Non-Asymptotic Analysis for Single-Loop (Natural) Actor-Critic with Compatible Function Approximation

Yudan Wang [1] **Yue Wang** [2] **Yi Zhou** [3] **Shaofeng Zou** [1 4]

Abstract

Actor-critic (AC) is a powerful method for learning an optimal policy in reinforcement learning, where the critic uses algorithms, e.g., temporal difference (TD) learning with function approximation, to evaluate the current policy and the actor updates the policy along an approximate gradient direction using information from the critic. This paper provides the *tightest* non-asymptotic convergence bounds for both the AC and natural AC (NAC) algorithms. Specifically, existing studies show that AC converges to an $\epsilon + \varepsilon_{\text{critic}}$ neighborhood of stationary points with the best known sample complexity of $\mathcal{O}(\epsilon^{-2})$ (up to a log factor), and NAC converges to an $\epsilon + \varepsilon_{\text{critic}} + \sqrt{\varepsilon_{\text{actor}}}$ neighborhood of the global optimum with the best known sample complexity of $\mathcal{O}(\epsilon^{-3})$, where $\varepsilon_{\text{critic}}$ is the approximation error of the critic and $\varepsilon_{\text{actor}}$ is the approximation error induced by the insufficient expressive power of the parameterized policy class. This paper analyzes the convergence of both AC and NAC algorithms with compatible function approximation. Our analysis eliminates the term $\varepsilon_{\text{critic}}$ from the error bounds while still achieving the best known sample complexities. Moreover, we focus on the challenging single-loop setting with a single Markovian sample trajectory. Our major technical novelty lies in analyzing the stochastic bias due to policy-dependent and time-varying compatible function approximation in the critic, and handling the non-ergodicity of the MDP due to the single Markovian sample trajectory. Numerical results are also provided in the appendix.

[1]Electrical Engineering, University at Buffalo [2]Electrical and Computer Engineering, University of Central Florida [3]Electrical and Computer Engineering, University of Utah [4]Computer Science & Engineering, University at Buffalo. Correspondence to: Shaofeng Zou <szou3@buffalo.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Actor-Critic (AC) (Barto et al., 1983; Konda & Tsitsiklis, 2003) is a reinforcement learning algorithm that combines the advantages of actor-only methods and critic-only methods by alternatively performing policy gradient update (actor) and action-value function estimation (critic) in an online fashion. Specifically, the critic uses a parameterized function to estimate the value function of the current policy, e.g., temporal difference (TD) (Sutton, 1988) and Q-learning (Watkins & Dayan, 1992). Then the actor updates the policy along an approximate gradient direction based on the estimate from the critic using approaches such as policy gradient (Sutton et al., 1999) and natural policy gradient (Kakade, 2001). In contrast to critic-only methods, AC methods, which are gradient based, usually have desirable convergence properties when combined with the approach of function approximation. However, critic-only methods may not converge or even diverge when applied together with function approximation (Baird, 1995; Gordon, 1996). Moreover, AC methods also enjoy a reduced variance due to the critic, and thus their convergence is typically more stable and faster than actor only methods.

While the asymptotic convergence for AC and NAC has been well understood in the literature, e.g., (Bhatnagar et al., 2009; Kakade, 2001; Konda & Tsitsiklis, 2003; Suttle et al., 2023), its non-asymptotic convergence analysis has been largely open until very recently. The non-asymptotic analysis is of great practical importance as it answers the questions that how many samples are needed and how to appropriately choose the different learning rates for the actor and the critic. Existing studies show that AC converges to an $\epsilon + \varepsilon_{\text{critic}}$ neighborhood of stationary points with the best known sample complexity of $\mathcal{O}\left(\epsilon^{-2}\right)$ (Chen et al., 2021; Olshevsky & Gharesifard, 2023; Xu et al., 2020a), and NAC converges to an $\epsilon + \varepsilon_{\text{critic}} + \sqrt{\varepsilon_{\text{actor}}}$ neighborhood of the global optimum with the best known sample complexity of $\mathcal{O}(\epsilon^{-3})$ (Chen et al., 2022; Xu et al., 2020a), where $\varepsilon_{\text{critic}}$ is the approximation error of the critic and $\varepsilon_{\text{actor}}$ is the approximation error induced by the insufficient expressive power of the parameterized policy class. In this paper, when presenting sample complexity, we omit the log factors. In these studies, the critic employs a fixed class of parameter-

Table 1. Comparison of sample complexity of AC

Reference	Single-loop	Sample size	Error	Comments
(Wang et al., 2020)	$\times$	$\mathcal{O}\left(\epsilon^{-6}\right)$	$\epsilon+\varepsilon_{\text{critic}}$	Critic: neural
(Zhou & Lu, 2023)	$\surd$	$\mathcal{O}\left(\epsilon^{-1}\right)$	ϵ	LQR
(Chen et al., 2023)	$\surd$	$\mathcal{O}\left(\epsilon^{-2.5}\right)$		
(Zhang et al., 2020b)	$\surd$	Asymptotic	$\epsilon+\varepsilon_{\text{critic}}$	Actor: non-linear, smooth Critic: linear function approx.
(Qiu et al., 2021)	$\times$	$\mathcal{O}\left(\epsilon^{-4}\right)$		
(Kumar et al., 2023)	$\times$	$\mathcal{O}(\epsilon^{-3})$		
(Kumar et al., 2023) (Xu et al., 2020b)	$\times$	$\mathcal{O}\left(\epsilon^{-2.5}\right)$		
(Xu et al., 2020a; Suttle et al., 2023)	$\times$	$\mathcal{O}\left(\epsilon^{-2}\right)$		
(Barakat et al., 2022) (Wu et al., 2020)	$\surd$	$\mathcal{O}\left(\epsilon^{-2.5}\right)$		
(Olshevsky & Gharesifard, 2023) (Chen et al., 2021)	$\surd$	$\mathcal{O}\left(\epsilon^{-2}\right)$		
Our Work	$\surd$	$\mathcal{O}(\epsilon^{-2})$	ϵ	

ized functions (typically linear function approximation with fixed feature), which may not satisfy the compatible condition (Sutton et al., 1999) (see Section 2 for details). This will result in a non-diminishing bias in the policy gradient estimate, and therefore, an additional error term $\varepsilon_{\text{critic}}$ is incurred in the overall error bound. Several works (Cayci et al., 2022; Wang et al., 2020) propose to use overparameterized neural networks in the critic to mitigate this issue, where $\varepsilon_{\text{critic}}$ diminishes as the network size increases. However, the convergence of the critic requires stringent conditions that are hard to verify (Cayci et al., 2022; Wang et al., 2020), and large neural network introduces expensive computational and memory costs. Actually, if the critic employs the approach of compatible function approximation, which is *linear*, then $\varepsilon_{\text{critic}}$ vanishes without introducing additional computational and memory costs (Sutton et al., 1999) (see details in Section 2). Moreover, for NAC applied with fixed function approximation in the critic, one needs to explicitly estimate the Fisher information matrix and compute its inverse, which will be computationally and memory expensive. Another advantage of compatible function approximation when applied with NAC is that the inverse of the Fisher information in the natural gradient will cancel out with the policy gradient (see Proposition 2), and thus there is no need to estimate the Fisher information matrix anymore.

1.1. Challenges and Contributions

Though AC and NAC with compatible function approximation enjoy no approximation error from the critic and no need of estimating the Fisher information matrix (for NAC), their non-asymptotic convergence analyses are much more challenging than the ones with fixed function approximation. To the best of the authors' knowledge, this paper develops the tightest non-asymptotic error bounds for AC and NAC algorithms, and our analyses are for the challenging case of single Markovian sample trajectory. We prove that AC with compatible function approximation converges to an ϵ stationary point with sample complexity $\mathcal{O}(\epsilon^{-2})$, and NAC with compatible function approximation converges to an $\epsilon+\sqrt{\varepsilon_{\text{actor}}}$ neighborhood of the globally optimal policy with sample complexity $\mathcal{O}(\epsilon^{-3})$. Our non-asymptotic error bounds outperform the best known AC and NAC bounds in the literature by a constant $\varepsilon_{\text{critic}}$ and achieve the same sample complexity: $\mathcal{O}(\epsilon^{-2})$ for AC and $\mathcal{O}(\epsilon^{-3})$ for NAC (see Tables 1 and 2). We note that this constant $\varepsilon_{\text{critic}}$ is due to the approximation error of the function class used by the critic, and does not diminish with time.

One of the biggest challenges in the analysis is due to the time-varying critic feature function. Specifically, the critic with compatible function approximation employs an ω-dependent linear function class, where ω is the policy parameter. As the actor updates the policy, the feature function of the critic also changes with ω. Therefore, the critic is using a linear function with time-varying ω-dependent feature to track the value function of the current policy π_ω, which is also time varying. This makes the analysis of the tracking error, i.e., the error between the ideal limit of the critic given the current policy and its current estimate, challenging. In this paper, we design a novel approach to explicitly bound this error. The central idea is to construct an auxiliary eligibility trace with fixed feature to approximate the eligibility trace with time-varying feature (in the critic, we use k-step TD with compatible function approximation).

In this paper, we focus on the challenging single-loop setting with a single Markovian sample trajectory. Some studies tried to decouple the updates of the actor and the critic using approaches, e.g., nested loop (Qiu et al., 2021; Agarwal et al., 2021; Chen et al., 2022; Xu et al., 2020a; Suttle et al., 2023), and to further develop the non-asymptotic analysis. Specifically, after the actor updates the policy, then the policy is fixed and the critic starts an inner loop to iterate sufficient number of steps until it gets a perfect evaluation of the current policy. This decoupling approach makes it easier to analyze as there is no need to analyze the interaction between the actor and the critic. However, this

Table 2. Comparison of sample complexity of NAC

Reference	Single-loop	Sample size	Error	Comments
(Khodadadian et al., 2022)	$\surd$	$\mathcal{O}\left(\epsilon^{-6}\right)$	ϵ	Tabular case
(Khodadadian et al., 2021)	$\times$	$\mathcal{O}\left(\epsilon^{-3}\right)$		
(Wang et al., 2020)	$\times$	$\mathcal{O}\left(\epsilon^{-6}\right)$	$\epsilon+\varepsilon_{\text{critic}}$ $+\sqrt{\varepsilon_{\text{actor}}}$	Critic: neural
(Cayci et al., 2022)	$\times$	$\mathcal{O}\left(\epsilon^{-3}\right)$		
(Agarwal et al., 2021)	$\times$	$\mathcal{O}\left(\epsilon^{-6}\right)$	$\epsilon+\varepsilon_{\text{critic}}$ $+\sqrt{\varepsilon_{\text{actor}}}$	Actor: non-linear, smooth Critic: linear function approx.
(Xu et al., 2020a)	$\times$	$\mathcal{O}\left(\epsilon^{-3}\right)$		
(Xu et al., 2020b)	$\times$	$\mathcal{O}\left(\epsilon^{-4}\right)$		
(Chen et al., 2022)	$\times$	$\mathcal{O}\left(\epsilon^{-3}\right)$		
Our Work	$\surd$	$\mathcal{O}(\epsilon^{-3})$	$\epsilon+\sqrt{\varepsilon_{\text{actor}}}$	

decoupling approach does not enjoy benefits from the two time-scale structure in the original AC and NAC algorithms (Konda & Tsitsiklis, 2003; Bhatnagar et al., 2009), e.g., algorithmic simplicity and statistical efficiency, and techniques therein cannot be generalized to analyze the single-loop single-trajectory two time-scale AC and NAC algorithms. Moreover, analyses therein require some kind of i.i.d. assumptions or require trajectories starting from any arbitrary state, which might be difficult to guarantee in practice. To develop the tightest bound, we develop a novel approach that bounds the tracking error as a function of the policy gradient norm (for AC) and the optimality gap (for NAC). We also note that our analysis for NAC does not need the smoothness assumption on the parameterized policy, which is typically required in existing NAC and AC analyses (Chen et al., 2021; Olshevsky & Gharesifard, 2023).

1.2. Related Work

In this section, we review recent relevant works on non-asymptotic analyses on reinforcement learning algorithms with function approximation. We provide a detailed comparison between our results and existing studies on AC and NAC in Tables 1 and 2. The "Sample complexity" in the table is the one needed to guarantee the gradient norm/optimality gap less than or equal to the "Error".

Actor-critic analyses. We list recent works on non-asymptotic analyses for AC in Table 1. Based on whether the updates of actor and critic are decoupled, the results can be grouped into "single-loop" and "nested-loop/decoupling" approaches. For a general MDP, the best known sample complexity for both single-loop and nested-loop approaches is $\mathcal{O}(\epsilon^{-2})$ (Chen et al., 2021; Olshevsky & Gharesifard, 2023; Xu et al., 2020a; Suttle et al., 2023). The only exception is (Zhou & Lu, 2023), which is due to the special structure of the LQR problem. These studies all use a fixed function class in the critic, and therefore, the convergence error consists of a non-diminishing constant term of $\varepsilon_{\text{critic}}$. In this paper, we analyze the AC with compatible function approximation, and we obtain a strictly tighter error bound without $\varepsilon_{\text{critic}}$. Our analysis is also much more challenging than the ones in the literature, which is mainly due to that the function class in the critic varies with the policy in the actor.

Natural actor-critic analyses. We list recent works on non-asymptotic analyses for NAC in Table 2. The best sample complexity for single-loop NAC is $\mathcal{O}\left(\epsilon^{-6}\right)$ and it is for the tabular case (Khodadadian et al., 2022), whereas the best sample complexity for nested-loop/decoupling NAC is $\mathcal{O}\left(\epsilon^{-3}\right)$ with an error of $\epsilon+\varepsilon_{\text{critic}}+\sqrt{\varepsilon_{\text{actor}}}$ (Chen et al., 2022; Xu et al., 2020a). There exists a gap of $\mathcal{O}\left(\epsilon^{-3}\right)$ between these two approaches, which is mainly due to the challenge in bounding the tracking error for NAC in the single-loop setting. In this paper, we close this gap and show that NAC in the single-loop setting can also achieve the sample complexity of $\mathcal{O}(\epsilon^{-3})$, and more importantly with a reduced error of $\epsilon+\sqrt{\varepsilon_{\text{actor}}}$.

Actor/critic only analyses. Non-asymptotic analyses for critic only methods have been extensively studied recently, e.g., TD (Srikant & Ying, 2019; Lakshminarayanan & Szepesvari, 2018; Bhandari et al., 2018; Cai et al., 2019; Sun et al., 2020; Xu & Gu, 2020), SARSA (Zou et al., 2019), gradient TD (GTD) method (Dalal et al., 2018; Xu et al., 2019; Wang et al., 2021; 2017; Liu et al., 2015; Gupta et al., 2019; Kaledin et al., 2020; Ma et al., 2020; 2021; Wang & Zou, 2020). There are also non-asymptotic analyses for actor only method, e.g., (Bhandari & Russo, 2021; 2024; Agarwal et al., 2021; Mei et al., 2020; Li et al., 2021; Laroche & des Combes, 2021; Zhang et al., 2022; Cen et al., 2021; Zhang et al., 2020a; Xiao, 2022). In this paper, we focus on AC and NAC algorithms, where how the errors in the actor and the critic affects the other needs to be analyzed.

2. Preliminaries

Markov Decision Processes Consider a general reinforcement learning setting, where an agent interacts with a stochastic environment modeled as a Markov decision process (MDP). An MDP can be represented by a tuple

$\langle \mathcal{S}, \mathcal{A}, P, R \rangle$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the discrete finite action space, $R(\cdot,\cdot): \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the reward function. The transition kernel $P(\cdot|s,a)$ denotes the distribution of the next state if taking action a at state s, $\forall s \in \mathcal{S}, a \in \mathcal{A}$.

A stationary policy π maps a state $s \in \mathcal{S}$ to a probability distribution $\pi(\cdot|s)$ over the action space $\mathcal{A}$. Then the expected long term average reward for a policy π is defined as follows:

$$J(\pi) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{t=0}^{N-1} R(s_t, a_t) | \pi \right] = \mathbb{E}_{s\sim d_\pi, a\sim\pi(\cdot|s)} \left[R(s,a) \right],$$

where we denote by d_π the stationary distribution

$$d_\pi(s) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathbb{P}\left(s_t = s | \pi\right).$$

Denote by $D_\pi = d_\pi \times \pi$ the state-action stationary distribution. We rewrite $R_t := R(s_t, a_t)$. For a given policy π and an initial state s, the relative value function is defined as

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} R_t - J(\pi) | s_0 = s, \pi \right], \forall s \in \mathcal{S}.$$

Given initial state s and action a, the relative action value function (Q function) for a given policy π is defined as

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} R_t - J(\pi) | s_0 = s, a_0 = a, \pi \right], \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

The relative advantage function is defined as

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s), \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

The goal is to find the optimal policy π^* that maximizes the long term average reward: $\max_\pi J(\pi)$.

(Natural) Actor-Critic with Compatible Function Approximation Consider a parameterized policy class $\Pi_\omega = \{\pi_\omega : \omega \in \mathcal{W}\}$, where $\mathcal{W} \subseteq \mathbb{R}^d$. Then the problem in Section 2 can be solved by optimizing over the parameter space $\mathcal{W}$. Specifically, the actor updates the policy via the approach of (natural) policy gradient, where the policy gradient is given by (Sutton et al., 1999)

$$\nabla J(\pi) = \mathbb{E}_{D_{\pi_\omega}} \left[Q^{\pi_\omega}(s,a) \, \phi_\omega(s,a) \right], \tag{1}$$

where $\phi_\omega(s,a) = \nabla_\omega \log \pi_\omega(a|s)$. We further let Φ_ω denote the feature matrix, which is the stack of all feature vectors. Specifically, $\Phi_\omega \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$ and the (s,a)-row of Φ_ω is $\phi_\omega^\top(s,a)$. On the other hand, the critic estimates the Q function in Equation (1) via the approach of TD learning, and the Q function is usually parameterized using linear function approximation in the existing literature, i.e., $\mathcal{Q} = \{Q_\theta(s,a) = \phi(s,a)^\top \theta, \theta \in \Theta\}$ where ϕ denotes the feature vector and $\Theta \subseteq \mathbb{R}^d$. However, as summarized in Tables 1 and 2, using a fixed ϕ introduces an additional non-vanishing error term $\varepsilon_{\text{critic}}$ to the gradient estimate.

To avoid the critic's function approximation error, (Sutton et al., 1999; Konda & Tsitsiklis, 2003) proposed a smart idea of compatible function approximation, which uses the compatible feature vector ϕ_ω that depends on the policy parameter ω. To explain, in order to approximate the value function Q^{π_ω} associated with policy π_ω, we can set the feature vector as $\phi_\omega(s,a) := \nabla_\omega \log \pi_\omega(a|s)$ and solve for the best linear approximation parameter $\bar{\theta}^*_\omega$ via the following optimization problem.

$$\bar{\theta}^*_\omega \in \arg\min_\theta \mathbb{E}_{D_{\pi_\omega}} \left[\left(Q^{\pi_\omega}(s,a) - \phi_\omega^\top(s,a)\theta \right)^2 \right]. \tag{2}$$

Proposition 1 ((Sutton et al., 1999))**.** *With compatible function approximation, the policy gradient $\nabla J(\pi_\omega)$ can be rewritten as:*

$$\nabla J(\pi_\omega) = \mathbb{E}_{D_{\pi_\omega}} \left[\nabla \log \pi_\omega(a|s) Q^{\pi_\omega}(s,a) \right] = \mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a)(\phi_\omega^\top(s,a)\, \bar{\theta}^*_\omega) \right]. \tag{3}$$

This implies that as long as we can solve the finite dimensional problem Equation (2), linear function approximation with the compatible feature ϕ_ω and parameter $\bar{\theta}^*_\omega$ does not induce any function approximation error. This approach is referred to as compatible function approximation (Sutton et al., 1999), i.e., estimating Q^{π_ω} using an ω-dependent linear function class: $\mathcal{Q}_\omega = \{\phi_\omega^\top(s,a)\theta, \theta \in \Theta\}$. To solve Equation (2) for the compatible function approximation parameter, we use the k-step TD algorithm with compatible feature ϕ_ω (Sutton et al., 1999).

The actor can also use the following natural policy gradient to update the policy (Kakade, 2001):

$$\widetilde{\nabla} J(\pi_\omega) = F_\omega^{-1} \nabla J(\pi_\omega),$$

where the matrix F_ω denotes the Fisher information matrix:

$$F_\omega = \mathbb{E}_{D_{\pi_\omega}} \left[\nabla \log \pi_\omega(a|s) \left(\nabla \log \pi_\omega(a|s) \right)^\top \right].$$

Proposition 2 ((Peters & Schaal, 2008))**.** *With compatible function approximation, natural policy gradient is reduced to:*

$$\widetilde{\nabla} J(\pi_\omega) = \bar{\theta}^*_\omega.$$

That is, there is no need to estimate the Fisher information matrix and compute its inverse, which is typically computationally expensive.

Algorithm 1 (Natural) Actor-Critic with Compatible Function Approximation

Initialization: k, $\eta_0, \omega_0, \pi_0 = \pi_{\omega_0}, \theta_0$, $\phi_0 = \nabla \log \pi_0$, $s_0, a_0 \sim \pi_0(\cdot|s_0), z_0 = 0$
for $t = 0, ..., T-1$ **do**
  Observe R_t; Sample $s_{t+1} \sim P(\cdot|s_t, a_t)$; $a_{t+1} \sim \pi_t(\cdot|s_{t+1})$
  $\phi_t(s,a) = \nabla_\omega \log \pi_t(a|s)$ /*Compatible function approximation*/
  Critic: $\delta_t(\theta_t) = R_t - \eta_t + \phi_t^\top(s_{t+1}, a_{t+1})\theta_t - \phi_t^\top(s_t, a_t)\theta_t$ /*TD error*/
    $z_t = \sum_{j=t-k}^{t} \phi_j(s_j, a_j)$ /*eligibility trace*/
    $\eta_{t+1} = \eta_t + \gamma_t(R_t - \eta_t)$ /*average reward update*/
    $\theta_{t+1} = \Pi_{2,B}(\theta_t + \alpha_t \delta_t(\theta_t) z_t)$ /*TD update*/
  Option I: $\omega_{t+1} = \omega_t + \beta_t \phi_t^\top(s_t, a_t)\theta_t \phi_t(s_t, a_t)$ /*Actor update in AC*/
  Option II: $\omega_{t+1} = \omega_t + \beta_t \theta_t$ /*Actor update in NAC*/
end for

3. Main Results

The detailed AC and NAC algorithms with compatible function approximation is summarized in Algorithm 1. In the critic update, α_t is the stepsize, and denote by $\Pi_{2,B}(v) = \arg\min_{\|\omega\|_2 \le B} \|v - \omega\|_2$ for any $v \in \mathbb{R}^d$ the project operator, and B is the radius. Next, we present the non-asymptotic bounds for the AC and NAC with compatible function approximation in Algorithm 1.

Assumption 1. *(Uniform Ergodicity) Consider the MDP with policy π_ω and transition kernel P, there exists constants $m > 0$, and $\rho \in (0,1)$ such that*

$$\sup_{s \in \mathcal{S}} \|\mathbb{P}(s_t \in \cdot | s_0 = s) - D_{\pi_\omega}(\cdot)\|_{\mathcal{TV}} \le m\rho^t.$$

Here $\|\cdot\|_{\mathcal{TV}}$ denotes the total variation distance between two distributions. Assumption 1 is widely used in the literature to handle the Markovian noise, e.g., (Srikant & Ying, 2019; Zou et al., 2019; Bhandari et al., 2018). We further assume that the d feature functions, $\phi_{\omega,i}, i = 1, ..., d$, are linearly independent, i.e., the feature matrix Φ_ω is full rank when $|\mathcal{S}||\mathcal{A}| \ge d$. This is also commonly used in the literature of analyzing RL algorithms with linear function approximation (Srikant & Ying, 2019; Zou et al., 2019; Bhandari et al., 2018).

3.1. Critic: k-step TD

Consider the critic update, where the TD method is used to learn the relative value function under the average-reward setting. It is known that the feature function needs to satisfy certain condition (Assumption 2 in (Tsitsiklis & Van Roy, 1999)) so that the limit of the TD method is unique. In the following proposition, we show that compatible function approximation automatically satisfy the assumption needed in (Tsitsiklis & Van Roy, 1999), and therefore guarantees the convergence of the critic without the need of any additional assumptions.

Proposition 3. *For any $\omega \in \mathcal{W}$ and $\theta \in \Theta$, $\Phi_\omega \theta \ne \mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^d$ is an all-one vector.*

We note that the results in (Wu et al., 2020) use a different assumption from the one in (Tsitsiklis & Van Roy, 1999) to guarantee the convergence of the critic in the average-reward setting (Assumption 4.1 in (Wu et al., 2020)): the matrix $\mathbb{E}[\phi(s)(\phi(s') - \phi(s))^\top]$ is negative definite, where ϕ is the fixed feature function, s is the current state and s' is the subsequent state.

As discussed in Section 2, we would like the critic to find the solution of Equation (2). However, the objective in Equation (2) requires the knowledge of Q^{π_ω}, which is unavailable. Therefore, in the critic, we propose to use the method of k-step TD, so that as k enlarges, the solution from the k-step TD converges to the solution of Equation (2). We present the k-step TD algorithm in Algorithm 2. Here, the AC and NAC algorithms in Algorithm 1 are single-loop, single sample trajectory and two time-scale. We introduce the k-step TD algorithm in Algorithm 2 only to illustrate the basic idea.

Based on Proposition 3, Assumption 1, and the assumption that Φ_ω is full rank, from (Tsitsiklis & Van Roy, 1999, Theorem 1), we can show that the k-step TD algorithm in Algorithm 2 has a unique solution, denoted by θ_ω^*:

$$\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega^\top(s,a)\left(\mathcal{T}_{\pi_\omega}^{(k)}\left(\phi_\omega^\top(s,a)\theta_\omega^*\right) - \phi_\omega^\top(s,a)\theta_\omega^*\right)\right] = \mathbf{0}, \tag{4}$$

where $\mathcal{T}_{\pi_\omega}^{(k)}(Q(s,a)) = \mathbb{E}\big[\sum_{j=0}^{k-1}(R_j - J(\pi_\omega)) + Q(s_k, a_k) | s_0 = s, a_0 = a, \pi_\omega\big]$.

Assume that $\mathbb{E}_{D_w}\left[\phi_w(s,a)\phi_w^\top(s,a)\right]$ is positive definite with the minimum eigenvalue $\lambda_{\min} > 0$. This is to guarantee that the solution to Equation (2) is unique. We can remove this assumption by adding a regularizer $\lambda\|\theta\|_2^2$ to Equation (2) to guarantee the solution to the regularized Equation (2) is unique, and bounding the difference.

Then we bound the difference between the solution to Equation (2) and the solution to the k-step TD algorithm in the following proposition.

Algorithm 2 Compatible k-step TD Algorithm

Initialization: $k,\ \eta, \theta_0,\ \phi = \nabla \log \pi_\omega,\ s_0, a_0 \sim \pi_\omega(\cdot|s_0), z_0 = 0$
for $t = 0, ..., T-1$ **do**
Observe R_t
$s_{t+1} \sim P(\cdot|s_t, a_t)$; $a_{t+1} \sim \pi_\omega(\cdot|s_{t+1})$
$\delta(\theta_t) = R_t - \eta + \phi^\top(s_{t+1}, a_{t+1})\theta_t - \phi^\top(s_t, a_t)\theta_t$ /*TD error*/
$z_t = \sum_{j=t-k}^{t} \phi(s_j, a_j)$ /*eligibility trace*/
$\eta = \eta + \gamma_t(R_t - \eta)$ /*average reward update*/
$\theta_{t+1} = \Pi_{2,B}(\theta_t + \alpha_t \delta(\theta_t) z_t)$ /*TD update*/
end for

Proposition 4. *For any $\omega \in \mathcal{W}$, the difference between θ^*_ω and $\bar{\theta}^*_\omega$ can be bounded as follows:*

$$\left\|\theta^*_\omega - \bar{\theta}^*_\omega\right\|_2 \le \frac{C_{gap} m\rho^k}{\lambda_{\min}},$$

where C_{gap} is a constant defined in Appendix A.2.

It can be seen that the bound diminishes exponentially with k. Therefore by picking a large k, the k-step TD is expected to solve Equation (2) to a desired accuracy.

3.2. Non-asymptotic Bound for AC

Assumption 2. *(Smoothness and Boundedness) For any $\omega, \omega' \in \mathbb{R}^d$ and any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, there exist positive constants L_ϕ, C_ϕ, C_π and L_δ such that*

1) $\|\phi_\omega(s,a) - \phi_{\omega'}(s,a)\|_2 \le L_\phi \|\omega - \omega'\|_2$;
2) $\|\pi_\omega(\cdot|s) - \pi_{\omega'}(\cdot|s)\|_{\mathcal{TV}} \le C_\pi \|\omega - \omega'\|_2$;
3) $\|\phi_\omega(s,a)\|_2 \le C_\phi$;
4) $\left\|\nabla^2 \pi_\omega(a|s)\right\|_2 \le C_\delta$;
5) $\left|\partial_{\omega_i}\partial_{\omega_j}\partial_{\omega_l}\pi_\omega(a|s)\right| \le L_\delta$, for $1 \le i, j, l \le n$.

The first three assumptions in Assumption 2 assume the policy and feature function ϕ_ω is smooth and bounded. The fourth and fifth assumptions in Assumption 2 are only needed for the AC analysis. For the NAC analysis, it is not necessary. We note that these assumptions can be easily satisfied by choosing a proper policy parameterization. For example, if the policy is parameterized using neural network, then these assumptions can be satisfied (Du et al., 2019; Miyato et al., 2018; Neyshabur, 2017) if the activation functions are analytic functions and have bounded each-order derivative, (e.g. logistic, hyperbolic tangent and softplus). With these proper policy parameterizations, the fifth one in Assumption 2 can be deduced by $\left\|\nabla^2\pi_\omega(a|s) - \nabla^2\pi_{\omega'}(a|s)\right\|_2 \le L_\delta \|\omega - \omega'\|_2$.

We first present the bound on the tracking error, which measures how the critic tracks its ideal limit:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta^*_t - \theta_t\|_2^2\right].$$

Here, θ_t is the critic parameter at time t of Algorithm 1, and we rewrite $\theta^*_t = \theta^*_{\omega_t}$ and $J(\omega_t) = J(\pi_{\omega_t})$ for convenience. In the AC algorithm, we set $\alpha_t = \alpha$, $\beta_t = \beta$, $\gamma_t = \gamma$, and $k = \mathcal{O}(\log T)$ such that $\gamma \ge \alpha \ge \beta \ge m\rho^k$. Note that we use a projection in Line 8 in Algorithm 1. In order for convergence and optimality, we require that all $\|\theta^*_\omega\| \le B$. A sufficient condition to guarantee this is to set $B = \frac{mR_{\max}C_\phi}{(1-\rho)\left(\lambda_{\min} - C_\phi^2 dm\rho^k\right)}$ (see Appendix A for the proof).

Proposition 5. *The tracking error of the AC algorithm in Algorithm 1 can be bounded as follows:*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta^*_t - \theta_t\|_2^2\right] \\
&\le \left(\frac{c_\alpha\beta}{\alpha} + \frac{c_\eta\beta}{\gamma}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \mathcal{O}\left(\frac{1}{T\alpha}\right) \\
&+ \mathcal{O}\left(\frac{\log^2 T}{T\gamma}\right) + \mathcal{O}\left(\alpha\log^2 T\right) + \mathcal{O}\left(\beta\log^3 T\right) \\
&+ \mathcal{O}\left(\gamma\log^3 T\right) + \mathcal{O}\left(\frac{\beta^2\log^2 T}{\alpha}\right) + \mathcal{O}\left(\frac{\beta^2\log T}{\gamma}\right) \\
&+ \mathcal{O}\left(\frac{\beta}{\alpha}(m\rho^k)\right) + \mathcal{O}\left((m\rho^k)\log^2 T\right) + \mathcal{O}\left(\frac{(m\rho^k)^2}{\alpha\beta}\right),
\end{aligned}$$

where c_α and c_η is a positive constant defined in Appendix B. Set $\gamma = \mathcal{O}(\frac{1}{\sqrt{T}})$, $\alpha = \mathcal{O}(\frac{1}{\sqrt{T}\log^2 T})$, $\beta = \mathcal{O}(\frac{1}{\sqrt{T}\log^2 T})$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\theta^*_t - \theta_t\|_2^2] \\
&\quad \le \frac{1}{4C_\phi^4 T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla J(\omega_t)\|_2^2] + \mathcal{O}\left(\frac{\log^3 T}{\sqrt{T}}\right). \qquad (5)
\end{aligned}$$

For simplicity, we only present the order of the bound here, and the detailed non-asymptotic bound can be found in the Appendix B. The key novelty in the analysis is that we bound the tracking error as a function of the policy gradient, and we also bound the policy gradient as a function of the tracking error. By applying the bound recursively, we get a tight bound on the tracking error in Proposition 5. Many existing studies in the two time-scale analysis upper bound the policy gradient in the tracking error using its maximum norm, which is constant-level. However, as we see in the following theorem, the policy gradient shall also decrease to zero. Therefore, the above approach does not obtain the tightest bound, and leads to a higher-order sample complexity.

Theorem 1. *Consider the AC algorithm in Algorithm 1. It can be shown that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \leq \frac{2C_\phi^4}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t^* - \theta_t\|_2^2\right] + \mathcal{O}\left(\frac{1}{T\beta}\right) + \mathcal{O}\left(\beta \log^2 T\right) + \mathcal{O}(m\rho^k).$$

Set $\gamma = \mathcal{O}(\frac{1}{\sqrt{T}}), \alpha, \beta = \mathcal{O}(\frac{1}{\sqrt{T}\log^2 T})$, then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \leq \mathcal{O}\left(\frac{\log^3 T}{\sqrt{T}}\right).$$

Theorem 1 implies that the AC algorithm with compatible function approximation converges to an ϵ-stationary point with sample complexity ϵ^{-2}. This improves the best known error bound by a constant $\varepsilon_{\text{critic}}$ (Wang et al., 2020; Zhang et al., 2020b; Qiu et al., 2021; Kumar et al., 2023; Xu et al., 2020b; Barakat et al., 2022; Wu et al., 2020; Chen et al., 2021; Olshevsky & Gharesifard, 2023; Xu et al., 2020a), and matches the best known sample complexity (Chen et al., 2021; Olshevsky & Gharesifard, 2023; Xu et al., 2020a; Suttle et al., 2023).

