Learning to Coordinate in Multi-Agent Systems: A Coordinated Actor-Critic Algorithm and Finite-Time Guarantees

Anonymous authors
Paper under double-blind review

Abstract

Multi-agent reinforcement learning (MARL) has attracted much research attention recently. However, unlike its single-agent counterpart, many theoretical and algorithmic aspects of MARL have not been well-understood. In this paper, we study the emergence of coordinated behavior by autonomous agents using an actor-critic (AC) algorithm. Specifically, we propose and analyze a class of coordinated actor-critic (CAC) algorithms in which individually parametrized policies have a *shared* part (which is jointly optimized among all agents) and a *personalized* part (which is only locally optimized). Such a kind of *partially personalized* policy allows agents to coordinate by leveraging peers' experience and adapt to individual tasks. The flexibility in our design allows the proposed CAC algorithm to be used in a *fully decentralized* setting, where the agents can only communicate with their neighbors, as well as in a *federated* setting, where the agents occasionally communicate with a server while optimizing their (partially personalized) local models. Theoretically, we show that under some standard regularity assumptions, the proposed CAC algorithm requires $\mathcal{O}(\epsilon^{-\frac{5}{2}})$ samples to achieve an ϵ-stationary solution (defined as the solution whose squared norm of the gradient of the objective function is less than ϵ). To the best of our knowledge, this work provides the first finite-sample guarantee for decentralized AC algorithm with partially personalized policies.

1 Introduction

We consider the multi-agent reinforcement learning (MARL) problem, in which a common environment is influenced by the joint actions of multiple autonomous agents, each aiming to optimize their own individual objective. The MARL (Zhang et al., 2019; Lee et al., 2020) has received significant attention recently due to their outstanding performance in many practical applications including robotics (Stone & Veloso, 2000), autonomous driving (Shalev-Shwartz et al., 2016) and video games (Tampuu et al., 2017). Many efficient algorithms have been proposed (Lowe et al., 2017; Espeholt et al., 2018; Rashid et al., 2018), but unlike its single-agent counterpart, the theoretical understanding of MARL is still very limited, especially in settings where there is no central controller to coordinate different agents, so that the information sharing is limited (Zhang et al., 2019).

An important subclass of MARL – the so-called *cooperative* MARL – has become popular recently due to its wide applications. In the cooperative MARL, the agents aim to collaborate with each other to learn and optimize a joint global objective. To this end, local information exchange and local communication may be used to jointly optimize a system-level performance measure (Zhang et al., 2018; Grosnit et al., 2021; Zhang et al., 2021; Lu et al., 2021). Next, we provide a brief survey about related works in cooperative MARL, and discuss their settings as well as theoretical guarantees.

Related Works. The systematic study of the cooperative MARL can be traced back to Claus & Boutilier (1998); Wolpert et al. (1999), which extended Q-learning algorithm (Watkins & Dayan, 1992) or its variants to the multi-agent setting. More recently, there are a number of works that characterize the theoretical performance of cooperative MARL algorithms in a fully observable, decentralized setting (Kar et al., 2012; Zhang et al., 2018; Doan et al., 2019; Wang et al., 2020). In such a setting, the agents are connected by a time-varying graph, and they can only communicate with their immediate neighbors. Each agent observes the global state of the networked system and independently executes an action based on its own policy. Based on the joint actions by all agents, the system will transit into the next state and the *local* rewards will be received. The goal of the agents is to cooperatively maximize certain global reward, by communicating local information with their neighbors. Under the above cooperative MARL setting, there are several lines of works which studied different problem formulations, proposed new algorithms and analyzed their theoretical performance.

The first line of works about the coorperative and fully observable MARL has focused on developing and analyzing policy evaluation algorithms, where the agents jointly estimate the global value function for a given policy. In Wai et al. (2018), a decentralized double averaging primal-dual optimization algorithm was proposed to solve the mean squared projected Bellman error minimization problem. It is shown that the proposed algorithm converges to the optimal solution at a global geometric rate. In Doan et al. (2019), the authors obtained a finite-sample analysis for decentralized TD(0) method. Their analysis is closely related to the theoretical results of decentralized stochastic gradient descent method on convex optimization problems (Nedic et al., 2010).

However, the problem becomes much more challenging when the agents are allowed to optimize their policies. A recent line of works has focused on applying and analyzing various policy optimization methods in the MARL setting. In Zhang et al. (2018), the authors extended the actor-critic (AC) algorithm (Konda & Tsitsiklis, 2000) to the cooperative MARL setting. The algorithm allows each agent to perform its local policy improvement step while approximating the global value function. A few more recent works have extended Zhang et al. (2018) in different directions. For example in Grosnit et al. (2021), the authors considered the continuous action spaces and obtained the asymptotic convergence guarantee under both off-policy and on-policy settings. Moreover, Zhang et al. (2021) considered a new decentralized formulation where all agents cooperate to maximize general utilities in the cooperative MARL system, it developed AC-type algorithms to fit this setting but still suffering from high sampling cost in estimating the occupancy measure for all states and the nested loop of optimization steps. A concurrent work (Chen et al., 2021) adopts large-batch updates in decentralized (natural) AC methods to improve sample and communication efficiency, whose convergence rate matches the analysis results of the corresponding centralized versions (Xu et al., 2020a). However, the proposed algorithms in Chen et al. (2021) needs to generate $\mathcal{O}(\epsilon^{-1} \ln \epsilon^{-1})$ samples to update critic parameter before performing each actor update. It is worth noting, that all the above mentioned works do not allow the agents to share their local policies.

Our Contributions. Although there have been a growing literature on analyzing theoretical aspects of cooperative MARL, many challenges still remain, even under the basic fully observed setting. For example, most of the cooperative policy optimization algorithms, assume relatively simple collaboration mechanism, where the agents collaborate by jointly estimating the global value function, while *independently* optimizing their local policies. Such a form of collaboration decouples the agents' policy optimization process, and it is relatively easy to analyze. However, it fails to capture some intrinsic aspects of cooperative MARL, in the sense that when the agents' local tasks are similar (a.k.a. the *homogeneous* setting), the agent's policy should also be closely related to each other. Such an intuition has been verified in MARL systems (Gupta et al., 2017; Terry et al., 2020b), multi-task RL systems (Omidshafiei et al., 2017; Zeng et al., 2020; Yu et al., 2020), Markov games (Vadori et al., 2020) and mean-field multi-agent reinforcement learning (Liu et al., 2020; Li et al., 2021), where parameter sharing scheme results in more stable convergence due to the benefit of learning homogeneity among different agents. However, it is not clear how to design and analyze more sophisticated collaboration schemes which enable the agents to (partially) share their local policies to help them leverage each other's experience and build better behavior strategies.

In this work, we aim at providing better theoretical and practical understandings about the cooperative MARL problem. In particular, we consider the setting where the agents are connected by a time-varying network, and they can access the common observations while having different reward functions. We propose a Coordinated Actor-Critic (CAC) algorithm, provide the finite-sample analysis, and conduct extensive numerical experiments. Our specific contributions are given below:

• **A Generic Formulation.** We develop a new formulation of the cooperative MARL problem, which allows the agents to *coordinately* optimize their individual actions, by parameterizing their individual policies into a *shared* part (which is jointly optimized among all agents) and a *personalized* part (which is only locally optimized). The proposed formulation is general, in the sense that it can be used to cover a number of MARL settings. It can be used in a *fully decentralized* setting, where the agents can only communicate with their neighbors, as well as a *federated* setting, where the agents occasionally communicate with a server while optimizing their (partially personalized) local models.

