Finite-Time Analysis of Fully Decentralized Single-Timescale Actor-Critic

Anonymous Author(s)
Affiliation
Address
email

Abstract

Decentralized Actor-Critic (AC) algorithms have been widely utilized for multi-agent reinforcement learning (MARL) and have achieved remarkable success. Apart from its empirical success, the theoretical convergence property of decentralized AC algorithms is largely unexplored. The existing finite-time convergence results are derived based on either double-loop update or two-timescale step sizes rule, which is not often adopted in real implementation. In this work, we introduce a fully decentralized AC algorithm, where actor, critic, and global reward estimator are updated in an alternating manner with step sizes being of the same order, namely, we adopt the *single-timescale* update. Theoretically, using linear approximation for value and reward estimation, we show that our algorithm has sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ under Markovian sampling, which matches the optimal complexity with double-loop implementation (here, $\tilde{\mathcal{O}}$ hides a log term). The sample complexity can be improved to $\mathcal{O}(\epsilon^{-2})$ under the i.i.d. sampling scheme. The central to establishing our complexity results is *the hidden smoothness of the optimal critic variable* we revealed. We also provide a local action privacy-preserving version of our algorithm and its analysis. Finally, we conduct experiments to show the superiority of our algorithm over the existing decentralized AC algorithms.

1 Introduction

Multi-agent reinforcement learning (MARL) [16, 30] has been very successful in various models of multi-agent systems, such as robotics [14], autonomous driving [37], Go [25], etc. MARL has been extensively explored in the past decades; see, e.g., [18, 20, 41, 26, 8, 22]. These works either focus on the setting where an central controller is available, or assuming a common reward function for all agents. Among the many cooperative MARL settings, the work [42] proposes the fully decentralized MARL with networked agents. In this setting, each agent maintains a private heterogeneous reward function, and agents can only access local/neighboring information through communicating with its neighboring agents on the network. Then, the objective of all agents is to jointly maximize the average long-term reward through interacting with environment modeled by multi-agent Markov decision process (MDP). They proposed the decentralized Actor-Critic (AC) algorithm to solve this MARL problem, and showed its impressive performance. However, the theoretical convergence properties of such class of decentralized AC algorithms are largely unexplored; see [41] for a comprehensive survey. In this work, our goal is to establish the strong finite-time convergence results under this fully decentralized MARL setting. We first review some recent progresses on this line of research below.

Related works and motivations. The first fully decentralized AC algorithm with provable convergence guarantee was proposed by [42], and they achieved asymptotic convergence results under two-time scale step sizes, which requires actor's step sizes to diminish in a faster scale than the critic's step sizes. The sample complexities of decentralized AC were established recently. In particular, [6]

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

and [11] independently propose two communication efficient decentralized AC algorithms with optimal sample complexity of $\mathcal{O}(\varepsilon^{-2}\log(\varepsilon^{-1}))$ under Markovian sampling scheme. Their analysis are based on *double-loop* implementation, where each policy optimization step follows a nearly accurate critic optimization step (a.k.a. policy evaluation), i.e., solving the critic optimization subproblem to ε-accuracy. Such a double-loop scheme requires careful tuning of two additional hyper-parameters, which are the batch size and inner loop size. In particular, the batch size and inner loop size need to be of order $\mathcal{O}(\varepsilon^{-1})$ and $\mathcal{O}(\log(\varepsilon^{-1}))$ in order to achieve their sample complexity results, respectively. In practice, single-loop algorithmic framework is often utilized, where one updates the actor and critic in an alternating manner by performing only one algorithmic iteration for both of the two subproblems; see, e.g., [23, 18, 15, 39]. The work [38] proposes a new decentralized AC algorithm based on such a single-loop alternative update. Nevertheless, they have to adopt *two-timescale* step sizes rule to ensure convergence, which requires actor's step sizes to diminish in a faster scale than the critic's step sizes. Due to the separation of the step sizes, the critic optimization sub-problem is solved exactly when the number of iterations tends to ∞. Such a restriction on the step size will slow down the convergence speed of the algorithm. As a consequence, they only obtain sub-optimal sample complexity of $\mathcal{O}(\varepsilon^{-\frac{5}{2}})$. In practice, most algorithms are implemented with *single-timescale* step size rule, where the step sizes for actor and critic updates are of the same order. Though there are some theoretical achievements for single-timescale update in other areas such as TDC [31] and bi-level optimization [4], similar theoretical understanding under AC setting is largely unexplored.

Indeed, even when reducing to single-agent setting, the convergence property of single-timescale AC algorithm is not well established. The works [9, 10] establish the finite-time convergence result under a special single-timescale implementation, where they attain the sample complexity of $\mathcal{O}(\varepsilon^{-2})$. However, their analysis is based on an algorithm where the critic optimization step is formulated as a least-square temporal difference (LSTD) at each iteration, where they need to sample the transition tuples for $\tilde{\mathcal{O}}(\varepsilon^{-1})$ times to form the data matrix in the LSTD problem. Then, they solve the LSTD problem in a closed-form fashion, which requires to invert a matrix of large size. Later, [4] obtains the same sample complexity using TD(0) update for critic variables under i.i.d. sampling. Nonetheless, their analysis highly relies on the assumption that the Jacobian of the stationary distribution is Lipschitz continuous, which is not justified in their work.

The above observations motivate us to ask the following question:

Can we establish finite-time convergence result for decentralized AC algorithm with single-timescale step sizes rule?[1]

Main contributions. By answering this question positively, we have the following contributions:

- We design a fully decentralized AC algorithm, which employs a *single-timescale* step sizes rule and adopts Markovian sampling scheme. The proposed algorithm allows communication between agents for every K_c iterations with K_c being any integer lies in $[1, \mathcal{O}(\varepsilon^{-\frac{1}{2}})]$, rather than communicating at each iteration as adopted by previous single-loop decentralized AC algorithms [38, 42].
- Using linear approximation for value and reward estimation, we establish the *finite-time* convergence result for such an algorithm under the standard assumptions. In particular, we show that the algorithm has the sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-2})$, which matches the optimal complexity up to a logarithmic term. In addition, we show that the logarithmic term can be removed under the i.i.d. sampling scheme. Note that these convergence results are valid for all the above mentioned choices for K_c.
- To preserve the privacy of local actions, we propose a variant of our algorithm which utilizes noisy local rewards for estimating global rewards. We show that such an algorithm will maintain the optimal sample complexity at the expense of communicating at each iteration.

The underlying principle for obtaining the above convergence results is that we reveal *the hidden smoothness of the optimal critic variable*, so that we can derive an approximate descent on the averaged critic's optimal gap at each iteration. Consequently, we can resort to the classic convergence analysis for alternating optimization algorithms to establish the approximate ascent property of the overall optimization process, which leads to the final sample complexity results.

[1] As convention [9], when we use "single-timescale", it means we utilize a single-loop algorithmic framework with single-timescale step sizes rule.

Another technical highlight is the Lyapunov function we construct for measuring the progress of our algorithm. Such a construction is motivated by [4], which analyzes bi-level optimization algorithm. However, our Lyapunov function is different from theirs as it involves the additional optimal gap of averaged critic and reward estimator, which is necessary for dealing with the decentralized setting.

We finish this section by remarking that our convergence results are even new for single agent AC algorithms under the setting of single-timescale step sizes rule.

2 Preliminary

In this section, we introduce the problem formulation and the policy gradient theorem, which serves as the preliminary for the analyzed decentralzed AC algorithm.

Suppose there are multiple agents aiming to independently optimize a common global objective, and each agent can communicate with its neighbors through a network. To model the topology, we define the graph as $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes with $|\mathcal{N}| = N$ and $\mathcal{E}$ is the set of edges with $|\mathcal{E}| = E$. In the graph, each node represents an agent, and each edge represents a communication link. The interaction between agents follows the networked multi-agent MDP.

2.1 Markov decision process

A networked multi-agent MDP is defined by a tuple $(\mathcal{G}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \mathcal{P}, \{r^i\}_{i\in[N]}, \gamma)$. $\mathcal{G}$ denotes the communication topology (the graph), $\mathcal{S}$ is the finite state space observed by all agents, $\mathcal{A}^i$ represents the finite action space of agent i. Let $\mathcal{A} := \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$ denote the joint action space and $\mathcal{P}(s'|s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ denote the transition probability from any state $s \in \mathcal{S}$ to any state $s' \in \mathcal{S}$ for any joint action $a \in \mathcal{A}$. $r^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the local reward function that determines the reward received by agent i given transition (s,a); $\gamma \in [0,1]$ is the discount factor.

For simplicity, we will use $a := [a^1, \cdots, a^N]$ to denote the joint action, and $\theta := [\theta^1, \cdots, \theta^N] \in \mathbb{R}^{d_\theta \times N}$ to denote joint parameters of all actors, with $\theta^i \in \mathbb{R}^{d_\theta}$. Note that different actors may have different number of parameters, which is assumed to be the same for our paper without loss of generality. The MDP goes as follows: For a given state s, each agent make its decision a^i based on its policy $a^i \sim \pi_{\theta^i}(\cdot|s)$. The state transits to the next state s' based on the joint action of all the agents: $s' \sim \mathcal{P}(\cdot|s,a)$. Then, each agent will receive its own reward $r^i(s,a)$. For the notation brevity, we assume that the reward function mapping is deterministic and does not depend on the next state without loss of generality. The stationary distribution induced by the policy π_θ and the transition kernel is denoted by $\mu_{\pi_\theta}(s)$.

Our objective is to find a set of policies that maximize the accumulated discounted mean reward received by agents

$$\theta^* = \arg\max_{\theta} J(\theta) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \bar{r}(s_k, a_k)\right]. \tag{1}$$

Here, k represents the time step. $\bar{r}(s_k, a_k) := \frac{1}{N}\sum_{i=1}^{N} r^i(s_k, a_k)$ is the mean reward among agents at time step k. The randomness of the expectation comes from the initial state distribution $\mu_0(s)$, the transition kernel $\mathcal{P}$, and the stochastic policy $\pi_{\theta^i}(\cdot|s)$.

2.2 Policy gradient Theorem

Under the discounted reward setting, the global state-value function, action-value function, and advantage function for policy set θ, state s, and action a, are defined as

$$\begin{aligned} V_{\pi_\theta}(s) &:= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \bar{r}(s_k, a_k) | s_0 = s\right] \\ Q_{\pi_\theta}(s,a) &:= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \bar{r}(s_k, a_k) | s_0 = s, a_0 = a\right] \\ A_{\pi_\theta}(s,a) &:= Q_{\pi_\theta}(s,a) - V_{\pi_\theta}(s). \end{aligned} \tag{2}$$

To maximize the objective function defined in (1), the policy gradient [28] can be computed as follow

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim d_{\pi_\theta}, a\sim\pi_\theta}\left[\frac{1}{1-\gamma}A_{\pi_\theta}(s,a)\psi_{\pi_\theta}(s,a)\right],$$

where $d_{\pi_\theta}(s) := (1-\gamma)\sum_{k=0}^{\infty}\gamma^k \mathbb{P}(s_k = s)$ is the discounted state visitation distribution under policy π_θ, and $\psi_{\pi_\theta}(s,a) := \nabla \log \pi_\theta(s,a)$ is the score function.

Following the derivation of [42], the policy gradient for each agent under discounted reward setting can be expressed as

$$\nabla_{\theta^i} J(\theta) = \mathbb{E}_{s\sim d_{\pi_\theta}, a\sim\pi_\theta}\left[\frac{1}{1-\gamma}A_{\pi_\theta}(s,a)\psi_{\pi_{\theta^i}}(s,a^i)\right]. \tag{3}$$

3 Decentralized single-timescale actor-critic

Algorithm 1: Decentralized single-timescale AC (reward estimator version)

Initialize: Actor parameter θ_0, critic parameter ω_0, reward estimator parameter λ_0, initial state s_0.
for $k = 0, \cdots, K-1$ **do**
  Option 1: i.i.d. sampling:
  $s_k \sim \mu_{\theta_k}(\cdot), a_k \sim \pi_{\theta_k}(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$.
  Option 2: Markovian sampling:
  $a_k \sim \pi_{\theta_k}(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$.

  Periodical consensus: Compute $\tilde{\omega}_k^i$ and $\tilde{\lambda}_k^i$ by (4) and (7).

  for $i = 0, \cdots, N$ **in parallel do**
    Reward estimator update: Update λ_{k+1}^i by (8).
    Critic update: Update ω_{k+1}^i by (5).
    Actor update: Update θ_{k+1}^i by (6).
  end for
end for

We introduce the decentralized single-timescale AC algorithm; see Algorithm 1. In the remaining parts of this section, we will explain the updates in the algorithm in details.

In fully-decentralized MARL, each agent can only observe its local reward and action, while trying to maximize the global reward (mean reward) defined in (1). The decentralized AC algorithm solves the problem by performing online updates in an alternative fashion. Specifically, we have N pairs of actor and critic. In order to maximize $J(\theta)$, each critic tries to estimate the *global* state-value function $V_{\pi_\theta}(s)$ defined in (2), and each actor then updates its policy parameter based on approximated policy gradient. We now provide more details about the algorithm.

Critics' update. We will use $\omega^i \in \mathbb{R}^{d_\omega}$ to denote the i_{th} critic's parameter and $\bar{\omega} := \frac{1}{N}\sum_{i=1}^N \omega^i$ to represent the averaged parameter of critic. The i_{th} critic approximates the global value function as $V_{\pi_\theta}(s) \approx \hat{V}_{\omega^i}(s)$.

As we will see, the critic's approximation error can be categorized into two parts, namely, the consensus error $\frac{1}{N}\sum_{i=1}^N \|\omega^i - \bar{\omega}\|$, which measures how close the critics' parameters are; and the approximation error $\|\bar{\omega} - \omega^*(\theta)\|$, which measures the approximation quality of averaged critic.

In order for critics to reach consensus, we perform the following update for all critics

$$\tilde{\omega}_k^i = \begin{cases} \sum_{j=1}^N W^{ij}\omega_k^j & \text{if } k \mod K_c = 0 \\ \omega_k^i & \text{otherwise.} \end{cases} \tag{4}$$

where $W \in \mathbb{R}^{n\times n}$ is a weight matrix for communication among agents, whose property will be specified in Assumption 5; K_c denotes the consensus frequency.

To reduce the approximation error, we will perform the local TD(0) update [29] as

$$\omega_{k+1}^i = \Pi_{R_\omega}(\tilde{\omega}_k^i + \beta_k g_c^i(\xi_k, \omega_k^i)), \tag{5}$$

where $\xi := (s, a, s')$ represents a transition tuple, $g_c^i(\xi, \omega) := \delta^i(\xi, \omega)\nabla\hat{V}_\omega(s)$ is the update direction, $\delta^i(\xi, \omega) := r^i(s, a) + \gamma\hat{V}_\omega(s') - \hat{V}_\omega(s)$ is the local temporal difference error (TD-error). β_k is the step size for critic at iteration k. Π_{R_ω} projects the parameter into a ball of radius of R_ω containing the optimal solution, which will be explained when discussing Assumption 1 and 2.

Actors' update. We will use stochastic gradient ascent to update the policy's parameter, and the stochastic gradient is calculated based on policy gradient theorem in (3). The advantage function $A_{\pi_\theta}(s, a)$ can be estimated by

$$\delta(\xi, \theta) := \bar{r}(s, a) + \gamma V(s') - V(s),$$

with a sampled from $\pi_\theta(\cdot|s)$. However, to preserve the privacy of each agents, the local reward cannot be shared to other agents under the fully decentralized setting. Thus, the averaged reward $\bar{r}(s_k, a_k)$ is not directly attainable. Consequently, we need a strategy to approximate the averaged reward. In this paper, we will adopt the strategy proposed in [42]. In particular, each agent i will have a local reward estimator with parameter $\lambda^i \in \mathbb{R}^{d_\lambda}$, which estimates the global averaged reward as $\bar{r}(s_k, a_k) \approx \hat{r}_{\lambda^i}(s_k, a_k)$.

Thus, the update of the i_{th} actor is given by

$$\theta_{k+1}^i = \theta_k^i + \alpha_k\hat{\delta}(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\psi_{\pi_{\theta_k^i}}(s_k, a_k^i), \tag{6}$$

where $\hat{\delta}(\xi, \omega, \lambda) := \hat{r}_\lambda(s, a) + \gamma\hat{V}_\omega(s') - \hat{V}_\omega(s)$ is the approximated advantage function. α_k is the step size for actor's update at iteration k.

Reward estimators' update. Similar to critic, each reward estimator's approximation error can be decomposed into consensus error and the approximation error.

For each local reward estimator, we perform the consensus step to minimize the consensus error as

$$\tilde{\lambda}_k^i = \begin{cases} \sum_{j=1}^N W^{ij}\lambda_k^j & \text{if } k \bmod K_c = 0 \\ \lambda_k^i & \text{otherwise.} \end{cases} \tag{7}$$

To reduce the approximation error, we perform a local update of stochastic gradient descent.

$$\lambda_{k+1}^i = \Pi_{R_\lambda}(\tilde{\lambda}_k^i + \eta_k g_r^i(\xi_k, \lambda_k^i)), \tag{8}$$

where $g_r^i(\xi, \lambda) := (r^i(s, a) - \hat{r}_\lambda(s, a))\nabla\hat{r}_\lambda(s, a)$ is the update direction. η_k is the step size for reward estimator at iteration k. Note the calculation of $g_r^i(\xi, \lambda)$ does not require the knowledge of s'; we use ξ in (8) just for notation brevity. Similar to critic's update, Π_{R_λ} projects the parameter into a ball of radius of R_λ containing the optimal solution.

In our Algorithm 1, we will use the same order for α_k, β_k, and η_k and hence, our algorithm is in *single-timescale*.

Linear approximation for analysis. In our analysis, we will use linear approximation for both critic and reward estimator variables, i.e. $\hat{V}_\omega(s) := \phi(s)^T\omega; \hat{r}_\lambda(s, a) := \varphi(s, a)^T\lambda$, where $\phi(s) : \mathcal{S} \to \mathbb{R}^{d_\omega}$ and $\varphi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_\lambda}$ are two feature mappings, whose property will be specified in the discussion of Assumption 1.

Algorithm for preserving the local action. Note that in Algorithm 1, the reward estimators need the knowledge of joint actions in order to estimate the global rewards. To preserve the privacy of local actions, we further propose a variant of Algorithm 1, which estimates the global rewards by communicating noisy local rewards; see [6] for the original idea. However, to maintain the optimal sample complexity, such an approach requires $\mathcal{O}(\log(\varepsilon^{-1}))$ communication rounds for each iteration. We postpone the detailed design and analysis of such an algorithm scheme into Appendix B.

Remarks on sampling scheme. The unbiased update for critic and actor variables requires sampling from μ_{π_θ} and d_{π_θ}, respectively. However, in practical implementations, states are usually collected from an online trajectory (Markovian sampling), whose distribution is generally different for μ_{π_θ} and d_{π_θ}. Such a distribution mismatch will inevitably cause biases during the update of critic and actor variables. One has to bound the corresponding error terms when analyzing the algorithm. In this work, we will provide the analysis for both sampling schemes.

4 Main Results

In this section, we first introduce the technical assumptions used for our analysis, which are standard in the literature. Then, we present the convergence results for both actor and critic variables under i.i.d. sampling and Markovian sampling.

4.1 Assumptions

Assumption 1 (bounded rewards and feature vectors)**.** *All the local rewards are uniformly bounded, i.e., there exists a positive constants* $r_{\max}$ *such that* $|\, r^i(s,a)\,| \le r_{\max}$*, for all feasible* (s,a) *and* $i \in [N]$*. The norm of feature vectors are bounded such that for all* $s \in \mathcal{S},\ a \in \mathcal{A}$*,* $\|\phi(s)\| \le 1, \|\varphi(s,a)\| \le 1$*.*

Assumption 1 is standard and commonly adopted; see, e.g., [3, 35, 38, 24, 21]. This assumption can be achieved via normalizing the feature vectors.

Assumption 2 (negative definiteness of $A_{\theta,\phi}$ and $A_{\theta,\varphi}$)**.** *There exists two positive constants* $\lambda_\phi, \lambda_\varphi$ *such that for all policy* θ*, the following two matrices are negative definite*

$$A_{\theta,\phi} := \mathbb{E}_{s\sim\mu_\theta(s)}[\phi(s)(\gamma\phi(s')^T - \phi(s)^T)]$$
$$A_{\theta,\varphi} := \mathbb{E}_{s\sim\mu_\theta(s),a\sim\pi_\theta(\cdot|s)}[-\varphi(s,a)\varphi(s,a)^T],$$

with $\lambda_{\max}(A_{\theta,\phi}) \le \lambda_\phi, \lambda_{\max}(A_{\theta,\varphi}) \le \lambda_\varphi$*, where* $\lambda_{\max}(\cdot)$ *represents the largest eigenvalue.*

Assumption 2 can be achieved when the matrices $\Phi_\phi := [\phi(s_1), \cdots, \phi(s_{|\mathcal{S}|})]$ and $\Phi_\varphi := [\varphi(s_1,a_1), \cdots, \varphi(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})]$ have full row rank, which ensures that the optimal critic and reward estimator are unique; see also [24, 34]. Together with Assumption 1, we can show that the norm of $\omega^*(\theta)$ and $\lambda^*(\theta)$ are bounded by some positive constant, which justifies the projection steps.

Assumption 3 (Lipschitz properties of policy)**.** *There exists constants* C_ψ, L_ψ, L_π *such that for all* $\theta, \theta', s \in \mathcal{S}$ *and* $a \in \mathcal{A}$*, we have* $(1).\ |\pi_\theta(a|s) - \pi_{\theta'}(a|s)| \le L_\pi\|\theta - \theta'\|; (2).\ \|\psi_\theta(s,a) - \psi_{\theta'}(s,a)\| \le L_\psi\|\theta - \theta'\|; (3).\ \|\psi_\theta(s,a)\| \le C_\psi$*.*

Assumption 3 is common for analyzing policy-based algorithms; see, e.g., [33, 32, 11]. The assumption ensures the smoothness of objective function $J(\theta)$. It holds for a large range of policy classes such as tabular softmax policy [1], Gaussian policy [7], and Boltzman policy [13].

Assumption 4 (irreducible and aperiodic Markov chain)**.** *The Markov chain under* π_θ *and transition kernel* $\mathcal{P}(\cdot|s,a)$ *is irreducible and aperiodic for any* θ*.*

Assumption 4 is a standard assumption, which holds for any uniformly ergodic Markov chains and any time-homogeneous Markov chains with finite-state space. It ensures that there exists constants $\kappa > 0$ and $\rho \in (0,1)$ such that

$$\sup_{s\in\mathcal{S}} d_{TV}(\mathbb{P}(s_k \in \cdot|s_0 = s, \pi_\theta), \mu_\theta) \le \kappa\rho^k,\ \forall k.$$

Assumption 5 (doubly stochastic weight matrix)**.** *The communication matrix* W *is doubly stochastic, i.e. each column/row sum up to 1. Moreover, the second largest singular value* ν *is smaller than 1.*

Assumption 5 is a common assumption in decentralized optimization and multi-agent reinforcement learning; see, e.g., [27, 5, 6]. It ensures the convergence of consensus error for critic and reward estimator variables.

4.2 Sample complexity under i.i.d. sampling

Theorem 1 (sample complexity under i.i.d. sampling)**.** *Suppose Assumptions 1-5 hold. Consider the update of Algorithm 1 under i.i.d. sampling. Let* $\alpha_k = \frac{\bar{\alpha}}{\sqrt{K}}$ *for some positive constant* $\bar{\alpha}$*,* $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k$*, and* $\eta_k = \frac{C_{10}}{2\lambda_\varphi}\alpha_k$*,* $K_c \le \mathcal{O}(K^{1/4})$*, where* K *denotes the total number of iterations. Then, we have*

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\left[\|\omega_k^i - \omega^*(\theta_k)\|^2\right] \le \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$$
$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta^i} F(\theta_k)\|^2\right] \le \mathcal{O}\left(\frac{1}{\sqrt{K}}\right) + \mathcal{O}(\varepsilon_{app} + \varepsilon_{sp}), \quad (9)$$

where C_9, C_{10} are positive constants defined in the proof.

The proof of Theorem 1 can found in Appendix E.1. It establishes the iteration complexity of $\mathcal{O}(1/\sqrt{K})$, or equivalently, sample complexity of $\mathcal{O}(\varepsilon^{-2})$ for Algorithm 1. Note that actors, critics, and reward estimators use the step sizes of the same order. The sample complexity matches the optimal rate of SGD for general non-convex optimization problem. To explain the errors in (9), let us define the approximation error as the following:

$$\varepsilon_{app} := \max_{\theta,a} \sqrt{\mathbb{E}_{s\sim\mu_\theta}\left[|V_{\pi_\theta}(s) - \hat{V}_{\omega^*(\theta)}(s)|^2 + |\bar{r}(s,a) - \hat{r}_{\lambda^*(\theta)}(s,a)|^2\right]}.$$

The error ε_{app} captures the approximation power of critic and reward estimator. Similar terms also appear in the literature (see e.g., [35, 1, 21]). Such an approximation error becomes zero in tabular case. The error ε_{sp} is inevitably caused by the mismatch between discounted state visitation distribution d_{π_θ} and stationary distribution μ_{π_θ}; see, e.g., [38, 24]. It is defined as

$$\varepsilon_{sp} := 2C_\theta(\log_\rho \kappa^{-1} + \frac{1}{\rho})(1-\gamma).$$

When γ is close to 1, the error becomes small. This is because d_{π_θ} approaches to μ_{π_θ} when γ goes to 1. In the literature, some works assume that sampling from d_{π_θ} is permitted, thus eliminate this error; see, e.g., [4].

4.3 Sample complexity under markovian sampling

Theorem 2 (sample complexity under Markovian sampling)**.** *Suppose Assumptions 1-5 hold. Consider the update of Algorithm 1 under Markovian sampling. Let $\alpha_k = \frac{\bar{\alpha}}{\sqrt{K}}$ for some positive constant $\bar{\alpha}$, $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k$, and $\eta_k = \frac{C_{10}}{2\lambda_\varphi}\alpha_k$, $K_c \leq \mathcal{O}(K^{1/4})$, where K is the total number of iterations. Then, we have*

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\left[\|\omega_k^i - \omega^*(\theta_k)\|^2\right] \leq \mathcal{O}\left(\frac{\log^2 K}{\sqrt{K}}\right)$$
$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta^i} F(\theta_k)\|^2\right] \leq \mathcal{O}\left(\frac{\log^2 K}{\sqrt{K}}\right) + \mathcal{O}(\varepsilon_{app} + \varepsilon_{sp}), \tag{10}$$

where C_9, C_{10} are positive constants defined in proof.

We put the proof of Theorem 2 in Appendix E.2. In Markovian sampling, the updates are biased for critics, actors, and reward estimators. The error will decrease as the Markov chain mixes, and the logarithmic term is due to the cost for mixing.

Theorem 2 establishes the iteration complexity of $\mathcal{O}(\log^2 K/\sqrt{K})$, or equivalently, sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ for Algorithm 1. It matches the state-of-the-art sample complexity of decentralized AC algorithms, which are implemented in double-loop fashion [11, 6].

4.4 Proof sketch

We present the main elements for the proof of Theorem 2, which helps in understanding the difference between classical two-timescale/double-loop analysis and our single-timescale analysis. The proof of Theorem 1 follows the same framework with simpler sampling scheme.

Under Markovian sampling, it is possible to show the following inequality, which characterizes the ascent of the objective.

$$\begin{aligned}\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) \geq \sum_{i=1}^{N}\Big[&\frac{\alpha_k}{2}\mathbb{E}\|\nabla_{\theta^i} J(\theta_k)\|^2 + \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\|^2 \\ &- 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 - 4C_\psi^2\alpha_k\mathbb{E}\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2\Big] \\ &- \mathcal{O}(\log^2(K)\alpha_k^2) - \mathcal{O}((\varepsilon_{app} + \varepsilon_{sp})\alpha_k).\end{aligned} \tag{11}$$

To analyze the errors of critic $\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2$ and reward estimator $\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2$, the two-timescale analysis requires $\mathcal{O}(\alpha_k) < \min\{\mathcal{O}(\beta_k), \mathcal{O}(\eta_k)\}$ in order for these two errors to converge. The double-loop approach runs lower-level update for $\mathcal{O}(\log(\varepsilon^{-1}))$ times with batch size $\mathcal{O}(\varepsilon^{-1})$ to drive these errors below ε and hence, they cannot allow inner loop size and bath size to be $\mathcal{O}(1)$ simultaneously. To obtain the convergence result for *single-timescale* update, the idea is to further upper bound these two lower-level errors by the quantity $\mathcal{O}(\alpha_k \mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\|^2)$ (through a series of derivations), and then eliminate these errors by the ascent term $\frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\|^2$.

We mainly focus on the analysis of critic's error through the proof sketch. The analysis for reward estimator's error follows similar procedure. We start by decomposing the error of critic as

$$\sum_{i=1}^{N} \|\omega_{k+1}^i - \omega^*(\theta_k)\|^2 = \sum_{i=1}^{N} (\|\omega_{k+1}^i - \bar{\omega}_{k+1}\|^2 + \|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2). \tag{12}$$

The first term represents the consensus error, which can be bounded by the next lemma.