3.3. Non-asymptotic Bound for NAC

In this section, we present the non-asymptotic bound for the NAC algorithm in Algorithm 1. It was shown in (Agarwal et al., 2021) that due to the parameter invariant property of the natural policy gradient update, natural policy gradient is able to converge to the globally optimal policy with a gap that depends on the capacity of the policy class. Define the compatible linear function approximation error

$$\varepsilon_{\text{actor}} = \max_{\omega\in\mathcal{W}}\left\{\min_{\theta}\mathbb{E}_{D_{\pi_\omega}}\left[\left\|A^{\pi_\omega}(s,a) - \phi_\omega^\top(s,a)\theta\right\|_2^2\right]\right\}.$$

This error represents the approximation error due to the insufficient expressive power of the policy parameterization, and shall decrease if a large neural network is used.

Using the same idea as the one in AC, we can also develop a tight bound on the tracking error: $\mathcal{O}\left(T^{-\frac{1}{3}}\right)$, where now we bound the tracking error as a function of the optimality gap instead of the gradient norm. We then also develop bound of the optimality gap as a function of the tracking error. Applying them recursively, we obtain the tightest bound on the tracking error and the tightest bound on the optimality gap in the following theorem. We set $\alpha_t = \alpha$, $\beta_t = \beta$, $\gamma_t = \gamma$, and $k = \mathcal{O}(\log T)$ such that $\gamma \geq \alpha \geq \beta \geq m\rho^k$.

Assumption 3. *There exist a constant $C_\infty < \infty$ such that* $\sup_{\omega\in\mathcal{W}}\left\|\frac{D_{\pi^*}(s,a)}{D_{\pi_\omega}(s,a)}\right\|_\infty \leq C_\infty$.

Assumption 3 guarantees that the policy is sufficiently exploratory, and is commonly used in NAC analyses, e.g., (Cayci et al., 2022; Xu et al., 2020a; Agarwal et al., 2021). Approaches to guarantee this assumption were also studied in (Khodadadian et al., 2021; 2022).

Theorem 2. *Consider the NAC algorithm in Algorithm 1. Then, we have that*

$$\begin{aligned}\min_{t\leq T}\mathbb{E}\left[J(\pi^*) - J(\omega_t)\right] &\leq \mathcal{O}\left(\frac{\log^2 T}{T\alpha}\right) + \mathcal{O}\left(\frac{\log T}{T\beta}\right)\\ &+ \mathcal{O}\left(\frac{\gamma\sqrt{\log T}}{\sqrt{\alpha}}\right) + \mathcal{O}\left(\frac{\beta}{\sqrt{\alpha}}\right) + \mathcal{O}\left(\sqrt{\alpha\log^3 T}\right)\\ &+ \mathcal{O}\left(\frac{\sqrt{\log^3 T}}{T\sqrt{\alpha\beta}}\right) + \mathcal{O}\left(\sqrt{\frac{\beta\log T}{T\alpha}}\right) + \mathcal{O}\left(\sqrt{\frac{\log T}{T}}\right)\\ &+ \mathcal{O}\left(\frac{\sqrt{\gamma\beta}\log T}{\sqrt{\alpha}}\right) + \mathcal{O}\left(\sqrt{\beta\log^3 T}\right)\\ &+ \mathcal{O}\left(\sqrt{\frac{(m\rho^k)\beta}{\alpha}}\right) + \mathcal{O}\left(\frac{m\rho^k}{\sqrt{\alpha\beta}}\right) + \mathcal{O}\left(\sqrt{\frac{(m\rho^k)\gamma}{\alpha}}\right)\\ &+ \mathcal{O}\left(\sqrt{(m\rho^k)\log T}\right) + \mathcal{O}\left(\sqrt{\varepsilon_{actor}}\right).\end{aligned} \tag{6}$$

If we set $\gamma = \mathcal{O}(T^{-\frac{2}{3}}\log T), \alpha = \mathcal{O}(T^{-\frac{2}{3}}\log^{-1} T), \beta = \mathcal{O}(T^{-\frac{2}{3}}\log^{-1} T)$, we have

$$\min_{t\leq T}\mathbb{E}[J(\pi^*) - J(\omega_t)] \leq \mathcal{O}\left(T^{-\frac{1}{3}}\log^3 T\right) + \mathcal{O}\left(\sqrt{\varepsilon_{actor}}\right).$$

Remark 1. *Unlike the results for AC in Theorem 1, Theorem 2 for NAC only needs the first three assumptions in Assumption 2. This is one advantage of using compatible function approximation in NAC. As we can see from Line 11 in Algorithm 1 and Proposition 2, the inverse of the Fisher information matrix is cancelled out. Therefore, there is no stochastic noise from using $\phi_\omega(s_t,a_t)\phi_\omega^\top(s_t,a_t)$ in the analysis of NAC. However, in AC, we need to handle this noise, and therefore, the fourth assumption in Assumption 2 is needed for the AC algorithm.*

Theorem 2 implies that NAC with compatible function approximation converges to an $\epsilon+\sqrt{\varepsilon_{\text{actor}}}$-neighborhood of the globally optimal policy π^* with sample complexity $\mathcal{O}(\epsilon^{-3})$. Compared to existing studies, our work eliminate the approximation error of the critic, $\varepsilon_{\text{critic}}$, from the overall error bound (Wang et al., 2020; Cayci et al., 2022; Agarwal et al., 2021; Xu et al., 2020a;b; Chen et al., 2022). Moreover, as summarized in Table 2, the best known sample complexity of NAC is ϵ^{-3}, which however is for the nested-loop NAC variant (Xu et al., 2020a; Chen et al., 2022). Our results achieves this sample complexity, and is for the challenging single-loop NAC algorithm with a single Markovian sample trajectory.

Here we provide a proof sketch for the NAC algorithm to highlight major challenges and our technical novelties. The analysis of NAC contains of most major technical novelty in the AC analysis.

Proof sketch. For simplicity of presentation, we set $\hat{t} = \mathcal{O}\left(\frac{\log T}{\alpha}\right)$ and $\widetilde{T} = \hat{t}\lceil \frac{T}{\hat{t}\log T}\rceil$. We denote by

$$M_t = \mathbb{E}[||\theta_t - \theta_t^*||_2^2] + \mathbb{E}[(\eta_t - J(\omega_t))^2]$$

the sum of the tracking error and the estimation error of the average reward. Denote by $\mathrm{KL}(\omega_t) = KL(\pi^*|\pi_t)$ the KL divergence between policy π^* and π_t.

Step 1 (Error decomposition): According to the smoothness property of $\mathrm{KL}(\omega)$ with respect to ω, we bound the performance gap between the current policy and the optimal policy (optimality gap) as follows:

$$\frac{1}{\widetilde{\widetilde{T}}}\sum_{j=t}^{t+\widetilde{T}-1}\mathbb{E}[J(\pi^*) - J(\omega_j)] \leq \frac{\mathrm{KL}(\omega_{t+\widetilde{T}}) - \mathrm{KL}(\omega_t)}{\widetilde{T}\beta} + \mathcal{O}\left(\sqrt{\frac{1}{\widetilde{\widetilde{T}}}\sum_{j=t}^{t+\widetilde{T}-1} M_j}\right) + \mathcal{O}(C_\infty\sqrt{\varepsilon_{\text{actor}}} + \beta + m\rho^k).$$

Step 2 (Estimation error in the average reward): In this step, we analyze estimation error in the average reward: $\eta_t - J(\omega_t)$. We provide a tight characterization of this error:

$$\begin{aligned}&\mathbb{E}[(\eta_{t+1} - J(\omega_{t+1}))^2]\\ &\quad\leq (1-\gamma)\mathbb{E}[(\eta_t - J(\omega_t))^2] + \mathcal{O}(\beta\mathbb{E}[||\nabla J(\omega_t)||_2^2])\\ &\qquad+ \mathcal{O}(m\rho^k\gamma + k^2\gamma^2 + k^2\gamma\beta + \beta^2).\end{aligned}$$

One of our key novelties lies in that we bound this estimation error using the gradient norm $\mathbb{E}[||\nabla J(\omega_t)||_2^2]$. The above bound itself is tighter than the existing one in (Wu et al., 2020).

Step 3 (Tracking error): In this step, we bound the tracking error in the critic: $||\theta_t - \theta_t^*||_2^2$. By the TD error step in Algorithm 1, we decompose the term $||\theta_{t+1} - \theta_{t+1}^*||_2^2$ as follows:

$$\begin{aligned}&||\theta_{t+1} - \theta_{t+1}^*||_2^2\\ &\quad\leq ||\theta_t - \theta_t^*||_2^2 + ||\theta_t^* - \theta_{t+1}^*||_2^2 + \alpha^2||\delta_t z_t||_2^2\\ &\qquad+ 2\alpha\langle\theta_t - \theta_t^*, \delta_t z_t\rangle + 2\alpha\langle\theta_t^* - \theta_{t+1}^*, \delta_t z_t\rangle\\ &\qquad+ 2\langle\theta_t - \theta_t^*, \theta_t^* - \theta_{t+1}^*\rangle.\end{aligned}$$

Another key challenge lies in how to bound the term $\mathbb{E}[\langle\theta_t - \theta_t^*, \delta_t z_t\rangle]$. We develop a novel technique of auxiliary Markov chain to decompose this error into two parts: 1) error due to time-varying feature function and 2) error due to time-varying policy. Specifically, consider the first Markov chain generated from the algorithm:

$$s_0, a_0 \overset{\pi_0\times P}{\rightarrow} s_1, a_1 \rightarrow ... \rightarrow s_t, a_t \overset{\pi_t\times P}{\rightarrow} s_{t+1}, a_{t+1},$$

where at each time j, the action is chosen according to π_j and the transition kernel is P. Here $z_t = \sum_{j=t-k}^{t}\phi_j(s_j, a_j)$ is the eligibility trace used in the algorithm. It can be seen that in z_t, the feature ϕ_j changes with j, and the distribution of s_j, a_j depends on the time-varying policy π_j. We then design an auxiliary eligibility trace $\hat{z}_t = \sum_{j=t-k}^{t}\phi_t(s_j, a_j)$, where the feature is fixed to be ϕ_t, and only the the distribution of s_j, a_j depends on the time-varying policy π_j. To further handle the time-varying distribution of s_j, a_j, we design an auxiliary Markov chain (denoted by $A1$) as follows:

$$A1 : (s_0, \widetilde{a}_0) \sim \pi_t \overset{\pi_t\times P}{\rightarrow} \widetilde{s}_1, \widetilde{a}_1 \rightarrow ... \widetilde{s}_t, \widetilde{a}_t \overset{\pi_t\times P}{\rightarrow} \widetilde{s}_{t+1}, \widetilde{a}_{t+1},$$

where the action at each time j is always chosen according to a fixed policy π_t. Based on this auxiliary Markov chain, we introduce another auxiliary eligibility trace $\widetilde{z}_t = \sum_{j=t-k}^{t}\phi_t(\widetilde{s}_j, \widetilde{a}_j)$, where it uses a fixed feature ϕ_t, and samples from this auxiliary Markov chain. Lastly, we design a second auxiliary Markov chain (denoted by $A2$):

$$A2 : (\bar{s}_0, \bar{a}_0) \sim D_t \overset{\pi_t\times P}{\rightarrow} \bar{s}_1, \bar{a}_1 \rightarrow ... \bar{s}_t, \bar{a}_t \overset{\pi_t\times P}{\rightarrow} \bar{s}_{t+1}, \bar{a}_{t+1},$$

where the only difference between A2 and A1 lies in the initial state distribution. Then we define the last auxiliary eligibility trace as $\bar{z}_t = \sum_{j=t-k}^{t}\phi_t(\bar{s}_j, \bar{a}_j)$.

The difference between z_t and $\hat{z}_t$ measures the error due to the time-varying compatible feature function. We bound this error using the Lipschitz continuity of the feature function. The difference between $\hat{z}_t$ and $\widetilde{z}_t$ measures the error due to the time-varying sampling policy. The difference between $\widetilde{z}_t$ and $\bar{z}_t$ measures the error due to the difference between the stationary distribution and the actual distribution of the samples, which can be bounded based on Assumption 1. By such a error decomposition, we can show that

$$\begin{aligned}&\mathbb{E}[||\theta_{t+1} - \theta_{t+1}^*||_2^2] \leq (1 - \bar{\lambda}_{\min}\alpha/2)\mathbb{E}[||\theta_t - \theta_t^*||_2^2]\\ &\quad+ \mathcal{O}(k^2\alpha\mathbb{E}[(\eta_t - J(\omega_t))^2]) + \mathcal{O}(\beta\mathbb{E}[||\nabla J(\omega_t)||_2^2])\\ &\quad+ \mathcal{O}\left(k^3\alpha^2 + k^3\alpha\beta + \beta^2 + m\rho^k\alpha + \frac{(m\rho^k)^2}{\beta}\right).\end{aligned}$$

Step 4 (Bound on gradient): As we can see from Steps 2 and 3, we bound the estimation error of the average reward and the tracking error using the gradient norm $\|\nabla J(\omega_t)\|_2^2$. Therefore, in order to derive the tightest bound, we further develop a novel bound on the gradient norm $\|\nabla J(\omega_t)\|_2^2$. Note that the idea is novel as it serves as a pivotal link connecting the analysis of the tracking error/estimation error in the average reward and the optimality gap. Specifically, we bound the gradient norm using the estimation error in the average reward and tracking error. By the smoothness of $J(\omega)$, we have that

$$\sum_{j=t}^{t+\widetilde{T}-1} \frac{\mathbb{E}[||\nabla J(\omega_j)||_2^2]}{\widetilde{T}} \leq 2C_\phi^2 \frac{J(\omega^*) - E[J(\omega_t)]}{\beta \widetilde{T}} + \mathcal{O}\left(\frac{1}{\widetilde{T}} \sum_{j=t}^{t+\widetilde{T}-1} \mathbb{E}[||\theta_j - \theta_j^*||_2^2]\right) + \mathcal{O}(m\rho^k + \beta).$$

We also note that we bound the gradient norm using the optimality gap, and this is of great importance to establish the tight bound in this paper. In previous works, this term $\mathbb{E}[J(\omega_{t+\widetilde{T}})] - E[J(\omega_t)]$ is bounded by a constant, and thus the overall complexity is not as tight.

Step 5: Combining steps 1-4, we conclude the proof. □

4. Conclusion

In this paper, we develop the tightest non-asymptotic convergence bounds for both the AC and NAC algorithms with compatible function approximation. For the AC algorithm, our results achieve the best sample complexity of ϵ^{-2} with a reduced error from $\epsilon + \varepsilon_{\text{critic}}$ to ϵ, where $\varepsilon_{\text{critic}}$ is a non-diminishing constant. For the NAC algorithm, our results is the first one in the literature that analyze the single-loop NAC with a single Markovian trajectory, and we achieve the best known sample complexity of ϵ^{-3} also with a reduced error of $\epsilon + \sqrt{\varepsilon_{\text{actor}}}$. Our results demonstrate the advantage of compatible function approximation when applied in AC and NAC algorithms, including relaxed technical condition to guarantee convergence, no need of estimating Fisher information matrix, and no approximation error from the critic. Our technical novelty lies in analyzing the error due to use of a time-varying and policy dependent feature in the critic.

Impact Statement

This paper presents work whose goal is to advance the field of Reinforcement Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

The work of Yudan Wang and Shaofeng Zou is supported by the National Science Foundation under Grants CCF-2007783, CCF-2106560 and ECCS-2337375 (CAREER). Yi Zhou's work is supported by the National Science Foundation under grants CCF-2106216, DMS-2134223, ECCS-2237830 (CAREER).

This material is based upon work supported under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 - National AI Institute for Exceptional Education. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

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

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings*, pp. 30–37. Elsevier, 1995.

Barakat, A., Bianchi, P., and Lehmann, J. Analysis of a target-based actor-critic algorithm with linear function approximation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 991–1040. PMLR, 2022.

Barto, A. G., Sutton, R. S., and Anderson, C. W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983. doi: 10.1109/TSMC.1983.6313077.

Bhandari, J. and Russo, D. On the linear convergence of policy gradient methods for finite MDPs. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pp. 2386–2394. PMLR, 2021.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *Operations Research*, 2024.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 1691–1692, 2018.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor–critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11312–11322, 2019.

Cayci, S., He, N., and Srikant, R. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *arXiv preprint arXiv:2206.00833*, 2022.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:25294–25307, 2021.

Chen, X., Duan, J., Liang, Y., and Zhao, L. Global convergence of two-timescale actor-critic for solving linear quadratic regulator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7087–7095, 2023.

Chen, Z., Khodadadian, S., and Maguluri, S. T. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. *Proceedings of Machine Learning Research*, 75:1–35, 2018.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1675–1685. PMLR, 2019.

Gordon, G. J. Chattering in SARSA (λ). *CMU Learning Lab Technical Report*, 1996.

Gupta, H., Srikant, R., and Ying, L. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4706–4715, 2019.

Heidergott, B. and Hordijk, A. Taylor series expansions for stationary markov chains. *Advances in Applied Probability*, 35(4):1046–1070, 2003.

Kakade, S. M. A natural policy gradient. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 14, pp. 1531–1538, 2001.

Kaledin, M., Moulines, E., Naumov, A., Tadic, V., and Wai, H.-T. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 2144–2203, 2020.

Khodadadian, S., Chen, Z., and Maguluri, S. T. Finite-sample analysis of off-policy natural actor-critic algorithm. In *International Conference on Machine Learning (ICML)*, pp. 5420–5431. PMLR, 2021.

Khodadadian, S., Doan, T. T., Romberg, J., and Maguluri, S. T. Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*, 2022.

Konda, V. R. and Tsitsiklis, J. N. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

Kumar, H., Koppel, A., and Ribeiro, A. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, pp. 1–35, 2023.

Lakshminarayanan, C. and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1347–1355, 2018.

Laroche, R. and des Combes, R. T. Dr jekyll & mr hyde: the strange case of off-policy policy updates. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Softmax policy gradient methods can take exponential time to converge. *Proceedings of Machine Learning Research vol*, 134:1–4, 2021.

Li, T., Guan, Z., Zou, S., Xu, T., Liang, Y., and Lan, G. Faster algorithm and sharper analysis for constrained markov decision process. *Operations Research Letters*, 54:107107, 2024.

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., and Petrik, M. Finite-sample analysis of proximal gradient td algorithms. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 504–513, 2015.

Ma, S., Zhou, Y., and Zou, S. Variance-reduced off-policy TDC learning: Non-asymptotic convergence analysis. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 14796–14806, 2020.

Ma, S., Chen, Z., Zhou, Y., and Zou, S. Greedy-gq with variance reduction: Finite-time analysis and improved complexity. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *Proc. International Conference on Machine Learning (ICML)*, pp. 6820–6829, 2020.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In

International Conference on Learning Representations (ICLR), 2018.

Neyshabur, B. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Olshevsky, A. and Gharesifard, B. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023.

Peters, J. and Schaal, S. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Qiu, S., Yang, Z., Ye, J., and Wang, Z. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation and TD learning. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 2803–2830, Jun. 2019.

Sun, J., Wang, G., Giannakis, G. B., Yang, Q., and Yang, Z. Finite-time analysis of decentralized temporal-difference learning with linear function approximation. *Proceedings of Machine Learning Research*, 108:4485–4495, 2020.

Suttle, W. A., Bedi, A., Patel, B., Sadler, B. M., Koppel, A., and Manocha, D. Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level monte carlo actor-critic. In *International Conference on Machine Learning (ICML)*, pp. 33240–33267. PMLR, 2023.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Sutton, R. S. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems (NIPS)*, 8, 1995.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems (NIPS)*, 12, 1999.

Tsitsiklis, J. N. and Van Roy, B. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations (ICLR)*, 2020.

Wang, Y. and Zou, S. Finite-sample analysis of Greedy-GQ with linear function approximation under Markovian noise. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pp. 11–20, 2020.

Wang, Y., Chen, W., Liu, Y., Ma, Z.-M., and Liu, T.-Y. Finite sample analysis of the gtd policy evaluation algorithms in Markov setting. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 5504–5513, 2017.

Wang, Y., Zou, S., and Zhou, Y. Non-asymptotic analysis for two time-scale tdc with general smooth function approximation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:9747–9758, 2021.

Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 17617–17628, 2020.

Xiao, L. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282): 1–36, 2022.

Xu, P. and Gu, Q. A finite-time analysis of Q-learning with neural network function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 10555–10565, 2020.

Xu, T., Zou, S., and Liang, Y. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10633–10643, 2019.

Xu, T., Wang, Z., and Liang, Y. Improving sample complexity bounds for (natural) actor-critic algorithms. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020a.

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

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4572–4583, 2020a.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. Provably convergent two-timescale off-policy actor-critic with function approximation. In *International Conference on Machine Learning (ICML)*, pp. 11204–11213. PMLR, 2020b.

Zhang, S., Des Combes, R. T., and Laroche, R. Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *Journal of Machine Learning Research*, 23(343):1–91, 2022.

Zhou, M. and Lu, J. Single timescale actor-critic method to solve the linear quadratic regulator with convergence guarantees. *Journal of Machine Learning Research*, 24 (222):1–34, 2023.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for SARSA with linear function approximation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8665–8675, 2019.

Appendix

Table of Contents

A. Supporting Lemmas and Proofs for Propositions 3 and 4

In this section, we provide a number of supporting lemmas, and proofs for Propositions 1 to 4. In the following proofs, $\|x\|_2$ denotes the ℓ_2 norm if x is a vector; and $\|X\|_2$ denotes the operator norm if X is a matrix.

A.1. Supporting Lemmas

For convenience, we denote $J(\omega) = J(\pi_\omega)$. We first prove a lemma showing that both $J(\omega)$ and $\nabla J(\omega)$ are Lipschitz in ω.

Lemma 1. *Under Assumptions 1 and 2, for any $\omega, \omega' \in \mathcal{W}$, we have that*

$$\|\nabla J(\omega)\|_2 \le C_J, \tag{7}$$

where $C_J = C_\phi^2 \left(B + \frac{C_{gap} m \rho^k}{\lambda_{\min}}\right)$, and

$$\|\nabla J(\omega) - \nabla J(\omega')\|_2 \le L_J \|\omega - \omega'\|_2, \tag{8}$$

where $L_J = \frac{m R_{\max}}{1-\rho}(4 L_\pi C_\phi + L_\phi)$ and $L_\pi = \frac{1}{2} C_\pi \left(1 + \lceil \log m^{-1} \rceil + \frac{1}{1-\rho}\right)$.

Recall Equation (2). The solution $\bar{\theta}^*_\omega$ given the feature function satisfies that

$$\bar{\theta}^*_\omega = \arg\min_\theta \mathbb{E}_{D_\omega}\left[\left\|Q^{\pi_\omega}(s,a) - \phi_\omega^\top(s,a)\theta\right\|_2^2\right]. \tag{9}$$

We show that the solution $\bar{\theta}^*_\omega$ is Lipschitz in ω in the following lemma.

Lemma 2. *For any $\omega, \omega' \in \mathcal{W}$, it holds that*

$$\left\|\bar{\theta}^*_\omega - \bar{\theta}^*_{\omega'}\right\|_2 \leq C_\Theta \left\|\omega - \omega'\right\|_2, \tag{10}$$

where $C_\Theta = \frac{C_J}{\lambda_{\min}^2}\left(2C_\phi L_\phi + C_\phi^2 L_\pi\right) + \frac{L_J}{\lambda_{\min}}$.

For any $\omega \in \mathcal{W}$, let

$$\begin{aligned} H_\omega(s,a) &= \mathbb{E}\left[\phi_\omega(s_0,a_0)\left(\phi_\omega(s_k,a_k) - \phi_\omega(s_0,a_0)\right)^\top | s_0 = s, a_0 = a, \pi_\omega\right], \\ H_\omega &= \mathbb{E}_{D_{\pi_\omega}}\left[H_\omega(s,a)\right]. \end{aligned} \tag{11}$$

Lemma 3. *For $k > \left\lceil \frac{\log(mdC_\phi^2) - \log\lambda_{\min}}{1-\rho}\right\rceil$, it holds that*

$$\lambda_{\max}\left(\frac{H_\omega + H_\omega^\top}{2}\right) \leq C_\phi^2 dm\rho^k - \lambda_{\min} = -\bar{\lambda}_{\min} < 0,$$

where $\lambda_{\max}(X)$ is the largest eigenvalue of symmetric matrix X.

When $k > \left\lceil \frac{\log(mdC_\phi^2) - \log\lambda_{\min}}{1-\rho}\right\rceil$, we have that

$$\begin{aligned} \bar{\lambda}_{\min} &= \lambda_{\min} - C_\phi^2 dm\rho^k \\ &> \lambda_{\min} - C_\phi^2 dm e^{-k(1-\rho)} \\ &\geq \lambda_{\min} - C_\phi^2 dm e^{-\log\left(\frac{mdC_\phi^2}{\lambda_{\min}}\right)} = 0, \end{aligned} \tag{12}$$

and therefore $\bar{\lambda}_{\min}$ is positive.

The following lemma bounds the distance between the stationary distribution induced by π_t and the distribution of s_t, a_t in Algorithm 1. Define $\mathcal{F}_j$ to be σ-field generated by all the randomness until the j-th time-step. For simplicity, we write D_{π_t} as D_t.

Lemma 4. *For any $0 \leq k \leq t$, it can be shown that*

$$\left\|\mathbb{P}\left(s_t, a_t | \mathcal{F}_{t-k}\right) - D_t\right\|_{\mathcal{TV}} \leq C_\pi \sum_{j=t-k}^{t-1} \left\|\omega_t - \omega_j\right\|_2 + m\rho^k. \tag{13}$$

We rewrite $\theta_t^* = \theta_{\omega_t}^*$, where θ_ω^* is the solution to Equation (4).

Lemma 5. *Consider the term $\mathbb{E}\left[\langle \theta_t - \theta_t^*, \delta_t z_t\rangle\right]$. It can be shown that*

$$\begin{aligned} \mathbb{E}\left[\langle \theta_t - \theta_t^*, \delta_t z_t\rangle\right] \leq &-\frac{\bar{\lambda}_{\min}}{2}\mathbb{E}\left[\left\|\theta_t - \theta_t^*\right\|_2^2\right] \\ &+ \frac{(k+1)^2 C_\phi^2}{2\bar{\lambda}_{\min}}\mathbb{E}\left[\left(J(\omega_t) - \eta_t\right)^2\right] + G_t^\delta, \end{aligned} \tag{14}$$

where $U_\delta = R_{\max} + 2C_\phi B$. For AC,

$$\begin{aligned} G_t^\delta = &\, 2B^2 C_\phi^2 U_\delta C_\pi \sum_{j=t-k}^{t}\sum_{i=j}^{t-1}\beta_i + 4BC_\phi U_\delta \sum_{j=t-k}^{t}\left(BC_\phi^2 C_\pi \sum_{i=j-k}^{j-1}\sum_{\iota=i}^{j-1}\beta_\iota + m\rho^k + BC_\phi^2 L_\pi \sum_{i=j}^{t-1}\beta_i\right) \\ &+ 2(k+1)C_\phi U_\delta\left((k+1)C_\phi U_\delta \sum_{j=t-2k}^{t-1}\alpha_j + C_\Theta C_\phi^2 B \sum_{j=t-2k}^{t-1}\beta_j + \frac{2C_{gap} m\rho^k}{\lambda_{\min}}\right), \end{aligned} \tag{15}$$

and for NAC,

$$G_t^\delta = 2B^2U_\delta C_\pi \sum_{j=t-k}^{t}\sum_{i=j}^{t-1}\beta_i + 4BC_\phi U_\delta \sum_{j=t-k}^{t}\left(BC_\pi \sum_{i=j-k}^{j-1}\sum_{\iota=i}^{j-1}\beta_\iota + m\rho^k + BL_\pi \sum_{i=j}^{t-1}\beta_i\right)$$
$$+ 2(k+1)C_\phi U_\delta\left((k+1)C_\phi U_\delta \sum_{j=t-2k}^{t-1}\alpha_j + C_\Theta B \sum_{j=t-2k}^{t-1}\beta_j + \frac{2C_{gap}m\rho^k}{\lambda_{\min}}\right). \tag{16}$$

In the following, we prove that θ_ω^* defined in Equation (4) is bounded.

Lemma 6. *The solution θ_ω^* to Equation* (4) *is bounded:*

$$\|\theta_\omega^*\|_2 \le \frac{1}{\lambda_{\min} - dC_\phi^2 m\rho^k}\frac{mC_\phi R_{\max}}{1-\rho} = \frac{mC_\phi R_{\max}}{\bar{\lambda}_{\min}(1-\rho)}. \tag{17}$$

Lemma 7. *Under Assumption 2 and 1, for any $\omega, \omega' \in \mathcal{W}$,*

$$\left\|\nabla^2 J(\omega) - \nabla^2 J(\omega')\right\|_2 \le L_\Theta \|\omega - \omega'\|_2, \tag{18}$$

where $L_\Theta = d^2\left(\frac{6C_\phi^3 m^3 e^{4R_{\max}}}{(1-\rho)^3} + \frac{6m^2 C_\phi C_\delta e^{3R_{\max}}}{(1-\rho)^2} + \frac{mL_\delta e^{2R_{\max}}}{1-\rho}\right)$.

The proof of above Lemmas could be found in Appendix D.

A.2. Proofs for Propositions 1 to 4

We include the proof of Proposition 1 and Proposition 2 for completeness.

Proof. By the Equation (2), $\bar{\theta}_\omega^*$ satisfies that

$$\mathbb{E}_{D_{\pi_\omega}}\left[(Q^{\pi_\omega}(s,a) - \phi_\omega^\top(s,a)\bar{\theta}_\omega^*)\phi_\omega(s,a)\right] = 0. \tag{19}$$

Since $\phi_\omega^\top(s,a)\bar{\theta}_\omega^*$ is a scalar, we can get that

$$\mathbb{E}_{D_{\pi_\omega}}\left[Q^{\pi_\omega}(s,a)\phi_\omega(s,a)\right] = \mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega^\top(s,a)\bar{\theta}_\omega^*\right] \tag{20}$$

For the policy gradient $\nabla J(\pi_\omega)$, we get that

$$\nabla J(\pi_\omega) = \mathbb{E}_{D_{\pi_\omega}}\left[\nabla \log \pi_\omega(a|s) Q^{\pi_\omega}(s,a)\right] = \mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)(\phi_\omega^\top(s,a)\,\bar{\theta}_\omega^*)\right]. \tag{21}$$

Furthermore, we have that

$$\begin{aligned}\widetilde{\nabla} J(\pi_\omega) &= F_\omega^{-1}\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega(s,a)^\top \bar{\theta}_\omega^*\right] \\ &= F_\omega^{-1}\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega(s,a)^\top\right]\bar{\theta}_\omega^* = \bar{\theta}_\omega^*.\end{aligned} \tag{22}$$

This conclude the proof. □

We present the proof of Proposition 3.

Proposition 6. *(Restatement of Proposition 3) For any $\omega \in \mathcal{W}$ and $\theta \in \Theta$, $\Phi_\omega\theta \neq \mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is an all-one vector.*

Proof. Assume that there exists $\theta_c \in \Theta$ such that $\Phi_\omega\theta_c = \mathbf{e}$, then $\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega^\top(s,a)\theta_c\right] = 1$.

However, note that

$$\begin{aligned}\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega^\top(s,a)\theta_c\right] &= \sum_s d_{\pi_\omega}(s)\sum_a \pi_\omega(a|s)\phi_\omega^\top(s,a)\theta_c\\ &= \sum_s d_{\pi_\omega}(s)\sum_a \pi_\omega(a|s)\nabla\log\pi_\omega(a|s)^\top\theta_c\\ &= \sum_s d_{\pi_\omega}(s)\sum_a \pi_\omega(a|s)\frac{\nabla_\omega\pi_\omega(a|s)^\top}{\pi_\omega(a|s)}\theta_c\\ &= \sum_s d_{\pi_\omega}(s)\sum_a \nabla_\omega\pi_\omega(a|s)^\top\theta_c\\ &= \sum_s d_{\pi_\omega}(s)\nabla_\omega\left(\sum_a\pi_\omega(a|s)\right)^\top\theta_c\\ &= 0, \end{aligned} \tag{23}$$

where the last equation is from the fact that $\sum_a \pi_\omega(a|s)=1$, and hence the gradient of it is 0. This hence results in a contradiction, which completes the proof. □

We then present the proof of Proposition 4.

Proposition 7. *(Restatement of Proposition 4) For any $\omega\in\mathcal{W}$, denote the fixed point of k-step TD operator by θ^*_ω, and the solution to Equation* (2) *by $\bar{\theta}^*_\omega$, then*

$$\left\|\theta^*_\omega-\bar{\theta}^*_\omega\right\|_2 \le \frac{C_{gap}m\rho^k}{\lambda_{\min}}, \tag{24}$$

where $C_{gap}=C_\phi^2B+\frac{C_\phi R_{\max}}{1-\rho}$.