• **Finite-Time Analysis.** We propose an algorithm for the generic problem setting, and show that it requires $\mathcal{O}(\epsilon^{-\frac{5}{2}})$ samples to achieve an ϵ-stationary solution. When being specialized to the decentralized setting where agents do not share the local policies, our result matches the performance bounds recently developed for centralized AC algorithms (Wu et al., 2020; Xu et al., 2020b). To the best of our knowledge, this is the first result that shows finite-sample guarantees for the decentralized AC algorithm with partially shared policy parameters.

• **Empirical Studies.** Finally, we conduct extensive numerical experiments, which demonstrate the effectiveness of the proposed algorithm. Our experiments suggest that in the situations where the agents' tasks are *homogeneous*, it is advantageous to partially personalize the policies for each agent; when the tasks' are heterogeneous, then the agents are able to achieve satisfactory convergence results by constructing their local policies without any parameter sharing.

Notation. $\|\cdot\|$ is used to denote the ℓ_2 norm for vectors and Frobenius norm for matrices. Further, we use $\mathbb{E}[\cdot]$ to denote expectation, $\mathcal{P}(\cdot)$ to denote probability. For a square matrix A, we define $c_2(A), c_{\max}(A)$ and $c_{\min}(A)$ as the second largest, the largest, and the smallest eigenvalues, respectively. Define $\mathbf{1}$ as an all one vector with appropriate size. For any matrix M, M^{ij} denotes the element in i-th row and j-th column of matrix M. For matrix $\boldsymbol{b} := [b_1^T; b_2^T; \cdots; b_N^T]$, we denote the average of all row vectors in matrix $\boldsymbol{b}$ as $\bar{b}^T := \frac{1}{N} \cdot \mathbf{1}^T \boldsymbol{b}$.

2 Preliminaries

In this section, we introduce the background and formulation of the cooperative, fully observable MARL in a decentralized system.

Suppose there are multiple agents aiming to independently learn and optimize a common global objective, and each agent can communicate with its neighbors in a network with time-varying topology. The common environment is observable by all the agents, and it is influenced by their joint actions. To model the communication pattern among the agents, let us define the time-varying graph $\mathcal{G}_t = (\mathcal{N}, \mathcal{E}_t)$ consisting of a set of $\mathcal{N}$ nodes and a set of $\mathcal{E}_t$ edges, with $|\mathcal{N}| = N$ and $|\mathcal{E}| = E$. Each node $i \in \mathcal{N}$ represents an agent and $\mathcal{E}_t$ represents the set of communication links at time t so that the agents are connected to their neighbors according to the links $\mathcal{E}_t$.

Consider the MARL problem, formulated as a discrete-time Markov Decision Process (MDP) $M := \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \eta, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is the finite space for global state s and $\mathcal{A}$ is the finite space for joint action $\boldsymbol{a} = \{a_i\}_{i=1}^N$; $\eta(s) : \mathcal{S} \to [0,1]$ denotes the initial state distribution; $\mathcal{P}(s' \mid s, \boldsymbol{a}) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ denotes the transition probability; $r_i(s, \boldsymbol{a}) : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ denotes the local reward function of agent i; $\gamma \in (0,1)$ is the discounted factor. Furthermore, suppose the policy of each agent i is parameterized by θ_i, then $\boldsymbol{\theta} := \{\theta_i\}_{i=1}^N$ denotes the collections of all policy parameters in the multi-agent system. Then $\mu_{\boldsymbol{\theta}}(s)$ denotes the stationary distribution of each state s under joint policy $\pi_{\boldsymbol{\theta}}$, and $d_{\boldsymbol{\theta}}(\cdot)$ denotes the discounted visitation measure where $d_{\boldsymbol{\theta}}(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{P}^{\pi_{\boldsymbol{\theta}}}(s_t = s \mid s_0 \sim \eta)$. Under the joint policy $\pi_{\boldsymbol{\theta}}$, the probability for choosing any joint action $\boldsymbol{a} := \{a_i\}_{i=1}^N$ could be expressed as $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|s) := \Pi_{i=1}^N \pi_i(a_i|s, \theta_i)$.

Consider the discrete-time MDP under infinite horizon, the policy $\pi_{\boldsymbol{\theta}}$ can generate a trajectory $\tau := (s_0, \boldsymbol{a}_0, s_1, \boldsymbol{a}_1, \cdots)$ based on the initial state s_0 sampled from $\eta(\cdot)$. In this work, we consider the discounted cumulative reward setting and the global value function is defined as below:

$$V_{\pi_{\boldsymbol{\theta}}}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot \bar{r}(s_t, \boldsymbol{a}_t) \mid s_0 = s\right], \tag{1}$$

where we define $\bar{r}(s_t, \boldsymbol{a}_t) := \frac{1}{N} \sum_{i=1}^N r_i(s_t, \boldsymbol{a}_t)$ and the expectation is taken over the trajectory τ generated from joint policy $\pi_{\boldsymbol{\theta}}$. When $\pi_{\boldsymbol{\theta}}$ is fixed, the value function $V_{\pi_{\boldsymbol{\theta}}}(s)$ will satisfy the Bellman equation (Bertsekas et al., 2000) for all states $s \in \mathcal{S}$:

$$V_{\pi_{\boldsymbol{\theta}}}(s) = \mathbb{E}_{\boldsymbol{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s), s' \sim \mathcal{P}(\cdot|s, \boldsymbol{a})} \left[\bar{r}(s, \boldsymbol{a}) + \gamma \cdot V_{\pi_{\boldsymbol{\theta}}}(s')\right]. \tag{2}$$

The objective of RL is to find the optimal policy parameter $\boldsymbol{\theta}^*$ which maximizes the expected discounted cumulative reward as below:

$$\max_{\boldsymbol{\theta}} \; J(\boldsymbol{\theta}) := \mathbb{E}_{s\sim\eta(\cdot)}[V_{\pi_{\boldsymbol{\theta}}}(s)] = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t \cdot \bar{r}(s_t,\boldsymbol{a}_t)\Big] = \mathbb{E}\Big[\sum_{t=0}^{\infty}\frac{\gamma^t}{N}\sum_{i=1}^{N} r_i(s_t,\boldsymbol{a}_t)\Big]. \tag{3}$$

In order to optimize $J(\boldsymbol{\theta})$, one can compute the policy gradient (Sutton et al., 2000), expressed below:

$$\nabla J(\boldsymbol{\theta}) := \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\boldsymbol{\theta}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s), s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})}\big[\big(\bar{r}(s,\boldsymbol{a}) + \gamma V_{\pi_{\boldsymbol{\theta}}}(s')\big)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|s)\big]. \tag{4}$$

3 The Proposed Coordinated Actor-Critic Algorithm

3.1 The Proposed Formulation

In this section, we describe our MARL formulation. Our proposed formulation is based upon (3), but with the key difference that we no longer require the agents to have independent policy parameters θ_i. Specifically, we assume that the agents can (partially) share their policy parameters with their neighbors. Hence, each agent will decompose its policy into $\theta_i := \{\theta_i^s, \theta_i^p\}$, where the shared part θ_i^s has the same dimension across all agents, and the personalized part θ_i^p will be kept locally. Although such a kind of partially personalized policy structure may be relatively more difficult to analyze, it has a number of potential advantages, as we list below:

• *A Generic Model*. We use the partial policy sharing as a generic setting, to cover the full spectrum of strategies ranging from no sharing case ($\theta_i^s = 0, \; \forall\, i$) to the full sharing case (($\theta_i^p = 0, \; \forall\, i$)). This generic model ensures that our subsequent algorithms and analysis can be directly used for all cases.