Lemma 1. *Suppose Assumptions 1 and 5 hold. Consider the sequence $\{\omega_k^i\}$ generated by Algorithm 1, then the following holds*

$$\|Q\boldsymbol{\omega}_{k+1}\| \le \nu^{\frac{k'}{K_c}} \|\boldsymbol{\omega}_0\| + 4\sum_{t=0}^{k} \nu^{\lceil \frac{k'-1-t}{K_c} \rceil} \beta_t \sqrt{N} C_\delta,$$

where $\boldsymbol{\omega}_0 := [\omega^1, \cdots, \omega^N]^T, Q := I - \frac{1}{N}\mathbf{1}\mathbf{1}^T, k' := \lfloor \frac{k}{K_c} \rfloor * K_c$. The constant $\nu \in (0,1)$ is the second largest singular value of W.

Based on Lemma 1 and follow the step size rule of Theorem 2, it is possible to show $\|Q\boldsymbol{\omega}_{k+1}\|_F^2 = \sum_{i=1}^N \|\omega_{k+1}^i - \bar{\omega}_{k+1}\|^2 = \mathcal{O}(K_c^2 \beta_k^2)$. Let $K_c = \mathcal{O}(\beta_k^{-\frac{1}{2}})$, we have $\|Q\boldsymbol{\omega}_{k+1}\|_F^2 = \mathcal{O}(\beta_k)$, which maintains the optimal rate.

To analyze the second term in (12), we first construct the following Lyapunov function

$$\mathbb{V}_k := -J(\theta_k) + \|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + \|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2. \tag{13}$$

Then, it remains to derive an approximate descent property of the term $\|\bar{\omega}_k - \omega^*(\theta_k)\|^2$ in (13). Towards that end, our key step lies in establishing the *smoothness of the optimal critic variables* shown in the next lemma.

Lemma 2 (smoothness of optimal critic)**.** *Suppose Assumptions 1-3 hold, under the update of Algorithm 1, there exists a positive constant $L_{\mu,1}$ such that for all θ, θ', it holds that*

$$\|\nabla\omega^*(\theta) - \nabla\omega^*(\theta')\| \le L_{\mu,1}\|\theta - \theta'\|,$$

where $\nabla\omega^*(\theta)$ denotes the Jacobian of $\omega^*(\theta)$ with respect to θ.

This smoothness property is essential for achieving our $\tilde{\mathcal{O}}(1/\sqrt{K})$ convergence rate.

To the best of our knowledge, the smoothness of $\omega^*(\theta)$ has not been justified in the literature. Equipped with Lemma 2, we are able to establish the following lemma.

Lemma 3 (Error of critic)**.** *Under Assumptions 1-5, consider the update of Algorithm 1. Then, it holds that*

$$\begin{aligned}\mathbb{E}[\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2] &\le (1 + C_9\alpha_k)\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 \\ &\quad + \frac{\alpha_k}{4}\sum_{i=1}^{N} \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \mathcal{O}(\alpha_k^2).\end{aligned} \tag{14}$$

$$\begin{aligned}\mathbb{E}[\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2] &\le (1 - 2\lambda_\phi\beta_k)\|\bar{\omega}_k - \omega^*(\theta_k)\|^2 \\ &\quad + C_{K_1}\beta_k\beta_{k-Z_K} + C_{K_2}\alpha_{k-Z_K}\beta_k.\end{aligned} \tag{15}$$

Here, $Z_K := \min\{z \in \mathbb{N}^+ | \kappa\rho^{z-1} \le \min\{\alpha_k, \beta_k, \eta_k\}\}$, C_9, λ_ϕ are constants specified in appendix, and C_{K_1} and C_{K_2} are of order $\mathcal{O}(\log(K))$ and $\mathcal{O}(\log^2(K))d$ respectively.

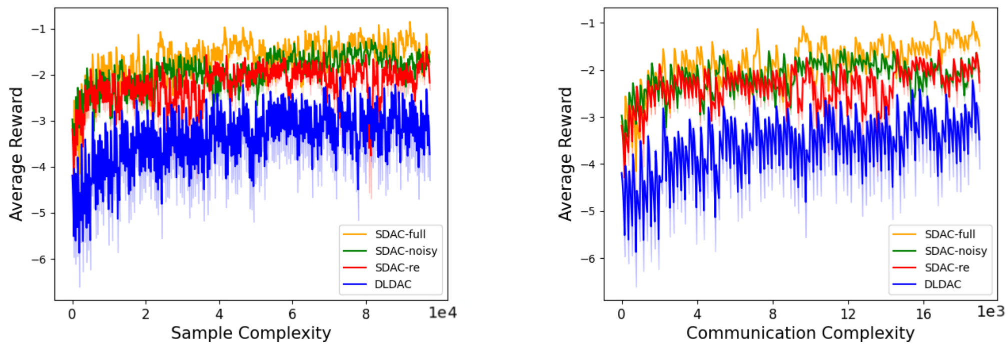


Figure 1: Averaged reward versus sample complexity and communication complexity. The vertical axis is the averaged reward over all the agents.

Plug (15) into (14), we can establish the approximate descent property of $\|\bar{\omega}_k - \omega^*(\theta_k)\|^2$ in (13):

$$\begin{aligned}\mathbb{E}[\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2] \le (1 + C_9\alpha_k)(1 - 2\lambda_\phi\beta_k)\|\bar{\omega}_k - \omega^*(\theta_k)\|^2 \\ + \frac{\alpha_k}{4}\sum_{i=1}^{N}\|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 \\ + \mathcal{O}(C_{K_1}\beta_k\beta_{k-Z_K} + C_{K_2}\alpha_{k-Z_K}\beta_k). \qquad (16)\end{aligned}$$

Finally, plugging (11), (14), and (16) into (13) gives the ascent of the Lyapunov function, which leads to our convergence result through steps of standard arguments.

5 Numerical results

In this section, our objective is to illustrate the empirical sample complexity and communication complexity of the proposed algorithms. We also implement the algorithm in [6] to serve as a baseline, which employs double-loop algorithmic framework. Our simulation is based on the grounded communication environment proposed in [19]; see Appendix A for detailed set up. Through the discussion, we refer the algorithm in [6] as "DLDAC", the Algorithm 1 as "SDAC-re", the Algorithm 2 as "SDAC-noisy" (see Appendix B). We also provide the result which assumes full reward is available to serve as baseline, which we refer as "SDAC-full". We set $K_r = 5$ for "SDAC-noisy"; $K_c = 1$ for "SDAC-re", "SDAC-noisy", and "SDAC-full". We choose $T_c = 5$ (loop size), $T_c' = 1$ (critic consensus number every iteration), $T' = 5$ (reward consensus number every iteration) for "DLDAC".

The sample complexity and communication complexity are shown in Figure 1. The results are averaged over 10 Monte Carlo runs. As we can see, the proposed two algorithms achieve significantly higher reward than "DLDAC" in terms of both sample complexity and communication complexity. Moreover, their performances approach the baseline "SDAC-full", where the global reward is assumed to be available, indicating that the reward approximation is nearly accurate. Due to space limit, we will put additional experiments on the comparison with existing decentralized AC algorithms and the ablation study of hyper-parameters to Appendix A.

6 Conclusion and future direction

In this paper, we studied the convergence of fully decentralized AC algorithm under practical single-timescale update for the first time. We designed such an algorithm which maintains the optimal sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ under less communications. We also proposed a variant to preserve the privacy of local actions by communicating noisy rewards. Extensive simulation results demonstrate the superiority of our algorithms' empirical performance over existing decentralized AC algorithms. One limitation of our work is that we only study the convergence to stationary point. Thus, we leave the research on the avoidance of saddle points and convergence to global optimum as promising future directions.

References

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

[2] J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[3] J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory (COLT)*, pages 1691–1692, 2018.

[4] T. Chen, Y. Sun, and W. Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[5] Z. Chen, Y. Zhou, and R. Chen. Multi-agent off-policy td learning: Finite-time analysis with near-optimal sample complexity and communication complexity. *arXiv preprint arXiv:2103.13147*, 2021.

[6] Z. Chen, Y. Zhou, R.-R. Chen, and S. Zou. Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis. In *International Conference on Machine Learning*, pages 3794–3834. PMLR, 2022.

[7] K. Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1):219–245, 2000.

[8] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pages 1407–1416. PMLR, 2018.

[9] Z. Fu, Z. Yang, and Z. Wang. Single-timescale actor-critic provably finds globally optimal policy. In *International Conference on Learning Representations*, 2021.

[10] H. Guo, Z. Fu, Z. Yang, and Z. Wang. Decentralized single-timescale actor-critic on zero-sum two-player stochastic games. In *International Conference on Machine Learning*, pages 3899–3909. PMLR, 2021.

[11] F. Hairi, J. Liu, and S. Lu. Finite-time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward. In *International Conference on Learning Representations*, 2022.

[12] S. M. Kakade. A natural policy gradient. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1531–1538, 2002.

[13] V. R. Konda and V. S. Borkar. Actor-critic–type learning algorithms for Markov decision processes. *SIAM Journal on Control and Optimization*, 38(1):94–123, 1999.

[14] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[15] Y. Lin, K. Zhang, Z. Yang, Z. Wang, T. Başar, R. Sandhu, and J. Liu. A communication-efficient multi-agent actor-critic algorithm for distributed reinforcement learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 5562–5567, 2019.

[16] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

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

[18] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[19] I. Mordatch and P. Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR, 2017.

[21] S. Qiu, Z. Yang, J. Ye, and Z. Wang. On the finite-time convergence of actor-critic algorithm. In *Optimization Foundations for Reinforcement Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[22] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4295–4304. PMLR, 2018.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[24] H. Shen, K. Zhang, M. Hong, and T. Chen. Asynchronous advantage actor critic: Non-asymptotic analysis and linear speedup. *ArXiv:2012.15511*, 2020.

[25] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[26] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5887–5896. PMLR, 2019.

[27] J. Sun, G. Wang, G. B. Giannakis, Q. Yang, and Z. Yang. Finite-sample analysis of decentralized temporal-difference learning with linear function approximation. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4485–4495, 2020.

[28] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1057–1063, 2000.

[29] J. N. Tsitsiklis and B. Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems (NIPS)*, pages 1075–1081, 1997.

[30] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[31] Y. Wang, S. Zou, and Y. Zhou. Non-asymptotic analysis for two time-scale TDC with general smooth function approximation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9747–9758. Curran Associates, Inc., 2021.

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

[33] P. Xu, F. Gao, and Q. Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

[34] T. Xu and Y. Liang. Sample complexity bounds for two timescale value-based reinforcement learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 811–819. PMLR, 2021.

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

[36] T. Xu, Z. Yang, Z. Wang, and Y. Liang. Doubly robust off-policy actor-critic: Convergence and optimality. *ArXiv:2102.11866*, 2021.

[37] C. Yu, X. Wang, X. Xu, M. Zhang, H. Ge, J. Ren, L. Sun, B. Chen, and G. Tan. Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs. *IEEE Transactions on Intelligent Transportation Systems*, 21(2):735–748, 2019.

[38] S. Zeng, T. Chen, A. Garcia, and M. Hong. Learning to coordinate in multi-agent systems: A coordinated actor-critic algorithm and finite-time guarantees. *arXiv preprint arXiv:2110.05597*, 2021.

[39] H. Zhang, W. Chen, Z. Huang, M. Li, Y. Yang, W. Zhang, and J. Wang. Bi-level actor-critic for multi-agent coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7325–7332, 2020.

[40] K. Zhang, A. Koppel, H. Zhu, and T. Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*, 2019.

[41] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

[42] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5872–5881. PMLR, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The limitation is written in an equivalent form as future works in the conclusion section; see Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We conduct research about the design and analysis of the fundamental actor-critic algorithm, which should not bring any negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

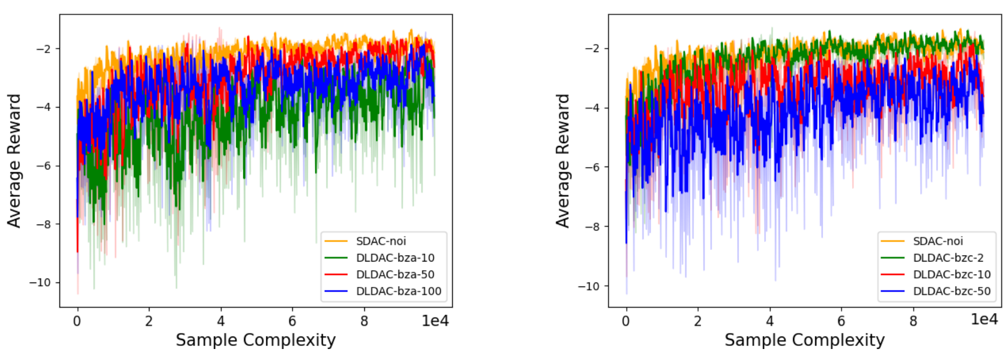
A Experiment settings and additional simulation results

In this section, we first introduce the experimental setting. Then, we present more experiments on the comparison between the proposed algorithms and existing decentralized AC algorithms. Additionally, we conduct ablation study on different consensus frequencies of the proposed algorithm.

Experiment setting. We adopt the grounded communication environment proposed in [19]. Our task consists of N agents and the corresponding N landmarks inhabited in a two-dimension world, where each agent can observe the relative position of other agents and landmarks. For every discrete time step, agents take actions to move along certain directions, and receive their rewards. Agents are rewarded based on the distance to their own landmark, and penalized if they collide with other agents. The objective is to maximize the long-term averaged reward over all agents. Since we focus on decentralized setting, each agent shall not know the target landmark of others, i.e., the reward function of others. To exchange information, each agent is allowed to send their local information via a fixed communication link. Through all the experiments, the agent number N is set to be 5, and the discount factor γ is set to be 0.95.

Comparison to double-loop decentralized AC under mini-batch update. Since the algorithm in [6] uses mini-batch update to reduce the variance during the update, we will compare the proposed algorithms with [6] under different choices of actor's batch sizes, critic's batch sizes, and inner loop sizes, respectively. Since their algorithm communicates noisy reward to achieve consensus, we will use "SDAC-noi" to serve as baseline.

1. **Actor's batch size.** We fix $T_c = 50$, $T_c' = 10$, $N_c = 10$, [2] which is adopted by [6]. We examine values of N in $\{10, 50, 100\}$. The results are in Figure 2a. We observe that the best choice of actor's batch size N is 50, and the proposed "SDAC-noi" converges faster than it in terms of sample complexity.
2. **Critic's batch size.** We fix $T_c = 50$, $T_c' = 10$, $N = 100$, which is adopted by [6]. We examine values of N_c in $\{2, 10, 50\}$. The results are shown in Figure 2b. As we can see, "DLDAC" with smaller critic's batch sizes can achieve better sample complexity, indicating that the variance of critic's update is relatively small and the mini-batch update is not needed for this task. Our proposed "SDAC-noi" achieves better convergence compared with the double-loop decentralized AC under different choices of N_c.
3. **Inner loop size.** We fix $T_c' = 10$, $N = 100$, $N_c = 10$, which is adopted by [6]. We examine values of T_c in $\{5, 20\}$. The results are shown in Figure 3. We can see that the proposed "SDAC-noi" enjoys a better convergence in terms of sample complexity.



(a) Different actor's batch sizes.　　(b) Different critic's batch sizes.

Figure 2: Comparison between the proposed algorithms and the double-loop decentralized AC algorithm that uses mini-batch update. The results are averaged over 10 Monte Carlo runs.

[2]Note that we adopt the notations in [6]. Here, T_c is the inner loop size, T_c' is the communication number for each outer loop, N is the batch size for actor's update, and N_c is the batch size for critic's update.

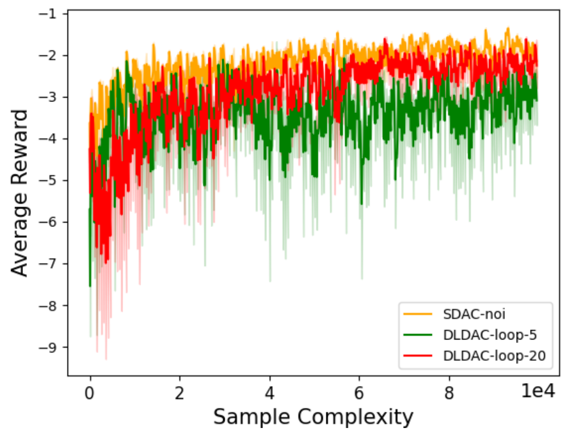


Figure 3: Comparison between the proposed algorithm and the double-loop decentralized AC algorithm under different inner loop sizes. The results are averaged over 10 Monte Carlo runs.

Comparison to two-timescale decentralized AC. Next, we compare the empirical performance between single-timescale and two-timescale implementations. The baseline we compare here is the existing decentralized two-timescale AC algorithm [38].

We use "TDAC-re" to denote the algorithm proposed in [38]. To compare with our proposed Algorithm 2, we also implement a noisy reward version of "TDAC-re" and denote it by "TDAC-noi". We fix $K_c = 1$, $K_r = 5$ for this experiment. We set $\alpha_k = 0.01(k+1)^{-0.5}$, $\beta_k = 0.1(k+1)^{-0.5}$, and $\eta_k = 0.1(k+1)^{-0.5}$ for "SDAC-re" and "SDAC-noi"; we set $\alpha_k = 0.01(k+1)^{-0.6}$, $\beta_k = 0.1(k+1)^{-0.4}$, and $\eta_k = 0.1(k+1)^{-0.4}$ for "TDAC-re" and "TDAC-noi". The sample complexity complexity is presented in Figure 4. We can see that the convergence speed of "TDAC-noi" is comparable to its single-timescale counterpart "SDAC-noi". However, when using reward estimator for the global reward estimation, we observe that "SDAC-re" has much more stable convergence behavior than "TDAC-re", and achieves significantly higher rewards.

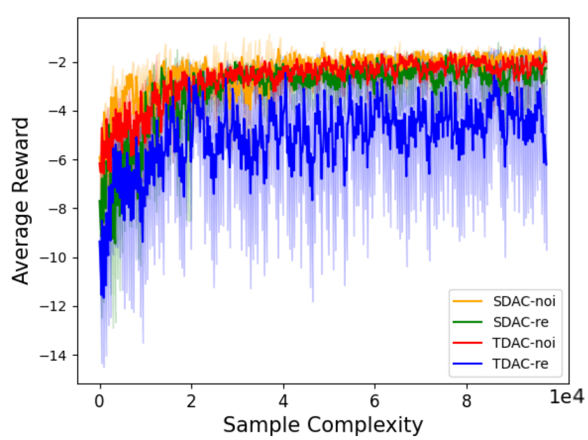
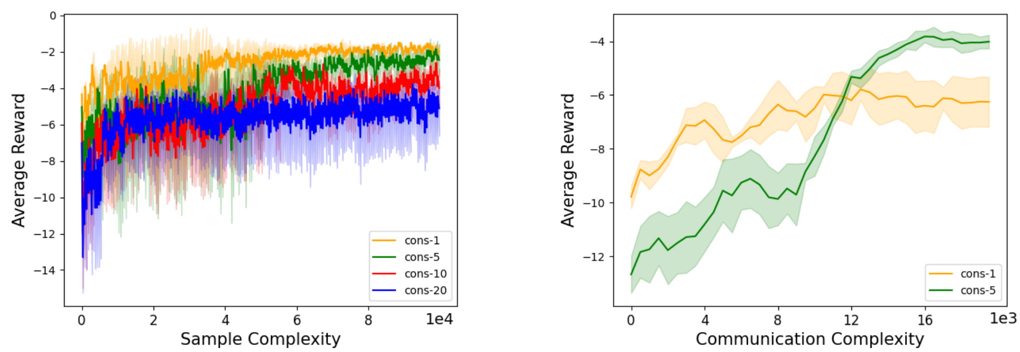


Figure 4: Comparison between the proposed algorithms and two-timescale decentralized AC algorithms [38]. The results are averaged over 10 Monte Carlo runs.

Ablation on different consensus periods. We compare the performance of "SDAC-noi" under different choices of consensus periods K_c. In particular, we let $\alpha_k = 0.01(k+1)^{-0.5}$, $\beta_k = 0.1(k+1)^{-0.5}$, $K_r = 1$ and examine the consensus periods K_c of 1, 5, 10, and 20, respectively.

The corresponding sample complexities and are summarized in Figure 5. Evidently, as the consensus period K_c increases, the convergence becomes slower and become relatively unstable. Therefore, when the communication cost is low, choosing a small K_c will yield a better performance. For this task, the consensus period K_c should be kept within 5 rounds in order to ensure a reasonable convergence. In Figure 5, we plot the communication complexity under the consensus periods of 1 and 5. We can see that the communication complexity of "cons-5" surpasses "cons-1" during the training, indicating that it requires less rounds of communications to achieve better performance. Thus, when the communication complexity is high, we may use large K_c to achieve better communication complexity. When extending the model to different tasks, we may try different values of K_c to balance the sample complexity and communication complexity.



(a) The sample complexity. (b) The communication complexity.

Figure 5: Ablation study on the consensus periods. The results are averaged over 10 Monte Carlo runs.

B Algorithm without local action

In this section, we introduce the variant of Algorithm 1 for preserving the privacy of local actions. The main difference is that instead of using a reward estimator to approximate the global reward, we now communicate the noisy local rewards for estimating the global rewards. Let r_k^i represents $r_k^i(s_k, a_k)$ for brevity. The reward estimation process goes as follow: for each agent i, we first produce a noisy local reward $\tilde{r}_k^i = r_k^i(1+z)$, with $z \sim \mathcal{N}(0, \sigma^2)$. Thus, the noise level is controlled by the variance σ^2, which is chosen artificially. To estimate the global reward, each agent i first initialize the estimation as $\tilde{r}_{t,0}^i = \tilde{r}_t^i$. Then, each agent i perform the following consensus step for K_r times, i.e.

$$\tilde{r}_{t,l+1}^i = \sum_{j=1}^{N} W^{ij} \tilde{r}_{t,l}^i, \quad l = 0, 1, \cdots, K_r - 1. \tag{17}$$

The reward $\tilde{r}_{k,K_r}^i$ will be used for estimating global reward for agent i. The error for the reward estimation, i.e. $|\bar{r}_k - \tilde{r}_{k,K_r}^i|$ will converge to 0 linearly. Therefore, to reduce the error to ε, we need $K_r = \mathcal{O}(\log(\varepsilon^{-1}))$ rounds of communications.

The following theorem establishes the sample complexity of Algorithm 2 under Markovian sampling.

Theorem 3. *Suppose Assumptions 1-5 hold. Consider the update of Algorithm 2 under Markovian sampling. Let* $\alpha_k = \frac{\bar{\alpha}}{\sqrt{K}}$ *for some positive constant* $\bar{\alpha}$, $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k$, $K_c = \mathcal{O}(\log(K^{1/4}))$, $K_r = \log(K^{1/2})$. *Then, we have*

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\left[\|\omega_k^i - \omega^*(\theta_k)\|^2\right] \le \mathcal{O}\left(\frac{\log^2 K}{\sqrt{K}}\right)$$
$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta^i} F(\theta_k)\|^2\right] \le \mathcal{O}\left(\frac{\log^2 K}{\sqrt{K}}\right) + \mathcal{O}(\varepsilon_{app} + \varepsilon_{sp}), \tag{18}$$

Algorithm 2: Decentralized single-timescale AC (noisy reward version)

Initialize: Actor parameter θ_0, critic parameter ω_0, initial state s_0.
for $k = 0, \cdots, K-1$ **do**
  Option 1: i.i.d. sampling:
  $s_k \sim \mu_{\theta_k}(\cdot), a_k \sim \pi_{\theta_k}(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$.
  Option 2: Markovian sampling:
  $a_k \sim \pi_{\theta_k}(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$.

  Periodical consensus: Compute $\tilde{\omega}_k^i$ by (4).

  for $i = 0, \cdots, N$ **in parallel do**
    Global reward estimation: Estimate $\bar{r}_k(s_k, a_k)$ by (17).
    Critic update: Update ω_{k+1}^i by (5).
    Actor update: Update θ_{k+1}^i by (6).
  end for
end for

where C_9 and C_{10} are positive constants defined in proof.

The Theorem 3 shows that Algorithm 2 has the same sample complexity as Algorithm 1; see Appendix E.3 for the proof. Algorithm 2 enjoys the advantage of preserving local actions and requiring less parameters since no reward estimator is needed. The cost is that we need to communicate $\mathcal{O}(\log(\varepsilon^{-1}))$ times for each iteration.

C Auxiliary lemmas

In this section, we provide some auxiliary lemmas, which serves as the preliminary for the proof of main theorems and lemmas.

Lemma 4 ([40], Lemma 3.2)**.** *Suppose Assumption 3 holds, then there exists a positive constant L such that for all $\theta, \theta' \in \mathbb{R}^{d_\theta}$, we have $\|\nabla J(\theta) - \nabla J(\theta')\| \le L\|\theta - \theta'\|$.*

Lemma 5 ([24], Lemma 1)**.** *Suppose Assumptions 4 holds, then there exists $\kappa > 0, \rho \in [0, 1]$ such that for any $\theta \in \mathbb{R}^{Nd_\theta}$ we have*

$$\sup_{s_0 \in \mathcal{S}} d_{TV}(\mathbb{P}((s_k, a_k, s_{k+1}) \in \cdot|s_0, \pi_\theta), \mu_\theta \otimes \pi_\theta, \mathcal{P}) \le \kappa\rho^k,$$

where μ_θ is the stationary distribution induced by π_θ and transition kernel $\mathcal{P}(\cdot|s, a)$.

Lemma 6 ([24], Lemma 2)**.** *Suppose Assumption 4 holds, then for any $\theta \in \mathbb{R}^d_\theta$, we have*

$$d_{TV}(d_\theta, \mu_\theta) \le 2(\log_\rho \kappa^{-1} + \frac{1}{1-\rho})(1-\gamma).$$

Lemma 7 ([24], Lemma 4)**.** *Suppose Assumption 3 holds, for any $\theta_1, \theta_2 \in \mathbb{R}^{d_\theta}$ and $s \in \mathcal{S}$, there exits a positive constant L_V such that*

$$\|\nabla V_{\pi_{\theta_1}}(s)\| \le L_V$$
$$|V_{\pi_{\theta_1}}(s) - V_{\pi_{\theta_2}}(s)| \le L_V\|\theta_1 - \theta_2\|.$$

Lemma 8 ([32], Lemma A.1)**.** *For any policy θ_1 and θ_2, it holds that*

$$d_{TV}(\mu_{\theta_1}, \mu_{\theta_2}) \le |\mathcal{A}|L_\pi(\log_\rho \kappa^{-1} + (1-\rho)^{-1})\|\theta_1 - \theta_2\|$$
$$d_{TV}(\mu_{\theta_1} \otimes \pi_{\theta_1}, \mu_{\theta_2} \otimes \pi_{\theta_2}) \le |\mathcal{A}|L_\pi(1 + \log_\rho \kappa^{-1} + (1-\rho)^{-1})\|\theta_1 - \theta_2\|$$
$$d_{TV}(\mu_{\theta_1} \otimes \pi_{\theta_1} \otimes \mathcal{P}, \mu_{\theta_2} \otimes \pi_{\theta_2} \otimes \mathcal{P}) \le |\mathcal{A}|L_\pi(1 + \log_\rho \kappa^{-1} + (1-\rho)^{-1})\|\theta_1 - \theta_2\|.$$

We will define $L_\mu := |\mathcal{A}|L_\pi(\log_\rho \kappa^{-1} + (1-\rho)^{-1})$ for the proof of main theorems and lemmas.