Proof. From the definition, it holds that

$$\theta^*_\omega = \left(\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega^\top(s,a)\right]\right)^{-1}\left(\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\left(\mathcal{T}^{(k)}_{\pi_\omega}\phi_\omega^\top(s,a)\theta^*_\omega\right)\right]\right), \tag{25}$$

and

$$\bar{\theta}^*_\omega = \left(\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega^\top(s,a)\right]\right)^{-1}\left(\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)Q^{\pi_\omega}(s,a)\right]\right). \tag{26}$$

Thus, we have that

$$\begin{aligned}&\left\|\theta^*_\omega-\bar{\theta}^*_\omega\right\|_2\\ =&\left\|\left(\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega^\top(s,a)\right]\right)^{-1}\left(\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\left(\mathcal{T}^{(k)}_{\pi_\omega}\phi_\omega^\top(s,a)\theta^*_\omega-Q^{\pi_\omega}(s,a)\right)\right]\right)\right\|_2\\ \le&\frac{1}{\lambda_{\min}}\left\|\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\left(\mathcal{T}^{(k)}_{\pi_\omega}\phi_\omega^\top(s,a)\theta^*_\omega-Q^{\pi_\omega}(s,a)\right)\right]\right\|_2\\ =&\frac{1}{\lambda_{\min}}\left\|\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\left(\mathbb{E}\left[\sum_{j=0}^{k-1}R_j-J(\omega)+\phi_\omega(s_k,a_k)^\top\theta^*_\omega|s_0=s,a_0=a,\pi_\omega\right]-Q^{\pi_\omega}(s,a)\right)\right]\right\|\\ =&\frac{1}{\lambda_{\min}}\left\|\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\left(\mathbb{E}\left[\sum_{j=0}^{k-1}R_j-J(\omega)+\phi_\omega(s_k,a_k)^\top\theta^*_\omega|s_0=s,a_0=a,\pi_\omega\right]\right.\right.\right.\\ &\left.\left.\left.-\mathbb{E}\left[\sum_{j=0}^{\infty}R_j-J(\omega)|s_0=s,a_0=a,\pi_\omega\right]\right)\right]\right\|_2\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\lambda_{\min}} \left\| \mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a) \left(\mathbb{E} \left[\sum_{j=k}^{\infty} R_j - J(\omega) + \phi_\omega(s_k,a_k)^\top \theta_\omega^* | s_0 = s, a_0 = a, \pi_\omega \right] \right) \right] \right\|_2 \\
&\overset{(a)}{\leq} \frac{1}{\lambda_{\min}} \left\| \mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a) \left(\mathbb{E} \left[\sum_{j=k}^{\infty} R_j - J(\omega) \Big| (s_k,a_k) \sim D_{\pi_\omega}, \pi_\omega \right] \right) \right] \right\|_2 \\
&\quad + \frac{1}{\lambda_{\min}} \left\| \mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a) \left(\mathbb{E} \left[\phi_\omega(s_k,a_k)^\top \theta_\omega^* \big| (s_k,a_k) \sim D_{\pi_\omega} \right] \right) \right] \right\|_2 \\
&\quad + \frac{1}{\lambda_{\min}} C_\phi^2 B \mathbb{E}_{D_{\pi_\omega}} \left[\| \mathbb{P} \left(s_k, a_k | s_0 = s, a_0 = a, \pi_\omega \right) - D_{\pi_\omega} \|_{\mathcal{TV}} \right] \\
&\quad + \frac{1}{\lambda_{\min}} C_\phi \sum_{j=k}^{\infty} R_{\max} \mathbb{E}_{D_{\pi_\omega}} \left[\| \mathbb{P}(s_j, a_j | s_0 = s, a_0 = s, \pi_\omega) - D_{\pi_\omega} \|_{\mathcal{TV}} \right] \\
&\overset{(b)}{\leq} \frac{1}{\lambda_{\min}} \left\| \mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a) \left(\mathbb{E} \left[\sum_{j=k}^{\infty} R_j - J(\omega) \Big| (s_k,a_k) \sim D_{\pi_\omega}, \pi_\omega \right] \right) \right] \right\|_2 \\
&\quad + \frac{1}{\lambda_{\min}} \left\| \mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a) \left(\mathbb{E} \left[\phi_\omega(s_k,a_k)^\top \theta_\omega^* \big| (s_k,a_k) \sim D_{\pi_\omega} \right] \right) \right] \right\|_2 \\
&\quad + \frac{C_\phi}{\lambda_{\min}} \left(C_\phi B m \rho^k + \sum_{j=k}^{\infty} R_{\max} m \rho^j \right) \\
&\overset{(c)}{\leq} \frac{1}{\lambda_{\min}} \left\| \mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a) \left(\mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a)^\top \theta_\omega^* \right] \right) \right] \right\|_2 + \frac{1}{\lambda_{\min}} C_\phi \left(C_\phi B m \rho^k + \sum_{j=k}^{\infty} R_{\max} m \rho^j \right) \\
&\overset{(d)}{\leq} \frac{1}{\lambda_{\min}} \left(C_\phi^2 B m \rho^k + C_\phi R_{\max} \frac{m \rho^k}{1-\rho} \right) \\
&= \frac{C_{\text{gap}} m \rho^k}{\lambda_{\min}},
\end{aligned} \tag{27}
$$

where $C_{\text{gap}} = C_\phi^2 B + C_\phi R_{\max} \frac{1}{1-\rho}$, (a) follows from the triangular inequality and the fact that for any probability distribution P_1 and P_2, and any random variable X, s.t. $|X| \leq X_{\max}$, $|\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X]| \leq X_{\max} \|P_1 - P_2\|_{\mathcal{TV}}$, (b) follows from Assumption 1, (c) follows from $J(\omega) = \mathbb{E}_{D_{\pi_\omega}}[R(s,a)]$, and (d) follows from $\mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a) \left(\mathbb{E}_{D_{\pi_\omega}} \left[\phi_\omega(s,a)^\top \theta_\omega^* \right] \right) \right] = 0$. □

B. AC Sample Complexity Analysis

In this section, we provide the sample complexity analysis for our single-loop AC algorithm.

B.1. Bound on Gradient Norm in AC

In this section, we first present a preliminary bound on the gradient norm $\|\nabla J(\omega)\|$.

Lemma 8. *It holds that*

$$
\frac{\beta_t}{2} \mathbb{E} \left[\|\nabla J(\omega_t)\|_2^2 \right] \leq \mathbb{E}[J(\omega_{t+1})] - \mathbb{E}[J(\omega_t)] + C_\phi^4 \beta_t \mathbb{E} \left[\|\theta_t - \theta_t^*\|_2^2 \right] + G_t^\omega, \tag{28}
$$

where

$$
\begin{aligned}
G_t^\omega =& \frac{L_J C_\phi^4 B^2 \beta_t^2}{2} + \left(\frac{C_\phi^4 C_{gap}^2 m \rho^k}{\lambda_{\min}^2} + C_J C_\phi^2 B \right) \beta_t m \rho^k + 2 C_\phi^4 B^2 L_J \beta_t \sum_{j=t-k}^{t-1} \beta_j \\
&+ C_J C_\phi^4 B^2 C_\pi \beta_t \sum_{j=t-k}^{t-1} \sum_{i=j}^{t-1} \beta_i.
\end{aligned} \tag{29}
$$

Proof. Recall that in the update of AC algorithm, $\omega_{t+1}-\omega_t=\beta_t\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)$. Following Lemma 1, it can be shown that

$$\begin{aligned}
J(\omega_{t+1}) &\geq J(\omega_t)+\langle\nabla J(\omega_t),\omega_{t+1}-\omega_t\rangle-\frac{L_J}{2}\|\omega_{t+1}-\omega_t\|_2^2\\
&=J(\omega_t)+\beta_t\left\langle\nabla J(\omega_t),\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)\right\rangle-\frac{L_J\beta_t^2}{2}\left\|\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)\right\|_2^2\\
&=J(\omega_t)+\beta_t\left\langle\nabla J(\omega_t),\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)-\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\rangle\\
&\quad+\beta_t\left\langle\nabla J(\omega_t),\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\left(\theta_t-\theta_t^*\right)\phi_t(s,a)\right]\right\rangle+\beta_t\langle\nabla J(\omega_t),\nabla J(\omega_t)\rangle\\
&\quad+\beta_t\left\langle\nabla J(\omega_t),\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t^*\phi_t(s,a)\right]-\nabla J(\omega_t)\right\rangle-\frac{L_J\beta_t^2}{2}\left\|\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)\right\|_2^2\\
&\geq J(\omega_t)+\beta_t\left\langle\nabla J(\omega_t),\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)-\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\rangle\\
&\quad+\beta_t\|\nabla J(\omega_t)\|_2^2-\frac{\beta_t}{4}\|\nabla J(\omega_t)\|_2^2-\beta_t\left\|\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\left(\theta_t-\theta_t^*\right)\phi_t(s,a)\right]\right\|_2^2\\
&\quad-\frac{\beta_t}{4}\|\nabla J(\omega_t)\|_2^2-\beta_t\left\|\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t^*\phi_t(s,a)\right]-\nabla J(\omega_t)\right\|_2^2-\frac{L_JC_\phi^4B^2}{2}\beta_t^2\\
&\overset{(a)}{=}J(\omega_t)+\beta_t\underbrace{\left\langle\nabla J(\omega_t),\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)-\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\rangle}_{\text{part I}}\\
&\quad+\frac{\beta_t}{2}\|\nabla J(\omega_t)\|_2^2-\beta_t\left\|\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\left(\theta_t-\theta_t^*\right)\phi_t(s,a)\right]\right\|_2^2-\frac{L_JC_\phi^4B^2}{2}\beta_t^2\\
&\quad-\beta_t\left\|\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)(\theta_t^*-\bar{\theta}_t^*)\phi_t(s,a)\right]\right\|_2^2,
\end{aligned}\tag{30}$$

where we write $\bar{\theta}_{\omega_t}^*$ as $\bar{\theta}_t^*$ for convenience, (a) follows from Equation (3) that $\nabla J(\omega_t)=\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\bar{\theta}_t^*\phi_t(s,a)\right]$.

We then bound part I in Equation (30). Note that

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\langle\nabla J(\omega_t),\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)-\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\rangle\right]\right|\\
&\quad\leq\left|\mathbb{E}\left[\left\langle\nabla J(\omega_t)-\nabla J(\omega_{t-k}),\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)-\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\rangle\right]\right|\\
&\qquad+\left|\mathbb{E}\left[\left\langle\nabla J(\omega_{t-k}),\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)-\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\rangle\right]\right|\\
&\quad\leq\mathbb{E}\left[\|\nabla J(\omega_t)-\nabla J(\omega_{t-k})\|_2\left\|\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)-\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\|_2\right]\\
&\qquad+\left|\mathbb{E}\left[\nabla^\top J(\omega_{t-k})\mathbb{E}\left[\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)-\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]|\mathcal{F}_{t-k}\right]\right]\right|\\
&\quad\overset{(a)}{\leq}2C_\phi^2BL_J\mathbb{E}\left[\|\omega_t-\omega_{t-k}\|_2\right]+C_J\mathbb{E}\left[\|\mathbb{P}\left(s_t,a_t|\mathcal{F}_{t-k}\right)-D_t\|_{\mathcal{TV}}\right]C_\phi^2B\\
&\quad\overset{(b)}{\leq}2C_\phi^4B^2L_J\sum_{j=t-k}^{t-1}\beta_j+C_JC_\phi^2B\left(C_\pi\sum_{j=t-k}^{t-1}\mathbb{E}\left[\|\omega_t-\omega_j\|_2\right]+m\rho^k\right)\\
&\quad\leq2C_\phi^4B^2L_J\sum_{j=t-k}^{t-1}\beta_j+C_JC_\phi^2B\left(C_\pi C_\phi^2B\sum_{j=t-k}^{t-1}\sum_{i=j}^{t-1}\beta_i+m\rho^k\right),
\end{aligned}\tag{31}$$

where (a) is from the L_J-smoothness of J, and (b) is from Lemma 4. On the other hand, from Proposition 4, we can show that

$$\left\|\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\left(\bar{\theta}_t^*-\theta_t^*\right)\phi_t(s,a)\right]\right\|_2^2\leq C_\phi^4\left\|\theta_t^*-\bar{\theta}_t^*\right\|_2^2\leq C_\phi^4\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2.\tag{32}$$

Thus, combining Equation (30),Equation (31) and Equation (32) completes the proof,

$$\mathbb{E}\left[J(\omega_{t+1})\right]\geq\mathbb{E}\left[J(\omega_t)\right]+\frac{\beta_t}{2}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right]-\frac{L_JC_\phi^4B^2}{2}\beta_t^2-C_\phi^4\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]$$

$$- C_\phi^4 \beta_t \left(\frac{C_{\text{gap}} m \rho^k}{\lambda_{\min}}\right)^2 - 2C_\phi^4 B^2 L_J \beta_t \sum_{j=t-k}^{t-1} \beta_j - C_J C_\phi^2 B \beta_t \left(C_\pi C_\phi^2 B \sum_{j=t-k}^{t-1} \sum_{i=j}^{t-1} \beta_i + m\rho^k \right). \tag{33}$$

□

B.2. Bound on $|\eta_t - J(\omega_t)|$ in AC

In this section, we bound the error between η_t and $J(\omega_t)$, where $J(\omega_t) = \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{t=0}^{N-1} R_t | \pi_{\omega_t}\right]$ is the average-reward for policy π_{ω_t}.

Lemma 9. *If $\gamma_t - \gamma_t^2 \geq \beta_t$, then it holds that*

$$\mathbb{E}\left[(\eta_{t+1} - J(\omega_{t+1}))^2\right] \leq (1-\gamma_t)\,\mathbb{E}\left[(\eta_t - J(\omega_t))^2\right] + C_\phi^4 B^2 \beta_t \mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + G_t^\eta, \tag{34}$$

where

$$G_t^\eta = 2\gamma_t \left(R_{\max}^2 C_\pi C_\phi^2 B \sum_{j=t-k}^{t-1}\sum_{i=j}^{t-1} \beta_j + R_{\max}^2 m\rho^k + R_{\max}^2 \sum_{j=t-k}^{t-1} \gamma_j + R_{\max} C_J C_\phi^2 B \beta_t + R_{\max} C_J C_\phi^2 B \sum_{j=t-k}^{t-1} \beta_j \right)$$
$$+ R_{\max}^2 \gamma_t^2 + C_J^2 C_\phi^4 B^2 \beta_t^2 + 2R_{\max} L_J C_\phi^4 B^2 \beta_t^2. \tag{35}$$

Proof. Recall the update rule in Algorithm 1. Then we have that

$$\eta_{t+1} - J(\omega_{t+1}) = \eta_t + \gamma_t (R_t - \eta_t) - J(\omega_t) + J(\omega_t) - J(\omega_{t+1}). \tag{36}$$

It then follows that

$$\begin{aligned}
(\eta_{t+1} - J(\omega_{t+1}))^2 &= ((1-\gamma_t)(\eta_t - J(\omega_t)) + \gamma_t (R_t - J(\omega_t)) + J(\omega_t) - J(\omega_{t+1}))^2 \\
&\leq (1-\gamma_t)^2 (\eta_t - J(\omega_t))^2 + \gamma_t^2 (R_t - J(\omega_t))^2 + (J(\omega_t) - J(\omega_{t+1}))^2 \\
&\quad + 2\gamma_t \underbrace{(R_t - J(\omega_t))(J(\omega_t) - J(\omega_{t+1}))}_{\text{I}} + 2\gamma_t (1-\gamma_t) \underbrace{(\eta_t - J(\omega_t))(R_t - J(\omega_t))}_{\text{II}} \\
&\quad + 2(1-\gamma_t) \underbrace{(\eta_t - J(\omega_t))(J(\omega_t) - J(\omega_{t+1}))}_{\text{III}}.
\end{aligned} \tag{37}$$

The term $(J(\omega_t) - J(\omega_{t+1}))^2$ can be bounded by Lemma 1:

$$|J(\omega_t) - J(\omega_{t+1})| \leq C_J \|\omega_t - \omega_{t+1}\|_2 \leq C_J C_\phi^2 B \beta_t. \tag{38}$$

Term I in Equation (37) can be bounded as follows:

$$\begin{aligned}
|\mathbb{E}[(R_t - J(\omega_t))(J(\omega_t) - J(\omega_{t+1}))]| &\leq \mathbb{E}[|R_t - J(\omega_t)|\,|J(\omega_t) - J(\omega_{t+1})|] \\
&\leq R_{\max} C_J \mathbb{E}[\|\omega_{t+1} - \omega_t\|_2] \\
&\leq R_{\max} C_J C_\phi^2 B \beta_t.
\end{aligned} \tag{39}$$

Term II in Equation (37) can be bounded as follows:

$$\begin{aligned}
&|\mathbb{E}[(\eta_t - J(\omega_t))(R_t - J(\omega_t))]| \\
&\leq |\mathbb{E}[(\eta_{t-k} - J(\omega_{t-k}))(R_t - J(\omega_t))]| + |\mathbb{E}[(\eta_t - \eta_{t-k} - J(\omega_t) + J(\omega_{t-k}))(R_t - J(\omega_t))]| \\
&\overset{(a)}{\leq} |\mathbb{E}[\mathbb{E}[(\eta_{t-k} - J(\omega_{t-k}))(R_t - J(\omega_t))|\mathcal{F}_{t-k}]] - \mathbb{E}_{D_t}[(\eta_{t-k} - J(\omega_{t-k}))(R(s,a) - J(\omega_t))]| \\
&\quad + \mathbb{E}[|\eta_t - \eta_{t-k} - J(\omega_t) + J(\omega_{t-k})|\,|R_t - J(\omega_t)|]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} R_{\max}^2 \mathbb{E}\left[\|\mathbb{P}\left(s_t, a_t | \mathcal{F}_{t-k}\right), D_t\|_{\mathcal{TV}}\right] + R_{\max}\mathbb{E}\left[|\eta_t - \eta_{t-k}| + |J(\omega_t) - J(\omega_{t-k})|\right] \\
&\overset{(c)}{\leq} R_{\max}^2 \left(C_\pi \sum_{j=t-k}^{t-1} \mathbb{E}\left[\|\omega_t - \omega_j\|_2\right] + m\rho^k \right) + R_{\max}\left(R_{\max} \sum_{j=t-k}^{t-1} \gamma_j + C_J \mathbb{E}\left[\|\omega_t - \omega_{t-k}\|_2\right] \right) \\
&\leq R_{\max}^2 C_\pi C_\phi^2 B \sum_{j=t-k}^{t-1} \sum_{i=j}^{t-1} \beta_j + R_{\max}^2 m\rho^k + R_{\max}^2 \sum_{j=t-k}^{t-1} \gamma_j + R_{\max} C_J C_\phi^2 B \sum_{j=t-k}^{t-1} \beta_j,
\end{aligned} \tag{40}
$$

where (a) follows from $\mathbb{E}_{D_t}\left[R(s,a) - J(\omega_t)\right] = 0$, (b) follows from that $0 \leq \eta_t \leq R_{\max}$, $0 \leq J(\omega_t) \leq R_{\max}$, $0 \leq R_t \leq R_{\max}$ and (c) follows from Lemma 4.

Term III in Equation (37) can be bounded as follows:

$$
\begin{aligned}
&|\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right)\left(J(\omega_t) - J(\omega_{t+1})\right)\right]| \\
&\overset{(a)}{\leq} \left|\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right)\left(\nabla^\top J(\omega_t)(\omega_{t+1} - \omega_t)\right)\right]\right| \\
&\qquad + \left|\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right)(\omega_{t+1} - \omega_t)^\top \frac{\nabla^2 J(\hat{\omega}_t)}{2} (\omega_{t+1} - \omega_t)\right]\right| \\
&= \beta_t \left|\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right) \nabla^\top J(\omega_t)(\phi_t^\top(s_t, a_t)\theta_t \phi_t(s_t, a_t))\right]\right| \\
&\qquad + \beta_t^2 \left|\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right)\left(\phi_t^\top(s_t, a_t)\theta_t\phi_t(s_t, a_t)\right)^\top \frac{\nabla^2 J(\hat{\omega}_t)}{2}\left(\phi_t^\top(s_t, a_t)\theta_t\phi_t(s_t, a_t)\right)\right]\right| \\
&\overset{(b)}{\leq} \frac{\beta_t}{2}\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right)^2\right] + \frac{C_\phi^4 B^2}{2}\beta_t \mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + R_{\max} L_J C_\phi^4 B^2 \beta_t^2,
\end{aligned} \tag{41}
$$

where (a) follows from the Lagrange's Mean Value Theorem and Lemma 1 for some $\hat{\omega}_t = \lambda\omega_t + (1-\lambda)\omega_{t+1}$ with $\lambda \in [0,1]$; (b) follows from $\langle a, b\rangle \leq \frac{\|a\|^2 + \|b\|^2}{2}$ and Lemma 1.

Combining Equation (37), Equation (39), Equation (40) and Equation (41) implies

$$
\begin{aligned}
\mathbb{E}\left[\left(\eta_{t+1} - J(\omega_{t+1})\right)^2\right] &\leq \left((1-\gamma_t)^2 + \beta_t\right)\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right)^2\right] + \beta_t C_\phi^4 B^2 \mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \\
&+ 2\gamma_t \left(R_{\max}^2 C_\pi C_\phi^2 B \sum_{j=t-k}^{t-1}\sum_{i=j}^{t-1} \beta_i + R_{\max}^2 m\rho^k + R_{\max}^2 \sum_{j=t-k}^{t-1}\gamma_j + R_{\max} C_J C_\phi^2 B \beta_t \right) \\
&+ R_{\max}^2 \gamma_t^2 + C_J^2 C_\phi^4 B^2 \beta_t^2 + 2R_{\max} C_J C_\phi^2 B \gamma_t \sum_{j=t-k}^{t-1} \beta_j + 2R_{\max} L_J C_\phi^4 B^2 \beta_t^2,
\end{aligned} \tag{42}
$$

which completes the proof. □

B.3. Tracking Error Analysis of AC

In this section, we bound the tracking error $\|\theta_t - \theta_t^*\|_2$. Recall that we write $\theta_t^* = \theta_{\omega_t}^*$ and θ_ω^* is the solution to Equation (4), i.e.,

$$
\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega^\top(s,a)\left(\mathcal{T}_{\pi_\omega}^{(k)}\left(\phi_\omega^\top(s,a)\theta_\omega^*\right) - \phi_\omega^\top(s,a)\theta_\omega^*\right)\right] = \mathbf{0}. \tag{43}
$$

We then present a recursive bound on the tracking error in the following lemma.

Lemma 10. *Set the step sizes such that*

$$
\frac{\bar{\lambda}_{\min}\alpha_t}{2} \geq \frac{BC_\phi^3\left(C_\phi L_\pi + 2L_\phi\right) + \lambda_{\min}(L_J + 2C_\phi) + 2\lambda_{\min}^2}{\lambda_{\min}^2}\beta_t, \tag{44}
$$

then it holds that

$$\mathbb{E}\left[\left\|\theta_{t+1}-\theta_{t+1}^*\right\|_2^2\right] \leq \left(1-\frac{\bar{\lambda}_{\min}\alpha_t}{2}\right)\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right] + \frac{L_J\lambda_{\min}+BC_\phi^3\left(C_\phi L_\pi+2L_\phi\right)}{\lambda_{\min}^2}\beta_t\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \frac{(k+1)^2C_\phi^2}{\bar{\lambda}_{\min}}\alpha_t\mathbb{E}\left[(\eta_t-J(\omega_t))^2\right] + G_t^\theta, \tag{45}$$

where

$$G_t^\theta = (k+1)^2U_\delta^2C_\phi^2\alpha_t^2 + \frac{4(k+1)BC_\phi^3U_\delta L_J}{\lambda_{\min}}\beta_t\sum_{j=t-k}^{t-1}\alpha_t + 2(k+1)C_\phi^3BU_\delta C_\Theta\alpha_t\beta_t + \left(2BC_\phi L_J(C_\phi L_\pi+2L_\phi)+2BL_\Theta\lambda_{\min}+C_\Theta\lambda_{\min}\right)\frac{4B^2C_\phi^4}{\lambda_{\min}^2}\beta_t\sum_{j=t-k}^{t}\beta_j + 2C_\phi^4B^2C_\Theta^2\beta_t^2 + \frac{4BC_\phi^2L_J\left(3C_{gap}+\lambda_{\min}\right)}{\lambda_{\min}^2}m\rho^k\beta_t + \frac{4(k+1)C_\phi U_\delta C_{gap}m\rho^k}{\lambda_{\min}}\alpha_t + \left(\frac{2}{\beta_t}+8\right)\left(\frac{C_{gap}m\rho^k}{\lambda_{\min}}\right)^2 + \frac{4B^3C_\pi C_\phi^4L_J}{\lambda_{\min}}\beta_t\sum_{j=t-k}^{t-1}\sum_{i=j}^{t-1}\beta_i + 2\alpha_tG_t^\delta. \tag{46}$$

Proof. From Algorithm 1, it holds that

$$\begin{aligned}\left\|\theta_{t+1}-\theta_{t+1}^*\right\|_2^2 &= \left\|\Pi_B\left(\theta_t+\alpha_t\delta_tz_t\right)-\theta_{t+1}^*\right\|_2^2\\ &\overset{(a)}{\leq} \left\|\theta_t+\alpha_t\delta_tz_t-\theta_{t+1}^*\right\|_2^2\\ &= \left\|\theta_t+\alpha_t\delta_tz_t-\theta_t^*+\theta_t^*-\theta_{t+1}^*\right\|_2^2\\ &= \|\theta_t-\theta_t^*\|_2^2+\alpha_t^2\|\delta_tz_t\|_2^2+\left\|\theta_t^*-\theta_{t+1}^*\right\|_2^2+2\alpha_t\left\langle\theta_t-\theta_t^*,\delta_tz_t\right\rangle\\ &\quad+2\alpha_t\left\langle\delta_tz_t,\theta_t^*-\theta_{t+1}^*\right\rangle+2\left\langle\theta_t-\theta_t^*,\theta_t^*-\theta_{t+1}^*\right\rangle,\end{aligned} \tag{47}$$

where (a) follows from the fact $\|\Pi_B(x)-y\|_2\leq\|x-y\|_2$ when $\|y\|_2\leq B$ and $\left\|\theta_{t+1}^*\right\|_2\leq B$.
Taking expectations on both sides further implies that

$$\begin{aligned}\mathbb{E}\left[\left\|\theta_{t+1}-\theta_{t+1}^*\right\|_2^2\right] &\leq \mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\alpha_t^2\mathbb{E}\left[\|\delta_tz_t\|_2^2\right]+\mathbb{E}\left[\left\|\theta_t^*-\theta_{t+1}^*\right\|_2^2\right]\\ &\quad+2\alpha_t\underbrace{\mathbb{E}\left[\left\langle\theta_t-\theta_t^*,\delta_tz_t\right\rangle\right]}_{\text{I}}+2\alpha_t\underbrace{\mathbb{E}\left[\left\langle\delta_tz_t,\theta_t^*-\theta_{t+1}^*\right\rangle\right]}_{\text{II}}+2\underbrace{\mathbb{E}\left[\left\langle\theta_t-\theta_t^*,\theta_t^*-\theta_{t+1}^*\right\rangle\right]}_{\text{III}}.\end{aligned} \tag{48}$$

Firstly, we can get that

$$\alpha_t^2\mathbb{E}\left[\|\delta_tz_t\|_2^2\right]\leq\alpha_t^2(k+1)^2C_\phi^2U_\delta^2. \tag{49}$$

The term $\left\|\theta_t^*-\theta_{t+1}^*\right\|_2$ can be bounded as follows:

$$\begin{aligned}\left\|\theta_t^*-\theta_{t+1}^*\right\|_2 &= \left\|\bar{\theta}_t^*-\bar{\theta}_{t+1}^*+\theta_t^*-\bar{\theta}_t^*-\theta_{t+1}^*+\bar{\theta}_{t+1}^*\right\|_2\\ &\leq \left\|\bar{\theta}_t^*-\bar{\theta}_{t+1}^*\right\|_2+\left\|\theta_t^*-\bar{\theta}_t^*\right\|_2+\left\|\theta_{t+1}^*-\bar{\theta}_{t+1}^*\right\|_2\\ &\overset{(a)}{\leq} \left\|\bar{\theta}_t^*-\bar{\theta}_{t+1}^*\right\|_2+\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}+\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\\ &\overset{(b)}{\leq} C_\Theta\|\omega_t-\omega_{t+1}\|_2+\frac{2C_{\text{gap}}m\rho^k}{\lambda_{\min}}\end{aligned}$$

$$
\begin{aligned}
&= \beta_t C_\Theta \left\|\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)\right\|_2 + \frac{2C_{\text{gap}} m\rho^k}{\lambda_{\min}} \\
&\leq C_\Theta C_\phi^2 B\beta_t + \frac{2C_{\text{gap}} m\rho^k}{\lambda_{\min}},
\end{aligned} \tag{50}
$$

where (a) follows from Proposition 4; and (b) follows from Lemma 2. Equation (50) further implies that

$$
\mathbb{E}\left[\left\|\theta_t^* - \theta_{t+1}^*\right\|_2^2\right] \leq 2C_\Theta^2 C_\phi^4 B^2\beta_t^2 + \frac{8C_{\text{gap}}^2 m^2\rho^{2k}}{\lambda_{\min}^2}. \tag{51}
$$

By Lemma 5, we can bound term I in Equation (48) as follows:

$$
\mathbb{E}\left[\langle\theta_t - \theta_t^*, \delta_t z_t\rangle\right] \leq -\frac{\bar{\lambda}_{\min}}{2}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{(k+1)^2 C_\phi^2}{2\bar{\lambda}_{\min}}\mathbb{E}\left[\left(J(\omega_t) - \eta_t\right)^2\right] + G_t^\delta. \tag{52}
$$

For term II in Equation (48), we have that

$$
\begin{aligned}
\mathbb{E}\left[\left\langle\delta_t z_t, \theta_t^* - \theta_{t+1}^*\right\rangle\right] &\leq \mathbb{E}\left[\|\delta_t z_t\|_2 \left\|\theta_t^* - \theta_{t+1}^*\right\|_2\right] \\
&\leq \mathbb{E}\left[\|\delta_t z_t\|_2 \left(\left\|\bar{\theta}_t^* - \bar{\theta}_{t+1}^*\right\|_2 + \left\|\theta_t^* - \bar{\theta}_t^*\right\|_2 + \left\|\theta_{t+1}^* - \bar{\theta}_{t+1}^*\right\|_2\right)\right] \\
&\leq (k+1)C_\phi U_\delta\left(C_\Theta C_\phi^2 B\beta_t + \frac{2C_{\text{gap}} m\rho^k}{\lambda_{\min}}\right) \\
&= (k+1)C_\phi^3 B U_\delta C_\Theta\beta_t + \frac{2(k+1)C_\phi U_\delta C_{\text{gap}} m\rho^k}{\lambda_{\min}},
\end{aligned} \tag{53}
$$

where the last inequality is from Lemma 2 and Proposition 4.