• *Better Models for Homogeneous Agents*. When the agents' local tasks have a high level of similarity (a.k.a. the *homogeneous* setting), partially sharing models' parameters could achieve better feature representation and guarantee that the agents' policies are closely related to each other. Additionally, the shared parameters could leverage more data (i.e., data drawn from all agents) compared with the personalized parameters, so the variance in the training process can be significantly reduced, potentially resulting in better training performance. Such an intuition has been verified empirically in reinforcement learning systems (Omidshafiei et al., 2017; Yu et al., 2020; Zeng et al., 2020), where sharing policies among different learners results in more stable convergence.

• *Approximate Common Knowledge.* A critical assumption often made in the analysis of multiagent systems is *common knowledge* (Aumann, 1976). Intuitively, this implies agents have a *shared* awareness of the underlying interaction. A key difficulty in MARL is that agents are simultaneously learning features of the underlying environment, thus common knowledge is not guaranteed. Thus notions of *approximate* common knowledge have been proposed for MARL (Schroeder de Witt et al., 2019). By relying on (partial) policy sharing mechanism, we hope to have some degree of approximate common knowledge and this is what facilitates coordination.

The above partially personalized policy structure leads to the following MARL formulation:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & J(\boldsymbol{\theta}) := \mathbb{E}_{s\sim\eta(\cdot)}\Big[V_{\pi_{\boldsymbol{\theta}}}(s)\Big] = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t \cdot \bar{r}(s_t,\boldsymbol{a}_t)\Big] \\ \text{s.t.} \quad & \theta_i^s = \theta_j^s \;\; \text{if } (i,j) \text{ are neighbors} \end{aligned} \tag{5}$$

where $\boldsymbol{\theta} := \{\theta_i^s, \theta_i^p\}_{i=1}^N$ is the collections of all local policy parameters $\theta_i := \{\theta_i^s, \theta_i^p\}$. To cast problem (5) into a more tractable form, we perform the following steps.

First, we approximate the global reward function for any $s \in \mathcal{S}$ and $\boldsymbol{a} \in \mathcal{A}$. Specifically, we use the following linear function $\widehat{r}(s,\boldsymbol{a};\lambda) := \varphi(s,\boldsymbol{a})^T\lambda$ to approximate the global reward $\bar{r}(s,\boldsymbol{a}) := \frac{1}{N}\sum_{i=1}^N r_i(s,\boldsymbol{a})$, where $\varphi(\cdot,\cdot) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^L$ is the feature mapping. Then the optimal parameter

$\lambda^*(\boldsymbol{\theta})$ can be found by solving the following problem:

$$\lambda^*(\boldsymbol{\theta}) \in \arg\min_{\lambda} \mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \Big[\big(\frac{1}{N}\sum_{i=1}^{N} r_i(s,\boldsymbol{a}) - \varphi(s,\boldsymbol{a})^T\lambda\big)^2\Big] \tag{6a}$$

$$= \arg\min_{\lambda} \sum_{i=1}^{N} \mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \Big[\big(r_i(s,\boldsymbol{a}) - \varphi(s,\boldsymbol{a})^T\lambda\big)^2\Big]. \tag{6b}$$

Second, we approximate the global value function $V_{\pi_{\boldsymbol{\theta}}}(s)$ for any $s\in\mathcal{S}$ under a fixed joint policy $\pi_{\boldsymbol{\theta}}$. Specifically, we use the following linear function $\widehat{V}(s;\omega) := \phi(s)^T\omega$ to approximate the global reward function $V_{\pi_{\boldsymbol{\theta}}}(s)$, where $\phi(\cdot):\mathcal{S}\to\mathbb{R}^K$ is a given feature mapping. Towards achieving the above approximation, we can solve the following mean squared Bellman error (MSBE) minimization problem (Tsitsiklis & Van Roy, 1997):

$$\omega^*(\boldsymbol{\theta}) \in \arg\min_{\omega} \mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})} \Big[\big(\frac{1}{N}\sum_{i=1}^{N} r_i(s,\boldsymbol{a}) + \gamma\widehat{V}(s';\omega) - \widehat{V}(s;\omega)\big)^2\Big] \tag{7a}$$

$$= \arg\min_{\omega} \sum\nolimits_{i=1}^{N} \mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})} \Big[\big(r_i(s,\boldsymbol{a}) + \gamma\widehat{V}(s';\omega) - \widehat{V}(s;\omega)\big)^2\Big]. \tag{7b}$$

To separate the objective into the sum of N terms (one for each agent), we introduce local copies of w and λ as $\{w_i\}_{i=1}^N$, $\{\lambda_i\}_{i=1}^N$, and define their vectorized versions $\boldsymbol{\omega} = [\omega_1,\cdots,\omega_N]^T$ and $\boldsymbol{\lambda} = [\lambda_1,\cdots,\lambda_N]^T$. Similarly, we also define $\boldsymbol{\omega}^*(\boldsymbol{\theta}) := [\omega_1^*(\boldsymbol{\theta}),\cdots,\omega_N^*(\boldsymbol{\theta})]^T$ and $\boldsymbol{\lambda}^*(\boldsymbol{\theta}) := [\lambda_1^*(\boldsymbol{\theta}),\cdots,\lambda_N^*(\boldsymbol{\theta})]^T$.

Summarizing the above discussion, problem (5) can be approximated using the following bi-level optimization problem:

$$\max_{\boldsymbol{\theta}} \quad \mathbb{E}_{\substack{s\sim\eta(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)\\ s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})}} \Big[\frac{1}{N}\sum_{i=1}^{N}\Big(\widehat{r}(s,\boldsymbol{a};\lambda_i^*(\boldsymbol{\theta})) + \gamma\cdot\widehat{V}(s';\omega_i^*(\boldsymbol{\theta}))\Big)\Big] \tag{8a}$$

$$s.t. \quad \boldsymbol{\omega}^*(\boldsymbol{\theta}) \in \arg\min_{\omega} \sum_{i=1}^{N} \mathbb{E}_{\substack{s\sim\mu_{\boldsymbol{\theta}}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)\\ s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})}} \Big[\big(r_i(s,\boldsymbol{a}) + \gamma\cdot\widehat{V}(s';\omega_i) - \widehat{V}(s;\omega_i)\big)^2\Big], \tag{8b}$$

$$\boldsymbol{\lambda}^*(\boldsymbol{\theta}) \in \arg\min_{\lambda} \sum_{i=1}^{N} \mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \Big[\big(r_i(s,\boldsymbol{a}) - \widehat{r}(s,\boldsymbol{a};\lambda_i)\big)^2\Big], \tag{8c}$$

$$\theta_i^s = \theta_j^s,\ \omega_i^*(\boldsymbol{\theta}) = \omega_j^*(\boldsymbol{\theta}),\ \lambda_i^*(\boldsymbol{\theta}) = \lambda_j^*(\boldsymbol{\theta}),\ \text{ if } (i,j) \text{ are neighbors.} \tag{8d}$$

In the subsequent discussion, we will refer to the problem of finding the optimal policy $\boldsymbol{\theta}$ as the *upper-level* problem, while referring to the problem of finding the optimal $\boldsymbol{\omega}^*(\boldsymbol{\theta})$ and $\boldsymbol{\lambda}^*(\boldsymbol{\theta})$ under a fixed policy parameters as the lower-level problem.

3.2 The Proposed Algorithm

In this subsection, we first present the assumptions related to network connectivity and communication protocols in the multi-agent systems. Then we describe the proposed Coordinated Actor-Critic (CAC) algorithm which is summarized in Algorithm 1.