Lemma 9 ([5], Lemma F.3)**.** *For a doubly stochastic matrix $W \in \mathbb{R}^{N \times N}$ and the difference matrix $Q := I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$, it holds that for any matrix $H \in \mathbb{R}^{N \times N}$, $\|W^k H\|_F \le \nu^k\|QH\|_F$, where ν is the second largest singular value of W.*

Lemma 10 (descent lemma in high dimension)**.** *Consider the mapping $F : \mathbb{R}^n \to \mathbb{R}^m$. If there exists a positive constant L such that*

$$\|\nabla F(x) - \nabla F(y)\|_F \le L\|x - y\|, \ \forall x, y \in dom(F), \tag{19}$$

then the following holds

$$\|F(y) - F(x) - \nabla F(x)(y - x)\| \le \frac{L_1}{2}\sqrt{m}\|y - x\|^2.$$

Proof. Observe that (19) directly implies the smoothness of each entry F_i:

$$\|\nabla F_i(x) - \nabla F_i(y)\| \le \|\nabla F(x) - \nabla F(y)\|_F \le L_1\|x - y\|.$$

Define

$$z_i(x, y) := F_i(y) - F_i(x) - \nabla F_i(x)^T(y - x).$$

We have

$$\begin{aligned}\|F(y) - F(x) - \nabla F(x)(y - x)\| &= \sqrt{\sum_{i=1}^{m} z_i(x, y)^2} \\ &\le \sqrt{m(\frac{L_1}{2}\|y - x\|^2)^2} \\ &= \frac{L_1}{2}\sqrt{m}\|y - x\|^2,\end{aligned}$$

where the inequality follows the descent lemma. □

Lemma 11 (Lipschitz property of multiplication)**.** *Suppose $f(x)$ and $g(x)$ are two functions bounded by C_f and C_g, and are L_f- and L_g-Lipschitz continuous, then $f(x)g(x)$ is $C_fL_g + C_gL_f$-Lipschitz continuous.*

Proof.

$$\begin{aligned}\|f(x_1)g(x_1) - f(x_2)g(x_2)\| &= \|f(x_1)g(x_1) - f(x_1)g(x_2) + f(x_1)g(x_2) - f(x_2)g(x_2)\| \\ &\le \|f(x_1)\|\|g(x_1) - g(x_2)\| + \|f(x_1) - f(x_2)\|\|g(x_2)\| \\ &\le (C_fL_g + C_gL_f)\|x_1 - x_2\|.\end{aligned}$$

□

Lemma 12 (invertible property of matrix)**.** *If a square matrix A satisfying $\lim_{t\to\infty} A^t = 0$, or equivalently, $|\lambda(A)| < 1$, then $I - A$ is invertible.*

Proof.

$$\begin{aligned}(I - A)\lim_{t\to\infty}\sum_{i=0}^{t} A^t &= \lim_{t\to\infty}[\sum_{i=0}^{t} A^t - \sum_{i=1}^{t+1} A^t] \\ &= I - \lim_{t\to\infty} A^{t+1} \\ &= I.\end{aligned}$$

Since I is invertible, by the rank inequality rank$(AB) \le \min(\text{rank}(A), \text{rank}(B))$, $I - A$ and $\lim_{t\to\infty}\sum_{i=0}^{t} A^t$ will be invertible. □

Lemma 13 (mismatch between Markovian sampling and stationary distribution)**.** *Consider the Markov chain:*

$$s_{k-z} \xrightarrow{\theta_{k-z}} a_{k-z} \xrightarrow{\mathcal{P}} s_{k-z+1} \xrightarrow{\theta_{k-z+1}} a_{k-z+1} \cdots \xrightarrow{\theta_{k-1}} a_{k-1} \xrightarrow{\mathcal{P}} s_k \xrightarrow{\theta_k} a_k \xrightarrow{\mathcal{P}} s_{k+1}.$$

Also consider the auxiliary Markov chain with fixed policy:

$$s_{k-z} \xrightarrow{\theta_{k-z}} a_{k-z} \xrightarrow{\mathcal{P}} s_{k-z+1} \xrightarrow{\theta_{k-z}} \tilde{a}_{k-z+1} \cdots \xrightarrow{\theta_{k-z}} \tilde{a}_{k-1} \xrightarrow{\mathcal{P}} \tilde{s}_k \xrightarrow{\theta_{k-z}} \tilde{a}_k \xrightarrow{\mathcal{P}} \tilde{s}_{k+1}.$$

Let $\xi_k := (s_k, a_k, s_{k+1})$ be sampled from chain 1, and $\tilde{\xi}_k := (s_k, a_k, s_{k+1})$ be sampled from chain 2. Then we have

$$d_{TV}(\mathbb{P}(\xi_k \in \cdot|\theta_{k-z}, s_{k-z+1}), \mathbb{P}(\tilde{\xi}_k \in \cdot|\theta_{k-z}, s_{k-z+1})) \leq \frac{1}{2}\sum_{m=0}^{z-1} |\mathcal{A}|L_\pi \|\theta_{k-m} - \theta_{k-z}\|.$$

Proof.

$$\begin{aligned}
&d_{TV}(\mathbb{P}(\xi_k \in \cdot), \mathbb{P}(\tilde{\xi}_k \in \cdot)) \\
&= \frac{1}{2}\int_{s\in\mathcal{S}}\int_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} |\mathbb{P}(s_k = ds, a_k = a, s_{k+1} = ds') - \mathbb{P}(\tilde{s}_k = ds, \tilde{a}_k = a, \tilde{s}_{k+1} = ds')| \\
&= \frac{1}{2}\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} |\mathbb{P}(s_k = ds, a_k = a) - \mathbb{P}(\tilde{s}_k = ds, \tilde{a}_k = a)| \int_{s'\in\mathcal{S}} \mathbb{P}(s_{k+1} = ds'|s_k = ds, a_k = a) \\
&= \frac{1}{2}\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} |\mathbb{P}(s_k = ds, a_k = a) - \mathbb{P}(\tilde{s}_k = ds, \tilde{a}_k = a)| \\
&= \frac{1}{2}\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} |\mathbb{P}(s_k = ds)\pi_{\theta_k}(a|ds) - \mathbb{P}(\tilde{s}_k = ds)\pi_{\theta_{k-z}}(a|ds)| \\
&\leq \frac{1}{2}\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} |\mathbb{P}(s_k = ds)\pi_{\theta_k}(a|ds) - \mathbb{P}(s_k = ds)\pi_{\theta_{k-z}}(a|ds)| \\
&\quad + \frac{1}{2}\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} |\mathbb{P}(s_k = ds)\pi_{\theta_{k-z}}(a|ds) - \mathbb{P}(\tilde{s}_k = ds)\pi_{\theta_{k-z}}(a|ds)| \\
&\leq \frac{1}{2}\int_{s\in\mathcal{S}} |\mathcal{A}|L_\pi\|\theta_k - \theta_{k-z}\|\mathbb{P}(s_k = ds) \\
&\quad + \frac{1}{2}\int_{s\in\mathcal{S}} |\mathbb{P}(s_k = ds) - \mathbb{P}(\tilde{s}_k = ds)| \sum_{a\in\mathcal{A}} \pi_{\theta_{k-z}}(a|ds) \\
&= \frac{1}{2}|\mathcal{A}|L_\pi\|\theta_k - \theta_{k-z}\| + d_{TV}(\mathbb{P}(s_k \in \cdot), \mathbb{P}(\tilde{s}_k \in \cdot)).
\end{aligned} \tag{20}$$

The second term can be bounded as

$$\begin{aligned}
&d_{TV}(\mathbb{P}(s_k \in \cdot), \mathbb{P}(\tilde{s}_k \in \cdot)) \\
&= \frac{1}{2}\int_{s'\in\mathcal{S}} |\mathbb{P}(s_k = ds) - \mathbb{P}(\tilde{s}_k = ds)| \\
&= \frac{1}{2}\int_{s'\in\mathcal{S}} |\sum_{a\in\mathcal{A}}\int_{s\in\mathcal{S}} \mathbb{P}(s_{k-1} = ds, a_{k-1} = a, s_k = ds') - \mathbb{P}(\tilde{s}_{k-1} = ds, \tilde{a}_{k-1} = a, \tilde{s}_k = ds')| \\
&\leq \frac{1}{2}\int_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\int_{s\in\mathcal{S}} |\mathbb{P}(s_{k-1} = ds, a_{k-1} = a, s_k = ds') - \mathbb{P}(\tilde{s}_{k-1} = ds, \tilde{a}_{k-1} = a, \tilde{s}_k = ds')| \\
&= d_{TV}(\mathbb{P}(\xi_{k-1} \in \cdot), \mathbb{P}(\tilde{\xi}_{k-1} \in \cdot)).
\end{aligned} \tag{21}$$

Combined (20) and (21), we obtain

$$d_{TV}(\mathbb{P}(\xi_k \in \cdot), \mathbb{P}(\tilde{\xi}_k \in \cdot)) \leq d_{TV}(\mathbb{P}(\xi_{k-1} \in \cdot), \mathbb{P}(\tilde{\xi}_{k-1} \in \cdot)) + \frac{1}{2}|\mathcal{A}|L_\pi|\theta_k - \theta_{k-z}\|.$$

Sum over $z - 1$ steps, we obtain

$$\begin{aligned}
d_{TV}(\mathbb{P}(\xi_k \in \cdot), \mathbb{P}(\tilde{\xi}_k \in \cdot)) &\leq d_{TV}(\mathbb{P}(\xi_{k-z} \in \cdot), \mathbb{P}(\tilde{\xi}_{k-z} \in \cdot)) + \frac{1}{2}\sum_{m=0}^{z-1} |\mathcal{A}|L_\pi\|\theta_{k-m} - \theta_{k-z}\| \\
&= \frac{1}{2}\sum_{m=0}^{z-1} |\mathcal{A}|L_\pi\|\theta_{k-m} - \theta_{k-z}\|.
\end{aligned}$$

□

D Supporting lemmas

Before proceeding to the analysis of critic variables, we firstly justify the uniqueness of optimal solution for critic and reward estimator variables. Define the following notations

$$
\begin{aligned}
A_{\theta,\phi} &:= \mathbb{E}[\phi(s)(\gamma\phi(s')^T - \phi(s)^T)] \\
A_{\theta,\varphi} &:= \mathbb{E}[\varphi(s,a)\varphi(s,a)^T] \\
b_{\theta,\phi} &:= \mathbb{E}[\phi(s)\bar{r}(s,a)] \\
b_{\theta,\varphi} &:= \mathbb{E}[\varphi(s,a)\bar{r}(s,a)],
\end{aligned} \tag{22}
$$

with expectation taken from $s \sim \mu_\theta(s), a \sim \pi_\theta, s' \sim \mathcal{P}$. The optimal critic and reward estimator variables given policy θ will satisfy $A_{\theta,\phi}\omega^*(\theta) + b_{\theta,\phi} = 0; A_{\theta,\varphi}\lambda^*(\theta) + b_{\theta,\varphi} = 0$. By Assumption 2, $A_{\theta,\phi}$ and $A_{\theta,\varphi}$ are negative definite with largest eigenvalue λ_ϕ and λ_φ, which ensures the unique solution $\omega^*(\theta) = -A_{\theta,\phi}^{-1}b_{\theta,\phi}; \lambda^*(\theta) = -A_{\theta,\phi}^{-1}b_{\theta,\phi}$. Let $R_\omega := \frac{r_{\max}}{\lambda_\phi}, R_\lambda := \frac{r_{\max}}{\lambda_\varphi}$. Then the norm of optimal solutions will be bounded as $\|\omega^*(\theta)\| \leq R_\omega,\ \|\lambda^*(\theta)\| \leq R_\lambda$, which justifies the projection step of the Algorithm 1.

To study the error of critic, we introduce the following notations

$$
\begin{aligned}
\delta^i(\xi,\theta) &:= r^i(s,a) + \gamma V_\theta(s') - V_\theta(s) \\
\delta(\xi,\theta) &:= \bar{r}(s,a) + \gamma V_\theta(s') - V_\theta(s) \\
\tilde{\delta}(\xi,\omega) &:= \bar{r}(s,a) + \gamma\phi(s')^T\omega - \phi(s)^T\omega \\
\hat{\delta}(\xi,\omega,\lambda) &:= \varphi(s,a)^T\lambda + \gamma\phi(s')^T\omega - \phi(s)^T\omega,
\end{aligned} \tag{23}
$$

where we overwrite V_{π_θ} as V_θ for simplicity.

For the ease of expression, we further define

$$
\begin{aligned}
g_a^i(\xi,\omega,\lambda) &:= \hat{\delta}(\xi,\omega,\lambda)\psi_{\theta^i}(s,a^i) \\
g_c^i(\xi,\omega) &:= \delta^i(\xi,\omega)\phi(s) \\
\bar{g}_c(\xi,\omega) &:= \tilde{\delta}(\xi,\omega)\phi(s) \\
g_c(\theta,\omega) &:= \mathbb{E}_{\xi\sim\mu_\theta}[\bar{g}_c(\xi,\omega)].
\end{aligned} \tag{24}
$$

We will start with the error of averaged critic parameter first. The following lemma characterizes the descent of averaged critic variables under i.i.d. sampling.

D.1 Error of critic

We first present several useful lemmas and propositions, which serves as the preliminary for establishing the approximate descent property of the critic variables' optimal gap.

Proposition 1 (Lipschitz continuity of $\omega^*(\theta)$ [32]). *Suppose Assumptions 2 and 4 hold, then there exists a positive constant L_ω such that for any $\theta_1, \theta_2 \in \mathbb{R}^{Nd_\theta}$, we have*

$$\|\omega^*(\theta_1) - \omega^*(\theta_2)\| \leq L_\omega\|\theta_1 - \theta_2\|.$$

Lemma 14 (smoothness of stationary distribution). *Suppose Assumptions 1, 3, and 4 hold, then for any $\theta, \theta' \in \mathbb{R}^d$, there exists a positive constant $L_{\mu,1}$ such that*

$$\|\nabla\mu_\theta(s) - \nabla\mu_{\theta'}(s)\| \leq L_{\mu,1}\|\theta - \theta'\|.$$

The proof of this Lemma consists of two main steps: 1) Derive the expression of the gradient and 2) establish that the gradient is Lipschitz continuous. For the first part, we follow the main idea in [2].

Proof. For a given policy π_θ, we define the transition probability $P_\theta(s|s') := \sum_a \pi_\theta(a|s')P(s|s',a)$. By the Assumption 4, there exists a stationary distribution $\mu_\theta(s)$ which satisfies for all state s

$$\mu_\theta(s) = \sum_{s'\in\mathcal{S}} \mu_\theta(s')P_\theta(s|s') \tag{25}$$

Define the following notations

$$\begin{aligned}
\mu_\theta &:= [\mu_\theta(s_1), \mu_\theta(s_2), \cdots, \mu_\theta(s_n)]^T & \mathbb{R}^{|\mathcal{S}|\times 1} \\
P_\theta(s) &:= [P_\theta(s|s_1), P_\theta(s|s_2), \cdots, P_\theta(s|s_n)]^T & \mathbb{R}^{|\mathcal{S}|\times 1} \\
P(\theta) &:= [P_\theta(s_1), P_\theta(s_2), \cdots, P_\theta(s_n)] & \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|} \\
\nabla\mu_\theta &:= [\nabla\mu_\theta(s_1), \nabla\mu_\theta(s_2), \cdots, \nabla\mu_\theta(s_n)] & \mathbb{R}^{d_\theta\times|\mathcal{S}|} \\
\nabla P_\theta(s) &:= [\nabla P_\theta(s|s_1), \nabla P_\theta(s|s_2), \cdots, \nabla P_\theta(s|s_n)] & \mathbb{R}^{d_\theta\times|\mathcal{S}|}
\end{aligned}$$

Upon taking derivative with respect to θ on both sides of (25), we have

$$\begin{aligned}
\nabla\mu_\theta(s) &= \sum_{s'\in\mathcal{S}} \nabla\mu_\theta(s')P_\theta(s|s') + \mu_\theta(s')\nabla_\theta P_\theta(s|s') \\
&= \nabla\mu_\theta P_\theta(s) + \nabla P_\theta(s)\mu_\theta
\end{aligned} \tag{26}$$

(26) can be written in compact form as

$$\nabla\mu_\theta = \nabla\mu_\theta P(\theta) + [\nabla P_\theta(s_1)\mu_\theta, \cdots, \nabla P_\theta(s_n)\mu_\theta] \tag{27}$$

Therefore, we have

$$\begin{aligned}
[\nabla P_\theta(s_1)\mu_\theta, \cdots, \nabla P_\theta(s_n)\mu_\theta] &= \nabla\mu_\theta(I - P(\theta)) \\
&= \nabla\mu_\theta(I - (P(\theta) - e\mu_\theta^T)),
\end{aligned}$$

where the second inequality is due to $\nabla\mu_\theta e = \nabla(\mu_\theta e) = \nabla 1 = 0$.

We now show that $I-(P(\theta)-e\mu_\theta^T)$ is invertible. The first step is to show $\lim_{t\to\infty}(P(\theta)-e\mu_\theta^T)^t = 0$. Let P, μ represent $P(\theta), \mu_\theta$ for simplicity, we first show $(P - e\mu^T)^t = P^t - P^{t-1}e\mu^T$ by induction. Observe that when $t = 1$, this is trivially satisfied. Suppose the equality holds for $t = k$, then

$$\begin{aligned}
(P - e\mu^T)^{k+1} &= (P^k - P^{k-1}e\mu^T)P - (P^k - P^{k-1}e\mu^T)e\mu^T \\
&= P^{k+1} - P^{k-1}e\mu^T - P^k e\mu^T + P^{k-1}(e\mu^T)^2 \\
&= P^{k+1} - P^k e\mu^T,
\end{aligned}$$

where the second equality is due to (25) such that $e\mu^T P = e\mu^T$ and the last equality is due to $\mu^T e = 1$.

Therefore, we have

$$\lim_{t\to\infty}(P(\theta) - e\mu_\theta^T)^t = \lim_{t\to\infty}(P(\theta)^t - P(\theta)^{t-1}e\mu_\theta^T) = e\mu_\theta^T - e\mu_\theta^T = 0,$$

which together with Lemma 12 justifies that $I - (P(\theta) - e\mu_\theta^T)$ is invertible. Thus, we have

$$\nabla\mu_\theta = (I - (P(\theta) - e\mu_\theta^T))^{-1}[\nabla P_\theta(s_1)\mu_\theta, \cdots, \nabla P_\theta(s_n)\mu_\theta]. \tag{28}$$

We will utilize Lemma 11 to prove the Lipschitz property of $\nabla\mu_\theta$. We first show the Lipschitz continuous of the first term. Let $A(\theta)$ to represent $I - (P(\theta) - e\mu_\theta^T)$, then we have

$$\begin{aligned}
\|A(\theta_1) - A(\theta_2)\| &= \|P(\theta_1) - P(\theta_2) + e(\mu_{\theta_2} - \mu_{\theta_1})^T\| \\
&\le \|P(\theta_1) - P(\theta_2)\| + \|e(\mu_{\theta_2} - \mu_{\theta_1})^T\| \\
&= \sqrt{\sum_{s,s'\in\mathcal{S}} |\sum_{a\in\mathcal{A}}(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s'))P(s|s',a)|^2} + \sqrt{|\mathcal{S}|}\|\mu_{\theta_2} - \mu_{\theta_1}\| \\
&\le \sqrt{\sum_{s,s'\in\mathcal{S}} (\sum_{a\in\mathcal{A}} |(\pi_{\theta_1}(a|s') - \pi_{\theta_2}(a|s'))P(s|s',a)|)^2} + \sqrt{|\mathcal{S}|}\|\mu_{\theta_2} - \mu_{\theta_1}\| \\
&\le \sqrt{\sum_{s'\in\mathcal{S}} |\mathcal{A}|^2 L_\pi^2 \|\theta_1 - \theta_2\|^2 \sum_{s\in\mathcal{S}} P(s|s',a)^2} + \sqrt{|\mathcal{S}|}L_\mu\|\theta_1 - \theta_2\| \\
&= \sqrt{|\mathcal{S}|}(|\mathcal{A}|L_\pi + L_\mu)\|\theta_1 - \theta_2\|.
\end{aligned}$$

where the second inequality uses triangle inequality. The last inequality is due to Lipschitz continuous of the policy specified in Assumption 3, and Lipschitz continuous of μ_θ implied by Lemma 7.

To see that $A^{-1}(\theta)$ is Lipschitz continuous and bounded, observe that

$$\begin{aligned}\|A^{-1}(\theta_1) - A^{-1}(\theta_2)\| &= \|A^{-1}(\theta_2)(A(\theta_2) - A(\theta_1))A^{-1}(\theta_1)\| \\ &\le \|A^{-1}(\theta_2)\|\|A^{-1}(\theta_1)\|\|A(\theta_2) - A(\theta_1)\| \\ &\le \sqrt{|\mathcal{S}|}(|\mathcal{A}|L_\pi + L_\mu)\|A^{-1}(\theta_2)\|\|A^{-1}(\theta_1)\|\|\theta_2 - \theta_1\|, \qquad (29)\end{aligned}$$

where the first inequality uses Cauchy-Schwartz inequality, and the last inequality uses the Lipschitz continuous of $A(\theta)$ in (29). Since $\|A(\theta)\|$ is bounded, $\|A^{-1}(\theta)\|$ is also bounded (due to invertibility), which justifies that the first term in (28) is Lipschitz continuous and bounded.

We now consider the second term in (28). For any state s

$$\begin{aligned}\|\nabla P_{\theta_1}(s)\mu_{\theta_1} - \nabla P_{\theta_2}(s)\mu_{\theta_2}\| &= \|\nabla P_{\theta_1}(s)(\mu_{\theta_1} - \mu_{\theta_2}) + (\nabla P_{\theta_1}(s) - \nabla P_{\theta_2}(s))\mu_{\theta_2}\| \\ &\le \|\nabla P_{\theta_1}(s)(\mu_{\theta_1} - \mu_{\theta_2})\| + \|(\nabla P_{\theta_1}(s) - \nabla P_{\theta_2}(s))\mu_{\theta_2}\| \\ &\le \|\nabla P_{\theta_1}(s)\|\|\mu_{\theta_1} - \mu_{\theta_2}\| + \|\nabla P_{\theta_1}(s) - \nabla P_{\theta_2}(s)\|\|\mu_{\theta_2}\| \\ &\le \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\|\nabla\pi_{\theta_1}(a|s')P(s|s',a)\|L_\mu\|\theta_1 - \theta_2\| \\ &\quad + \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\|(\nabla\pi_{\theta_1}(a|s') - \nabla\pi_{\theta_2}(a|s'))P(s|s',a)\| \\ &\le |\mathcal{S}||\mathcal{A}|(C_\pi L_\mu + L_\pi)\|\theta_1 - \theta_2\|,\end{aligned}$$

which justifies the Lipschitz continuous of $\nabla P_\theta(s)\mu_\theta$. Define $B(\theta) := [\nabla P_\theta(s_1)\mu_\theta, \cdots, \nabla P_\theta(s_n)\mu_\theta]$, we have

$$\|B(\theta_1) - B(\theta_2)\| \le |\mathcal{S}|^{3/2}|\mathcal{A}|(C_\pi L_\mu + L_\pi)\|\theta_1 - \theta_2\|.$$

Since $\nabla\mu_\theta = A^{-1}(\theta)B(\theta)$, with $A^{-1}(\theta)$ and $B(\theta)$ being Lipschitz continuous and bounded. Therefore, according to Lemma 11, there exists a positive constant $L_{\mu,1}$ which satisfies

$$\|\nabla\mu_{\theta_1} - \nabla\mu_{\theta_2}\| \le L_{\mu,1}\|\theta_1 - \theta_2\|.$$

□

Proposition 2 (Lipschitz continuity of $\nabla_\theta\omega^*(\theta)$ [4])**.** *Suppose Assumptions 1-4 hold, then there exists a positive constant $L_{\omega,2}$ such that*

$$\|\nabla_\theta\omega^*(\theta_1) - \nabla_\theta\omega^*(\theta_2)\|_F \le L_{\omega,2}\|\theta_1 - \theta_2\|.$$

Proof. The proof follows the derivation of Proposition 8 of [4]. However, they make assumption that $\mu_\theta(s)$ is Lipschitz continuous, which we have justified in Lemma 14. We present the proof for the completeness.

We have $\omega^*(\theta) = -A_{\theta,\phi}^{-1}b_{\theta,\phi}$, where $A_{\theta,\phi}$ is defined in (22). The Jacobian of $\omega^*(\theta)$ can be calculated as

$$\begin{aligned}\nabla_\theta\omega^*(\theta) &= -\nabla_\theta(A_{\theta,\phi}^{-1}b_{\theta,\phi}) \\ &= -A_{\theta,\phi}^{-1}(\nabla_\theta A_{\theta,\phi})A_{\theta,\phi}^{-1}b_{\theta,\phi} - A_{\theta,\phi}(\nabla_\theta b_{\theta,\phi}). \qquad (30)\end{aligned}$$

We can utilize Lemma 11 to show the Lipschitz continuity of $\nabla\omega^*(\theta)$. We have to verify the Lipschitz continuity and boundedness of $A_{\theta,\phi}^{-1}$, $b_{\theta,\phi}$, $\nabla_\theta A_{\theta,\phi}$, and $\nabla_\theta b_{\theta,\phi}$.

The Lipschitz continuity and boundedness of $A_{\theta,\phi}^{-1}$ has been shown in (29. Let b_1 and b_2 represent $b_{\theta_1,\phi}$, $b_{\theta_2,\phi}$, we have

$$\begin{aligned}\|b_1 - b_2\| &= \|\mathbb{E}[\bar{r}(s,a,s')\phi(s)] - \mathbb{E}[r(\tilde{s},\tilde{a},\tilde{s}')\phi(\tilde{s})]\| \\ &\le \sup_{s,a,s'}\|r(s,a,s')\phi(s)\|\|\mathbb{P}((s,a,s'\in\cdot)) - \mathbb{P}((\tilde{s},\tilde{a},\tilde{s}'\in\cdot))\|_{TV} \\ &\le r_{\max}\|\mathbb{P}((s,a,s'\in\cdot)) - \mathbb{P}((\tilde{s},\tilde{a},\tilde{s}'\in\cdot))\|_{TV} \\ &\le 2|\mathcal{A}|L_\pi(1 + \log_\rho\kappa^{-1} + (1-\rho)^{-1})\|\theta_1 - \theta_2\|,\end{aligned}$$

where the last inequality follows Lemma 8.

We now analyze $\nabla_\theta A_{\theta,\phi}$. We first define

$$A(s,s') := \phi(s)(\gamma\phi(s') - \phi(s))^T, \quad b(s,a,s') := r(s,a,s')\phi(s).$$

as

$$\begin{aligned}\nabla_\theta A_{\theta,\phi} &= \nabla_\theta\left(\sum_{s,a,s'}\mu_\theta(s)\pi_\theta(a|s)P(s'|s,a)A(s,s')\right)\\ &= \sum_{s,a,s'}\left[\nabla_\theta\mu_\theta(s)\pi_\theta(a|s)P(s'|s,a)A(s,s') + \mu_\theta\nabla_\theta\pi_\theta(a|s)P(s'|s,a)A(s,s')\right].\end{aligned}$$

By Lemma 14 and Lemma 8, and Assumption 3, $\mu_\theta(s)$, $\pi_\theta(a|s)$, $\nabla_\theta\mu_\theta(s)$, $\nabla_\theta\pi_\theta(a|s)$ are Lipschitz continuous and bounded. Therefore, $\nabla_\theta A_{\theta,\phi}$ is Lipschitz and bounded.

Finally, we analyze $\nabla_\theta b_{\theta,\phi}$ by following the same technique.

$$\begin{aligned}\nabla_\theta b_{\theta,\phi} &= \nabla_\theta\left(\sum_{s,a,s'}\mu_\theta(s)\pi_\theta(a|s)P(s'|s,a)b(s,a,s')\right)\\ &= \sum_{s,a,s'}\left[\nabla_\theta\mu_\theta(s)\pi_\theta(a|s)P(s'|s,a)b(s,a,s') + \mu_\theta(s)\nabla_\theta\pi_\theta(a|s)P(s'|s,a)b(s,a,s')\right].\end{aligned}$$

By Lemma 14 and Lemma 8, and Assumption 3, $\mu_\theta(s)$, $\pi_\theta(a|s)$, $\nabla_\theta\mu_\theta(s)$, $\nabla_\theta\pi_\theta(a|s)$ are Lipschitz continuous and bounded. Thus, $\nabla_\theta b_{\theta,\phi}$ is bounded and Lipschitz continuous.

We have shown the Lipschitz continuity and boundedness of $A^{-1}_{\theta,\phi}$, $b_{\theta,\phi}$, $\nabla_\theta A_{\theta,\phi}$, and $\nabla_\theta b_{\theta,\phi}$. Therefore, by applying Lemma 11, we conclude that there exists a positive constant $L_{\omega,2}$ such that $\nabla_\theta\omega^*(\theta)$ in (30) is $L_{\omega,2}$-Lipschitz continuous. □

Lemma 15 (descent of critic's optimal gap (i.i.d. sampling))**.** *Suppose Assumptions 1-4 hold, with ω_{k+1} generated by Algorithm 1 given ω_k and θ_k under i.i.d. sampling, then the following holds*

$$\begin{aligned}\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 \le{}& (1 + 4L^2_{\omega,2}N\alpha_k + \frac{L^2_{\omega,2}}{2}C^2_\theta N^2\alpha^2_k)\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2\\ &+ (\frac{L^2_{\omega,2}}{2}C^2_\theta N^2 + L^2_\omega C^2_\theta N^2)\alpha^2_k + \frac{\alpha_k}{4}\sum_{i=1}^N\|\mathbb{E}[g^i_a(\xi_k,\omega^i_{k+1},\lambda^i_{k+1})]\|^2.\end{aligned} \tag{31}$$

$$\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 \le (1 - 2\lambda_\phi\beta_k)\mathbb{E}\|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + C^2_\delta\beta^2_k. \tag{32}$$

Proof. We begin with the optimality gap of averaged critic variables

$$\begin{aligned}&\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2\\ &= \|\bar{\omega}_{k+1} - \omega^*(\theta_k) + \omega^*(\theta_k) - \omega^*(\theta_{k+1})\|^2\\ &= \|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + \|\omega^*(\theta_k) - \omega^*(\theta_{k+1})\|^2 + 2\langle\bar{\omega}_{k+1} - \omega^*(\theta_k), \omega^*(\theta_k) - \omega^*(\theta_{k+1})\rangle\\ &\le \|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + N^2L^2_\omega C^2_\theta\alpha^2_k + 2\langle\bar{\omega}_{k+1} - \omega^*(\theta_k), \nabla\omega^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle\\ &\quad + 2\langle\bar{\omega}_{k+1} - \omega^*(\theta_k), \omega^*(\theta_k) - \omega^*(\theta_{k+1}) - \nabla\omega^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle,\end{aligned} \tag{33}$$

where the inequality is due to

$$\begin{aligned}\|\omega^*(\theta_k) - \omega^*(\theta_{k+1})\|^2 &\le L_\omega\|\theta_k - \theta_{k+1}\|^2,\\ \|\theta_k - \theta_{k+1}\|^2 &= \|\sum_{i=1}^N\hat{\delta}(\xi_k,\omega^i_k,\lambda^i_k)\psi_{\theta^i_k}(s_k,a^i_k)\|^2 \le N^2\alpha^2_k C^2_\theta,\end{aligned} \tag{34}$$

with $C_\theta := C_\delta C_\psi$.

The third term in (33) can be bounded as

$$
\begin{aligned}
&\langle \bar{\omega}_{k+1} - \omega^*(\theta_k), \nabla \omega^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle \\
&\le \|\bar{\omega}_{k+1} - \omega^*(\theta_k)\| \|\nabla \omega^*(\theta_k)^T(\theta_k - \theta_{k+1})\| \\
&\le L_{\omega,2}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\| \|\theta_k - \theta_{k+1}\| \\
&\le \sum_{i=1}^{N} L_{\omega,2}\alpha_k \|\bar{\omega}_{k+1} - \omega^*(\theta_k)\| \|g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\| \\
&\le \sum_{i=1}^{N} (2L_{\omega,2}\alpha_k \|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + \frac{\alpha_k}{8}\|g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\|^2), \qquad (35)
\end{aligned}
$$

where the second inequality follows Proposition 1, the third inequality uses triangle inequality, and the last inequality uses Young's inequality.

