Order-Optimal Global Convergence for Actor-Critic with General Policy and Neural Critic Parametrization

Swetha Ganesh[1,2] **Jiayu Chen**[3] **Washim Uddin Mondal**[4] **Vaneet Aggarwal**[1]

[1]Purdue University
[2]Indian Institute of Science
[3]Carnegie Mellon University
[4]Indian Institute of Technology, Kanpur

Abstract

This paper addresses the challenge of achieving optimal sample complexity in reinforcement learning for Markov Decision Processes (MDPs) with general policy parameterization and multi-layer neural network critics. Existing approaches either fail to achieve the optimal rate or require impractical assumptions, such as access to knowledge of mixing times or the linearity of the critic. We introduce the Natural Actor-Critic with Data Drop (NAC-DD) algorithm, which integrates Natural Policy Gradient methods with a Data Drop technique to mitigate statistical dependencies inherent in Markovian sampling. NAC-DD achieves an optimal sample complexity of $\tilde{\mathcal{O}}(1/\epsilon^2)$, marking a significant improvement over the previous state-of-the-art guarantee of $\tilde{O}(1/\epsilon^3)$. The algorithm employs a multi-layer neural network critic with differentiable activation functions, aligning with real-world applications where tabular policies and linear critics are insufficient. Our work represents the first to achieve order-optimal sample complexity for actor-critic methods with neural function approximation, continuous state and action spaces, and Markovian sampling. Empirical evaluations on benchmark tasks confirm the theoretical findings, demonstrating the practical efficacy of the proposed method.

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a powerful framework with broad applications across various domains such as robotics [Gonzalez et al., 2023], transportation [Al-Abbasi et al., 2019], communication networks [Agarwal et al., 2022], and healthcare [Tamboli et al., 2024], where autonomous systems learn optimal decision-making strategies through interaction with their environment. However, unlike many machine learning scenarios, the temporal dependence inherent in RL violates the assumption of independent and identically distributed (i.i.d.) samples, complicating theoretical analysis and convergence guarantees. Among RL approaches, actor-critic methods have garnered attention for their scalability and adaptability, yet they generally fall short in achieving optimal convergence rates. This paper aims to address this gap by analyzing sample complexity for discounted reward Markov Decision Processes (MDPs) with general parametrized policies and neural critic. The current state of the art in this area, [Gaur et al., 2024], reaches a sample complexity of $\tilde{O}(1/\epsilon^3)$ under Markovian sampling. This brings forth a central question:

Can we achieve an ϵ-globally optimal solution with a sample complexity of $\tilde{O}(1/\epsilon^2)$ in the Markovian sampling setting, using general parameterized policies and a multi-layer neural network parameterized critic?

In this paper, we answer this question in the affirmative by proposing an algorithm called Natural Actor-Critic with Data Drop (NAC-DD). We observe that the general policy and neural critic parametrization we consider are widely used in practice. In contrast, while tabular policies and linear critics have been extensively studied, they find limited practical application. Our work focuses on neural critics with differentiable activation functions (such as Sigmoid, ELU, and GeLU), which smoothly approximate ReLU and are commonly employed in real-world settings.

1.1 RELATED WORKS

Policy Gradient Approaches: Recent studies have established an optimal sample complexity of $\tilde{O}(1/\epsilon^2)$ for policy gradient approaches with general parameterizations, as seen in [Fatkhullin et al., 2023, Mondal and Aggarwal, 2024], though these methods rely on independent sampling for gradient estimation. Thus, their approaches for policy gra-

Table 1: This table summarizes the features of different actor-critic convergence results. Our result is the first to provide order-optimal sample complexity results of AC for an MDP setting with general/multi-layer neural network parametrization for the actor-critic, continuous state and action space, and Markovian sampling.

References	Global Optimality	Continuous State Action Space	Multi Layer NN AC	Markovian Sampling	Sample Complexity
[Xu et al., 2020b]	✓	✓	✗	✓	$\tilde{\mathcal{O}}(\epsilon^{-4})$
[Khodadadian et al., 2021]	✓	✗	✗	✓	$\tilde{\mathcal{O}}(\epsilon^{-3})$
[Xu et al., 2020a]	✓	✓	✗	✓	$\tilde{\mathcal{O}}(\epsilon^{-3})$
[Xu et al., 2021]	✓	✓	✗	✓	$\tilde{\mathcal{O}}(\epsilon^{-4})$
[Wang et al., 2019]	✓	✗	✗	✗	$\tilde{\mathcal{O}}(\epsilon^{-4})$
[Cayci et al., 2024]	✓	✗	✗	✗	$\tilde{\mathcal{O}}(\epsilon^{-4})$
[Fu et al., 2021]	✓	✗	✓	✗	$\tilde{\mathcal{O}}(\epsilon^{-6})$
[Tian et al., 2023]	✗	✗	✓	✓	$\tilde{\mathcal{O}}(\epsilon^{-2})$
[Gaur et al., 2024]	✓	✓	✓	✓	$\tilde{\mathcal{O}}(\epsilon^{-3})$
This work	✓	✓	✓	✓	$\tilde{\mathcal{O}}(\epsilon^{-2})$

dient estimation are not directly extendable to actor-critic framework with Markovian sampling.

Actor-Critic Approaches: For actor-critic algorithms, however, no existing work has yet achieved this optimal sample complexity when using multi-layer neural network parameterizations for both actor and critic. Table 1 provides a summary of key actor-critic approaches, categorizing algorithms by their achievement of global optimality, compatibility with continuous state and action spaces, use of general/multi-layer neural network parameterizations, and reliance on Markovian sampling. The current state of the art in this area, [Gaur et al., 2024], reaches a sample complexity of $\tilde{O}(1/\epsilon^3)$ under Markovian sampling.

Neural Policy Evaluation: A recent paper has established an optimal-order result for Q-learning with a fixed sample-generating policy and neural function approximation [Ke et al., 2024]. This work also provides valuable intermediate results on Q-value approximation, which we use in our analysis. However, in order to achieve order-optimal global convergence, we also require a bound on the bias of the Q-function estimates, which requires a substantially different analysis.

1.2 MAIN CONTRIBUTIONS AND CHALLENGES

In this paper, we propose an algorithm that integrates the Natural Actor-Critic method with a Data Drop (NAC-DD) technique that involves selecting only one out of every t_{mix} samples for updates, thereby reducing correlation among samples. We show that this approach achieves optimal sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ (Theorem 1).

To motivate why we use DD, we first take a look at a generic recursion:

$$x_{t+1} = x_t - \alpha(g(x_t) + M_t),$$

where $g(x_t)$ is linear in x_t and $\{M_t\}_{t\geq 0}$ is an ergodic Markov chain. Denote the solution of this update as x^*. It is known that with linear function approximation, Q-learning (with a fixed policy) is of this form. It is known that the *bias* of such an update, $\|\mathbb{E}[x_t] - x^*\|$, can be constant [Nagaraj et al., 2020]. However, by modifying this update by applying data drop, it is shown in the same paper that an optimal sample complexity can be achieved.

However, we note that when the neural approximations are used instead, the update becomes non-linear and the bias becomes much more difficult to analyze. In the linear case, bias can be characterized through a recursive formulation, facilitating a precise analysis (e.g., (29) in [Mondal and Aggarwal, 2024]). In contrast, for non-linear critics, such a direct characterization is not feasible. To address this, we adopt a linearized update approach, commonly employed in neural critic analysis. This approximation is justified when the neural network width is sufficiently large, aligning with the popular Neural Tangent Kernel (NTK) theory [Jacot et al., 2018]. Nevertheless, ensuring that the error introduced by this linearization does not accumulate requires a refined analytical approach.

A further challenge arises from the use of a projection operator in the critic update. This projection operator is essential when employing NTK-based analysis. Although the use of projections does not typically complicate the proof, we note that our analysis also requires the bias of the critic to decrease sufficiently fast to achieve an improved sample complexity. Standard arguments based on the non-expansiveness property of projections are insufficient to guarantee an order-optimal bias. To overcome this limitation, we provide a careful analysis, which provides sharp bounds for the convergence rate and bias of the critic (Lemmas 4 and 5).

Finally, we provide empirical evaluations to validate our theoretical findings and demonstrate the practical efficacy of the proposed NAC-DD algorithm (Section 6).

2 SETUP

This paper addresses an infinite-horizon, discounted reward reinforcement learning problem formulated as a Markov Decision Process (MDP), represented by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \rho, \gamma)$. In this framework, $\mathcal{S}$ indicates a general state space, $\mathcal{A}$ is the action space, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ the reward function. When an agent takes action a in state s, it transitions to a subsequent state s' with a probability $P(s'|s, a)$. The initial state distribution is specified by ρ, and γ represents the discount factor. A (stationary) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ defines the distribution over actions given the current state. This induces a transition function $P^\pi : \mathcal{S} \to \Delta(\mathcal{S})$, given by $P^\pi(s, s') = \sum_{a\in\mathcal{A}} P(s'|s, a)\pi(a|s)$ for all states $s, s' \in \mathcal{S}$. Under any policy π, the resulting state sequence forms a Markov chain. We also consider a parameterized family of policies Π, consisting of all policies π_θ with parameters $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^d$.

The objective of the agent is to find a parameter θ that maximizes the long-term reward function, defined as $J(\theta) := \mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|\pi_\theta\right]$ where the expectation is over the distribution of π_θ-induced trajectories emanating from the initial distribution, ρ. For notational simplicity, we ignore the dependence on ρ. This work employs an actor-critic method to optimize $J(\cdot)$. Before delving into the optimization process, we first introduce several key concepts.

The action-value (Q^{π_θ}) function corresponding to π_θ is defined $\forall(s, a) \in \mathcal{S} \times \mathcal{A}$ as

$$Q^{\pi_\theta}(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\middle| s_0 = s, a_0 = a, \pi_\theta\right] \quad (1)$$

We can then further define the state value function as

$$V^{\pi_\theta}(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s, a)],\ \forall s \in \mathcal{S}. \quad (2)$$

For any $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define the Bellman operator $\mathcal{T}^{\pi_\theta}$ for all (s, a) as

$$\mathcal{T}^{\pi_\theta} Q(s, a) := r(s, a) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a), a'\sim\pi_\theta(\cdot|s')}[Q(s', a')].$$

It is known that $\mathcal{T}^{\pi_\theta}$ is a γ-contraction under the infinity norm and Q^{π_θ} is the unique fixed point.

We assume the following throughout the paper.

Assumption 1. The Markov chain $\{s_t\}_{t\geq 0}$, induced by an arbitrary policy $\pi \in \Pi$ is ergodic.

It is well-established that if $\mathcal{M}$ is ergodic, then $\forall\theta \in \Theta$, there exists a unique stationary ρ-independent distribution, denoted as $d^{\pi_\theta} \in \Delta(\mathcal{S})$, which obeys $(P^{\pi_\theta})^\top d^{\pi_\theta} = d^{\pi_\theta}$. With this notation in place, we define the mixing time of an MDP.

The mixing time of an MDP $\mathcal{M}$ with respect to a policy parameter θ is defined as

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

where $\|\cdot\|_{\text{TV}}$ denotes the total variation distance.

Let $\tilde{d}^{\pi_\theta}_{\gamma,\rho}(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr\big(s_t = s|s_0 \sim \rho,\ \pi_\theta\big)$ be the discounted state-visitation frequency under policy π_θ with initial distribution ρ and discount factor γ. Define the modified transition kernel $\widetilde{P}(s'|s, a) := \gamma\, P(s'|s, a) + (1-\gamma)\,\rho(s')$, which corresponds to sampling $s' \sim P(\cdot|s, a)$ with probability γ and $s' \sim \rho$ otherwise.

It is known that if the MDP is ergodic, then $\tilde{d}^{\pi_\theta}_{\gamma,\rho}$ is the stationary distribution of the Markov chain with the π_θ-induced transition kernel $\widetilde{P}^{\pi_\theta}(\cdot|s) := \gamma P^{\pi_\theta}(\cdot|s) + (1-\gamma)\rho(\cdot)$ [Konda, 2000]. We define

$$\tilde{t}^\theta_{\text{mix}} := \min\left\{t \geq 1\middle| \|(\widetilde{P}^{\pi_\theta})^t(s, \cdot) - \tilde{d}^{\pi_\theta}\|_{\text{TV}} \leq \frac{1}{4}, \forall s \in \mathcal{S}\right\}$$

When the state space is finite, $\tilde{t}^\theta_{\text{mix}} = \mathcal{O}\big((1-\gamma)^{-1}\big)$.To see this, observe that $\widetilde{P}^{\pi_\theta}$ is a convex combination of P^{π_θ} and a rank 1 matrix. The bound follows using Corollary 1 of [Nussbaum, 2003] to bound the spectral gap, which provides a bound for the mixing time [Levin and Peres, 2017]. For convenience, we introduce

$$t_{\text{mix}} = \sup_{\theta\in\Theta} \max\{t^\theta_{\text{mix}}, \tilde{t}^\theta_{\text{mix}}\},$$

which serves as a uniform upper bound on both quantities.

3 NATURAL ACTOR–CRITIC WITH DATA DROP (NAC–DD)

Policy Gradient (PG)-type algorithms typically maximize the long-term reward function $J(\cdot)$ by updating θ along the gradient of $J(\cdot)$, which can be expressed in the following form using the well-known policy gradient theorem [Sutton et al., 1999a].