Assumption 1 (Network Connectivity). *There exists an integer B such that the union of the consecutive B graphs is connected for all positive integers ℓ. That is, the following graph is connected:*

$$\Big(\mathcal{N}, \mathcal{E}(\ell\cdot B)\cup\mathcal{E}(\ell\cdot B+1)\cdots\cup\mathcal{E}((\ell+1)B-1)\Big),\ \forall\,\ell\geq 1$$

where $\mathcal{N}$ denotes the vertice set and $\mathcal{E}(t)$ denotes the set of active edges at time t.

Assumption 2 (Weight Matrices). *There exists a positive constant c such that $W_t = [W_t^{ij}] \in \mathcal{R}^{N\times N}$ is doubly stochastic and $W_t^{ii} \geq c$ for all $i\in\mathcal{N}$. Moreover, $W_t^{ij}\in[c,1)$ if $(i,j)\in\mathcal{E}(t)$, otherwise $W_t^{ij}=0$ for all $i,j\in\mathcal{N}$.*

Algorithm 1 *Coordinated Actor-Critic (CAC) Algorithm*

```
1: Input: Parameters $\{\alpha_t\}_{t=0}^{T-1}$, $\{\beta_t\}_{t=0}^{T-1}$, $\{\zeta_t\}_{t=0}^{T-1}$. Initialize $\theta_{i,0}, \omega_{i,0}, \lambda_{i,0}$ for all $i \in \mathcal{N}$
2: for $t = 0, 1, \ldots, T-1$ do
3:     Data Sampling: $s_t \sim \mu_{\boldsymbol{\theta}_t}(\cdot)$, $\boldsymbol{a}_t := \{a_{i,t} \sim \pi_i(\cdot|s_t, \theta_{i,t})\}_{i=1}^N$, $s_{t+1} \sim \mathcal{P}(\cdot|s_t, \boldsymbol{a}_t)$
4:     Consensus Step: $\widetilde{\boldsymbol{\omega}}_t = W_t \cdot \boldsymbol{\omega}_t$, $\widetilde{\boldsymbol{\lambda}}_t = W_t \cdot \boldsymbol{\lambda}_t$ and $\widetilde{\boldsymbol{\theta}}_t^s := W_t \cdot \boldsymbol{\theta}_t^s$
5:     for $i \in \mathcal{N}$ do
6:         Construct $\widetilde{\theta}_{i,t} = \{\widetilde{\theta}_{i,t}^s, \theta_{i,t}^p\}$ and update $\delta_{i,t} = r_{i,t} + \gamma \cdot \phi(s_{t+1})^T \omega_{i,t} - \phi(s_t)^T \omega_{i,t}$
7:         $\omega_{i,t+1} = \Pi_{R_\omega}\big(\widetilde{\omega}_{i,t} + \beta_t \cdot \delta_{i,t} \cdot \phi(s_t)\big)$
8:         $\lambda_{i,t+1} = \Pi_{R_\lambda}\Big(\widetilde{\lambda}_{i,t} + \zeta_t \cdot \big(r_{i,t} - \varphi(s_t, \boldsymbol{a}_t)^T \lambda_{i,t}\big) \cdot \varphi(s_t, \boldsymbol{a}_t)\Big)$
9:         $\theta_{i,t+1} = \widetilde{\theta}_{i,t} + \alpha_t \Big(\varphi(s_t, \boldsymbol{a}_t)^T \lambda_{i,t} + \gamma \phi(s_{t+1})^T \omega_{i,t} - \phi(s_t)^T \omega_{i,t}\Big) \nabla_{\theta_i} \log \pi_i(a_{i,t}|s_t, \theta_{i,t})$
10:    end for
11: end for
```

Assumption 1 ensures that the graph sequence is sufficiently connected for each agent to have repeated influence on other agents. Assumption 2 is standard in developing decentralized algorithms (Nedic et al., 2009), which could guarantee consensus results for shared parameter in each agent converging to a common vector.

After presenting the assumptions related to the network topology in the decentralized system, we are able to introduce the proposed CAC algorithm. The CAC algorithm takes two main steps, the policy optimization step (which optimizes $\boldsymbol{\theta}$), and policy evaluation step (which approximately solves the lower-level problem in (8)), as we describe below. For simplicity, we denote $\bar{r}(s_t, \boldsymbol{a}_t)$ as $\bar{r}_t$ and $r_i(s_t, \boldsymbol{a}_t)$ as $r_{i,t}$.

Policy Optimization. In this step, the agents optimize their local policy parameters, while trying to make sure that the shared parameters are not too far from their neighbors.

Towards this end, each agent i first produces a *locally averaged* shared parameter by linearly combining with its neighbors' current shared parameters. Such an operation can be expressed as

$$\widetilde{\boldsymbol{\theta}}_t^s := W_t \cdot \boldsymbol{\theta}_t^s \tag{9}$$

where $\boldsymbol{\theta}_t^s := [\theta_{1,t}^s, \theta_{2,t}^s, \cdots; \theta_{N,t}^s]^T \in \mathbb{R}^{N \times H}$ is a matrix which stores all parameters $\{\theta_{i,t}^s\}_{i=1}^N$, and $\widetilde{\boldsymbol{\theta}}_t^s$ is defined similarly. In the decentralized setting, the global reward $\bar{r}_t$ and the global value function $V_{\pi_{\boldsymbol{\theta}_t}}(\cdot)$ are not available for each agent i. Instead, the agents can locally estimate the global reward and the global value function using some linear approximation, evaluated on their local variables, as described in the previous subsection. As shown in line 11 of Algorithm 1, in a decentralized system, we consider the policy optimization step for each agent as below:

$$\theta_{i,t+1} := \widetilde{\theta}_{i,t} + \alpha_t \cdot \widehat{\delta}_{i,t} \cdot \nabla_{\theta_i} \log \pi_i(a_{i,t}|s_t, \theta_{i,t}), \;\; \forall i \in \mathcal{N} \tag{10}$$

$$\text{where} \quad \widehat{\delta}_{i,t} := \widehat{r}(s_t, \boldsymbol{a}_t; \lambda_{i,t}) + \gamma \cdot \widehat{V}(s_{t+1}; \omega_{i,t}) - \widehat{V}(s_t; \omega_{i,t}). \tag{11}$$

Policy Evaluation. Next, we update the local parameters $\lambda_{i,t}$ and $\omega_{i,t}$, which parameterize the global reward function and global value function. Towards this end, the parameters $\lambda_{i,t}$ and $\omega_{i,t}$ will be updated by first averaging over their neighbors, then performing one stochastic gradient descent step to minimize the local objectives, which are defined as in (8b) - (8c) and under consensus constraints (8d). That is, we have the following updates for $\boldsymbol{\lambda}_t$ and $\boldsymbol{\omega}_t$:

$$\widetilde{\boldsymbol{\lambda}}_t = W_t \cdot \boldsymbol{\lambda}_t, \quad \lambda_{i,t+1} = \Pi_{R_\lambda}\Big(\widetilde{\lambda}_{i,t} + \zeta_t \cdot \big(r_{i,t} - \widehat{r}(s_t, \boldsymbol{a}_t; \lambda_{i,t})\big) \cdot \nabla_{\lambda_i} \widehat{r}(s_t, \boldsymbol{a}_t; \lambda_{i,t})\Big), \tag{12}$$

$$\widetilde{\boldsymbol{\omega}}_t = W_t \cdot \boldsymbol{\omega}_t, \quad \omega_{i,t+1} = \Pi_{R_\omega}\Big(\widetilde{\omega}_{i,t} + \beta_t \cdot \delta_{i,t} \cdot \nabla_{\omega_i} \widehat{V}(s_t; \omega_{i,t})\Big), \quad \forall i \in \mathcal{N} \tag{13}$$

where we define $\delta_{i,t} := r_{i,t} + \gamma \cdot \widehat{V}(s_{t+1}; \omega_{i,t}) - \widehat{V}(s_t; \omega_{i,t})$. Moreover, $\Pi_{R_\omega}(\cdot)$ and $\Pi_{R_\lambda}(\cdot)$ are the projection operators, with R_ω and R_λ being the predetermined projection radii which are used to stabilize the update process (Tsitsiklis & Van Roy, 1997). Please see lines 8-10 in Algorithm 1.

