\right)^H \mathbb{E}_{\theta_k}\|\tilde{\xi}^k_{g,0} - \xi^k_{g,*}\|^2 + \frac{4}{\gamma_\xi \lambda}\Bigg(\frac{4\gamma_\xi}{\lambda}\left(8\lambda^{-2}\Lambda^2_{\mathbf{b}}\Delta^2_{\mathbf{A}} + \Delta^2_{\mathbf{b}}\right) \\
&\quad + \mathcal{O}\left(\gamma_\xi \log(T_{\max}) C_2 R^2 m^{-1/2}\sqrt{\log(H/\delta)}\right) + \mathcal{O}\left(\gamma_\xi^2\left(c_\gamma^2 + C_1'^2 C_1^2\right)\log(H/\delta)\tau_{\min}\log T_{\max}\right)\Bigg) \\
&\leq \exp\left(-\frac{H\gamma_\xi\lambda}{4}\right)\mathbb{E}_{\theta_k}\|\tilde{\xi}^k_{g,0} - \xi^k_{g,*}\|^2 + \frac{16}{\lambda^2}\left(8\lambda^{-2}\Lambda^2_{\mathbf{b}}\Delta^2_{\mathbf{A}} + \Delta^2_{\mathbf{b}}\right) \\
&\quad + \mathcal{O}\left(\lambda^{-1}\log(T_{\max}) C_2 R^2 m^{-1/2}\sqrt{\log(H/\delta)}\right) + \mathcal{O}\left(\lambda^{-1}\gamma_\xi\left(c_\gamma^2 + C_1'^2 C_1^2\right)\log(H/\delta)\tau_{\min}\log T_{\max}\right)\Bigg) \\
&\leq \mathcal{O}\left(\frac{\mathbb{E}_{\theta_k}\|\tilde{\xi}^k_{g,0} - \xi^k_{g,*}\|^2}{T^2}\right) + \mathcal{O}\left(\lambda^{-4}(c_\gamma^2 + C_1^2 C_1'^2)\log(H/\delta)\left(c_\gamma^2 + C_1^4\right)\tau_{\min} T^{-1}_{\max}\right) \\
&\quad + \mathcal{O}\left(\lambda^{-1}\log(T_{\max}) C_2 R^2 m^{-1/2}\sqrt{\log(H/\delta)}\right) + \mathcal{O}\left(\frac{\log T\left(c_\gamma^2 + C_1'^2 C_1^2\right)\log(H/\delta)\tau_{\min}\log T_{\max}}{\lambda^2 H}\right)
\end{aligned}$$

Combining this with Lemma C.8 yields

$$\mathbb{E}_{\theta_k}\left[\|\xi^k_{g,H} - \xi^k_{g,*}\|^2\right] \leq 2\mathbb{E}_{\theta_k}\left[\|\tilde{\xi}^k_{g,H} - \xi^k_{g,*}\|^2\right] + 2\mathbb{E}_{\theta_k}\left[\|\tilde{\xi}^k_{g,H} - \xi^k_{g,H}\|^2\right] \tag{142}$$

$$\begin{aligned}
&\leq \mathcal{O}\Bigg(\frac{\mathbb{E}_{\theta_k}\|\tilde{\xi}^k_{g,0} - \xi^k_{g,*}\|^2}{T^2} + \frac{c_\gamma^4 \tau_{\min}\log(H/\delta) R^2 \log T}{\lambda^4 T_{\max}} + \frac{R^2\sqrt{\log(H/\delta)}\log T\log T_{\max}}{\lambda m^{-1/2}} + \frac{R^2 T_{\max}\log(H/\delta)\log T}{H\lambda m} \\
&\quad + \frac{c_\gamma^2\log(H/\delta)\tau_{\min}\log T_{\max}\log T}{\lambda^2 H}\Bigg)
\end{aligned} \tag{143}$$

□

D NPG ANALYSIS

We now provide the analysis for the convergence of the Natural Policy Gradient (NPG) subroutine in Algorithm 1.

Lemma D.1. *Consider the single sample estimates for the Fisher information matrix and the policy gradient $\hat{F}(\theta_k; z)$ and $\hat{\nabla}_\theta J_g(\theta_k, \xi^k_g; z)$. If Assumptions 5.1 and 5.4 hold, then the following bounds on the squared bias and variance hold:*

$$\left\|\mathbb{E}_{\theta_k}\left[\hat{F}(\theta_k; z)\right] - F(\theta_k)\right\|^2 \leq \delta_F^2 = 0 \tag{144}$$

$$\left\|\mathbb{E}_{k}\left[\hat{F}(\theta_k; z)\right] - F(\theta_k)\right\|^2 \leq \bar{\delta}_F^2 = 0 \tag{145}$$

$$\mathbb{E}_{\theta_k}\left[\left\|\hat{F}(\theta_k; z) - F(\theta_k)\right\|^2\right] \leq \sigma_F^2 = 2G_1^4 \tag{146}$$

$$\mathbb{E}_{k}\left[\left\|\hat{F}(\theta_k; z) - F(\theta_k)\right\|^2\right] \leq \bar{\sigma}_F^2 = 2G_1^4 \tag{147}$$

$$\left\|\mathbb{E}_{\theta_k}\left[\hat{\nabla}_\theta J_g(\theta_k; z)\right] - \nabla_\theta J_g(\theta_k)\right\|^2 \leq \delta_J^2 = \mathcal{O}\left(C_1^2 G_1^2\|\xi^k_g - \xi^k_{g,*}\|^2 + G_1^2 C_4^2\log(H/\delta)m^{-1} + G_1^2\tau^2_{\min}\epsilon_{\text{app}}\right) \tag{148}$$

$$\left\|\mathbb{E}_{k}\left[\hat{\nabla}_\theta J_g(\theta_k; z)\right] - \nabla_\theta J_g(\theta_k)\right\|^2 \leq \bar{\delta}_J^2 = \mathcal{O}\left(C_1^2 G_1^2\|\mathbb{E}_k\left[\xi^k_g\right] - \xi^k_{g,*}\|^2 + G_1^2 C_4^2\log(H/\delta)m^{-1} + G_1^2\tau^2_{\min}\epsilon_{\text{app}}\right) \tag{149}$$

$$\mathbb{E}_{\theta_k}\left\|\left[\hat{\nabla}_\theta J_g(\theta_k; z)\right] - \nabla_\theta J_g(\theta_k)\right\|^2 \leq \sigma_J^2 = \mathcal{O}\left(G_1^2 C_1'^2\log(H/\delta)\right) \tag{150}$$

$$\mathbb{E}_{k}\left\|\left[\hat{\nabla}_\theta J_g(\theta_k; z)\right] - \nabla_\theta J_g(\theta_k)\right\|^2 \leq \bar{\sigma}_J^2 = \mathcal{O}\left(G_1^2 C_1'^2\log(H/\delta)\right) \tag{151}$$