The last term in (33) can be bounded as

$$
\begin{aligned}
&\mathbb{E}\langle \bar{\omega}_{k+1} - \omega^*(\theta_k), \omega^*(\theta_k) - \omega^*(\theta_{k+1}) - \nabla \omega^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle \\
&\le \frac{L_{\omega,2}^2}{2}\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\| \|\theta_{k+1} - \theta_k\|^2 \\
&\le \frac{L_{\omega,2}^2}{4}\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 \|\theta_{k+1} - \theta_k\|^2 + \frac{L_{\omega,2}^2}{4}\|\theta_{k+1} - \theta_k\|^2 \\
&\le \frac{L_{\omega,2}^2}{4}N^2 C_\theta^2 \alpha_k^2 \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + \frac{L_{\omega,2}^2}{4}N^2 C_\theta^2 \alpha_k^2. \qquad (36)
\end{aligned}
$$

The first inequality uses Lemma 10, and the second inequality is induced by Young's inequality. The last inequality follows (34).

Plug (35) and (36) into (33) will yield (31).

We now prove (32).

$$
\begin{aligned}
\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 &= \|\prod_{R_\omega}(\bar{\omega}_k + \beta_k \bar{g}_c(\xi_k, \bar{\omega}_k)) - \prod_{R_\omega}\omega^*(\theta_k)\|^2 \\
&\le \|\bar{\omega}_k + \beta_k \bar{g}_c(\xi, \bar{\omega}_k) - \omega^*(\theta_k)\|^2 \\
&= \|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + \beta_k^2 \|\bar{g}_c(\xi_k, \bar{\omega}_k)\|^2 + 2\beta_k \mathbb{E}[\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k)\rangle] \\
&\le \|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + \beta_k^2 C_\delta^2 + 2\beta_k \langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k)\rangle. \qquad (37)
\end{aligned}
$$

The first inequality is due to the non-expansiveness of projection to convex set. The last inequality follows

$$
\|\bar{g}_c(\xi, \omega)\| \le |r(s,a) + \gamma\phi(s')^T\omega - \phi(s)^T\omega| \le r_{\max} + (1+\gamma)R_\omega := C_\delta.
$$

Let $\xi \sim \mu_\theta$ to represent $s \sim \mu_{\pi_\theta}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)$, the last term in (37) can be bounded as

$$
\begin{aligned}
&\mathbb{E}[\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k)\rangle] \\
&= \langle \bar{\omega}_k - \omega^*(\theta_k), \mathbb{E}[\bar{g}_c(\xi_k, \bar{\omega}_k) - g_c(\theta_k, \omega^*(\theta_k))]\rangle \\
&= \beta_k \langle \bar{\omega}_k - \omega^*(\theta_k), \mathbb{E}_{\xi \sim \mu_{\theta_k}}[\phi(s)(\gamma\phi(s') - \phi(s))^T|\theta_k](\bar{\omega}_k - \omega^*(\theta_k))\rangle \\
&= \beta_k \langle \bar{\omega}_k - \omega^*(\theta_k), A_{\theta_k,\phi}(\bar{\omega}_k - \omega^*(\theta_k))\rangle \\
&\le -\lambda_\phi \beta_k \|\bar{\omega}_k - \omega^*(\theta_k)\|^2. \qquad (38)
\end{aligned}
$$

Here the first equality is due to critic's optimality condition $g_c(\theta_k, \omega^*(\theta_k)) = \mathbb{E}_{\xi_k \sim \mu_{\theta_k}}[\bar{g}_c(\xi_k, \omega^*(\theta_k))|\theta_k] = 0$. The last inequality uses the negative definiteness of $A_{\theta_k,\phi}$. Plug (38) into (37) gives us (36). □

The next lemma describes the descent property of averaged critic variables under Markovian sampling.

Lemma 16 (descent of critic's optimal gap (Markovian sampling)). *Under Assumptions 1-4, with ω_{k+1} generated by Algorithm 1 given ω_k and θ_k under Markovian sampling, then the following holds*

$$\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 \leq (1 + 4L_{\omega,2}^2 N\alpha_k + \frac{L_{\omega,2}^2}{2} C_\theta^2 N^2 \alpha_k^2)\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + (\frac{L_{\omega,2}^2}{2} C_\theta^2 N^2 + L_\omega^2 C_\theta^2 N^2)\alpha_k^2 + \frac{\alpha_k}{4}\sum_{i=1}^{N} \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2. \tag{39}$$

$$\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 \leq (1 - 2\lambda_\phi\beta_k)\mathbb{E}\|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + C_{K_1}\beta_k\beta_{k-Z_K} + C_{K_2}\alpha_{k-Z_K}\beta_k. \tag{40}$$

where $C_{K_1} := 4C_2 C_\delta Z_K + C_\delta^2,\ C_{K_2} := 4C_1 C_\theta Z_K + 2C_3 C_\theta Z_K^2 + C_8, Z_K := \min\{z \in \mathbb{N}^+ | \kappa\rho^{z-1} \leq \min\{\alpha_k, \beta_k, \eta_k\}\}$.

Proof. (39) has already been derived in the proof of i.i.d. sampling setting, please check the derivation of (31).

We now prove (40). Follow the derivation of (37), we have

$$\begin{aligned}
\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 &\leq \|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + \beta_k^2 C_\delta^2 + 2\beta_k \mathbb{E}[\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k)\rangle] \\
&= \|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + \beta_k^2 C_\delta^2 + 2\beta_k \mathbb{E}\langle \bar{\omega}_k - \omega^*(\theta_k), g_c(\theta_k, \bar{\omega}_k)\rangle \\
&\quad + 2\beta_k \mathbb{E}\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k) - g_c(\theta_k, \bar{\omega}_k)\rangle \\
&\leq (1 - 2\lambda_\phi\beta_k)\|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + \beta_k^2 C_\delta^2 \\
&\quad + 2\beta_k \mathbb{E}\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k) - g_c(\theta_k, \bar{\omega}_k)\rangle.
\end{aligned} \tag{41}$$

Here, the last inequality bound the third term using the same technique of (38).

We now bound the last term in (41). By Lemma 17, for any $z \in \mathbb{N}^+$, we have

$$\begin{aligned}
&\mathbb{E}\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k) - g_c(\theta_k, \bar{\omega}_k)\rangle \\
&\leq C_1\mathbb{E}\|\theta_k - \theta_{k-z}\| + C_2\mathbb{E}\|\bar{\omega}_k - \bar{\omega}_{k-z}\| + C_3\sum_{m=0}^{z-1}\mathbb{E}\|\theta_{k-m} - \theta_{k-z}\| + C_8\kappa\rho^{z-1} \\
&\overset{(i)}{\leq} C_1\sum_{n=1}^{z}\mathbb{E}\|\theta_{k-n+1} - \theta_{k-n}\| + C_2\sum_{n=1}^{z}\mathbb{E}\|\bar{\omega}_{k-n+1} - \bar{\omega}_{k-n}\| \\
&\quad + C_3\sum_{m=0}^{z-1}\sum_{n=1}^{z-m}\mathbb{E}\|\theta_{k-m-n+1} - \theta_{k-m-n}\| + C_8\kappa\rho^{z-1} \\
&\leq 2C_1C_\theta\sum_{n=1}^{z}\alpha_{k-n} + 2C_2C_\delta\sum_{n=1}^{z}\beta_{k-n} + C_3C_\theta\sum_{m=0}^{z-1}\sum_{n=1}^{z-m}\alpha_{k-m-n} + C_8\kappa\rho^{z-1} \\
&\overset{(ii)}{\leq} 2C_1C_\theta z\alpha_{k-z} + 2C_2C_\delta z\beta_{k-z} + C_3C_\theta z(z-1)\alpha_{k-z} + C_8\kappa\rho^{z-1},
\end{aligned} \tag{42}$$

where the (i) uses triangle inequality, (ii) uses the non-increasing property of step sizes.

Let $z = Z_K := \min\{z \in \mathbb{N}^+ | \kappa\rho^{z-1} \leq \min\{\alpha_k, \beta_k, \eta_k\}\}$, we have

$$\begin{aligned}
&\mathbb{E}\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k) - g_c(\theta_k, \bar{\omega}_k)\rangle \\
&\leq 2C_1C_\theta Z_K\alpha_{k-Z_K} + 2C_2C_\delta Z_K\beta_{k-Z_K} + C_3C_\theta Z_K^2\alpha_{k-Z_K} + C_8\alpha_{k-Z_K}.
\end{aligned} \tag{43}$$

Plug (43) into (41) will yield

$$\begin{aligned}
\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 &\leq (1 - 2\lambda_\phi\beta_k)\|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + C_\delta^2\beta_k^2 \\
&\quad + 4C_1C_\theta Z_K\alpha_{k-Z_K} + 4C_2C_\delta Z_K\beta_{k-Z_K} + 2C_3C_\theta Z_K^2\alpha_{k-Z_K} + 2C_8\alpha_{k-Z_K}.
\end{aligned}$$

By defining $C_{K_1} := 4C_2C_\delta Z_K + C_\delta^2,\ C_{K_2} := 4C_1C_\theta Z_K + 2C_3C_\theta Z_K^2 + C_8$, we complete the proof. □

Lemma 17. *Consider the sequence generated by Algorithm 1, for any $z \in \mathbb{N}^+$, we have*

$$\mathbb{E}\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k) - g_c(\theta_k, \bar{\omega}_k)\rangle \leq C_1\|\theta_k - \theta_{k-z}\| + C_2\|\bar{\omega}_k - \bar{\omega}_{k-z}\| + C_3\sum_{m=0}^{z-1}\|\theta_{k-m} - \theta_{k-z}\| + C_8\kappa\rho^{z-1},$$

where $C_1 := 4R_\omega C_\delta|\mathcal{A}|L_\pi(1+\log_\rho \kappa^{-1} + (1-\rho)^{-1}) + 2C_\delta L_\omega,\ C_2 := 4(1+\gamma)R_\omega + 2C_\delta,\ C_3 := 4R_\omega C_\delta|\mathcal{A}|L_\pi,\ C_8 := 8R_\omega C_\delta$.

Proof. Consider the Markov chain since timestep $k-z$:

$$s_{k-z} \xrightarrow{\theta_{k-z}} a_{k-z} \xrightarrow{\mathcal{P}} s_{k-z+1} \xrightarrow{\theta_{k-z+1}} a_{k-z+1} \cdots \xrightarrow{\theta_{k-1}} a_{k-1} \xrightarrow{\mathcal{P}} s_k \xrightarrow{\theta_k} a_k \xrightarrow{\mathcal{P}} s_{k+1}.$$

Also consider the auxiliary Markov chain with fixed policy since timestep $k-z$:

$$s_{k-z} \xrightarrow{\theta_{k-z}} a_{k-z} \xrightarrow{\mathcal{P}} s_{k-z+1} \xrightarrow{\theta_{k-z}} \tilde{a}_{k-z+1} \cdots \xrightarrow{\theta_{k-z}} \tilde{a}_{k-1} \xrightarrow{\mathcal{P}} \tilde{s}_k \xrightarrow{\theta_{k-z}} \tilde{a}_k \xrightarrow{\mathcal{P}} \tilde{s}_{k+1}.$$

Throughout the proof of this lemma, we will use $\theta, \theta', \bar{\omega}, \bar{\omega}', \xi, \tilde{\xi}$ as shorthand notations of $\theta_k, \theta_{k-z}, \bar{\omega}_k, \bar{\omega}_{k-z}, \xi_k, \tilde{\xi}_k$.

For the ease of expression, define

$$\Delta_1(\xi, \theta, \omega) := \langle \omega - \omega^*(\theta), \bar{g}_c(\xi, \omega) - g_c(\theta, \omega)\rangle.$$

Therefore, we have

$$\begin{aligned}\langle \bar{\omega}_k - \omega^*(\theta_k), \bar{g}_c(\xi_k, \bar{\omega}_k) - g_c(\theta_k, \bar{\omega}_k)\rangle &= \Delta_1(\xi, \theta, \bar{\omega}) \\ &= \underbrace{\Delta_1(\xi, \theta, \bar{\omega}) - \Delta_1(\xi, \theta', \bar{\omega})}_{I_1} + \underbrace{\Delta_1(\xi, \theta', \bar{\omega}) - \Delta_1(\xi, \theta', \bar{\omega}')}_{I_2} \\ &\quad + \underbrace{\Delta_1(\xi, \theta', \bar{\omega}') - \Delta_1(\tilde{\xi}, \theta', \bar{\omega}')}_{I_3} + \underbrace{\Delta_1(\tilde{\xi}, \theta', \bar{\omega}')}_{I_4}. \quad (44)\end{aligned}$$

I_1 can be expressed as

$$\begin{aligned}I_1 &= \langle \bar{\omega} - \omega^*(\theta), \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta, \bar{\omega})\rangle - \langle \bar{\omega} - \omega^*(\theta'), \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta', \bar{\omega})\rangle \\ &= \langle \bar{\omega} - \omega^*(\theta), \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta, \bar{\omega})\rangle - \langle \bar{\omega} - \omega^*(\theta), \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta', \bar{\omega})\rangle \\ &\quad + \langle \omega^*(\theta) - \omega^*(\theta'), \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta', \bar{\omega})\rangle \\ &\leq \|\bar{\omega} - \omega^*(\theta)\|\|g_c(\theta', \bar{\omega}) - g_c(\theta, \bar{\omega})\| + \|\omega^*(\theta) - \omega^*(\theta')\|\|\bar{g}_c(\xi, \bar{\omega}) - g_c(\theta', \bar{\omega})\|. \quad (45)\end{aligned}$$

The first term can be bounded as

$$\begin{aligned}\|\bar{\omega} - \omega^*(\theta)\|\|g_c(\theta', \bar{\omega}) - g_c(\theta, \bar{\omega})\| &\leq 2R_\omega\|\mathbb{E}_{\xi\sim\mu'_\theta}[\bar{g}_c(\xi, \bar{\omega})] - \mathbb{E}_{\xi\sim\mu_\theta}[\bar{g}_c(\xi, \bar{\omega})]\| \\ &\leq 4R_\omega \sup_\xi \|\bar{g}_c(\xi, \bar{\omega})\| d_{TV}(\mu'_\theta \otimes \pi'_\theta \otimes \mathcal{P}, \mu_\theta \otimes \pi_\theta \otimes \mathcal{P}) \\ &\leq 4R_\omega C_\delta d_{TV}(\mu'_\theta \otimes \pi'_\theta \otimes \mathcal{P}, \mu_\theta \otimes \pi_\theta \otimes \mathcal{P}) \\ &\leq 4R_\omega C_\delta|\mathcal{A}|L_\pi(1 + \log_\rho \kappa^{-1} + (1-\rho)^{-1})\|\theta - \theta'\|, \quad (46)\end{aligned}$$

where the first inequality follows the projection update of each critic step, the third inequality is due to $\|\bar{g}_c(\xi, \bar{\omega})\| \leq C_\delta$, and the last inequality follows Lemma 8.

By the Lipschitz conitinuous of $\omega^*(\theta)$ proposed in Proposition 1, the second term can be bounded as

$$\|\omega^*(\theta) - \omega^*(\theta')\|\|\bar{g}_c(\xi, \bar{\omega}) - g_c(\theta, \bar{\omega})\| \leq 2C_\delta L_\omega\|\theta - \theta'\| \quad (47)$$

Plug (46) and (47) into (45), we can bound I_1 as

$$I_1 \leq (4R_\omega C_\delta|\mathcal{A}|L_\pi(1 + \log_\rho \kappa^{-1} + (1-\rho)^{-1}) + 2C_\delta L_\omega)\|\theta - \theta'\|. \quad (48)$$

Next we bound I_2 as

$$\begin{aligned} I_2 &= \langle \bar{\omega} - \omega^*(\theta'), \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta', \bar{\omega}) \rangle - \langle \bar{\omega}' - \omega^*(\theta'), \bar{g}_c(\xi, \bar{\omega}') - g_c(\theta', \bar{\omega}') \rangle \\ &= \langle \bar{\omega} - \omega^*(\theta'), \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta', \bar{\omega}) \rangle - \langle \bar{\omega}' - \omega^*(\theta'), \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta', \bar{\omega}) \rangle \\ &\quad + \langle \bar{\omega}' - \omega^*(\theta'), \bar{g}_c(\xi, \bar{\omega}) - \bar{g}_c(\xi, \bar{\omega}') - g_c(\theta', \bar{\omega}) + g_c(\theta', \bar{\omega}') \rangle. \end{aligned}$$

The first two terms can be bounded as

$$\langle \bar{\omega} - \bar{\omega}', \bar{g}_c(\xi, \bar{\omega}) - g_c(\theta', \bar{\omega}) \rangle \leq 2C_\delta \|\bar{\omega} - \bar{\omega}'\|. \tag{49}$$

The last term can be bounded as

$$\begin{aligned} &\langle \bar{\omega}' - \omega^*(\theta'), \bar{g}_c(\xi, \bar{\omega}) - \bar{g}_c(\xi, \bar{\omega}') - g_c(\theta', \bar{\omega}) + g_c(\theta', \bar{\omega}') \rangle \\ &\leq \|\bar{\omega} - \omega^*(\theta')\| (\|\bar{g}_c(\xi, \bar{\omega}) - \bar{g}_c(\xi, \bar{\omega}')\| + \|g_c(\theta', \bar{\omega}') - g_c(\theta', \bar{\omega})\|) \\ &\leq 2R_\omega (\|\bar{g}_c(\xi, \bar{\omega}) - \bar{g}_c(\xi, \bar{\omega}')\| + \|g_c(\theta', \bar{\omega}') - g_c(\theta', \bar{\omega})\|) \\ &\leq 4R_\omega (1 + \gamma) \|\bar{\omega} - \bar{\omega}'\|, \end{aligned} \tag{50}$$

where the second inequality follows the projection of each critic step. The last inequality is due to

$$\begin{aligned} \|\bar{g}_c(\xi, \bar{\omega}) - \bar{g}_c(\xi, \bar{\omega}')\| &= \|\phi(s)(\gamma \phi(s')^T (\bar{\omega} - \bar{\omega}') - \phi(s)^T (\bar{\omega} - \bar{\omega}'))\| \\ &\leq \gamma \|\phi(s')^T (\bar{\omega} - \bar{\omega}')\| + \|\phi(s)^T (\bar{\omega} - \bar{\omega}')\| \\ &\leq (1 + \gamma) \|\bar{\omega} - \bar{\omega}'\|. \end{aligned}$$

Combine (49) and (50), we can bound I_2 as

$$I_2 \leq (4(1 + \gamma) R_\omega + 2C_\delta) \|\bar{\omega} - \bar{\omega}'\|. \tag{51}$$

We bound I_3 as

$$\begin{aligned} \mathbb{E}[I_3 | \theta', s_{k-z+1}] &= \mathbb{E}[\Delta_1(\xi, \theta', \bar{\omega}') - \Delta_1(\tilde{\xi}, \theta', \bar{\omega}') | \theta', s_{k-z+1}] \\ &\leq 2 \sup_\xi |\Delta_1(\xi, \theta', \bar{\omega}')| \; d_{TV}(\mathbb{P}(\xi \in \cdot | \theta', s_{k-z+1}), \mathbb{P}(\tilde{\xi} \in \cdot | \theta', s_{k-z+1})) \\ &\leq 8R_\omega C_\delta d_{TV}(\mathbb{P}(\xi \in \cdot | \theta', s_{k-z+1}), \mathbb{P}(\tilde{\xi} \in \cdot | \theta', s_{k-z+1})) \\ &\leq 4R_\omega C_\delta |\mathcal{A}| L_\pi \sum_{m=0}^{z-1} \|\theta_{k-m} - \theta_{k-z}\|. \end{aligned} \tag{52}$$

Here, the second inequality is due to $\|\Delta_1(\xi, \theta', \bar{\omega}')\| \leq \|\omega' - \omega^*(\theta')\| \|\bar{g}_c(\xi, \omega') - g_c(\theta', \omega')\| \leq 4R_\omega C_\delta$, and the last inequality is according to Lemma 13.

We now bound I_4

$$\begin{aligned} \mathbb{E}[I_4 | \theta', \bar{\omega}', s_{k+z-1}] &= \mathbb{E}[\Delta_1(\tilde{\xi}, \theta', \bar{\omega}') | \theta', \bar{\omega}', s_{k-z+1}] \\ &\leq \sup_\xi |\Delta_1(\xi, \theta', \bar{\omega}')| \|\mathbb{P}(\xi \in \cdot | \theta', s_{k-z+1}) - \mu_{\theta'} \otimes \pi_{\theta'} \otimes \mathcal{P}\| \\ &\leq 8R_\omega C_\delta d_{TV}(\mathbb{P}(\tilde{x} \in \cdot | \theta', s_{t-z+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \mathcal{P}) \\ &\leq 8R_\omega C_\delta \kappa \rho^{z-1}, \end{aligned} \tag{53}$$

where the last inequality follows Lemma 5.

Plug (48), (51), (52), and (53) into (44), we get

$$\begin{aligned} \mathbb{E}[\Delta_1(\xi, \theta, \bar{\omega})] &\leq (4R_\omega C_\delta |\mathcal{A}| L_\pi (1 + \log_\rho \kappa^{-1} + (1 - \rho)^{-1}) + 2C_\delta L_\omega) \mathbb{E}\|\theta_k - \theta_{k-z}\| \\ &\quad + (4(1 + \gamma) R_\omega + 2C_\delta) \mathbb{E}\|\bar{\omega}_k - \bar{\omega}_{k-z}\| \\ &\quad + (4R_\omega C_\delta |\mathcal{A}| L_\pi) \sum_{m=0}^{z-1} \mathbb{E}\|\theta_{k-m} - \theta_{k-z}\| \\ &\quad + (8R_\omega C_\delta) \kappa \rho^{z-1}, \end{aligned}$$

which completes the proof. □

D.2 Error of reward estimator

The analysis for the error of reward estimator is similar to critic. To see this, we only need to change $\bar{g}_c(\xi, \bar{\omega})$ into $\bar{g}_r(\xi, \bar{\lambda}) := (r(s,a) - \varphi(s,a)^T \bar{\lambda})\varphi(s,a)$ to recover most of the proofs. We provide the reward estimator's analysis for the completeness. For the ease of discussion, we define

$$\begin{aligned} g_r^i(\xi, \lambda) &:= \varphi(s,a)(r^i(s,a) - \varphi(s,a)^T \lambda), \\ \bar{g}_r(\xi, \lambda) &:= \varphi(s,a)(\bar{r}(s,a) - \varphi(s,a)^T \lambda), \\ g_r(\theta, \lambda) &:= \mathbb{E}_{\xi \sim \mu_\theta}[\bar{g}_r(\xi, \lambda)]. \end{aligned}$$

Note here $g_r^i(\xi, \lambda)$ and $\bar{g}_r(\xi, \lambda)$ do not depend on the next state s'. We use ξ for notational convience.

The following lemma is the counter part of Lemma 15 for reward estimator.

Lemma 18 (descent of reward estimator's optimal gap (i.i.d. sampling))**.** *Suppopse Assumptions 1-4 hold, with λ_{k+1} generated by Algorithm 1 given λ_k and θ_k under i.i.d. sampling, then the following holds*

$$\begin{aligned} \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_{k+1})\|^2 \le (1 + 4L_{\lambda,2}^2 N\alpha_k + \frac{L_{\lambda,2}^2}{2} C_\theta^2 N^2 \alpha_k^2)\mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 \\ + (\frac{L_{\lambda,2}^2}{2} C_\theta^2 N^2 + L_\lambda^2 C_\theta^2 N^2)\alpha_k^2 + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \lambda_{k+1}^i, \lambda_{k+1}^i)]\|^2. \end{aligned} \tag{54}$$

$$\mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 \le (1 - 2\eta_k \lambda_\varphi)\|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2 + \eta_k^2 C_\lambda^2. \tag{55}$$

Proof. We begin with the optimal gap

$$\begin{aligned} &\|\bar{\lambda}_{k+1} - \lambda^*(\theta_{k+1})\|^2 \\ &= \|\bar{\lambda}_{k+1} - \lambda^*(\theta_k) + \lambda^*(\theta_k) - \lambda^*(\theta_{k+1})\|^2 \\ &= \|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 + \|\lambda^*(\theta_k) - \lambda^*(\theta_{k+1})\|^2 + 2\langle \bar{\lambda}_{k+1} - \lambda^*(\theta_k), \lambda^*(\theta_k) - \lambda^*(\theta_{k+1})\rangle \\ &\le \|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 + N^2 L_\lambda^2 C_\theta^2 \alpha_k^2 + 2\langle \bar{\lambda}_{k+1} - \lambda^*(\theta_k), \nabla\lambda^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle \\ &\quad + 2\langle \bar{\lambda}_{k+1} - \lambda^*(\theta_k), \lambda^*(\theta_k) - \lambda^*(\theta_{k+1}) - \nabla\lambda^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle \\ &\le \|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 + N^2 L_\lambda^2 C_\theta^2 \alpha_k^2 + 2\alpha_k L_{\lambda,2}\sum_{i=1}^N \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|\|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\| \\ &\quad + 2\langle \bar{\lambda}_{k+1} - \lambda^*(\theta_k), \lambda^*(\theta_k) - \lambda^*(\theta_{k+1}) - \nabla\lambda^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle \\ &\le \|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 + N^2 L_\lambda^2 C_\theta^2 \alpha_k^2 + 4\alpha_k N L_{\lambda,2}^2 \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 \\ &\quad + 2\langle \bar{\lambda}_{k+1} - \lambda^*(\theta_k), \lambda^*(\theta_k) - \lambda^*(\theta_{k+1}) - \nabla\lambda^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle. \end{aligned} \tag{56}$$

where the first inequality uses the Lipschitz continuous of $\lambda^*(\theta)$ and $\|\theta_k - \theta_{k+1}\|^2 \le N^2\alpha_k^2 C_\theta^2$. The second inequality uses triangle inequality and the Lemma 2. The last inequality is due to Young's inequality.

The last term in (56) can be bounded as

$$\begin{aligned} &\mathbb{E}\langle \bar{\lambda}_{k+1} - \lambda^*(\theta_k), \lambda^*(\theta_k) - \lambda^*(\theta_{k+1}) - \nabla\lambda^*(\theta_k)^T(\theta_k - \theta_{k+1})\rangle \\ &\le \frac{L_{\lambda,2}^2}{2}\mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|\|\theta_{k+1} - \theta_k\|^2 \\ &\le \frac{L_{\lambda,2}^2}{4}\mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2\|\theta_{k+1} - \theta_k\|^2 + \frac{L_{\lambda,2}^2}{4}\|\theta_{k+1} - \theta_k\|^2 \\ &\le \frac{L_{\lambda,2}^2}{4}N^2 C_\theta^2 \alpha_k^2 \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 + \frac{L_{\lambda,2}^2}{4}N^2 C_\theta^2 \alpha_k^2. \end{aligned} \tag{57}$$

The first inequality uses Lemma 10, and the second inequality is induced by Young's inequality. Plug (57) into (56) will yield (54).

We now prove (55)

$$\begin{aligned}\|\bar{\lambda}_{k+1}-\lambda^*(\theta_k)\|^2 &= \|\prod_{R_\lambda}(\bar{\lambda}_k-\eta_k\bar{g}_r(\xi_k,\bar{\lambda}_k))-\prod_{R_\lambda}\lambda^*(\theta_k)\|^2\\ &\le \|\bar{\lambda}_k-\eta_k\bar{g}_r(\xi_k,\bar{\lambda}_k)-\lambda^*(\theta_k)\|^2\\ &\le \|\bar{\lambda}_k-\lambda^*(\theta_k)\|^2+\eta_k^2\|\bar{g}_r(\xi_k,\bar{\lambda}_k)\|^2+2\eta_k\mathbb{E}[\langle\bar{\lambda}_k-\lambda^*(\theta_k),\bar{g}_r(\xi_k,\bar{\lambda}_k)\rangle]\\ &\le \|\bar{\lambda}_k-\lambda^*(\theta_k)\|^2+C_\lambda\eta_k^2-2\eta_k\mathbb{E}[\langle\bar{\lambda}_k+\lambda^*(\theta_k),\bar{g}_r(s_k,a_k,\bar{\lambda}_k)\rangle],\end{aligned}\tag{58}$$

where the last inequality is due to $\|\bar{g}_r(\xi_k,\bar{\lambda}_k)\|\le|r(s,a)-\varphi(s,a)^T\lambda|\le r_{\max}+R_\lambda:=C_\lambda$.