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\tilde{d}^{\pi_\theta}_{\gamma,\rho}, a\sim\pi_\theta(\cdot|s)}\Big[Q^{\pi_\theta}(s, a)\nabla_\theta \log \pi_\theta(a|s)\Big] \quad (3)$$

Natural Policy Gradient (NPG) methods, however, update θ along the NPG ω^*_θ instead, where

$$\omega^*_\theta = F(\theta)^\dagger \nabla_\theta J(\theta), \quad (4)$$

$\dagger$ denotes the Moore-Penrose pseudo-inverse and $F(\theta)$ is the Fisher information matrix as defined as:

$$F(\theta) = \mathbb{E}_{s\sim\tilde{d}^{\pi_\theta}_{\gamma,\rho}}\, \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s)(\nabla_\theta \log \pi_\theta(a|s))^\top\right]$$

The precoder $F(\theta)$ takes the change of the parameterized policy with respect to θ into account, thereby preventing overshooting or slow updates of θ. Note that ω^*_θ can be written as the minimizer of the function $L_{\pi_\theta}(\cdot, \theta)$ where

$$L_{\pi_\theta}(\omega, \theta) = \frac{1}{2}\,\mathbb{E}_{s\sim\tilde{d}^{\pi_\theta}_{\gamma,\rho}(\cdot), a\sim\pi(\cdot|s)}$$

$$\left[\left((1-\gamma)Q^{\pi_\theta}(s,a) - \omega^\top \nabla_\theta \log \pi_\theta(a|s)\right)^2\right] \quad (5)$$

for all $\omega \in \mathbb{R}^{\mathrm{d}}$. This is essentially a convex optimization that can be iteratively solved utilizing a gradient-based method. One can show that

$$\nabla_\omega L_{\pi_\theta}(\omega,\theta) = F(\theta)\omega - \nabla_\theta J(\theta) \quad (6)$$

Note that $\nabla_\omega L_{\pi_\theta}(\omega,\theta)$ is not exactly computable since the transition function P and hence the stationary distribution, $\tilde{d}^{\pi_\theta}_{\gamma,\rho}$, and the state-action value function, $Q^{\pi_\theta}(\cdot,\cdot)$ are typically unknown in most practical cases.

To estimate the policy gradient, we introduce a parameterized critic $Q(\phi(s,a);\zeta)$ in place of the true action-value function $Q^{\pi_\theta}(s,a)$. Here $\phi\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}^n$ is a fixed feature map and $\zeta\in\mathbb{R}^m$ the critic parameters.

In this paper, we consider a neural temporal difference learning method where the action-value function $Q^{\pi_{\theta_k}}(\cdot,\cdot)$ is parameterized by some multi-layer neural network. Let us define a feedforward neural network by the following recursion:

$$x^{(l)} = \frac{1}{\sqrt{m}}\sigma\left(W_l x^{(l-1)}\right), \quad l\in\{1,2,\cdots,L\}, \quad (7)$$

where $W_1\in\mathbb{R}^{m\times n}$, $W_l\in\mathbb{R}^{m\times m}$ for $2\le l\le L$ are the weight matrices of the network, $\sigma(\cdot)$ is an activation function, and $x^{(0)}=\phi(s,a)\in\mathbb{R}^n$. Using $x^{(L)}$ computed above, the approximate action-value function $Q(\phi(s,a);\zeta)$ can be computed as

$$Q(\phi(s,a);\zeta) = \frac{1}{\sqrt{m}} b^\top x^{(L)}, \quad (8)$$

where the parameter $\zeta = (\mathrm{Vec}(W_1);\cdots;\mathrm{Vec}(W_L))$ denotes the collection of all weight matrices, and b is given by a random initialization. The parameter b will not be optimized during training. Note that the RHS of the above equation depends on ζ via $x^{(L)}$. $\mathrm{Vec}(\cdot)$ stands for the vectorization operator that reshapes a matrix to a column vector by stacking its columns one by one and the ";" separator in ζ stands for the vertical stacking of the elements. That is, we reshape ζ to a long column vector for the notational convenience.

Assumption 2. The activation function $\sigma(\cdot)$ is L_1-Lipschitz and L_2-smooth, i.e. , for $\forall y_1, y_2\in\mathbb{R}$:

$$|\sigma(y_1)-\sigma(y_2)| \le L_1|y_1-y_2|$$

and

$$|\sigma'(y_1)-\sigma'(y_2)| \le L_2|y_1-y_2|.$$

Assumption 2 indicates that our results below are not based on the popular ReLU activation function. However, we primarily focus on some twice-differentiable activation functions (such as Sigmoid, ELU, GeLU, etc.), which are smooth approximations of the ReLU function and are frequently utilized in practical problems [Devlin et al., 2018, Godfrey, 2019]. Such a setup aligns with [Liu et al., 2020a], and provides a $\mathcal{O}(m^{-\frac{1}{2}})$-smooth property for the neural Q-function.

Let $\zeta_0 = \left(\mathrm{Vec}(W_1^0);\cdots;\mathrm{Vec}(W_L^0)\right)$ be the initial solution. For each l, we initialize the weights of W_l^0 element-wise from a normal distribution $\mathcal{N}(0,1)$ and each element of b is drawn uniformly from $\{-1,+1\}$. For regularity purpose, we would like to restrict the iterations to a bounded set around ζ_0, which is defined as

$$S_R := \Big\{\zeta = \left(\mathrm{Vec}(W_1);\ldots;\mathrm{Vec}(W_L)\right) \,\Big|\, \|\zeta-\zeta_0\|_2 \le R, \quad 1\le l\le L\Big\} \quad (9)$$

and denote the projection onto S_R as Π_R.

We now define the local linearization function class of the multi-layer Q network (8) at the random initialization ζ_0:

$$\mathcal{F}_{R,m} := \left\{\widehat{Q}(\cdot\,;\zeta) = Q(\cdot\,;\zeta_0) + \langle\nabla_\zeta Q(\cdot\,;\zeta_0), \zeta-\zeta_0\rangle\right\} \quad (10)$$

for any $\zeta\in S_R$.

For each policy parameter θ, define the mean-squared Bellman error under the on-policy state–action distribution by

$$E(\theta,\zeta) := \frac{1}{2}\sum_{s,a} d^{\pi_\theta}(s)\,\pi_\theta(a|s)\left[Q^{\pi_\theta}(s,a) - \widehat{Q}\big(\phi(s,a);\zeta\big)\right]^2$$

The critic parameter ζ_*^θ is then chosen (not necessarily uniquely) to minimize this error:

$$\zeta_*^\theta \in \arg\min_{\zeta\in\mathbb{R}^m} E(\theta,\zeta)\,.$$

A direct computation of ζ_*^θ is infeasible and instead, we perform stochastic gradient descent. Noting

$$\nabla_\zeta \tfrac{1}{2}\left[Q^{\pi_\theta}(s,a) - \widehat{Q}(\phi(s,a);\zeta)\right]^2 = \left[\widehat{Q}(\phi(s,a);\zeta) - Q^{\pi_\theta}(s,a)\right]\nabla_\zeta\widehat{Q}(\phi(s,a);\zeta), \quad (11)$$

we obtain the batch gradient

$$\nabla_\zeta E(\theta,\zeta) = \sum_{s,a} d^{\pi_\theta}(s)\pi_\theta(a|s)\cdot \left[Q^{\pi_\theta}(s,a) - \widehat{Q}(\phi(s,a);\zeta)\right]\nabla_\zeta\widehat{Q}(\phi(s,a);\zeta). \quad (12)$$

In practice, samples obtained from a contiguous trajectory induced by π_θ are used to form unbiased estimates of this gradient. The details of these estimates are given below.

3.1 ALGORITHMIC DESCRIPTION

We divide each outer epoch into two phases of equal length N. Within each phase we process data in contiguous blocks

of size $M = \kappa t_{\text{mix}} \lfloor \log_2 T \rfloor$, where $T = KN$ is the time horizon and $\kappa \geq 1$ is a user-chosen integer. Each phase thus comprises $H = \lceil N/M \rceil$ blocks.

Notation. Let s_t and a_t denote the state and action at time t. We write

$$x_h^k = \phi\big(s_{hM}^k, a_{hM}^k\big), \quad x_h'^{\,k} = \phi\big(s_{hM+1}^k, a_{hM+1}^k\big).$$

The temporal-difference error on block h of epoch k is

$$\Delta_h^k = Q(x_h^k; \zeta_h^k) - \big[r_{hM}^k + \gamma\, Q(x_h'^{\,k}; \zeta_h^k)\big], \tag{13}$$

and its gradient contribution

$$g_h^k(x_h^k; \zeta_h^k) = \Delta_h^k\, \nabla_\zeta Q(x_h^k; \zeta_h^k).$$

g_h^k serves as an estimate of $\nabla_\zeta E(\theta_k, \zeta_h^k)$.

For the NPG update, let $\bar{s}_h^k = s_{hM}^k$ and $\bar{a}_h^k = a_{hM}^k$. Define

$$\begin{aligned}\widehat{\nabla}_\omega L_h^k = {} & \nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)\, \big[\nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)\big]^\top \omega_{H+h}^k \\ & - Q\big(\phi(\bar{s}_h^k, \bar{a}_h^k); \zeta_N^k\big)\, \nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k).\end{aligned} \tag{14}$$

It can be seen that $\widehat{\nabla}_\omega L_h^k$ is an estimate of $\nabla_\omega L_h^k$.

Block updates. Putting it all together, for each block h in epoch k we perform the neural TD updates

$$\zeta_{h+1}^k = \Pi_R\big(\zeta_h^k - \beta\, g_h^k\big), \tag{15}$$

in the first phase followed by the NPG updates

$$\omega_{H+h+1}^k = \omega_{H+h}^k - \eta\, \widehat{\nabla}_\omega L_h^k, \tag{16}$$

and then finalize with the policy update $\theta_{k+1} = \theta_k + \alpha\, \omega_{2N}^k$.

Algorithm 1 Natural Actor–Critic with Data Drop (NAC–DD)

Input: Initial parameters θ_0, $\{\omega_H^k\}$, $\{\zeta_0^k\}$; policy step size α; NPG step size η; critic step size β; initial state $s_0 \sim \rho$; time horizon T; outer loops K; inner loop length H; discount factor γ; drop number $M = \kappa t_{\text{mix}} \lfloor \log_2 T \rfloor$.
Critic Initialization ζ_0: Sample each entry of $W_l^0 \sim \mathcal{N}(0,1)$ for $l = 1, \ldots, L$, and each entry of $b \sim \text{Unif}\{-1, +1\}$.
for $k = 0, \ldots, K-1$ **do**
  Set s_0^k to the final state of epoch $k-1$.
  for $h = 0, \ldots, H-1$ **do** ▷ Critic phase
    for $m = 0, \ldots, M-1$ **do**
      Sample $a_{hM+m}^k \sim \pi_{\theta_k}(\cdot|s_{hM+m}^k)$
      Sample $s_{hM+m+1}^k \sim P(\cdot|s_{hM+m}^k, a_{hM+m}^k)$
    end for
    Compute Δ_h^k and update $\zeta_{h+1}^k = \Pi_R\big(\zeta_h^k - \beta\, g_h^k\big)$
  end for
  for $h = 0, \ldots, H-1$ **do** ▷ NPG phase
    for $m = 0, \ldots, M-1$ **do**
      Sample $a_{hM+m}^k \sim \pi_{\theta_k}(\cdot|s_{hM+m}^k)$
      Sample $s_{hM+m+1}^k \sim \widetilde{P}(\cdot|s_{hM+m}^k, a_{hM+m}^k)$
    end for
    Compute $\widehat{\nabla}_\omega L_h^k$
    Update $\omega_{H+h+1}^k \leftarrow \omega_{H+h}^k - \eta\, \widehat{\nabla}_\omega L_h^k$
  end for
  Set $\omega_k \leftarrow \omega_{2H}^k$.
  Update $\theta_{k+1} \leftarrow \theta_k + \alpha\, \omega_k$. ▷ Policy update
end for

4 SAMPLE COMPLEXITY OF THE PROPOSED ALGORITHM

We first state some assumptions that we will be using before proceeding to the main result.

Assumption 3. The critic approximation error defined as

$$\epsilon_{\text{app}} := \sup_{\theta \in \Theta} \mathbb{E}[(Q^{\pi_\theta}(s,a) - \Pi_{\mathcal{F}_{R,m}} Q^{\pi_\theta}(s,a))^2], \tag{17}$$

where the expectation is over $s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)$, is assumed to be finite.

Assumption 4. There exist $\lambda_0 > 0$ such that $\forall \theta$

$$\mathbb{E}[\nabla_\zeta Q(\phi(s,a);\zeta_0) \nabla_\zeta Q(\phi(s,a);\zeta_0)^\top] \succcurlyeq \lambda_0 I.$$

where the expectation is over $s \sim d_{\gamma,\rho}^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)$. We will denote $\mathbb{E}[\nabla_\zeta Q(\phi(s,a);\zeta_0) \nabla_\zeta Q(\phi(s,a);\zeta_0)^\top]$ as Σ_{π_θ} henceforth.

Assumption 5. For any θ, the *transferred compatible function approximation error*, $L_{\pi^*}(\omega_\theta^*; \theta)$, satisfies the following inequality.

$$\begin{aligned}L_{\pi^*}(\omega_\theta^*; \theta) := \mathbb{E}_{s \sim d_{\gamma,\rho}^{\pi^*}, a \sim \pi^*(\cdot|s)} \Big[& (1-\gamma) A^{\pi_\theta}(s,a) \\ & - (\omega_\theta^*)^\top \nabla_\theta \log \pi_\theta(a|s) \Big]^2 \leq \epsilon_{\text{bias}},\end{aligned} \tag{18}$$

where π^* is an optimal policy for the discounted MDP $\mathcal{M}$ and ω_θ^* is the exact NPG direction at θ.

Assumption 6. For all $\theta, \theta_1, \theta_2 \in \Theta$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, the following statements hold:

(a) $\|\nabla_\theta \log \pi_\theta(a|s)\| \leq G_1$

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

Assumption 7 (Fisher non-degenerate policy)**.** There exists a constant $\mu > 0$ such that $F(\theta) - \mu I_d$ is positive semidefinite where I_d denotes an identity matrix.

Comments on Assumptions 3-4: Assumption 3 ensures that a class of neural networks can approximate the function obtained by applying the Bellman operator to another neural network within the same class. Similar assumptions have been considered in [Fu et al., 2021, Wang et al., 2019, Ke et al., 2024, Gaur et al., 2024]. In works such as [Cayci et al., 2024], even stronger assumptions are made, where the function class used for critic parameterization is assumed to approximate any smooth function.

Assumption 4 has been employed in prior works [Zou et al., 2019, Xu and Gu, 2020] and is closely related to the state regularity assumption, which similarly ensures a strong convexity-type property in the critic update [Tian et al., 2023, Gaur et al., 2024]. It can also be viewed as a generalization of the positive definite feature covariance matrix assumption in the analysis of linear Q-learning [Xu et al., 2019, Ganesh et al., 2025].