Proof. By the definition of the single sample estimate for the Fisher Information Matrix, we have

$$\left\|\mathbb{E}_{\theta_k}\left[\hat{F}(\theta_k;z)\right]-F(\theta_k)\right\|^2=0 \tag{152}$$

as the estimate is unbiased. We also have

$$\begin{aligned}\left\|\hat{F}(\theta_k;z)-F(\theta_k)\right\|^2 &\le 2\|\hat{F}(\theta_k;z)\|^2 &(153)\\ &\le 2\|\nabla_{\theta_k}\log\pi_\theta(a|s)\|^4 &(154)\\ &\le 2G_1^4 &(155)\end{aligned}$$

For the policy gradient estimate $\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)$, we have

$$\begin{aligned}\|\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)\|^2 &= \hat{A}_g^{\pi_{\theta_k}}(\xi_g^k;z)^2\|\nabla_\theta\log\pi_{\theta_k}(a|s)\|^2\\ &\le G_1^2\left(|g(s,a)|+|\eta_g^k|+\left|Q_g(\phi(s',a');\zeta_g^k)-Q_g(\phi(s,a);\zeta_g^k)\right|\right)^2 &(156)\\ &\le \mathcal{O}\left(G_1^2C_1'^2\log(H/\delta)\right) &(157)\end{aligned}$$

where the last inequality follows from the boundedness of the neural network output in the NTK regime (Lemma C.1). Thus, the single-sample variance is bounded by a constant independent of $T_{\max}$. This results in the following bound

$$\mathbb{E}_{\theta_k}\left[\left\|\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)-\nabla_\theta J_g(\theta_k)\right\|^2\right]\le\mathcal{O}\left(G_1^2C_1'^2\log(H/\delta)\right) \tag{158}$$

We now analyze the bias of the estimator. By definition, we have $\xi_{g,*}^k=(\eta_{g,*}^k,\zeta_{g,*}^k)$ as the optimal parameters in the linearized function class $\mathcal{F}_{R,m}$. We decompose the expected error $\mathbb{E}_{\theta_k}[\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)]-\nabla_\theta J_g(\theta_k)$ into four terms $T_0,T_{0,\text{lin}},T_1,T_2$:

$$\begin{aligned}&\mathbb{E}_{\theta_k}\left[\hat{\nabla}_\theta J_g(\theta_k,\xi_g^k;z)\right]-\nabla_\theta J_g(\theta_k)\\ &=\mathbb{E}_{\theta_k}\Bigg[\Bigg\{\underbrace{(\eta_{g,*}^k-\eta_g^k)+\left(\hat{Q}_g(\phi(s',a');\zeta_g^k)-\hat{Q}_g(\phi(s',a');\zeta_{g,*}^k)\right)-\left(\hat{Q}_g(\phi(s,a);\zeta_g^k)-\hat{Q}_g(\phi(s,a);\zeta_{g,*}^k)\right)}_{T_0:\text{Optimization Error (Linearized)}}\\ &+\underbrace{\left(Q_g(\phi(s',a');\zeta_g^k)-\hat{Q}_g(\phi(s',a');\zeta_g^k)\right)-\left(Q_g(\phi(s,a);\zeta_g^k)-\hat{Q}_g(\phi(s,a);\zeta_g^k)\right)}_{T_{0,\text{lin}}:\text{Linearization Residuals}}\\ &+\underbrace{Q_g^{\pi_{\theta_k}}(s',a')-\hat{Q}_g(\phi(s',a');\zeta_{g,*}^k)-\left(Q_g^{\pi_{\theta_k}}(s,a)-\hat{Q}_g(\phi(s,a);\zeta_{g,*}^k)\right)}_{T_1:\text{Function Approximation Error}}\\ &+\underbrace{g(s,a)-\eta_{g,*}^k+Q_g^{\pi_{\theta_k}}(s',a')-Q_g^{\pi_{\theta_k}}(s,a)}_{T_2:\text{True Advantage / Bellman Error}}\Bigg\}\nabla_\theta\log\pi_{\theta_k}(a|s)\Bigg]-\nabla_\theta J_g(\theta_k)\end{aligned} \tag{159}$$

Using the definition of the advantage function $A_g^\pi(s,a)$ and the Policy Gradient Theorem:

$$\begin{aligned}&\mathbb{E}_{\theta_k}\left[T_2\nabla_\theta\log\pi_{\theta_k}(a|s)\right]-\nabla_\theta J_g(\theta_k)\\ &=\mathbb{E}_{s,a\sim d^{\pi_{\theta_k}}}\left[\mathbb{E}_{s'}\left[g(s,a)-\eta_{g,*}^k+Q_g^{\pi_{\theta_k}}(s')-Q_g^{\pi_{\theta_k}}(s)\;\middle|\;s,a\right]\nabla_\theta\log\pi_{\theta_k}(a|s)\right]-\nabla_\theta J_g(\theta_k) &(160)\\ &\overset{(a)}{=}\mathbb{E}_{s,a\sim d^{\pi_{\theta_k}}}\left[A_g^{\pi_{\theta_k}}(s,a)\nabla_\theta\log\pi_{\theta_k}(a|s)\right]-\nabla_\theta J_g(\theta_k) &(161)\\ &=0 &(162)\end{aligned}$$

where (a) uses the Bellman equation and that $J_g(\theta_k)=\eta_{g,*}^k$ by definition. Using the linearity of $\hat{Q}_g$ with respect to ζ (defined via inner product with $\nabla_\zeta Q_g(\cdot;\zeta_{g,0})$):

$$T_0=(\eta_{g,*}^k-\eta_g^k)+\left(\hat{Q}_g(\phi(s',a');\zeta_g^k)-\hat{Q}_g(\phi(s',a');\zeta_{g,*}^k)\right)-\left(\hat{Q}_g(\phi(s,a);\zeta_g^k)-\hat{Q}_g(\phi(s,a);\zeta_{g,*}^k)\right)$$

$$= (\eta_{g,*}^k - \eta_g^k) + \left\langle \nabla_\zeta Q_g(\phi(s',a');\zeta_0) - \nabla_\zeta Q_g(\phi(s',a');\zeta_0),\ \zeta_g^k - \zeta_{g,*}^k \right\rangle \tag{163}$$

$$\leq 2C_1 \|\xi_g^k - \xi_{g,*}^k\| \tag{164}$$

where the last inequality uses Lemma C.1. Applying the Cauchy-Schwarz inequality and taking the squared expectation:

$$\|\mathbb{E}_{\theta_k}[T_0 \nabla_\theta \log \pi]\|^2 \leq G_1^2 \mathbb{E}_{\theta_k}\left[|T_0|^2\right] \leq \mathcal{O}\left(C_1^2 G_1^2 \|\xi_g^k - \xi_{g,*}^k\|^2\right) \tag{165}$$