The last term can be bounded as

$$\begin{aligned}\mathbb{E}[\langle\bar{\lambda}_k-\lambda^*(\theta_k),\bar{g}_r(\xi_k,\bar{\lambda}_k)\rangle] &= \langle\bar{\lambda}_k-\lambda^*(\theta_k),\mathbb{E}[\bar{g}_r(\xi_k,\bar{\lambda}_k)-g_r(\theta_k,\lambda^*(\theta_k))]\rangle\\ &= \langle\bar{\lambda}_k-\lambda^*(\theta_k),\mathbb{E}_{\xi\sim\mu_{\theta_k}}[\varphi(s_k,a_k)\varphi(s_k,a_k)^T|\bar{\lambda}_k](\lambda^*(\theta_k)-\bar{\lambda}_k)\rangle\\ &= \langle\bar{\lambda}_k-\lambda^*(\theta_k),A_{\theta,\varphi}(\lambda^*(\theta_k)-\bar{\lambda}_k)\rangle\\ &\le -\lambda_\varphi\|\bar{\lambda}_k-\lambda^*(\theta_k)\|^2,\end{aligned}\tag{59}$$

where the first equality is according to the optimality condition of reward estimator

$$\mathbb{E}_{\xi\sim\mu_{\theta_k}}[\varphi(s,a)(r(s,a)-\varphi(s,a)^T\lambda^*(\theta_k))]=0.$$

Plug (59) into (58) will give us (55), which completes the proof. □

Lemma 19 (descent of reward estimator's optimal gap (Markovian sampling))**.** *Suppose Assumptions 1-4 hold, with λ_{k+1} generated by Algorithm 1 given λ_k and θ_k under Markovian sampling, then the following holds*

$$\begin{aligned}\mathbb{E}\|\bar{\lambda}_{k+1}-\lambda^*(\theta_{k+1})\|^2 &\le (1+4L_{\lambda,2}^2N\alpha_k+\frac{L_{\lambda,2}^2}{2}C_\theta^2N^2\alpha_k^2)\mathbb{E}\|\bar{\lambda}_{k+1}-\lambda^*(\theta_k)\|^2\\ &\quad+(\frac{L_{\lambda,2}^2}{2}C_\theta^2N^2+L_\lambda^2C_\theta^2N^2)\alpha_k^2+\frac{\alpha_k}{4}\sum_{i=1}^N\|\mathbb{E}[g_a^i(\xi_k,\lambda_{k+1}^i,\lambda_{k+1}^i)]\|^2.\end{aligned}\tag{60}$$

$$\mathbb{E}\|\bar{\lambda}_{k+1}-\lambda^*(\theta_k)\|^2\le(1-2\eta_k\lambda_\varphi)\|\bar{\lambda}_k-\lambda^*(\theta_k)\|^2+C_{K_3}\eta_k\eta_{k-Z_K}+C_{K_4}\eta_k\alpha_{k-Z_K},\tag{61}$$

where $C_{K_3}:=4C_6C_\lambda Z_K+C_\lambda^2,\ C_{K_4}:=4C_5C_\theta Z_K+2C_7C_\theta Z_K^2+C_8, Z_K:=\min\{z\in\mathbb{N}^+|\kappa\rho^{z-1}\le\min\{\alpha_k,\eta_k,\eta_k\}\}$.

Proof. Since analysis of (60) does not involve the update of $\bar{\lambda}_k$, it can be directly recovered from (54).

We now prove (61). Following the derivation of (58), we obtain

$$\begin{aligned}\|\bar{\lambda}_{k+1}-\lambda^*(\theta_k)\|^2 &\le \|\bar{\lambda}_k-\lambda^*(\theta_k)\|^2+C_\lambda^2\eta_k^2+2\eta_k\mathbb{E}[\langle\bar{\lambda}_k-\lambda^*(\theta_k),\bar{g}_r(\xi_k,\bar{\lambda}_k)\rangle]\\ &= \|\bar{\lambda}_k-\lambda^*(\theta_k)\|^2+C_\lambda^2\eta_k^2+2\eta_k\mathbb{E}[\langle\bar{\lambda}_k-\lambda^*(\theta_k),g_r(\theta_k,\bar{\lambda}_k)\rangle]\\ &\quad+2\eta_k\mathbb{E}[\langle\bar{\lambda}_k-\lambda^*(\theta_k),\bar{g}_r(\xi_k,\bar{\lambda}_k)-g_r(\theta_k,\bar{\lambda}_k)\rangle]\\ &\le (1-2\lambda_\varphi\eta_k)\|\bar{\lambda}_k-\lambda^*(\theta_k)\|^2+C_\lambda^2\eta_k^2\\ &\quad+2\eta_k\mathbb{E}[\langle\bar{\lambda}_k-\lambda^*(\theta_k),\bar{g}_r(\xi_k,\bar{\lambda}_k)-g_r(\theta_k,\bar{\lambda}_k)\rangle],\end{aligned}\tag{62}$$

where the last inequality is obatined by (61).

We now bound the last term. By Lemma 20, for any $z\in\mathbb{N}^+$, we have

$$
\begin{aligned}
&\mathbb{E}\langle\bar{\lambda}_k - \lambda^*(\theta_k), \bar{g}_r(\xi_k, \bar{\lambda}_k) - g_r(\theta_k, \bar{\lambda}_k)\rangle \\
&\le C_5\mathbb{E}\|\theta_k - \theta_{k-z}\| + C_6\mathbb{E}\|\bar{\lambda}_k - \bar{\lambda}_{k-z}\| + C_7\sum_{m=0}^{z-1}\mathbb{E}\|\theta_{k-m} - \theta_{k-z}\| + C_8\kappa\rho^{z-1} \\
&\overset{(i)}{\le} C_5\sum_{n=1}^{z}\mathbb{E}\|\theta_{k-n+1} - \theta_{k-n}\| + C_6\sum_{n=1}^{z}\mathbb{E}\|\bar{\lambda}_{k-n+1} - \bar{\lambda}_{k-n}\| \\
&\quad + C_7\sum_{m=0}^{z-1}\sum_{n=1}^{z-m}\mathbb{E}\|\theta_{k-m-n+1} - \theta_{k-m-n}\| + C_8\kappa\rho^{z-1} \\
&\le 2C_5C_\theta\sum_{n=1}^{z}\alpha_{k-n} + 2C_6C_\lambda\sum_{n=1}^{z}\eta_{k-n} + C_7C_\theta\sum_{m=0}^{z-1}\sum_{n=1}^{z-m}\alpha_{k-m-n} + C_8\kappa\rho^{z-1} \\
&\overset{(ii)}{\le} 2C_5C_\theta z\alpha_{k-z} + 2C_6C_\lambda z\eta_{k-z} + C_7C_\theta z(z-1)\alpha_{k-z} + C_8\kappa\rho^{z-1},
\end{aligned}
\tag{63}
$$

where the (i) uses triangle inequality, (ii) uses the non-increasing property of step sizes.

Let $z = Z_K$, recall $Z_K := \min\{z \in \mathbb{N}^+ | \kappa\rho^{z-1} \le \min\{\alpha_k, \eta_k, \eta_k\}\}$, we have

$$
\begin{aligned}
&\mathbb{E}\langle\bar{\lambda}_k - \lambda^*(\theta_k), \bar{g}_r(\xi_k, \bar{\lambda}_k) - g_r(\theta_k, \bar{\lambda}_k)\rangle \\
&\le 2C_5C_\theta Z_K\alpha_{k-Z_K} + 2C_6C_\lambda Z_K\eta_{k-Z_K} + C_7C_\theta Z_K^2\alpha_{k-Z_K} + C_8\alpha_{k-Z_K}.
\end{aligned}
\tag{64}
$$

Plug (64) into (62) will yield

$$
\begin{aligned}
\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 &\le (1 - 2\lambda_\phi\eta_k)\|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2 + C_\lambda^2\eta_k^2 \\
&\quad + 4C_5C_\theta Z_K\alpha_{k-Z_K} + 4C_6C_\lambda Z_K\eta_{k-Z_K} + 2C_7C_\theta Z_K^2\alpha_{k-Z_K} + 2C_8\alpha_{k-Z_K}.
\end{aligned}
$$

By defining $C_{K_3} := 4C_6C_\lambda Z_K + C_\lambda^2,\ C_{K_4} := 4C_5C_\theta Z_K + 2C_7C_\theta Z_K^2 + C_8$, we complete the proof.

□

Lemma 20. *Consider the sequence generated by Algorithm 1, for any $z \in \mathbb{N}^+$, we have*

$$
\begin{aligned}
\mathbb{E}[\langle\bar{\lambda}_k - \lambda^*(\theta), \bar{g}_r(\xi_k, \bar{\lambda}_k) - g_r(\theta_k, \bar{\lambda}_k)\rangle] &\le C_5\|\theta_k - \theta_{k-z}\| + C_6\|\lambda_k - \lambda_{k-z}\| \\
&\quad + C_7\sum_{m=0}^{z-1}\|\theta_{k-m} - \theta_{k-z}\| + C_8\kappa\rho^{z-1},
\end{aligned}
\tag{65}
$$

where $C_5 := 4R_\lambda C_\lambda|\mathcal{A}|L_\pi(1 + \log_\rho\kappa^{-1} + (1-\rho)^{-1}) + 2C_\lambda L_\lambda, C_6 := 4R_\lambda + 2C_\lambda, C_7 := 4R_\lambda C_\lambda|\mathcal{A}|L_\pi, C_8 := 8R_\lambda C_\lambda$.

Proof. Consider the Markov chain since timestep $k - z$:

$$
s_{k-m} \xrightarrow{\theta_{k-m}} a_{k-m} \xrightarrow{\mathcal{P}} s_{k-m+1} \xrightarrow{\theta_{k-m+1}} a_{k-m+1} \cdots \xrightarrow{\theta_{k-1}} a_{k-1} \xrightarrow{\mathcal{P}} s_k \xrightarrow{\theta_k} a_k \xrightarrow{\mathcal{P}} s_{k+1}.
$$

Also consider the auxiliary Markov chain with fixed policy since timestep $k - z$:

$$
s_{k-m} \xrightarrow{\theta_{k-m}} a_{k-m} \xrightarrow{\mathcal{P}} s_{k-m+1} \xrightarrow{\theta_{k-m}} \tilde{a}_{k-m+1} \cdots \xrightarrow{\theta_{k-m}} \tilde{a}_{k-1} \xrightarrow{\mathcal{P}} \tilde{s}_k \xrightarrow{\theta_{k-m}} \tilde{a}_k \xrightarrow{\mathcal{P}} \tilde{s}_{k+1}.
$$

Throughout the proof, we will use $\theta, \theta', \bar{\lambda}, \bar{\lambda}', \xi, \tilde{\xi}$ to represent $\theta_k, \theta_{k-z}, \bar{\lambda}_k, \bar{\lambda}_{k-z}, \xi_k, \xi_{k-z}$, respectively.

For the ease of expression, define

$$
\Delta_2(\xi, \lambda, \theta) := \langle\lambda - \lambda^*(\theta), \bar{g}_r(\xi, \lambda) - g_r(\theta, \lambda)\rangle.
$$

We have

$$
\begin{aligned}
\langle\bar{\lambda}_k - \lambda^*(\theta), \bar{g}_r(\xi_k, \bar{\lambda}_k) - g_r(\theta_k, \bar{\lambda}_k)\rangle &= \Delta_2(\xi, \bar{\lambda}, \theta) \\
&= \underbrace{\Delta_2(\xi, \bar{\lambda}, \theta) - \Delta_2(\xi, \bar{\lambda}, \theta')}_{I_1} + \underbrace{\Delta_2(\xi, \bar{\lambda}, \theta') - \Delta_2(\xi, \bar{\lambda}', \theta')}_{I_2} \\
&\quad + \underbrace{\Delta_2(\xi, \bar{\lambda}', \theta') - \Delta_2(\tilde{\xi}, \bar{\lambda}', \theta')}_{I_3} + \underbrace{\Delta_2(\tilde{\xi}, \bar{\lambda}', \theta')}_{I_4}.
\end{aligned}
$$

I_1 can be expressed as

$$
\begin{aligned}
I_1 &= \langle \bar{\lambda} - \lambda^*(\theta), \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta, \bar{\lambda}) \rangle - \langle \bar{\lambda} - \lambda^*(\theta'), \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta', \bar{\lambda}) \rangle \\
&= \langle \bar{\lambda} - \lambda^*(\theta), \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta, \bar{\lambda}) \rangle - \langle \bar{\lambda} - \lambda^*(\theta), \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta', \bar{\lambda}) \rangle \\
&\quad + \langle \lambda^*(\theta) - \lambda^*(\theta'), \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta', \bar{\lambda}) \rangle \\
&\le \|\bar{\lambda} - \lambda^*(\theta)\| \|g_r(\theta', \bar{\lambda}) - g_r(\theta, \bar{\lambda})\| + \|\lambda^*(\theta) - \lambda^*(\theta')\| \|\bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta', \bar{\lambda})\|.
\end{aligned} \tag{66}
$$

The first term can be bounded as

$$
\begin{aligned}
\|\bar{\lambda} - \lambda^*(\theta)\| \|g_r(\theta', \bar{\lambda}) - g_r(\theta, \bar{\lambda})\| &\le 2R_\lambda \|\mathbb{E}_{\xi \sim \mu'_\theta}[\bar{g}_r(\xi, \bar{\lambda})] - \mathbb{E}_{\xi \sim \mu_\theta}[\bar{g}_r(\xi, \bar{\lambda})]\| \\
&\le 4R_\lambda \sup_\xi \|\bar{g}_r(\xi, \bar{\lambda})\| d_{TV}(\mu'_\theta \otimes \pi'_\theta \otimes \mathcal{P}, \mu_\theta \otimes \pi_\theta \otimes \mathcal{P}) \\
&\le 4R_\lambda C_\lambda d_{TV}(\mu'_\theta \otimes \pi'_\theta \otimes \mathcal{P}, \mu_\theta \otimes \pi_\theta \otimes \mathcal{P}) \\
&\le 4R_\lambda C_\lambda |\mathcal{A}| L_\pi (1 + \log_\rho \kappa^{-1} + (1-\rho)^{-1}) \|\theta - \theta'\|,
\end{aligned} \tag{67}
$$

where the first inequality follows the projection update of each lambda step, the third inequality is due to $\|\bar{g}_r(\xi, \bar{\lambda})\| \le C_\lambda$, and the last inequality follows Lemma 8.

The second term can be bounded as

$$
\|\lambda^*(\theta) - \lambda^*(\theta')\| \|\bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta, \bar{\lambda})\| \le 2C_\lambda L_\lambda \|\theta - \theta'\| \tag{68}
$$

Plug (67) and (68) into (66), we can bound I_1 as

$$
I_1 \le (4R_\lambda C_\lambda |\mathcal{A}| L_\pi (1 + \log_\rho \kappa^{-1} + (1-\rho)^{-1}) + 2C_\lambda L_\lambda) \|\theta - \theta'\|. \tag{69}
$$

Next we bound I_2 as

$$
\begin{aligned}
I_2 &= \langle \bar{\lambda} - \lambda^*(\theta'), \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta', \bar{\lambda}) \rangle - \langle \bar{\lambda}' - \lambda^*(\theta'), \bar{g}_r(\xi, \bar{\lambda}') - g_r(\theta', \bar{\lambda}') \rangle \\
&= \langle \bar{\lambda} - \lambda^*(\theta'), \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta', \bar{\lambda}) \rangle - \langle \bar{\lambda}' - \lambda^*(\theta'), \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta', \bar{\lambda}) \rangle \\
&\quad + \langle \bar{\lambda}' - \lambda^*(\theta'), \bar{g}_r(\xi, \bar{\lambda}) - \bar{g}_r(\xi, \bar{\lambda}') - g_r(\theta', \bar{\lambda}) + g_r(\theta', \bar{\lambda}') \rangle.
\end{aligned}
$$

The first two terms can be bounded as

$$
\langle \bar{\lambda} - \bar{\lambda}', \bar{g}_r(\xi, \bar{\lambda}) - g_r(\theta', \bar{\lambda}) \rangle \le 2C_\lambda \|\bar{\lambda} - \bar{\lambda}'\|. \tag{70}
$$

The last term can be bounded as

$$
\begin{aligned}
&\langle \bar{\lambda}' - \lambda^*(\theta'), \bar{g}_r(\xi, \bar{\lambda}) - \bar{g}_r(\xi, \bar{\lambda}') - g_r(\theta', \bar{\lambda}) + g_r(\theta', \bar{\lambda}') \rangle \\
&\le \|\bar{\lambda} - \lambda^*(\theta')\| (\|\bar{g}_r(\xi, \bar{\lambda}) - \bar{g}_r(\xi, \bar{\lambda}')\| + \|g_r(\theta', \bar{\lambda}') - g_r(\theta', \bar{\lambda})\|) \\
&\le 2R_\lambda (\|\bar{g}_r(\xi, \bar{\lambda}) - \bar{g}_r(\xi, \bar{\lambda}')\| + \|g_r(\theta', \bar{\lambda}') - g_r(\theta', \bar{\lambda})\|) \\
&\le 4R_\lambda \|\bar{\lambda} - \bar{\lambda}'\|,
\end{aligned} \tag{71}
$$

where the second inequality follows the projection of each lambda step. The last inequality is due to

$$
\begin{aligned}
\|\bar{g}_r(\xi, \bar{\lambda}) - \bar{g}_r(\xi, \bar{\lambda}')\| &= \|\varphi(s,a)(\varphi(s,a)^T(\bar{\lambda} - \bar{\lambda}'))\| \\
&\le \|\bar{\lambda} - \bar{\lambda}'\|
\end{aligned}
$$

Combine (70) and (71), we can bound I_2 as

$$
I_2 \le (4R_\lambda + 2C_\lambda) \|\bar{\lambda} - \bar{\lambda}'\|. \tag{72}
$$

We bound I_3 as

$$
\begin{aligned}
\mathbb{E}[I_3 | \theta', s_{k-z+1}] &= \mathbb{E}[\Delta_2(\xi, \theta', \bar{\lambda}') - \Delta_2(\tilde{\xi}, \theta', \bar{\lambda}') | \theta', s_{k-z+1}] \\
&\le 2 \sup_\xi |\Delta_2(\xi, \theta', \bar{\lambda}')| \, d_{TV}(\mathbb{P}(\xi \in \cdot | \theta', s_{k-z+1}), \mathbb{P}(\tilde{\xi} \in \cdot | \theta', s_{k-z+1})) \\
&\le 8R_\lambda C_\lambda d_{TV}(\mathbb{P}(\xi \in \cdot | \theta', s_{k-z+1}), \mathbb{P}(\tilde{\xi} \in \cdot | \theta', s_{k-z+1})) \\
&\le 4R_\lambda C_\lambda |\mathcal{A}| L_\pi \sum_{m=0}^{z-1} \|\theta_{k-m} - \theta_{k-z}\|.
\end{aligned} \tag{73}
$$

Here, the second inequality is due to $\|\Delta_2(\xi,\theta',\bar{\lambda}')\| \leq \|\lambda' - \lambda^*(\theta')\|\|\bar{g}_r(\xi,\lambda') - g_r(\theta',\lambda')\| \leq 4R_\lambda C_\lambda$, and the last inequality is according to Lemma 13.

We now bound I_4

$$
\begin{aligned}
\mathbb{E}[I_4|\theta',\bar{\lambda}',s_{k+z-1}] &= \mathbb{E}[\Delta_2(\tilde{\xi},\theta',\bar{\lambda}')|\theta',\bar{\lambda}',s_{k-z+1}] \\
&\leq \sup_\xi |\Delta_2(\xi,\theta',\bar{\lambda}')\|\mathbb{P}(\xi \in \cdot|\theta', s_{k-z+1}) - \mu_{\theta'} \otimes \pi_{\theta'} \otimes \mathcal{P}\| \\
&\leq 8R_\lambda C_\lambda d_{TV}(\mathbb{P}(\tilde{x} \in \cdot|\theta', s_{t-z+1}), \mu_{\theta'} \otimes \pi_{\theta'} \otimes \mathcal{P}) \\
&\leq 8R_\lambda C_\lambda \kappa \rho^{z-1},
\end{aligned} \tag{74}
$$

where the last inequality follows Lemma 5.

Plug (69), (72), (73), and (74) into (65), we get

$$
\begin{aligned}
\mathbb{E}[\Delta_2(\xi,\theta,\bar{\lambda})] \leq\ & (4R_\lambda C_\lambda|\mathcal{A}|L_\pi(1+\log_\rho \kappa^{-1} + (1-\rho)^{-1}) + 2C_\lambda L_\lambda)\mathbb{E}\|\theta_k - \theta_{k-z}\| \\
&+ (4R_\lambda + 2C_\lambda)\mathbb{E}\|\bar{\lambda}_k - \bar{\lambda}_{k-z}\| \\
&+ 4R_\lambda C_\lambda |\mathcal{A}| L_\pi \sum_{m=0}^{z-1} \mathbb{E}\|\theta_{k-m} - \theta_{k-z}\| \\
&+ 8R_\lambda C_\lambda \kappa \rho^{z-1},
\end{aligned}
$$

which completes the proof. □

D.3 Consensus error

Lemma 21 (bound of consensus error)**.** *Suppose Asssumptions 1 and 5 hold. Let $\boldsymbol{\omega_k}, \boldsymbol{\lambda_k}$ be the sequence generated by the algorithm 1, then for $k \geq 1$, the following hold*

$$
\sum_{i=1}^N \|\omega_k^i - \bar{\omega}_k\|^2 \leq \nu^{2k}\|\boldsymbol{\omega}_0\|_F + \frac{16NC_\delta^2}{1-\nu}\beta_k^2 + \frac{8\sqrt{N}C_\delta\|\boldsymbol{\omega}_0\|_F}{1-\nu}\nu^k\beta_k. \tag{75}
$$

$$
\sum_{i=1}^N \|\lambda_k^i - \bar{\lambda}_k\|^2 \leq \nu^{2k}\|\boldsymbol{\lambda}_0\|_F + \frac{16NC_\lambda^2}{1-\nu}\eta_k^2 + \frac{8\sqrt{N}C_\lambda\|\boldsymbol{\lambda}_0\|_F}{1-\nu}\nu^k\eta_k, \tag{76}
$$

where $\nu \in [0,1]$ is the second largest singular value of W. $\boldsymbol{\omega_k}, \boldsymbol{\lambda_k}$ are defined as

$$
\boldsymbol{\omega_k} := \begin{bmatrix} (\omega_k^1)^T \\ \vdots \\ (\omega_k^N)^T \end{bmatrix}, \qquad \boldsymbol{\lambda_k} := \begin{bmatrix} (\lambda_k^1)^T \\ \vdots \\ (\lambda_k^N)^T \end{bmatrix}.
$$

Proof. We will prove the bound in (75) for critic variables. The analysis for reward estimator in (76) follows the same routine. To simplify the notation, we will use g_k^i to represent $g_c^i(\xi_k, \omega_k^i)$ throughout the proof of this lemma. We also use e_k^i to represent the projection error $e_k^i := \prod_{R_\omega}(\omega_k^i - \beta_k g_k^i) - (\omega_k^i - \beta_k g_k^i)$. Also define $\bar{g}_k := \frac{1}{N}\sum_{i=1}^N g_k^i$; $\bar{e}_k := \frac{1}{N}\sum_{i=1}^N e_k^i$. The corresponding matrix exressions are

$$
G_k := \begin{bmatrix} (g_k^1)^T, \\ \vdots \\ (g_k^N)^T \end{bmatrix}, E_k := \begin{bmatrix} (e_k^1)^T, \\ \vdots \\ (e_k^N)^T \end{bmatrix}.
$$

Then the following equality holds by the update rule of critic variables

$$
\boldsymbol{\omega}_{k+1} = \begin{cases} W\boldsymbol{\omega}_k - \beta_k G_k + E_k, & \text{if} \quad k \bmod K_c = 0 \\ \boldsymbol{\omega}_k - \beta_k G_k + E_k, & \text{otherwise.} \end{cases} \tag{77}
$$

Let $Q := I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$, then the consensus error can be expressed as $\|\boldsymbol{\omega}_k - \mathbf{1}\bar{\omega}_k^T\|_F = \|Q\boldsymbol{\omega}_k\|_F$.

We bound the consensus error of critic's first

$$\|QG_k\| = \sqrt{\sum_{i=1}^{N} \|g_k^i - \bar{g}_k\|} \overset{(i)}{\leq} \sqrt{\sum_{i=1}^{N} 2\|g_k^i\|^2 + 2\|\bar{g}_k\|^2} \leq 2\sqrt{N}C_\delta. \tag{78}$$

$$\|QE_k\| = \sqrt{\sum_{i=1}^{N} \|e_t^i - \bar{e}_t\|} \leq \sqrt{\sum_{i=1}^{N} 2\|e_k^i\|^2 + 2\|\bar{e}_k\|^2} \overset{(ii)}{\leq} \sqrt{\sum_{i=1}^{N} 2\|g_k^i\|^2 + 2\|\bar{g}_k\|^2} \leq 2\beta_k\sqrt{N}C_\delta, \tag{79}$$

where (i) is due to $\|g_k^i\| \leq C_\delta$, (ii) is ensured by the convexity of the projection set.

We now study the consensus error of critic variables. Let $k' = \lfloor \frac{k}{K_c} \rfloor * K_c$. Without loss of generality, assume $k \bmod K_c \neq 0$. We have

$$\begin{aligned} Q\boldsymbol{\omega}_{k+1} &= QW\boldsymbol{\omega}_k - \beta_k QG_k + QE_k \\ &= WQ\boldsymbol{\omega}_k + \beta_k QG_k + QE_k \\ &= W^{k+1}Q\boldsymbol{\omega}_0 + \sum_{t=0}^{k} \beta_t W^{k-t} QG_t + \sum_{t=0}^{k} W^{k-t} QE_k, \end{aligned} \tag{80}$$

where the first equality follows (77). The second equality is due to the doubly stochasticity of matrix W (see Assumption 5): $QW = W - \frac{1}{N}\mathbf{1}\mathbf{1}^T W = W - \frac{1}{N}W\mathbf{1}\mathbf{1}^T = WQ$. The last equality expands the recursion of the second equation.

Take Frobenius norm on each side of (80) and apply triangle inequality, we get

$$\begin{aligned} \|Q\boldsymbol{\omega}_{k+1}\|_F &\leq \|W^k\boldsymbol{\omega}_0\|_F + \sum_{t=0}^{k} \beta_t \|W^{k-t}QG_t\|_F + \sum_{t=0}^{k} \|W^{k-t}QE_k\|_F \\ &\leq \nu^k \|\boldsymbol{\omega}_0\|_F + 4\sum_{t=0}^{k} \beta_t \nu^{k-t}\sqrt{N}C_\delta \\ &\leq \nu^k \|\boldsymbol{\omega}_0\|_F + \frac{4\sqrt{N}C_\delta \beta_k}{1-\nu}. \end{aligned} \tag{81}$$

The ν in (81) denotes the second largest singular value of W, which satisfies $\nu < 1$ as specified by Assumption 5. The second inequality uses (78), (79) and Lemma 9.

Take square on each side, we obtain

$$\|Q\boldsymbol{\omega}_{k+1}\|_F^2 \leq \nu^{2k}\|\boldsymbol{\omega}_0\|_F + \frac{16NC_\delta^2}{1-\nu}\beta_k^2 + \frac{8\sqrt{N}C_\delta\|\boldsymbol{\omega}_0\|_F}{1-\nu}\nu^k\beta_k$$

which completes the proof for (75). The proof of (76) follows similar procedure, we leave it as an exercise to reader.

□

D.4 Error of actor

Lemma 22. *Consider the sequence generated by Algorithm 1, for any $z \geq 1$ we have*

$$\begin{aligned} &\|\mathbb{E}_{\xi\sim\mu_{\theta_k}}[\delta(\xi,\theta_k)\psi_{\theta_k^i}(s_k,a_k^i)] - \mathbb{E}[\delta(\xi_k,\theta_k)\psi_{\theta_k^i}(s_k,a_k^i)]\| \\ &\leq 2C_\theta\kappa\rho^{z-1} + C_{12}\sum_{m=0}^{z-1}\|\theta_{k-m} - \theta_{k-z}\| + C_{13}\|\theta_k - \theta_{k-z}\| + C_{14}\|\theta_k^i - \theta_{k-z}^i\|, \end{aligned} \tag{82}$$

where $C_{12} := 2C_\theta|\mathcal{A}|L_\pi,\ C_{13} := |\mathcal{A}|L(\log_\rho \kappa^{-1} + (1-\rho)^{-1})C_\theta + 2(1+\gamma)L_V,\ C_{14} := 2C_\delta L_\psi$.

Proof. Consider the Markov chain since timestep $k - z$:

$$s_{k-z} \xrightarrow{\theta_{k-z}} a_{k-z} \xrightarrow{\mathcal{P}} s_{k-z+1} \xrightarrow{\theta_{k-z+1}} a_{k-z+1} \cdots \xrightarrow{\theta_{k-1}} a_{k-1} \xrightarrow{\mathcal{P}} s_k \xrightarrow{\theta_k} a_k \xrightarrow{\mathcal{P}} s_{k+1}.$$

Also consider the auxiliary Markov chain with fixed policy since timestep $k-z$:

$$s_{k-z} \xrightarrow{\theta_{k-z}} a_{k-z} \xrightarrow{\mathcal{P}} s_{k-z+1} \xrightarrow{\theta_{k-z}} \tilde{a}_{k-z+1} \cdots \xrightarrow{\theta_{k-z}} \tilde{a}_{k-1} \xrightarrow{\mathcal{P}} \tilde{s}_k \xrightarrow{\theta_{k-z}} \tilde{a}_k \xrightarrow{\mathcal{P}} \tilde{s}_{k+1}.$$

Throughout the proof of this lemma, we wil use ψ_{θ^i} to represent $\psi_{\theta^i}(s_k, a_k^i)$ for brevity.