To convenience, we rewrite $F_t = F_{\omega_t}$. To bound term III in Equation (48), note that

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle\theta_t - \theta_t^*, \theta_t^* - \theta_{t+1}^*\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, \bar{\theta}_t^* - \bar{\theta}_{t+1}^*\right\rangle\right] + \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, \theta_t^* - \bar{\theta}_t^*\right\rangle\right] + \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, \theta_{t+1}^* - \bar{\theta}_{t+1}^*\right\rangle\right] \\
&\leq \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, F_t^{-1}\nabla J(\omega_t) - F_{t+1}^{-1}\nabla J(\omega_{t+1})\right\rangle\right] + \frac{\beta_t}{2}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{1}{2\beta_t}\mathbb{E}\left[\left\|\theta_t^* - \bar{\theta}_t^*\right\|_2^2\right] \\
&\quad + \frac{\beta_t}{2}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{1}{2\beta_t}\mathbb{E}\left[\left\|\theta_{t+1}^* - \bar{\theta}_{t+1}^*\right\|_2^2\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, F_{t+1}^{-1}\left(\nabla J(\omega_t) - \nabla J(\omega_{t+1})\right)\right\rangle\right] + \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, \left(F_t^{-1} - F_{t+1}^{-1}\right)\nabla J(\omega_t)\right\rangle\right] \\
&\quad + \beta_t\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{1}{\beta_t}\left(\frac{C_{\text{gap}} m\rho^k}{\lambda_{\min}}\right)^2 \\
&\overset{(b)}{=} \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t)\left(\omega_t - \omega_{t+1}\right)\right\rangle\right] + \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, \left(F_t^{-1} - F_{t+1}^{-1}\right)\nabla J(\omega_t)\right\rangle\right] \\
&\quad + \beta_t\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{1}{\beta_t}\left(\frac{C_{\text{gap}} m\rho^k}{\lambda_{\min}}\right)^2 \\
&\overset{(c)}{\leq} \mathbb{E}\left[\left\langle\theta_t - \theta_t^*, F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t)\left(\omega_t - \omega_{t+1}\right)\right\rangle\right] + \frac{1}{\lambda_{\min}^2}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2 \|F_t - F_{t+1}\|_2 \|\nabla J(\omega_t)\|_2\right] \\
&\quad + \beta_t\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{1}{\beta_t}\left(\frac{C_{\text{gap}} m\rho^k}{\lambda_{\min}}\right)^2 \\
&\overset{(d)}{\leq} -\beta_t\mathbb{E}\left[\left\langle\theta_t - \theta_t^*, F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t)\left(\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)\right)\right\rangle\right] + \frac{BC_\phi^3\left(C_\phi L_\pi + 2L_\phi\right)}{2\lambda_{\min}^2}\beta_t\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] \\
&\quad + \frac{BC_\phi^3\left(C_\phi L_\pi + 2L_\phi\right)}{2\lambda_{\min}^2}\beta_t\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \beta_t\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{1}{\beta_t}\left(\frac{C_{\text{gap}} m\rho^k}{\lambda_{\min}}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
&= \underbrace{-\beta_t \mathbb{E}\left[\left\langle \theta_t - \theta_t^*, F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t)\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\bar{\theta}_t^*\phi_t(s,a)\right]\right\rangle\right]}_{(i)} \\
&\quad - \underbrace{\beta_t \mathbb{E}\left[\left\langle \theta_t - \theta_t^*, F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t)\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)(\theta_t^* - \bar{\theta}_t^*)\phi_t(s,a)\right]\right\rangle\right]}_{(ii)} \\
&\quad - \underbrace{\beta_t \mathbb{E}\left[\left\langle \theta_t - \theta_t^*, F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t)\mathbb{E}_{D_t}\left[\phi_t^\top(s,a)(\theta_t - \theta_t^*)\phi_t(s,a)\right]\right\rangle\right]}_{(iii)} \\
&\quad - \underbrace{\beta_t \mathbb{E}\left[\left\langle \theta_t - \theta_t^*, F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t)\left(\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t) - \mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right)\right\rangle\right]}_{(iv)} \\
&\quad + \frac{BC_\phi^3\left(C_\phi L_\pi + 2L_\phi\right)}{2\lambda_{\min}^2}\beta_t \mathbb{E}\left[\left\|\nabla J(\omega_t)\right\|_2^2\right] + \frac{BC_\phi^3\left(C_\phi L_\pi + 2L_\phi\right) + 2\lambda_{\min}^2}{2\lambda_{\min}^2}\beta_t \mathbb{E}\left[\left\|\theta_t - \theta_t^*\right\|_2^2\right] \\
&\quad + \frac{1}{\beta_t}\left(\frac{C_{\text{gap}} m\rho^k}{\lambda_{\min}}\right)^2. \qquad (54)
\end{aligned}
$$

where (a) follows from Proposition 4, (b) follows from Lagrange's Mean Value for some $\lambda \in [0,1]$, such that $\hat{\omega}_t = \lambda\omega_t + (1-\lambda)\omega_{t+1}$, (c) follows from the facts that for positive definite matrices X and Y,

$$
\begin{aligned}
\left\|X^{-1} - Y^{-1}\right\|_2 &\le \left\|X^{-1}\left(X - Y\right)Y^{-1}\right\|_2 \\
&\le \left\|X^{-1}\right\|_2 \left\|X - Y\right\|_2 \left\|Y^{-1}\right\|_2, \qquad (55)
\end{aligned}
$$

and (d) is from that

$$
\begin{aligned}
\left\|F_{t+1} - F_t\right\|_2 &= \left\|\mathbb{E}_{D_{t+1}}\left[\phi_{t+1}(s,a)\phi_{t+1}^\top(s,a)\right] - \mathbb{E}_{D_t}\left[\phi_t(s,a)\phi_t^\top(s,a)\right]\right\|_2 \\
&\le \left\|\mathbb{E}_{D_{t+1}}\left[\phi_{t+1}(s,a)\phi_{t+1}^\top(s,a)\right] - \mathbb{E}_{D_t}\left[\phi_{t+1}(s,a)\phi_{t+1}^\top(s,a)\right]\right\|_2 \\
&\qquad + \left\|\mathbb{E}_{D_t}\left[\phi_{t+1}(s,a)\phi_{t+1}^\top(s,a)\right] - \mathbb{E}_{D_t}\left[\phi_t(s,a)\phi_t^\top(s,a)\right]\right\|_2 \\
&\le C_\phi^2 \left\|D_{t+1} - D_t\right\|_{\mathcal{TV}} + \mathbb{E}\left[\left(\left\|\phi_t(s,a)\right\|_2 + \left\|\phi_{t+1}(s,a)\right\|_2\right)\left\|\phi_t(s,a) - \phi_{t+1}(s,a)\right\|_2\right] \\
&\overset{(a)}{\le} C_\phi^2 L_\pi \left\|\omega_{t+1} - \omega_t\right\|_2 + 2C_\phi L_\phi \left\|\omega_{t+1} - \omega_t\right\|_2 \\
&= \left(C_\phi^2 L_\pi + 2C_\phi L_\phi\right)\beta_t \left\|\phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t)\right\|_2 \\
&\le BC_\phi^3\left(C_\phi L_\pi + 2L_\phi\right)\beta_t, \qquad (56)
\end{aligned}
$$

where (a) follows from (Zou et al., 2019) and Theorem 1 in (Li et al., 2024), where $L_\pi = \frac{1}{2}C_\pi\left(1 + \lceil \log m^{-1}\rceil + \frac{1}{1-\rho}\right)$, and

$$
\left\|D_{t+1} - D_t\right\|_{\mathcal{TV}} \le L_\pi \left\|\omega_{t+1} - \omega_t\right\|_2. \qquad (57)
$$

We then consider the term (i),

$$
\begin{aligned}
(i) &\le \beta_t \mathbb{E}\left[\left\|\theta_t - \theta_t^*\right\|_2 \left\|F_{t+1}^{-1}\right\|_2 \left\|\nabla^2 J(\hat{\omega}_t)\right\|_2 \left\|\nabla J(\omega_t)\right\|_2\right] \\
&\le \frac{L_J\beta_t}{\lambda_{\min}}\left(\frac{1}{2}\mathbb{E}\left[\left\|\theta_t - \theta_t^*\right\|_2^2\right] + \frac{1}{2}\mathbb{E}\left[\left\|\nabla J(\omega_t)\right\|_2^2\right]\right), \qquad (58)
\end{aligned}
$$

where the last inequality follows from Lemma 1.

Next, we consider the term (ii),

$$
\begin{aligned}
(ii) &\le \beta_t C_\phi^2 \mathbb{E}\left[\left\|\theta_t - \theta_t^*\right\|_2 \left\|F_{t+1}^{-1}\right\|_2 \left\|\nabla^2 J(\hat{\omega}_t)\right\|_2 \left\|\theta_t^* - \bar{\theta}_t^*\right\|_2\right] \\
&\le \frac{2BC_\phi^2 C_{\text{gap}} L_J m\rho^k\beta_t}{\lambda_{\min}^2}, \qquad (59)
\end{aligned}
$$

where the last inequality follows from Lemma 1 and Proposition 4.

Then, consider the term (iii),

$$
\begin{aligned}
(iii) &\leq \beta_t C_\phi^2 \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2 \left\|F_{t+1}^{-1}\right\|_2 \left\|\nabla^2 J(\hat{\omega}_t)\right\|_2 \|\theta_t - \theta_t^*\|_2\right] \\
&\leq \frac{C_\phi^2 L_J \beta_t}{\lambda_{\min}} \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right].
\end{aligned}
\tag{60}
$$

Consider the term (iv),

$$
\begin{aligned}
&(iv) = -\beta_t \mathbb{E}\left[\left\langle (F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t))^\top (\theta_t - \theta_t^*), \phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t) - \mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\rangle\right] \\
&= -\beta_t \mathbb{E}[(F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t))^\top (\theta_t - \theta_t^*) - (F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top (\theta_{t-k} - \theta_{t-k}^*), \\
&\qquad \phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t) - \mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\rangle] \\
&\quad - \beta_t \mathbb{E}\left[\left\langle (F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top (\theta_{t-k} - \theta_{t-k}^*), \phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t) - \mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\right\rangle\right] \\
&\leq 2C_\phi^2 B\beta_t \mathbb{E}\left[\left\|(F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t))^\top (\theta_t - \theta_t^*) - (F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top (\theta_{t-k} - \theta_{t-k}^*)\right\|_2\right] \\
&\quad + \beta_t \mathbb{E}[\langle (F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top (\theta_{t-k} - \theta_{t-k}^*), \phi_t^\top(s_t,a_t)\theta_t\phi_t(s_t,a_t) - \mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\theta_t\phi_t(s,a)\right]\rangle] \\
&\overset{(a)}{\leq} 2C_\phi^2 B\beta_t \mathbb{E}\left[\left\|(F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t))^\top (\theta_t - \theta_t^*) - (F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top (\theta_{t-k} - \theta_{t-k}^*)\right\|_2\right] \\
&\quad + \beta_t \mathbb{E}\left[\frac{2B^2 C_\phi^2 L_J}{\lambda_{\min}} \|\mathbb{P}(s_t, a_t|\mathcal{F}_{t-k}) - D_t\|_{\mathcal{TV}}\right] \\
&\overset{(b)}{\leq} 2C_\phi^2 B\beta_t \mathbb{E}\left[\left\|(F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t))^\top (\theta_t - \theta_t^*) - (F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top (\theta_{t-k} - \theta_{t-k}^*)\right\|_2\right] \\
&\quad + \frac{2B^2 C_\phi^2 L_J \beta_t}{\lambda_{\min}} \left(C_\pi \sum_{j=t-k}^{t-1} \mathbb{E}\left[\|\omega_t - \omega_j\|_2\right] + m\rho^k\right) \\
&\leq 2C_\phi^2 B\beta_t \mathbb{E}\left[\left\|(F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t))^\top (\theta_t - \theta_t^*) - (F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top (\theta_{t-k} - \theta_{t-k}^*)\right\|_2\right] \\
&\quad + \frac{2B^2 C_\phi^2 L_J \beta_t}{\lambda_{\min}} \left(C_\pi C_\phi^2 B \sum_{j=t-k}^{t-1} \sum_{i=j}^{t-1} \beta_i + m\rho^k\right),
\end{aligned}
\tag{61}
$$

where (a) follows from Lemma 1 and (b) follows from Lemma 4.

Consider the first term in Equation (61) and we have that

$$
\begin{aligned}
&\left\|(F_{t+1}^{-1}\nabla^2 J(\hat{\omega}_t))^\top (\theta_t - \theta_t^*) - (F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top (\theta_{t-k} - \theta_{t-k}^*)\right\|_2 \\
&\leq \left\|\left(F_{t+1}^{-1} - F_{t-k}^{-1}\right)\nabla^2 J(\hat{\omega}_t))^\top (\theta_t - \theta_t^*)\right\|_2 + \left\|F_{t-k}^{-1}\left(\nabla^2 J(\hat{\omega}_t) - \nabla^2 J(\hat{\omega}_{t-k-1})\right)(\theta_t - \theta_t^*)\right\|_2 \\
&\qquad + \left\|(F_{t-k}^{-1}\nabla^2 J(\hat{\omega}_{t-k-1}))^\top \left(\theta_t - \theta_{t-k} - \theta_t^* + \theta_{t-k}^*\right)\right\|_2 \\
&\leq 2BL_J \left\|F_{t+1}^{-1} - F_{t-k}^{-1}\right\|_2 + \frac{2B}{\lambda_{\min}} \left\|\nabla^2 J(\hat{\omega}_t) - \nabla^2 J(\hat{\omega}_{t-k-1})\right\|_2 \\
&\qquad + \frac{L_J}{\lambda_{\min}} \left(\|\theta_t - \theta_{t-k}\|_2 + \left\|\theta_t^* - \theta_{t-k}^*\right\|_2\right) \\
&\overset{(a)}{\leq} \frac{2B^2 L_J C_\phi^3 (C_\phi L_\pi + 2L_\phi)}{\lambda_{\min}^2} \sum_{j=t-k}^{t} \beta_j + \frac{2B}{\lambda_{\min}} \left\|\nabla^2 J(\hat{\omega}_t) - \nabla^2 J(\hat{\omega}_{t-k-1})\right\|_2 \\
&\qquad + \frac{L_J}{\lambda_{\min}} \left(\|\theta_t - \theta_{t-k}\|_2 + \left\|\bar{\theta}_t^* - \bar{\theta}_{t-k}^*\right\|_2 + \left\|\bar{\theta}_t^* - \theta_t^*\right\|_2 + \left\|\theta_{t-k}^* - \bar{\theta}_{t-k}^*\right\|_2\right) \\
&\overset{(b)}{\leq} \frac{2B^2 L_J C_\phi^3 (C_\phi L_\pi + 2L_\phi)}{\lambda_{\min}^2} \sum_{j=t-k}^{t} \beta_j + \frac{2BL_\Theta}{\lambda_{\min}} \|\hat{\omega}_t - \hat{\omega}_{t-k-1}\|_2
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{L_J}{\lambda_{\min}} \left(\left\| \theta_t - \theta_{t-k} \right\|_2 + C_\Theta \left\| \omega_t - \omega_{t-k} \right\|_2 + \frac{2C_{\text{gap}} m \rho^k}{\lambda_{\min}} \right) \\
\leq\ & \frac{2B^2 L_J C_\phi^3 \left(C_\phi L_\pi + 2L_\phi \right)}{\lambda_{\min}^2} \sum_{j=t-k}^{t} \beta_j + \frac{2B^2 C_\phi^2 L_\Theta}{\lambda_{\min}} \sum_{j=t-k}^{t} \beta_t \\
&+ \frac{L_J}{\lambda_{\min}} \left((k+1) C_\phi U_\delta \sum_{j=t-k}^{t-1} \alpha_j + B C_\phi^2 C_\Theta \sum_{j=t-k}^{t-1} \beta_j + \frac{2C_{\text{gap}} m \rho^k}{\lambda_{\min}} \right) \\
=\ & \frac{(k+1) C_\phi U_\delta L_J}{\lambda_{\min}} \sum_{j=t-k}^{t-1} \alpha_j + \frac{2C_{\text{gap}} L_J m \rho^k}{\lambda_{\min}^2} \\
&+ \frac{B C_\phi^2}{\lambda_{\min}^2} \left(2B C_\phi L_J (C_\phi L_\pi + 2L_\phi) + 2B L_\Theta \lambda_{\min} + C_\Theta \lambda_{\min} \right) \sum_{j=t-k}^{t} \beta_j,
\end{aligned} \tag{62}
$$

where (a) follows from Equation (55) and Equation (56) and (b) follows from Lemma 7.

Therefore, the term (iv) can be bounded as:

$$
\begin{aligned}
(iv) \leq\ & \frac{2(k+1) B C_\phi^3 U_\delta L_J}{\lambda_{\min}} \beta_t \sum_{j=t-k}^{t-1} \alpha_j \\
&+ \frac{2B C_\phi^2 L_J \left(2C_{\text{gap}} + B\lambda_{\min} \right)}{\lambda_{\min}^2} m \rho^k \beta_t + \frac{2B^3 C_\pi C_\phi^4 L_J}{\lambda_{\min}} \beta_t \sum_{j=t-k}^{t-1} \sum_{i=j}^{t-1} \beta_i \\
&+ \frac{2B^2 C_\phi^4}{\lambda_{\min}^2} \left(2B C_\phi L_J (C_\phi L_\pi + 2L_\phi) + 2B L_\Theta \lambda_{\min} + C_\Theta \lambda_{\min} \right) \beta_t \sum_{j=t-k}^{t} \beta_j.
\end{aligned} \tag{63}
$$

Combining the above bounds on terms (i), (ii), (iii) and (iv), we have that

$$
\begin{aligned}
&\mathbb{E}\left[\langle \theta_t - \theta_t^*, \theta_t^* - \theta_{t+1}^* \rangle \right] \leq \frac{B C_\phi^3 \left(C_\phi L_\pi + 2L_\phi \right) + \lambda_{\min}(L_J + 2C_\phi) + 2\lambda_{\min}^2}{2\lambda_{\min}^2} \beta_t \mathbb{E}\left[\left\| \theta_t - \theta_t^* \right\|_2^2 \right] \\
&+ \frac{L_J \lambda_{\min} + B C_\phi^3 \left(C_\phi L_\pi + 2L_\phi \right)}{2\lambda_{\min}^2} \beta_t \mathbb{E}\left[\left\| \nabla J(\omega_t) \right\|_2^2 \right] + \frac{2B C_\phi^2 L_J \left(3C_{\text{gap}} + \lambda_{\min} \right)}{\lambda_{\min}^2} m \rho^k \beta_t \\
&+ \frac{2(k+1) B C_\phi^3 U_\delta L_J}{\lambda_{\min}} \beta_t \sum_{j=t-k}^{t-1} \alpha_j + \frac{2B^3 C_\pi C_\phi^4 L_J}{\lambda_{\min}} \beta_t \sum_{j=t-k}^{t-1} \sum_{i=j}^{t-1} \beta_i + \frac{1}{\beta_t} \left(\frac{C_{\text{gap}} m \rho^k}{\lambda_{\min}} \right)^2 \\
&+ \frac{2B^2 C_\phi^4}{\lambda_{\min}^2} \left(2B C_\phi L_J (C_\phi L_\pi + 2L_\phi) + 2B L_\Theta \lambda_{\min} + C_\Theta \lambda_{\min} \right) \beta_t \sum_{j=t-k}^{t} \beta_j.
\end{aligned} \tag{64}
$$

This bounds term III in Equation (48).

Plugging bounds on terms I, II, III in Equation (48) further implies that

$$
\begin{aligned}
E\left[\left\| \theta_{t+1} - \theta_{t+1}^* \right\|_2^2 \right] \leq\ & \left(1 - \bar{\lambda}_{\min} \alpha_t + \frac{B C_\phi^3 \left(C_\phi L_\pi + 2L_\phi \right) + \lambda_{\min}(L_J + 2C_\phi) + 2\lambda_{\min}^2}{\lambda_{\min}^2} \beta_t \right) \mathbb{E}\left[\left\| \theta_t - \theta_t^* \right\|_2^2 \right] \\
&+ \frac{L_J \lambda_{\min} + B C_\phi^3 \left(C_\phi L_\pi + 2L_\phi \right)}{\lambda_{\min}^2} \beta_t \mathbb{E}\left[\left\| \nabla J(\omega_t) \right\|_2^2 \right] + \frac{(k+1)^2 C_\phi^2}{\bar{\lambda}_{\min}} \alpha_t \mathbb{E}\left[\left(\eta_t - J(\omega_t) \right)^2 \right] \\
&+ (k+1)^2 U_\delta^2 C_\phi^2 \alpha_t^2 + \frac{4(k+1) B C_\phi^3 U_\delta L_J}{\lambda_{\min}} \beta_t \sum_{j=t-k}^{t-1} \alpha_j + 2(k+1) C_\phi^3 B U_\delta C_\Theta \alpha_t \beta_t
\end{aligned}
$$

$$
\begin{aligned}
&+ (2BC_\phi L_J(C_\phi L_\pi + 2L_\phi) + 2BL_\Theta\lambda_{\min} + C_\Theta\lambda_{\min}) \frac{4B^2C_\phi^4}{\lambda_{\min}^2}\beta_t \sum_{j=t-k}^{t} \beta_j + 2C_\phi^4 B^2 C_\Theta^2 \beta_t^2 \\
&+ \frac{4BC_\phi^2 L_J (3C_{\text{gap}} + \lambda_{\min})}{\lambda_{\min}^2} m\rho^k \beta_t + \frac{4(k+1)C_\phi U_\delta C_{\text{gap}} m\rho^k}{\lambda_{\min}}\alpha_t + \left(\frac{2}{\beta_t} + 8\right)\left(\frac{C_{\text{gap}} m\rho^k}{\lambda_{\min}}\right)^2 \\
&+ \frac{4B^3 C_\pi C_\phi^4 L_J}{\lambda_{\min}}\beta_t \sum_{j=t-k}^{t-1}\sum_{i=j}^{t-1}\beta_i + 2\alpha_t G_t^\delta. \qquad (65)
\end{aligned}
$$

Set the stepsize such that $\frac{\bar{\lambda}_{\min}\alpha_t}{2} \geq \frac{BC_\phi^3(C_\phi L_\pi + 2L_\phi) + \lambda_{\min}(L_J + 2C_\phi) + 2\lambda_{\min}^2}{\lambda_{\min}^2}\beta_t$, and the proof ends. □

B.4. Sample Complexity of AC

We first present our proof of Proposition 5. For convenience, we set $\alpha_t = \alpha$, $\beta_t = \beta$, $\gamma_t = \gamma$ in following.

Proposition 8. *(Restatement of Proposition 5) With the constant step sizes such that $k = \mathcal{O}(\log T)$ and $\gamma \geq \alpha \geq \beta$, the tracking error of the AC algorithm in Algorithm 1 can be bounded as follows:*

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t^* - \theta_t\|_2^2\right] &\leq \left(\frac{c_\alpha\beta}{\alpha} + \frac{c_\eta\beta}{\gamma}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \mathcal{O}\left(\frac{1}{T\alpha}\right) + \mathcal{O}\left(\frac{\log^2 T}{T\gamma}\right) \\
&+ \mathcal{O}\left(\alpha\log^3 T\right) + \mathcal{O}\left(\beta\log^3 T\right) + \mathcal{O}\left(\gamma\log^3 T\right) + \mathcal{O}\left(\frac{\beta^2\log^2 T}{\alpha}\right) + \mathcal{O}\left(\frac{\beta^2\log^2 T}{\gamma}\right) \\
&+ \mathcal{O}\left(\frac{\beta}{\alpha}(m\rho^k)\right) + \mathcal{O}\left((m\rho^k)\log^2 T\right) + \mathcal{O}\left(\frac{(m\rho^k)^2}{\alpha\beta}\right), \qquad (66)
\end{aligned}
$$

where $c_\alpha = \frac{2L_J\lambda_{\min} + 2BC_\phi^3(C_\phi L_\pi + 2L_\phi)}{\lambda_{\min}^2\bar{\lambda}_{\min}}$ and $c_\eta = \frac{2B^2(k+1)^2C_\phi^6}{(\bar{\lambda}_{\min})^2}$.

Proof. Recall that in Lemma 10, we showed

$$
\begin{aligned}
\mathbb{E}\left[\|\theta_{t+1} - \theta_{t+1}^*\|_2^2\right] &\leq \left(1 - \frac{\bar{\lambda}_{\min}\alpha}{2}\right)\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{L_J\lambda_{\min} + BC_\phi^3(C_\phi L_\pi + 2L_\phi)}{\lambda_{\min}^2}\beta\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \\
&+ \frac{(k+1)^2C_\phi^2}{\bar{\lambda}_{\min}}\alpha\mathbb{E}\left[(\eta_t - J(\omega_t))^2\right] + G_t^\theta. \qquad (67)
\end{aligned}
$$

Apply this inequality recursively and we have that

$$
\begin{aligned}
&\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] \\
&\leq \left(1 - \frac{\bar{\lambda}_{\min}\alpha}{2}\right)^t\mathbb{E}\left[\|\theta_0 - \theta_0^*\|_2^2\right] + \frac{L_J\lambda_{\min} + BC_\phi^3(C_\phi L_\pi + 2L_\phi)}{\lambda_{\min}^2}\beta\sum_{j=0}^{t-1}\left(1 - \frac{\bar{\lambda}_{\min}\alpha}{2}\right)^{t-j-1}\mathbb{E}\left[\|\nabla J(\omega_j)\|_2^2\right] \\
&+ \frac{(k+1)^2C_\phi^2}{\bar{\lambda}_{\min}}\alpha\sum_{j=0}^{t-1}\left(1 - \frac{\bar{\lambda}_{\min}\alpha}{2}\right)^{t-j-1}\mathbb{E}\left[(J(\omega_j) - \eta_j)^2\right] + \sum_{j=0}^{t-1}\left(1 - \frac{\bar{\lambda}_{\min}\alpha}{2}\right)^{t-j-1}G_j^\theta \\
&= (1-q)^t\,\mathbb{E}\left[\|\theta_0 - \theta_0^*\|_2^2\right] + \frac{L_J\lambda_{\min} + BC_\phi^3(C_\phi L_\pi + 2L_\phi)}{\lambda_{\min}^2}\beta\sum_{j=0}^{t-1}(1-q)^{t-j-1}\,\mathbb{E}\left[\|\nabla J(\omega_j)\|_2^2\right] \\
&+ \frac{(k+1)^2C_\phi^2}{\bar{\lambda}_{\min}}\alpha\sum_{j=0}^{t-1}(1-q)^{t-j-1}\,\mathbb{E}\left[(J(\omega_j) - \eta_j)^2\right] + \sum_{j=0}^{t-1}(1-q)^{t-j-1}G_j^\theta, \qquad (68)
\end{aligned}
$$

where we let $q = \frac{\bar{\lambda}_{\min}\alpha}{2}$.

Summing the inequality in Equation (68) above w.r.t. t from 0 to $T-1$ further implies that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right] \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}(1-q)^t\,\mathbb{E}\left[\|\theta_0-\theta_0^*\|_2^2\right]+\frac{L_J\lambda_{\min}+BC_\phi^3\left(C_\phi L_\pi+2L_\phi\right)}{\lambda_{\min}^2}\frac{\beta}{T}\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}(1-q)^{t-j-1}\,\mathbb{E}\left[\|\nabla J(\omega_j)\|_2^2\right] \\
&\quad+\frac{(k+1)^2C_\phi^2}{\bar{\lambda}_{\min}}\alpha\frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}(1-q)^{t-j-1}\,\mathbb{E}\left[\left(J(\omega_j)-\eta_j\right)^2\right]+\frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}(1-q)^{t-j-1}\,G_j^\theta \\
&\leq \frac{4B^2}{Tq}+\frac{L_J\lambda_{\min}+BC_\phi^3\left(C_\phi L_\pi+2L_\phi\right)}{\lambda_{\min}^2}\frac{\beta}{Tq}\sum_{j=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_j)\|_2^2\right]+\frac{(k+1)^2C_\phi^2}{\bar{\lambda}_{\min}}\frac{\alpha}{Tq}\sum_{j=0}^{T-1}\mathbb{E}\left[\left(J(\omega_j)-\eta_j\right)^2\right] \\
&\quad+\frac{1}{Tq}\sum_{t=0}^{T-1}G_t^\theta, \qquad (69)
\end{aligned}$$

where the last inequality is from the double-sum trick: $\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}y^{t-j-1}X_j\leq(\sum_{t=0}^{T-1}X_t)(\sum_{t=0}^{T-1}y^t)\leq\frac{\sum_{t=0}^{T-1}X_t}{1-y}$ for $X_j\geq 0, j=0,1,2,...,T-1$ and $y\in(0,1)$.

Recall that we showed in Lemma 9 that

$$\mathbb{E}\left[\left(\eta_{t+1}-J(\omega_{t+1})\right)^2\right]\leq(1-\gamma)\,\mathbb{E}\left[\left(\eta_t-J(\omega_t)\right)^2\right]+C_\phi^4B^2\beta\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right]+G_t^\eta. \qquad (70)$$

Recursively applying this inequality implies that

$$\begin{aligned}
&\mathbb{E}\left[\left(\eta_t-J(\omega_t)\right)^2\right] \\
&\leq(1-\gamma)^t\left(\eta_0-J(\omega_0)\right)^2+C_\phi^4B^2\beta\sum_{j=0}^{t-1}(1-\gamma)^{t-j-1}\,\mathbb{E}\left[\|\nabla J(\omega_j)\|_2^2\right]+\sum_{j=0}^{t-1}(1-\gamma)^{t-j-1}\,G_j^\eta \\
&\leq R_{\max}^2\,(1-\gamma)^t+C_\phi^4B^2\beta\sum_{j=0}^{t-1}(1-\gamma)^{t-j-1}\,\mathbb{E}\left[\|\nabla J(\omega_j)\|_2^2\right]+\sum_{j=0}^{t-1}(1-\gamma)^{t-j-1}\,G_j^\eta. \qquad (71)
\end{aligned}$$

We then sum the above inequality w.r.t. t from 0 to $T-1$, and have that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\eta_t-J(\omega_t)\right)^2\right] \\
&\leq\frac{R_{\max}^2}{T}\sum_{t=0}^{T-1}(1-\gamma)^t+C_\phi^4B^2\frac{\beta}{T}\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}(1-\gamma)^{t-j-1}\,\mathbb{E}\left[\|\nabla J(\omega_j)\|_2^2\right]+\frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}(1-\gamma)^{t-j-1}\,G_j^\eta \\
&=\frac{R_{\max}^2}{T\gamma}+C_\phi^4B^2\frac{\beta}{T\gamma}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right]+\frac{1}{T\gamma}\sum_{t=0}^{T-1}G_t^\eta, \qquad (72)
\end{aligned}$$

where we use the double-sum trick below Equation (69) again.

Plugging Equation (72) in Equation (69) further implies that

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]&\overset{(a)}{\leq}\frac{8B^2}{\bar{\lambda}_{\min}T\alpha}+\frac{2L_J\lambda_{\min}+2BC_\phi^3\left(C_\phi L_\pi+2L_\phi\right)}{\lambda_{\min}^2\bar{\lambda}_{\min}}\frac{\beta}{T\alpha}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \\
&\quad+\frac{2(k+1)^2C_\phi^2}{(\bar{\lambda}_{\min})^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(J(\omega_t)-\eta_t\right)^2\right]+\frac{2}{\bar{\lambda}_{\min}T\alpha}\sum_{t=0}^{T-1}G_t^\theta
\end{aligned}$$

$$\leq \left(\frac{2L_J\lambda_{\min} + 2BC_\phi^3\left(C_\phi L_\pi + 2L_\phi\right)}{\lambda_{\min}^2\bar{\lambda}_{\min}}\frac{\beta}{\alpha} + \frac{2(k+1)^2B^2C_\phi^6\beta}{(\bar{\lambda}_{\min})^2\gamma}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right]$$
$$+ \frac{8B^2}{\bar{\lambda}_{\min}T\alpha} + \frac{2}{\bar{\lambda}_{\min}T\alpha}\sum_{t=0}^{T-1}G_t^\theta + \frac{2(k+1)^2C_\phi^2R_{\max}^2}{(\bar{\lambda}_{\min})^2T\gamma} + \frac{2(k+1)^2C_\phi^2}{(\bar{\lambda}_{\min})^2}\frac{1}{T\gamma}\sum_{t=0}^{T-1}G_t^\eta$$
$$= \left(\frac{c_\alpha\beta}{\alpha} + \frac{c_\eta\beta}{\gamma}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \frac{8B^2}{\bar{\lambda}_{\min}T\alpha} + \frac{2}{\bar{\lambda}_{\min}T\alpha}\sum_{t=0}^{T-1}G_t^\theta$$
$$+ \frac{2(k+1)^2C_\phi^2R_{\max}^2}{(\bar{\lambda}_{\min})^2T\gamma} + \frac{2(k+1)^2C_\phi^2}{(\bar{\lambda}_{\min})^2}\frac{1}{T\gamma}\sum_{t=0}^{T-1}G_t^\eta, \tag{73}$$

where (a) follows from that $q = \frac{\bar{\lambda}_{\min}\alpha}{2}$, $c_\alpha = \frac{2L_J\lambda_{\min}+2BC_\phi^3(C_\phi L_\pi+2L_\phi)}{\lambda_{\min}^2\bar{\lambda}_{\min}}$ and $c_\eta = \frac{2(k+1)^2B^2C_\phi^6}{(\bar{\lambda}_{\min})^2}$. By Equation (73), if we set the stepsize $\beta = \min\left\{\frac{\lambda_{\min}^2\bar{\lambda}_{\min}\alpha}{16C_\phi^4\left(L_J\lambda_{\min}+BC_\phi^3(C_\phi L_\pi+2L_\phi)\right)}, \frac{(\bar{\lambda}_{\min})^2\gamma}{16(k+1)^2C_\phi^{10}B^2}\right\}$, then we can get that $\left(\frac{c_\alpha\beta}{\alpha} + \frac{c_\eta\beta}{\gamma}\right) \leq \frac{1}{4C_\phi^4}$. This completes the proof of Proposition 5. □

We are now ready to prove Theorem 1.