Using Lemma C.1, which bounds the difference between the neural network and its linearization in the NTK regime by $\mathcal{O}(m^{-1/2})$:

$$\begin{aligned} |T_{0,\text{lin}}| &= \left| \left(Q_g(\phi(s',a');\zeta_g^k) - \hat{Q}_g(\phi(s',a');\zeta_g^k) \right) - \left(Q_g(\phi(s,a);\zeta_g^k) - \hat{Q}_g(\phi(s,a);\zeta_g^k) \right) \right| \\ &\leq \left| Q_g(\phi(s',a');\zeta_g^k) - \hat{Q}_g(\phi(s',a');\zeta_g^k) \right| + \left| Q_g(\phi(s,a);\zeta_g^k) - \hat{Q}_g(\phi(s,a);\zeta_g^k) \right| \\ &\leq \mathcal{O}\left(C_4 m^{-1/2} \sqrt{\log(H/\delta)} \right) \end{aligned} \tag{166}$$

Thus, the squared error contribution is:

$$\|\mathbb{E}_{\theta_k}[T_{0,\text{lin}} \nabla_\theta \log \pi]\|^2 \leq G_1^2 \mathbb{E}_{\theta_k}[|T_{0,\text{lin}}|^2] \leq \mathcal{O}\left(\frac{G_1^2 C_4^2 \log(H/\delta)}{m} \right) \tag{167}$$

Using Assumption 3.4 (Neural Critic Approximation Error), which bounds the expected squared Bellman error of the optimal linearized critic $\xi_{g,*}^k$:

$$\begin{aligned} \mathbb{E}_{\theta_k}[T_1^2] &= \mathbb{E}_{\theta_k}\left[\left(\left(Q_g^\pi(s',a') - \hat{Q}_g(\phi(s',a');\zeta_{g,*}^k) \right) - \left(Q_g^\pi(s,a) - \hat{Q}_g(\phi(s,a);\zeta_{g,*}^k) \right) \right)^2 \right] \\ &\leq \mathcal{O}\left(\tau_{\text{mix}}^2 \epsilon_{\text{app}} \right) \end{aligned} \tag{168}$$

using the average reward analogue of A.3 in Ke et al. [2024]. We have

$$\|\mathbb{E}_{\theta_k}[T_1 \nabla_\theta \log \pi]\|^2 \leq G_1^2 \mathbb{E}_{\theta_k}[T_1^2] \leq \mathcal{O}\left(G_1^2 \tau_{\text{mix}}^2 \epsilon_{\text{app}} \right) \tag{169}$$

Combining these terms, we have

$$\left\| \mathbb{E}_{\theta_k}\left[\hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z) \right] - \nabla_\theta J_g(\theta_k) \right\|^2 \leq \mathcal{O}\left(C_1^2 G_1^2 \|\xi_g^k - \xi_{g,*}^k\|^2 + G_1^2 C_4^2 \log(H/\delta) m^{-1} + G_1^2 \tau_{\text{mix}}^2 \epsilon_{\text{app}} \right) \tag{170}$$

Applying the same argument yields

$$\left\| \mathbb{E}_k \mathbb{E}_{\theta_k}\left[\hat{\nabla}_\theta J_g(\theta_k, \xi_g^k; z) \right] - \nabla_\theta J_g(\theta_k) \right\|^2 \leq \mathcal{O}\left(C_1^2 G_1^2 \|\mathbb{E}\left[\xi_g^k\right] - \xi_{g,*}^k\|^2 + G_1^2 C_4^2 \log(H/\delta) m^{-1} + G_1^2 \tau_{\text{mix}}^2 \epsilon_{\text{app}} \right) \tag{171}$$

□

Lemma D.2. *We denote the MLMC estimators for the Fisher Information Matrix and the policy gradient as*

$$F_{kh}^{\text{MLMC}} = F_{kh}^0 + 2^{Q_h^k} \left(F_{kh}^{Q_h^k} - F_{kh}^{Q_h^k - 1} \right) \mathbb{I}\left(2^{Q_h^k} \leq T_{\max} \right) \tag{172}$$

$$\nabla_\theta J_{g,kh}^{\text{MLMC}} = \nabla_\theta J_{g,kh}^0 + 2^{Q_h^k} \left(\nabla_\theta J_{g,kh}^{Q_h^k} - \nabla_\theta J_{g,kh}^{Q_h^k - 1} \right) \mathbb{I}\left(2^{Q_h^k} \leq T_{\max} \right) \tag{173}$$

where $Q_h^k \sim \text{Geom}(1/2)$, $F_{kh}^j = \frac{1}{2^j} \sum_{i=0}^{2^j-1} \hat{F}(\theta_k; z_{kh}^i)$, *and* $\nabla_\theta J_{g,kh}^j = \frac{1}{2^j} \sum_{i=0}^{2^j-1} \hat{\nabla}_\theta J_g(\theta_k; \xi_g^k; z_{kh}^i)$. *Then the following bounds hold:*

$$\left\| \mathbb{E}_{k,h}\left[F_{kh}^{\text{MLMC}} \right] - F(\theta_k) \right\|^2 \leq \mathcal{O}\left(G_1^4 \tau_{\text{mix}} T_{\max}^{-1} \right) \tag{174}$$

$$\mathbb{E}_{k,h}\left[\left\| F_{kh}^{\text{MLMC}} - F(\theta_k) \right\|^2 \right] \leq \mathcal{O}\left(G_1^4 \tau_{\text{mix}} \log T_{\max} \right) \tag{175}$$

$$\left\| \mathbb{E}_{k,h}\left[\nabla_\theta J_{g,kh}^{\text{MLMC}} \right] - \nabla_\theta J_g(\theta_k, \xi_g^k) \right\|^2 \leq \mathcal{O}\Bigg(G_1^2 \log(H/\delta) \tau_{\text{mix}} T_{\max}^{-1} + G_1^2 \|\xi_g^k - \xi_{g,*}^k\|^2$$

$$+ G_1^2 \log(H/\delta) m^{-1} + G_1^2 \tau_{\mathrm{mix}}^2 \epsilon_{\mathrm{app}} \Bigg) \tag{176}$$

$$\mathbb{E}_{k,h}\left[\left\|\nabla_\theta J_{g,kh}^{\mathrm{MLMC}} - \nabla_\theta J_g(\theta_k, \xi_g^k)\right\|^2\right] \leq \mathcal{O}\Bigg(G_1^2 \log(H/\delta)\tau_{\mathrm{mix}} \log T_{\max} + G_1^2 \|\xi_g^k - \xi_{g,*}^k\|^2 + G_1^2 \log(H/\delta) m^{-1} + G_1^2 \tau_{\mathrm{mix}}^2 \epsilon_{\mathrm{app}} \Bigg) \tag{177}$$

$$\left\|\mathbb{E}_k\left[\nabla_\theta J_{g,kh}^{\mathrm{MLMC}}\right] - \nabla_\theta J_g(\theta_k, \xi_g^k)\right\|^2 \leq \mathcal{O}\Bigg(G_1^2 \log(H/\delta)\tau_{\mathrm{mix}} T_{\max}^{-1} + G_1^2 \|\mathbb{E}_k\left[\xi_g^k\right] - \xi_{g,*}^k\|^2 + G_1^2 \log(H/\delta) m^{-1} + G_1^2 \tau_{\mathrm{mix}}^2 \epsilon_{\mathrm{app}} \Bigg) \tag{178}$$