We define the following notation for the ease of discussion

$$\Delta_3(\xi, \theta) := \mathbb{E}_{\xi \sim \mu_\theta}[\delta(\xi, \theta)\psi_{\theta^i}] - \delta(\xi, \theta)\psi_{\theta^i}].$$

Then our objective is to bound

$$\mathbb{E}[\|\Delta_3(\xi_k, \theta_k)\| \| \, \theta_{k-z}].$$

We decompose $\|\Delta_3(\xi_k, \theta_k)\|$ by applying triangle inequality

$$\begin{aligned} \|\Delta_3(\xi_k, \theta_k)\| \leq & \underbrace{\|\Delta_3(\xi_k, \theta_k) - \Delta_3(\xi_k, \theta_{k-z})\|}_{I_1} \\ & + \underbrace{\|\Delta_3(\xi_k, \theta_{k-z}) - \Delta_3(\tilde{\xi}_k, \theta_{k-z})\|}_{I_2} \\ & + \underbrace{\|\Delta_3(\tilde{\xi}_k, \theta_{k-z})\|}_{I_3}. \end{aligned} \tag{83}$$

We apply triangle inequality again to bound I_1 as

$$\begin{aligned} I_1 \leq & \underbrace{\|\delta(\xi_k, \theta_{k-z})\psi_{\theta^i_{k-z}} - \delta(\xi_k, \theta_k)\psi_{\theta^i_k}\|}_{I_1^{(1)}} \\ & + \underbrace{\|\mathbb{E}_{\xi \sim \mu_{\theta_k}}[\delta(\xi, \theta_k)\psi_{\theta^i_k}] - \mathbb{E}_{\xi \sim \mu_{\theta_{k-z}}}[\delta(\xi, \theta_{k-z})\psi_{\theta^i_{k-z}}]\|}_{I_1^{(2)}} \end{aligned} \tag{84}$$

$I_1^{(1)}$ can be bounded as

$$\begin{aligned} I_1^{(1)} &= \|\delta(\xi_k, \theta_{k-z})\psi_{\theta^i_{k-z}} - \delta(\xi_k, \theta_k)\psi_{\theta^i_k}\| \\ &\leq \|\delta(\xi_k, \theta_{k-z})\psi_{\theta^i_{k-z}} - \delta(\xi_k, \theta_k)\psi_{\theta^i_{k-z}}\| \\ &\quad + \|\delta(\xi_k, \theta_k)\psi_{\theta^i_{k-z}} - \delta(\xi_k, \theta_k)\psi_{\theta^i_k}\| \\ &\leq \||\gamma(V_{\theta_{k-z}}(s') - V_{\theta_k}(s')) + (V_{\theta_{k-z}}(s) - V_{\theta_{k-z}}(s'))|\psi^i_{k-z}\| \\ &\quad + \|\delta(\xi_k, \theta_k)\psi_{\theta^i_{k-z}} - \delta(\xi_k, \theta_k)\psi_{\theta^i_k}\| \\ &\leq (1+\gamma)L_V\|\theta_k - \theta_{k-z}\| + \|\delta(\xi_k, \theta_k)\psi_{\theta^i_{k-z}} - \delta(\xi_k, \theta_k)\psi_{\theta^i_k}\| \\ &\leq (1+\gamma)L_V\|\theta_k - \theta_{k-z}\| + C_\delta L_\psi\|\theta^i_k - \theta^i_{k-z}\|, \end{aligned} \tag{85}$$

where the second last inequality follows the Lipschitz continuous of value function in Lemma 7, and the last inequality uses Lipschitz continuous of ψ_{θ^i}.

$I_1^{(2)}$ can be bounded as

$$\begin{aligned} I_1^{(2)} &= \|\mathbb{E}_{\xi \sim \mu_{\theta_k}}[\delta(\xi, \theta_k)\psi_{\theta^i_k}] - \mathbb{E}_{\xi \sim \mu_{\theta_{k-z}}}[\delta(\xi, \theta_{k-z})\psi_{\theta^i_{k-z}}]\| \\ &= \|\mathbb{E}_{\xi \sim \mu_{\theta_k}}[\delta(\xi, \theta_{k-z})\psi_{\theta^i_{k-z}}] - \mathbb{E}_{\xi \sim \mu_{\theta_{k-z}}}[\delta(\xi, \theta_{k-z})\psi_{\theta^i_{k-z}}] \\ &\quad + \mathbb{E}_{\xi \sim \mu_{\theta_k}}[\delta(\xi, \theta_k)\psi_{\theta^i_k} - \delta(\xi, \theta_{k-z})\psi_{\theta^i_{k-z}}]\| \\ &\leq |\mathcal{A}|L(\log_\rho \kappa^{-1} + (1-\rho)^{-1})C_\theta\|\theta_k - \theta_{k-z}\| \\ &\quad + \|\mathbb{E}_{\xi \sim \mu_{\theta_k}}[\delta(\xi, \theta_k)\psi_{\theta^i_k} - \delta(\xi, \theta_{k-z})\psi_{\theta^i_{k-z}}]\| \\ &\leq |\mathcal{A}|L(\log_\rho \kappa^{-1} + (1-\rho)^{-1})C_\theta\|\theta_k - \theta_{k-z}\| \\ &\quad + (1+\gamma)L_V\|\theta_k - \theta_{k-z}\| + C_\delta L_\psi\|\theta^i_k - \theta^i_{k-z}\|, \end{aligned} \tag{86}$$

where the first inequality applies Lemma 8, and the last inequality uses the derivation in (85).

Combine (85) and (86), we have

$$\begin{aligned} I_1 \le\ & |\mathcal{A}|L(\log_\rho \kappa^{-1} + (1-\rho)^{-1})C_\theta \|\theta_k - \theta_{k-z}\| \\ & + 2(1+\gamma)L_V\|\theta_k - \theta_{k-z}\| + 2C_\delta L_\psi \|\theta_k^i - \theta_{k-z}^i\| \end{aligned} \tag{87}$$

We now bound I_2 as

$$\begin{aligned} \mathbb{E}[I_2] &= \mathbb{E}\|\delta(\tilde{\xi}_k, \theta_{k-z})\psi^i_{\theta_{k-z}} - \delta(\xi_k, \theta_{k-z})\psi^i_{\theta_{k-z}}\| \\ &\le 2\sup_\xi \|\delta(\xi, \theta_{k-z})\psi_{\theta^i_{k-z}}\| d_{TV}(P(\tilde{\xi}_k \in \cdot|\theta_{k-z}, s_{k-z}), P(\xi_k \in \cdot|\theta_{k-z}, s_{k-z})) \\ &\le 2C_\theta \sum_{m=0}^{z-1} |\mathcal{A}|L_\pi \|\theta_{k-m} - \theta_{k-z}\|, \end{aligned} \tag{88}$$

where the last inequality follows Lemma 13.

I_3 can be bounded as

$$\begin{aligned} I_3 &= \mathbb{E}\|\mathbb{E}_{\xi\sim\mu_{\theta_{k-z}}}[\delta(\xi, \theta_{k-z})\psi^i_{k-z}] - \delta(\tilde{\xi}_k, \theta_{k-z}\psi^i_{\theta_{k-z}})\| \\ &\le 2\sup_\xi \|\delta(\xi, \theta_{k-z})\psi^i_{\theta_{k-z}}\| d_{TV}(P(\tilde{\xi} \in \cdot|\theta_{k-z}, s_{k-z}), \mu_{\theta_{k-z}} \otimes \pi_{\theta_{k-z}} \otimes \mathcal{P}) \\ &\le 2C_\theta \kappa \rho^{z-1}, \end{aligned} \tag{89}$$

where the last inequality follows Lemma 5.

Plug (87), (88), and (89), we have

$$\begin{aligned} & \|\mathbb{E}_{\xi\sim\mu_{\theta_k}}[\delta(\xi, \theta_k)\psi_{\theta^i_k}(s_k, a^i_k)] - \mathbb{E}[\delta(\xi_k, \theta_k)\psi_{\theta^i_k}(s_k, a^i_k)]\| \\ & \le 2C_\theta\kappa\rho^{z-1} + 2C_\delta L_\psi\|\theta^i_k - \theta^i_{k-z}\| + 2C_\theta \sum_{m=0}^{z-1} |\mathcal{A}|L_\pi\|\theta_{k-m} - \theta_{k-z}\| \\ & \quad + (|\mathcal{A}|L(\log_\rho \kappa^{-1} + (1-\rho)^{-1})C_\theta + 2(1+\gamma)L_V)\|\theta_k - \theta_{k-z}\|, \end{aligned}$$

which completes the proof.

□

E Proof of main results

E.1 Proof of Theorem 1

In this section, we provide the analysis for i.i.d. sampling. By Lemma 4, we have

$$\begin{aligned}
\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) &\geq \mathbb{E}[\langle \nabla J(\theta_k), \theta_{k+1} - \theta_k \rangle] - \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= \sum_{i=1}^{N} \mathbb{E}[\langle \nabla_{\theta^i} J(\theta_k), \theta_{k+1}^i - \theta_k^i \rangle] - \frac{L}{2}\sum_{i=1}^{N} \|\theta_{k+1}^i - \theta_k^i\|^2 \\
&= \sum_{i=1}^{N} \mathbb{E}[\alpha_k \langle \nabla_{\theta^i} J(\theta_k), g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i) \rangle] - \frac{L}{2}\alpha_k^2 \sum_{i=1}^{N} \mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\|^2 \\
&\geq \sum_{i=1}^{N} [\frac{\alpha_k}{2}\|\nabla_{\theta^i} J(\theta_k)\|^2 + \frac{\alpha_k}{2}\|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 \\
&\quad - \frac{\alpha_k}{2}\|\nabla_{\theta^i} J(\theta_k) - \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2] - \frac{L}{2} N C_\theta^2 \alpha_k^2,
\end{aligned} \tag{90}$$

where the last inequality is due to $\|g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\| = \|\hat{\delta}(\xi_k, \omega_k^i, \lambda_k^i)\psi_{\theta_k^i}(s_k, a_k^i)\| \leq C_\delta C_\psi := C_\theta$.

For brevity, we will use $\psi_{\theta_k^i}$ to represent $\psi_{\theta_k^i}(s_k, a_k^i)$. The gradient bias can be bounded as

$$\begin{aligned}
&\|\nabla_{\theta^i} J(\theta_k) - \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)|\omega_{k+1}^i, \lambda_{k+1}^i]\|^2 \\
&\leq 4\underbrace{\|\nabla_{\theta^i} J(\theta_k) - \mathbb{E}[\delta(\xi_k, \theta_k)\psi_{\theta_k^i}]\|^2}_{I_1} \\
&\quad + 4\underbrace{\|\mathbb{E}[(\delta(\xi_k, \theta_k) - \tilde{\delta}(\xi_k, \omega^*(\theta_k)))\psi_{\theta_k^i}]\|^2}_{I_2} \\
&\quad + 4\underbrace{\|\mathbb{E}[(\tilde{\delta}(\xi_k, \omega^*(\theta_k)) - \tilde{\delta}(\xi_k, \omega_{k+1}^i))\psi_{\theta_k^i}]\|^2}_{I_3} \\
&\quad + 4\underbrace{\|\mathbb{E}[(\tilde{\delta}(\xi_k, \omega_{k+1}^i) - \hat{\delta}(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i))\psi_{\theta_k^i}]\|^2}_{I_4},
\end{aligned} \tag{91}$$

where the inequality uses $\|a + b + c + c\|^2 \leq 4\|a\|^2 + 4\|b\|^2 + 4\|c\|^2 + 4\|d\|^2$.

From now on, we will use $\xi \sim d_\theta$ to denote $s \sim d_{\pi_\theta}, a \sim \pi(\cdot|s), s' \sim \mathcal{P}$ for notational simplicity.

I_1 reflects the sampling error under perfect value function estimation of critic. It can be bounded as

$$\begin{aligned}
\mathbb{E}[I_1|\theta_k] &= \|\nabla_{\theta^i} J(\theta_k) - \mathbb{E}[\delta(\xi_k, \theta_k)\psi_{\theta_k^i}|\theta_k]\|^2 \\
&= \|\mathbb{E}_{\xi \sim d_{\theta_k}}[\delta(\xi, \theta_k)\psi_{\theta_k^i}|\theta_k] - \mathbb{E}_{\xi \sim \mu_{\theta_k}}[\delta(\xi, \theta_k)\psi_{\theta_k^i}|\theta_k]\|^2 \\
&\leq (2\sup_{\xi}|\bar{r}(s,a) + \gamma V_{\theta_k}(s') - V_{\theta_k}(s)|\, d_{TV}(\mu_{\theta_k} \otimes \pi_{\theta_k} \otimes \mathcal{P}, d_{\theta_k} \otimes \pi_{\theta_k} \otimes \mathcal{P}))^2 \\
&\leq (2 r_{\max} C_\psi d_{TV}(\mu_{\theta_k}, d_{\theta_k}))^2 \\
&\leq 16 C_\theta^2 (\log_\rho \kappa^{-1} + \frac{1}{\rho})^2 (1 - \gamma^2),
\end{aligned}$$

where the last inequality follows Lemma 6.

Define $\varepsilon_{sp} := 4C_\theta^2(\log_\rho \kappa^{-1} + \frac{1}{\rho})^2(1 - \gamma^2)$, then I_1 can be bounded as

$$I_1 \leq 4\varepsilon_{sp}. \tag{92}$$

The term I_2 describe the approximation quality of linear function class, it can be bounded as

$$
\begin{aligned}
I_2 &= \|\mathbb{E}[(\delta(\xi_k,\theta_k) - \tilde{\delta}(\xi_k,\omega^*(\theta_k)))\psi_{\theta_k^i}]\|^2 \\
&\overset{(i)}{\le} \mathbb{E}[|\delta(\xi_k,\theta_k) - \tilde{\delta}(\xi_k,\omega^*(\theta_k))|^2\|\psi_{\theta_k^i}\|^2] \\
&\overset{(ii)}{\le} C_\psi^2\mathbb{E}[|\gamma(V_{\theta_k}(s_{k+1}) - \phi(s_{k+1})^T\omega^*(\theta_k)) + (V_{\theta_k}(s_k) - \phi(s_k)^T\omega^*(\theta_k))|^2] \\
&\overset{(iii)}{\le} C_\psi^2(2\mathbb{E}[\gamma^2(V_{\theta_k}(s_{k+1}) - \phi(s_{k+1})^T\omega^*(\theta_k))^2] + 2\mathbb{E}[(V_{\theta_k}(s_k) - \phi(s_k)^T\omega^*(\theta_k))^2]) \\
&\overset{(iiii)}{\le} 2C_\psi^2(1+\gamma^2)\varepsilon_{app}^c \le 4C_\psi^2\varepsilon_{app}^c.
\end{aligned} \tag{93}
$$

where (i) applies triangle inequality and Cauchy Schwarz inequality, (ii) follows Assumption 3, (iii) uses $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, and $(iiii)$ follows the definition of $\varepsilon_{app}^c := \max_{\theta,a}\sqrt{\mathbb{E}_{s\sim\mu_\theta}[|V_{\pi_\theta}(s) - \hat{V}_{\omega^*(\theta)}(s)|^2]}$.

I_3 can be bounded as

$$
\begin{aligned}
\mathbb{E}[I_3] &= \|\mathbb{E}[(\tilde{\delta}(\xi_k,\omega^*(\theta_k)) - \tilde{\delta}(\xi_k,\omega_{k+1}^i))\psi_{\theta_k^i}]\|^2 \\
&\le \mathbb{E}[|\tilde{\delta}(\xi_k,\omega^*(\theta_k)) - \tilde{\delta}(\xi_k,\omega_{k+1}^i)|^2\|\psi_{\theta_k^i}\|^2] \\
&\le C_\psi^2\mathbb{E}[|\gamma\phi(s_k+1)^T(\omega^*(\theta_k) - \omega_{k+1}^i) - \phi(s_k)^T(\omega^*(\theta_k) - \omega_{k+1}^i)|^2] \\
&\le C_\psi^2(2\mathbb{E}[|\gamma\phi(s_{k+1})^T(\omega^*(\theta_k) - \omega_{k+1}^i)|^2] + 2\mathbb{E}[|\phi(s_k)^T(\omega^*(\theta_k - \omega_{k+1}^i))|^2]) \\
&\le C_\psi^2(2\gamma^2\mathbb{E}[\|\phi(s_{k+1})\|^2\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2] + 2\mathbb{E}[\|\phi(s_k)\|^2\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2]) \\
&\overset{(i)}{\le} 2C_\psi^2(1+\gamma^2)\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 \le 4C_\psi^2\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2.
\end{aligned} \tag{94}
$$

where the last inequality is due to $\|\phi(s)\| \le 1$, as specified by Assumption 1.

I_4 can be bounded as

$$
\begin{aligned}
\mathbb{E}[I_4] &= \|\mathbb{E}[(\tilde{\delta}(\xi_k,\omega_{k+1}^i) - \hat{\delta}(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i))\psi_{\theta_k^i}|\lambda_{k+1}^i]\|^2 \\
&\le \mathbb{E}[|\tilde{\delta}(\xi_k,\omega_{k+1}^i) - \hat{\delta}(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i)|^2\|\psi_{\theta_k^i}\|^2|\lambda_{k+1}^i] \\
&\le C_\psi^2\mathbb{E}[|\bar{r}(s_k,a_k) - \varphi(s_k,a_k)^T\lambda_{k+1}^i|^2|\lambda_{k+1}^i] \\
&\le C_\psi^2(2\mathbb{E}[|\bar{r}(s_k,a_k) - \varphi(s_k,a_k)^T\lambda^*(\theta_k)|^2] + 2\mathbb{E}[|\varphi(s_k,a_k)^T\lambda^*(\theta_k) - \varphi(s_k,a_k)^T\lambda_{k+1}^i|^2|\lambda_{k+1}^i]) \\
&\le 2C_\psi^2\varepsilon_{app}^r + 2C_\psi^2\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2
\end{aligned} \tag{95}
$$

Thus, the gradient bias for i_{th} agent can be bounded as

$$
\begin{aligned}
&\|\nabla_{\theta^i}F(\theta_k) - \mathbb{E}[g_a^i(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i)]\|^2 \\
&\le 16\varepsilon_{sp} + 16C_\psi^2\varepsilon_{app}^c + 16C_\psi^2\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 \\
&\quad + 8C_\psi^2\varepsilon_{app}^r + 8C_\psi^2\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2 \\
&\le 16(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}) + 16C_\psi^2\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 + 8C_\psi^2\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2,
\end{aligned} \tag{96}
$$

where the last inequality follows the definition of ε_{app}.

Plug (96) into (90) gives us

$$
\begin{aligned}
\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) \ge \sum_{i=1}^{N}(&\frac{\alpha_k}{2}\mathbb{E}\|\nabla_{\theta^i}J(\theta_k)\|^2 + \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i)\|^2 \\
&- 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 - 4C_\psi^2\alpha_k\mathbb{E}\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2) \\
&- \frac{L}{2}NC_\theta^2\alpha_k^2 - 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app})N\alpha_k.
\end{aligned} \tag{97}
$$

Consider the Lyapunov function

$$
\mathbb{V}_k := -J(\theta_k) + \|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + \|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2.
$$

The difference between two Lyapunov functions will be

$$\begin{aligned}
\mathbb{E}[\mathbb{V}_{k+1}] - \mathbb{E}[\mathbb{V}_k] &= \mathbb{E}[J(\theta_k)] - \mathbb{E}[J(\theta_{k+1})] + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\omega}_k - \omega^*(\theta_k)\|^2 \\
&\quad + \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 - \mathbb{E}\|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2 \\
&\le \sum_{i=1}^N (-\frac{\alpha_k}{2}\|\nabla_{\theta^i} J(\theta_k)\|^2 - \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i)\|^2) + \frac{L}{2} N C_\theta^2 \alpha_k^2 + 8(\varepsilon_{sp} + C_\psi^2 \varepsilon_{app}) N \alpha_k \\
&\quad + \underbrace{\sum_{i=1}^N 8C_\psi^2 \alpha_k \mathbb{E}\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\omega}_k - \omega^*(\theta_k)\|^2}_{I_5} \\
&\quad + \underbrace{\sum_{i=1}^N 4C_\psi^2 \alpha_k \mathbb{E}\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2 + \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2}_{I_6}
\end{aligned} \tag{98}$$

The first two terms of I_5 can be bounded as

$$\begin{aligned}
&\sum_{i=1}^N 8C_\psi^2 \alpha_k \mathbb{E}\|\omega^*(\theta_k) - \bar{\omega}_{k+1} + \bar{\omega}_{k+1} - \omega_{k+1}^i\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 \\
&= \sum_{i=1}^N 8C_\psi^2 \alpha_k \mathbb{E}\|\bar{\omega}_{k+1} - \omega_{k+1}^i\|^2 + 8C_\psi^2 \alpha_k \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 \\
&\le 8C_\psi^2 \alpha_k (\nu^{2k}\|\boldsymbol{\omega}_0\|_F + \frac{16NC_\delta^2}{1-\nu}\beta_k^2 + \frac{8\sqrt{N}C_\delta\|\boldsymbol{\omega}_0\|}{1-\nu}\nu^k \beta_k) \\
&\quad + 8C_\psi^2 \alpha_k \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2,
\end{aligned} \tag{99}$$

where the second equality is due to

$$\sum_{i=1}^N \langle \omega^*(\theta_k) - \bar{\omega}_{k+1}, \bar{\omega}_{k+1} - \omega_{k+1}^i \rangle = \langle \omega^*(\theta_k) - \bar{\omega}_{k+1}, \bar{\omega}_{k+1} - \bar{\omega}_{k+1} \rangle = 0,$$

and the last inequality follows the Lemma 21.

For the ease of expression, we define

$$M_{k_1} := 8C_\psi^2 (\nu^{2k}\|\boldsymbol{\omega}_0\|_F + \frac{16NC_\delta^2}{1-\nu}\beta_k^2 + \frac{8\sqrt{N}C_\delta\|\boldsymbol{\omega}_0\|_F}{1-\nu}\nu^k \beta_k). \tag{100}$$

Plug (100) into (99), we have

$$\begin{aligned}
I_5 &\le 8C_\psi^2 \alpha_k \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 + \alpha_k M_{k_1} \\
&\le (1 + 4L_{\omega,2}^2 N \alpha_k + 8C_\psi^2 \alpha_k + \frac{L_{\omega,2}^2}{2} C_\theta^2 N^2 \alpha_k^2) \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 \\
&\quad + (\frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2)\alpha_k^2 + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_1},
\end{aligned} \tag{101}$$

where the second inequality follows (31) in Lemma 15.

Let $C_9 := \min\{c \mid 4L_{\omega,2}^2 N\alpha_k + 8C_\psi^2\alpha_k + \frac{L_{\omega,2}^2}{2}C_\theta^2 N^2\alpha_k^2 \le c\alpha_k\}$. Plug the definition into (101), we get

$$\begin{aligned}
I_5 &\le (1+C_9\alpha_k)\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + (\frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2)\alpha_k^2 \\
&\quad + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_1} \\
&\le (1+C_9\alpha_k)(1-2\lambda_\phi\beta_k)\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + (1+C_9\alpha_k)C_\delta^2\beta_k^2 \\
&\quad + (\frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2)\alpha_k^2 + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_1}, \qquad (102)
\end{aligned}$$

where the last inequality follows (32) in Lemma 15.

By letting $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k$, we can ensure

$$(1+C_9\alpha_k)(1-2\lambda_\phi\beta_k) < 0.$$

Therefore, I_5 can be bounded as

$$I_5 \le (1+C_9\alpha_k)C_\delta^2\beta_k^2 + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_1} + (\frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2)\alpha_k^2. \qquad (103)$$

By applying Lemma 18 and following the similar procedure, we can bound I_6 as

$$I_6 \le (1+C_{10}\alpha_k)C_\lambda^2\eta_k^2 + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_2} + (\frac{L_{\lambda,2}^2 C_\theta^2 N^2}{2} + L_\lambda^2)\alpha_k^2, \qquad (104)$$

with $\eta_k = \frac{C_{10}}{2\lambda_\varphi}\alpha_k$ and

$$\begin{aligned}
C_{10} &:= \min\{c \mid 4\frac{L_{\lambda,2}^2}{2}C_\theta^2\alpha_k + 8C_\psi^2\alpha_k + \frac{L_{\lambda,2}^2 C_\delta^2}{2}\alpha_k^2 \le c\alpha_k\}, \\
M_{k_2} &:= 8C_\psi^2(\nu^{2k}\|\boldsymbol{\lambda}_0\|_F + \frac{16NC_\lambda^2}{1-\nu}\eta_k^2 + \frac{8\sqrt{N}C_\lambda\|\boldsymbol{\lambda}_0\|_F}{1-\nu}\nu^k\eta_k). \qquad (105)
\end{aligned}$$

Plug (103) and (104) into (98), we have

$$\begin{aligned}
\mathbb{E}[\mathbb{V}_{k+1}] - \mathbb{E}[\mathbb{V}_k] &\le \sum_{i=1}^N (-\frac{\alpha_k}{2}\|\nabla_{\theta^i}J(\theta_k)\|^2 - \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i)\|^2) + \frac{\alpha_k}{2}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 \\
&\quad + (1+C_9\alpha_k)C_\delta^2\beta_k^2 + (1+C_{10}\alpha_k)C_\lambda^2\eta_k^2 + (\frac{L}{2}NC_\theta^2 + C_{11})\alpha_k^2 \\
&\quad + (M_{k_1} + M_{k_2})\alpha_k + 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}N)\alpha_k, \\
&= \sum_{i=1}^N (-\frac{\alpha_k}{2}\|\nabla_{\theta^i}J(\theta_k)\|^2) + (M_{k_1} + M_{k_2})\alpha_k + 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}N)\alpha_k \\
&\quad + (1+C_9\alpha_k)C_\delta^2\beta_k^2 + (1+C_{10}\alpha_k)C_\lambda^2\eta_k^2 + (\frac{L}{2}NC_\theta^2 + C_{11})\alpha_k^2, \qquad (106)
\end{aligned}$$

where $C_{11} := \frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + \frac{L_{\lambda,2}^2 C_\theta^2 N^2}{2} + L_\omega^2 + L_\lambda^2$.

By telescoping (106), we get

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^K \sum_{i=1}^N \mathbb{E}\|\nabla_{\theta^i}J(\theta_k)\|^2 &\le \frac{2\mathbb{E}[\mathbb{V}_0]}{K\alpha_k} + 16(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}N) + \frac{2}{K}\sum_{k=0}^K (M_{k_1} + M_{k_2}) \\
&\quad + (1+C_9\alpha_k)C_\delta^2\frac{\beta_k^2}{\alpha_k} + (1+C_{10}\alpha_k)C_\lambda^2\frac{\eta_k^2}{\alpha_k} + (\frac{L}{2}NC_\theta^2 + C_{11})\alpha_k. \qquad (107)
\end{aligned}$$

The third term can be bounded as

$$
\begin{aligned}
&\frac{2}{K}\sum_{k=0}^{K}(M_{k_1}+M_{k_2}) \\
&= \frac{16C_\psi^2}{K}(\|\boldsymbol{\omega}_0\|_F+\|\boldsymbol{\lambda}_0\|_F)\sum_{k=1}^{K}\nu^{2k}+\frac{256NC_\psi^2}{(1-\nu)K}\sum_{k=0}^{K}(C_\delta^2\beta_k^2+C_\lambda^2\eta_k^2) \\
&\quad+\frac{128\sqrt{N}C_\psi^2}{(1-\nu)K}(\sum_{k=1}^{K}C_\delta\|\boldsymbol{\omega}_0\|_F\nu^k\beta_k+\sum_{k=1}^{K}C_\lambda\|\boldsymbol{\lambda}_0\|_F\nu^k\eta_k) \\
&\le \frac{16C_\psi^2}{K(1-\nu^2)}(\|\boldsymbol{\omega}_0\|_F+\|\boldsymbol{\lambda}_0\|_F)+\frac{256NC_\psi^2}{(1-\nu)}(C_\delta^2\beta_k^2+C_\lambda^2\eta_k^2) \\
&\quad+\frac{128\sqrt{N}C_\psi^2}{(1-\nu)^2K}(C_\delta\|\boldsymbol{\omega}_0\|_F\beta_k+C_\lambda\|\boldsymbol{\lambda}_0\|_F\eta_k) \\
&= o(\frac{1}{\sqrt{K}}),
\end{aligned}
\tag{108}
$$

where we use $\sum_{k=0}^{K}\nu^k\le\frac{1}{1-\nu}$ for the inequality.

Plug (108) back into (107) and let $\alpha_k=\frac{\bar\alpha}{\sqrt{K}}$ for some positive constant $\bar\alpha$, $\beta_k=\frac{C_9}{2\lambda_\phi}\alpha_k, \eta_k=\frac{C_{10}}{2\lambda_\varphi}\alpha_k$, we obtain the desired result.