4 Theoretical Results

In this section, we first present Assumptions 3 - 4 about reward function and linear approximations for policy evaluation. Then we show our theoretical results for the proposed CAC algorithm.

Assumption 3 (Bounded Reward). *All the local rewards $r_i(s,a)$ are uniformly bounded, i.e., there exist constants $R_{\max}$, for all $i \in \mathcal{N}$ and $s \in \mathcal{S}$ such that $|r_i(s,a)| \leq R_{\max}$.*

Assumption 4 (Function Approximation). *For each agent i, the value function and the global reward function are both parameterized by the class of linear functions, i.e., $\widehat{V}(s;\omega_i) := \phi(s)^T\omega_i$ and $\widehat{r}(s,\boldsymbol{a};\lambda_i) := \varphi(s,\boldsymbol{a})^T\lambda_i$ where we denote $\phi(s) := [\phi_1(s),\cdots,\phi_K(s)]^T \in \mathbb{R}^K$ and $\varphi(s,\boldsymbol{a}) = [\varphi_1(s,\boldsymbol{a}),\cdots,\varphi_L(s,\boldsymbol{a})]^T \in \mathbb{R}^L$ are the feature vector associated with s and $(s,\boldsymbol{a})$, respectively. The feature vectors $\phi(s)$ and $\varphi(s,\boldsymbol{a})$ are uniformly bounded for any $s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$, i.e., $\|\phi(s)\| \leq 1$ and $\|\varphi(s,\boldsymbol{a})\| \leq 1$. Furthermore, constructing the feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}|\times K}$ which has $[\phi_k(s), s \in \mathcal{S}]^T$ as its k-th column for any $k \in K$. Also constructing the feature matrix $\Psi \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|\times L}$ which has $[\varphi_l(s), s \in \mathcal{S}]^T$ as its ℓ-th column for any $\ell \in L$. Then, we further assume both Φ and Ψ have full column ranks.*

Assumption 3 - 4 are common in analyzing TD with linear function approximation; see e.g., Konda & Tsitsiklis (2000); Bhandari et al. (2018); Wu et al. (2020). With global observability, each agent could construct linear function approximations of the global value function and global reward function. Under these assumptions, it is guaranteed that there exist unique optimal solutions $\lambda^*(\boldsymbol{\theta})$ and $\omega^*(\boldsymbol{\theta})$ to approximate the global reward function in (6) and the global value function in (7) with linear functions. It is crucial to have the properties of unique optimal solutions in $\lambda^*(\boldsymbol{\theta})$ and $\omega^*(\boldsymbol{\theta})$ for constructing the convergence analysis of policy parameters $\boldsymbol{\theta}$.

Due to space limitation, we relegate remaining technical assumptions (i.e., Assumptions 5 - 6) to Appendix C and technical lemmas to Appendix D. We first present the convergence speed of the variables $\{\boldsymbol{\omega}_t\}$ and $\{\boldsymbol{\lambda}_t\}$ for the policy evaluation problem defined in (8b) - (8d). Please see Appendix G for the detailed proof.

Proposition 1. *Suppose Assumptions 1 - 6 hold. For any iteration t, by selecting stepsizes*

$$\alpha_t = \frac{\alpha_0}{T^{\sigma_1}},\ \beta_t = \frac{\beta_0}{T^{\sigma_2}},\ \zeta_t = \frac{\zeta_0}{T^{\sigma_2}}$$

where $0 < \sigma_2 < \sigma_1 < 1$ and $\alpha_0, \beta_0, \zeta_0 > 0$ are some fixed constants, the following holds:

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left(\mathbb{E}\left[\|\omega_{i,t}-\omega^*(\boldsymbol{\theta}_t)\|^2\right]+\mathbb{E}\left[\|\lambda_{i,t}-\lambda^*(\boldsymbol{\theta}_t)\|^2\right]\right)\\ =&\mathcal{O}(T^{-1+\sigma_2})+\mathcal{O}(T^{-\sigma_2})+\mathcal{O}\left(T^{\sigma_2-2\sigma_1}\right)+\mathcal{O}(T^{-2\sigma_1+2\sigma_2})+\mathcal{O}(T^{-2+2\sigma_2})+\mathcal{O}(T^{-2\sigma_2})\end{aligned}$$

where the expectation is taken over the data sampling procedure as shown in line 3 of Algorithm 1.

Compared with existing works (Wai et al., 2018; Doan et al., 2019) which established finite-time convergence guarantees for decentralized policy evaluation problems under the fixed policy, our results in Proposition 1 are analyzed in a more challenging situation where both policies and critics are updated in an alternating manner. Here, we must set $\sigma_1 > \sigma_2$ to ensure that the relation above is useful. This is reasonable since the optimal critic parameter $\omega^*(\boldsymbol{\theta}_t)$ is constantly drifting as the policy parameters $\boldsymbol{\theta}_t$ changes at each iteration, so the actor should update slowly compared with the critic.

Next, we study the convergence rate of policy parameters. We define $Q := I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ and define the average gradient of shared policy parameters as $\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta})} := \frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta_i^s} J(\boldsymbol{\theta})$. We will show that after averaging over the iterations, the expected stationarity condition violation for the policy optimization problem defined in (8a) is small. Please see Appendix H for the proof.

Proposition 2. *Under the same setting as Proposition 1, there exist two constant error term $\epsilon_{app} > 0$ and $\epsilon_{sp} > 0$. Algorithm 1 generates a sequence of policies $\{\boldsymbol{\theta}_t\}$, which satisfies the following:*

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|Q\cdot\boldsymbol{\theta}_t^s\|^2\right]+N\cdot\mathbb{E}\left[\|\overline{\nabla_{\theta^s} J(\boldsymbol{\theta}_t)}\|^2\right]+\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\right]\right)\\ =&\mathcal{O}(T^{-1+\sigma_1})+\mathcal{O}(T^{-\sigma_1})+\mathcal{O}(T^{-1+\sigma_2})+\mathcal{O}(T^{-\sigma_2})+\mathcal{O}\left(T^{\sigma_2-2\sigma_1}\right)+\mathcal{O}(T^{-2\sigma_1+2\sigma_2})\\ &+\mathcal{O}(T^{-2+2\sigma_2})+\mathcal{O}(T^{-2\sigma_2})+\mathcal{O}\left(\epsilon_{app}+\epsilon_{sp}\right).\end{aligned}$$

Algorithm 2 *Double Sampling Procedures*

```
Input: Parameters $\{\omega_{i,t}\}_{i=1}^N$, $\{\lambda_{i,t}\}_{i=1}^N$, $\{\theta_{i,t}\}_{i=1}^N$.
Double i.i.d. Sampling:
1) Sample $s_t \sim \mu_{\boldsymbol{\theta}_t}(\cdot)$, $\boldsymbol{a}_t := \{a_{i,t} \sim \pi_i(\cdot|s_t, \theta_{i,t})\}_{i=1}^N$, $s_{t+1} \sim \mathcal{P}(\cdot|s_t, \boldsymbol{a}_t)$
2) Sample $\tilde{s}_t \sim d_{\boldsymbol{\theta}_t}(\cdot)$, $\tilde{\boldsymbol{a}}_t := \{\tilde{a}_{i,t} \sim \pi_i(\cdot|\tilde{s}_t, \theta_{i,t})\}_{i=1}^N$, $\tilde{s}_{t+1} \sim \mathcal{P}(\cdot|\tilde{s}_t, \tilde{\boldsymbol{a}}_t)$
```