Proof. Using Lemma D.1 with Lemma B.2, we have

$$\left\|\mathbb{E}_{k,h}\left[F_{kh}^{\mathrm{MLMC}}\right] - F(\theta_k)\right\|^2 \leq \mathcal{O}\left(\sigma_F^2 \tau_{\mathrm{mix}} T_{\max}^{-1} + \delta_F^2\right) = \mathcal{O}\left(G_1^4 \tau_{\mathrm{mix}} T_{\max}^{-1}\right) \tag{179}$$

$$\mathbb{E}_{k,h}\left[\left\|F_{kh}^{\mathrm{MLMC}} - F(\theta_k)\right\|^2\right] \leq \mathcal{O}\left(\sigma_F^2 \tau_{\mathrm{mix}} \log T_{\max} + \delta_F^2\right) = \mathcal{O}\left(G_1^4 \tau_{\mathrm{mix}} \log T_{\max}\right) \tag{180}$$

Similarly, we have

$$\left\|\mathbb{E}_{k,h}\left[\nabla_\theta J_{g,kh}^{\mathrm{MLMC}}\right] - \nabla_\theta J_g(\theta_k, \xi_g^k)\right\|^2 \leq \mathcal{O}\left(\sigma_J^2 \tau_{\mathrm{mix}} T_{\max}^{-1} + \delta_J^2\right) \tag{181}$$

$$= \mathcal{O}\Bigg(G_1^2 \log(H/\delta)\tau_{\mathrm{mix}} T_{\max}^{-1} + G_1^2 \|\xi_g^k - \xi_{g,*}^k\|^2 + G_1^2 \log(H/\delta) m^{-1} + G_1^2 \tau_{\mathrm{mix}}^2 \epsilon_{\mathrm{app}} \Bigg) \tag{182}$$

$$\mathbb{E}_{k,h}\left[\left\|\nabla_\theta J_{g,kh}^{\mathrm{MLMC}} - \nabla_\theta J_g(\theta_k, \xi_g^k)\right\|^2\right] \leq \mathcal{O}\left(\sigma_J^2 \tau_{\mathrm{mix}} \log T_{\max} + \delta_J^2\right) \tag{183}$$

$$= \mathcal{O}\Bigg(G_1^2 \log(H/\delta)\tau_{\mathrm{mix}} \log T_{\max} + G_1^2 \|\xi_g^k - \xi_{g,*}^k\|^2 + G_1^2 \log(H/\delta) m^{-1} + G_1^2 \tau_{\mathrm{mix}}^2 \epsilon_{\mathrm{app}} \Bigg) \tag{184}$$

$$\left\|\mathbb{E}_k\left[\nabla_\theta J_{g,kh}^{\mathrm{MLMC}}\right] - \nabla_\theta J_g(\theta_k, \xi_g^k)\right\|^2 \leq \mathcal{O}\left(\bar{\sigma}_J^2 \tau_{\mathrm{mix}} T_{\max}^{-1} + \bar{\delta}_J^2\right) \tag{185}$$

$$= \mathcal{O}\Bigg(G_1^2 \log(H/\delta)\tau_{\mathrm{mix}} T_{\max}^{-1} + G_1^2 \|\mathbb{E}_k\left[\xi_g^k\right] - \xi_{g,*}^k\|^2 + G_1^2 \log(H/\delta) m^{-1} + G_1^2 \tau_{\mathrm{mix}}^2 \epsilon_{\mathrm{app}} \Bigg) \tag{186}$$

□

Theorem D.3 (NPG Estimation Error Bounds)**.** *Suppose the assumptions in Section 5 hold. Let the NPG step size be set as $\gamma_\omega = \frac{2\log T}{\mu H}$ and the initialization be $\omega_{g,0}^k = 0$. Let $\omega_{g,*}^k = F(\theta_k)^{-1}\nabla_\theta J_g(\theta_k)$ denote the true Natural Policy Gradient direction. Then, after H inner iterations, the mean-squared error of the estimator ω_g^k satisfies:*

$$\mathbb{E}_k\left[\left\|\omega_g^k - \omega_{g,*}^k\right\|^2\right] \leq \frac{G_1^2 \tau_{\mathrm{mix}}^2}{\mu^2 T^2} + \tilde{\mathcal{O}}\left(\frac{G_1^6 \tau_{\mathrm{mix}}^3}{\mu^4}\left(\frac{1}{H} + \frac{1}{T_{\max}}\right) + \frac{G_1^2}{\mu^2\sqrt{m}} + \frac{G_1^2 \tau_{\mathrm{mix}}^2}{\mu^2}\epsilon_{\mathrm{app}}\right), \tag{187}$$

and the squared bias of the estimator satisfies:

$$\left\|\mathbb{E}_k[\omega_g^k] - \omega_{g,*}^k\right\|^2 \leq \frac{G_1^2 \tau_{\mathrm{mix}}^2}{\mu^2 T^2} + \tilde{\mathcal{O}}\left(\frac{G_1^6 \tau_{\mathrm{mix}}^3}{\mu^4 T_{\max}} + \frac{G_1^2}{\mu^2\sqrt{m}} + \frac{G_1^2 \tau_{\mathrm{mix}}^2}{\mu^2}\epsilon_{\mathrm{app}}\right), \tag{188}$$

where $\tilde{\mathcal{O}}$ hides polylogarithmic factors in $H, T, T_{\max}$, and $1/\delta$.

Proof. We note that by Lemma D.2, that the conditions of Equation 36 in Theorem B.1 in Xu et al. [2025] are satisfied. Additionally, for any policy parameter θ, we have

$$\mu \leq \|F(\theta)\| \leq \mathcal{O}\left(G_1\right), \quad \|\nabla_\theta J(\theta)\| \leq \mathcal{O}\left(G_1 \tau_{\text{mix}}\right) \tag{189}$$

from Theorem 4.7 in Xu et al. [2025]. Thus, all conditions of Theorem B.1 in Xu et al. [2025] are satisfied, and we have

$$\mathbb{E}_k\left[\left\|\omega_g^k - \omega_{g,*}^k\right\|^2\right] \leq \exp\left(-H\gamma_\omega\mu\right)\left\|\omega_0 - \omega_{g,*}^k\right\|^2 + \mathcal{O}(\gamma_\omega R_0 + R_1)$$