Comments on Assumptions 5-7: We would like to highlight that all these assumptions are commonly found in PG literature [Liu et al., 2020b, Agarwal et al., 2021, Papini et al., 2018, Xu et al., 2019, Fatkhullin et al., 2023]. We elaborate more on these assumptions below.

The term ϵ_{bias} captures the expressivity of the parameterized policy class. If the policy class is complete such as in the case of softmax parametrization, we have $\epsilon_{\text{bias}} = 0$ [Agarwal et al., 2021]. However, for restricted parametrization which may not contain all stochastic policies, we have $\epsilon_{\text{bias}} > 0$. It is known that ϵ_{bias} is insignificant for rich neural parametrization [Wang et al., 2019]. Assumption 6 requires that the score function is bounded and Lipschitz continuous. This assumption is widely used in the analysis of PG based methods [Liu et al., 2020b, Agarwal et al., 2021, Papini et al., 2018, Xu et al., 2019, Fatkhullin et al., 2023]. Assumption 7 requires that the eigenvalues of the Fisher information matrix can be bounded from below and is commonly used in obtaining global complexity bounds for PG based methods [Liu et al., 2020b, Zhang et al., 2021, Bai et al., 2022, Fatkhullin et al., 2023]. Assumptions 6-7 were shown to hold for various examples recently including Gaussian policies with linearly parameterized means and certain neural parametrizations [Liu et al., 2020b, Fatkhullin et al., 2023].

Theorem 1. *Consider Algorithm 1 with $K = \frac{1}{\epsilon}$, $H = \frac{1}{2t_{\text{mix}}\lfloor \log_2(1/\epsilon)\rfloor \epsilon}$ and $M = 2t_{\text{mix}}\lfloor \log_2(1/\epsilon)\rfloor$. If Assumptions 1-7 hold then there exists a choice of parameters such that the following holds for sufficiently small ϵ:*

$$J^* - \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1} \mathbb{E}[J(\theta_k)|\zeta_0] \leq \mathcal{O}\left(\frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} + \frac{\sqrt{\epsilon_{\text{app}}}}{1-\gamma} + \frac{t_{\text{mix}}\log^3\left(\frac{1}{\epsilon\delta}\right)}{(1-\gamma)^3}\cdot\epsilon + \frac{1}{m^{1/4}(1-\gamma)^{1/2}}\right).$$

with probability $1-2\delta-2L\exp(-Cm)$, for some constant $C > 0$. Here, m and L denote the width and depth of the critic neural network, respectively.

5 PROOF OUTLINE

We structure our analysis into three parts: policy update, NPG estimation, and critic analysis.

5.1 POLICY UPDATE ANALYSIS

We begin with a useful lemma from [Mondal and Aggarwal, 2024].

Lemma 1. *Consider any policy update rule of the form*

$$\theta_{k+1} = \theta_k + \alpha\omega_k. \tag{19}$$

If Assumptions 5 and 6 hold, then the following inequality is satisfied:

$$\begin{aligned} J^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[J(\theta_k)] &\leq \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} \\ &+ \frac{G_1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\,\mathbb{E}[\omega_k|\theta_k] - \omega_k^*\right\| + \frac{\alpha G_2}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\omega_k\right\|^2 \\ &+ \frac{1}{\alpha K}\mathbb{E}_{s\sim d^{\pi^*}}\left[\text{KL}(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right], \end{aligned} \tag{20}$$

where $\text{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence, ω_k^* is the NPG direction $F(\theta_k)^{-1}\nabla J(\theta_k)$, π^* is the optimal policy, and J^* is the optimal value of the function $J(\cdot)$.

The last term above is of order $\mathcal{O}(1/K)$ since

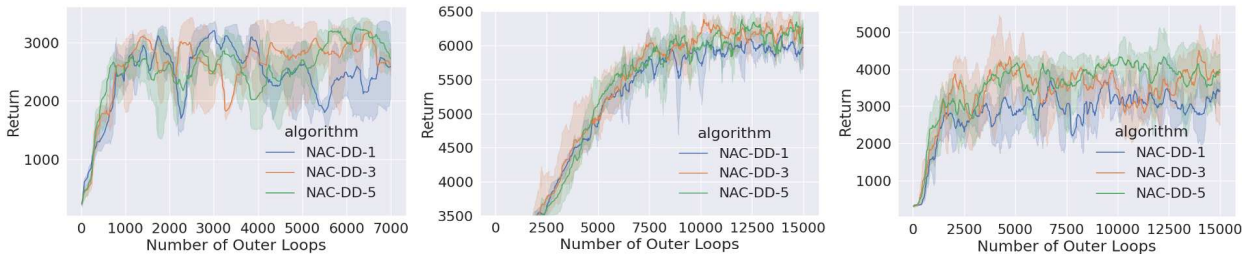
$$\mathbb{E}_{s\sim d^{\pi^*}}\left[\text{KL}(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))\right]$$

is constant. The term $\mathbb{E}\left\|\omega_k\right\|^2$ is further decomposed as:

$$\begin{aligned} \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\omega_k\right\|^2 &\leq \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\omega_k - \omega_k^*\right\|^2 \\ &+ \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\omega_k^*\right\|^2 \\ &\overset{(a)}{\leq} \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\omega_k - \omega_k^*\right\|^2 \\ &+ \frac{2\mu^{-2}}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla_\theta J(\theta_k)\right\|^2, \end{aligned} \tag{21}$$

where (a) follows from Assumption 7 and the definition $\omega_k^* = F(\theta_k)^{-1}\nabla_\theta J(\theta_k)$.

Thus, we can obtain a global convergence bound by bounding the terms $\mathbb{E}\left\|\omega_k - \omega_k^*\right\|^2$, $\mathbb{E}\left\|\,\mathbb{E}[\omega_k|\theta_k] - \omega_k^*\right\|$, and $\mathbb{E}\left\|\nabla_\theta J(\theta_k)\right\|^2$. The first two terms represent the second-order error and bias of the NPG estimator ω_k, and the third term indicates the local convergence rate. Since $\mathbb{E}\left\|\nabla_\theta J(\theta_k)\right\|^2$ can be expressed in terms of $\mathbb{E}\left\|\omega_k - \omega_k^*\right\|^2$, we now briefly describe how to bound these terms.



(a) Hopper-v3 (b) HalfCheetah-v3 (c) Walker2d-v3

Figure 1: Performance of NAC-DD on MuJoCo locomotion tasks with varying drop numbers (M). The results demonstrate that NAC-DD consistently achieves better performance when the drop number exceeds 1.

5.2 NPG ANALYSIS

In this section, we derive bounds on the second-order error and bias of the NPG estimator ω_k. For any policy π_{θ_k}, the critic subroutine's fixed point need not be unique. Let

$$Z_k = \left\{\zeta : \widehat{Q}(\cdot;\zeta) \text{ is a fixed point of } \Pi_{\mathcal{F},m}\mathcal{T}^{\pi_{\theta_k}}\right\},$$

and let ζ_*^k be the projection of the initial critic parameter ζ_0 onto Z_k. We will show that the algorithm's iterates closely track ζ_*^k. Finally, denote by $\mathbb{E}_k[\cdot]$ the expectation conditioned on θ_k.

Lemma 2 (Second-order error of NPG estimator)**.** *Consider the NPG-finding recursion* (16) *with* $\eta = \frac{2\log H}{\mu H}$*. If all assumptions in Theorem 1 hold, then for sufficiently large* H*,*

$$\mathbb{E}_k[\|\omega_k - \omega_k^*\|^2|\zeta_0] \leq \mathcal{O}\left(\frac{G_1^2(C_1')^2\log(H/\delta)}{H\mu^2(1-\gamma)^4} + \mu^{-2}m^{-1/2}\right.$$
$$\left. + \mu^{-2}G_1^2\,\mathbb{E}_k[\|\zeta_H^k - \zeta_*^k\|^2|\zeta_0] + \frac{G_1^2\epsilon_{\text{app}}}{\mu^2(1-\gamma)^2}\right)$$

Lemma 3 (Bias of NPG estimator)**.** *Consider the NPG-finding recursion* (16) *with* $\eta = \frac{2\log H}{\mu H}$*. If all assumptions in Theorem 1 hold, then for sufficiently large* H*, we have the following bound with probability* $1 - 2\delta - 2L\exp(-Cm)$*, for some constant* $C > 0$

$$\|\,\mathbb{E}_k[\omega_k|\zeta_0] - \omega_k^*\|^2 \leq \mathcal{O}\left(\frac{G_1^2(C_1')^2G_1^2\log(H/\delta)}{T^\kappa}\right.$$
$$\left. + \|\,\mathbb{E}_k[\zeta_H^k|\zeta_0] - \zeta_*^k\|^2 + \frac{G_1^2\epsilon_{\text{app}}}{\mu^2(1-\gamma)^2}\right)$$

The proof of this result can be found in Appendix B. Since the NPG estimator ω_k uses the critic values, the above bounds depend on the second-order error and bias of the critic estimator. The bounds for these quantities are provided in the next section.

5.3 CRITIC ANALYSIS

In this section, we focus on providing bounds for the second-order error and bias of the critic estimator ζ_H^k. A second-order error bound of $\mathcal{O}(\frac{1}{\epsilon})$ for Q-learning with neural approximation was recently studied in [Ke et al., 2024], without requiring strict positive definiteness as in Assumption 3. Instead, we present an alternative analysis of this result that enables us to also derive a bound on the critic's bias. The proof of this result can be found in Appendix A.1.

Lemma 4 (Second-order error of the Critic)**.** *Consider Algorithm 1 and let* $\beta = \frac{2\log H}{\lambda H}$*. If all assumptions of Theorem 1 hold, then for sufficiently large* H*,*

$$\mathbb{E}[\|\zeta_H^k - \zeta_*^k\|^2|\zeta_0] \leq \mathcal{O}\left(\frac{\mathbb{E}\,\|\zeta_0 - \zeta_*^k\|^2}{H^2} + \right.$$
$$\left.\frac{\log^2(H/\delta)}{\lambda_0^2(1-\gamma)^2H} + \frac{\log(H/\delta)}{\lambda_0(1-\gamma)m^{1/2}} + \frac{1}{(1-\gamma)^4\lambda_0^4T^\kappa}\right)$$

with probability $1 - 2\delta - 2L\exp(-Cm)$*, for some constant* $C > 0$*.*

Analyzing the bias forms a key challenge due to the nonlinearity of the critic update due to the neural network and due to the presence of the projection operator. The proof details of the following result can be found in Appendix A.2.

Lemma 5 (Bias of the Critic estimator)**.** *Consider Algorithm 1 and let* $\beta = \frac{2\log H}{\lambda H}$*. If all assumptions of Theorem 1 hold, then the following is true for sufficiently large* H*.*

$$\|\,\mathbb{E}[\zeta_{h+1}^k|\zeta_0] - \zeta_*^k\|^2 \leq \mathcal{O}\left(\frac{\|\,\mathbb{E}[\zeta_0] - \zeta_*^k\|^2}{\lambda_0^2(1-\gamma)^2H^2}\right.$$
$$\left. + \frac{\log^4(H/\delta)}{\lambda_0^6(1-\gamma)^6H^2} + \frac{\sqrt{\log(H/\delta)}}{\lambda_0(1-\gamma)m^{1/2}} + \frac{1}{(1-\gamma)^{10}\lambda_0^{10}T^{2\kappa}}\right)$$

with probability $1 - 2\delta - 2L\exp(-Cm)$*, for some constant* $C > 0$*.*

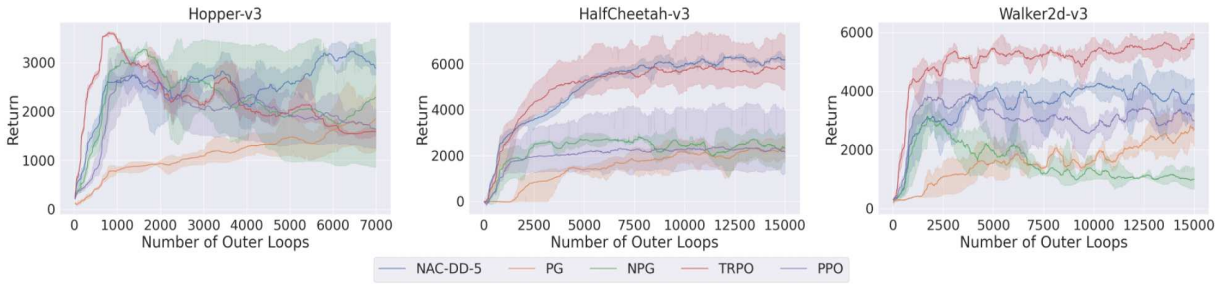


Figure 2: Comparison of NAC-DD and standard policy gradient algorithms on various MuJoCo locomotion tasks. Here, NAC-DD-5 represents NAC-DD with a drop number of 5. Our algorithm achieves the best performance on two out of three tasks and ranks second on the remaining task.

Table 2: Training time (in hours) for each algorithm on MuJoCo benchmarks, measured on a single NVIDIA GeForce RTX 2080 Ti GPU

Algorithm	NAC-DD-1	NAC-DD-3	NAC-DD-5	PG	NPG	TRPO	PPO
Hopper	5.44 ± 0.39	12.0 ± 0.37	17.1 ± 0.51	2.95 ± 0.02	3.30 ± 0.15	3.34 ± 0.03	4.22 ± 0.09
HalfCheetah	10.5 ± 0.36	19.1 ± 0.33	27.7 ± 0.34	6.68 ± 0.02	7.00 ± 0.04	7.49 ± 0.06	9.82 ± 0.04
Walker2d	12.4 ± 0.31	23.7 ± 0.30	35.1 ± 0.41	6.94 ± 0.08	7.04 ± 0.13	8.57 ± 0.14	9.74 ± 0.35

It can be seen that substituting Lemmas 4 and 5 in Lemmas 2 and 3 with the bound on the policy update in Lemma 1 yields Theorem 1. Based on Lemmas 2, 3, 4 and 5, we observe that it is sufficient to set $\kappa = 2$ to obtain the desired result.