E.2 Proof of Theorem 2

Following the proof under i.i.d. sampling in (90), we have

$$
\begin{aligned}
&\mathbb{E}[J(\theta_{k+1})]-J(\theta_k) \\
&\ge \sum_{i=1}^{N}[\frac{\alpha_k}{2}\|\nabla_{\theta^i}J(\theta_k)\|^2+\frac{\alpha_k}{2}\|\mathbb{E}[g_a^i(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i)]\|^2 \\
&\quad-\frac{\alpha_k}{2}\|\nabla_{\theta^i}J(\theta_k)-\mathbb{E}[g_a^i(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i)]\|^2-\frac{L}{2}NC_\theta^2\alpha_k^2.
\end{aligned}
\tag{109}
$$

By following the derivation of (91), the gradient bias can be bounded as (crf. $\psi_{\theta_k^i}:=\psi_{\theta_k^i}(s_k,a_k^i)$)

$$
\begin{aligned}
&\|\nabla_{\theta^i}J(\theta_k)-\mathbb{E}[g_a^i(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i)|\omega_{k+1}^i,\lambda_{k+1}^i]\|^2 \\
&\le 4\underbrace{\|\nabla_{\theta^i}J(\theta_k)-\mathbb{E}[\delta(\xi_k,\theta_k)\psi_{\theta_k^i}]\|^2}_{I_1} \\
&\quad+4\underbrace{\|\mathbb{E}[(\delta(\xi_k,\theta_k)-\tilde\delta(\xi_k,\omega^*(\theta_k)))\psi_{\theta_k^i}]\|^2}_{I_2} \\
&\quad+4\underbrace{\|\mathbb{E}[(\tilde\delta(\xi_k,\omega^*(\theta_k))-\tilde\delta(\xi_k,\omega_{k+1}^i))\psi_{\theta_k^i}]\|^2}_{I_3} \\
&\quad+4\underbrace{\|\mathbb{E}[(\tilde\delta(\xi_k,\omega_{k+1}^i)-\hat\delta(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i))\psi_{\theta_k^i}]\|^2}_{I_4},
\end{aligned}
\tag{110}
$$

We bound I_1 as

$$\begin{aligned} I_1 &= \|\nabla_{\theta^i} J(\theta_k) - \mathbb{E}[\delta(\xi_k,\theta_k)\psi_{\theta_k^i}|\theta_k]\|^2 \\ &= \|\mathbb{E}_{\xi\sim d_{\theta_k}}[\delta(\xi,\theta_k)\psi_{\theta_k^i}|\theta_k] - \mathbb{E}[\delta(\xi_k,\theta_k)\psi_{\theta_k^i}|\theta_k]\|^2 \\ &\le 2\underbrace{\|\mathbb{E}_{\xi\sim d_{\theta_k}}[\delta(\xi,\theta_k)\psi_{\theta_k^i}|\theta_k] - \mathbb{E}_{\xi\sim \mu_{\theta_k}}[\delta(\xi,\theta_k)\psi_{\theta_k^i}|\theta_k]\|^2}_{I_1^{(1)}} \\ &\quad + 2\underbrace{\|\mathbb{E}_{\xi\sim \mu_{\theta}}[\delta(\xi,\theta_k)\psi_{\theta_k^i}|\theta_k] - \mathbb{E}[\delta(\xi_k,\theta_k)\psi_{\theta_k^i}|\theta_k]\|^2}_{I_1^{(2)}} \end{aligned} \tag{111}$$

Follow the derivation of (92), we have

$$I_1^{(1)} \le 4\varepsilon_{sp}.$$

By Lemma 22, $I_1^{(2)}$ can be bounded as

$$\begin{aligned} I_1^{(2)} &\le (2C_\theta\kappa\rho^{z-1} + C_{12}\sum_{m=0}^{z-1}\|\theta_{k-m}-\theta_{k-z}\| + C_{13}\|\theta_k - \theta_{k-z}\| + C_{14}\|\theta_k^i - \theta_{k-z}^i\|)^2 \\ &\le (2C_\theta\kappa\rho^{z-1} + C_{12}\sum_{m=0}^{z-1}\sum_{n=1}^{z-m}\|\theta_{k-m-n+1}-\theta_{k-m}\| + C_{13}\sum_{n=1}^{z}\|\theta_{k-n+1} - \theta_{k-n}\| + C_{14}\sum_{n=1}^{z}\|\theta_{k-n+1}^i - \theta_{k-n}^i\|)^2 \\ &\le (2C_\theta\kappa\rho^{z-1} + C_{12}NC_\theta\frac{z(z+1)}{2}\alpha_{k-z} + C_{13}NzC_\theta\alpha_{k-z} + C_{14}zC_\theta\alpha_{k-z})^2 \\ &\le 16C_\theta^2\kappa^2\rho^{2z-2} + 2C_{12}^2C_\theta^2z^2\alpha_{k-z}^2 + 4C_{13}^2N^2z^2C_\theta^2\alpha_{k-z}^2 + 4C_{14}^2z^2C_\theta^2\alpha_{k-z}^2, \end{aligned} \tag{112}$$

where the second inequality uses triangle inequality, and the last inequality applies $(a+b+c+d)^2 \le 4a^2+4b^2+4c^2+4d^2$.

Let $z = Z_K$. Recall Z_K is defined as $Z_K := \min\{z\in\mathbb{N}^+|\kappa\rho^{z-1} \le \min\{\alpha_k,\beta_k,\eta_k\}\}$. Then we have

$$I_1^{(2)} \le C_{K_5}\alpha_{k-Z_K}^2, \tag{113}$$

where we define $C_{K_5} := 16C_\theta^2 + 2C_{12}^2C_\theta^2Z_K^2 + 4C_{13}^2N^2Z_K^2C_\theta^2 + 4C_{14}^2Z_K^2C_\theta^2$.

Thus, we have

$$I_1 \le 4\varepsilon_{sp} + C_{K_5}\alpha_{k-Z_K}^2. \tag{114}$$

The bound of I_2, I_3, and I_4 follows the analysis under i.i.d. sampling. Plug in (93), (94), and (95) will give us the bound of gradient bias

$$\begin{aligned} &\|\nabla_{\theta^i}F(\theta_k) - \mathbb{E}[g_a^i(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i)]\|^2 \\ &\le 16(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}) + 16C_\psi^2\|\omega^*(\theta_k)-\omega_{k+1}^i\|^2 \\ &\quad + 8C_\psi^2\|\lambda^*(\theta_k)-\lambda_{k+1}^i\|^2 + 4C_{K_5}\alpha_{k-Z_K}^2. \end{aligned}$$

Thus, we have

$$\begin{aligned} \mathbb{E}[J(\theta_{k+1})] - J(\theta_k) \ge \sum_{i=1}^{N}&(\frac{\alpha_k}{2}\mathbb{E}\|\nabla_{\theta^i}J(\theta_k)\|^2 + \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k,\omega_{k+1}^i,\lambda_{k+1}^i)\|^2 \\ &- 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k)-\omega_{k+1}^i\|^2 - 4C_\psi^2\alpha_k\mathbb{E}\|\lambda^*(\theta_k)-\lambda_{k+1}^i\|^2) \\ &- \frac{L}{2}NC_\theta^2\alpha_k^2 - 2NC_{K_5}\alpha_{k-Z_K}^2 - 8(\varepsilon_{sp}+C_\psi^2\varepsilon_{app})N\alpha_k. \end{aligned} \tag{115}$$

Consider the Lyapunov function

$$\mathbb{V}_k := -J(\theta_k) + \|\bar{\omega}_k - \omega^*(\theta_k)\|^2 + \|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2. \tag{116}$$

The difference between two Lyapunov functions will be

$$\begin{aligned}
\mathbb{E}[\mathbb{V}_{k+1}] - \mathbb{E}[\mathbb{V}_k] &= \mathbb{E}[J(\theta_k)] - \mathbb{E}[J(\theta_{k+1})] + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\omega}_k - \omega^*(\theta_k)\|^2 \\
&\quad + \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_k)\|^2 - \mathbb{E}\|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2 \\
&\le \sum_{i=1}^N (-\frac{\alpha_k}{2}\|\nabla_{\theta^i} J(\theta_k)\|^2 - \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i)\|^2) \\
&\quad + 2NC_{K_5}\alpha_{k-Z_K} + \frac{L}{2}NC_\theta^2\alpha_k^2 + 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app})N\alpha_k \\
&\quad + \underbrace{\sum_{i=1}^N 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\omega}_k - \omega^*(\theta_k)\|^2}_{I_5} \\
&\quad + \underbrace{\sum_{i=1}^N 4C_\psi^2\alpha_k\mathbb{E}\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2 + \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2}_{I_6}.
\end{aligned} \tag{117}$$

The first two terms of I_5 can be bounded as

$$\begin{aligned}
&\sum_{i=1}^N 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k) - \bar{\omega}_{k+1} + \bar{\omega}_{k+1} - \omega_{k+1}^i\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 \\
&= \sum_{i=1}^N 8C_\psi^2\alpha_k\mathbb{E}\|\bar{\omega}_{k+1} - \omega_{k+1}^i\|^2 + 8C_\psi^2\alpha_k\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 \\
&\le 8C_\psi^2\alpha_k\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 + \alpha_k M_{k_1} \\
&\le (1 + 4L_{\omega,2}^2 N\alpha_k + 8C_\psi^2\alpha_k + \frac{L_{\omega,2}^2}{2}C_\theta^2 N^2\alpha_k^2)\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 \\
&\quad + (\frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2)\alpha_k^2 + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_1},
\end{aligned} \tag{118}$$

where the equality is due to

$$\sum_{i=1}^N \langle \omega^*(\theta_k) - \bar{\omega}_{k+1}, \bar{\omega}_{k+1} - \omega_{k+1}^i \rangle = \langle \omega^*(\theta_k) - \bar{\omega}_{k+1}, \bar{\omega}_{k+1} - \bar{\omega}_{k+1} \rangle = 0.$$

The first inequality follows the Lemma 21, with M_{k_1} is defined in (100). The last inequality follows (39) in Lemma 16.

Plug (118) into (117), and recall $C_9 := \min\{c \mid 4L_{\omega,2}^2 N\alpha_k + 8C_\psi^2\alpha_k + \frac{L_{\omega,2}^2}{2}C_\theta^2 N^2\alpha_k^2 \le c\alpha_k\}$, we get

$$\begin{aligned}
I_5 &\le (1 + C_9\alpha_k)\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 + (\frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2)\alpha_k^2 \\
&\quad + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_1} \\
&\le (1 + C_9\alpha_k)(1 - 2\lambda_\phi\beta_k)\mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_k)\|^2 \\
&\quad + (1 + C_9\alpha_k)(C_{K_1}\beta_k\beta_{k-Z_K} + C_{K_2}\beta_k\alpha_{k-Z_K}) \\
&\quad + (\frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2)\alpha_k^2 + \frac{\alpha_k}{4}\sum_{i=1}^N \|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_1},
\end{aligned} \tag{119}$$

where the last inequality follows (40) in Lemma 16.

By letting $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k$, we can ensure

$$(1 + C_9\alpha_k)(1 - 2\lambda_\phi\beta_k) < 0.$$

Therefore, I_5 can be bounded as

$$\begin{aligned} I_5 \le \frac{\alpha_k}{4}\sum_{i=1}^{N}\|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_1} + (\frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2)\alpha_k^2 \\ + (1 + C_9\alpha_k)(C_{K_1}\beta_k\beta_{k-Z_K} + C_{K_2}\beta_k\alpha_{k-Z_K}). \end{aligned} \tag{120}$$

By applying Lemma 19 and following the similar procedure, we can bound I_6 as

$$\begin{aligned} I_6 \le \frac{\alpha_k}{4}\sum_{i=1}^{N}\|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k M_{k_2} + (\frac{L_{\lambda,2}^2 C_\theta^2 N^2}{2} + L_\lambda^2)\alpha_k^2 \\ + (1 + C_{10}\alpha_k)(C_{K_3}\eta_k\eta_{k-Z_K} + C_{K_4}\eta_k\alpha_{k-Z_K}). \end{aligned} \tag{121}$$

with $\eta_k = \frac{C_{10}}{2\lambda_\varphi}\alpha_k$, and M_{k_2} defined in (105).

Plug (120) and (121) into (117), we have

$$\begin{aligned} \mathbb{E}[\mathbb{V}_{k+1}] - \mathbb{E}[\mathbb{V}_k] \le \sum_{i=1}^{N} -\frac{\alpha_k}{2}\|\nabla_{\theta^i}J(\theta_k)\|^2 + (M_{k_1} + M_{k_2})\alpha_k \\ + (1 + C_9\alpha_k)(C_{K_1}\beta_k\beta_{k-Z_K} + C_{K_2}\beta_k\alpha_{k-Z_K}) \\ + (1 + C_{10}\alpha_k)(C_{K_3}\eta_k\eta_{k-Z_K} + C_{K_4}\eta_k\alpha_{k-Z_K}) \\ + (\frac{L}{2}NC_\theta^2 + C_{11})\alpha_k^2 + 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}N)\alpha_k, \end{aligned} \tag{122}$$

where we recall $C_{11} := \frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + \frac{L_{\lambda,2}^2 C_\theta^2 N^2}{2} + L_\omega^2 + L_\lambda^2$.

By letting $\alpha_k = \frac{\bar{\alpha}}{\sqrt{K}}$ for some positive constant $\bar{\alpha}$, and recall $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k, \eta_k = \frac{C_{10}}{2\lambda_\varphi}\alpha_k$, we can telescope (122) as

$$\begin{aligned} \frac{1}{K}\sum_{k=0}^{K}\sum_{i=1}^{N}\mathbb{E}\|\nabla_{\theta^i}J(\theta_k)\|^2 \le \frac{2\mathbb{E}[\mathbb{V}_0]}{K\alpha_k} + 16(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}N) + \frac{2}{K}\sum_{k=0}^{K}(M_{k_1} + M_{k_2}) \\ + (1 + C_9\alpha_k)(C_{K_1}\frac{\beta_k}{\alpha_k}\beta_{k-Z_K} + C_{K_2}\frac{\beta_k}{\alpha_k}\alpha_{k-Z_K}) \\ + (1 + C_{10}\alpha_k)(C_{K_3}\frac{\eta_k}{\alpha_k}\eta_{k-Z_K} + C_{K_4}\frac{\eta_k}{\alpha_k}\alpha_{k-Z_K}) \\ + (\frac{L}{2}NC_\theta^2 + C_{11})\alpha_k. \end{aligned} \tag{123}$$

The third term can be bounded as

$$\begin{aligned} \frac{2}{K}\sum_{k=0}^{K}(M_{k_1} + M_{k_2}) &= \frac{16C_\psi^2}{K}(\|\boldsymbol{\omega}_0\|_F + \|\boldsymbol{\lambda}_0\|_F)\sum_{k=1}^{K}\nu^{2k} + \frac{256NC_\psi^2}{(1-\nu)K}\sum_{k=0}^{K}(C_\delta^2\beta_k^2 + C_\lambda^2\eta_k^2) \\ &\quad + \frac{128\sqrt{N}C_\psi^2}{(1-\nu)K}(\sum_{k=1}^{K}C_\delta\|\boldsymbol{\omega}_0\|_F\nu^k\beta_k + \sum_{k=1}^{K}C_\lambda\|\boldsymbol{\lambda}_0\|_F\nu^k\eta_k) \\ &\le \frac{16C_\psi^2}{K(1-\nu^2)}(\|\boldsymbol{\omega}_0\|_F + \|\boldsymbol{\lambda}_0\|_F) + \frac{256NC_\psi^2}{(1-\nu)}(C_\delta^2\beta_k^2 + C_\lambda^2\eta_k^2) \\ &\quad + \frac{128\sqrt{N}C_\psi^2}{(1-\nu)^2K}(C_\delta\|\boldsymbol{\omega}_0\|_F\beta_k + C_\lambda\|\boldsymbol{\lambda}_0\|_F\eta_k) \\ &= o(\frac{1}{\sqrt{K}}), \end{aligned} \tag{124}$$

where we use $\sum_{k=0}^{K} \nu^k \leq \frac{1}{1-\nu}$ for the inequality.

Plug (124) back into (123). By noticing $C_{K_1} = \mathcal{O}(\log \frac{1}{\alpha_k}), C_{K_2} = \mathcal{O}(\log^2 \frac{1}{\alpha_k}), C_{K_3} = \mathcal{O}(\log \frac{1}{\alpha_k}), C_{K_4} = \mathcal{O}(\log^2 \frac{1}{\alpha_k})$, we obtain the desired result.

E.3 Proof of Theorem 3

Define the update of actor i using the noisy reward as

$$g_a^i(\epsilon_k, \omega_{k+1}^i) := \tilde{r}_{k,K_r}^i(s_k, a_k) + \gamma\phi(s')^T\omega_{k+1}^i - \phi(s)^T\omega_{k+1}^i. \tag{125}$$

Following the derivation of (90), we have

$$\begin{aligned}\mathbb{E}[J(\theta_{k+1}] - J(\theta_k) \geq \sum_{i=1}^{N}[\frac{\alpha_k}{2}\|\nabla_{\theta^i}J(\theta_k)\|^2 + \frac{\alpha_k}{2}\|\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i)]\|^2 \\ - \frac{\alpha_k}{2}\|\nabla_{\theta^i}J(\theta_k) - \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i)]\|^2] - \frac{L}{2}NC_\theta^2\alpha_k^2.\end{aligned} \tag{126}$$

Similarly to the proof of Theorem 1 and 2, the gradient bias term can be decomposed as as

$$\begin{aligned}\|\nabla_{\theta^i}J(\theta_k) - \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i)]\|^2 \leq 4\underbrace{\|\nabla_{\theta^i}J(\theta_k) - \mathbb{E}[\delta(\xi_k, \theta_k)\psi_{\theta_k^i}]\|^2}_{I_1} \\ + 4\underbrace{\|\mathbb{E}[(\delta(\xi_k, \theta_k) - \tilde{\delta}(\xi_k, \omega^*(\theta_k)))\psi_{\theta_k^i}]\|^2}_{I_2} \\ + 4\underbrace{\|\mathbb{E}[(\tilde{\delta}(\xi_k, \omega^*(\theta_k)) - \tilde{\delta}(\xi_k, \omega_{k+1}^i))\psi_{\theta_k^i}]\|^2}_{I_3} \\ + 4\underbrace{\|\mathbb{E}[(\bar{r}_k(s_k, a_k) - \tilde{r}_{k,K_r}(s_k, a_k))\psi_{\theta_k^i}]\|^2}_{I_4}\end{aligned} \tag{127}$$

I_1, I_2, I_3 can be bounded following the derivation of (114), (91), and (96), respectively. Plug these bounds into (127), we have

$$\begin{aligned}\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) \geq \sum_{i=1}^{N}(\frac{\alpha_k}{2}\mathbb{E}\|\nabla_{\theta^i}J(\theta_k)\|^2 + \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i)\|^2 - 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2) \\ - \sum_{i=1}^{N}\frac{\alpha_k}{2}C_\psi^2\|\bar{r}_k(s_k, a_k) - \tilde{r}_{k,K_r}^i(s_k, a_k)\|^2 - \frac{L}{2}NC_\theta^2\alpha_k^2 \\ - 2NC_{K_5}\alpha_{k-Z_K}^2 - 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app})N\alpha_k.\end{aligned} \tag{128}$$

Define $\tilde{r}_{k,K_r} := [r_{k,K_r}^1, \cdots, r_{k,K_r}^N]^T$. The reward bias can be bounded as

$$\begin{aligned}\sum_{i=1}^{N}\|\bar{r}_k(s_k, a_k) - \tilde{r}_{k,K_r}^i(s_k, a_k)\|^2 &= \|Q\tilde{r}_{k,K_r}\|^2 \\ &= \|QW^{K_r}\tilde{r}_{k,0}(s_k, a_k)\|^2 \\ &\leq \nu^{2K_r}\|\tilde{r}_{k,0}(s_k, a_k)\|^2 \\ &= \nu^{2K_r}\sum_{i=1}^{N}(\|\tilde{r}_{k,0}^i(s_k, a_k) - \bar{r}_k(s_k, a_k)\|^2 + \|\bar{r}_k(s_k, a_k)\|^2) \\ &\leq \nu^{2K_r}N(\sigma^2 + r_{\max}),\end{aligned} \tag{129}$$

where σ^2 is the variance of the reward noise. Let $K_r = \frac{1}{2}\log_\nu \alpha_k$ and define $C_{15} := \sigma^2 + r_{\max}^2$. Plug (128) back to (127), we have

$$\begin{aligned}\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) \geq \sum_{i=1}^{N}(\frac{\alpha_k}{2}\mathbb{E}\|\nabla_{\theta^i}J(\theta_k)\|^2 + \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i)\|^2 - 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2) \\ + \frac{N}{2}(C_{15} + C_\theta^2 L)\alpha_k^2 - 2NC_{K_5}\alpha_{k-Z_K}^2 - 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app})N\alpha_k.\end{aligned}$$

Consider the Lyapunov function

$$\mathbb{V}_k := -J(\theta_k) + \|\bar{\omega}_k - \omega^*(\theta_k)\|^2.$$

The difference between two Lyapunov functions is

$$\begin{aligned}\mathbb{E}[\mathbb{V}_{k+1}] - \mathbb{E}[\mathbb{V}_k] \leq \sum_{i=1}^{N}&(-\frac{\alpha_k}{2}\|\nabla_{\theta^i} J(\theta_k)\|^2 - \frac{\alpha_k}{2}\mathbb{E}\|g_a^i(\xi_k, \omega_{k+1}^i)\|^2) \\ &+ \frac{N}{2}C_{16}\alpha_k^2 - 2NC_{K_5}\alpha_{k-Z_K}^2 - 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app})N\alpha_k \\ &+ \underbrace{\sum_{i=1}^{N} 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\omega}_k - \omega^*(\theta_k)\|^2}_{I_5}.\end{aligned}$$

I_5 can be bounded by following the derivation of (120). Thus, we have

$$\begin{aligned}&\mathbb{E}[\mathbb{V}_{k+1}] - \mathbb{E}[\mathbb{V}_k] \\ &\leq \sum_{i=1}^{N} -\frac{\alpha_k}{2}\|\nabla_{\theta^i} J(\theta_k)\|^2 + \frac{N}{2}C_{16}\alpha_k^2 - 2NC_{K_5}\alpha_{k-Z_K}^2 - 8(\varepsilon_{sp} + C_\psi^2\varepsilon_{app})N\alpha_k \\ &\quad + (1 + C_9\alpha_k)(C_{K_1}\beta_k\beta_{k-Z_K} + C_{K_2}\beta_k\alpha_{k-Z_K}) + M_{k_1}\alpha_k, \end{aligned} \tag{130}$$

where $C_{16} := C_{15} + C_\theta^2 L + \frac{L_{\omega,2}^2 C_\theta^2 N^2}{2} + L_\omega^2$.

Telescoping (130), we have

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K}\sum_{i=1}^{N}\mathbb{E}\|\nabla_{\theta^i} J(\theta_k)\|^2 \leq \frac{2\mathbb{E}[\mathbb{V}_0]}{K\alpha_k} + 16(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}N) + \frac{2}{K}\sum_{k=0}^{K} M_{k_1} + C_{16}\alpha_k \\ + (1 + C_9\alpha_k)(C_{K_1}\frac{\beta_k}{\alpha_k}\beta_{k-Z_K} + C_{K_2}\frac{\beta_k}{\alpha_k}\alpha_{k-Z_K}).\end{aligned}$$

The term $\frac{2}{K}\sum_{k=0}^{K} M_{k_1}$ has been bounded in (124). Let $\alpha_k = \frac{\bar{\alpha}}{\sqrt{K}}$ for some positive constant $\bar{\alpha}$, $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k$ will yield the desired rate.

F Natural AC variant and its convergence

In this section, we propose a natural Actor-Critic variant of Algorithm 1, where the approach of calculating the natural policy graident under the decentralized setting is mainly inspired by [6]. We show that the gradient norm square of such an algorithm will convergence with the optimal sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-3})$. Moreover, the algorithm will converge to the *global optimum* with the sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-4})$. In the rest of this section, we first explain the update of the algorithm, and then prove its convergence.

F.1 Decentralized natural Actor-Critic

The natural policy gradient (NPG) algorithm [12] can be viewed as a preconditioned policy gradient algorithm, which updates as follow:

$$\theta_{k+1} = \theta_k - \alpha_k F(\theta_k)^{-1}\nabla J(\theta_k), \tag{131}$$

where $F(\theta) := \mathbb{E}_{s\sim d_{\pi_\theta}, a\sim\pi_\theta}\left[\psi_\theta(s,a)\psi_\theta(s,a)^T\right]$ is the Fisher information matrix (FIM).[3] The natural Actor-Critic (NAC) uses the critic variable to estimate the gradient. The main challenge for implementing NAC lies in the estimation of the inverse matrix-vector product $F(\theta_k)^{-1}\nabla J(\theta_k)$,

[3] Throughout the discussion, we assume that FIM is invertible and thus positive-definite.

Algorithm 3: Decentralized single-timescale NAC

Initialize: Actor parameter θ_0, critic parameter ω_0, reward estimator parameter λ_0, initial state s_0, natural policy gradient estimation $h_{k,0}$.
for $k = 0, \cdots, K-1$ **do**
  Option 1: i.i.d. sampling:
  $s_k \sim \mu_{\theta_k}(\cdot), a_k \sim \pi_{\theta_k}(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$.
  Option 2: Markovian sampling:
  $a_k \sim \pi_{\theta_k}(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$.

  Periodical consensus: Compute $\tilde{\omega}_k^i$ and $\tilde{\lambda}_k^i$ by (4) and (7).

  for $i = 0, \cdots, N$ **in parallel do**
    Reward estimator update: Update λ_{k+1}^i by (8).
    Critic update: Update ω_{k+1}^i by (5).
    Actor update:
      Collect N_a transition samples based on Markovian/i.i.d sampling.
      for $k' = 1, \cdots, K_a$ **do**
        Estimate $\bar{z}_{k',n}, \ \forall n \in [N_a]$ using (133).
        Update $h_{k,k'+1}$ by (135).
      end for
      Update θ_{k+1}^i by (136).
  end for
end for

especially under the decentralized setting. The work [6] proposes to solve the following strongly convex problem in order to estimate the product in a decentralized way

$$h(\theta_k) = \arg\min_h f_{\theta_k}(h) := \frac{1}{2} h^T F(\theta_k) h - \nabla J(\theta_k)^T h. \tag{132}$$

Such a problem can be solved by using (stochastic) gradient descent, where the gradient is calculated by $F(\theta_k)h - \nabla J(\theta_k)$. For the centralized setting, the gradient w.r.t. each agent can be approximated as $\frac{1}{N_a}\sum_{n=1}^{N_a} \psi_{\theta_k}^i(s_n, a_n^i)\psi_{\theta_k}(s_n, a_n)^T h - g_a^i(\xi_n, \omega_{k+1}, \lambda_{k+1})$. However, when considering the decentralized setting, the term $\bar{z}_n := \psi_{\theta_k}(s_n, a_n)^T h = \sum_{i=1}^N \psi_{\theta_k}^i(s_n, a_n)^T h^i$ is not accessible for each agent. Therefore, to approximate this value, agents compute $z_{n,0}^i := \psi_{\theta_k}^i(s_n, a_n)^T h^i$ locally and then perform the following communication step for K_z steps

$$z_{n,k'+1}^i = \sum_{j=1}^N W^{ij} z_{n,k'}^i, \ \forall n \in [N_a], \ k' = 0, \cdots, K_z - 1. \tag{133}$$

As we will see, $N z_{n,k'}^i$ converges to $\bar{z}_n$ linearly. Thus, the gradient of agent i can be approximated as

$$\widetilde{\nabla} f_{\theta_k}^i(h_{k,k'}) := \frac{N}{N_a}\sum_{n=1}^{N_a} \psi_{\theta_k}^i(s_n, a_n^i) z_{n,K_z}^i - g_a^i(\xi_k, \omega_{k+1}, \lambda_{k+1}). \tag{134}$$

Then, each agent i performs the following update for K_a steps to estimate the natural policy gradient direction as

$$h_{k,k'+1}^i = \Pi_{C_h}(h_{k,k'}^i - \varrho \widetilde{\nabla} f_{\theta_k}^i(h_{k,k'})), \tag{135}$$

where ϱ is a positive constant step size. Since the norm of optimal direction is bounded by $C_h := \lambda_{\max}(F(\theta)^{-1})C_\theta$, we project the vector into a ball of norm C_h for each update. Finally, we perform the approximate natural policy gradient step as

$$\theta_{k+1}^i = \theta_k^i - \alpha_k h_{k,K_a}^i. \tag{136}$$

F.2 Convergence of natural Actor-Critic

In this section, we establish the sample complexity of Algorithm 3. We first introduce an additional assumption.

Assumption 6. *(invertible FIM) There exists a positive constant λ_F such that for all policy θ, $\lambda_{\min}(F(\theta)) \geq \lambda_F$.*

Assumption 6 ensures that $F(\theta)$ is positive definite so that the problem (132) is strongly convex. Such an assumption is commonly adopted; see [6, 36, 17].

We now show the sample complexity of the Alrgroithm 3 in terms of gradient norm square and the global optimal gap. We consider the i.i.d. sampling to simplify the proof. We remark that the proof for Markovian sampling follows the similar analysis, with additional $\mathcal{O}(\log(\varepsilon^{-1}))$ error terms caused by Markov chain mixing.

Theorem 4. *Suppose Assumptions 1-6 hold. Consider the update of Algorithm 3 under i.i.d. sampling. Let $\alpha_k = \frac{\bar{\alpha}}{\sqrt{K}}$ for some positive constant $\bar{\alpha}$, $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k$, $\varrho \leq \frac{1}{2C_\psi^2}$, $N_a = \mathcal{O}(\sqrt{K})$, $K_a = \mathcal{O}(\log(K^{1/2}))$, $K_c = \mathcal{O}(\log(K^{1/4}))$. Then, the following hold*

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta^i}F(\theta_k)\|^2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right) + \mathcal{O}(\varepsilon_{app} + \varepsilon_{sp}) \tag{137}$$

$$\frac{1}{K}\sum_{k=0}^{K} J(\theta^*) - J(\theta_k) \leq \mathcal{O}\left(\frac{1}{K^{1/4}}\right) + \mathcal{O}(\varepsilon_{app} + \varepsilon_{sp} + \varepsilon_{actor}). \tag{138}$$

Based on Theorem 4, Algorithm 3 needs $K = \mathcal{O}(\varepsilon^{-2})$ iterations to achieve ε-error for gradient norm square, and thus attains sample complexity of $KN_aK_a = \widetilde{\mathcal{O}}(\varepsilon^{-3})$, which matches the best existing sample complexity of NAC [35, 6]. In terms of the global optimality gap, the algorithm requires $K = \mathcal{O}(\varepsilon^{-4})$ iterations to achieve ε-error, and thus has $KN_aK_a = \widetilde{\mathcal{O}}(\varepsilon^{-6})$ sample complexity. Such a sample complexity is much worse than the best existing sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ [35, 6].