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

where

$$R_0 = \tilde{\mathcal{O}}\left(\mu^{-3}G_1^6\tau_{\text{mix}}^3 + \mu^{-1}G_1^2\log(H/\delta)(\tau_{\text{mix}} + m^{-1}) + \mu^{-1}G_1^2\mathbb{E}_k\|\xi_g^k - \xi_{g,*}^k\|^2 + \mu^{-1}G_1^2\tau_{\text{mix}}^2\epsilon_{\text{app}}\right) \tag{190}$$

$$R_1 = \mathcal{O}\left(\mu^{-4}G_1^6\tau_{\text{mix}}^3 T_{\max}^{-1} + \mu^{-2}G_1^2\log(H/\delta)(\tau_{\text{mix}}T_{\max}^{-1} + m^{-1}) + \mu^{-2}G_1^2\mathbb{E}_k\|\xi_g^k - \xi_{g,*}^k\|^2 + \mu^{-2}G_1^2\tau_{\text{mix}}^2\epsilon_{\text{app}}\right) \tag{191}$$

$$\bar{R}_1 = \mathcal{O}\left(\mu^{-2}G_1^6\tau_{\text{mix}}^3 T_{\max}^{-1} + G_1^2\log(H/\delta)(\tau_{\text{mix}}T_{\max}^{-1} + m^{-1}) + G_1^2\|\mathbb{E}_k\left[\xi_g^k\right] - \xi_{g,*}^k\|^2 + G_1^2\tau_{\text{mix}}^2\epsilon_{\text{app}}\right) \tag{192}$$

and we set $\gamma_\omega = \frac{2\log T}{\mu H}$. Using $\omega_0 = 0$ and $\|\omega_{g,*}^k\|^2 = \|F(\theta_k)^{-1}\nabla_\theta J(\theta_k)\|^2 \leq \mathcal{O}\left(\mu^{-1}G_1\tau_{\text{mix}}\right)$, this yields

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\omega_g^k - \omega_{g,*}^k\right\|^2\right] &\leq \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2} + \tilde{\mathcal{O}}\Bigg(\left(H^{-1} + T_{\max}^{-1}\right)\left(\mu^{-4}G_1^6\tau_{\text{mix}}^3 + \mu^{-2}G_1^2\log(H/\delta)\tau_{\text{mix}}\right) \\
&\qquad + \mu^{-2}G_1^2 m^{-1/2} + \mu^{-2}G_1^2\mathbb{E}_k\|\xi_g^k - \xi_{g,*}^k\|^2 + \mu^{-2}G_1^2\tau_{\text{mix}}^2\epsilon_{\text{app}}\Bigg) \\
\left\|\mathbb{E}_k[\omega_g^k] - \omega_{g,k}^*\right\|^2 &\leq \left(\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2} + \frac{G_1^6\tau_{\text{mix}}^3}{\mu^4 T_{\max}}\right) + \tilde{\mathcal{O}}\Bigg(\mu^{-4}G_1^6\tau_{\text{mix}}^3 T_{\max}^{-1} + \mu^{-2}G_1^2\log(H/\delta)\tau_{\text{mix}}T_{\max}^{-1} + \mu^{-2}G_1 m^{-1/2} \\
&\qquad + \mu^{-2}G_1^2\|\mathbb{E}_k\left[\xi_g^k\right] - \xi_{g,*}^k\|^2 + \mu^{-2}G_1^2\tau_{\text{mix}}^2\epsilon_{\text{app}} + \frac{G_1^6\tau_{\text{mix}}^2}{\mu^4 T_{\max}}\mathbb{E}_k\|\xi_g^k - \xi_{g,*}^k\|^2\Bigg) \\
&\leq \left(\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2} + \frac{G_1^6\tau_{\text{mix}}^3}{\mu^4 T_{\max}}\right) + \tilde{\mathcal{O}}\Bigg(\mu^{-4}G_1^6\tau_{\text{mix}}^3 T_{\max}^{-1} + \mu^{-2}G_1^2\log(H/\delta)\tau_{\text{mix}}T_{\max}^{-1} + \mu^{-2}G_1 m^{-1/2} \\
&\qquad + \mu^{-2}G_1^2\mathbb{E}_k\left[\|\xi_g^k - \xi_{g,*}^k\|^2\right] + \mu^{-2}G_1^2\tau_{\text{mix}}^2\epsilon_{\text{app}} + \frac{G_1^6\tau_{\text{mix}}^2}{\mu^4 T_{\max}}\mathbb{E}_k\|\xi_g^k - \xi_{g,*}^k\|^2\Bigg) \\
&\leq \left(\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2} + \frac{G_1^6\tau_{\text{mix}}^3}{\mu^4 T_{\max}}\right) + \tilde{\mathcal{O}}\Bigg(\mu^{-4}G_1^6\tau_{\text{mix}}^3 T_{\max}^{-1} + \mu^{-2}G_1^2\log(H/\delta)\tau_{\text{mix}}T_{\max}^{-1} + \mu^{-2}G_1 m^{-1/2} \\
&\qquad + \mu^{-2}G_1^2\mathbb{E}_k\left[\|\xi_g^k - \xi_{g,*}^k\|^2\right] + \mu^{-2}G_1^2\tau_{\text{mix}}^2\epsilon_{\text{app}}\Bigg)
\end{aligned}$$

Finally, we substitute the value for $\mathbb{E}_k\left[\|\xi_g^k - \xi_{g,*}^k\|^2\right]$ to yield

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\omega_g^k - \omega_{g,*}^k\right\|^2\right] &\leq \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2} + \tilde{\mathcal{O}}\Bigg(\left(\frac{1}{H} + \frac{1}{T_{\max}}\right)\left(\frac{G_1^6\tau_{\text{mix}}^3}{\mu^4} + \frac{G_1^2\log(H/\delta)\tau_{\text{mix}}}{\mu^2}\right) \\
&\qquad + \frac{G_1^2}{\mu^2\sqrt{m}} + \frac{G_1^2 c_\gamma^2\log(H/\delta)\tau_{\text{mix}}}{\lambda^2\mu^2 H} + \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\epsilon_{\text{app}}\Bigg) \qquad (193) \\
\left\|\mathbb{E}_k[\omega_g^k] - \omega_{g,k}^*\right\|^2 &\leq \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2} + \tilde{\mathcal{O}}\Bigg(\frac{G_1^6\tau_{\text{mix}}^3}{T_{\max}\mu^4} + \frac{G_1^2\log(H/\delta)\tau_{\text{mix}}}{\mu^2 T_{\max}} + \frac{G_1^2}{\mu^2\sqrt{m}} \\
&\qquad + \frac{G_1^2 c_\gamma^2\log(H/\delta)}{\mu^2\lambda^2 H} + \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\epsilon_{\text{app}}\Bigg) \qquad (194)
\end{aligned}$$

□

E CMDP ANALYSIS

In this section, we formally analyze the convergence rate of the overall CMDP.

Theorem E.1. *Suppose the assumptions in Section 5 hold. Let the step sizes be $\alpha = \beta = \Theta(T^{-1/4})$, the number of outer iterations be $K = \Theta(T^{1/2})$, and the number of inner iterations be $H = \Theta(T^{1/2})$. Let the MLMC truncation level be $T_{\max} = T$. Then, the sequence of policies $\{\theta_k\}_{k=0}^{K-1}$ generated by Algorithm 1 satisfies the following bounds on the average reward gap and constraint violation:*

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\left(J_r^{\pi^*} - J_r(\theta_k)\right)\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + T^{-1/4} + m^{-1/4}\right), \tag{195}$$

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1} -J_c(\theta_k)\right] \leq \tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + T^{-1/4} + m^{-1/4}\right), \tag{196}$$

where $\tilde{\mathcal{O}}$ hides polylogarithmic factors in T.