The approximation error ϵ_{app} and sampling error ϵ_{sp} are defined in Appendix E. A few remarks about the above results follow. First, one challenge in analyzing the convergence of Actor-Critic algorithms is that the actor and critic updates are typically sampled from different distributions (i.e., the distribution mismatch problem). To see this, note that to obtain an unbiased estimator for the policy gradient in (4), one needs to sample from the discounted visitation measure $d_{\boldsymbol{\theta}}(\cdot)$, while to obtain an unbiased estimator for the gradient of the MSBE in (7) (which is utilized to update the critic parameters), one needs to sample from the stationary distribution $\mu_{\boldsymbol{\theta}}(\cdot)$. However, standard implementations for AC methods in practice only use one sampling procedure for both actor and critic updates (Mnih et al., 2016; Shen et al., 2020). Therefore, the mismatch between the two sampling distributions inevitably introduces constant biases, and this is where the error term ϵ_{sp} comes from.

Second, at each local agent i, the value function $V_{\pi_{\boldsymbol{\theta}}}(s)$ is approximated by $\phi(s)^T\omega_i$ and the global reward function is approximated by $\varphi(s,\boldsymbol{a})^T\lambda_i$. Due to the linear approximation, the approximation error is inevitable in the convergence analysis. Here, we use a constant term ϵ_{app} to quantify the approximation error due to utilizing linear function for policy evaluation.

By combining previous Propositions, and by properly selecting the stepsize parameters σ_1 and σ_2, we show the main result as below. In Appendix E, we will present more discussion about a special case where there is no policy parameter sharing.

Theorem 1. *(Convergence of the CAC Algorithm) Suppose Assumptions 1 - 6 hold. Consider Algorithm 1 with partially shared policy parameters $\boldsymbol{\theta} := \cup_{i=1}^N\{\theta_i^s, \theta_i^p\}$. Let $\sigma_1 = \frac{3}{5}$ and $\sigma_2 = \frac{2}{5}$, it holds that:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left(\mathbb{E}\left[\|\omega_{i,t} - \omega^*(\boldsymbol{\theta}_t)\|^2\right] + \mathbb{E}\left[\|\lambda_{i,t} - \lambda^*(\boldsymbol{\theta}_t)\|^2\right]\right) = \mathcal{O}(T^{-\frac{2}{5}}),$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\right] + N \cdot \mathbb{E}\left[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\right] + \sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\right]\right) = \mathcal{O}(T^{-\frac{2}{5}}) + \mathcal{O}(\epsilon_{app} + \epsilon_{sp}).$$

As mentioned before, the sampling error ϵ_{sp} arises because there is a mismatch between the way that estimators of the actor's and the critics' updates are obtained. To remove the sampling error, one can implement separate sampling protocols for the critic and the actor. More specifically, we can use two different i.i.d. samples at each iteration step t: 1) $x_t := (s_t, \boldsymbol{a}_t, s_{t+1})$ where $s_t \sim \mu_{\boldsymbol{\theta}}(\cdot)$, $\boldsymbol{a}_t \sim \pi_{\boldsymbol{\theta}}(\cdot \mid s_t)$ and $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, \boldsymbol{a}_t)$; 2) $\tilde{x}_t := (\tilde{s}_t, \tilde{\boldsymbol{a}}_t, \tilde{s}_{t+1})$ where $\tilde{s}_t \sim d_{\boldsymbol{\theta}}(\cdot)$, $\tilde{\boldsymbol{a}}_t \sim \pi_{\boldsymbol{\theta}}(\cdot \mid s_t)$ and $\tilde{s}_{t+1} \sim \mathcal{P}(\cdot \mid \tilde{s}_t, \tilde{\boldsymbol{a}}_t)$; see Algorithm 2. Then x_t and $\tilde{x}_t$ will be utilized in policy evaluation and policy optimization, respectively. The corollary below shows the convergence result for the modified CAC algorithm. Please see Appendix I for the proof.

Corollary 1. *(Convergence under double sampling) Under the same setting as Theorem 1, consider CAC with the double sampling procedures in Algorithm 2. The following result holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left(\mathbb{E}\left[\|\omega_{i,t} - \omega^*(\boldsymbol{\theta}_t)\|^2\right] + \mathbb{E}\left[\|\lambda_{i,t} - \lambda^*(\boldsymbol{\theta}_t)\|^2\right]\right) = \mathcal{O}(T^{-\frac{2}{5}}),$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\right] + N \cdot \mathbb{E}\left[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\right] + \sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\right]\right) = \mathcal{O}(T^{-\frac{2}{5}}) + \mathcal{O}(\epsilon_{app}).$$

5 Numerical Results

In this section, we present our simulation results on two environments: 1) the coordination game (Osborne & Rubinstein, 1994); 2) the pursuit-evasion game (Gupta et al., 2017), which is built on the PettingZoo platform (Terry et al., 2020a). Detailed experiment settings are present in Appendix A.

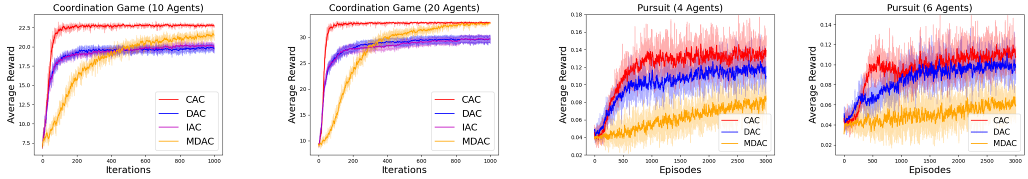

Figure 1: **Simulation Results.** The averaged reward versus the learning process. We present the algorithm performance on the coordination game and on the pursuit-evasion game. The performance is averaged over 10 Monte Carlo runs.

Coordination Game: In this setting, there are N agents staying at a static state and they choose their actions simultaneously at each time. After actions are executed at each time t, each agent i receives its reward as: $r_{i,t} = (a_{i,t} - 3.5)^2 + \sum_{j \neq i} I_{\{a_{j,t}=a_{i,t}\}} + \epsilon_{i,t}$ where the action space is $\{0, 1, 2, \cdots, 7\}$, $I_{\{a_{j,t}=a_{i,t}\}}$ is an indicator function and $\epsilon_{i,t}$ is a random payoff following standard Gumbel distribution. In this coordination game, there are multiple Nash equilibria where two optimal equilibria are that all agents select $a = \{0\}$ or $a = \{7\}$ simultaneously. In order to obtain high rewards and achieve efficient equilibria, it is crucial for agents to coordinate with others while only having limited communications.

Here, the communication graph $\mathcal{G}_t$ between the agents is a complete graph every 5 iterations, and is not connected for the rest of time. We compare the performance of CAC with three benchmark algorithms: independent Actor-Critic (IAC); decentralized Actor-Critic (DAC) in Zhang et al. (2018); mini-batch decentralized Actor-Critic (MDAC) in Chen et al. (2021). For each algorithm, we set the actor stepsize and critic stepsize as 0.05 and 0.1. Theoretically, MDAC needs $\mathcal{O}(\epsilon^{-1} \ln \epsilon^{-1})$ batch size in its inner loop to update critic parameters before each update in policy parameters, which is inefficient in practice. Here, we set small batch $B = 5$ in the inner loop for MDAC to achieve fast convergence. The simulation results on this coordination game are present in Fig.1 (two left figures). According to the simulations, compared with the benchmarks, we see that the CAC algorithm converges faster and has higher probability to achieve efficient equilibria due to the use of policy sharing and coordination.