We now explain the intuition of the gap for the sample complexity. Mimicking the analysis of [6] allows to establish the following inequality

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K} J(\theta^*) - \mathbb{E}[J(\theta_k)] \leq{}& \mathcal{O}\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{E}[\|\nabla_{\theta^i}J(\theta_k)\|^2]\right) \\ &+ \mathcal{O}\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\|\omega_k^i - \omega^*(\theta_k)\|\right) + \mathcal{O}\left(\frac{1}{K\alpha_k}\right).\end{aligned}$$

While our analysis can obtain the iteration complexity of $\mathcal{O}(\frac{1}{\sqrt{K}})$ for $\|\nabla J(\theta_k)\|^2$, we can only achieve $\mathcal{O}(\frac{1}{K^{1/4}})$ iteration complexity for critic's error $\|\omega_k - \omega^*(\theta_k)\|$. This is because our algorithm uses single-timescale update, where the critic's error inevitably converges slower than that of double-loop based algorithms which have $\mathcal{O}(\frac{1}{\sqrt{K}})$ complexity for the critic's error at each iteration. Therefore, the sample complexity in terms of global optimality gap of our single-timescale NAC is dominated by this critic's error term, resulting in the final complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-6})$.

We remark that this sample complexity result is based on a straightforward application of the analysis of [6], which is designed for double-loop algorithm. Therefore, such a proof technique may not be the tightest one for our single-timescale NAC (intuitively, the result is not tight). We leave the research on the improvement of such highly suboptimal results of single-timescale NAC as a future work.

F.3 Proof of Theorem 4

By Lemma 4, we have

$$\begin{aligned}\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) &\geq \sum_{i=1}^{N}\mathbb{E}\langle\nabla_{\theta^i}J(\theta_k), \theta_{k+1}^i - \theta_k^i\rangle - \frac{L}{2}\sum_{i=1}^{N}\|\theta_{k+1}^i - \theta_k^i\|^2 \\ &\overset{(i)}{\geq} \sum_{i=1}^{N}\alpha_k\mathbb{E}\langle\nabla_{\theta^i}J(\theta_k), h_k^i\rangle - \frac{L}{2}NC_h^2\alpha_k^2\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} [\alpha_k \mathbb{E}\langle \nabla_{\theta^i} J(\theta_k), F(\theta_k)^{-1} g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\rangle \\
&\quad + \alpha_k \mathbb{E}\langle \nabla_{\theta^i} J(\theta_k), h_k^i - F(\theta_k)^{-1} g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\rangle] - \frac{L}{2} N C_h^2 \alpha_k^2 \\
&\overset{(ii)}{=} \sum_{i=1}^{N} [\alpha_k \mathbb{E}\langle F(\theta_k)^{-1/2} \nabla_{\theta^i} J(\theta_k), F(\theta_k)^{-1/2} g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\rangle \\
&\quad + \alpha_k \mathbb{E}\langle \nabla_{\theta^i} J(\theta_k), h_k^i - F(\theta_k)^{-1} g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\rangle] - \frac{L}{2} N C_h^2 \alpha_k^2 \\
&= \sum_{i=1}^{N} [\frac{\alpha_k}{2} \|F(\theta_k)^{-1/2} \nabla_{\theta^i} J(\theta_k)\|^2 + \frac{\alpha_k}{2} \|F(\theta_k)^{-1/2} \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 \\
&\quad - \frac{\alpha_k}{2} \|F(\theta_k)^{-1/2} \nabla_{\theta^i} J(\theta_k) - F(\theta_k)^{-1/2} \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 \\
&\quad + \alpha_k \mathbb{E}\langle \nabla_{\theta^i} J(\theta_k), h_k^i - F(\theta_k)^{-1} g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\rangle] - \frac{L}{2} N C_\theta^2 \alpha_k^2 \\
&\overset{(iii)}{\geq} \sum_{i=1}^{N} [\frac{\alpha_k}{4} C_\psi^{-2} \|\nabla_{\theta^i} J(\theta_k)\|^2 + \frac{\alpha_k}{2} \lambda_F \|F(\theta_k)^{-1} \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 \\
&\quad - \frac{\alpha_k}{2} \lambda_F^{-1} \underbrace{\|\nabla_{\theta^i} J(\theta_k) - \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2}_{I_1} \\
&\quad - \alpha_k C_\psi^2 \underbrace{\|\mathbb{E}[h_k^i] - F(\theta_k)^{-1} \mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2}_{I_2}] - \frac{L}{2} N C_\theta^2 \alpha_k^2,
\end{aligned} \tag{139}
$$

where (i) is due to $\|\theta_{k+1}^i - \theta_k^i\| \leq C_h := \lambda_F C_\theta$. Note that we use h_k^i to represent h_{k,K_a}^i for simplifying the notation. (ii) uses decomposition of positive definite (PD) matrix. Specifically, let A be PD matrix, then by eigenvalue decomposition, $A = V \Lambda V^T$ for some orthonormal matrix V. Define $A^{-1/2} := V \Lambda^{1/2} V^T$, then $\langle x, Ay\rangle = \langle A^{1/2} x, A^{1/2} y\rangle$ for any x and y. (iii) uses $\lambda_F \leq \lambda(F(\theta)) \leq C_\psi^2$, $\forall \theta$.

I_1 represents the error of gradient bias, which we have bounded when analyzing the error of AC. By (96), we have

$$
I_1 \leq 16(\varepsilon_{sp} + C_\psi^2 \varepsilon_{app}) + 16 C_\psi^2 \|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 + 8 C_\psi^2 \|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2. \tag{140}
$$

To bound I_2, we need to bound the error of $h_{k,k'}$. We start with the gradient bias when estimating $h_{k,k'}$. Define $\overline{\nabla} f_{k,k'}(h_{k,k'}) := \nabla F(\theta_k) h_{k,k'} - \mathbb{E}[g_a(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]$, then it is easy to see that $\overline{\nabla} f_{k,k'}(h_{k,k'})$ is the unbiased gradient of the following problem

$$
\frac{1}{2} h_{k,k'}^T \nabla F(\theta_k) h_{k,k'} - \mathbb{E}[g_a(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]^T h_{k,k'}.
$$

Define the following notation for the ease of expression

$$
\begin{aligned}
\widehat{\nabla} f_{k,k'}^i(h_{k,k'}) &:= \frac{1}{N_a} \sum_{n=1}^{N_a} \psi_{\theta_k^i}(s_n, a_n^i) \psi_{\theta_k}(s_n, a_n)^T h_{k,k'} - g_a^i(\xi_{k,k'}, \omega_{k+1}^i, \lambda_{k+1}^i) \\
\widehat{\nabla} f_{k,k'}(h_{k,k'}) &:= [\widehat{\nabla} f_{k,k'}^1(h_{k,k'}), \cdots, \widehat{\nabla} f_{k,k'}^N(h_{k,k'})] \\
\widetilde{\nabla} f_{k,k'}^i(h_{k,k'}) &:= \frac{N}{N_a} \sum_{n=1}^{N_a} \psi_{\theta_k^i}(s_n, a_n^i) z_{n,K_z}^i - g_a^i(\xi_{k,k'}, \omega_{k+1}^i, \lambda_{k+1}^i) \\
\widetilde{\nabla} f_{k,k'}(h_{k,k'}) &:= [\widetilde{\nabla} f_{k,k'}^1(h_{k,k'}), \cdots, \widetilde{\nabla} f_{k,k'}^N(h_{k,k'})].
\end{aligned}
$$

We now analyze the error at outer-loop iteration k. For notational simplicity, we omit the subscript k for the prementioned notations, e.g. we use $\widehat{\nabla} f_{k'}^i(h_{k'})$, $\widehat{\nabla} f_{k'}(h_{k'})$, $\widetilde{\nabla} f_{k'}^i(h_{k'})$, $\widetilde{\nabla} f_{k'}(h_{k'})$ to represent the above notations, respectively.

$$\|\overline{\nabla} f_{k'}(h_{k'}) - \widetilde{\nabla} f_{k'}(h_{k'})\|^2 \leq 2\underbrace{\|\overline{\nabla} f_{k'}(h_{k'}) - \widehat{\nabla} f_{k'}(h_{k'})\|^2}_{I_3} + 2\underbrace{\|\widehat{\nabla} f_{k'}(h_{k'}) - \widetilde{\nabla} f_{k'}(h_{k'})\|^2}_{I_4}.$$

I_3 can be bounded as

$$\begin{aligned} I_3 &= \|\sum_{n=1}^{N_a}(\frac{1}{N_a}\psi_\theta(s_n,a_n)\psi_\theta(s_n,a_n)^T - F(\theta))h_{k'}\|^2 \\ &\leq \|\sum_{n=1}^{N_a}(\frac{1}{N_a}\psi_\theta(s_n,a_n)\psi_\theta(s_n,a_n)^T - F(\theta))\|^2 C_h^2 \\ &\leq \frac{1}{N_a}C_\psi^4 C_h^2. \end{aligned} \tag{141}$$

I_4 can be bounded as

$$\begin{aligned} I_4 &= \sum_{i=1}^{N}\left\|\psi_{\theta^i}(s_n,a_n^i)\left(\frac{1}{N_a}\sum_{n=1}^{N_a} N z_{n,K_z}^i - \psi_\theta(s_n,a_n)^T h_{k'}\right)\right\|^2 \\ &\leq \frac{1}{N_a}NC_\psi^2\sum_{i=1}^{N}\sum_{n=1}^{N_a}\|z_{n,K_z}^i - \bar{z}_{n,K_z}\|^2 \\ &= \frac{NC_\psi^2}{N_a}\sum_{n=}^{N_a}\|QW^{K_z}z_{n,0}\|^2 \\ &\leq \frac{NC_\psi^2}{N_a}\sum_{n=1}^{N_a}\nu^{K_z}\|z_{n,0}\|^2 \leq NC_\psi^4C_h^2\nu^{K_z}. \end{aligned} \tag{142}$$

Let $K_z = \min\{c \in \mathbb{N}^+ | \nu^c \leq \frac{4}{N_a N}\}$, then $K_z = \mathcal{O}(\log\frac{1}{N_a})$. Combine (141) and (142) gives us

$$\|\overline{\nabla} f_{k'}(h_{k'}) - \widetilde{\nabla} f_{k'}(h_{k'})\|^2 \leq \frac{4C_\psi^4 C_h^2}{N_a}.$$

We now analyze the error of $h_{k,k'}$. Note that we omit the subscript k here for simplifying notation. Define

$$h^* = \arg\min_h \bar{f}_\theta(h) := h^T F(\theta) h := -\mathbb{E}_{\xi\sim\mu_\theta}[g_a(\xi,\omega,\lambda)]^T h. \tag{143}$$

It is easy to see that the function on the RHS is strongly convex, since $F(\theta)$ is positive definite w.r.t. h. We bound the optimal gap by

$$\begin{aligned} \mathbb{E}\|h_{k'+1} - h^*\|^2 &= \mathbb{E}\|h_{k'} - \varrho\widetilde{\nabla} f_{k'}(h_{k'}) - h^*\|^2 \\ &= \mathbb{E}\|h_{k'} - h^*\|^2 - 2\varrho\mathbb{E}\langle h_{k'} - h^*, \widetilde{\nabla} f_{k'}(h_{k'})\rangle + \varrho^2\|\widetilde{\nabla} f_{k'}(h_{k'})\|^2 \\ &\leq \mathbb{E}\|h_{k'} - h^*\|^2 - 2\varrho\mathbb{E}\langle h_{k'} - h^*, \overline{\nabla} f_{k'}(h_{k'})\rangle + 2\varrho\mathbb{E}\langle h_{k'} - h^*, \overline{\nabla} f_{k'}(h_{k'}) - \widetilde{\nabla} f_{k'}(h_{k'})\rangle \\ &\quad + 2\varrho^2\|\overline{\nabla} f_{k'}(h_{k'})\|^2 + 2\varrho^2\|\widetilde{\nabla} f_{k'}(h_{k'}) - \overline{\nabla} f_{k'}(h_{k'})\|^2 \\ &\overset{(i)}{\leq} (1-\varrho\lambda_F)\mathbb{E}\|h_{k'} - h^*\|^2 - 2\varrho(f_{k'}(h_{k'}) - \overline{f}^*) + 2\varrho\mathbb{E}\langle h_{k'} - h^*, \overline{\nabla} f_{k'}(h_{k'}) - \widetilde{\nabla} f_{k'}(h_{k'})\rangle \\ &\quad + 2\varrho^2\|\overline{\nabla} f_{k'}(h_{k'})\|^2 + 2\varrho^2\|\widetilde{\nabla} f_{k'}(h_{k'}) - \overline{\nabla} f_{k'}(h_{k'})\|^2 \\ &\overset{(ii)}{\leq} (1-\varrho\lambda_F)\mathbb{E}\|h_{k'} - h^*\|^2 - 2\varrho(1 - 2\varrho C_\psi^2)(f_{k'}(h_{k'}) - \overline{f}^*) \\ &\quad + 2\varrho\mathbb{E}\langle h_{k'} - h^*, \overline{\nabla} f_{k'}(h_{k'}) - \widetilde{\nabla} f_{k'}(h_{k'})\rangle + 2\varrho^2\|\widetilde{\nabla} f_{k'}(h_{k'}) - \overline{\nabla} f_{k'}(h_{k'})\|^2 \\ &\overset{(iii)}{\leq} (1-\varrho\lambda_F)\mathbb{E}\|h_{k'} - h^*\|^2 + 2\varrho\mathbb{E}\langle h_{k'} - h^*, \overline{\nabla} f_{k'}(h_{k'}) - \widetilde{\nabla} f_{k'}(h_{k'})\rangle \\ &\quad + 2\varrho^2\|\widetilde{\nabla} f_{k'}(h_{k'}) - \overline{\nabla} f_{k'}(h_{k'})\|^2 \\ &\overset{(iiii)}{\leq} (1 - \frac{\varrho\lambda_F}{2})\mathbb{E}\|h_{k'} - h^*\|^2 + (\frac{2\varrho}{\lambda_F} + 2\varrho^2)\|\widetilde{\nabla} f_{k'}(h_{k'}) - \overline{\nabla} f_{k'}(h_{k'})\|^2, \end{aligned}$$

where $\overline{f}^*$ is the optimal value of $\overline{f}(h)$ defined in (143), and the inequality follows the property of λ_F-strongly convex function: $\overline{f}(h_2) \geq \overline{f}(h_1) + \langle \nabla \overline{f}(h_1), h_2 - h_2 \rangle + \frac{\lambda_F}{2}\|h_1 - h_2\|^2$, $\forall h_1, h_2$. (ii) uses the PL condition implied by λ_F-strong convexity: $\overline{f}(h^*) - \overline{f}(h) \leq -\frac{1}{2\lambda_F}\|\nabla \overline{f}(h)\|^2$, $\forall h$. (iii) is due to step size rule that $\varrho \leq \frac{1}{2C_\psi^2}$. $(iiii)$ applies Young's inequality.

Use the above induction, we have

$$\begin{aligned}\mathbb{E}\|h_{K_a} - h^*\|^2 &\leq (1 - \frac{\varrho\lambda_F}{2})^{K_a}\|h_0 - h^*\|^2 + \sum_{t=0}^{K_a}(1 - \frac{\varrho\lambda_F}{2})^t(\frac{2\varrho}{\lambda_F} + 2\varrho^2)\|\overline{\nabla} f_{K_a - t}(h_{K_a}) - \widetilde{\nabla} f_{K_a}(h_{K_a})\|^2 \\ &\leq 4C_h^2(1 - \frac{\varrho\lambda_F}{2})^{K_a} + (\frac{4\varrho}{\varrho\lambda_F^2} + \frac{4\varrho}{\lambda_F})C_\psi^4 C_h^2 \frac{4}{N_a}.\end{aligned}$$

Let $K_a = \min\{c \in \mathbb{N}^+ | 4C_h^2(1 - \frac{\varrho\lambda_F}{2})^c = (\frac{4\varrho}{\varrho\lambda_F^2} + \frac{4\varrho}{\lambda_F})C_\psi^4 C_h^2 \frac{1}{N_a}\}$, then $K_a = \mathcal{O}(\log(\frac{1}{N_a}))$. Define $C_{18} := (\frac{16\varrho}{\varrho\lambda_F^2} + \frac{16\varrho}{\lambda_F})C_\psi^4 C_h^2$, we have

$$I_2 = \mathbb{E}\|h_{K_a} - h^*\|^2 \leq \frac{2C_{18}}{N_a}. \tag{144}$$

Plug (140) and (144) back to (139), we have

$$\begin{aligned}\mathbb{E}[J(\theta_{k+1})] - J(\theta_k) \geq \sum_{i=1}^N [&\frac{\alpha_k}{4}C_\psi^{-2}\|\nabla_{\theta^i} J(\theta_k)\|^2 + \frac{\alpha_k}{2}\lambda_F\|F(\theta_k)^{-1}\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k C_\psi^2 \frac{2C_{18}}{N_a} \\ &+ 8\lambda_F^{-1}(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}) + 8\lambda_F^{-1}C_\psi^2\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 + 4\lambda_F^{-1}C_\psi^2\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2]\end{aligned}$$

Consider the Lyapunov function

$$\mathbb{V}^k = -J(\theta_k) + \lambda_F^{-1}(\|\omega_k - \omega^*(\theta_k)\|^2 + \|\lambda_k - \lambda^*(\theta_k)\|^2).$$

The difference of the Lyapunov function is

$$\begin{aligned}\mathbb{E}[\mathbb{V}^{k+1}] - \mathbb{E}[\mathbb{V}^k] &= \mathbb{E}[J(\theta_k)] - \mathbb{E}[J(\theta_{k+1})] + \lambda_F^{-1}(\mathbb{E}\|\omega_{k+1} - \omega^*(\theta_{k+1})\|^2 - \mathbb{E}\|\omega_k - \omega^*(\theta_k)\|^2 \\ &\quad + \mathbb{E}\|\lambda_{k+1} - \lambda^*(\theta_{k+1})\|^2 - \mathbb{E}\|\lambda_k - \lambda^*(\theta_k)\|^2) \\ &\leq \sum_{i=1}^N \left[\frac{\alpha_k}{4}C_\psi^{-2}\mathbb{E}\|\nabla_{\theta^i} J(\theta_k)\|^2 + \frac{\alpha_k}{2}\lambda_F\|F(\theta_k)^{-1}\mathbb{E}[g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)]\|^2 + \alpha_k C_\psi^2 \frac{2C_{18}}{N_a}\right] \\ &\quad + \lambda_F^{-1}\underbrace{\left[\sum_{i=1}^N 8C_\psi^2\alpha_k\mathbb{E}\|\omega^*(\theta_k) - \omega_{k+1}^i\|^2 + \mathbb{E}\|\bar{\omega}_{k+1} - \omega^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\omega}_k - \omega^*(\theta_k)\|^2\right]}_{I_5} \\ &\quad + \lambda_F^{-1}\underbrace{\left[\sum_{i=1}^N 4C_\psi^2\alpha_k\mathbb{E}\|\lambda^*(\theta_k) - \lambda_{k+1}^i\|^2 + \mathbb{E}\|\bar{\lambda}_{k+1} - \lambda^*(\theta_{k+1})\|^2 - \mathbb{E}\|\bar{\lambda}_k - \lambda^*(\theta_k)\|^2\right]}_{I_6} \\ &\quad + 8N\lambda_F^{-1}(\varepsilon_{sp} + C_\psi^2\varepsilon_{app}).\end{aligned} \tag{145}$$

By following the similar procedures through (98) to (106), we can bound I_5 and I_6 as

$$I_5 \leq (1 + C_{19}\alpha_k)C_\delta^2\beta_k^2 + \frac{\alpha_k}{4}\lambda_F^{-1}\sum_{i=1}^N \mathbb{E}\|F(\theta_k)^{-1}g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\|^2 + \alpha_k M_{k_1} + C_{20}\alpha_k^2 \tag{146}$$

$$I_6 \leq (1 + C_{21}\alpha_k)C_\lambda^2\eta_k^2 + \frac{\alpha_k}{4}\lambda_F^{-1}\sum_{i=1}^N \mathbb{E}\|F(\theta_k)^{-1}g_a^i(\xi_k, \omega_{k+1}^i, \lambda_{k+1}^i)\|^2 + \alpha_k M_{k_2} + C_{22}\alpha_k^2, \tag{147}$$

where $C_{19}, C_{20}, C_{21}, C_{22}$ are some positive constants. Plug (146) and (147) back to (145), we have

$$\mathbb{E}[\mathbb{V}^{k+1}] - \mathbb{E}[\mathbb{V}^k] \leq \sum_{i=1}^{N} [\frac{\alpha_k}{4} C_\psi^{-2} \mathbb{E}\|\nabla_{\theta^i} J(\theta_k)\|^2 + \alpha_k C_\psi^2 \frac{2C_{18}}{N_a} + \mathcal{O}(\alpha_k^2 + \beta_k^2 + \eta_k^2) + (M_{k_1} + M_{k_2})\alpha_k + \mathcal{O}(\varepsilon_{sp} + \varepsilon_{app})\alpha_k]. \tag{148}$$

By telescoping (148), we can get

$$\frac{1}{K}\sum_{k=0}^{K}\sum_{i=1}^{N} \mathbb{E}\|\nabla_{\theta^i} J(\theta_k)\|^2 \leq \frac{4C_\psi^2 \mathbb{V}_0}{K\alpha_k} + \mathcal{O}(\varepsilon_{sp} + \varepsilon_{app}) + \frac{8C_\psi^2 C_{18}}{N_a} + \mathcal{O}(\alpha_k + \frac{\beta_k^2}{\alpha_k} + \frac{\eta_k^2}{\alpha_k}) + 4C_\psi^2(M_{k_1} + M_{k_2})$$

By (108), $M_{k_1} + M_{k_2} = \mathcal{O}(\frac{1}{\sqrt{K}})$ when $K_c \leq \mathcal{O}(K^{1/4})$. Therefore, let $C, \bar{\alpha}$ be some positive constants. Set $N_a = C\sqrt{K}$, $\alpha_k = \frac{\bar{\alpha}}{\sqrt{K}}$, $\beta_k = \frac{C_9}{2\lambda_\phi}\alpha_k$, $\eta_k = \frac{C_{10}}{2\lambda_\varphi}\alpha_k$, we obtain the desired result of (137).

We now prove (138). Let $\mathbb{E}_{\theta^*}$ denote the expectation over $s \sim d_{\pi_{\theta^*}}, a \sim \pi_{\theta^*}(\cdot|s)$. We begin with the descent of policy gap as

$$\begin{aligned}
&\mathbb{E}_{\theta^*}[\log \pi_{\theta_{k+1}}(a|s) - \log \pi_{\theta_k}(a|s)] \\
&\geq \alpha_k \mathbb{E}_{\theta^*}[\psi_{\theta_k}(s,a)^T h_k] - \frac{L_\psi \alpha_k^2}{2} C_h^2 \\
&\geq \alpha_k \mathbb{E}_{\theta^*}[\psi_{\theta_k}(s,a)^T (h_k - h^*(\theta_k))] + \alpha_k \mathbb{E}_{\theta^*}[\psi_{\theta_k}(s,a)^T h^*(\theta_k) - A_{\theta_k}(s,a)] \\
&\quad + \alpha_k \mathbb{E}_{\theta^*}[A_{\theta_k}(s,a)] - \frac{L_\psi \alpha_k^2}{2} C_h^2 \\
&\geq -\alpha_k C_\psi \|h_k - h^*(\theta_k)\| - \alpha_k \sqrt{\varepsilon_{actor}} + \alpha_k (J(\theta^*) - J(\theta_k)) - \frac{L_\psi \alpha_k^2}{2} C_h^2.
\end{aligned}$$

By telescoping the above inequality and rearranging terms, we have

$$\frac{1}{K}\sum_{k=1}^{K}(J(\theta^*) - J(\theta_k)) \leq \frac{1}{K\alpha_k}\mathbb{E}_{\theta^*}[\log \pi_K(a|s) - \log \pi_0(a|s)] + \sqrt{\varepsilon_{actor}} + \frac{1}{K}\sum_{k=1}^{K} C_\psi \|h_k - h^*(\theta_k)\| + \frac{1}{K}\sum_{k=1}^{K}\frac{L_\psi \alpha_k}{2}.$$

The term $\|h_k - h^*(\theta_k)\| \leq \|h_k - F(\theta_k)^{-1}\mathbb{E}[g_a(\xi_k, \omega_{k+1}, \lambda_{k+1}]\| + \|\mathbb{E}[g_a(\xi_k, \omega_{k+1}, \lambda_{k+1}] - F^{-1}\nabla J(\theta_k)\|$. Since by the (144) and (96), these two terms are of order $\mathcal{O}(\frac{1}{N_a^{1/2}})$ and $\mathcal{O}(\|\omega_k - \omega_{k+1}\| + \varepsilon_{app})$, respectively, we conclude that $\|h_k - h^*(\theta_k)\|$ is of order $\mathcal{O}(\|\omega_k - \omega^*(\theta_k)\| + \varepsilon_{app})$. By following the step size rule as suggested by Theorem 4, we obtain the desired result.

G Overview of communication complexity

The Table 1 compares related works in terms of sample complexity and communication complexity.

Setting	Paper	Update	Sampling	Sample complexity	Communication complexity
Single-agent AC	[32]	Two-timescale	Markovian	$\widetilde{\mathcal{O}}(\varepsilon^{-\frac{5}{2}})$	-
	[35]	Double-loop	Markovian	$\widetilde{\mathcal{O}}(\varepsilon^{-2})$	-
Decentralized AC	[42]	Two-timescale	Markovian	Asymptotic	-
	[38]	Two-timescale	i.i.d.	$\mathcal{O}(\varepsilon^{-\frac{5}{2}})$	$\mathcal{O}(\varepsilon^{-\frac{5}{2}})$
	[6]	Double-loop	Markovian	$\widetilde{\mathcal{O}}(\varepsilon^{-2})$	$\widetilde{\mathcal{O}}(\varepsilon^{-1})$
	[11]	Double-loop	Markovian	$\widetilde{\mathcal{O}}(\varepsilon^{-2})$	$\widetilde{\mathcal{O}}(\varepsilon^{-1})$
	This work	**Single-timescale**	Markovian	$\widetilde{\mathcal{O}}(\varepsilon^{-2})$	$\widetilde{\mathcal{O}}(\varepsilon^{-\frac{3}{2}})$

Table 1: Comparison of some existing sample complexity results. The symbol $\tilde{\mathcal{O}}(\cdot)$ hides the logarithmic terms.

H Policy gradient theorem

The following derivation establishes the policy gradient update of our algorithm.

$$
\begin{aligned}
\nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi^\theta}(s_0)] &= \mathbb{E}_{s_0 \sim \mu_0}\left[\nabla \sum_{a_0} \pi_\theta\left(a_0|s_0\right) Q_{\pi_\theta}(s_0, a_0)\right] \\
&= \mathbb{E}_{s_0 \sim \mu_0}\left[\underbrace{\sum_{a_0} \nabla \pi_\theta(a_0|s_0) Q_{\pi_\theta}(s_0, a_0)}_{\text{1st term on RHS of (7)}} + \underbrace{\sum_{a_0} \pi_\theta(a_0|s_0) \nabla Q_{\pi_\theta}(s_0, a_0)}_{\text{2nd term on RHS of (7)}}\right] \\
&= \mathbb{E}_{s_0 \sim \mu_0}\left[\sum_{a_0} \pi_\theta(a_0|s_0) \nabla \log \pi_\theta(a_0|s_0) Q_{\pi_\theta}(s_0, a_0)\right] \\
&\quad + \mathbb{E}_{s_0 \sim \mu_0}\left[\sum_{a_0} \pi_\theta(a_0|s_0) \nabla \left(r(s_0, a_0) + \gamma \sum_{s_1} P(s_1|s_0, a_0) V_{\pi_\theta}(s_1)\right)\right] \\
&= \mathbb{E}_{s_0 \sim \mu_0}\left[\sum_{a_0} \pi_\theta(a_0|s_0) \nabla \log \pi_\theta(a_0|s_0) Q_{\pi_\theta}(s_0, a_0) + \gamma \sum_{a_0, s_1} \pi_\theta(a_0|s_0) \nabla V_{\pi_\theta}(s_1)\right] \\
&= \mathbb{E}_\tau\left[Q_{\pi_\theta}(s_0, a_0) \nabla \log \pi_\theta(a_0|s_0)\right] + \gamma \mathbb{E}_\tau[\nabla V_{\pi_\theta}(s_1)],
\end{aligned}
$$

where the (7) in the second inequality refers to equation (7) of [4], and the expectation on τ is taken over a trajectory: $a_0 \sim \pi_\theta(\cdot|s_0), s_1 \sim P(s_1|s_0, a_0), \cdots$. By expanding the above recursion, we can derive the policy gradient

$$
\begin{aligned}
\nabla \mathbb{E}_{s_0 \sim \mu_0}[V_{\pi_\theta}(s_0)] &= \mathbb{E}_\tau\left[\sum_{k=0}^{\infty} \gamma^k Q_{\pi_\theta}(s_k, a_k) \nabla \log \pi_\theta(a_k, s_k)\right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\pi_\theta}, a \sim \pi_\theta}\left[Q_{\pi_\theta}(s, a) \nabla \log \pi_\theta(a|s)\right].
\end{aligned}
$$