Proof. Combining Theorem D.3 with the fact that $\lambda_k \in [0, 2/\delta_s]$, we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{E}_k\left[\omega_k\right] - \omega_k^*\|\right] \leq \left(1 + \frac{2}{\delta_s}\right)\tilde{\mathcal{O}}\Bigg(&\frac{G_1\tau_{\text{mix}}}{\mu T} + \frac{G_1^3\tau_{\text{mix}}^2}{\sqrt{T_{\max}}\mu^2} + \frac{G_1\sqrt{\log(H/\delta)}\tau_{\text{mix}}}{\mu\sqrt{T_{\max}}} + \frac{G_1}{\mu m^{1/4}} \\ &+ \frac{G_1 c_\gamma\sqrt{\log(H/\delta)}}{\mu\lambda\sqrt{H}} + \frac{G_1\tau_{\text{mix}}}{\mu}\sqrt{\epsilon_{\text{app}}}\Bigg)\end{aligned} \tag{197}$$

$$\begin{aligned}\mathbb{E}_k\left[\left\|\omega^k - \omega_k^*\right\|^2\right] \leq \left(1 + \frac{4}{\delta_s^2}\right)\tilde{\mathcal{O}}\Bigg(&\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2 T^2} + \left(\frac{1}{H} + \frac{1}{T_{\max}}\right)\left(\frac{G_1^6\tau_{\text{mix}}^3}{\mu^4} + \frac{G_1^2\log(H/\delta)\tau_{\text{mix}}}{\mu^2}\right) \\ &+ \frac{G_1^2}{\mu^2\sqrt{m}} + \frac{G_1^2 c_\gamma^2\log(H/\delta)\tau_{\text{mix}}}{\lambda^2\mu^2 H} + \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\epsilon_{\text{app}}\Bigg)\end{aligned} \tag{198}$$

Using Lemma 4.6 and Lemma G.2 in Xu et al. [2025] and setting $T_{\max} = T$, we have

$$\begin{aligned}&\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\left(\mathcal{L}(\pi^*, \lambda_k) - \mathcal{L}(\theta_k, \lambda_k)\right)\right] \\ &\leq \sqrt{\epsilon_{\text{bias}}} + \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right] \\ &\quad + \left(1 + \frac{2}{\delta_s}\right)\tilde{\mathcal{O}}\left(\frac{G_1^3\tau_{\text{mix}}^2}{\sqrt{T}\mu^2} + \frac{G_1\sqrt{\log(H/\delta)}\tau_{\text{mix}}}{\mu\sqrt{T}} + \frac{G_1}{\mu m^{1/4}} + \frac{G_1\tau_{\text{mix}}}{\mu}\sqrt{\epsilon_{\text{app}}} + \frac{G_1 c_\gamma\sqrt{\log(H/\delta)}}{\mu\lambda\sqrt{H}}\right) \\ &\quad + \alpha G_2\left(1 + \frac{4}{\delta_s^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^6\tau_{\text{mix}}^3}{H\mu^4} + \frac{G_1^2\log(H/\delta)\tau_{\text{mix}}}{H\mu^2} + \frac{G_1^2}{\mu^2\sqrt{m}} + \frac{G_1^2 c_\gamma^2\log(H/\delta)\tau_{\text{mix}}}{\lambda^2\mu^2 H} + \frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\epsilon_{\text{app}}\right) \\ &\quad + \alpha\mathcal{O}\left(\left(1 + \frac{4}{\delta_s^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right)\end{aligned} \tag{199}$$

By the definition of the Lagrange function, we have

$$\begin{aligned}&\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\left(J_r^{\pi^*} - J_r(\theta_k)\right)\right] \\ &\leq \sqrt{\epsilon_{\text{bias}}} + \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right] \\ &\quad + \left(1 + \frac{2}{\delta_s}\right)\tilde{\mathcal{O}}\left(\frac{G_1^3\tau_{\text{mix}}^2}{\sqrt{T}\mu^2} + \frac{G_1\sqrt{\log(H/\delta)}\tau_{\text{mix}}}{\mu\sqrt{T}} + \frac{G_1}{\mu m^{1/4}} + \frac{G_1\tau_{\text{mix}}}{\mu}\sqrt{\epsilon_{\text{app}}} + \frac{G_1 c_\gamma\sqrt{\log(H/\delta)}}{\mu\lambda\sqrt{H}}\right)\end{aligned}$$

$$+\alpha G_2\left(1+\frac{4}{\delta_s^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^6\tau_{\text{mix}}^3}{H\mu^4}+\frac{G_1^2\log(H/\delta)\tau_{\text{mix}}}{H\mu^2}+\frac{G_1^2}{\mu^2\sqrt{m}}+\frac{G_1^2c_\gamma^2\log(H/\delta)\tau_{\text{mix}}}{\lambda^2\mu^2H}+\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\epsilon_{\text{app}}\right)$$
$$+\alpha\mathcal{O}\left(\left(1+\frac{4}{\delta_s^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right)-\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\lambda_k\left(J_c^{\pi^*}-J_c(\theta_k)\right)\right] \tag{200}$$

From Equation 75 in Xu et al. [2025], we have

$$-\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\lambda_k\left(J_c^{\pi^*}-J_c(\theta_k)\right)\right]\leq\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}|\lambda_k|\left(\eta_c^*(\theta_k)-\mathbb{E}\left[\eta_c^k\right]\right)\right]+\beta \tag{201}$$
$$\leq\tilde{\mathcal{O}}\left(\frac{G_1c_\gamma\sqrt{\log(H/\delta)}}{\mu\lambda\sqrt{H}}+\frac{G_1}{\mu m^{1/4}}+\beta\right) \tag{202}$$

Thus,

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\left(J_r^{\pi^*}-J_r(\theta_k)\right)\right]$$
$$\leq\sqrt{\epsilon_{\text{bias}}}+\frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right]$$
$$+\left(1+\frac{2}{\delta_s}\right)\tilde{\mathcal{O}}\left(\frac{G_1^3\tau_{\text{mix}}^2}{\sqrt{T}\mu^2}+\frac{G_1\sqrt{\log(H/\delta)}\tau_{\text{mix}}}{\mu\sqrt{T}}+\frac{G_1}{\mu m^{1/4}}+\frac{G_1\tau_{\text{mix}}}{\mu}\sqrt{\epsilon_{\text{app}}}+\frac{G_1c_\gamma\sqrt{\log(H/\delta)}}{\mu\lambda\sqrt{H}}\right)$$
$$+\alpha G_2\left(1+\frac{4}{\delta_s^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^6\tau_{\text{mix}}^3}{H\mu^4}+\frac{G_1^2\log(H/\delta)\tau_{\text{mix}}}{H\mu^2}+\frac{G_1^2}{\mu^2\sqrt{m}}+\frac{G_1^2c_\gamma^2\log(H/\delta)\tau_{\text{mix}}}{\lambda^2\mu^2H}+\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\epsilon_{\text{app}}\right)$$
$$+\alpha\mathcal{O}\left(\left(1+\frac{4}{\delta_s^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right)+\tilde{\mathcal{O}}\left(\frac{G_1c_\gamma\sqrt{\log(H/\delta)}}{\mu\lambda\sqrt{H}}+\beta\right) \tag{203}$$
$$\leq\tilde{\mathcal{O}}\left(\sqrt{\epsilon_{\text{bias}}}+\sqrt{\epsilon_{\text{app}}}+\frac{1}{\alpha K}+\alpha+\beta+\frac{\sqrt{\log(H/\delta)}}{\sqrt{H}}+m^{-1/4}\right) \tag{204}$$