Pursuit-Evasion Game: there are two groups of nodes, pursuers (agents) and evaders. The pursuers aim to obtain reward through catching evaders. In a two-dimensional environment, an evader is considered caught if two pursuers simultaneously arrive at the evader's location. In order to catch an evader, each pursuer should learn to cooperate with other pursuers to catch the evaders. From this perspective, the pursuers share some similarities with each other since they need to follow similar strategies to achieve their local tasks: simultaneously catching a same evader with other pursuers.

In Figure 1 (two right figures), we compare the numerical performance of the proposed CAC algorithm and two benchmarks: decentralized Actor-Critic (DAC) in Zhang et al. (2018); mini-batch decentralized Actor-Critic (MDAC) in Chen et al. (2021). Each agent maintains two convolutional neural networks (CNNs), one for the actor and one for the critic. Please see Figure 2 in Appendix for the structure diagrams of actor network and critic network being used. In the CAC, two convolutional layers of actor network will be regarded as shared policy parameters, and the output layer is personalized (thus not shared).

The two sets of numerical results suggest that, when local tasks share a certain degree of similarity / homogeneity, CAC algorithm with (partial) parameter sharing could achieve more stable convergence.

6 Conclusion

This paper develops a novel collaboration mechanism for designing robust MARL systems. Further, it develops and analyzes a novel multi-agent AC method, where agents are allowed to (partially) share their policy parameters with the neighbors to learn from different agents. To our knowledge, this is the first non-asymptotic convergence result for two-timescale multi-agent AC methods. We leave the extensions of our proposed algorithm to partially observable Markov decision process as the future work.

REFERENCES

Alekh Agarwal, Nan Jiang, and Sham M Kakade. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Robert Aumann. *Annals of Statistics*, 4(6):1236–1239, 1976.

Dimitri P Bertsekas et al. *Dynamic programming and optimal control: Vol. 1*. Athena scientific Belmont, 2000.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pp. 1691–1692. PMLR, 2018.

Tianyi Chen, Kaiqing Zhang, Georgios B Giannakis, and Tamer Başar. Communication-efficient policy gradient methods for distributed reinforcement learning. *arXiv preprint arXiv:1812.03239*, 2018.

Ziyi Chen, Yi Zhou, Rongrong Chen, and Shaofeng Zou. Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis. *arXiv preprint arXiv:2109.03699*, 2021.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998(746-752):2, 1998.

Thinh Doan, Siva Maguluri, and Justin Romberg. Finite-time analysis of distributed td (0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 1626–1635. PMLR, 2019.

Kenji Doya. Reinforcement learning in continuous time and space. *Neural computation*, 12(1): 219–245, 2000.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pp. 1407–1416. PMLR, 2018.

Antoine Grosnit, Desmond Cai, and Laura Wynter. Decentralized deterministic multi-agent reinforcement learning. *arXiv preprint arXiv:2102.09745*, 2021.

Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pp. 66–83. Springer, 2017.

Soummya Kar, José MF Moura, and H Vincent Poor. Qd-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus. *arXiv preprint arXiv:1205.0047*, 2012.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014. Citeseer, 2000.

Vijaymohan R Konda and Vivek S Borkar. Actor-critic–type learning algorithms for markov decision processes. *SIAM Journal on control and Optimization*, 38(1):94–123, 1999.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Donghwan Lee, Niao He, Parameswaran Kamalaruban, and Volkan Cevher. Optimization for reinforcement learning: From a single agent to cooperative agents. *IEEE Signal Processing Magazine*, 37(3):123–135, 2020.

Yan Li, Lingxiao Wang, Jiachen Yang, Ethan Wang, Zhaoran Wang, Tuo Zhao, and Hongyuan Zha. Permutation invariant policy optimization for mean-field multi-agent reinforcement learning: A principled approach. *arXiv preprint arXiv:2105.08268*, 2021.

Lewis Liu, Zhuoran Yang, Yuchen Lu, and Zhaoran Wang. Decentralized policy gradient method for mean-field linear quadratic regulator with global convergence. 2020.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017.

Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Basar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8767–8775, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Angelia Nedic, Alex Olshevsky, Asuman Ozdaglar, and John N Tsitsiklis. On distributed averaging algorithms and quantization effects. *IEEE Transactions on automatic control*, 54(11):2506–2517, 2009.

Angelia Nedic, Asuman Ozdaglar, and Pablo A Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938, 2010.

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pp. 2681–2690. PMLR, 2017.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 4295–4304. PMLR, 2018.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Christian Schroeder de Witt, Jakob Foerster, Gregory Farquhar, Philip Torr, Wendelin Bohmer, and Shimon Whiteson. Multi-agent common knowledge reinforcement learning. In *Advances in neural information processing systems*, pp. 1008–1014. Citeseer, 2019.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Han Shen, Kaiqing Zhang, Mingyi Hong, and Tianyi Chen. Asynchronous advantage actor critic: Non-asymptotic analysis and linear speedup. *arXiv preprint arXiv:2012.15511*, 2020.

Peter Stone and Manuela Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3):345–383, 2000.

Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. *arXiv preprint arXiv:1809.04216*, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4):e0172395, 2017.

Justin K Terry, Benjamin Black, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis Santos, Clemens Dieffendahl, Niall L Williams, Yashas Lokesh, Caroline Horsch, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *arXiv preprint arXiv:2009.14471*, 2020a.

Justin K Terry, Nathaniel Grammel, Ananth Hari, Luis Santos, and Benjamin Black. Revisiting parameter sharing in multi-agent deep reinforcement learning. *arXiv preprint arXiv:2005.13625*, 2020b.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.

Nelson Vadori, Sumitra Ganesh, Prashant Reddy, and Manuela Veloso. Calibration of shared equilibria in general sum partially observable markov games. *arXiv preprint arXiv:2006.13085*, 2020.

Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. *arXiv preprint arXiv:1806.00877*, 2018.

Gang Wang, Songtao Lu, Georgios Giannakis, Gerald Tesauro, and Jian Sun. Decentralized td tracking with linear function approximation and its finite-time analysis. *Advances in Neural Information Processing Systems*, 33, 2020.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

David H Wolpert, Kevin R Wheeler, and Kagan Tumer. General principles of learning-based multi-agent systems. In *Proceedings of the third annual conference on Autonomous Agents*, pp. 77–83, 1999.

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

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

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

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pp. 1094–1100. PMLR, 2020.

Sihan Zeng, Aqeel Anwar, Thinh Doan, Justin Romberg, and Arijit Raychowdhury. A decentralized policy gradient approach to multi-task reinforcement learning. *arXiv preprint arXiv:2006.04338*, 2020.

Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. Marl with general utilities via decentralized shadow reward actor-critic. *arXiv preprint arXiv:2106.00543*, 2021.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pp. 5872–5881. PMLR, 2018.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020.

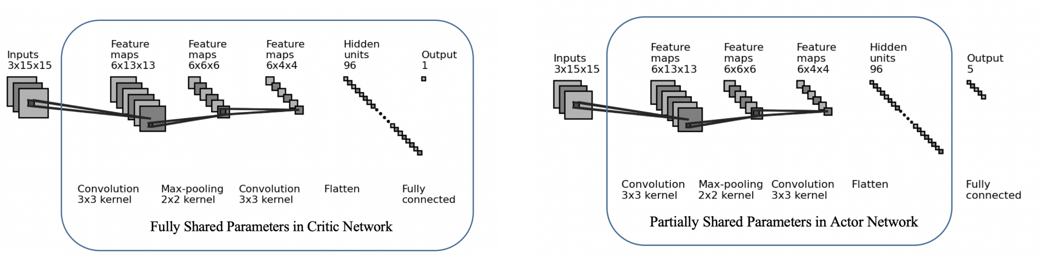

Figure 2: **Neural Network Architecture Diagrams for the CAC Algorithm.** The architecture diagrams for actor network and critic network of algorithm CAC in the pursuit-evasion Game. (Left) The diagram of critic network. (Right) The diagram of actor network.