Theorem 3. *(Restatement of Theorem 1) Consider the AC algorithm in Algorithm 1 with constant step sizes $\alpha_t = \alpha, \beta_t = \beta, \gamma_t = \gamma$. Then, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \leq 2C_\phi^4\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t^* - \theta_t\|_2^2\right] + \mathcal{O}\left(\frac{1}{T\beta}\right) + \mathcal{O}\left(\beta\log^2 T\right) + \mathcal{O}(m\rho^k). \tag{74}$$

If we further set $k = \left\lceil\frac{\log T}{1-\rho}\right\rceil, \gamma = \mathcal{O}(\frac{1}{\sqrt{T}}), \alpha = \mathcal{O}(\frac{1}{\sqrt{T}\log^2 T}), \beta = \mathcal{O}(\frac{1}{\sqrt{T}\log^2 T})$, then we have that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \leq \mathcal{O}\left(\frac{\log^3 T}{\sqrt{T}}\right). \tag{75}$$

Proof. Recall Equation (28), and we have that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \leq \frac{2\left(\mathbb{E}\left[J(\omega_T)\right] - J(\omega_0)\right)}{T\beta} + \frac{2C_\phi^4}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{2}{T\beta}\sum_{t=0}^{T-1}G_t^\omega. \tag{76}$$

Plug Equation (76) in Equation (73), and we have that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right]$$
$$\leq \frac{1}{4C_\phi^4}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \frac{8B^2}{\bar{\lambda}_{\min}T\alpha} + \frac{2}{\bar{\lambda}_{\min}T\alpha}\sum_{t=0}^{T-1}G_t^\theta + \frac{2(k+1)^2C_\phi^2R_{\max}^2}{(\bar{\lambda}_{\min})^2T\gamma} + \frac{2(k+1)^2C_\phi^2}{(\bar{\lambda}_{\min})^2}\frac{1}{T\gamma}\sum_{t=0}^{T-1}G_t^\eta$$
$$= \frac{1}{2C_\phi^4T\beta}\left(\mathbb{E}\left[J(\omega_T)\right] - J(\omega_0)\right) + \frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{8B^2}{\bar{\lambda}_{\min}T\alpha} + \frac{1}{2C_\phi^4T\beta}\sum_{t=0}^{T-1}G_t^\omega$$
$$+ \frac{2(k+1)^2C_\phi^2}{(\bar{\lambda}_{\min})^2}\left(\frac{R_{\max}^2}{T\gamma} + \frac{1}{T\gamma}\sum_{t=0}^{T-1}G_t^\eta\right) + \frac{2}{\bar{\lambda}_{\min}T\alpha}\sum_{t=0}^{T-1}G_t^\theta. \tag{77}$$

This further implies that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] \leq \frac{1}{C_\phi^4T\beta}\left(\mathbb{E}\left[J(\omega_T)\right] - J(\omega_0)\right)$$

$$+\frac{16B^2}{\bar{\lambda}_{\min}T\alpha}+\frac{1}{C_\phi^4 T\beta}\sum_{t=0}^{T-1}G_t^\omega+\frac{4(k+1)^2C_\phi^2}{(\bar{\lambda}_{\min})^2}\left(\frac{R_{\max}^2}{T\gamma}+\frac{1}{T\gamma}\sum_{t=0}^{T-1}G_t^\eta\right)+\frac{4}{\bar{\lambda}_{\min}T\alpha}\sum_{t=0}^{T-1}G_t^\theta. \tag{78}$$

Next, we choose the stepsizes to minimize the tracking error and the gradient norm. We choose $\gamma=\frac{1}{\sqrt{T}}$ and $k=\left\lceil\frac{\log T}{1-\rho}\right\rceil$. Consider $T\geq\frac{2C_\phi^2 dm}{\lambda_{\min}}$, then we can get that $\bar{\lambda}_{\min}\geq\frac{\lambda_{\min}}{2}$. We set α and β such that $\alpha=\frac{1}{(k+1)^2}\gamma$ and $\beta=\min\left\{\frac{\lambda_{\min}^3}{4\left(L_J\lambda_{\min}+BC_\phi^3(C_\phi L_\pi+2L_\phi)+2\lambda_{\min}^2\right)}\alpha,\frac{\lambda_{\min}^3}{32C_\phi^4\left(L_J\lambda_{\min}+BC_\phi^3(C_\phi L_\pi+2L_\phi)\right)}\alpha,\frac{\lambda_{\min}^2}{64(k+1)^2C_\phi^6B^2}\gamma\right\}$. It holds that

$$\gamma=\mathcal{O}\left(\frac{1}{\sqrt{T}}\right),\alpha=\mathcal{O}\left(\frac{1}{\sqrt{T}\log^2 T}\right),\beta=\mathcal{O}\left(\frac{1}{\sqrt{T}\log^2 T}\right),q=\mathcal{O}\left(\frac{1}{\sqrt{T}\log^2 T}\right). \tag{79}$$

According to Equation (29), Equation (35) and Equation (46), the orders of the following terms are as follows:

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}G_t^\omega=\mathcal{O}\left((m\rho^k)\beta+k^2\beta^2\right)=\mathcal{O}\left(\frac{1}{T\log^2 T}\right);\\&\frac{1}{T}\sum_{t=0}^{T-1}G_t^\eta=\mathcal{O}\left((m\rho^k)\gamma+\beta^2+k^2\beta\gamma+k\gamma^2\right)=\mathcal{O}\left(\frac{\log T}{T}\right);\\&\frac{1}{T}\sum_{t=0}^{T-1}G_t^\theta=\mathcal{O}\left(k(m\rho^k)\alpha+(m\rho^k)\beta+\frac{(m\rho^k)^2}{\beta}+k^3\alpha^2+k^3\alpha\beta+k^2\beta^2\right)=\mathcal{O}\left(\frac{1}{T\log T}\right).\end{aligned} \tag{80}$$

Then Equation (78) can be bounded as follows:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]=\mathcal{O}\left(\frac{\log^3 T}{\sqrt{T}}\right). \tag{81}$$

Plugging Equation (78) in Equation (76) implies that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right]\leq&\frac{32C_\phi^4B^2}{\bar{\lambda}_{\min}T\alpha}+\frac{8(k+1)^2C_\phi^6}{(\bar{\lambda}_{\min})^2}\left(\frac{R_{\max}^2}{T\gamma}+\frac{1}{T\gamma}\sum_{t=0}^{T-1}G_t^\eta\right)+\frac{8C_\phi^4}{\bar{\lambda}_{\min}T\alpha}\sum_{t=0}^{T-1}G_t^\theta\\&+\frac{4}{T\beta}\left(\mathbb{E}\left[J(\omega_T)\right]-J(\omega_0)\right)+\frac{4}{T\beta}\sum_{t=0}^{T-1}G_t^\omega.\end{aligned} \tag{82}$$

Plugging in the step-sizes above, we have that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right]=\mathcal{O}\left(\frac{\log^3 T}{\sqrt{T}}\right), \tag{83}$$

which completes the proof. □

C. NAC Sample Complexity Analysis

In this section, we provide the sample complexity analysis for NAC.

C.1. Bound on Gradient Norm in NAC

Recall that in Algorithm 1, NAC updates the policy parameter as follows: $\omega_{t+1}-\omega_t=\beta_t\theta_t$, which directly implies that

$$\|\omega_{t+1}-\omega_t\|_2\leq\beta_t\|\theta_t\|_2\leq B\beta_t. \tag{84}$$

Consider the largest eigenvalue of the matrix $F_t = \mathbb{E}_{D_t}\left[\phi_t(s,a)\phi_t^\top(s,a)\right]$. Note that for any vector $x \in \mathbb{R}^d$

$$\left\|F_t x\right\|_2 = \left\|\mathbb{E}_{D_t}\left[\phi_t(s,a)\phi_t^\top(s,a)x\right]\right\|_2 \leq \left\|\phi_t(s,a)\right\|_2 \left\|\phi_t(s,a)\right\|_2 \left\|x\right\|_2 \leq C_\phi^2 \left\|x\right\|_2. \tag{85}$$

Thus, $\lambda_{\max}(F_t) \leq C_\phi^2$. Then, by Lemma 1, we can show that

$$\begin{aligned}
&J(\omega_{t+1}) \geq J(\omega_t) + \langle \nabla J(\omega_t), \omega_{t+1} - \omega_t \rangle - \frac{L_J}{2}\left\|\omega_{t+1} - \omega_t\right\|_2^2 \\
&\geq J(\omega_t) + \beta_t \langle \nabla J(\omega_t), \theta_t \rangle - \frac{\beta_t^2 L_J}{2}\left\|\theta_t\right\|_2^2 \\
&\geq J(\omega_t) + \beta_t \left\langle \nabla J(\omega_t), \bar{\theta}_t^* \right\rangle + \beta_t \left\langle \nabla J(\omega_t), \theta_t - \theta_t^* \right\rangle + \beta_t \left\langle \nabla J(\omega_t), \theta_t^* - \bar{\theta}_t^* \right\rangle - \frac{L_J B^2 \beta_t^2}{2} \\
&\overset{(a)}{\geq} J(\omega_t) + \beta_t \left\langle \nabla J(\omega_t), \left(\mathbb{E}_{D_t}\left[\phi_t(s,a)\phi_t^\top(s,a)\right]\right)^{-1} \nabla J(\omega_t) \right\rangle - \frac{\beta_t}{4C_\phi^2}\left\|\nabla J(\omega_t)\right\|_2^2 - C_\phi^2 \beta_t \left\|\theta_t - \theta_t^*\right\|_2^2 \\
&\quad - \frac{\beta_t}{4C_\phi^2}\left\|\nabla J(\omega_t)\right\|_2^2 - C_\phi^2 \beta_t \left\|\theta_t^* - \bar{\theta}_t^*\right\|_2^2 - \frac{L_J B^2 \beta_t^2}{2} \\
&\overset{(b)}{\geq} J(\omega_t) + \frac{\beta_t}{C_\phi^2}\left\|\nabla J(\omega_t)\right\|_2^2 - \frac{\beta_t}{2C_\phi^2}\left\|\nabla J(\omega_t)\right\|_2^2 - C_\phi^2 \beta_t \left\|\theta_t - \theta_t^*\right\|_2^2 - \frac{C_{\text{gap}} C_\phi^2 m \rho^k \beta_t}{\lambda_{\min}} - \frac{L_J B^2 \beta_t^2}{2} \\
&= J(\omega_t) + \frac{\beta_t}{2C_\phi^2}\left\|\nabla J(\omega_t)\right\|_2^2 - C_\phi^2 \beta_t \left\|\theta_t - \theta_t^*\right\|_2^2 - \frac{C_{\text{gap}} C_\phi^2 m \rho^k \beta_t}{\lambda_{\min}} - \frac{L_J B^2 \beta_t^2}{2},
\end{aligned} \tag{86}$$

where (a) follows from that $\langle \nabla J(\omega_t), \theta_t - \theta_t^* \rangle \geq -\frac{1}{4C_\phi^2}\left\|\nabla J(\omega_t)\right\|_2^2 - C_\phi^2 \left\|\theta_t - \theta_t^*\right\|_2^2$ and (b) follows from that $\lambda_{\max}(F_t) \leq C_\phi^2$, $\left\langle \nabla J(\omega_t), \left(\mathbb{E}_{D_t}\left[\phi_t(s,a)\phi_t^\top(s,a)\right]\right)^{-1} \nabla J(\omega_t) \right\rangle = \nabla J(\omega_t)^\top (F_t)^{-1} \nabla J(\omega_t) \geq \frac{1}{C_\phi^2}\left\|\nabla J(\omega_t)\right\|_2^2$, and Proposition 4. Taking the expectation on both sides of Equation (86), we have that

$$\mathbb{E}\left[\left\|\nabla J(\omega_t)\right\|_2^2\right] \leq 2C_\phi^2 \frac{\mathbb{E}\left[J(\omega_{t+1})\right] - \mathbb{E}\left[J(\omega_t)\right]}{\beta_t} + 2C_\phi^4 \mathbb{E}\left[\left\|\theta_t - \theta_t^*\right\|_2^2\right] + \frac{2C_\phi^4 C_{\text{gap}} m \rho^k}{\lambda_{\min}} + L_J B^2 C_\phi^2 \beta_t. \tag{87}$$

C.2. Bound on $|\eta_t - J(\omega_t)|$ in NAC

In this section, we bound the term $\eta_t - J(\omega_t)$ for the NAC algorithm.

Lemma 11. *If we denote*

$$\begin{aligned}
\widetilde{G}_t^\eta = 2\gamma_t &\left(B R_{\max}^2 C_\pi \sum_{j=t-k}^{t-1} \sum_{i=j}^{t-1} \beta_i + R_{\max}^2 m \rho^k + R_{\max}^2 \sum_{j=t-k}^{t-1} \gamma_j + B C_J R_{\max} \sum_{j=t-k}^{t-1} \beta_j \right) \\
&+ R_{\max}^2 \gamma_t^2 + C_J^2 B^2 \beta_t^2 + 2 B C_J R_{\max} \beta_t \gamma_t + R_{\max} L_J B^2 \beta_t^2,
\end{aligned} \tag{88}$$

and set $\gamma_t - \gamma_t^2 \geq \beta_t$, then it holds that

$$\mathbb{E}\left[\left(\eta_{t+1} - J(\omega_{t+1})\right)^2\right] \leq (1-\gamma_t)\,\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right)^2\right] + \beta_t B^2 \mathbb{E}\left[\left\|\nabla J(\omega_t)\right\|_2^2\right] + \widetilde{G}_t^\eta. \tag{89}$$

Proof. Similar to the AC analysis in Equation (37) in Appendix B.2, we have that

$$\begin{aligned}
\left(\eta_{t+1} - J(\omega_{t+1})\right)^2 &= \left((1-\gamma_t)\left(\eta_t - J(\omega_t)\right) + \gamma_t (R_t - J(\omega_t)) + J(\omega_t) - J(\omega_{t+1})\right)^2 \\
&= (1-\gamma_t)^2 \left(\eta_t - J(\omega_t)\right)^2 + \gamma_t^2 \left(R_t - J(\omega_t)\right)^2 + \left(J(\omega_t) - J(\omega_{t+1})\right)^2 \\
&\quad + 2\gamma_t \underbrace{\left(R_t - J(\omega_t)\right)\left(J(\omega_t) - J(\omega_{t+1})\right)}_{\text{I}} + 2\gamma_t (1-\gamma_t) \underbrace{\left(\eta_t - J(\omega_t)\right)\left(R_t - J(\omega_t)\right)}_{\text{II}}
\end{aligned}$$

$$+ 2\left(1-\gamma_t\right)\underbrace{\left(\eta_t - J(\omega_t)\right)\left(J(\omega_t) - J(\omega_{t+1})\right)}_{\text{III}}. \tag{90}$$

The term $|J(\omega_t) - J(\omega_{t+1})|$ can be bounded using its Lipschitz smoothness in Lemma 1:

$$|J(\omega_t) - J(\omega_{t+1})| \leq C_J \left\|\omega_t - \omega_{t+1}\right\|_2 \leq C_J B \beta_t. \tag{91}$$

Term I in Equation (90) can be bounded as follows:

$$|\mathbb{E}\left[(R_t - J(\omega_t))(J(\omega_t) - J(\omega_{t+1}))\right]| \leq \mathbb{E}\left[|R_t - J(\omega_t)|\,|J(\omega_t) - J(\omega_{t+1})|\right] \leq BC_J R_{\max}\beta_t. \tag{92}$$

Term II in Equation (90) can be bounded as follows:

$$\begin{aligned}
&|\mathbb{E}\left[(\eta_t - J(\omega_t))(R_t - J(\omega_t))\right]| \\
&\leq |\mathbb{E}\left[(\eta_{t-k} - J(\omega_{t-k}))(R_t - J(\omega_t))\right]| + |\mathbb{E}\left[(\eta_t - \eta_{t-k} - J(\omega_t) + J(\omega_{t-k}))(R_t - J(\omega_t))\right]| \\
&\leq |\mathbb{E}\left[\mathbb{E}\left[(\eta_{t-k} - J(\omega_{t-k}))(R_t - J(\omega_t))|\mathcal{F}_{t-k}\right]\right]| + \mathbb{E}\left[|\eta_t - \eta_{t-k} - J(\omega_t) + J(\omega_{t-k})|\,|R_t - J(\omega_t)|\right] \\
&\overset{(a)}{\leq} \left|\mathbb{E}\left[\mathbb{E}\left[(\eta_{t-k} - J(\omega_{t-k}))R_t|\mathcal{F}_{t-k}\right] - \mathbb{E}_{(s,a)\sim D_t}\left[(\eta_{t-k} - J(\omega_{t-k}))R(s,a)|\mathcal{F}_{t-k}\right]\right]\right| \\
&\quad + \mathbb{E}\left[|\eta_t - \eta_{t-k} - J(\omega_t) + J(\omega_{t-k})|\,|R_t - J(\omega_t)|\right] \\
&\overset{(b)}{\leq} R_{\max}^2 \mathbb{E}\left[\|\mathbb{P}\left((s_t,a_t)|\mathcal{F}_{t-k}\right), D_t\|_{\mathcal{TV}}\right] + R_{\max}\mathbb{E}\left[|\eta_t - \eta_{t-k}| + |J(\omega_t) - J(\omega_{t-k})|\right] \\
&\overset{(c)}{\leq} R_{\max}^2\left(C_\pi \sum_{j=t-k}^{t-1} \mathbb{E}\left[\|\omega_t - \omega_j\|_2\right] + m\rho^k\right) + R_{\max}\left(R_{\max}\sum_{j=t-k}^{t-1}\gamma_j + C_J \mathbb{E}\left[\|\omega_t - \omega_{t-k}\|_2\right]\right) \\
&\leq BR_{\max}^2 C_\pi \sum_{j=t-k}^{t-1}\sum_{i=j}^{t-1}\beta_i + R_{\max}^2 m\rho^k + R_{\max}^2\sum_{j=t-k}^{t-1}\gamma_j + R_{\max}C_J B\sum_{j=t-k}^{t-1}\beta_j,
\end{aligned} \tag{93}$$

where (a) follows from $\mathbb{E}_{(s,a)\sim D_t}\left[R(s,a) - J(\omega_t)|\mathcal{F}_t\right] = 0$, (b) follows from $0 \leq \eta_t \leq R_{\max}$, $0 \leq J(\omega_t) \leq R_{\max}$, $0 \leq R_t \leq R_{\max}$, and (c) follows from Lemma 4.

We then bound term III as follows:

$$\begin{aligned}
&|\mathbb{E}\left[(\eta_t - J(\omega_t))\left(J(\omega_t) - J(\omega_{t+1})\right)\right]| \\
&\quad= \left|\mathbb{E}\left[(\eta_t - J(\omega_t))\left(-\nabla^\top J(\omega_t)(\omega_{t+1} - \omega_t) + (\omega_{t+1} - \omega_t)^\top \frac{\nabla^2 J(\hat{\omega}_t)}{2}(\omega_{t+1} - \omega_t)\right)\right]\right| \\
&\quad\leq \left|\mathbb{E}\left[(\eta_t - J(\omega_t))\nabla^\top J(\omega_t)(\omega_{t+1} - \omega_t)\right]\right| \\
&\qquad + \left|\mathbb{E}\left[(\eta_t - J(\omega_t))(\omega_{t+1} - \omega_t)^\top \frac{\nabla^2 J(\hat{\omega}_t)}{2}(\omega_{t+1} - \omega_t)\right]\right| \\
&\quad= \beta_t\left|\mathbb{E}\left[(\eta_t - J(\omega_t))\nabla^\top J(\omega_t)\theta_t\right]\right| + \beta_t^2\left|\mathbb{E}\left[|\eta_t - J(\omega_t)|\frac{\left\|\nabla^2 J(\hat{\omega}_t)\right\|_2}{2}\|\theta_t\|_2^2\right]\right| \\
&\quad\leq \frac{\beta_t}{2}\mathbb{E}\left[(\eta_t - J(\omega_t))^2\right] + \frac{B^2\beta_t}{2}\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \frac{R_{\max}L_J B^2}{2}\beta_t^2,
\end{aligned} \tag{94}$$

where the first equation is from the Lagrange's Mean Value theorem for some $\hat{\omega}_t = \lambda\omega_t + (1-\lambda)\omega_{t+1}$, $\lambda \in (0,1)$.

Plug Equation (92), Equation (93), and Equation (94) in Equation (90), and we have that

$$\begin{aligned}
\mathbb{E}\left[(\eta_{t+1} - J(\omega_{t+1}))^2\right] &\leq \left((1-\gamma_t)^2 + \beta_t\right)\mathbb{E}\left[(\eta_t - J(\omega_t))^2\right] + B^2\beta_t\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \\
&+ 2\gamma_t\left(BR_{\max}^2 C_\pi \sum_{j=t-k}^{t-1}\sum_{i=j}^{t-1}\beta_i + R_{\max}^2 m\rho^k + R_{\max}^2\sum_{j=t-k}^{t-1}\gamma_j + R_{\max}C_J B\sum_{j=t-k}^{t-1}\beta_j\right)
\end{aligned}$$

$$+ R_{\max}^2\gamma_t^2 + C_J^2B^2\beta_t^2 + 2BC_JR_{\max}\beta_t\gamma_t + R_{\max}L_JB^2\beta_t^2. \tag{95}$$

In Equation (95), we use the fact that $1-\gamma_t \le 1$. This completes the proof. □

C.3. Tracking Error Analysis of NAC

In this section, we bound the tracking error $\theta_t - \theta_t^*$ for NAC. Define

$$\widetilde{G}_t^\theta = \left(8 + \frac{2}{\beta_t} + \beta_t C_\Theta\right)\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2 + (k+1)^2U_\delta^2C_\phi^2\alpha_t^2 + \frac{4(k+1)C_\phi U_\delta C_{\text{gap}}m\rho^k}{\lambda_{\min}}\alpha_t$$
$$+ 2(k+1)C_\phi BU_\delta C_\Theta\alpha_t\beta_t + 2\alpha_t G_t^\delta + 2C_\Theta^2B^2\beta_t^2. \tag{96}$$

Lemma 12. *If we set the step size satisfies that $\beta_t \le \frac{\bar{\lambda}_{\min}}{4(2C_\Theta+1)}\alpha_t$, then it holds that*

$$\mathbb{E}\left[\left\|\theta_{t+1}-\theta_{t+1}^*\right\|_2^2\right] \le \left(1 - \frac{\bar{\lambda}_{\min}\alpha_t}{2}\right)\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right]$$
$$+ \frac{C_\Theta}{\lambda_{\min}^2}\beta_t\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \frac{(k+1)^2C_\phi^2}{\bar{\lambda}_{\min}}\alpha_t\mathbb{E}\left[(J(\omega_t)-\eta_t)^2\right] + \widetilde{G}_t^\theta. \tag{97}$$

Proof. From the update rule of Algorithm 1, we have that

$$\begin{aligned}
\left\|\theta_{t+1}-\theta_{t+1}^*\right\|_2^2 &= \left\|\Pi_B\left(\theta_t + \alpha_t\delta_t z_t\right) - \theta_{t+1}^*\right\|_2^2 \\
&\overset{(a)}{\le} \left\|\theta_t + \alpha_t\delta_t z_t - \theta_{t+1}^*\right\|_2^2 \\
&\le \left\|\theta_t + \alpha_t\delta_t z_t - \theta_t^* + \theta_t^* - \theta_{t+1}^*\right\|_2^2 \\
&= \|\theta_t - \theta_t^*\|_2^2 + \alpha_t^2\|\delta_t z_t\|_2^2 + \left\|\theta_t^* - \theta_{t+1}^*\right\|_2^2 + 2\alpha_t\langle\theta_t-\theta_t^*, \delta_t z_t\rangle \\
&\quad + 2\alpha_t\left\langle\delta_t z_t, \theta_t^* - \theta_{t+1}^*\right\rangle + 2\left\langle\theta_t - \theta_t^*, \theta_t^* - \theta_{t+1}^*\right\rangle,
\end{aligned} \tag{98}$$

where (a) follows from the fact $\|\Pi_B(x)-y\|_2 \le \|x-y\|_2$ when $\|y\|_2 \le B$ and $\left\|\theta_{t+1}^*\right\|_2 \le B$.

Taking expectations on both sides of Equation (98), we have that

$$\mathbb{E}\left[\left\|\theta_{t+1}-\theta_{t+1}^*\right\|_2^2\right] \le \mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right] + \alpha_t^2\mathbb{E}\left[\|\delta_t z_t\|_2^2\right] + \mathbb{E}\left[\left\|\theta_t^*-\theta_{t+1}^*\right\|_2^2\right]$$
$$+ 2\alpha_t\underbrace{\mathbb{E}\left[\langle\theta_t-\theta_t^*, \delta_t z_t\rangle\right]}_{\text{I}} + 2\alpha_t\underbrace{\mathbb{E}\left[\left\langle\delta_t z_t, \theta_t^*-\theta_{t+1}^*\right\rangle\right]}_{\text{II}} + 2\underbrace{\mathbb{E}\left[\left\langle\theta_t-\theta_t^*, \theta_t^*-\theta_{t+1}^*\right\rangle\right]}_{\text{III}}. \tag{99}$$

For the term $\left\|\theta_t^* - \theta_{t+1}^*\right\|_2$, we have that

$$\begin{aligned}
\left\|\theta_t^*-\theta_{t+1}^*\right\|_2 &= \left\|\bar{\theta}_t^* - \bar{\theta}_{t+1}^* + \theta_t^* - \bar{\theta}_t^* - \theta_{t+1}^* + \bar{\theta}_{t+1}^*\right\|_2 \\
&\le \left\|\bar{\theta}_t^*-\bar{\theta}_{t+1}^*\right\|_2 + \left\|\theta_t^*-\bar{\theta}_t^*\right\|_2 + \left\|\theta_{t+1}^*-\bar{\theta}_{t+1}^*\right\|_2 \\
&\overset{(a)}{\le} C_\Theta\|\omega_t-\omega_{t+1}\|_2 + \frac{2C_{\text{gap}}m\rho^k}{\lambda_{\min}} \\
&= \beta_t C_\Theta\|\theta_t\|_2 + \frac{2C_{\text{gap}}m\rho^k}{\lambda_{\min}} \\
&\le C_\Theta B\beta_t + \frac{2C_{\text{gap}}m\rho^k}{\lambda_{\min}},
\end{aligned} \tag{100}$$

where (a) follows from Lemma 2 and Proposition 4. Hence we have that

$$\mathbb{E}\left[\left\|\theta_t^*-\theta_{t+1}^*\right\|_2^2\right] \le 2C_\Theta^2B^2\beta_t^2 + \frac{8C_{\text{gap}}^2m^2\rho^{2k}}{\lambda_{\min}^2}. \tag{101}$$

By Lemma 5, term I in Equation (99) can be bounded as

$$\mathbb{E}\left[\langle\theta_t-\theta_t^*,\delta_t z_t\rangle\right]\le-\frac{\bar{\lambda}_{\min}}{2}\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{(k+1)^2C_\phi^2}{2\bar{\lambda}_{\min}}\mathbb{E}\left[\left(J(\omega_t)-\eta_t\right)^2\right]+G_t^\delta. \tag{102}$$

For term II in Equation (99), we have that

$$\begin{aligned}
\mathbb{E}\left[\left\langle\delta_t z_t,\theta_t^*-\theta_{t+1}^*\right\rangle\right] &\le \mathbb{E}\left[\|\delta_t z_t\|_2\left\|\theta_t^*-\theta_{t+1}^*\right\|_2\right]\\
&\le (k+1)C_\phi U_\delta\mathbb{E}\left[\left\|\bar{\theta}_{t+1}^*-\bar{\theta}_t^*\right\|_2+\left\|\theta_t^*-\bar{\theta}_t^*\right\|_2+\left\|\theta_{t+1}^*-\bar{\theta}_{t+1}^*\right\|_2\right]\\
&\overset{(a)}{\le} (k+1)C_\phi U_\delta C_\Theta\mathbb{E}\left[\|\omega_{t+1}-\omega_t\|_2\right]+\frac{2(k+1)C_\phi U_\delta C_{\text{gap}}m\rho^k}{\lambda_{\min}}\\
&\le (k+1)C_\phi BU_\delta C_\Theta\beta_t+\frac{2(k+1)C_\phi U_\delta C_{\text{gap}}m\rho^k}{\lambda_{\min}},
\end{aligned} \tag{103}$$

where (a) follows from Lemma 2 and Proposition 4.

For term III in Equation (99), we have that

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\theta_t-\theta_t^*,\theta_t^*-\theta_{t+1}^*\right\rangle\right]\\
&=\mathbb{E}\left[\left\langle\theta_t-\theta_t^*,\bar{\theta}_t^*-\bar{\theta}_{t+1}^*\right\rangle\right]+\mathbb{E}\left[\left\langle\theta_t-\theta_t^*,\theta_t^*-\bar{\theta}_t^*\right\rangle\right]+\mathbb{E}\left[\left\langle\theta_t-\theta_t^*,\bar{\theta}_{t+1}^*-\theta_{t+1}^*\right\rangle\right]\\
&\le\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2\left\|\bar{\theta}_t^*-\bar{\theta}_{t+1}^*\right\|_2\right]+\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2\left\|\bar{\theta}_t^*-\theta_t^*\right\|_2\right]+\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2\left\|\bar{\theta}_{t+1}^*-\theta_{t+1}^*\right\|_2\right]\\
&\overset{(a)}{\le} C_\Theta\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2\|\omega_{t+1}-\omega_t\|_2\right]+\frac{\beta_t}{2}\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{2\beta_t}\mathbb{E}\left[\left\|\theta_t^*-\bar{\theta}_t^*\right\|_2^2\right]\\
&\quad+\frac{\beta_t}{2}\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{2\beta_t}\mathbb{E}\left[\left\|\theta_{t+1}^*-\bar{\theta}_{t+1}^*\right\|_2^2\right]\\
&\overset{(b)}{\le} C_\Theta\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2\|\omega_{t+1}-\omega_t\|_2\right]+\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{\beta_t}\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2\\
&=C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2\|\theta_t\|_2\right]+\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{\beta_t}\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2\\
&\le C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2\left\|\bar{\theta}_t^*\right\|_2\right]+C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2\left\|\theta_t^*-\bar{\theta}_t^*\right\|_2\right]\\
&\quad+\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{\beta_t}\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2\\
&\le\frac{1}{2}C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{2}C_\Theta\beta_t\mathbb{E}\left[\left\|\bar{\theta}_t^*\right\|_2^2\right]+C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{2}C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]\\
&\quad+\frac{1}{2}C_\Theta\beta_t\mathbb{E}\left[\left\|\theta_t^*-\bar{\theta}_t^*\right\|_2^2\right]+\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{\beta_t}\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2\\
&\overset{(c)}{\le}\frac{1}{2}C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{2}C_\Theta\beta_t\mathbb{E}\left[\left\|(F_t)^{-1}\nabla J(\omega_t)\right\|_2^2\right]+C_\Theta\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]\\
&\quad+\frac{C_\Theta}{2}\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{C_\Theta\beta_t}{2}\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2+\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{1}{\beta_t}\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2\\
&\overset{(d)}{\le}\left(2C_\Theta+1\right)\beta_t\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{C_\Theta}{2\lambda_{\min}^2}\beta_t\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right]+\left(\frac{\beta_t C_\Theta}{2}+\frac{1}{\beta_t}\right)\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2,
\end{aligned} \tag{104}$$

where (a) follows from Lemma 2; (b) follows from Proposition 4; (c) follows from Proposition 1, Proposition 2, and Proposition 4; (d) follows from that $\left\|F_t^{-1}\right\|_2\le\frac{1}{\lambda_{\min}}$.

Plugging the above bounds in Equation (99), we have that

$$\mathbb{E}\left[\left\|\theta_{t+1}-\theta_{t+1}^*\right\|_2^2\right]$$

$$
\begin{aligned}
&\leq \left(1-\bar{\lambda}_{\min}\alpha_t + (4C_\Theta+2)\beta_t\right)\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right] + \frac{C_\Theta}{\lambda_{\min}^2}\beta_t\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] \\
&+ \frac{(k+1)^2C_\phi^2}{\bar{\lambda}_{\min}}\alpha_t\mathbb{E}\left[(J(\omega_t)-\eta_t)^2\right] + \left(8+\frac{2}{\beta_t}+\beta_t C_\Theta\right)\left(\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right)^2 + (k+1)^2U_\delta^2C_\phi^2\alpha_t^2 \\
&+ \frac{4(k+1)C_\phi U_\delta C_{\text{gap}}m\rho^k}{\lambda_{\min}}\alpha_t + 2(k+1)C_\phi BU_\delta C_\Theta\alpha_t\beta_t + 2\alpha_t G_t^\delta + 2C_\Theta^2B^2\beta_t^2. \qquad (105)
\end{aligned}
$$

This hence completes the proof. □

C.4. Sample Complexity of NAC

In the following, we set $\alpha_t=\alpha$, $\beta_t=\beta$, and $\gamma_t=\gamma$ for any $t\geq 0$. We denote by $\text{KL}(\omega_t) = -\mathbb{E}_{D_{\pi^*}}\left[\log\frac{\pi_{\omega_t}(a|s)}{\pi^*(a|s)}\right]$. Recall that $\hat{t}=\lceil\frac{3\log T}{\lambda_{\min}\alpha}\rceil$. We first have the following lemma.