6 EVALUATION

To demonstrate the effectiveness of the proposed algorithm, NAC-DD, we compare its performance against several standard policy gradient methods, including Vanilla Policy Gradient (PG) [Sutton et al., 1999b], Natural Policy Gradient (NPG) [Kakade, 2001], TRPO [Schulman et al., 2015], and PPO [Schulman et al., 2017]. For the baseline implementations, we utilized the open-source repository available at https://github.com/reinforcement-learning-kr/pg_travel. The evaluation is conducted on three benchmark MuJoCo locomotion tasks: HalfCheetah-v3, Hopper-v3, and Walker2d-v3, all of which are continuous control problems. Notably, the codes for reproducing all results in this paper is available at https://github.com/LucasCJYSDL/NAC-DD.

We begin by evaluating the impact of the key hyperparameter—the drop number M in Algorithm 1—on the performance of NAC-DD. As illustrated in Figure 1, we set M to 1, 3, and 5, and record the training progress of these variants on the MuJoCo tasks. Each experiment is repeated three times with different random seeds, with the means and 95% confidence intervals shown as solid lines and shaded areas, respectively. The results indicate that NAC-DD consistently achieves better performance when the drop number exceeds one. This improvement is attributed to the fact that dropping training samples helps mitigate the statistical dependency between samples from different time steps, aligning with the theoretical requirements. The performance of the algorithm with drop numbers of three and five is comparable. However, we anticipate that a larger drop number could be more beneficial for addressing more challenging control tasks (than MuJoCo).

In Figure 2, we position our algorithm by comparing it against standard policy gradient methods on various MuJoCo tasks. The results show that natural-policy-gradient-based methods consistently outperform the vanilla policy gradient approach[1] Additionally, actor-critic methods (i.e., NAC-DD, PPO, and TRPO) generally outperform pure policy gradient methods (i.e., NPG and PG). Notably, our algorithm achieves the best performance on two out of three tasks and ranks second on the third task. Thus, while this is a theory-focused paper with the algorithm built on solid theoretical foundations, its strong practical performance in challenging continuous control tasks further demonstrates its effectiveness and applicability.

Finally, for completeness, Table 2 reports the training time for each algorithm on the Hopper, HalfCheetah, and Walker2d benchmarks using a single NVIDIA GeForce RTX 2080 Ti GPU. The table above reports the training time (in hours) for each algorithm on each benchmark, using a single NVIDIA GeForce RTX 2080 Ti GPU. NAC-DD-5 requires more training time because it discards 80% of the collected

[1]For practical implementation, we estimate the natural policy gradient, as defined in Equation (4), and update the actor accordingly using the conjugate gradient method combined with backtracking line search.

samples, using only the remaining 20% for training. To ensure the total number of training samples is comparable to other methods, NAC-DD must collect more transitions. However, this additional sampling can be parallelized using a vectorized environment setup. In terms of computation time for policy and critic updates, NAC-DD is comparable to TRPO and PPO, as demonstrated by the results of NAC-DD-1 in relation to the other methods.

7 CONCLUSIONS

In this work, we address the challenge of achieving optimal sample complexity in reinforcement learning for Markov Decision Processes (MDPs) with general policy parameterization and multi-layer neural network critics. Existing methods either fall short of achieving the optimal rate or rely on linear critic approximations. To overcome these limitations, we introduce Natural Actor-Critic with Data Drop (NAC-DD) algorithm, which integrates Natural Policy Gradient methods with a Data Drop technique to mitigate statistical dependencies inherent in Markovian sampling. By achieving an optimal sample complexity of $\tilde{O}(1/\epsilon^2)$, our approach significantly improves upon the previous state-of-the-art guarantee of $\tilde{O}(1/\epsilon^3)$, marking a pivotal advancement in the field.

ACKNOWLEDGEMENT

This work was supported in part by the Anusandhan National Research Foundation (ANRF), India, through the Overseas Visiting Doctoral Fellowship and the U.S. National Science Foundation under grant CCF-2149588.

References

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

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

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.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:3682–3689, Jun. 2022. doi: 10.1609/aaai.v36i4.20281.

Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*, April 2024. URL https://openreview.net/forum?id=BkEqk7pS1I.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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. In *9th International Conference on Learning Representations, ICLR*, 2021.

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

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

Luke B Godfrey. An evaluation of parametric activation functions for deep learning. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3006–3011. IEEE, 2019.

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.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Zhifa Ke, Zaiwen Wen, and Junyu Zhang. An improved finite-time analysis of temporal difference learning with deep neural networks. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=dqdctVbSfs.

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.

Vijay Konda. *Actor-critic algorithms*. PhD thesis, University of Alberta, 2000.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33:15954–15964, 2020a.

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, 2020b.

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, 2024.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pages 16666–16676, 2020.

Roger Nussbaum. Notes on the second eigenvalue of the google matrix, 2003. URL https://arxiv.org/abs/math/0307056.

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.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015. URL http://arxiv.org/abs/1502.05477.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

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, 1999a.

Richard S. Sutton, David A. McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1057–1063. The MIT Press, 1999b.

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 Ioannis Paschalidis. Convergence of actor-critic with multi-layer neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 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*, 2019.

Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10555–10565. PMLR, 13–18 Jul 2020.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2019.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020a.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020b.

Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. In *International Conference on Machine Learning*, pages 11581–11591. PMLR, 2021.

Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. In

A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

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

A NEURAL CRITIC ANALYSIS

Let $\hat{g}(x_h^k;\zeta) \coloneqq (\widehat{Q}(x_h^k;\zeta) - (r_h^k + \gamma\widehat{Q}({x'}_h^k;\zeta)))\nabla_\zeta Q(x_h^k;\zeta_0)$ be the linearization of $g(x_h^k;\zeta)$ at ζ_0. For brevity, we denote $\phi(s_h^k, a_h^k)$ by x_h^k. It can be seen that

$$\hat{g}(x_h^k;\zeta) = A(x_h^k)\zeta - b(x_h^k), \tag{22}$$

where

$$A(x_h^k) = \nabla_\zeta Q(x_h^k;\zeta_0)(\nabla_\zeta Q(x_h^k;\zeta_0) - \gamma\nabla_\zeta Q({x'}_h^k;\zeta_0))^\top \tag{23}$$

and

$$b(x_h^k) = (r_h^k + \gamma Q({x'}_h^k;\zeta_0) - Q(x_h^k;\zeta_0))\nabla_\zeta Q(x_h^k;\zeta_0). \tag{24}$$

Define $A_k \coloneqq \mathbb{E}_{s\sim d^{\pi_{\theta_k}}, a\sim\pi_{\theta_k}(\cdot|s)}[A(x_h^k)|\zeta_0]$ and $b_k \coloneqq \mathbb{E}_{s\sim d^{\pi_{\theta_k}}, a\sim\pi_{\theta_k}(\cdot|s)}[b(x_h^k)|\zeta_0]$. For notational convenience, we henceforth drop the conditional expectation $\mathbb{E}[\cdot|\zeta_0]$ and instead write denote $\mathbb{E}[\cdot]$ instead.

We now state a useful lemma summarizing various properties of the Q-value estimator below:

Lemma 6. *Fix an outer iteration index k. Let $\zeta_h^k \in S_R$ for all $h \in \{0, 1, 2, \ldots, H\}$, where the radius R satisfies $R = \mathcal{O}(1)$. Then, for all $h \in \{1, 2, \ldots, H\}$, there exist positive constants C, C_1' and $\{C_i\}_{i=1,2,\cdots,5}$ such that the following statements hold with a probability of at least $1 - \delta - 2L\exp(-Cm)$.*

(a) $\left\|\nabla_\zeta Q(x_h^k;\zeta_h^k)\right\| \le C_1$, $|Q(x_h^k;\zeta_h^k)| \le C_1'\sqrt{\log(H/\delta)}$

(b) $\|g(x_h^k;\zeta_h^k) - \hat{g}(x_h^k;\zeta_h^k)\| \le C_2 m^{-\frac{1}{2}}\sqrt{\log(H/\delta)}$

(c) $\left|\left\langle g\left(x_h^k;\zeta_h^k\right) - \hat{g}\left(x_h^k;\zeta_h^k\right), \zeta_h^k - \zeta_*\right\rangle\right| \le C_3 m^{-\frac{1}{2}}\sqrt{\log(H/\delta)}$

(d) $|\widehat{Q}(x_h^k;\zeta_h^k) - Q(x_h^k;\zeta_h^k)| \le C_4 m^{-\frac{1}{2}}\sqrt{\log(H/\delta)}$

(e) $\|\nabla_\zeta Q(x_h^k;\zeta_0) - \nabla_\zeta Q(x_h^k;\zeta_h^k)\| \le C_5 m^{-\frac{1}{2}}\sqrt{\log(H/\delta)}$

Statements (a)-(e) follow from results in Ke et al. [2024]: Statement (a) from Lemmas D.2 and D.3, Statements (b) and (c) from Lemma D.5, and Statements (d) and (e) from Lemma D.4. Building on the above result, we obtain the following bounds.

Lemma 7. *There exist positive constants $c_1, c_2 > 0$ such that the following bounds for each h, k hold under the assumptions stated in Theorem 1 with a probability of at least $1 - \delta - 2L\exp(-Cm)$.*

1. $\|A(x_h^k)\| \le c_1$
2. $\|b(x_h^k)\| \le c_2\sqrt{\log(H/\delta)}$
3. $\|\mathbb{E}[A(x_h^k)] - A_k\| \le c_1 T^{-\kappa}$
4. $\|\mathbb{E}[b(x_h^k)] - b_k\| \le c_2\sqrt{\log(H/\delta)}T^{-\kappa}$

Proof. Note that from Lemma 6(a)

$$\|A(x_h^k)\| \le \|\nabla_\zeta Q(x_h^k;\zeta_0) - \gamma\nabla_\zeta Q({x'}_h^k;\zeta_0)\|\|\nabla_\zeta Q(x_h^k;\zeta_0)\| \le (1+\gamma)C_1^2. \tag{25}$$

Statement 1 follows by setting $c_1 = (1+\gamma)C_1^2$. Again, from Lemma 6(a)

$$\|b(x_h^k)\| \le |r_h^k + Q(x_h^k;\zeta_0) - \gamma Q({x'}_h^k;\zeta_0)|\|\nabla_\zeta Q(x_h^k;\zeta_0)\| \le (1 + C_1' + \gamma C_1')C_1\sqrt{\log(H/\delta)}.$$

and by setting $c_2 = (1 + C_1' + \gamma C_1')C_1$, Statement 2 follows. For Statement 3, observe that

$$\mathbb{E}[A(x_h^k)] - A_k = \sum_{x_h^k} A(x_h^k)((P^{\pi_{\theta_k}})^M(s_{(h-1)M}^k, s_{hM}^k) - d^{\pi_{\theta_k}}(s_{hM}^k))\pi(a_{hM}^k|s_{hM}^k). \tag{26}$$

Since $M = \kappa t_{\text{mix}} \lceil \log_2 T \rceil$

$$\| \mathbb{E}[A(x_h^k)] - A_k \| \leq c_1 \sum_{x_h^k} |(P^{\pi_{\theta_k}})^M (s_{(h-1)M}^k, s_{hM}) - d^{\pi_{\theta_k}}(s_{hM})| \pi(a_{hM}^k | s_{hM}) \leq \frac{c_1}{T^\kappa}. \tag{27}$$

Statement 4 follows along similar lines. □

Lemma 8. *Fix k and let assumptions in Theorem 1 hold. Then the following holds $\forall \zeta \in \ker(A_k)^\perp$*

$$\zeta^\top A_k \zeta \geq (1-\gamma)\lambda_0 \|\zeta\|^2 \tag{28}$$

Proof.

$$\begin{aligned}
\zeta^\top A_k \zeta &= \zeta^\top \mathbb{E}[\nabla_\zeta Q(x_h^k;\zeta_0) \nabla_\zeta Q(x_h^k;\zeta_0)^\top - \gamma \nabla_\zeta Q(x_h'^k;\zeta_0) \nabla_\zeta Q(x_h^k;\zeta_0)^\top]\zeta \\
&= \mathbb{E}[(\nabla_\zeta Q(x_h^k;\zeta_0)^\top \zeta)^2] - \gamma \, \mathbb{E}[(\nabla_\zeta Q(x_h'^k;\zeta_0)^\top \zeta)(\nabla_\zeta Q(x_h^k;\zeta_0)^\top \zeta)] \\
&\overset{(a)}{\geq} \mathbb{E}[(\nabla_\zeta Q(x_h^k;\zeta_0)^\top \zeta)^2] - \gamma (\mathbb{E}[(\nabla_\zeta Q(x_h'^k;\zeta_0)^\top \zeta)^2] \, \mathbb{E}[(\nabla_\zeta Q(x_h^k;\zeta_0)^\top \zeta)^2])^{1/2} \\
&\overset{(b)}{=} (1-\gamma) \, \mathbb{E}[(\nabla_\zeta Q(x_h^k;\zeta_0)^\top \zeta)^2] \\
&\overset{(c)}{\geq} (1-\gamma)\lambda_0 \|\zeta\|^2
\end{aligned} \tag{29}$$

where (a) follows from Cauchy-Schwartz inequality, (b) follows since x_h^k and $x_h'^k$ have the same marginal distribution and (c) follows from Assumption 3. □

A.1 PROOF OF LEMMA 4

We begin by introducing some notation. Let Λ_A, Λ_b, δ_A, and δ_b be positive constants such that $|\mathbb{E}[A(x_h^k)] - A_k| \leq \delta_A$, $|\mathbb{E}[b(x_h^k)] - b_k| \leq \delta_b$, $|A(x_h^k)| \leq \Lambda_A$, and $|b(x_h^k)| \leq \Lambda_b$. The values of these quantities are provided in Lemma 7.

We introduce an auxiliary sequence $\{\tilde{\zeta}_h^k\}_{h \geq 0}$ that replaces the neural update g with its linear approximation $\hat{g}$. Specifically, define

$$\tilde{\zeta}_0^k = \zeta_0^k \equiv \zeta_0, \quad \tilde{\zeta}_{h+1}^k = \Pi_R \left(\tilde{\zeta}_h^k - \beta \hat{g}(x_h^k; \tilde{\zeta}_h^k) \right),$$

where Π_R denotes the projection onto the ball of radius R centered at ζ_0, and $\beta > 0$ is a step-size parameter.