From Equation 79 in Xu et al. [2025],

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}J_c(\theta_k)\left(\lambda_k-\frac{2}{\delta_s}\right)\right]\leq\frac{2}{\delta_s^2\beta K}+\frac{\beta}{2} \tag{205}$$

Since $\lambda_k J_c^{\pi^*}\geq 0$, we have

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\left(J_r^{\pi^*}-J_r(\theta_k)\right)\right]+\frac{2}{\delta_s K}\mathbb{E}\left[\sum_{k=0}^{K-1}-J_c(\theta_k)\right]$$
$$\leq\sqrt{\epsilon_{\text{bias}}}+\frac{2}{\delta_s^2\beta K}+\frac{\beta}{2}+\frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}\left[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right]$$
$$+\left(1+\frac{2}{\delta_s}\right)\tilde{\mathcal{O}}\left(\frac{G_1^3\tau_{\text{mix}}^2}{\sqrt{T}\mu^2}+\frac{G_1\sqrt{\log(H/\delta)}\tau_{\text{mix}}}{\mu\sqrt{T}}+\frac{G_1}{\mu m^{1/4}}+\frac{G_1\tau_{\text{mix}}}{\mu}\sqrt{\epsilon_{\text{app}}}+\frac{G_1c_\gamma\sqrt{\log(H/\delta)}}{\mu\lambda\sqrt{H}}\right)$$
$$+\alpha G_2\left(1+\frac{4}{\delta_s^2}\right)\tilde{\mathcal{O}}\left(\frac{G_1^6\tau_{\text{mix}}^3}{H\mu^4}+\frac{G_1^2\log(H/\delta)\tau_{\text{mix}}}{H\mu^2}+\frac{G_1^2}{\mu^2\sqrt{m}}+\frac{G_1^2c_\gamma^2\log(H/\delta)\tau_{\text{mix}}}{\lambda^2\mu^2H}+\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\epsilon_{\text{app}}\right)$$
$$+\alpha\mathcal{O}\left(\left(1+\frac{4}{\delta_s^2}\right)\frac{G_1^2\tau_{\text{mix}}^2}{\mu^2}\right) \tag{206}$$

Then by Lemma G.6 in Xu et al. [2025], we have

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}-J_c(\theta_k)\right]$$

$$
\begin{aligned}
&\leq \sqrt{\epsilon_{\text{bias}}} + \frac{1}{\delta_s \beta K} + \frac{\delta_s \beta}{4} + \frac{\delta_s}{2\alpha K} \mathbb{E}_{s\sim d^{\pi^*}} \left[KL(\pi^*(\cdot|s) \| \pi_{\theta_0}(\cdot|s)) \right] \\
&\quad + \left(1 + \frac{\delta_s}{2}\right) \tilde{\mathcal{O}} \left(\frac{G_1^3 \tau_{\text{mix}}^2}{\sqrt{T}\mu^2} + \frac{G_1 \sqrt{\log(H/\delta)} \tau_{\text{mix}}}{\mu\sqrt{T}} + \frac{G_1}{\mu m^{1/4}} + \frac{G_1 \tau_{\text{mix}}}{\mu} \sqrt{\epsilon_{\text{app}}} + \frac{G_1 c_\gamma \sqrt{\log(H/\delta)}}{\mu\lambda\sqrt{H}} \right) \\
&\quad + \alpha G_2 \left(\frac{\delta_s}{2} + \frac{2}{\delta_s} \right) \tilde{\mathcal{O}} \left(\frac{G_1^6 \tau_{\text{mix}}^3}{H\mu^4} + \frac{G_1^2 \log(H/\delta) \tau_{\text{mix}}}{H\mu^2} + \frac{G_1^2}{\mu^2\sqrt{m}} + \frac{G_1^2 c_\gamma^2 \log(H/\delta) \tau_{\text{mix}}}{\lambda^2 \mu^2 H} + \frac{G_1^2 \tau_{\text{mix}}^2}{\mu^2} \epsilon_{\text{app}} \right) \\
&\quad + \alpha \mathcal{O} \left(\left(\frac{\delta_s}{2} + \frac{2}{\delta_s} \right) \frac{G_1^2 \tau_{\text{mix}}^2}{\mu^2} \right) \qquad (207) \\
&\leq \tilde{\mathcal{O}} \left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + \frac{1}{\alpha K} + \frac{1}{\beta K} + \alpha + \beta + \frac{\sqrt{\log(H/\delta)}}{\sqrt{H}} + m^{-1/4} \right) \qquad (208)
\end{aligned}
$$

Using $K = H = \Theta(T^{1/2})$, $\alpha = \beta = \Theta(T^{-1/4})$, we have

$$
\frac{1}{K} \mathbb{E} \left[\sum_{k=0}^{K-1} \left(J_r^{\pi^*} - J_r(\theta_k) \right) \right] \leq \tilde{\mathcal{O}} \left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + T^{-1/4} + m^{-1/4} \right) \qquad (209)
$$

$$
\frac{1}{K} \mathbb{E} \left[\sum_{k=0}^{K-1} -J_c(\theta_k) \right] \leq \tilde{\mathcal{O}} \left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{app}}} + T^{-1/4} + m^{-1/4} \right) \qquad (210)
$$

□

F EXPERIMENTS

F.1 VARIANCE STABILITY OF THE MLMC ESTIMATOR NEAR CONSTRAINT BOUNDARIES

To explicitly verify the stability of the MLMC estimator in practice, we conduct variance stability experiments ($N = 500$ independent trials) on `Hopper-v4`.

Experimental Setup. The constraint is $J_c(\theta) \geq 0$, where we define the cost as $c(s, a) = 0.25 - \text{mean}(a^2)$. The true constraint boundary is thus exactly at the average cost $\eta_c = 0$. We fix the variance of the MLMC gradient norm estimator over 500 independent draws at three η values near the boundary.

Table 2: MLMC estimator variance near the constraint boundary on `Hopper-v4`.

Regime	η_c	MLMC Variance	MC Variance	Ratio (MLMC/MC)
Interior	$+0.10$	9.33×10^{-5}	1.35×10^{-5}	$6.89\times$
Boundary	0.00	7.76×10^{-5}	1.18×10^{-5}	$6.57\times$
Exterior	-0.10	6.90×10^{-5}	1.26×10^{-5}	$5.47\times$

Key Observations and Theoretical Justification.