Lemma 13. *Denote by $\widetilde{T}=\left\lceil\frac{T}{\hat{t}\log T}\right\rceil\hat{t}\geq\frac{T}{\log T}$, then for any $t'\leq T-\widetilde{T}$, it holds that*

$$
\begin{aligned}
&\min_{t\leq T}\mathbb{E}\left[J(\pi^*)-J(\omega_t)\right] \\
&\leq \frac{KL(\omega_{t'})-KL(\omega_{t'+\widetilde{T}})}{\beta\widetilde{T}} + \frac{C_\phi C_{gap}m\rho^k}{\lambda_{\min}} + 2C_\infty\sqrt{\varepsilon_{actor}} + \sqrt{2}C_\phi\left(\frac{2eC_\phi^3C_M\hat{t}}{\widetilde{T}}\right) + \sqrt{2}C_\phi\sqrt{\frac{4B^2+R_{\max}^2}{\widetilde{T}}} \\
&+\sqrt{2}C_\phi\sqrt{\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\sum_{j=t-\hat{t}}^{t-1}(1-q)^{t-j-1}\left(\widetilde{G}_j^\omega+\widetilde{G}_j^\theta+\widetilde{G}_j^\eta\right)} + 2C_\phi\sqrt{C_\phi^2C_M\frac{e\hat{t}}{\widetilde{T}}\frac{KL(\omega_{t'})-KL(\omega_{t'+\widetilde{T}})}{\beta\widetilde{T}}} \\
&+2C_\phi\sqrt{\frac{e\hat{t}C_\phi^2C_M}{\widetilde{T}}\left(\frac{C_\phi C_{gap}m\rho^k}{\lambda_{\min}}+\frac{B^2L_\phi}{2}\beta\right)} + \frac{B^2L_\phi}{2}\beta + \frac{2R_{\max}}{\sqrt{T}}. \qquad (106)
\end{aligned}
$$

Proof. Recall that $\pi^*=\arg\max_\pi J(\pi)$ and $A^{\pi_t}=Q^{\pi_t}(s,a)-V^{\pi_t}(s)$. We first have that

$$
\begin{aligned}
&\text{KL}(\omega_t)-\text{KL}(\omega_{t+1}) = \mathbb{E}_{D_{\pi^*}}\left[\log\pi_{t+1}(a|s)-\log\pi_t(a|s)\right] \\
&\overset{(a)}{\geq} \mathbb{E}_{D_{\pi^*}}\left[\nabla^\top\log\pi_t(a|s)\right](\omega_{t+1}-\omega_t) - \frac{L_\phi}{2}\|\omega_{t+1}-\omega_t\|_2^2 \\
&= \beta\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)\theta_t\right] - \frac{L_\phi\beta^2}{2}\|\theta_t\|_2^2 \\
&\geq \beta\mathbb{E}_{D_{\pi^*}}\left[A^{\pi_t}(s,a)\right] + \beta\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)\theta_t^*-A^{\pi_t}(s,a)\right] + \beta\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)(\theta_t-\theta_t^*)\right] - \frac{B^2L_\phi\beta^2}{2} \\
&\overset{(b)}{=} \beta\left(J(\pi^*)-J(\pi_t)\right) + \beta\underbrace{\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)\theta_t^*-A^{\pi_t}(s,a)\right]}_{\text{I}} + \beta\underbrace{\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)(\theta_t-\theta_t^*)\right]}_{\text{II}} - \frac{B^2L_\phi\beta^2}{2}, \qquad (107)
\end{aligned}
$$

where (a) follows from that

$$
\begin{aligned}
&\|\nabla\mathbb{E}_{D_{\pi^*}}[\log\pi_\omega(a|s)] - \nabla\mathbb{E}_{D_{\pi^*}}[\log\pi_{\omega'}(a|s)]\|_2 \\
&= \|\mathbb{E}_{D_{\pi^*}}[\phi_\omega(s,a)-\phi_{\omega'}(s,a)]\|_2 \leq \mathbb{E}_{D_{\pi^*}}[L_\phi\|\omega-\omega'\|_2] = L_\phi\|\omega-\omega'\|_2, \qquad (108)
\end{aligned}
$$

and (b) follows from the fact that

$$
\begin{aligned}
\mathbb{E}_{D_{\pi^*}}\left[A^{\pi_t}(s,a)\right] &= \mathbb{E}_{D_{\pi^*}}\left[Q^{\pi_t}(s,a)-V^{\pi_t}(s)\right] \\
&= \mathbb{E}_{D_{\pi^*}}\left[R(s,a)-J(\pi_t)+V^{\pi_t}(s')-V^{\pi_t}(s)\right] \\
&= \mathbb{E}_{D_{\pi^*}}\left[R(s,a)\right]-J(\pi_t)+\mathbb{E}_{(s,a)\sim D_{\pi^*},s\sim P(\cdot|s,a)}\left[V^{\pi_t}(s')\right]-\mathbb{E}_{D_{\pi^*}}\left[V^{\pi_t}(s)\right] \\
&= J(\pi^*)-J(\pi_t)+\mathbb{E}_{D_{\pi^*}}\left[V^{\pi_t}(s')\right]-\mathbb{E}_{D_{\pi^*}}\left[V^{\pi_t}(s)\right]
\end{aligned}
$$

$$= J(\pi^*) - J(\pi_t). \tag{109}$$

To bound Term I, we first have that

$$\begin{aligned}
&\left|\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)\theta_t^* - A^{\pi_t}(s,a)\right]\right| \\
&\le \left|\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)\bar{\theta}_t^* - A^{\pi_t}(s,a)\right]\right| + \left|\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)(\bar{\theta}_t^* - \theta_t^*)\right]\right| \\
&\le \left\|\frac{D_{\pi^*}}{D_t}\right\|_\infty \mathbb{E}_{D_t}\left[\left|\phi_t^\top(s,a)\bar{\theta}_t^* - A^{\pi_t}(s,a)\right|\right] + C_\phi \left\|\bar{\theta}_t^* - \theta_t^*\right\|_2 \\
&\overset{(a)}{\le} \left\|\frac{D_{\pi^*}}{D_t}\right\|_\infty \sqrt{\mathbb{E}_{D_t}\left[\left(\phi_t^\top(s,a)\bar{\theta}_t^* - A^{\pi_t}(s,a)\right)^2\right]} + \frac{C_\phi C_{\text{gap}} m\rho^k}{\lambda_{\min}} \\
&\overset{(b)}{\le} \left\|\frac{D_{\pi^*}}{D_t}\right\|_\infty \sqrt{\varepsilon_{\text{actor}}} + \frac{C_\phi C_{\text{gap}} m\rho^k}{\lambda_{\min}},
\end{aligned} \tag{110}$$

where (a) follows from that proposition 4, and (b) follows from the definition of $\bar{\theta}_t^*$ in Equation (2), the definition of $\varepsilon_{\text{actor}}$ and the facts that

$$\mathbb{E}_{D_t}\left[\left(\phi_t^\top(s,a)\bar{\theta}_t^* - A^{\pi_t}(s,a)\right)^2\right] \le \varepsilon_{\text{actor}}. \tag{111}$$

For term II, we have that

$$\left|\mathbb{E}_{D_{\pi^*}}\left[\phi_t^\top(s,a)(\theta_t - \theta_t^*)\right]\right| \le C_\phi \left\|\theta_t - \theta_t^*\right\|_2. \tag{112}$$

Plug the two bounds on terms I and II in Equation (107), and we have that

$$\begin{aligned}
\mathbb{E}\left[\text{KL}(\omega_t)\right] - \mathbb{E}\left[\text{KL}(\omega_{t+1})\right] \ge& \beta\left(J(\pi^*) - \mathbb{E}\left[J(\omega_t)\right]\right) - \beta\left\|\frac{D_{\pi^*}}{D_t}\right\|_\infty \sqrt{\varepsilon_{\text{actor}}} - \beta\frac{C_\phi C_{\text{gap}} m\rho^k}{\lambda_{\min}} \\
&- \beta C_\phi \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2\right] - \frac{L_\phi}{2}\beta^2 B^2,
\end{aligned} \tag{113}$$

which implies

$$\begin{aligned}
\beta\left(J(\pi^*) - \mathbb{E}\left[J(\omega_t)\right]\right) \le& \mathbb{E}\left[\text{KL}(\omega_t)\right] - \mathbb{E}\left[\text{KL}(\omega_{t+1})\right] + \beta\left\|\frac{D_{\pi^*}}{D_t}\right\|_\infty \sqrt{\varepsilon_{\text{actor}}} + \beta\frac{C_\phi C_{\text{gap}} m\rho^k}{\lambda_{\min}} \\
&+ \beta C_\phi \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2\right] + \frac{B^2 L_\phi}{2}\beta^2.
\end{aligned} \tag{114}$$

Set $M_{t+1} = \mathbb{E}\left[\left\|\theta_{t+1} - \theta_{t+1}^*\right\|_2^2\right] + \mathbb{E}\left[\left(\eta_{t+1} - J(\omega_{t+1})\right)^2\right]$, and we now aim to bound M_t. Combine the bounds we obtained in Equation (89) and Equation (97), and we have that

$$\begin{aligned}
M_{t+1} \le& \left(1 - \frac{1}{2}\bar{\lambda}_{\min}\alpha\right)\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{C_\Theta}{\lambda_{\min}^2}\beta\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \frac{(k+1)^2 C_\phi^2}{\bar{\lambda}_{\min}}\alpha\mathbb{E}\left[\left(J(\omega_t) - \eta_t\right)^2\right] \\
&+ (1-\gamma)\,\mathbb{E}\left[\left(\eta_t - J(\omega_t)\right)^2\right] + B^2\beta\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \widetilde{G}_t^\theta + \widetilde{G}_t^\eta \\
=& \left(1 - \frac{1}{2}\bar{\lambda}_{\min}\alpha\right)\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{(k+1)^2 C_\phi^2}{\bar{\lambda}_{\min}}\alpha\mathbb{E}\left[\left(J(\omega_t) - \eta_t\right)^2\right] + (1-\gamma)\,\mathbb{E}\left[\left(J(\omega_t) - \eta_t\right)^2\right] \\
&+ \left(\frac{C_\Theta}{\lambda_{\min}^2} + B^2\right)\beta\mathbb{E}\left[\|\nabla J(\omega_t)\|_2^2\right] + \widetilde{G}_t^\theta + \widetilde{G}_t^\eta \\
\overset{(a)}{\le}& \left(1 - \frac{1}{2}\bar{\lambda}_{\min}\alpha\right)\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{(k+1)^2 C_\phi^2}{\bar{\lambda}_{\min}}\alpha\mathbb{E}\left[\left(J(\omega_t) - \eta_t\right)^2\right] + (1-\gamma)\,\mathbb{E}\left[\left(J(\omega_t) - \eta_t\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \left(\frac{C_\Theta}{\lambda_{\min}^2} + B^2\right) \beta \left(\frac{2C_\phi^2}{\beta} \left(\mathbb{E}\left[J(\omega_{t+1})\right] - \mathbb{E}\left[J(\omega_t)\right]\right) + 2C_\phi^4 \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right]\right) \\
&+ \left(\frac{C_\Theta}{\lambda_{\min}^2} + B^2\right) \left(2C_{\text{gap}} \frac{C_\phi^4 m \rho^k \beta}{\lambda_{\min}} + L_J C_\phi^2 B^2 \beta^2\right) + \widetilde{G}_t^\theta + \widetilde{G}_t^\eta \\
\leq& \left(1 - \frac{1}{2}\bar{\lambda}_{\min}\alpha + 2C_\phi^4 C_M \beta\right) \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + 2C_\phi^2 C_M \left(\mathbb{E}\left[J(\omega_{t+1})\right] - \mathbb{E}\left[J(\omega_t)\right]\right) \\
&+ \left(1 - \gamma + \frac{(k+1)^2 C_\phi^2}{\bar{\lambda}_{\min}}\alpha\right) \mathbb{E}\left[\left(J(\omega_t) - \eta_t\right)^2\right] + C_M \left(\frac{2C_{\text{gap}} C_\phi^4 m \rho^k \beta}{\lambda_{\min}} + L_J C_\phi^2 B^2 \beta^2\right) \\
&+ \widetilde{G}_t^\theta + \widetilde{G}_t^\eta, \qquad (115)
\end{aligned}
$$

where (a) is obtained by Equation (87), and $C_M = \frac{C_\Theta}{\lambda_{\min}^2} + B^2$. For convenience, we set

$$
\widetilde{G}_t^\omega = C_M \left(\frac{2C_{\text{gap}} C_\phi^4 m \rho^k \beta}{\lambda_{\min}} + L_J C_\phi^2 B^2 \beta^2\right). \qquad (116)
$$

We set $k = \lceil \frac{\log T}{1-\rho} \rceil$ and $T \geq \frac{2C_\phi^2 dm}{\lambda_{\min}}$ such that $\bar{\lambda}_{\min} \geq \frac{1}{2}\lambda_{\min}$. Furthermore, we set the step sizes such that $\frac{1}{6}\bar{\lambda}_{\min}\alpha \geq 2C_\phi^4 C_M \beta$ and $\frac{\gamma}{2} \geq \max\left\{\frac{1}{3}\bar{\lambda}_{\min}\alpha, \frac{(k+1)^2 C_\phi^2}{\bar{\lambda}_{\min}}\alpha\right\}$. Let $q = \frac{1}{3}\bar{\lambda}_{\min}\alpha$ and $\frac{\gamma}{2} \geq q$. Then the inequality in Equation (115) can be written as

$$
M_{t+1} \leq (1-q) M_t + 2C_\phi^2 C_M \left(\mathbb{E}\left[J(\omega_{t+1})\right] - \mathbb{E}\left[J(\omega_t)\right]\right) + \widetilde{G}_t^\omega + \widetilde{G}_t^\theta + \widetilde{G}_t^\eta. \qquad (117)
$$

Recall that $\hat{t} = \left\lceil \frac{1}{q}\log T \right\rceil = \left\lceil \frac{3\log T}{\bar{\lambda}_{\min}\alpha} \right\rceil$. For $t \geq 2k + \hat{t}$, we recursively apply Equation (117) for $\hat{t}$ times, and we have that

$$
\begin{aligned}
M_t \leq& (1-q)^{\hat{t}} M_{t-\hat{t}} + 2C_\phi^2 C_M \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\mathbb{E}\left[J(\omega_{j+1})\right] - \mathbb{E}\left[J(\omega_j)\right]\right) \\
&+ \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta\right) \\
\overset{(a)}{\leq}& \frac{4B^2 + R_{\max}^2}{T} + 2C_\phi^2 C_M \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\mathbb{E}\left[J(\omega_{j+1})\right] - \mathbb{E}\left[J(\omega_j)\right]\right) \\
&+ \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta\right), \qquad (118)
\end{aligned}
$$

where (a) follows from $(1-q)^{\hat{t}} \leq e^{-q\hat{t}} \leq e^{-\log T} = \frac{1}{T}$ and $M_i = \mathbb{E}\left[\|\theta_i - \theta_i^*\|_2^2\right] + \mathbb{E}\left[\left(\eta_i - J(\omega_i)\right)^2\right] \leq 4B^2 + R_{\max}^2$ for $i = 0, 1, ..., T$.

Denote the time length $\widetilde{T} = \left\lceil \frac{T}{\hat{t}\log T} \right\rceil \hat{t} \geq \frac{T}{\log T}$. For any $t' \leq T - \widetilde{T}$, together with Equation (114) we have that

$$
\begin{aligned}
&\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} J(\pi^*) - \mathbb{E}\left[J(\omega_t)\right] \\
\leq& \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta\widetilde{T}} + \frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + \frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \left\|\frac{D_{\pi^*}}{D_t}\right\|_\infty \sqrt{\varepsilon_{\text{actor}}}
\end{aligned}
$$

$$
\begin{aligned}
&+ C_\phi \frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2\right] + \frac{B^2 L_\phi}{2}\beta \\
\overset{(a)}{\leq}\ & \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}} + \frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + \frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \left\| \frac{D_{\pi^*}}{D_t} \right\|_\infty \sqrt{\varepsilon_{\text{actor}}} \\
&+ C_\phi \sqrt{\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right]} + \frac{B^2 L_\phi}{2}\beta \\
\overset{(b)}{\leq}\ & \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}} + \frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + C_\infty \sqrt{\varepsilon_{\text{actor}}} \\
&+ C_\phi \sqrt{\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} M_t} + \frac{B^2 L_\phi}{2}\beta, \qquad (119)
\end{aligned}
$$

where (a) follows from the rearrangement inequality and the fact for any random variable X, $\|\mathbb{E}[X]\|_2^2 \leq \mathbb{E}\left[\|X\|_2^2\right]$ and (b) follows from Assumption 3.

Moreover, for $2k + \hat{t} \leq t' \leq T - \widetilde{T}$, summing Equation (118) w.r.t. t from t' to $t' + \tilde{T} - 1$ implies

$$
\begin{aligned}
&\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} M_t \\
\leq\ & 2C_\phi^2 C_M \frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\mathbb{E}[J(\omega_{j+1})] - \mathbb{E}[J(\omega_j)]\right) \\
&+ \frac{4B^2 + R_{\max}^2}{\widetilde{T}} + \frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta\right) \\
\overset{(a)}{=}\ & 2C_\phi^2 C_M \frac{1}{\widetilde{T}} \sum_{i=0}^{\hat{t}-1} \sum_{t=\hat{t}-i}^{t'+\widetilde{T}-1} (1-q)^i \left(\mathbb{E}[J(\omega_{t-i})] - \mathbb{E}[J(\omega_{t-i-1})]\right) \\
&+ \frac{4B^2 + R_{\max}^2}{\widetilde{T}} + \frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta\right) \\
\overset{(b)}{=}\ & 2C_\phi^2 C_M \frac{1}{\widetilde{T}} \sum_{i=0}^{\hat{t}-1} (1-q)^i \left(\mathbb{E}\left[J(\omega_{t'+\widetilde{T}-i-1})\right] - \mathbb{E}[J(\omega_{t'-i-1})]\right) \\
&+ \frac{4B^2 + R_{\max}^2}{\widetilde{T}} + \frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta\right) \\
\overset{(c)}{\leq}\ & 2C_\phi^2 C_M \frac{1}{\widetilde{T}} \sum_{i=0}^{\hat{t}-1} (1-q)^i \left(J(\pi^*) - \mathbb{E}[J(\omega_{t'-i-1})]\right) \\
&+ \frac{4B^2 + R_{\max}^2}{\widetilde{T}} + \frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta\right) \\
\leq\ & 2C_\phi^2 C_M \frac{1}{\widetilde{T}} \sum_{i=0}^{\hat{t}-1} \left(J(\pi^*) - \mathbb{E}[J(\omega_{t'-i-1})]\right)
\end{aligned}
$$

$$+\frac{4B^2+R_{\max}^2}{\widetilde{T}}+\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\sum_{j=t-\hat{t}}^{t-1}(1-q)^{t-j-1}\left(\widetilde{G}_j^{\omega}+\widetilde{G}_j^{\theta}+\widetilde{G}_j^{\eta}\right), \tag{120}$$

where (a) follows from that we set $i=t-j-1$, (b) follows from the fact that $\sum_{t=\hat{t}-i}^{t'+\widetilde{T}-i-1}\left(\mathbb{E}\left[J(\omega_{t-i})\right]-\mathbb{E}\left[J(\omega_{t-i-1})\right]\right)=\mathbb{E}\left[J(\omega_{t'+\widetilde{T}-i-1})\right]-\mathbb{E}\left[J(\omega_{t'-i-1})\right]$ and (c) follows from $J(\pi^*)\geq J(\pi_{t'+\widetilde{T}-i-1})$.

Lemma 14. *We denote by* $X_{t'}=\frac{1}{\hat{t}}\sum_{t=t'-\hat{t}}^{t'-1}\left(J(\pi^*)-\mathbb{E}\left[J(\omega_t)\right]\right)$ *and* $Y_{t'}=\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\left(J(\pi^*)-\mathbb{E}\left[J(\omega_t)\right]\right)$. *We have that there must exist* $t'+\hat{t}\leq t''\leq t'+\widetilde{T}$ *s.t.* $X_{t''}\leq Y_{t'}$.

Proof. We prove the lemma by contraction. Assume that there does not exist $t'+\hat{t}\leq t''\leq t'+\widetilde{T}$ s.t. $X_{t''}\leq Y_{t'}$. Then, for any $t'+\hat{t}\leq t''\leq t'+\widetilde{T}$, $X_{t''}>Y_{t'}$.

Next, recall $\widetilde{T}=\left\lceil\frac{T}{\hat{t}\log T}\right\rceil\hat{t}$, then we have that

$$\begin{aligned}Y_{t'}&=\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\left(J(\pi^*)-\mathbb{E}\left[J(\omega_t)\right]\right)=\frac{1}{\left\lceil\frac{T}{\hat{t}\log T}\right\rceil}\sum_{\tau=1}^{\left\lceil\frac{T}{\hat{t}\log T}\right\rceil}\frac{1}{\hat{t}}\sum_{t=t'+(\tau-1)\hat{t}}^{t'+\tau\hat{t}-1}\left(J(\pi^*)-\mathbb{E}\left[J(\omega_t)\right]\right)\\&=\frac{1}{\left\lceil\frac{T}{\hat{t}\log T}\right\rceil}\sum_{\tau=1}^{\left\lceil\frac{T}{\hat{t}\log T}\right\rceil}X_{t'+\tau\hat{t}}\overset{(a)}{>}\frac{1}{\left\lceil\frac{T}{\hat{t}\log T}\right\rceil}\sum_{\tau=1}^{\left\lceil\frac{T}{\hat{t}\log T}\right\rceil}Y_{t'}=Y_{t'},\end{aligned} \tag{121}$$

where (a) follows from $t'+\hat{t}\leq t'+\tau\hat{t}\leq t'+\widetilde{T}$, $X_{t+\tau\hat{t}}>Y_{t'}$ for $\tau=1,...,\left\lceil\frac{T}{\hat{t}\log T}\right\rceil$. This hence results in a contradiction, which completes the proof. □

We then discuss the following two cases.

CASE 1:

[For any $2k+\hat{t}\leq t'\leq T-\widetilde{T}$, it holds that $eY_{t'}<X_{t'}$.]

Then, for $\widetilde{t}_0=2k+\hat{t}$, we have $X_{\widetilde{t}_0}>eY_{\widetilde{t}_0}$. Recall that $\widetilde{T}=\left\lceil\frac{T}{\hat{t}\log T}\right\rceil\hat{t}\leq\left(\frac{T}{\hat{t}\log T}+1\right)\hat{t}\leq\frac{T}{\log T}+\hat{t}\leq\frac{2T}{\log T}$ for large T. Then, $\frac{T}{\widetilde{T}}\geq\frac{\log T}{2}\geq\lfloor\frac{\log T}{2}\rfloor$. Thus, there exist $\widetilde{t}_0+\hat{t}\leq\widetilde{t}_1\leq\widetilde{t}_0+\widetilde{T}$, s.t. $X_{\widetilde{t}_1}\overset{(a)}{\leq}Y_{\widetilde{t}_0}\overset{(b)}{<}\frac{1}{e}X_{\widetilde{t}_0}$, where (a) follows from Lemma 14 and (b) follows from the condition of case 1. Recursively applying the above argument, we have that for $j=0,1,...,\lfloor\frac{\log T}{2}\rfloor$, there exists $\widetilde{t}_j+\hat{t}\leq\widetilde{t}_{j+1}\leq\widetilde{t}_j+\widetilde{T}$, such that $X_{\widetilde{t}_{j+1}}\leq Y_{\widetilde{t}_j}<\frac{1}{e}X_{\widetilde{t}_j}$. This further implies that

$$X_{\widetilde{t}_0}>eY_{\widetilde{t}_0}\geq eX_{\widetilde{t}_1}>e^2Y_{\widetilde{t}_1}\geq...\geq e^jY_{\widetilde{t}_j}>e^{j+1}Y_{\widetilde{t}_j}\geq...\geq e^{\lfloor\frac{\log T}{2}\rfloor}X_{\widetilde{t}_{\lfloor\frac{\log T}{2}\rfloor}}>e^{\lfloor\frac{\log T}{2}\rfloor+1}Y_{\widetilde{t}_{\lfloor\frac{\log T}{2}\rfloor}}. \tag{122}$$

Then, by Equation (122), we can conclude that

$$Y_{\widetilde{t}_{\lfloor\frac{\log T}{2}\rfloor}}:=\frac{1}{\widetilde{T}}\sum_{t=\widetilde{t}_{\lfloor\frac{\log T}{2}\rfloor}}^{\widetilde{t}_{\lfloor\frac{\log T}{2}\rfloor}+\widetilde{T}-1}\left(J(\pi^*)-\mathbb{E}\left[J(\omega_t)\right]\right)\leq\frac{1}{e^{\lfloor\frac{\log T}{2}\rfloor+1}}X_{\widetilde{t}_0}\leq\frac{1}{\sqrt{T}}X_{\widetilde{t}_0}. \tag{123}$$

Note that $X_{\widetilde{t}}\leq J(\pi^*)\leq R_{\max}$, hence we have that

$$Y_{\widetilde{t}_{\lfloor\frac{\log T}{2}\rfloor}}=\frac{1}{\widetilde{T}}\sum_{t=\widetilde{t}_{\lfloor\frac{\log T}{2}\rfloor}}^{\widetilde{t}_{\lfloor\frac{\log T}{2}\rfloor}+\widetilde{T}-1}\left(J(\pi^*)-\mathbb{E}\left[J(\omega_t)\right]\right)\leq\frac{R_{\max}}{\sqrt{T}}. \tag{124}$$

This further implies that

$$\min_{t\leq T}\mathbb{E}\left[J(\pi^*)-J(\omega_t)\right]\leq\frac{R_{\max}}{\sqrt{T}}. \tag{125}$$

This hence completes the proof of Theorem 2 under Case 1.

CASE 2:

[There exists some $2k+\hat{t}\leq t'\leq T-\widetilde{T}$ s.t. $X_{t'}\leq eY_{t'}$.]

Define $Z_{t'}=\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}M_t$. From Equation (120), we obtain that

$$\begin{aligned}
Z_{t'}&=\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}M_t\leq 2C_\phi^2C_M\frac{1}{\widetilde{T}}\sum_{i=0}^{\hat{t}-1}\left(J(\pi^*)-\mathbb{E}\left[J(\omega_{t'-i-1})\right]\right)\\
&\quad+\frac{4B^2+R_{\max}^2}{\widetilde{T}}+\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\sum_{j=t-\hat{t}}^{t-1}(1-q)^{t-j-1}\left(\widetilde{G}_j^\omega+\widetilde{G}_j^\theta+\widetilde{G}_j^\eta\right)\\
&=2C_\phi^2C_M\frac{\hat{t}}{\widetilde{T}}X_{t'}+\frac{4B^2+R_{\max}^2}{\widetilde{T}}+\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\sum_{j=t-\hat{t}}^{t-1}(1-q)^{t-j-1}\left(\widetilde{G}_j^\omega+\widetilde{G}_j^\theta+\widetilde{G}_j^\eta\right)\\
&\overset{(a)}{\leq}2C_\phi^2C_M\frac{e\hat{t}}{\widetilde{T}}Y_{t'}+\frac{4B^2+R_{\max}^2}{\widetilde{T}}+\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\sum_{j=t-\hat{t}}^{t-1}(1-q)^{t-j-1}\left(\widetilde{G}_j^\omega+\widetilde{G}_j^\theta+\widetilde{G}_j^\eta\right),
\end{aligned}\tag{126}$$

where (a) follows from the condition of Case 2.

Next, from Equation (119), we have that

$$\begin{aligned}
Y_{t'}&=\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}J(\pi^*)-\mathbb{E}\left[J(\omega_t)\right]\\
&\leq\frac{\mathrm{KL}(\omega_{t'})-\mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta\widetilde{T}}+\frac{C_\phi C_{\text{gap}}m\rho^k}{\lambda_{\min}}+C_\infty\sqrt{\varepsilon_{\text{actor}}}+C_\phi\sqrt{\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}M_t}+\frac{B^2L_\phi}{2}\beta\\
&=\frac{\mathrm{KL}(\omega_{t'})-\mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta\widetilde{T}}+\frac{C_\phi C_{\text{gap}}m\rho^k}{\lambda_{\min}}+C_\infty\sqrt{\varepsilon_{\text{actor}}}+C_\phi\sqrt{Z_{t'}}+\frac{B^2L_\phi}{2}\beta.
\end{aligned}\tag{127}$$

Plug Equation (127) in Equation (126), and we have that

$$\begin{aligned}
Z_{t'}&=\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}M_t\\
&\leq\frac{4B^2+R_{\max}^2}{\widetilde{T}}+\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\sum_{j=t-\hat{t}}^{t-1}(1-q)^{t-j-1}\left(\widetilde{G}_j^\omega+\widetilde{G}_j^\theta+\widetilde{G}_j^\eta\right)+\frac{2eC_\phi^2C_M\hat{t}}{\widetilde{T}}\left(\frac{C_\phi C_{\text{gap}}m\rho^k}{\lambda_{\min}}+\frac{B^2L_\phi}{2}\beta\right)\\
&\quad+2C_\phi^2C_M\frac{e\hat{t}}{\widetilde{T}}\left(\frac{\mathrm{KL}(\omega_{t'})-\mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta\widetilde{T}}+C_\infty\sqrt{\varepsilon_{\text{actor}}}+C_\phi\sqrt{Z_{t'}}\right)\\
&\overset{(a)}{\leq}\frac{4B^2+R_{\max}^2}{\widetilde{T}}+\frac{1}{\widetilde{T}}\sum_{t=t'}^{t'+\widetilde{T}-1}\sum_{j=t-\hat{t}}^{t-1}(1-q)^{t-j-1}\left(\widetilde{G}_j^\omega+\widetilde{G}_j^\theta+\widetilde{G}_j^\eta\right)+\frac{2eC_\phi^2C_M\hat{t}}{\widetilde{T}}\left(\frac{C_\phi C_{\text{gap}}m\rho^k}{\lambda_{\min}}+\frac{B^2L_\phi}{2}\beta\right)\\
&\quad+\left(\frac{2eC_\phi^3C_M\hat{t}}{\widetilde{T}}\right)^2+\frac{1}{2}Z_{t'}+\frac{C_\infty^2\varepsilon_{\text{actor}}}{2C_\phi^2}+2C_\phi^2C_M\frac{e\hat{t}}{\widetilde{T}}\left(\frac{\mathrm{KL}(\omega_{t'})-\mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta\widetilde{T}}\right),
\end{aligned}\tag{128}$$

where (a) follows from $xy \leq \frac{x^2+y^2}{2}$

Thus, it follows that

$$Z_{t'} \leq \frac{8B^2 + 2R_{\max}^2}{\widetilde{T}} + \frac{2}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta \right) + 2 \left(\frac{2eC_\phi^3 C_M \hat{t}}{\widetilde{T}} \right)^2 + \frac{C_\infty^2 \varepsilon_{\text{actor}}}{C_\phi^2}$$
$$+ 4C_\phi^2 C_M \frac{e\hat{t}}{\widetilde{T}} \left(\frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}} \right) + \frac{4eC_\phi^2 C_M \hat{t}}{\widetilde{T}} \left(\frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + \frac{B^2 L_\phi}{2} \beta \right). \tag{129}$$

Then, we get that

$$Y_{t'} \leq \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}} + \frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + C_\infty \sqrt{\varepsilon_{\text{actor}}} + C_\phi \sqrt{\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} M_t} + \frac{B^2 L_\phi}{2} \beta$$
$$\leq \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}} + \frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + 2C_\infty \sqrt{\varepsilon_{\text{actor}}} + C_\phi \sqrt{2} \left(\frac{2eC_\phi^3 C_M \hat{t}}{\widetilde{T}} \right)$$
$$+ C_\phi \sqrt{2} \sqrt{\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta \right)} + C_\phi \sqrt{2} \sqrt{\frac{4B^2 + R_{\max}^2}{\widetilde{T}}}$$
$$+ 2C_\phi \sqrt{C_\phi^2 C_M \frac{e\hat{t}}{\widetilde{T}} \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}}} + 2C_\phi \sqrt{\frac{e\hat{t} C_\phi^2 C_M}{\widetilde{T}} \left(\frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + \frac{B^2 L_\phi}{2} \beta \right)} + \frac{B^2 L_\phi}{2} \beta, \tag{130}$$

which proves the claim under Case 2.

Thus, combine the Case 1 result in Equation (124) and the Case 2 result in Equation (130), we have that

$$\min_{t \leq T} \mathbb{E} \left[J(\pi^*) - J(\omega_t) \right]$$
$$\leq \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}} + \frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + 2C_\infty \sqrt{\varepsilon_{\text{actor}}} + C_\phi \sqrt{2} \left(\frac{2eC_\phi^3 C_M \hat{t}}{\widetilde{T}} \right) + C_\phi \sqrt{2} \sqrt{\frac{4B^2 + R_{\max}^2}{\widetilde{T}}}$$
$$+ C_\phi \sqrt{2} \sqrt{\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta \right)} + 2C_\phi \sqrt{C_\phi^2 C_M \frac{e\hat{t}}{\widetilde{T}} \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}}}$$
$$+ 2C_\phi \sqrt{\frac{e\hat{t} C_\phi^2 C_M}{\widetilde{T}} \left(\frac{C_\phi C_{\text{gap}} m \rho^k}{\lambda_{\min}} + \frac{B^2 L_\phi}{2} \beta \right)} + \frac{B^2 L_\phi}{2} \beta + \frac{2R_{\max}}{\sqrt{T}}. \tag{131}$$

which completes the proof. □

Next, we prove Theorem 2.

Theorem 4. *(Restatement of Theorem 2) Consider the NAC algorithm in Algorithm 1 with constant step sizes that $\gamma \geq \alpha \geq \beta$, then it holds that*

$$\min_{t \leq T} \mathbb{E} \left[J(\pi^*) - J(\omega_t) \right] \leq \mathcal{O} \left(\frac{\log^2 T}{T\alpha} \right) + \mathcal{O} \left(\frac{\log T}{T\beta} \right) + \mathcal{O} \left(\frac{\sqrt{\log^3 T}}{T \sqrt{\alpha\beta}} \right) + \mathcal{O} \left(\sqrt{\frac{\beta \log T}{T\alpha}} \right) + \mathcal{O} \left(\sqrt{\frac{\log T}{T}} \right)$$
$$+ \mathcal{O} \left(\frac{\sqrt{\gamma\beta \log^2 T}}{\sqrt{\alpha}} \right) + \mathcal{O} \left(\frac{\gamma \sqrt{\log T}}{\sqrt{\alpha}} \right) + \mathcal{O} \left(\frac{\beta}{\sqrt{\alpha}} \right) + \mathcal{O} \left(\sqrt{\alpha \log^3 T} \right) + \mathcal{O} \left(\sqrt{\beta \log^3 T} \right)$$
$$+ \mathcal{O} \left(\sqrt{(m\rho^k) \log T} \right) + \mathcal{O} \left(\sqrt{\frac{(m\rho^k)\beta}{\alpha}} \right) + \mathcal{O} \left(\sqrt{\frac{(m\rho^k)\gamma}{\alpha}} \right) + \mathcal{O} \left(\frac{m\rho^k}{\sqrt{\alpha\beta}} \right) + \mathcal{O} \left(\sqrt{\varepsilon_{actor}} \right).$$

If we set $\gamma = \mathcal{O}(T^{-\frac{2}{3}} \log T), \alpha = \mathcal{O}(T^{-\frac{2}{3}} \log^{-1} T), \beta = \mathcal{O}(T^{-\frac{2}{3}} \log^{-1} T)$*, we have*

$$\min_{t \le T} J(\pi^*) - J(\omega_t) \le \mathcal{O}\left(T^{-\frac{1}{3}} \log^3 T\right) + \mathcal{O}\left(\sqrt{\varepsilon_{actor}}\right). \tag{132}$$

Proof. From Lemma 13, we have that

$$\begin{aligned}
&\min_{t \le T} \mathbb{E}\left[J(\pi^*) - J(\omega_t)\right] \\
&\le \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}} + \frac{C_\phi C_{\mathrm{gap}} m \rho^k}{\lambda_{\min}} + 2C_\infty \sqrt{\varepsilon_{\mathrm{actor}}} + C_\phi \sqrt{2} \left(\frac{2eC_\phi^3 C_M \hat{t}}{\widetilde{T}}\right) + C_\phi \sqrt{2} \sqrt{\frac{4B^2 + R_{\max}^2}{\widetilde{T}}} \\
&+ C_\phi \sqrt{2} \sqrt{\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta\right)} + 2C_\phi \sqrt{C_\phi^2 C_M \frac{e\hat{t}}{\widetilde{T}} \frac{\mathrm{KL}(\omega_{t'}) - \mathrm{KL}(\omega_{t'+\widetilde{T}})}{\beta \widetilde{T}}} \\
&+ 2C_\phi \sqrt{\frac{e\hat{t} C_\phi^2 C_M}{\widetilde{T}} \left(\frac{C_\phi C_{\mathrm{gap}} m \rho^k}{\lambda_{\min}} + \frac{B^2 L_\phi}{2} \beta\right)} + \frac{B^2 L_\phi}{2} \beta + \frac{2R_{\max}}{\sqrt{T}}.
\end{aligned} \tag{133}$$

We then set the stepsize as follows:

$$\begin{aligned}
\gamma &= T^{-\frac{2}{3}} \log T; \\
\alpha &= \min\left\{\frac{\bar{\lambda}_{\min}}{(k+1)^2 C_\phi^2} \gamma, \frac{3}{2\bar{\lambda}_{\min}}\right\} = \mathcal{O}\left(T^{-\frac{2}{3}} \log^{-1} T\right); \\
\beta &= \min\left\{\frac{\bar{\lambda}_{\min}}{4(2C_\Theta + 1)} \alpha, \frac{\bar{\lambda}_{\min}}{12 C_\phi^4 C_M} \alpha\right\} = \mathcal{O}\left(T^{-\frac{2}{3}} \log^{-1} T\right).
\end{aligned} \tag{134}$$

Recall that

$$\begin{aligned}
k &= \lceil \frac{\log T}{1-\rho} \rceil; \\
q &= \frac{\bar{\lambda}_{\min} \alpha}{2} = \mathcal{O}(\alpha); \\
\hat{t} &= \left\lceil \frac{1}{q} \log T \right\rceil = \mathcal{O}\left(\frac{\log T}{\alpha}\right); \\
\widetilde{T} &= \left\lceil \frac{T}{\hat{t} \log T} \right\rceil \hat{t} = \mathcal{O}\left(\frac{T}{\log T}\right).
\end{aligned} \tag{135}$$

Applying the above stepsizes in Equation (134), for $t \le T$, we can have that

$$\begin{aligned}
\widetilde{G}_t^\eta &= \mathcal{O}\left((m\rho^k)\gamma + \beta^2 + k^2 \beta \gamma + k \gamma^2\right) = \mathcal{O}\left(T^{-\frac{4}{3}} \log^3 T\right); \\
\widetilde{G}_t^\theta &= \mathcal{O}\left(k(m\rho^k)\alpha + \frac{(m\rho^k)^2}{\beta} + k^3 \alpha^2 + k^3 \alpha \beta + \beta^2\right) = \mathcal{O}\left(T^{-\frac{4}{3}} \log T\right); \\
\widetilde{G}_t^\omega &= \mathcal{O}\left((m\rho^k)\beta + \beta^2\right) = \mathcal{O}\left(T^{-\frac{4}{3}} \log^{-2} T\right).
\end{aligned} \tag{136}$$

Besides, we have that $\mathrm{KL}(\omega_i) = \mathbb{E}_{D_{\pi^*}}\left[\log\left(\frac{\pi^*(a|s)}{\pi_{\omega_i}(a|s)}\right)\right] \le \mathbb{E}_{D_{\pi^*}}\left[\log\left(\frac{D_{\pi^*}(s,a)}{D_{\pi_{\omega_i}}(s,a)}\right)\right] \overset{(a)}{\le} \log C_\infty$, where (a) follows from Assumption 3.