Let $\Pi_\perp$ denote the orthogonal projection onto $\ker(A_k)^\perp$. We now bound the expected discrepancy between the auxiliary and original iterates:

$$\begin{aligned}
\mathbb{E} \left\| \tilde{\zeta}_{h+1}^k - \zeta_{h+1}^k \right\|^2 &= \mathbb{E} \left\| \Pi_R \left(\zeta_h^k - \beta g(x_h^k;\zeta_h^k) \right) - \Pi_R \left(\tilde{\zeta}_h^k - \beta \hat{g}(x_h^k;\tilde{\zeta}_h^k) \right) \right\|^2 && (30) \\
&\leq \mathbb{E} \left\| \zeta_h^k - \beta g(x_h^k;\zeta_h^k) - \left(\tilde{\zeta}_h^k - \beta \hat{g}(x_h^k;\tilde{\zeta}_h^k) \right) \right\|^2 && (31) \\
&\leq \mathbb{E} \left\| \zeta_h^k - \beta \hat{g}(x_h^k;\zeta_h^k) - \left(\tilde{\zeta}_h^k - \beta \hat{g}(x_h^k;\tilde{\zeta}_h^k) \right) \right\|^2 + \beta C_2 m^{-1/2} \sqrt{\log(H/\delta)} && (32) \\
&= \mathbb{E} \left\| \left(\zeta_h^k - \tilde{\zeta}_h^k \right) - \beta A(x_h^k) \left(\zeta_h^k - \tilde{\zeta}_h^k \right) \right\|^2 + \beta C_2 m^{-1/2} \sqrt{\log(H/\delta)} && (33) \\
&= \mathbb{E} \left\| \zeta_h^k - \tilde{\zeta}_h^k \right\|^2 - 2\beta \, \mathbb{E} \left\langle \zeta_h^k - \tilde{\zeta}_h^k, A(x_h^k)(\zeta_h^k - \tilde{\zeta}_h^k) \right\rangle && (34) \\
&\quad + \beta^2 \, \mathbb{E} \left\| A(x_h^k)(\zeta_h^k - \tilde{\zeta}_h^k) \right\|^2 + \beta C_2 m^{-1/2} \sqrt{\log(H/\delta)} && (35) \\
&\leq \mathbb{E} \left\| \zeta_h^k - \tilde{\zeta}_h^k \right\|^2 - 2\beta \, \mathbb{E} \left\langle \zeta_h^k - \tilde{\zeta}_h^k, A_k(\zeta_h^k - \tilde{\zeta}_h^k) \right\rangle && (36) \\
&\quad + 2\beta\delta_A \, \mathbb{E} \left\| \zeta_h^k - \tilde{\zeta}_h^k \right\|^2 + \beta^2 \, \mathbb{E} \left\| A(x_h^k)(\zeta_h^k - \tilde{\zeta}_h^k) \right\|^2 + \beta C_2 m^{-1/2} \sqrt{\log(H/\delta)} && (37)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left\|\zeta_h^k - \tilde{\zeta}_h^k\right\|^2 - 2\beta\,\mathbb{E}\left\langle \Pi_\perp(\zeta_h^k - \tilde{\zeta}_h^k), A_k \Pi_\perp(\zeta_h^k - \tilde{\zeta}_h^k)\right\rangle && (38)\\
&\quad + 2\beta R^2 \delta_A + \beta^2 \Lambda_A^2\,\mathbb{E}\left\|\Pi_\perp(\zeta_h^k - \tilde{\zeta}_h^k)\right\|^2 + \beta C_2 m^{-1/2}\sqrt{\log(H/\delta)} && (39)\\
&\leq \mathbb{E}\left\|\zeta_h^k - \tilde{\zeta}_h^k\right\|^2 + (\beta^2\Lambda_A^2 - 2\beta\mu_A)\,\mathbb{E}\left\|\Pi_\perp(\zeta_h^k - \tilde{\zeta}_h^k)\right\|^2 && (40)\\
&\quad + 2\beta R^2 \delta_A + \beta C_2 m^{-1/2}\sqrt{\log(H/\delta)} && (41)\\
&\leq \mathbb{E}\left\|\zeta_h^k - \tilde{\zeta}_h^k\right\|^2 + 2\beta R^2 \delta_A + \beta C_2 m^{-1/2}\sqrt{\log(H/\delta)} && (42)\\
&\leq \mathbb{E}\left\|\zeta_0^k - \tilde{\zeta}_0^k\right\|^2 + 2\beta(h+1) R^2 \delta_A + \beta(h+1) C_2 m^{-1/2}\sqrt{\log(H/\delta)} && (43)\\
&= 2(h+1)\beta R^2 \delta_A + \beta(h+1) C_2 m^{-1/2}\sqrt{\log(H/\delta)} && (44)\\
&\leq 2(h+1)\beta R^2 c_1 T^{-\kappa} + \beta(h+1) C_2 m^{-1/2}\sqrt{\log(H/\delta)}, && (45)
\end{aligned}
$$

where we used the non-expansiveness of Π_R and the approximation error bound between g and $\hat{g}$. Substituting $\beta = \frac{2\log H}{\lambda_0(1-\gamma)H}$, we obtain the following result for all $h \in \{1, 2, \cdots, H\}$:

$$
\mathbb{E}\left\|\tilde{\zeta}_h^k - \zeta_h^k\right\|^2 \leq \mathcal{O}\left(\frac{R^2 c_1 \log H}{\lambda_0(1-\gamma)T^\kappa} + C_2 m^{-1/2}\sqrt{\log(H/\delta)} + \right).
$$

Lemma 9. *Let $Z_k := \{z \in \mathbb{R}^d : z = A_k^\dagger b_k + v,\ v \in \ker(A_k)\}$ denote the set of minimum-norm least-squares solutions to $A_k z \approx b_k$, and let ζ_*^k be the projection of a fixed point $\zeta_0 \in \mathbb{R}^d$ onto Z_k. Then, under the update rule*

$$
\tilde{\zeta}_h^k = \Pi_R\left(\tilde{\zeta}_{h-1}^k - \beta\hat{g}(x_h^k; \tilde{\zeta}_h^k)\right),
$$

where $\hat{g}(x_h^k; \tilde{\zeta}_h^k) \in \ker(A_k)^\perp$, it holds that

$$
\tilde{\zeta}_h^k - \zeta_*^k \in \ker(A_k)^\perp, \quad \textit{for all } h \geq 0.
$$

Proof. The set $Z_k = A_k^\dagger b_k + \ker(A_k)$ is an affine subspace, and ζ_*^k is the projection of ζ_0 onto Z_k. By the projection theorem for affine spaces, we have:

$$
\zeta_0 - \zeta_*^k \in \ker(A_k)^\perp.
$$

We proceed by induction on h.

Base case ($h = 0$): By initialization, $\tilde{\zeta}_0^k = \zeta_0$, hence

$$
\tilde{\zeta}_0^k - \zeta_*^k = \zeta_0 - \zeta_*^k \in \ker(A_k)^\perp.
$$

Inductive step: Assume that $\tilde{\zeta}_{h-1}^k - \zeta_*^k \in \ker(A_k)^\perp$. Define the intermediate iterate:

$$
\hat{\zeta}_h^k := \tilde{\zeta}_{h-1}^k - \beta\hat{g}(x_h^k; \tilde{\zeta}_h^k).
$$

Since $\hat{g}(x_h^k; \tilde{\zeta}_h^k) \in \ker(A_k)^\perp$ and $\tilde{\zeta}_{h-1}^k - \zeta_*^k \in \ker(A_k)^\perp$ by the inductive hypothesis, we conclude:

$$
\hat{\zeta}_h^k - \zeta_*^k = (\tilde{\zeta}_{h-1}^k - \zeta_*^k) - \beta\hat{g}(x_h^k; \tilde{\zeta}_h^k) \in \ker(A_k)^\perp.
$$

Now consider the projection operator Π_R, defined as:

$$\Pi_R(\zeta) = \begin{cases} \zeta, & \text{if } \|\zeta - \zeta_0\| \le R, \\ \zeta_0 + R \cdot \dfrac{\zeta - \zeta_0}{\|\zeta - \zeta_0\|}, & \text{otherwise.} \end{cases}$$

This operation returns a point on the line segment between ζ_0 and ζ, and since both $\zeta_0 - \zeta_*^k$ and $\hat{\zeta}_h^k - \zeta_*^k$ lie in $\ker(A_k)^\perp$, which is a linear subspace (and hence convex), we have:

$$\Pi_R(\hat{\zeta}_h^k) - \zeta_*^k \in \ker(A_k)^\perp.$$

Therefore, by definition of the update:

$$\tilde{\zeta}_h^k = \Pi_R(\hat{\zeta}_h^k),$$

we conclude that:

$$\tilde{\zeta}_h^k - \zeta_*^k \in \ker(A_k)^\perp.$$

This completes the proof.

□

To derive the second-order error bound, we first note the following relations.

$$\begin{aligned}
&\|\tilde{\zeta}_{h+1}^k - \tilde{\zeta}_*^k\|^2 \\
&= \|\Pi_R(\tilde{\zeta}_h^k - \beta g(x_h^k; \tilde{\zeta}_h^k)) - \Pi_R(\tilde{\zeta}_*^k)\|^2 \\
&\le \|\tilde{\zeta}_h^k - \beta g(x_h^k; \tilde{\zeta}_h^k) - \tilde{\zeta}_*^k\|^2 \\
&= \|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, g(x_h^k; \tilde{\zeta}_h^k)\rangle + \beta^2\|g(x_h^k; \tilde{\zeta}_h^k)\|^2 \\
&= \|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, \hat{g}(x_h^k; \tilde{\zeta}_h^k)\rangle - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, g(x_h^k; \tilde{\zeta}_h^k) - \hat{g}(x_h^k; \tilde{\zeta}_h^k)\rangle + \beta^2\|g(x_h^k; \tilde{\zeta}_h^k)\|^2 \\
&\le \|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, \hat{g}(x_h^k; \tilde{\zeta}_h^k)\rangle - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, g(x_h^k; \tilde{\zeta}_h^k) - \hat{g}(x_h^k; \tilde{\zeta}_h^k)\rangle + \beta^2 C_1^2 \log(H/\delta) \\
&\overset{(a)}{\le} \|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, A_k(\tilde{\zeta}_h^k - \tilde{\zeta}_*^k)\rangle - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, \hat{g}(x_h^k; \tilde{\zeta}_h^k) - A_k(\tilde{\zeta}_h^k - \tilde{\zeta}_*^k)\rangle \\
&\quad + C_3 m^{-1/2}\log(H/\delta) + \beta^2 C_1^2 \log(H/\delta) \\
&\overset{(b)}{\le} \|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 - 2\beta\lambda_0(1-\gamma)\|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, \hat{g}(x_h^k; \tilde{\zeta}_h^k) - A_k(\tilde{\zeta}_h^k - \tilde{\zeta}_*^k)\rangle \\
&\quad + 2\beta C_3 m^{-1/2}\log(H/\delta) + \beta^2 C_1^2 \log(H/\delta)
\end{aligned}$$

where (a) follows from Lemma 6(a), (b) follows from the fact that $A_k \succcurlyeq \lambda_0(1-\gamma)I$ and Lemma 6 (a). Taking conditional expectation $\mathbb{E}_h$ on both sides, we obtain

$$\begin{aligned}
\mathbb{E}_h\left[\left\|\tilde{\zeta}_{h+1}^k - \tilde{\zeta}_*^k\right\|^2\right] &\le (1 - 2\beta\lambda_0(1-\gamma))\|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 - 2\beta\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, \mathbb{E}_h\left[\hat{g}(x_h^k; \tilde{\zeta}_h^k) - A_k(\tilde{\zeta}_h^k - \tilde{\zeta}_*^k)\right]\rangle \\
&\quad + 2\beta C_3 m^{-1/2}\log(H/\delta) + \beta^2 C_1^2 \log(H/\delta)
\end{aligned} \tag{46}$$

The second term in (46) can be bounded as

$$\begin{aligned}
&-\langle \tilde{\zeta}_h^k - \tilde{\zeta}_*^k, \mathbb{E}_h\left[\hat{g}(x_h^k; \tilde{\zeta}_h^k) - A_k(\tilde{\zeta}_h^k - \tilde{\zeta}_*^k)\right]\rangle \\
&\le \frac{\lambda_0(1-\gamma)}{4}\|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 + \frac{1}{\lambda_0(1-\gamma)}\left\|\mathbb{E}_h[\hat{g}(x_h^k; \tilde{\zeta}_h^k) - A_k(\tilde{\zeta}_h^k - \tilde{\zeta}_*^k)]\right\|^2 \\
&\le \frac{\lambda_0(1-\gamma)}{4}\|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 + \frac{1}{\lambda_0(1-\gamma)}\left\|\left\{\mathbb{E}_h[A(x_h^k)] - A_k\right\}\tilde{\zeta}_h^k + \left\{b_k - \mathbb{E}_h\left[b(z_h^k)\right]\right\}\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\lambda_0(1-\gamma)}{4}\|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 + \frac{2\delta_A^2\|\tilde{\zeta}_h^k\|^2 + 2\delta_b^2}{\lambda_0(1-\gamma)} \\
&\leq \frac{\lambda_0(1-\gamma)}{4}\|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 + \frac{4\delta_A^2\|\tilde{\zeta}_h - \tilde{\zeta}_*^k\|^2 + 4\delta_A^2\lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2 + 2\delta_b^2}{\lambda_0(1-\gamma)}
\end{aligned} \tag{47}
$$

where the last inequality follows from $\left\|\tilde{\zeta}_*^k\right\|^2 = \left\|A^{-1}b\right\|^2 \leq \lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2$. Substituting the above bounds in (46),

$$
\begin{aligned}
&\mathbb{E}_h\left[\|\tilde{\zeta}_{h+1}^k - \tilde{\zeta}_*^k\|^2\right] \\
&\leq \left(1 - \frac{3\beta\lambda_0(1-\gamma)}{2} + \frac{8\beta\delta_A^2}{\lambda_0(1-\gamma)}\right)\|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 + \frac{4\beta}{\lambda_0(1-\gamma)}\left[2\delta_A^2\lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2 + \delta_b^2\right] \\
&\quad + 2\beta C_3 m^{-1/2}\log(H/\delta) + \beta^2 C_1^2\log(H/\delta)
\end{aligned}
$$

For $\delta_A \leq \lambda_0(1-\gamma)/4$, we can modify the above inequality to the following.