1. **No variance blowup at the boundary.** As η_c passes through zero, the MLMC estimator variance decreases monotonically rather than spiking. If the MLMC telescoping correction were sensitive to the constraint surface, we would expect amplified variance precisely at $\eta_c = 0$, where the gradient of the TD error with respect to η_c changes sign. We observe no such effect.
2. **Stable MLMC/MC ratio.** The variance ratio remains in a tight band ($5.47\times$ to $6.89\times$) across all regimes, confirming that the MLMC correction overhead is not adversely amplified near the boundary.
3. **Structural stability.** The estimator is structurally stable near $\eta_c = 0$ because the MLMC critic gradient enters η_c linearly through two terms (see Eq. (20)): the average-cost gradient $c_\gamma(\eta_c - c(s, a))$ and the TD error $\delta_c = Q_c(s, a) + \eta - c(s, a) - Q_c(s', a')$. Both are smooth and affine in η; there is no discontinuity, indicator function, or non-smooth projection in the gradient path as η_c crosses zero.

4. **Independent critics.** Furthermore, our algorithm trains completely separate critics for the reward (r) and cost (c). Thus, even when the dual variable λ fluctuates near the boundary, the target values for the critics remain bounded in $[-1, 1]$, inherently preventing the severe variance spikes commonly seen in single-critic Lagrangian methods.

F.2 COMPARISON WITH CRPO ON SAFE PENDULUM

We compare PDNAC-NC with CRPO [Xu et al., 2021], a representative primal baseline for average-reward CMDPs, on a continuous-control safety task. The environment is a safety-constrained pendulum: reward $r(s, a) = \frac{1}{2}(\cos\theta + 1) \in [0, 1]$ rewards uprightness, and the constraint $J_c(\theta) \geq 0$ enforces a speed limit $|\dot{\theta}| \leq 3.5$ through the cost signal $c(s, a) = \text{clip}(3.5 - |\dot{\theta}|, -1, 1)$. Constraint satisfaction is in direct tension with reward maximization, since large angular velocities are required to swing the pole upright.

Experimental Setup. Both methods use the same actor architecture (single-hidden-layer Gaussian MLP with 32 units, diagonal covariance), the same NTK critic parameterization (width $m = 64$, depth $L = 2$), and the same average-reward on-policy sampling scheme (32 parallel chains rolled for 32 steps per outer iteration). We run $K = 200$ outer iterations and select hyperparameters from a grid over the primal step size α, dual step size β, and NPG step size γ_ω.

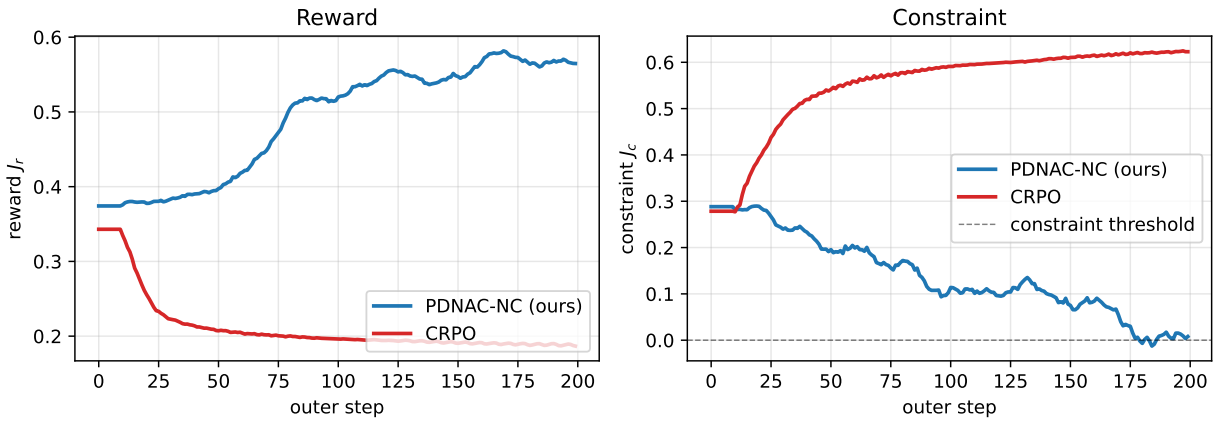

Figure 1: Reward (J_r, left) and constraint value (J_c, right) versus outer iteration on the safe pendulum task. PDNAC-NC (blue) tightens toward the constraint boundary while steadily increasing reward; CRPO (red) collapses to an overly conservative policy that satisfies the constraint by margin but achieves low reward. Curves are smoothed with a 10-step moving average.

Discussion. Figure 1 shows the two methods' training curves. PDNAC-NC increases J_r from 0.37 at initialization to a mean of 0.57 over the last 20 iterations while driving J_c down to a small violation of ≈ 0.011 near the feasibility boundary, i.e., it learns to operate close to the constraint surface, where the highest reward lies. CRPO's discrete switch between reward and cost updates makes it sensitive to step-size selection: with standard hyperparameters the switch becomes noise-dominated near the boundary and the policy retreats into the safe interior, yielding a final mean reward of 0.19 with $J_c \approx 0.62$. This illustrates the practical benefit of our continuous dual update, which smoothly balances the primal and safety directions rather than relying on a hard indicator.

G LIMITATIONS AND FUTURE WORK

While our work establishes strong theoretical guarantees for neural actor-critic methods in CMDPs, it presents several limitations that open avenues for future research:

Beyond the NTK Regime Our theoretical analysis heavily relies on the Neural Tangent Kernel (NTK) regime, which requires the neural network's width m to be highly overparameterized to ensure the parameters remain close to their initialization. While this allows for mathematically tractable linearized dynamics, it limits the network's ability to perform deep feature representation learning. A vital future direction is to extend this analysis beyond the lazy training regime,

potentially utilizing mean-field theory or representation learning frameworks to capture the true feature-learning capabilities of deep RL.

Ergodicity and Unichain Assumptions Our theoretical analysis relies on the uniform ergodicity assumption Assumption 3.1, which necessitates that the Markov chain induced by every parameterized policy is irreducible and aperiodic. In many practical safe RL domains, such as autonomous driving or robotics, policies may naturally lead to absorbing states (e.g., system failures) or possess disjoint transient regions. Currently, only MDPs with linear critics have been developed in the unichain setting [Ganesh and Aggarwal, 2025, Satheesh and Aggarwal, 2026]. Extending neural critic approaches to these settings remains an open problem.

Order-Optimal Convergence Rates We established a global convergence rate and cumulative constraint violation of $\tilde{\mathcal{O}}(T^{-1/4})$, which is not order-optimal natural actor-critic methods under Markovian sampling. The main bottleneck for CMDPs is the NTK analysis for the squared bias due to the projection operator. Extending our results by alleviating this bottleneck remains an interesting open problem.