Thus, it holds that

$$\sqrt{\frac{1}{\widetilde{T}} \sum_{t=t'}^{t'+\widetilde{T}-1} \sum_{j=t-\hat{t}}^{t-1} (1-q)^{t-j-1} \left(\widetilde{G}_j^\omega + \widetilde{G}_j^\theta + \widetilde{G}_j^\eta\right)} = \mathcal{O}\left(T^{-\frac{1}{3}} \log^2 T\right). \tag{137}$$

Plugging the above equations to Equation (131) , we have that

$$\min_{t\leq T}\mathbb{E}\left[J(\pi^*)-J(\omega_t)\right]\leq\left(T^{-\frac{1}{3}}\log^3 T\right)+\mathcal{O}\left(\sqrt{\varepsilon_{\text{actor}}}\right). \tag{138}$$

This concludes the proof. □

D. Proof of Lemmas

Proof of Lemma 1. Recall the definition of $\nabla J(\omega)$ in Equation (1):

$$\nabla J(\omega)=\mathbb{E}_{D_{\pi_\omega}}\left[Q^{\pi_\omega}(s,a)\phi_\omega(s,a)\right], \tag{139}$$

which implies that

$$\begin{aligned}\left\|\nabla J(\omega)\right\|_2&=\left\|\mathbb{E}_{D_{\pi_\omega}}\left[Q^{\pi_\omega}(s,a)\phi_\omega(s,a)\right]\right\|_2\\&\overset{(a)}{=}\left\|\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega^\top(s,a)\bar{\theta}^*_\omega\phi_\omega(s,a)\right]\right\|_2\\&\leq C_\phi^2\left(\left\|\theta^*_\omega\right\|_2+\left\|\bar{\theta}^*_\omega-\theta^*_\omega\right\|_2\right)\\&\overset{(b)}{\leq} C_\phi^2\left(B+C_{\text{gap}}\frac{m\rho^k}{\lambda_{\min}}\right)=C_J,\end{aligned} \tag{140}$$

where (a) follows from Equation (3) and (b) follows from Proposition 4. It hence proves the first claim. The second claim is proved in Lemma A.1 in (Wu et al., 2020). □

Proof of Lemma2. Recall that $F_\omega=\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega^\top(s,a)\right]$. From the definition of $\bar{\theta}^*_\omega$ in Equation (2), it can be verified that

$$\bar{\theta}^*_\omega=F_\omega^{-1}\nabla J(\omega). \tag{141}$$

Hence

$$\begin{aligned}&\left\|\bar{\theta}^*_\omega-\bar{\theta}^*_{\omega'}\right\|_2\\&\quad=\left\|F_\omega^{-1}\nabla J(\omega)-F_{\omega'}^{-1}\nabla J(\omega')\right\|_2\\&\quad\leq\left\|F_\omega^{-1}\nabla J(\omega)-F_{\omega'}^{-1}\nabla J(\omega)\right\|_2+\left\|F_{\omega'}^{-1}\nabla J(\omega)-F_{\omega'}^{-1}\nabla J(\omega')\right\|_2\\&\quad\overset{(a)}{\leq}\left\|F_\omega^{-1}\right\|_2\left\|F_{\omega'}^{-1}\right\|_2\left\|F_\omega-F_{\omega'}\right\|_2\left\|\nabla J(\omega)\right\|_2+\left\|(F_{\omega'})^{-1}\right\|_2\left\|\nabla J(\omega)-\nabla J(\omega')\right\|_2,\end{aligned} \tag{142}$$

where (a) follows from Equation (55). Note that F_ω can be shown to be Lipschitz as follows:

$$\begin{aligned}\left\|F_\omega-F_{\omega'}\right\|_2&=\left\|\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega^\top(s,a)\right]-\mathbb{E}_{D_{\pi_{\omega'}}}\left[\phi_{\omega'}(s,a)\phi_{\omega'}^\top(s,a)\right]\right\|_2\\&\leq\left\|\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\phi_\omega^\top(s,a)\right]-\mathbb{E}_{D_{\pi_\omega}}\left[\phi_{\omega'}(s,a)\phi_{\omega'}^\top(s,a)\right]\right\|_2\\&\quad+\left\|\mathbb{E}_{D_{\pi_\omega}}\left[\phi_{\omega'}(s,a)\phi_{\omega'}^\top(s,a)\right]-\mathbb{E}_{D_{\pi_{\omega'}}}\left[\phi_{\omega'}(s,a)\phi_{\omega'}^\top(s,a)\right]\right\|_2\\&\leq 2C_\phi L_\phi\left\|\omega'-\omega\right\|_2+C_\phi^2\left\|D_{\pi_\omega}-D_{\pi_{\omega'}}\right\|_{\mathcal{TV}}\\&\overset{(a)}{\leq} 2C_\phi L_\phi\left\|\omega'-\omega\right\|_2+C_\phi^2 L_\pi\left\|\omega-\omega'\right\|_2,\end{aligned} \tag{143}$$

where (a) follows from (Zou et al., 2019) and Theorem 1 in (Li et al., 2024), that

$$\left\|D_{\pi_\omega}-D_{\pi_{\omega'}}\right\|_{\mathcal{TV}}\leq L_\pi\left\|\omega-\omega'\right\|_2. \tag{144}$$

Hence combining Equation (142), Equation (143) and Lemma 1, we obtain that

$$\left\|\bar{\theta}_{\omega}^{*}-\bar{\theta}_{\omega'}^{*}\right\|_{2} \leq\left(\frac{C_{J}}{\lambda_{\min }^{2}}\left(2 C_{\phi} L_{\phi}+C_{\phi}^{2} L_{\pi}\right)+\frac{L_{J}}{\lambda_{\min }}\right)\left\|\omega-\omega'\right\|_{2} . \tag{145}$$

This completes the proof.

□

Proof of Lemma 3. From the definition of H_ω in Equation (11), we first have that

$$\begin{aligned} H_{\omega}= & \mathbb{E}_{D_{\pi_{\omega}}}\left[\mathbb{E}\left[\phi_{\omega}\left(s_{0}, a_{0}\right)\left(\phi_{\omega}\left(s_{k}, a_{k}\right)-\phi_{\omega}\left(s_{0}, a_{0}\right)\right)^{\top} \mid s_{0}=s, a_{0}=a, \pi_{\omega}\right]\right] \\ \stackrel{(a)}{=} & \mathbb{E}_{D_{\pi_{\omega}}}\left[\phi_{\omega}(s, a)\left(\mathbb{E}\left[\phi_{\omega}^{\top}\left(s_{k}, a_{k}\right) \mid s_{0}=s, a_{0}=a, \pi_{\omega}\right]-\mathbb{E}_{D_{\pi_{\omega}}}\left[\phi_{\omega}^{\top}(s, a)\right]\right)\right] \\ & -\mathbb{E}_{D_{\pi_{\omega}}}\left[\phi_{\omega}(s, a) \phi_{\omega}^{\top}(s, a)\right], \end{aligned} \tag{146}$$

where (a) follows from $\phi_\omega(s,a)=\nabla\log\pi_\omega(a|s)$ and $\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega^\top(s,a)f(s)\right]=0$, where $f(s)$ is the function which is not determined by action a.

Define $\Delta H_\omega=\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega(s,a)\left(\mathbb{E}\left[\phi_\omega^\top(s_k,a_k)|s_0=s,a_0=a,\pi_\omega\right]-\mathbb{E}_{D_{\pi_\omega}}\left[\phi_\omega^\top(s,a)\right]\right)\right]$. Thus,

$$\frac{H_{\omega}+H_{\omega}^{\top}}{2}=\frac{\Delta H_{\omega}+\Delta H_{\omega}^{\top}}{2}-\mathbb{E}_{D_{\pi_{\omega}}}\left[\phi_{\omega}(s, a) \phi_{\omega}^{\top}(s, a)\right] . \tag{147}$$

For any symmetric matrices X and Y, $\lambda_{\max}(X+Y)\leq\lambda_{\max}(X)+\lambda_{\max}(Y)$. Thus, we have that

$$\begin{aligned} \lambda_{\max }\left(\frac{H_{\omega}+H_{\omega}^{\top}}{2}\right) & \leq \lambda_{\max }\left(\frac{\Delta H_{\omega}+\Delta H_{\omega}^{\top}}{2}\right)+\lambda_{\max }\left(-\mathbb{E}_{D_{\pi_{\omega}}}\left[\phi_{\omega}(s, a) \phi_{\omega}^{\top}(s, a)\right]\right) \\ & \leq C_{\phi}^{2} \mathbb{E}\left[\left\|P\left(s_{k}, a_{k} \mid s_{0}=s, a_{0}=s, \pi_{\omega}\right), D_{\pi_{\omega}}\right\|_{\mathcal{TV}}\right]-\lambda_{\min } \\ & \leq d C_{\phi}^{2} m \rho^{k}-\lambda_{\min }=-\bar{\lambda}_{\min }, \end{aligned} \tag{148}$$

where last inequality follows from Assumption 1. □

Proof of Lemma 4. Conditioned on (s_{t-k},a_{t-k}), the sample trajectory in Algorithm 1 is generated according to the following Markov chain:

$$\left(s_{t-k}, a_{t-k}\right) \xrightarrow{\pi_{t-k} \times P}\left(s_{t-k+1}, a_{t-k+1}\right) \xrightarrow{\pi_{t-k+1} \times P} \ldots\left(s_{t}, a_{t}\right) \xrightarrow{\pi_{t} \times P}\left(s_{t+1}, a_{t+1}\right) . \tag{149}$$

Using the technique in (Zou et al., 2019), we construct an auxiliary Markov chain as follows. Before time $t-k$, the states and actions are generated according to Algorithm 1; and after time $t-k$, all the subsequent state-action pairs, denoted by $(\widetilde{s}_l,\widetilde{a}_l)$, are generated according to a fixed policy π_t and transition kernel P:

$$\left(s_{t-k}, a_{t-k}\right) \xrightarrow{\pi_{t} \times P}\left(\widetilde{s}_{t-k+1}, \widetilde{a}_{t-k+1}\right) \xrightarrow{\pi_{t} \times P} \ldots\left(\widetilde{s}_{t}, \widetilde{a}_{t}\right) \xrightarrow{\pi_{t} \times P}\left(\widetilde{s}_{t+1}, \widetilde{a}_{t+1}\right) . \tag{150}$$

Denote by $\widetilde{\mathcal{F}}_t$ the filtration corresponding to the auxiliary Markov chain designed in Equation (150).

Then follow steps similar to those in (Zou et al., 2019, Appendix B) and (Li et al., 2024, Lemma 6), it can be shown that

$$\left\|\mathbb{P}\left(s_{t}, a_{t} \mid \mathcal{F}_{t-k}\right)-D_{t}\right\|_{\mathcal{TV}} \leq m \rho^{k}+\sum_{j=t-k}^{t} C_{\pi}\left\|\omega_{t}-\omega_{j}\right\|_{2} . \tag{151}$$

□

Proof of Lemma 5. Define the sum of the feature along the trajectory as follows:

$$z_t = \sum_{j=t-k}^{t} \phi_j(s_j, a_j), \quad \hat{z}_t = \sum_{j=t-k}^{t} \phi_t(s_j, a_j) \text{ and } \widetilde{z}_t = \sum_{j=t-k}^{t} \phi_t(\widetilde{s}_j, \widetilde{a}_j). \tag{152}$$

For every policy π_t, we construct another auxiliary Markov chain, denoted by $\{(\bar{s}_j, \bar{a}_j)\}_{j=0}^{\infty}$, which is under the stationary distribution induced by policy π_t and transition kernel P, i.e.,

$$(\bar{s}_0, \bar{a}_0) \sim D_t, \tag{153}$$

and all the subsequent actions are generated by π_t. Define

$$\bar{z}_t = \sum_{j=t-k}^{t} \phi_t(\bar{s}_j, \bar{a}_j). \tag{154}$$

Denote by $\bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t, \omega_t) = R(\bar{s}_t, \bar{a}_t) - J(\omega_t) + \phi_t^\top(\bar{s}_{t+1}, \bar{a}_{t+1})\theta_t - \phi_t^\top(\bar{s}_t, \bar{a}_t)\theta_t$.

Lemma 15. *It holds that*

$$\mathbb{E}\left[\bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta, \omega_t) | \pi_t\right] = H_{\omega_t}\theta + b_{\omega_t}. \tag{155}$$

From the definition in Equation (4), θ_t^* is the fixed point of the k-step TD operator $\mathcal{T}_{\pi_t}^{(k)}$. Then, it follows that

$$H_{\omega_t}\theta_t^* + b_{\omega_t} = \mathbb{E}_{D_t}\left[\phi_t^\top(s,a)\left(\mathcal{T}_{\pi_t}^{(k)}\left(\phi_t^\top(s,a)\theta_t^*\right) - \phi_t^\top(s,a)\theta_t^*\right)\right] = \mathbf{0}. \tag{156}$$

Together with Lemma 15, we have that

$$\mathbb{E}\left[\bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t^*, \omega_t)\right] = 0. \tag{157}$$

Thus we have that

$$\begin{aligned}
\mathbb{E}\left[\langle \theta_t - \theta_t^*, \bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t, \omega_t)\rangle\right] &= \mathbb{E}\left[\langle \theta_t - \theta_t^*, \bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t, \omega_t) - \bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t^*, \omega_t)\rangle\right] \\
&= \mathbb{E}\left[\langle \theta_t - \theta_t^*, H_{\omega_t}(\theta_t - \theta_t^*)\rangle\right] \\
&\leq \lambda_{\max}\left(\frac{H_{\omega_t} + H_{\omega_t}^\top}{2}\right)\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] \\
&\overset{(a)}{\leq} -\bar{\lambda}_{\min}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right],
\end{aligned} \tag{158}$$

where (a) follows from Lemma 3.

Then, recall $\hat{z}_t = \sum_{j=t-k}^{t} \phi_t(s_j, a_j)$. Denote by $\hat{\delta}_t = R(s_t, a_t) - J(\omega_t) + \phi_t^\top(s_{t+1}, a_{t+1})\theta_t - \phi_t^\top(s_t, a_t)\theta_t$, we have that

$$\begin{aligned}
\mathbb{E}\left[\langle \theta_t - \theta_t^*, \delta_t z_t\rangle\right] &= \mathbb{E}\left[\langle \theta_t - \theta_t^*, \bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t, \omega_t)\rangle\right] + \mathbb{E}\left[\left\langle \theta_t - \theta_t^*, z_t \hat{\delta}_t - \bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t, \omega_t)\right\rangle\right] \\
&\quad + \mathbb{E}\left[\left\langle \theta_t - \theta_t^*, z_t \delta_t - \hat{z}_t \hat{\delta}_t\right\rangle\right] \\
&\overset{(a)}{\leq} -\bar{\lambda}_{\min}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \mathbb{E}\left[\left\langle \theta_t - \theta_t^*, z_t \hat{\delta}_t - \bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t)\right\rangle\right] \\
&\quad + \mathbb{E}\left[\langle \theta_t - \theta_t^*, z_t\left(J(\omega_t) - \eta_t\right)\rangle\right] \\
&\leq -\bar{\lambda}_{\min}\mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \mathbb{E}\left[\left\langle \theta_t - \theta_t^*, z_t \hat{\delta}_t - \bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta_t)\right\rangle\right]
\end{aligned}$$

$$+\frac{\bar{\lambda}_{\min}}{2}\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right]+\frac{(k+1)^2C_\phi^2}{2\bar{\lambda}_{\min}}\mathbb{E}\left[(J(\omega_t)-\eta_t)^2\right], \tag{159}$$

where (a) follows from Equation (158).

Consider the term $\mathbb{E}\left[\left\langle\theta_t-\theta_t^*, z_t\hat{\delta}_t-\bar{z}_t\bar{\delta}_t(\bar{s}_t,\bar{a}_t;\theta_t,\omega_t)\right\rangle\right]$, and we have that

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\theta_t-\theta_t^*, z_t\hat{\delta}_t-\bar{z}_t\bar{\delta}_t(\bar{s}_t,\bar{a}_t;\theta_t,\omega_t)\right\rangle\right]\\
&=\mathbb{E}\left[\left\langle\theta_t-\theta_{t-2k}-\theta_t^*+\theta_{t-2k}^*, z_t\hat{\delta}_t-\bar{z}_t\bar{\delta}_t(\bar{s}_t,\bar{a}_t;\theta_t,\omega_t)\right\rangle\right]\\
&\qquad+\mathbb{E}\left[\left\langle\theta_{t-2k}-\theta_{t-2k}^*, z_t\hat{\delta}_t-\bar{z}_t\bar{\delta}_t(\bar{s}_t,\bar{a}_t;\theta_t,\omega_t)\right\rangle\right]\\
&\le\underbrace{\mathbb{E}\left[\|\theta_t-\theta_{t-2k}\|_2+\left\|\theta_t^*-\theta_{t-2k}^*\right\|_2\left(\|z_t\|_2\left\|\hat{\delta}_t\right\|_2+\|\bar{z}_t\|_2\left\|\bar{\delta}_t(\bar{s}_t,\bar{a}_t;\theta_t,\omega_t)\right\|_2\right)\right]}_{(i)}\\
&\quad+\underbrace{\mathbb{E}\left[\left\langle\theta_{t-2k}-\theta_{t-2k}^*, z_t\hat{\delta}_t-\hat{z}_t\hat{\delta}_t\right\rangle\right]}_{(ii)}\\
&\quad+\underbrace{\mathbb{E}\left[\left\langle\theta_{t-2k}-\theta_{t-2k}^*, \hat{z}_t\hat{\delta}_t-\bar{z}_t\bar{\delta}_t(\bar{s}_t,\bar{a}_t;\theta_t,\omega_t)\right\rangle\right]}_{(iii)}.
\end{aligned}\tag{160}$$

In AC algorithm, recall $U_\delta=R_{\max}+2C_\phi B$. Then, consider $\|\theta_t-\theta_{t-2k}\|_2$ and $\left\|\theta_t^*-\theta_{t-2k}^*\right\|_2$, we have that

$$\|\theta_t-\theta_{t-2k}\|_2\le\left\|\sum_{j=t-2k}^{t-1}\alpha_j\delta_jz_j\right\|_2\le\sum_{j=t-2k}^{t-1}\alpha_j|\delta_j|\,\|z_j\|_2\overset{(a)}{\le}(k+1)C_\phi U_\delta\sum_{j=t-2k}^{t-1}\alpha_j, \tag{161}$$

where (a) follows from the fact that $\|z_t\|_2\le(k+1)C_\phi$.

Then, it can be shown that

$$\begin{aligned}
\left\|\theta_t^*-\theta_{t-2k}^*\right\|_2&=\left\|\bar{\theta}_t^*-\bar{\theta}_{t-2k}^*+\theta_t^*-\bar{\theta}_t^*+\bar{\theta}_{t-2k}^*-\theta_{t-2k}^*\right\|_2\\
&\le\left\|\bar{\theta}_t^*-\bar{\theta}_{t-2k}^*\right\|_2+\left\|\theta_t^*-\bar{\theta}_t^*\right\|_2+\left\|\bar{\theta}_{t-2k}^*-\theta_{t-2k}^*\right\|_2\\
&\le C_\Theta\|\omega_t-\omega_{t-2k}\|_2+\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}+\frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}}\\
&\le C_\Theta\left\|\sum_{j=t-2k}^{t-1}\beta_j\phi_j^\top(s_j,a_j)\theta_j\phi_j(s_j,a_j)\right\|_2+\frac{2C_{\text{gap}}m\rho^k}{\lambda_{\min}}\\
&\le C_\Theta C_\phi^2B\sum_{j=t-2k}^{t-1}\beta_j+\frac{2C_{\text{gap}}m\rho^k}{\lambda_{\min}}.
\end{aligned}\tag{162}$$

Thus, from Equation (162), the term (i) in Equation (160) can be bounded as follows:

$$(i)\le2(k+1)C_\phi U_\delta\left((k+1)C_\phi U_\delta\sum_{j=t-2k}^{t-1}\alpha_j+C_\Theta C_\phi^2B\sum_{j=t-2k}^{t-1}\beta_j+2C_{\text{gap}}\frac{m\rho^k}{\lambda_{\min}}\right). \tag{163}$$

Then, for term (ii) in Equation (160), it can be bounded as follows:

$$(ii)\le\mathbb{E}\left[\left\|\theta_{t-2k}-\theta_{t-2k}^*\right\|_2\|z_t-\hat{z}_t\|_2|\delta_t(\theta_t)|\right]\le2BU_\delta\left\|\sum_{j=t-k}^{t}\phi_j(s_j,a_j)-\phi_t(s_j,a_j)\right\|_2$$

$$
\begin{aligned}
&\leq 2BU_\delta \sum_{j=t-k}^{t} C_\pi \mathbb{E}\left[\|\omega_t - \omega_j\|_2\right] \leq 2BU_\delta C_\pi \sum_{j=t-k}^{t}\sum_{i=j}^{t} \mathbb{E}\left[\left\|\beta_i \phi_i^\top(s_i,a_i)\theta_i\phi_i(s_i,a_i)\right\|_2\right] \\
&\leq 2B^2C_\phi^2 U_\delta C_\pi \sum_{j=t-k}^{t}\sum_{i=j}^{t}\beta_i.
\end{aligned} \tag{164}
$$

Next, for term (iii) in Equation (160), we can show that

$$
\begin{aligned}
(iii) &= \mathbb{E}\left[\left\langle \theta_{t-2k} - \theta_{t-2k}^*, \hat{z}_t\hat{\delta}_t - \bar{z}_t\bar{\delta}_t(\bar{s}_t,\bar{a}_t;\theta_t)\right\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle \theta_{t-2k} - \theta_{t-2k}^*, \hat{z}_t\hat{\delta}_t - \bar{z}_t\bar{\delta}_t(\bar{s}_t,\bar{a}_t;\theta_t)\right\rangle|\mathcal{F}_{t-2k}\right]\right] \\
&\leq 4BC_\phi U_\delta \sum_{j=t-k}^{t} \mathbb{E}\left[\|\mathbb{P}(s_j,a_j|\mathcal{F}_{t-2k}) - D_t\|_{\mathcal{TV}}\right] \\
&\leq 4BC_\phi U_\delta \sum_{j=t-k}^{t} \mathbb{E}\left[\|\mathbb{P}(s_j,a_j|\mathcal{F}_{t-2k}) - D_j\|_{\mathcal{TV}} + \|D_j - D_t\|_{\mathcal{TV}}\right] \\
&\leq 4BC_\phi U_\delta \sum_{j=t-k}^{t}\left(C_\pi \sum_{i=j-k}^{j-1}\mathbb{E}\left[\|\omega_i-\omega_j\|_2\right] + m\rho^k + L_\pi\mathbb{E}\left[\|\omega_t-\omega_j\|_2\right]\right) \\
&\leq 4BC_\phi U_\delta \sum_{j=t-k}^{t}\left(C_\pi \sum_{i=j-k}^{j-1}\sum_{\iota=i}^{j-1}\beta_\iota C_\phi^2 B + m\rho^k + L_\pi \sum_{i=j}^{t-1}\beta_i C_\phi^2 B\right).
\end{aligned} \tag{165}
$$

Thus, combining Equation (163), Equation (164) and Equation (165), we can bound term as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\langle \theta_t - \theta_t^*, \delta_t z_t\rangle\right] \leq -\frac{\bar{\lambda}_{\min}}{2}\mathbb{E}\left[\|\theta_t-\theta_t^*\|_2^2\right] + \frac{(k+1)^2C_\phi^2}{2\bar{\lambda}_{\min}}\mathbb{E}\left[(J(\omega_t)-\eta_t)^2\right] \\
&+ 2B^2C_\phi^2U_\delta C_\pi \sum_{j=t-k}^{t}\sum_{i=j}^{t}\beta_i + 4BC_\phi U_\delta \sum_{j=t-k}^{t}\left(BC_\phi^2C_\pi \sum_{i=j-k}^{j-1}\sum_{\iota=i}^{j-1}\beta_\iota + m\rho^k + BC_\phi^2 L_\pi \sum_{i=j}^{t-1}\beta_i\right) \\
&+ 2(k+1)C_\phi U_\delta\left((k+1)C_\phi U_\delta \sum_{j=t-2k}^{t-1}\alpha_j + C_\Theta C_\phi^2 B \sum_{j=t-2k}^{t-1}\beta_j + \frac{2C_{\text{gap}}m\rho^k}{\lambda_{\min}}\right).
\end{aligned} \tag{166}
$$

In NAC algorithm, terms $\|\theta_t - \theta_{t-2k}\|_2$ and $\left\|\theta_t^* - \theta_{t-2k}^*\right\|_2$ can be bounded as follows:

$$
\|\theta_t - \theta_{t-2k}\|_2 \leq \left\|\sum_{j=t-2k}^{t-1}\alpha_j\delta_j z_j\right\|_2 \leq (k+1)C_\phi U_\delta \sum_{j=t-2k}^{t-1}\alpha_j, \tag{167}
$$

and

$$
\begin{aligned}
\left\|\theta_t^* - \theta_{t-2k}^*\right\|_2 &= \left\|\bar{\theta}_t^* - \bar{\theta}_{t-2k}^* + \theta_t^* - \bar{\theta}_t^* + \bar{\theta}_{t-2k}^* - \theta_{t-2k}^*\right\|_2 \\
&\leq \left\|\bar{\theta}_t^* - \bar{\theta}_{t-2k}^*\right\|_2 + \left\|\theta_t^* - \bar{\theta}_t^*\right\|_2 + \left\|\bar{\theta}_{t-2k}^* - \theta_{t-2k}^*\right\|_2 \\
&\leq C_\Theta \left\|(\omega_t - \omega_{t-2k})\right\|_2 + \frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}} + \frac{C_{\text{gap}}m\rho^k}{\lambda_{\min}} \\
&\leq C_\Theta \left\|\sum_{j=t-2k}^{t-1}\beta_j\theta_j\right\|_2 + \frac{2C_{\text{gap}}m\rho^k}{\lambda_{\min}}
\end{aligned}
$$

$$\leq C_\Theta B \sum_{j=t-2k}^{t-1} \beta_j + \frac{2C_{\text{gap}} m\rho^k}{\lambda_{\min}}. \tag{168}$$

Thus, using Equation (168) and Equation (167), term (i) in Equation (160) can be bounded as

$$(i) \leq 2(k+1)C_\phi U_\delta \left((k+1)C_\phi U_\delta \sum_{j=t-2k}^{t-1} \alpha_j + C_\Theta B \sum_{j=t-2k}^{t-1} \beta_j + 2C_{\text{gap}} \frac{m\rho^k}{\lambda_{\min}} \right). \tag{169}$$

Next, we bound the term (ii) in Equation (160) as follows:

$$(ii) \leq \mathbb{E}\left[\left\|\theta_{t-2k} - \theta_{t-2k}^*\right\|_2 \|z_t - \hat{z}_t\|_2 \left|\hat{\delta}_t\right| \right] \leq 2BU_\delta \mathbb{E}\left[\left\| \sum_{j=t-k}^{t} \phi_j(s_j, a_j) - \phi_t(s_j, a_j) \right\|_2 \right]$$

$$\leq 2BU_\delta \sum_{j=t-k}^{t} C_\pi \mathbb{E}\left[\|\omega_t - \omega_j\|_2\right] \leq 2BU_\delta C_\pi \sum_{j=t-k}^{t} \sum_{i=j}^{t-1} \|\beta_i \theta_i\|_2 \leq 2B^2 U_\delta C_\pi \sum_{j=t-k}^{t} \sum_{i=j}^{t-1} \beta_i. \tag{170}$$

Term (iii) in Equation (160) can be bounded as

$$\begin{aligned}
(iii) &= \mathbb{E}\left[\left\langle \theta_{t-2k} - \theta_{t-2k}^*, \hat{z}_t \delta_t(\theta_t) - z_t' \delta_t'(\theta_t) \right\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle \theta_{t-2k} - \theta_{t-2k}^*, \hat{z}_t \delta_t(\theta_t) - z_t' \delta_t'(\theta_t) \right\rangle |\mathcal{F}_{t-2k}\right]\right] \\
&\leq 2B2C_\phi U_\delta \sum_{j=t-k}^{t} \mathbb{E}\left[\|\mathbb{P}(s_j, a_j|\mathcal{F}_{t-2k}) - D_t\|_{\mathcal{TV}}\right] \\
&\leq 4BC_\phi U_\delta \sum_{j=t-k}^{t} \mathbb{E}\left[\|\mathbb{P}(s_j, a_j|\mathcal{F}_{t-2k}) - D_j\|_{\mathcal{TV}} + \|D_j - D_t\|_{\mathcal{TV}}\right] \\
&\leq 4BC_\phi U_\delta \sum_{j=t-k}^{t} \left(C_\pi \sum_{i=j-k}^{j-1} \mathbb{E}\left[\|\omega_i - \omega_j\|_2\right] + m\rho^k + L_\pi \mathbb{E}\left[\|\omega_t - \omega_j\|_2\right] \right) \\
&\leq 4B^2 C_\phi U_\delta \sum_{j=t-k}^{t} \left(C_\pi \sum_{i=j-k}^{j-1} \sum_{\iota=i}^{j-1} \beta_\iota + m\rho^k + L_\pi \sum_{i=j}^{t-1} \beta_i \right).
\end{aligned} \tag{171}$$

Combining the above bounds on terms $(i), (ii), (iii)$, term can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\langle \theta_t - \theta_t^*, \delta_t z_t \rangle\right] \\
&\leq -\frac{\bar{\lambda}_{\min}}{2} \mathbb{E}\left[\|\theta_t - \theta_t^*\|_2^2\right] + \frac{(k+1)^2 C_\phi^2}{2\bar{\lambda}_{\min}} \mathbb{E}\left[(J(\omega_t) - \eta_t)^2\right] + 2B^2 U_\delta C_\pi \sum_{j=t-k}^{t} \sum_{i=j}^{t-1} \beta_i \\
&\quad + 2(k+1)C_\phi U_\delta \left((k+1)C_\phi U_\delta \sum_{j=t-2k}^{t-1} \alpha_j + C_\Theta B \sum_{j=t-2k}^{t-1} \beta_j + \frac{2C_{\text{gap}} m\rho^k}{\lambda_{\min}} \right) \\
&\quad + 4BC_\phi U_\delta \sum_{j=t-k}^{t} \left(BC_\pi \sum_{i=j-k}^{j-1} \sum_{\iota=i}^{j-1} \beta_\iota + m\rho^k + BL_\pi \sum_{i=j}^{t-1} \beta_i \right).
\end{aligned} \tag{172}$$

This completes the proof. □

Proof of Lemma 7. From $\nabla^2 J(\omega) = \sum_{s,a} \nabla^2 D_{\pi_\omega}(s,a) R(s,a)$, we can get that for any $\omega, \omega' \in \mathbb{R}^d$

$$\left\|\nabla^2 J(\omega) - \nabla^2 J(\omega')\right\|_2 = \left\| \sum_{s,a} \left(\nabla^2 D_{\pi_\omega}(s,a) R(s,a) - \nabla^2 D_{\pi_{\omega'}}(s,a) R(s,a) \right) \right\|_2. \tag{173}$$

Denote by $\boldsymbol{\sigma}(A)$ the spectral radius of matrix $A \in \mathbb{R}^{n\times n}$. Recall the fact that $\boldsymbol{\sigma}(A) \le \|A\|_\infty = \max_i \left\{\sum_j |a_{ij}|\right\}$. If A is symmetric matrix, $\|A\|_2 = \boldsymbol{\sigma}(A)$, and thus $\|A\|_2 \le \max_i \left\{\sum_j |a_{ij}|\right\}$.