$$
\begin{aligned}
&\mathbb{E}_h[\|\tilde{\zeta}_{h+1}^k - \tilde{\zeta}_*^k\|^2] \\
&\leq (1-\beta\lambda_0(1-\gamma))\|\tilde{\zeta}_h^k - \tilde{\zeta}_*^k\|^2 + \frac{4\beta}{\lambda_0(1-\gamma)}\left[2\delta_A^2\lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2 + \delta_b^2\right] + 2\beta C_3 m^{-1/2}\log(H/\delta) \\
&+ \beta^2 C_1^2\log(H/\delta)
\end{aligned}
$$

Taking expectation on both sides and unrolling the recursion yields

$$
\begin{aligned}
&\mathbb{E}[\|\tilde{\zeta}_H^k - \tilde{\zeta}_*^k\|^2] \\
&\leq (1-\beta\lambda_0(1-\gamma))^H\,\mathbb{E}\|\tilde{\zeta}_0 - \tilde{\zeta}_*^k\|^2 \\
&+ \sum_{h=0}^{H-1}(1-\beta\lambda_0(1-\gamma))^h\left\{\frac{4\beta}{\lambda_0(1-\gamma)}\left[2\delta_A^2\lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2 + \delta_b^2\right] + 2\beta C_3 m^{-1/2}\log(H/\delta)\right. \\
&\left. + \beta^2 C_1^2\log(H/\delta)\right\} \\
&\leq \exp\left(-H\beta\lambda_0(1-\gamma)\right)\mathbb{E}\|\tilde{\zeta}_0 - \tilde{\zeta}_*^k\|^2 \\
&+ \frac{1}{\beta\lambda_0(1-\gamma)}\left\{\frac{4\beta}{\lambda_0(1-\gamma)}\left[2\delta_A^2\lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2 + \delta_b^2\right] + 2\beta C_3 m^{-1/2}\log(H/\delta) + \beta^2 C_1^2\log(H/\delta)\right\} \\
&= \exp\left(-H\beta\lambda_0(1-\gamma)\right)\mathbb{E}\|\tilde{\zeta}_0 - \tilde{\zeta}_*^k\|^2 \\
&+ \left\{4\lambda_0^{-2}(1-\gamma)^{-2}\left[2\delta_A^2\lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2 + \delta_b^2\right] + 2\lambda_0(1-\gamma)^{-1}C_3 m^{-1/2}\log(H/\delta)\right. \\
&\left. + \beta\lambda_0(1-\gamma)^{-1}C_1^2\log(H/\delta)\right\}
\end{aligned}
$$

Substituting $\beta = \frac{2\log H}{\lambda_0(1-\gamma)H}$ and using Lemma 7 yields

$$
\mathbb{E}[\|\zeta_H^k - \zeta_*^k\|^2] \leq \mathcal{O}\left(\frac{\mathbb{E}\|\tilde{\zeta}_0 - \tilde{\zeta}_*^k\|^2}{H^2} + \frac{\log^2(H/\delta)}{\lambda_0^2(1-\gamma)^2 H} + \frac{\log(H/\delta)}{\lambda_0(1-\gamma)m^{1/2}} + \frac{1}{(1-\gamma)^4\lambda_0^4 T^\kappa}\right)
$$

A.2 PROOF OF LEMMA 5

Consider the critic update

$$
\zeta_{h+1}^k = \Pi_R(\zeta_h^k - \beta g(x_h^k;\zeta_h^k)) \tag{48}
$$

This can be rewritten as

$$
\zeta_{h+1}^k = \zeta_h^k - \beta g(x_h^k;\zeta_h^k) + \epsilon_h^k \tag{49}
$$

where $\epsilon_h^k = \Pi_R(\zeta_h^k - \beta g(\zeta_h^k)) - (\zeta_h^k - \beta g(z_h^k;\zeta_h^k))$. Note that if $\zeta_h^k - \beta g(x_h^k;\zeta_h^k) \in S_R$ then $\epsilon_h^k = 0$ and if $\zeta_h^k - \beta g(x_h^k;\zeta_h^k) \notin S_R$, we have the following

$$\|\epsilon_h^k\| = \|\Pi_R(\zeta_h^k - \beta g(x_h^k;\zeta_h^k)) - (\zeta_h^k - \beta g(x_h^k;\zeta_h^k))\| \overset{(a)}{\leq} \|\zeta_h^k - (\zeta_h^k - \beta g(x_h^k;\zeta_h^k))\| \leq \beta\|g(x_h^k;\zeta_h^k)\|, \tag{50}$$

where (a) follows since $\Pi_R(\zeta_h^k - \beta g(x_h^k;\zeta_h^k))$ is the closest point to $\zeta_h^k - \beta g(x_h^k;\zeta_h^k)$ in the set S_R and $\zeta_h^k \in S_R$. This yields,

$$\|\epsilon_h^k\| \leq \beta\|g(x_h^k;\zeta_h^k)\|\mathbf{1}_{\{\zeta_h^k - \beta g(x_h^k;\zeta_h^k)\notin S_R\}}, \tag{51}$$

where $\mathbf{1}_A$ denotes the indicator function for event A. Taking expectation on both sides gives us

$$\mathbb{E}\,\|\epsilon_h^k\| \leq \beta C_1\sqrt{\log(H/\delta)}\,\Pr(\zeta_h^k - \beta g(x_h^k;\zeta_h^k) \notin S_R) \overset{(a)}{\leq} \beta C_1\sqrt{\log(H/\delta)}\,\Pr\left(\zeta_h^k \notin S_{R-\beta C_1\sqrt{\log(H/\delta)}}\right), \tag{52}$$

where (a) follows from the fact that the event $\{\zeta_h^k - \beta g(x_h^k;\zeta_h^k) \notin S_R\}$ is contained in the event $\{\zeta_h^k \notin S_{R-\beta C_1\sqrt{\log(H/\delta)}}\}$. To see this, observe that if $\zeta_h^k - \beta g(x_h^k;\zeta_h^k) \notin S_R$, then

$$\|\zeta_h^k - \zeta_0\| \geq \|\zeta_h^k - \zeta_0 - \beta g(x_h^k;\zeta_h^k)\| - \|\beta g(x_h^k;\zeta_h^k)\| \geq R - \beta C_1\sqrt{\log(H/\delta)} \tag{53}$$

We now bound $\Pr\left(\zeta_h^k \in S_{R-\beta C_1\sqrt{\log(H/\delta)}}\right)$ using Markov's inequality combined with the bound on $\mathbb{E}\,\|\zeta_h^k - \zeta_*^k\|^2$ obtained earlier.

$$\begin{aligned}\Pr\left(\|\zeta_h^k - \zeta_0\| \geq R - \beta C_1\sqrt{\log(H/\delta)}\right) &\leq \Pr\left(\|\zeta_h^k - \zeta_*^k\| + \|\zeta_0 - \zeta_*^k\| \geq R - \beta C_1\sqrt{\log(H/\delta)}\right)\\ &\leq \Pr(\|\zeta_h^k - \zeta_*^k\| \geq \bar{R})\end{aligned} \tag{54}$$

where $\bar{R} := (R/2) - \beta C_1\sqrt{\log(H/\delta)}$ and R is chosen such that $\zeta_*^k \in S_{R/2}$. Since $\Pr(\|\zeta_h^k - \zeta_*^k\| \geq \bar{R}) \leq \frac{1}{\bar{R}^2}\cdot\mathbb{E}\,\|\zeta_h^k - \zeta_*^k\|^2$, it follows that

$$\|\,\mathbb{E}[\epsilon_h^k]\| \leq \mathbb{E}\,\|\epsilon_h^k\| \leq \mathcal{O}\left(\frac{\beta\,\mathbb{E}\,\|\zeta_0 - \zeta_*^k\|^2}{H^2} + \frac{\beta\log^2(H/\delta)}{\lambda_0^2(1-\gamma)^2 H} + \frac{\beta\log(H/\delta)}{\lambda_0(1-\gamma)m^{1/2}} + \frac{\beta}{(1-\gamma)^4\lambda_0^4 T^\kappa}\right) \tag{55}$$

The term ϵ_h^k arising from the projection operator can now be viewed as a small error term. Taking the expectation given the policy parameter θ and the square norm in (49), we obtain

$$\begin{aligned}&\|\,\mathbb{E}[\zeta_{h+1}^k] - \zeta_*^k\|^2\\ &= \|\,\mathbb{E}[\zeta_h^k] - \zeta_*^k - \beta\,\mathbb{E}[g(x_h^k;\zeta_h^k)]\|^2 + \|\,\mathbb{E}[\epsilon_h^k]\|^2 + 2\langle\mathbb{E}[\zeta_h^k] - \zeta_*^k - \beta\,\mathbb{E}[g(x_h^k;\zeta_h^k)], \mathbb{E}[\epsilon_h^k]\rangle\end{aligned} \tag{56}$$

Note that

$$\begin{aligned}&2\langle\mathbb{E}[\zeta_h^k] - \zeta_*^k - \beta\,\mathbb{E}[g(\zeta_h^k)], \mathbb{E}[\epsilon_h^k]\rangle\\ &\qquad\leq \frac{\beta\lambda_0(1-\gamma)}{2}\|\,\mathbb{E}[\zeta_h^k] - \zeta_*^k - \beta\,\mathbb{E}[g(x_h^k;\zeta_h^k)]\|^2 + \frac{2}{\beta\lambda_0(1-\gamma)}\|\,\mathbb{E}[\epsilon_h^k]\|^2\end{aligned} \tag{57}$$

Thus, combining the above inequalities

$$\begin{aligned}&\|\,\mathbb{E}[\zeta_{h+1}^k] - \zeta_*^k\|^2\\ &\leq \left(1 + \frac{\beta\lambda_0(1-\gamma)}{2}\right)\|\,\mathbb{E}[\zeta_h^k] - \zeta_*^k - \beta\,\mathbb{E}[g(x_h^k;\zeta_h^k)]\|^2 + \left(1 + \frac{2}{\beta\lambda_0(1-\gamma)}\right)\|\,\mathbb{E}[\epsilon_h^k]\|^2\end{aligned} \tag{58}$$

We now focus on bounding $\|\mathbb{E}[\zeta_h^k] - \zeta_*^k - \beta\,\mathbb{E}[g(x_h^k;\zeta_h^k)]\|^2$. Observe the following

$$
\begin{aligned}
&\|\mathbb{E}[\zeta_h^k] - \zeta_*^k - \beta\,\mathbb{E}[g(x_h^k;\zeta_h^k)]\|^2 \\
&= \|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}[g(x_h^k;\zeta_h^k)]\rangle + \beta^2\|\mathbb{E}[g(x_h^k;\zeta_h^k)]\|^2 \\
&\le \|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}[A(x_h^k)]\rangle - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}[g(x_h^k;\zeta_h^k)] - \mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\rangle \\
&\quad + \beta^2\|\mathbb{E}[g(x_h^k;\zeta_h^k)] - \mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\|^2 + \beta^2\|\mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\|^2 \\
&\le \|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\rangle - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}[g(x_h^k;\zeta_h^k)] - \mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\rangle \\
&\quad + 2\beta^2\|\mathbb{E}[g(x_h^k;\zeta_h^k)] - \mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\|^2 + 2\beta^2\|\mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\|^2 \\
&\overset{(a)}{\le} \|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\rangle + 2\beta C_3\sqrt{\log(H/\delta)}m^{-1/2} \\
&\quad + 2\beta^2 C_2^2\log(H/\delta)m^{-1} + 2\beta^2\|\mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)]\|^2 \\
&\le \|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, A_k(\mathbb{E}[\zeta_h^k] - \zeta_*^k)\rangle - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)] - A_k(\mathbb{E}[\zeta_h^k] - \zeta_*^k)\rangle \\
&\quad + 2\beta C_3\sqrt{\log(H/\delta)}m^{-1/2} + 2\beta^2 C_2^2\log(H/\delta)m^{-1} + 2\beta^2\|A_k(\mathbb{E}[\zeta_h^k] - \zeta_*^k)\|^2 \\
&\quad + 2\beta^2\|\mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)] - A_k(\mathbb{E}[\zeta_h^k] - \zeta_*^k)\|^2 \\
&\overset{(b)}{\le} (1 - 2\beta\lambda_0(1-\gamma) + 2\Lambda_A^2\beta^2)\|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 - 2\beta\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)] - A_k(\mathbb{E}[\zeta_h^k] - \zeta_*^k)\rangle \\
&\quad + 2\beta C_3\sqrt{\log(H/\delta)}m^{-1/2} + 2\beta^2 C_2^2\log(H/\delta)m^{-1} + 2\beta^2\|\mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)] - A_k(\mathbb{E}[\zeta_h^k] - \zeta_*^k)\|^2
\end{aligned}
\tag{59}
$$

where (a) follows from Lemma 6, while (b) follows from the fact that $\|A_k\| \le \Lambda_A$ and $A_k \succcurlyeq \lambda_0(1-\gamma)I$. The last term in the last line of (59) can be bounded as follows.

$$
\begin{aligned}
&\|\mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)] - (A_k\,\mathbb{E}[\zeta_h^k] - b_k)\|^2 \\
&= \left\|\mathbb{E}\left[(\mathbb{E}[A(x_h^k)] - A_k)(\zeta_h^k - \zeta_*^k)\right] + (\mathbb{E}[A(x_h^k)] - A_k)\zeta_*^k + (b_k - \mathbb{E}[b(x_h^k)])\right\|^2 \\
&\le 3\,\mathbb{E}\left[\|\mathbb{E}[A(x_h^k)] - A_k\|^2\|\zeta_h^k - \zeta_*^k\|^2\right] + 3\,\mathbb{E}\left[\|\mathbb{E}[A(x_h^k)] - A_k)\|^2\right]\|\zeta_*^k\|^2 + 3\left\|b_k - \mathbb{E}[b(x_h^k)]\right\|^2 \\
&\le 3\delta_A^2\,\mathbb{E}\left[\|\zeta_h^k - \zeta_*^k\|^2\right] + 3\lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2\delta_A^2 + 3\bar{\delta}_b^2
\end{aligned}
$$

The second term in the last line of (59) can be bounded as follows.

$$
\begin{aligned}
&-\langle \mathbb{E}[\zeta_h^k] - \zeta_*^k, \mathbb{E}_h\left[\mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)] - A_k(\mathbb{E}[\zeta_h^k] - \zeta_*^k)\right]\rangle \\
&\qquad \le \frac{\lambda_0(1-\gamma)}{4}\|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 + \frac{1}{\lambda_0(1-\gamma)}\left\|\mathbb{E}[\hat{g}(x_h^k;\zeta_h^k)] - A_k(\mathbb{E}[\zeta_h^k] - \zeta_*^k)\right\|^2 \\
&\qquad \le \frac{\lambda_0(1-\gamma)}{4}\|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 + \frac{3}{\lambda_0(1-\gamma)}\left[\delta_A^2\,\mathbb{E}\,\|\zeta_h^k - \zeta_*^k\|^2 + \lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2\delta_A^2 + \bar{\delta}_b^2\right]
\end{aligned}
$$

Substituting the above bounds in (59), we obtain the following bound

$$
\begin{aligned}
&\|\mathbb{E}[\zeta_h^k] - \zeta_*^k - \beta\,\mathbb{E}[g(x_h^k;\zeta_h^k)]\|^2 \\
&\le \left(1 - \frac{3\beta\lambda_0(1-\gamma)}{2} + 2\Lambda_A^2\beta^2\right)\|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 + 2\beta C_3\sqrt{\log(H/\delta)}m^{-1/2} \\
&+ 2\beta^2 C_2^2\log(H/\delta)m^{-1} + 6\beta\left(\beta + \frac{1}{\lambda_0(1-\gamma)}\right)\left[\delta_A^2\,\mathbb{E}\,\|\zeta_h^k - \zeta_*^k\|^2 + \lambda_0^{-2}(1-\gamma)^{-2}\Lambda_b^2\delta_A^2 + \bar{\delta}_b^2\right]
\end{aligned}
$$

Combining (58) with the above bound yields the following result.

$$
\begin{aligned}
&\|\mathbb{E}[\zeta_{h+1}^k] - \zeta_*^k\|^2 \\
&\le \left(1 - \frac{3\beta\lambda_0(1-\gamma)}{2} + 2\Lambda_A^2\beta^2\right)\left(1 + \frac{\beta\lambda_0(1-\gamma)}{2}\right)\|\mathbb{E}[\zeta_h^k] - \zeta_*^k\|^2 \\
&+ \left(2\beta + \beta^2\lambda_0(1-\gamma)\right)C_3\sqrt{\log(H/\delta)}m^{-1/2} \\
&\quad + 6\left(1 + \frac{\beta\lambda_0(1-\gamma)}{2}\right)\left(\beta^2 + \frac{\beta}{\lambda_0(1-\gamma)}\right)\Big[\delta_A^2\,\mathbb{E}\,\|\zeta_h^k - \zeta_*^k\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \lambda_0^{-2}(1-\gamma)^{-2}\left(1+\frac{\beta\lambda_0(1-\gamma)}{2}\right)\Lambda_b^2\delta_A^2+\bar{\delta}_b^2\Big] \\
&\quad + \left(1+\frac{2}{\beta\lambda_0(1-\gamma)}\right)\|\,\mathbb{E}[\epsilon_h^k]\|^2 + 2\left(1+\frac{\beta\lambda_0(1-\gamma)}{2}\right)\beta^2 C_2^2\log(H/\delta)m^{-1} \\
&\leq \left(1-\beta\lambda_0(1-\gamma)+\Lambda_A^2\lambda_0(1-\gamma)\beta^3\right)\|\,\mathbb{E}[\zeta_h^k]-\zeta_*^k\|^2 + 2\left(1+\frac{\beta\lambda_0(1-\gamma)}{2}\right)\beta C_3\sqrt{\log(H/\delta)}m^{-1/2} \\
&\quad + 6\left(1+\frac{\beta\lambda_0(1-\gamma)}{2}\right)\beta\left(\beta+\frac{1}{\lambda_0(1-\gamma)}\right)\left[\delta_A^2R^2+\lambda_0^{-2}(1-\gamma)^{-2}\left(1+\frac{\beta\lambda_0(1-\gamma)}{2}\right)\Lambda_b^2\delta_A^2+\bar{\delta}_b^2\right] \\
&\quad + \left(1+\frac{2}{\beta\lambda_0(1-\gamma)}\right)\|\,\mathbb{E}[\epsilon_h^k]\|^2 + 2\left(1+\frac{\beta\lambda_0(1-\gamma)}{2}\right)\beta^2 C_2^2\log(H/\delta)m^{-1} \\
&:= \left(1-\beta\lambda_0(1-\gamma)+\Lambda_A^2\lambda_0(1-\gamma)\beta^3\right)\|\,\mathbb{E}[\zeta_h^k]-\zeta_*^k\|^2 + \Delta_h^k
\end{aligned}
$$

If $\beta < 1/(2\Lambda_A)$, the above bound implies the following.

$$
\|\,\mathbb{E}[\zeta_{h+1}^k]-\zeta_*^k\|^2 \leq \left(1-\frac{\beta\lambda_0(1-\gamma)}{2}\right)\|\,\mathbb{E}[\zeta_h^k]-\zeta_*^k\|^2 + \Delta_h^k
$$

Unrolling the recursion, we obtain the following result.

$$
\begin{aligned}
&\|\,\mathbb{E}[\zeta_{h+1}^k]-\zeta_*^k\|^2 \\
&\leq \left(1-\frac{\beta\lambda_0(1-\gamma)}{2}\right)^H\|\,\mathbb{E}[\zeta_0]-\zeta_*^k\|^2 + \sum_{i=0}^{H}\left(1-\frac{\beta\lambda_0(1-\gamma)}{2}\right)^{H-h}\Delta_h^k \\
&\leq \left(1-\frac{\beta\lambda_0(1-\gamma)}{2}\right)^H\|\,\mathbb{E}[\zeta_0]-\zeta_*^k\|^2 + \frac{2}{\beta\lambda_0(1-\gamma)}\Delta_h^k \\
&\leq \exp\left(\frac{\beta\lambda_0(1-\gamma)H}{2}\right)\|\,\mathbb{E}[\zeta_0]-\zeta_*^k\|^2 + \frac{2}{\beta\lambda_0(1-\gamma)}\Delta_h^k \\
&\leq \exp\left(\frac{\beta\lambda_0(1-\gamma)H}{2}\right)\|\,\mathbb{E}[\zeta_0]-\zeta_*^k\|^2 + \mathcal{O}\left(\frac{2}{\beta\lambda_0(1-\gamma)}\Delta_h^k\right)
\end{aligned}
$$

Substituting $\beta = \frac{2\log H}{\lambda_0(1-\gamma)H}$, it follows that

$$
\|\,\mathbb{E}[\zeta_{h+1}^k]-\zeta_*^k\|^2 \leq \mathcal{O}\left(\frac{\|\,\mathbb{E}[\zeta_0]-\zeta_*^k\|^2}{\lambda_0^2(1-\gamma)^2H^2} + \frac{\log^4(H/\delta)}{\lambda_0^6(1-\gamma)^6H^2} + \frac{\sqrt{\log(H/\delta)}}{\lambda_0(1-\gamma)m^{1/2}} + \frac{1}{(1-\gamma)^{10}\lambda_0^{10}T^{2\kappa}}\right) \tag{60}
$$

B PROOF OF LEMMAS 2 AND 3

Recall that the block-indexed NAC-DD algorithm (Algorithm 1) uses the NPG estimator evaluated once every M transitions. The NPG updates can be written as follows

$$
\omega_{H+h+1}^k = \omega_{H+h}^k - \eta\big(X(x_h^k)\,\omega_{H+h}^k - y(x_h^k)\big), \tag{61}
$$

where

$$
X(x_h^k) := \nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)\nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)\big)^\top, \qquad y(x_h^k) := Q\big(\phi(\bar{a}_h^k|\bar{s}_h^k);\zeta_H^k\big)\,\nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k), \tag{62}
$$

with $x_h^k = \phi(\bar{s}_h^k, \bar{a}_h^k)$. Throughout, the conditional expectation $\mathbb{E}_{k,h}[\cdot]$ is over all the randomness from the hth block in epoch k given the entire history prior to this block. Whereas, $\mathbb{E}_k[\cdot]$ denotes the expectation given θ_k. For notational convenience, we henceforth denote $\mathbb{E}[\cdot]$, $\mathbb{E}_{k,h}[\cdot]$ and $\mathbb{E}_k[\cdot]$ in place of $\mathbb{E}[\cdot|\zeta_0]$, $\mathbb{E}_{k,h}[\cdot|\zeta_0]$ and $\mathbb{E}_k[\cdot|\zeta_0]$, respectively. Recall that $X(x_h^k)$ serves as an estimate of $F(\theta_k)$ and $X(x_h^k)$ serves as an estimate of $\nabla_\theta J(\theta_k)$.

The Fisher information matrix satisfies $\mu I \preccurlyeq F(\theta_k)$ and $\|F(\theta_k)\| \leq G_1^2$ from Assumptions 7 and 6, respectively. Furthermore, the norm of the policy gradient norm is bounded by $\|\nabla_\theta J(\theta_k)\| \leq G_1(1-\gamma)^2$ [Liu et al., 2020b].

In this Section, we establish that, for each block h, there exists positive constants σ_X^2, δ_X^2, σ_y^2, δ_y^2, $\bar{\delta}_y^2$, Λ_X, Λ_y such that the following bounds hold:

$$\mathbb{E}_{k,h}\left\|X(x_h^k) - F(\theta_k)\right\|^2 \le \sigma_X^2, \quad \left\|\mathbb{E}_{k,h}[X(x_h^k)] - F(\theta_k)\right\|^2 \le \delta_X^2, \quad \|X(x_h^k)\| \le \Lambda_X, \quad \|y(x_h^k)\| \le \Lambda_y$$

$$\mathbb{E}_{k,h}\left\|y(x_h^k) - \nabla_\theta J(\theta_k)\right\|^2 \le \sigma_y^2, \quad \left\|\mathbb{E}_{k,h}[y(x_h^k)] - \nabla_\theta J(\theta_k)\right\|^2 \le \delta_y^2, \quad \left\|\mathbb{E}_k[y(x_h^k)] - \nabla_\theta J(\theta_k)\right\|^2 \le \bar{\delta}_y^2.$$

Using these bounds, Theorem 2 of Ganesh et al. [2025] can then be applied with

$$P = F(\theta_k), \quad q = \nabla_\theta J(\theta_k), \quad \hat{P}_h = X(x_h^k), \quad \hat{q}_h = y(x_h^k),$$

and step size $\eta = \frac{2\log H}{\mu H}$, yielding the desired mean-square and bias guarantees for the iterates ω_H^k.

From Assumption 7, eigenvalues of $F(\theta_k)$ are bounded below by μ. Whereas

Lemma 10. *Under the assumptions of Theorem 1, for every epoch k and block h:*

1. $\|X(x_h^k)\| \le G_1^2$.
2. $\|\mathbb{E}_{k,h}[X(x_h^k)] - F(\theta_k)\|^2 \le G_1^4 T^{-2\kappa}$.

Proof. Part (1) follows directly from Assumption 6. Part (2) follows by bounding the bias bound as in Lemma 7. □

Since $\|X(x_h^k)\|, \|F(\theta_k)\| \le G_1^2$, it follows that $\mathbb{E}_{k,h}\,\|X(x_h^k) - F(\theta_k)\|^2 \le 2G_1^4$. Separately, using Lemma 6, we obtain $\|y(x_h^k)\| \le C_1' G_1 \sqrt{\log(H/\delta)}$. Next, we bound the bias and second-order error of $y(x_h^k)$, which also carries the critic approximation error.

Lemma 11. *Fix epoch k. Under the assumptions of Theorem 1, for each block h:*

$$\text{(a)} \quad \left\|\mathbb{E}_{k,h}[\,y(x_h^k)\,] - \nabla_\theta J(\theta_k)\right\|^2 \;\le\; \tilde{\mathcal{O}}\Big(\tfrac{\sigma_y^2}{T^\kappa} + \delta_y^2\Big),$$

$$\text{(b)} \quad \mathbb{E}_{k,h}\big[\|y(x_h^k) - \nabla_\theta J(\theta_k)\|^2\big] \;\le\; \tilde{\mathcal{O}}\Big(\sigma_y^2 + G_1^2 (C_1')^2 \sqrt{\log(H/\delta)}\Big),$$

where

$$\sigma_y^2 = \tilde{\mathcal{O}}\Big(\tfrac{G_1^2}{(1-\gamma)^4}\Big), \quad \delta_y^2 = \tilde{\mathcal{O}}\Big(G_1^2\|\zeta_H^k - \zeta_*^k\|^2 + m^{-1/2} + \tfrac{G_1^2\epsilon_{\text{app}}}{(1-\gamma)^2}\Big).$$

Moreover,

$$\text{(c)} \quad \left\|\mathbb{E}_k[\,y(x_h^k)\,] - \nabla_\theta J(\theta_k)\right\|^2 \;\le\; \tilde{\mathcal{O}}\Big(\tfrac{\sigma_y^2}{T^\kappa} + \bar{\delta}_y^2\Big),$$

with $\bar{\delta}_y^2 = \tilde{\mathcal{O}}\Big(G_1^2\|\mathbb{E}_k[\zeta_H^k] - \zeta_*^k\|^2 + m^{-1/2} + G_1^2\epsilon_{\text{app}}/(1-\gamma)^2\Big)$.

Proof. We expand

$$\begin{aligned}
&\mathbb{E}_{k,h}[y(x_h^k)] - \nabla_\theta J(\theta_k) \\
&= \mathbb{E}_{k,h}[Q(x_h^k;\zeta_H^k)\,\nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)] - \nabla_\theta J(\theta_k) \\
&= \mathbb{E}_{k,h}[(Q(x_h^k;\zeta_H^k) - Q(\bar{s}_h^k,\bar{a}_h^k))\nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)] + \mathbb{E}_{k,h}[Q(\bar{s}_h^k,\bar{a}_h^k)\nabla_\theta \log \pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)] - \nabla_\theta J(\theta_k),
\end{aligned}$$

and decompose

$$Q(x_h^k;\zeta_H^k) - Q(\bar{s}_h^k,\bar{a}_h^k) \;= T_0 + T_1 + T_2 + T_3,$$

where

$$\begin{aligned}
T_0 &= Q(x_h^k;\zeta_*^k) - Q(\bar{s}_h^k,\bar{a}_h^k) \quad & T_1 &= Q(x_h^k;\zeta_H^k) - \hat{Q}(x_h^k;\zeta_H^k), \\
T_2 &= \hat{Q}(x_h^k;\zeta_*^k) - Q(x_h^k;\zeta_*^k), \quad & T_3 &= \hat{Q}(x_h^k;\zeta_H^k) - \hat{Q}(x_h^k;\zeta_*^k).
\end{aligned}$$

We have $\|T_1\|, \|T_2\| = \mathcal{O}(m^{-1/2})$, which follows from the bounds on the linearization error, while $\|T_3\| = \mathcal{O}(C_1\|\zeta_H^k - \zeta_*^k\|)$.