It is clear that $\nabla^2 D_{\pi_\omega}(s,a)$ is symmetric, and therefore $\sum_{s,a}(\nabla^2 D_{\pi_\omega}(s,a) - \nabla^2 D_{\pi_{\omega'}}(s,a))R(s,a)$ is also symmetric. It then follows that

$$\begin{aligned}
&\left\|\sum_{s,a}\left(\nabla^2 D_{\pi_\omega}(s,a) - \nabla^2 D_{\pi_{\omega'}}(s,a)\right) R(s,a)\right\|_2 \\
&\le \max_i \left\{\sum_j \left|\sum_{s,a}\left(\partial_{\omega_i}\partial_{\omega_j} D_{\pi_\omega}(s,a) - \partial_{\omega_i}\partial_{\omega_j} D_{\pi_{\omega'}}(s,a)\right) R(s,a)\right|\right\} \\
&\overset{(a)}{=} \max_i \left\{\sum_j \left|\left(\nabla_\omega \sum_{s,a} \partial_{\omega_i}\partial_{\omega_j} D_{\pi_{\hat\omega}}(s,a)R(s,a)\right)^\top (\omega-\omega')\right|\right\} \\
&\le \max_i \left\{\sum_j \left\|\nabla_\omega \sum_{s,a} \partial_{\omega_i}\partial_{\omega_j} D_{\pi_{\hat\omega}}(s,a)R(s,a)\right\|_2 \|\omega-\omega'\|_2\right\} \\
&\overset{(b)}{\le} \max_i \left\{\sum_j \sum_l \left|\sum_{s,a} \partial_{\omega_i}\partial_{\omega_j}\partial_{\omega_l} D_{\pi_{\hat\omega}}(s,a)R(s,a)\right| \|\omega-\omega'\|_2\right\} \\
&\le \max_{i,j,l} \left\{d^2 \left|\sum_{s,a} \partial_{\omega_i}\partial_{\omega_j}\partial_{\omega_l} D_{\pi_{\hat\omega}}(s,a)R(s,a)\right|\right\} \|\omega-\omega'\|_2,
\end{aligned} \tag{174}$$

where (a) follows from the fact that D_{π_ω} is n times differentiable as long as the Theorem 4 in (Heidergott & Hordijk, 2003) and the Lagrange's Mean Value Theorem for some $\hat\omega = \lambda_{ij}\omega + (1-\lambda_{ij})\omega'$ with $\lambda_{ij} \in [0,1]$, and (b) follows from that for a vector a, $\|a\|_2 \le \|a\|_1$.

Define a function $v : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, denote by $\|f\|_v = \sup_{s,a} \frac{|f(s,a)|}{|v(s,a)|}$ the finite v-norm of function f. Set $v(s,a) = e^{R(s,a)}$ and we can get that

$$\sup_{s,a} \frac{|R(s,a)|}{|v(s,a)|} \le 1. \tag{175}$$

Moreover, $v(s,a) = e^{R(s,a)} \le e^{R_{\max}}$ and $v(s,a) \ge 1$. If $\|f\|_v \le 1$, it implies that

$$\sup_{s,a} |f(s,a)| \le e^{R_{\max}}. \tag{176}$$

For a (signed) measure μ, the associated norm is

$$\|\mu\|_v = \sup_{\|f\|_v \le 1} \left|\sum_{s,a} \mu(s,a) f(s,a)\right|. \tag{177}$$

For a kernel $P(s',a'|s,a)$, its associated norm is

$$\|P\|_v = \sup_{s,a} \sup_{\|f\|_v\le 1} \frac{\left|\sum_{s',a'} f(s',a')P(s',a'|s,a)\right|}{|v(s,a)|}. \tag{178}$$

From the fact that $\sup_{s,a} \frac{|R(s,a)|}{|v(s,a)|} \le 1$, we can get that $||R||_v \le 1$. This further implies that

$$\left|\sum_{s,a} \partial_{\omega_i}\partial_{\omega_j}\partial_{\omega_l} D_{\pi_{\hat\omega}}(s,a)R(s,a)\right| \le \sup_{\|f\|_v\le 1} \left|\sum_{s,a} \partial_{\omega_i}\partial_{\omega_j}\partial_{\omega_l} D_{\pi_{\hat\omega}}(s,a)f(s,a)\right|$$

$$= \left\|\partial_{\omega_i}\partial_{\omega_j}\partial_{\omega_l} D_{\pi_{\hat{\omega}}}\right\|_v . \tag{179}$$

Furthermore, we define $\mathcal{K}^{(1)}_{\omega_i}(s,a)$, $\mathcal{K}^{(2)}_{\omega_{i,j}}(s,a)$ and $\mathcal{K}^{(3)}_{\omega_{i,j,l}}(s,a)$ as follows:

$$\begin{aligned}
\mathcal{K}^{(1)}_{\omega_i}(s,a|s',a') &= \sum_{\iota=0}^{\infty} \partial_{\omega_i}\pi_\omega(a|s)\left(\mathbb{P}\left(s_\iota = s, a_\iota = a|s_0 = s', a_0 = a', \pi_\omega\right) - D_{\pi_\omega}(s,a)\right);\\
\mathcal{K}^{(2)}_{\omega_{i,j}}(s,a|s',a') &= \sum_{\iota=0}^{\infty} \partial_{\omega_i}\partial_{\omega_j}\pi_\omega(a|s)\left(\mathbb{P}\left(s_\iota = s, a_\iota = a|s_0 = s', a_0 = a', \pi_\omega\right) - D_{\pi_\omega}(s,a)\right);\\
\mathcal{K}^{(3)}_{\omega_{i,j,l}}(s,a|s',a') &= \sum_{\iota=0}^{\infty} \partial_{\omega_i}\partial_{\omega_j}\partial_{\omega_l}\pi_\omega(a|s)\left(\mathbb{P}\left(s_\iota = s, a_\iota = a|s_0 = s', a_0 = a', \pi_\omega\right) - D_{\pi_\omega}(s,a)\right).
\end{aligned} \tag{180}$$

Then, define the kernel $\Gamma_\omega(s',a'|s,a)$, s.t., $\Gamma_\omega(s',a'|s,a) = D_{\pi_\omega}(s',a')$ for any $s \in \mathcal{S}$, $a \in \mathcal{A}$. It then follows that

$$\begin{aligned}
\|\Gamma_\omega\|_v &= \sup_{s,a}\sup_{\|f\|_v \le 1} \frac{\left|\sum_{s',a'} f(s',a')\Gamma_\omega(s',a'|s,a)\right|}{|v(s,a)|}\\
&= \sup_{s,a}\sup_{\|f\|_v \le 1} \frac{\left|\sum_{s',a'} f(s',a') D_{\pi_\omega}(s',a')\right|}{|v(s,a)|} = \sup_{s,a}\frac{\|D_{\pi_\omega}\|_v}{|v(s,a)|}.
\end{aligned} \tag{181}$$

By Theorem 3 and proof in (Heidergott & Hordijk, 2003), we can get

$$\partial_{\omega_j}\mathcal{K}^{(1)}_{\omega_i} = \mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_j} + \mathcal{K}^{(2)}_{\omega_{ij}}; \qquad \partial_{\omega_l}\mathcal{K}^{(2)}_{\omega_{ij}} = \mathcal{K}^{(2)}_{\omega_{ij}}\mathcal{K}^{(1)}_{\omega_l} + \mathcal{K}^{(3)}_{\omega_{ijl}}. \tag{182}$$

Combining with Theorem 4 and Section 4 in (Heidergott & Hordijk, 2003), we can have

$$\begin{aligned}
\partial_{\omega_i}\Gamma_\omega &= \Gamma_\omega\mathcal{K}^{(1)}_{\omega_i};\\
\partial_{\omega_j}\partial_{\omega_i}\Gamma_\omega &= \left(\partial_{\omega_j}\Gamma_\omega\right)\mathcal{K}^{(1)}_{\omega_i} + \Gamma_\omega\partial_{\omega_i}\mathcal{K}^{(1)}_{\omega_j} = \Gamma_\omega\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_j} + \Gamma_\omega\left(\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_j} + \mathcal{K}^{(2)}_{\omega_{ij}}\right)\\
&= 2\Gamma_\omega\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_j} + \Gamma_\omega\mathcal{K}^{(2)}_{\omega_{ij}};\\
\partial_{\omega_l}\partial_{\omega_j}\partial_{\omega_i}\Gamma_\omega &= \partial_{\omega_l}\left(2\Gamma_\omega\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_j} + \Gamma_\omega\mathcal{K}^{(2)}_{\omega_{ij}}\right)\\
&= 2\left(\partial_{\omega_l}\Gamma_\omega\right)\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_j} + 2\Gamma_\omega\left(\partial_{\omega_l}\mathcal{K}^{(1)}_{\omega_i}\right)\mathcal{K}^{(1)}_{\omega_j} + 2\Gamma_\omega\mathcal{K}^{(1)}_{\omega_i}\left(\partial_{\omega_l}\mathcal{K}^{(1)}_{\omega_j}\right)\\
&\quad + \left(\partial_{\omega_l}\Gamma_\omega\right)\mathcal{K}^{(2)}_{\omega_{ij}} + \Gamma_\omega\left(\partial_{\omega_l}\mathcal{K}^{(2)}_{\omega_{ij}}\right)\\
&= 2\Gamma_\omega\mathcal{K}^{(1)}_{\omega_l}\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_j} + 2\Gamma_\omega\left(\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_l} + \mathcal{K}^{(2)}_{\omega_{il}}\right)\mathcal{K}^{(1)}_{\omega_j} + 2\Gamma_\omega\mathcal{K}^{(1)}_{\omega_i}\left(\mathcal{K}^{(1)}_{\omega_j}\mathcal{K}^{(1)}_{\omega_l} + \mathcal{K}^{(2)}_{\omega_{jl}}\right)\\
&\quad + \Gamma_\omega\mathcal{K}^{(1)}_{\omega_l}\mathcal{K}^{(2)}_{\omega_{ij}} + \Gamma_\omega\left(\mathcal{K}^{(2)}_{\omega_{ij}}\mathcal{K}^{(1)}_{\omega_l} + \mathcal{K}^{(3)}_{\omega_{ijl}}\right)\\
&= 6\Gamma_\omega\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(1)}_{\omega_j}\mathcal{K}^{(1)}_{\omega_l} + 2\Gamma_\omega\mathcal{K}^{(1)}_{\omega_i}\mathcal{K}^{(2)}_{\omega_{jl}} + 2\Gamma_\omega\mathcal{K}^{(1)}_{\omega_j}\mathcal{K}^{(2)}_{\omega_{il}} + 2\Gamma_\omega\mathcal{K}^{(1)}_{\omega_l}\mathcal{K}^{(2)}_{\omega_{ij}} + \Gamma_\omega\mathcal{K}^{(3)}_{\omega_{ijl}}.
\end{aligned} \tag{183}$$

Then, according to the discussion in Section 4 in (Heidergott & Hordijk, 2003), it can be shown that

$$\begin{aligned}
\sup_{s,a}\frac{\|D_{\pi_{\hat{\omega}}}\|_v}{|v(s,a)|} &= \left\|\partial_{\omega_i}\partial_{\omega_j}\partial_{\omega_l}\Gamma_{\pi_{\hat{\omega}}}\right\|_v\\
&\le \|\Gamma_{\pi_{\hat{\omega}}}\|_v\left\|\mathcal{K}^{(3)}_{\hat{\omega}_{ijl}}\right\|_v + 2\|\Gamma_{\pi_{\hat{\omega}}}\|_v\left\|\mathcal{K}^{(2)}_{\hat{\omega}_{ij}}\right\|_v\left\|\mathcal{K}^{(1)}_{\hat{\omega}_l}\right\|_v\\
&\quad + 2\|\Gamma_{\pi_{\hat{\omega}}}\|_v\left\|\mathcal{K}^{(2)}_{\hat{\omega}_{il}}\right\|_v\left\|\mathcal{K}^{(1)}_{\hat{\omega}_j}\right\|_v + 2\|\Gamma_{\pi_{\hat{\omega}}}\|_v\left\|\mathcal{K}^{(2)}_{\hat{\omega}_{lj}}\right\|_v\left\|\mathcal{K}^{(1)}_{\hat{\omega}_i}\right\|_v\\
&\quad + 6\|\Gamma_{\pi_{\hat{\omega}}}\|_v\left\|\mathcal{K}^{(1)}_{\hat{\omega}_i}\right\|_v\left\|\mathcal{K}^{(1)}_{\hat{\omega}_j}\right\|_v\left\|\mathcal{K}^{(1)}_{\hat{\omega}_l}\right\|_v
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{s,a} \frac{1}{|v(s,a)|} \big(\left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(3)}_{\hat{\omega}_{ijl}} \right\|_v + 2 \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(2)}_{\hat{\omega}_{ij}} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_l} \right\|_v \\
&\quad + 2 \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(2)}_{\hat{\omega}_{il}} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_j} \right\|_v + 2 \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(2)}_{\hat{\omega}_{lj}} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_i} \right\|_v \\
&\quad + 6 \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_i} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_j} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_l} \right\|_v \big).
\end{aligned} \tag{184}
$$

Note that $R_{\max} \geq R(s,a) \geq 0$ and $\sup_{s,a} \frac{1}{|v(s,a)|} \geq e^{-R_{\max}}$. Then we have that

$$
\begin{aligned}
\left\| \partial_{\omega_i} \partial_{\omega_j} \partial_{\omega_l} D_{\pi_{\hat{\omega}}} \right\|_v \leq & \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(3)}_{\hat{\omega}_{ijl}} \right\|_v + 2 \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(2)}_{\hat{\omega}_{ij}} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_l} \right\|_v \\
& + 2 \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(2)}_{\hat{\omega}_{il}} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_j} \right\|_v + 2 \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(2)}_{\hat{\omega}_{lj}} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_i} \right\|_v \\
& + 6 \left\| D_{\pi_{\hat{\omega}}} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_i} \right\|_v \left\| \mathcal{K}^{(1)}_{\hat{\omega}_j} \right\|_v \left\| \mathcal{K}^{(1)}_{(\hat{\omega})_l} \right\|_v .
\end{aligned} \tag{185}
$$

Next, we bound the term $\left\| D_{\pi_{\hat{\omega}}} \right\|_v$ as follows:

$$
\begin{aligned}
\left\| D_{\pi_{\hat{\omega}}} \right\|_v &= \sup_{\|f\|_v \leq 1} \left| \sum_{s,a} f(s,a) D_{\pi_{\hat{\omega}}}(s,a) \right| \leq \sup_{\|f\|_v \leq 1} \left| \sum_{s,a} |f(s,a)| \left| D_{\pi_{\hat{\omega}}}(s,a) \right| \right| \\
&\overset{(a)}{\leq} \left| \sum_{s,a} e^{R_{\max}} D_{\pi_{\hat{\omega}}}(s,a) \right| \leq e^{R_{\max}},
\end{aligned} \tag{186}
$$

where (a) follows from the Equation (176).

Then, we bound the terms in Equation (185) as follows:

$$
\begin{aligned}
&\left\| \mathcal{K}^{(1)}_{(\hat{\omega})_i} \right\|_v \\
&= \sup_{s',a'} \sup_{\|f\|_v \leq 1} \frac{\left| \sum_{\iota=0}^{\infty} \sum_{s,a} \partial_{\omega_i} \pi_\omega(a|s) \left(\mathbb{P}\left(s_\iota = s, a_\iota = a | s_0 = s', a_0 = a', \pi_\omega\right) - D_{\pi_\omega}(s,a) \right) f(s,a) \right|}{|v(s',a')|} \\
&\overset{(a)}{\leq} \sup_{s',a'} \sup_{\|f\|_v \leq 1} \left| \sum_{\iota=0}^{\infty} \sum_{s,a} \partial_{\omega_i} \pi_\omega(a|s) \left(\mathbb{P}\left(s_\iota = s, a_\iota = a | s_0 = s', a_0 = a', \pi_\omega\right) - D_{\pi_\omega}(s,a) \right) f(s,a) \right| \\
&\leq \sup_{s',a'} \sup_{\|f\|_v \leq 1} \left| \sum_{\iota=0}^{\infty} \sum_{s,a} \left| \partial_{\omega_i} \pi_\omega(a|s) \right| \left| \mathbb{P}\left(s_\iota = s, a_\iota = a | s_0 = s', a_0 = a', \pi_\omega\right) - D_{\pi_\omega}(s,a) \right| |f(s,a)| \right| \\
&\overset{(b)}{\leq} \sup_{s',a'} \sum_{\iota=0}^{\infty} \max_{s,a} \left\{ \left| \partial_{\omega_i} \pi_\omega(a|s) \right| \right\} \left\| \mathbb{P}\left(s_\iota, a_\iota | s_0 = s', a_0 = a', \pi_\omega\right) - D_{\pi_\omega} \right\|_{\mathcal{TV}} e^{R_{\max}} \\
&\overset{(c)}{\leq} \max_{s,a} \left\{ \left| \partial_{\omega_i} \pi_\omega(a|s) \right| \right\} \frac{m e^{R_{\max}}}{1-\rho},
\end{aligned} \tag{187}
$$

where (a) follows from that $|v(s,a)| \geq 1$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, (b) follows from Equation (176) and (c) follows from Assumption 1.

Similar to Equation (187), by Assumption 1 and Assumption 2, we can further bound $\left\| \mathcal{K}^{(1)}_{\omega_i} \right\|_v$, $\left\| \mathcal{K}^{(2)}_{\omega_{ij}} \right\|_v$ and $\left\| \mathcal{K}^{(3)}_{\omega_{ijl}} \right\|_v$ as follows:

$$
\left\| \mathcal{K}^{(1)}_{\omega_i} \right\|_v \leq \frac{m C_\phi e^{R_{\max}}}{1-\rho}; \left\| \mathcal{K}^{(2)}_{\omega_{ij}} \right\|_v \leq \frac{m C_\delta e^{R_{\max}}}{1-\rho}; \left\| \mathcal{K}^{(3)}_{\omega_{ijl}} \right\|_v \leq \frac{m L_\delta e^{R_{\max}}}{1-\rho}. \tag{188}
$$

Plug Equation (186) and Equation (188) in Equation (185), and we have that

$$
\left\| \partial_{\omega_i} \partial_{\omega_j} \partial_{\omega_l} D_{\pi_{\hat{\omega}}} \right\|_v \leq \frac{6 C_\phi^3 m^3 e^{4R_{\max}}}{(1-\rho)^3} + \frac{6 m^2 C_\phi C_\delta e^{3R_{\max}}}{(1-\rho)^2} + \frac{m L_\delta e^{2R_{\max}}}{1-\rho}. \tag{189}
$$

Thus, we can further get that

$$\left\|\nabla^2 J(\omega) - \nabla^2 J(\omega')\right\|_2 \le d^2 \left(\frac{6C_\phi^3 m^3 e^{4R_{\max}}}{(1-\rho)^3} + \frac{6m^2 C_\phi C_\delta e^{3R_{\max}}}{(1-\rho)^2} + \frac{mL_\delta e^{2R_{\max}}}{1-\rho} \right) \|\omega - \omega'\|_2 . \tag{190}$$

□

Proof of Lemma 15. Consider the probability $\mathbb{P}\left(\bar{s}_j, \bar{a}_j, \bar{s}_{t-k}, \bar{a}_{t-k}\right)$ and term $\mathbb{E}\left[\bar{z}_t \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta, \omega_t) | \pi_t\right]$. We have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=t-k}^{t} \phi_t^\top(\bar{s}_j, \bar{a}_j) \left(R(\bar{s}_t, \bar{a}_t) - J(\omega_t) + \phi_t^\top(\bar{s}_{t+1}, \bar{a}_{t+1})\theta - \phi_t^\top(\bar{s}_t, \bar{a}_t)\theta \right) \Big| \pi_t \right] \\
&= \mathbb{E}\left[\sum_{j=t-k}^{t} \phi_t^\top(\bar{s}_j, \bar{a}_j) \bar{\delta}_t(\bar{s}_t, \bar{a}_t; \theta, \omega_t) \Big| \pi_t \right] \\
&= \sum_{s,a} \left[\sum_{j=t-k}^{t} \mathbb{P}\left(\bar{s}_j, \bar{a}_j, \bar{s}_t, \bar{a}_t\right) \phi_t^\top(\bar{s}_j, \bar{a}_j) \delta_t(\bar{s}_t, \bar{a}_t; \theta, \omega_t) \Big| \pi_t \right] \\
&= \sum_{s,a} \left[\sum_{j=t-k}^{t} \mathbb{P}\left(\bar{s}_t, \bar{a}_t, \bar{s}_{2t-j}, \bar{a}_{2t-j}\right) \phi_t^\top(\bar{s}_t, \bar{a}_t) \delta_t(\bar{s}_{2t-j}, \bar{a}_{2t-j}; \theta, \omega_t) \Big| \pi_t \right] \\
&\overset{(a)}{=} \sum_{s,a} \left[\sum_{i=0}^{k} \mathbb{P}\left(\bar{s}_t, \bar{a}_t, \bar{s}_{t+i}, \bar{a}_{t+i})\right) \phi_t^\top(\bar{s}_t, \bar{a}_t) \delta_t(\bar{s}_{t+i}, \bar{a}_{t+i}; \theta, \omega_t) \Big| \pi_t \right] \\
&= \mathbb{E}_{(\bar{s}_t, \bar{a}_t) \sim D_t} \left[\phi_t^\top(\bar{s}_t, \bar{a}_t) \sum_{i=0}^{k} \delta_t(\bar{s}_{t+i}, \bar{a}_{t+i}; \theta, \omega_t) \Big| \pi_t \right] \\
&= \mathbb{E}_{D_t} \left[\phi_t^\top(s,a) \left(\mathcal{T}^{(k)} \left(\phi_t^\top(s,a)\theta \right) - \phi_t^\top(s,a)\theta \right) \right] \\
&= H_{\omega_t}\theta + b_{\omega_t},
\end{aligned} \tag{191}$$

where (a) follows from the fact that $\mathbb{P}\left(\bar{s}_j, \bar{a}_j\right) = \mathbb{P}\left(\bar{s}_t, \bar{a}_t\right) \sim D_t$, and thus,

$$\begin{aligned}
\mathbb{P}\left(\bar{s}_j, \bar{a}_j, \bar{s}_t, \bar{a}_t\right) &= \mathbb{P}\left(\bar{s}_t, \bar{a}_t | \bar{s}_j, \bar{a}_j\right) \mathbb{P}\left(\bar{s}_j, \bar{a}_j\right) \\
&= \mathbb{P}\left(\bar{s}_{2t-j}, \bar{a}_{2t-j} | \bar{s}_t, \bar{a}_t\right) \mathbb{P}\left(\bar{s}_t, \bar{a}_t\right) = \mathbb{P}\left(\bar{s}_t, \bar{a}_t, \bar{s}_{2t-j}, \bar{a}_{2t-j}\right).
\end{aligned} \tag{192}$$

□

Proof of Lemma 6. Recall $b_\omega = \mathbb{E}\left[\sum_{j=0}^{k} \phi_\omega^\top(s_0, a_0)(R(s_j, a_j) - J(\omega)) | (s_0, a_0) \sim D_{\pi_\omega}, \pi_\omega\right]$. Recall the definition of H_ω in Equation (11). Then, the solution to Equation (4) can be written as

$$\theta^*_\omega = -H_\omega^{-1} b_\omega. \tag{193}$$

First, b_ω can be bounded as follows:

$$\begin{aligned}
\|b_\omega\|_2 &= \left\| \mathbb{E}\left[\sum_{j=0}^{k} \phi_\omega^\top(s_0, a_0)(R(s_j, a_j) - J(\omega)) | (s_0, a_0) \sim D_{\pi_\omega}, \pi_\omega \right] \right\|_2 \\
&= \left\| \sum_{j=0}^{k} \mathbb{E}\left[\phi_\omega^\top(s_0, a_0)(R(s_j, a_j) - J(\omega)) | (s_0, a_0) \sim D_{\pi_\omega}, \pi_\omega \right] \right\|_2
\end{aligned}$$

$$= \left\| \sum_{j=0}^{k} \mathbb{E}\left[\phi_\omega^\top(s_0,a_0)\left(R(s_j,a_j) - \mathbb{E}_{D_{\pi_\omega}}\left[R(s,a)\right]\right) |(s_0,a_0) \sim D_{\pi_\omega}, \pi_\omega\right] \right\|_2$$

$$\overset{(a)}{\leq} \sum_{j=0}^{k} C_\phi R_{\max} \mathbb{E}\left[\left\| D_{\pi_\omega} - \mathbb{P}\left(s_j, a_j | s_0, a_0, \pi_\omega\right)\right\|_{\mathcal{TV}} |(s_0,a_0) \sim D_{\pi_\omega}\right]$$

$$\overset{(b)}{\leq} C_\phi R_{\max} \sum_{j=0}^{k} m\rho^k \leq \frac{C_\phi R_{\max} m}{1-\rho}, \tag{194}$$

where (a) follows from the triangular inequality and the fact that for any probability distribution P_1 and P_2, and any random variable X, s.t. $|X| \leq X_{\max}$, $\left|\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X]\right| \leq X_{\max} \left\|P_1 - P_2\right\|_{\mathcal{TV}}$, (b) follows from Assumption 1.

From the following equation:

$$\theta_\omega^{*\top} H_\omega \theta_\omega^* = \theta_\omega^{*\top} b_\omega = \left(\theta_\omega^{*\top} b_\omega\right)^\top = \theta_\omega^{*\top} H_\omega^\top \theta_\omega^*, \tag{195}$$

it holds that

$$\lambda_{\max}\left(\frac{H_\omega + H_\omega^\top}{2}\right) \left\|\theta_\omega^*\right\|_2^2 \geq \theta_\omega^{*\top} \frac{H_\omega + H_\omega^\top}{2} \theta_\omega^* = \theta_\omega^{*\top} b_\omega \geq -\left\|\theta_\omega^*\right\|_2 \left\|b_\omega\right\|_2. \tag{196}$$

Thus, we can bound θ_ω^* as follows:

$$\left\|\theta_\omega^*\right\|_2 \overset{(a)}{\leq} -\frac{1}{\lambda_{\max}\left(\frac{H_\omega + H_\omega^\top}{2}\right)} \left\|b_\omega\right\|_2 \overset{(b)}{\leq} -\frac{1}{\lambda_{\max}\left(\frac{H_\omega + H_\omega^\top}{2}\right)} \frac{mC_\phi R_{\max}}{1-\rho}$$

$$\overset{(c)}{\leq} \frac{1}{\lambda_{\min} - dC_\phi^2 m\rho^k} \frac{mC_\phi R_{\max}}{1-\rho} = \frac{mC_\phi R_{\max}}{\bar{\lambda}_{\min}(1-\rho)}, \tag{197}$$

where (a) follows from Equation (196), (b) follows from Equation (194) and (c) follows from Lemma 3. □

E. Experiments

In this section, we conduct experiments to numerically verify our AC/NAC with compatible function approximation. We test our algorithms in the Acrobot environment (Sutton, 1995). The environment involves a two-link linear chain with one end anchored and a joint that can be actuated. The goal is to apply torques at this joint to swing the unanchored end of the chain to a certain height from an initial position of hanging down. We parameterize our policy using a neural network and use compatible function approximation in the critic part. We compare the performance between our AC/NAC with compatible function approximation and the standard AC/NAC with linear function approximation.

We first compare the performance of AC algorithms. We run vanilla AC, 1-step AC with compatible function approximation, and k-step AC with compatible function approximation (shortened as AC, 1-step CAC and k-step CAC); And then we compare three NAC algorithms: vanilla NAC, 1-step NAC with compatible function approximation and k-step NAC with compatible function approximation (shortened as NAC, 1-step CNAC and k-step CNAC).

In our experiment setup, we set $k = 128$, and design a 2-layer neural network with 16 hidden neurons to represent the policy, which contains 163 parameters. We run the algorithms for 20 times. At each time step, after we obtain the policies from the algorithms, we evaluate them and plot the average reward in Figure 1 and Figure 2 among 20 runs. We also plot the 90 and 10 percentiles of the 20 curves as the upper and lower envelopes of the curves.

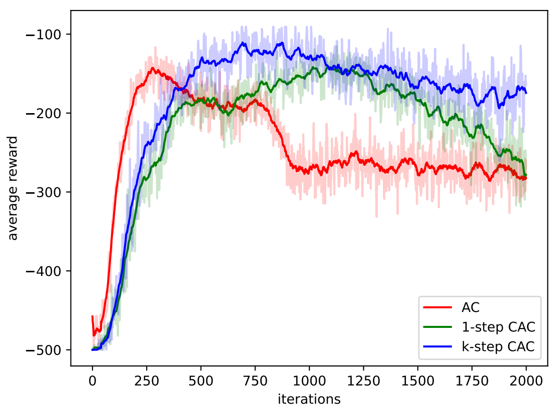

Figure 1. Vanilla AC with fixed feature function v.s. One-step AC with compatible feature function v.s. 128-step AC with compatible feature function.

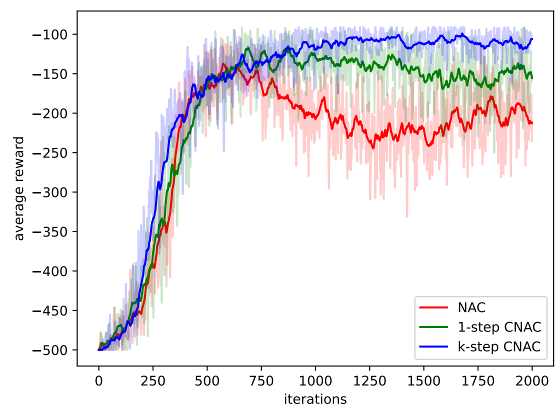

Figure 2. Vanilla NAC with fixed feature function v.s. One-step NAC with compatible feature function v.s. 128-step NAC with compatible feature function.

We observe that for both AC and NAC, our algorithms with compatible function approximation lead to better performances. As our theoretical results showed, this performance improvement is due to the fact that the compatible function approximation can avoid critic error, whereas the linear function approximation results in an inaccurate estimation of the value functions in the critic part.

F. Symbol Reference

Constants	First Appearance	
$B = \frac{R_{\max} C_\phi}{(1-\rho)\left(\lambda_{\min} - C_\phi^2 dm\rho^k\right)}$	Section 3.2	
$c_\alpha = \frac{2L_J\lambda_{\min} + 2BC_\phi^3(C_\phi L_\pi + 2L_\phi)}{\lambda_{\min}^2 \bar{\lambda}_{\min}}$	Proposition 5	expression in Equation (73)
$c_\eta = \frac{2(k+1)^2 B^2 C_\phi^6}{(\bar{\lambda}_{\min})^2}$	Proposition 5	expression in Equation (73)
$C_{\text{gap}} = C_\phi^2 B + \frac{C_\phi R_{\max}}{1-\rho}$	Proposition 7	
C_∞	Assumption 3	
C_π	Assumption 2	
C_ϕ	Assumption 2	
C_δ	Assumption 2	
L_ϕ	Assumption 2	
L_δ	Assumption 2	
$\lambda_{\min}$	Section 3.1	

Constants	First Appearance
$C_J = C_\phi^2 \left(B + C_{\text{gap}} \frac{m\rho^k}{\lambda_{\min}}\right)$	Lemma 1
$C_M = \frac{C_\Theta}{\lambda_{\min}^2} + \frac{1}{2}B^2$	Equation (115)
$C_\Theta = \frac{C_J}{\lambda_{\min}^2}\left(2C_\phi L_\phi + C_\phi^2 L_\pi\right) + \frac{L_J}{\lambda_{\min}}$	Lemma 2
$L_J = \frac{mR_{\max}}{1-\rho}\left(4L_\pi C_\phi + L_\phi\right)$	Lemma 1
$L_\pi = \frac{1}{2}C_\pi\left(1 + \lceil \log m^{-1} \rceil + \frac{1}{1-\rho}\right)$	Lemma 1
$L_\Theta = d^2\left(\frac{6C_\phi^3 m^3 e^{4R_{\max}}}{(1-\rho)^3} + \frac{6m^2 C_\phi C_\delta e^{3R_{\max}}}{(1-\rho)^2} + \frac{mL_\delta e^{2R_{\max}}}{1-\rho}\right)$	Lemma 7
$\bar{\lambda}_{\min} = \lambda_{\min} - dC_\phi^2 m\rho^k$	Lemma 10
$U_\delta = R_{\max} + 2C_\phi B$	Lemma 5

Variable	Appearance	Order (set $\alpha_t \equiv \alpha$, $\beta_t \equiv \beta$, $\gamma_t \equiv \gamma$)
G_t^δ	Lemma 5	$\mathcal{O}(k^3\alpha + k^3\beta + k(m\rho^k))$
G_t^ω	Lemma 8	$\mathcal{O}\left((m\rho^k)\beta + k^2\beta^2\right)$
G_t^η	Lemma 9	$\mathcal{O}\left((m\rho^k)\gamma + \beta^2 + k^2\beta\gamma + k\gamma^2\right)$
G_t^θ	Lemma 10	$\mathcal{O}\left(k(m\rho^k)\alpha + (m\rho^k)\beta + \frac{(m\rho^k)^2}{\beta} + k^3\alpha^2 + k^3\alpha\beta + k^2\beta^2\right)$
$\widetilde{G}_t^\eta$	Lemma 11	$\mathcal{O}\left((m\rho^k)\gamma + \beta^2 + k^2\beta\gamma + k\gamma^2\right)$
$\widetilde{G}_t^\theta$	Lemma 12	$\mathcal{O}\left(k(m\rho^k)\alpha + \frac{(m\rho^k)^2}{\beta} + k^3\alpha^2 + k^3\alpha\beta + \beta^2\right)$
$\widetilde{G}_t^\omega$	Equation (116)	$\mathcal{O}\left((m\rho^k)\beta + \beta^2)\right)$
q	Equation (68)	$\mathcal{O}\left(\alpha\right)$
$\hat{t} = \left\lceil \frac{1}{q}\log T \right\rceil$	Equation (118)	$\mathcal{O}\left(\frac{\log T}{\alpha}\right)$
$\widetilde{T} = \left\lceil \frac{T}{\hat{t}\log T} \right\rceil \hat{t}$	Equation (119)	$\mathcal{O}\left(\frac{T}{\log T}\right)$
M_t	Equation (115)	-