To bound $\|T_0\|$, first note that

$$(Q(x_h^k;\zeta_*^k) - Q(\bar{s}_h^k,\bar{a}_h^k))^2 \leq 2(Q(x_h^k;\zeta_*^k) - \widehat{Q}(x_h^k;\zeta_*^k))^2 + 2(\widehat{Q}(x_h^k;\zeta_*^k) - Q(\bar{s}_h^k,\bar{a}_h^k))^2 \tag{63}$$

We have $(Q(x_h^k;\zeta_*^k) - \widehat{Q}(x_h^k;\zeta_*^k))^2 \leq \mathcal{O}(m^{-1})$. Furthermore, from Appendix A.3 in Ke et al. [2024], we have

$$\mathbb{E}\,\|\widehat{Q}(x_h^k;\zeta_*^k) - Q(\bar{s}_h^k,\bar{a}_h^k)\|^2 \leq \frac{1}{(1-\gamma)^2}\,\mathbb{E}\,\|\Pi_{\mathcal{F}_{R,m}} Q(\bar{s}_h^k,\bar{a}_h^k) - Q(\bar{s}_h^k,\bar{a}_h^k)\|^2 \leq \frac{\epsilon_{\text{app}}}{(1-\gamma)^2} \tag{64}$$

Using arguments as in Lemma 7, we obtain $\|\mathbb{E}_{k,h}[Q(\bar{s}_h^k,\bar{a}_h^k)\nabla\log\pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)] - \nabla_\theta J(\theta_k)\| \leq G_1((1-\gamma)T^\kappa)^{-1}$, which yields part (a). Part (b) follows easily using the bounds on $\|\nabla_\theta J(\theta_k)\|$ and $\|y(x_h^k)\|$. For part (c), note that

$$\begin{aligned}
&\mathbb{E}_k[y(x_h^k)] - \nabla_\theta J(\theta_k) \\
&= \mathbb{E}_k[(T_0+T_1+T_2+T_3)\nabla_\theta\log\pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)] + \mathbb{E}_k[Q(\bar{s}_h^k,\bar{a}_h^k)\nabla_\theta\log\pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)] - \nabla_\theta J(\theta_k)
\end{aligned}$$

and $\|\,\mathbb{E}_k[T_3\nabla_\theta\log\pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)]\| = \|\,\mathbb{E}_{k,h}[\mathbb{E}_k[\hat{Q}(x_h^k;\zeta_H^k) - \hat{Q}(x_h^k;\zeta_*^k)]\nabla_\theta\log\pi_{\theta_k}(\bar{a}_h^k|\bar{s}_h^k)]\| \leq C_1G_1\|\,\mathbb{E}_k[\zeta_H^k] - \zeta_*^k\|$. The bounds for the remaining terms follow from the bounds in part (a). □

We now can invoke Theorem 2 in Ganesh et al. [2025] to obtain

$$\mathbb{E}_k\,\|\omega_k - \omega_k^*\|^2 \leq \mathcal{O}\left(\frac{G_1^4}{H\mu^4(1-\gamma)^4} + \frac{G_1^2(C_1')^2\log(H/\delta)}{H\mu^2(1-\gamma)^4} + \mu^{-2}G_1^2\,\mathbb{E}\,\|\zeta_H^k - \zeta_*^k\|^2 + \mu^{-2}m^{-1/2} + \frac{G_1^2\epsilon_{\text{app}}}{\mu^2(1-\gamma)^2}\right) \tag{65}$$

and

$$\|\,\mathbb{E}_k[\omega_k] - \omega_k^*\|^2 \leq \mathcal{O}\left(\frac{G_1^2(C_1')^2G_1^2\log(H/\delta)}{T^\kappa} + \|\,\mathbb{E}[\zeta_H^k] - \zeta_*^k\|^2 + \frac{G_1^2\epsilon_{\text{app}}}{\mu^2(1-\gamma)^2}\right) \tag{66}$$

C PROOF OF THEOREM 1

Recall that the global convergence of any update of form $\theta_{k+1} = \theta_k + \alpha\omega_k$ can be bounded as

$$\begin{aligned}
J^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[J(\theta_k)] \leq \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} + \frac{G_1}{K}\sum_{k=0}^{K-1}\mathbb{E}\,\|(\mathbb{E}\,[\omega_k|\theta_k] - \omega_k^*)\| + \frac{\alpha G_2}{K}\sum_{k=0}^{K-1}\mathbb{E}\,\|\omega_k - \omega_k^*\|^2 \\
+ \frac{\alpha\mu^{-2}}{K}\sum_{k=0}^{K-1}\mathbb{E}\,\|\nabla_\theta J(\theta_k)\|^2 + \frac{1}{\alpha K}\,\mathbb{E}_{s\sim d^{\pi^*}}[\text{KL}(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))].
\end{aligned} \tag{67}$$

We note that our algorithm updates θ at each iteration k using ω_H and ζ_H obtained after H iterations of the NPG and critic estimation inner loops. Therefore, we use ω_H and ζ_H instead of ω_k and ζ_H^k. We begin by deriving a bound for $\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla_\theta J(\theta_k)\|^2$. It is known that J is L_J-smooth with $L_J := \frac{G_2}{(1-\gamma)^2} + \frac{2G_1^2}{(1-\gamma)^3}$ [Lemma B.1, Liu et al. [2020b]].

With this, we obtain:

$$
\begin{aligned}
&J(\theta_{k+1}) \\
&\geq J(\theta_k) + \langle \nabla_\theta J(\theta_k), \theta_{k+1} - \theta_k \rangle - \frac{L_J}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= J(\theta_k) + \alpha \langle \nabla_\theta J(\theta_k), \omega_k \rangle - \frac{\alpha^2 L_J}{2}\|\omega_k\|^2 \\
&= J(\theta_k) + \alpha \langle \nabla_\theta J(\theta_k), \omega_k^* \rangle + \alpha \langle \nabla_\theta J(\theta_k), \omega_k - \omega_k^* \rangle - \frac{\alpha^2 L_J}{2}\|\omega_k - \omega_k^* + \omega_k^*\|^2 \\
&\overset{(a)}{\geq} J(\theta_k) + \alpha \left\langle \nabla_\theta J(\theta_k), F(\theta_k)^{-1}\nabla_\theta J(\theta_k) \right\rangle + \alpha \langle \nabla_\theta J(\theta_k), \omega_k - \omega_k^* \rangle \\
&\quad - \alpha^2 L_J \|\omega_k - \omega_k^*\|^2 - \alpha^2 L_J \|\omega_k^*\|^2 \\
&\overset{(b)}{\geq} J(\theta_k) + \frac{\alpha}{G_1^2}\|\nabla_\theta J(\theta_k)\|^2 + \alpha \langle \nabla_\theta J(\theta_k), \omega_k - \omega_k^* \rangle - \alpha^2 L_J \|\omega_k - \omega_k^*\|^2 - \alpha^2 L_J \|\omega_k^*\|^2 \\
&= J(\theta_k) + \frac{\alpha}{2G_1^2}\|\nabla_\theta J(\theta_k)\|^2 + \frac{\alpha}{2G_1^2}\left[\|\nabla_\theta J(\theta_k)\|^2 + 2G_1^2 \langle \nabla_\theta J(\theta_k), \omega_k - \omega_k^* \rangle + G_1^4 \|\omega_k - \omega_k^*\|^2\right] \\
&\quad - \left(\frac{\alpha G_1^2}{2} + \alpha^2 L_J\right)\|\omega_k - \omega_k^*\|^2 - \alpha^2 L_J \|\omega_k^*\|^2 \\
&= J(\theta_k) + \frac{\alpha}{2G_1^2}\|\nabla_\theta J(\theta_k)\|^2 + \frac{\alpha}{2G_1^2}\|\nabla_\theta J(\theta_k) + G_1^2(\omega_k - \omega_k^*)\|^2 - \left(\frac{\alpha G_1^2}{2} + \alpha^2 L_J\right)\|\omega_k - \omega_k^*\|^2 \\
&\quad - \alpha^2 L_J \|\omega_k^*\|^2 \\
&\geq J(\theta_k) + \frac{\alpha}{2G_1^2}\|\nabla_\theta J(\theta_k)\|^2 - \left(\frac{\alpha G_1^2}{2} + \alpha^2 L_J\right)\|\omega_k - \omega_k^*\|^2 - \alpha^2 L_J \|F(\theta_k)^{-1}\nabla_\theta J(\theta_k)\|^2 \\
&\overset{(c)}{\geq} J(\theta_k) + \left(\frac{\alpha}{2G_1^2} - \frac{\alpha^2 L_J}{\mu^2}\right)\|\nabla_\theta J(\theta_k)\|^2 - \left(\frac{\alpha G_1^2}{2} + \alpha^2 L_J\right)\|\omega_k - \omega_k^*\|^2
\end{aligned}
\tag{68}
$$

where (a) utilizes the Cauchy-Schwarz inequality and the definition that $\omega_k^* = F(\theta_k)^{-1}\nabla_\theta J(\theta_k)$. Inequalities (b), and (c) follow from Assumption 6(a) and 7 respectively. We take the above inequality, sum over $k = 0, \cdots, K-1$, rearrange the terms and substitute $\alpha = \frac{\mu^2}{4G_1^2 L_J}$, to obtain:

$$
\begin{aligned}
\frac{\mu^2}{16G_1^4 L_J}\left(\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla_\theta J(\theta_k)\|^2\right) &\leq \frac{J(\theta_K) - J(\theta_0)}{K} + \left(\frac{\mu^2}{8L_J} + \frac{\mu^4}{16G_1^4 L_J}\right)\left(\frac{1}{K}\sum_{k=0}^{K-1}\|\omega_k - \omega_k^*\|^2\right) \\
&\overset{(a)}{\leq} \frac{2}{(1-\gamma)K} + \left(\frac{\mu^2}{8L_J} + \frac{\mu^4}{16G_1^4 L_J}\right)\left(\frac{1}{K}\sum_{k=0}^{K-1}\|\omega_k - \omega_k^*\|^2\right)
\end{aligned}
\tag{69}
$$

where (a) uses the fact that $J(\cdot)$ is absolutely bounded above by $(1-\gamma)^{-1}$. Using (69), we obtain

$$
\frac{\mu^{-2}}{K}\left(\sum_{k=0}^{K-1}\|\nabla_\theta J(\theta_k)\|^2\right) \leq \frac{32 L_J G_1^4}{\mu^4 K} + \left(\frac{2G_1^4}{\mu^2} + 1\right)\left(\frac{1}{K}\sum_{k=0}^{K-1}\|\omega_k - \omega_k^*\|^2\right) \tag{70}
$$

Substituting Lemma 4 in Lemma 2, we obtain

$$
\begin{aligned}
\mathbb{E}_k\left\|\omega_k - \omega_k^*\right\|^2 \leq \mathcal{O}\Bigg(&\frac{G_1^4}{H\,\mu^4(1-\gamma)^4} + \frac{G_1^2 (C_1')^2 \log(H/\delta)}{H\,\mu^2(1-\gamma)^4} + \frac{1}{\mu^2\, m^{1/2}} + \frac{G_1^2\,\epsilon_{\text{app}}}{\mu^2\,(1-\gamma)^2} \\
&+ \frac{G_1^2}{\mu^2}\frac{\mathbb{E}\,\|\tilde{\zeta}_0 - \tilde{\zeta}_*^k\|^2}{H^2} + \frac{G_1^2}{\mu^2}\frac{\log^2(H/\delta)}{\lambda_0^2(1-\gamma)^2\,H} + \frac{G_1^2}{\mu^2}\frac{\log(H/\delta)}{\lambda_0(1-\gamma)\,m^{1/2}} \\
&+ \frac{G_1^2}{\mu^2}\frac{1}{\lambda_0^4(1-\gamma)^4\,T^\kappa}\Bigg),
\end{aligned}
\tag{71}
$$

$$\left\|\mathbb{E}_k[\omega_k] - \omega_k^*\right\|^2 \leq \mathcal{O}\Bigg(\frac{G_1^4\,(C_1')^2\,\log(H/\delta)}{T^\kappa} \;+\; \frac{\|\,\mathbb{E}[\zeta_0] - \zeta_*^k\|^2}{\lambda_0^2(1-\gamma)^2\,H^2} \;+\; \frac{\log^4(H/\delta)}{\lambda_0^6(1-\gamma)^6\,H^2} \\ +\; \frac{\sqrt{\log(H/\delta)}}{\lambda_0(1-\gamma)\,m^{1/2}} \;+\; \frac{1}{\lambda_0^{10}(1-\gamma)^{10}\,T^{2\kappa}} \;+\; \frac{G_1^2\,\epsilon_{\text{app}}}{\mu^2\,(1-\gamma)^2}\Bigg). \tag{72}$$

Now combining (70), (71) and (72) with (67), and substituting $K = \sqrt{T}$, $M = 2t_{\text{mix}}\lfloor \log T \rfloor$ and $H = (\sqrt{T})/M$, we obtain the following bound

$$J^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[J(\theta_k)] \leq \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} + \frac{G_1^2\sqrt{\epsilon_{\text{app}}}}{\mu(1-\gamma)} + \mathcal{O}\left(\frac{t_{\text{mix}}\,\log^3(T/\delta)}{(1-\gamma)^3\sqrt{T}} + \frac{1}{m^{1/4}(1-\gamma)^{1/2}} + \frac{\mathbb{E}_{s\sim d^{\pi^*}}\left[\text{KL}(\pi^*\|\pi_{\theta_0})\right]}{\sqrt{T}}\right).$$