Appendix

A Experiment Details

In this section, we will present the experiment details on the pursuit-evasion game.

A.1 pursuit-evasion Game

The 'capture' reward for each agent is set to be 5 when a pursuer successfully catches an evader. Moreover, the pursuer will receive a small reward signal which is set to be 0.1 when the pursuer encounters an evader at its current location. The environment is set to be a 15×15 grid and this 2D grid contains obstacles where the agents cannot pass through. Hence, the global state of the pursuit-evasion game consists of three images (binary matrices) of the size of 15×15. Hence, the dimension of the global state is $3 \times 15 \times 15$. These three images (binary matrices) respectively present the location of the pursuers, evaders and obstacles in the two-dimensional grid. Only given the 3-channel images as the global state, it is difficult for each pursuer (agent) to distinguish itself with other pursuers since the 3-channel images (global state) does not directly show the ID for each pursuer in the pursuit-evasion Game. To tackle this challenging, we center each agent's observation at its own location. With a large observation radius, each agent could observe the global information in the environment.

Considering the observation of each agent is a 3-channel image, each agent respectively maintains two convolutional neural networks (CNNs) with two convolutional layers, one max-pooling layer and one fully connected layer for the actor and the critic. Please see Figure 2 for the structure diagrams of actor network and critic network in algorithm CAC. The communication graph $\mathcal{G}_t$ between the agents is a complete graph every 20 iterations, and is not connected for the rest of time. Hence, for CAC algorithm, the global averaging step will be performed on the entire critic networks and the two CNN layers of actor networks every 20 iterations. The RuLU activation function is utilized in each hidden layer of actor network and critic network. The output of critic network approximates the value function $V_{\pi_{\boldsymbol{\theta}}}(s)$ for all $s \in \mathcal{S}$ and the dimension of the output layer is 1. Furthermore, the output dimension of actor network is 5 which corresponds to the number of possible actions. In each CNN, the raw images (3-channel location matrices), whose dimension is $3 \times 15 \times 15$, are processed by two convolutional layers and one max-pooling layer first and then pass through a fully connected layer as the output layer. We utilize the RMSprop optimizer (Ruder, 2016) to train neural networks, which is a common choice in training neural networks for reinforcement learning problems (Mnih et al., 2013). For each algorithm, we set the actor stepsize and critic stepsize as 1×10^{-4} and 1×10^{-3}. For algorithm MDAC to achieve quick convergence, we tune its batch size and set small batch $B = 5$ in its inner loop. The discount factor γ is set to be 0.95 in this simulation.

B Discussion: Application in Multi-Agent Settings

Here, we discuss how the proposed CAC algorithm can be used in two popular multi-agent settings:

- **Fully personalized multi-agent RL.** The CAC algorithm can be applied to the special setting where the agents *do not* share their policy parameters with neighbors, and no cooperation is considered in generating the local policies. This setting has been proposed and studied in a number of existing works, such as Zhang et al. (2018); Chen et al. (2018).

- **Federated RL.** The proposed algorithm can be applied to a general federated (reinforcement) learning setting, where the agents jointly optimize a common objective. To see the connection, let us first describe a standard federated learning (FL) setting (Konečný et al., 2016; Arivazhagan et al., 2019): a central controller coordinates a few agents, where the agents continuously optimize their local parameters and perform occasional averaging steps over the parameters. It can be shown that this protocol corresponds to a dynamic setting where $\mathcal{G}_t$ is a *complete* graph every fixed number of iterations, while it is *not connected at all* for the rest of times (which is a special case of our setting, see Assumption 1). When the network is connected, each agent could gather other agents' models and perform the averaging; when the network is not connected, then each agent just performs local updates. We generalize the above FL setting to MARL in CAC algorithm.

C Technical Assumptions

Assumption 5. *Define the score function $\psi_{\boldsymbol{\theta}}(s,\boldsymbol{a}) := \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid s)$. For any policy parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, and any state-action $(s,\boldsymbol{a})$, the following holds:*

$$\|\psi_{\boldsymbol{\theta}}(s,\boldsymbol{a}) - \psi_{\boldsymbol{\theta}'}(s,\boldsymbol{a})\| \leq L_\psi \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \quad \|\psi_{\boldsymbol{\theta}}(s,\boldsymbol{a})\| \leq C_\psi \tag{14a}$$

$$|\pi_{\boldsymbol{\theta}}(\boldsymbol{a} \mid s) - \pi_{\boldsymbol{\theta}'}(\boldsymbol{a} \mid s)| \leq L_\pi \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \tag{14b}$$

where L_π, L_ψ, C_ψ are some constants.

Assumption 5 has been often used in analyzing policy gradient-type algorithms, for example see Zhang et al. (2020); Agarwal et al. (2020). Many policy parameterization methods such as tabular softmax policy Agarwal et al. (2020), Gaussian policy Doya (2000) and Boltzmann policy Konda & Borkar (1999) satisfy this assumption.

Assumption 6. *For any policy parameters $\boldsymbol{\theta}$, the markov chain under policy $\pi_{\boldsymbol{\theta}}$ and transition kernel $\mathcal{P}(\cdot|s,\boldsymbol{a})$ is irreducible and aperiodic. Then there exist constants $\kappa > 0$ and $\tau \in (0,1)$ such that*

$$\sup_{s\in\mathcal{S}} d_{TV}\left(\mathcal{P}^{\pi_{\boldsymbol{\theta}}}(s_t \in \cdot | s_0 = s), \mu_{\boldsymbol{\theta}}(\cdot)\right) \leq \kappa \cdot \tau^t, \quad \forall t \tag{15}$$

where $d_{TV}(\cdot)$ is the total variation (TV) norm; μ_θ is the stationary state distribution under $\pi_{\boldsymbol{\theta}}$.

Assumption 6 assumes the Markov chain mixes at a geometric rate; see also Bhandari et al. (2018); Sun et al. (2018).

D Auxiliary Lemmas

Lemma 1. *((Nedic et al., 2010, Lemma 1)) Let $\mathcal{X}$ be a nonempty closed convex set in $\mathbb{R}^K$, then the following holds:*

$$(\Pi_{\mathcal{X}}[x] - x)^T (x - y) \leq -\|\Pi_{\mathcal{X}}[x] - x\|^2, \ \forall\, x \in \mathbb{R}^K, \ y \in \mathcal{X} \tag{16a}$$

$$\|\Pi_{\mathcal{X}}[x] - y\|^2 \leq \|x - y\|^2 - \|\Pi_{\mathcal{X}}[x] - x\|^2, \ \forall\, x \in \mathbb{R}^K, \ y \in \mathcal{X} \tag{16b}$$

where $\Pi_{\mathcal{X}}[\cdot]$ denotes the projection operator on to the convex set $\mathcal{X}$.

Lemma 2. *((Zhang et al., 2020, Lemma 3.2)). Suppose Assumption 5 holds. Then the following holds:*

$$\|\nabla J(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta}')\| \leq L_J \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \quad \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^{N\times D} \tag{17}$$

where $J(\cdot)$ is the objective function defined in (3)*; $L_J := \frac{R_{\max}\cdot L_\psi}{(1-\gamma)^2} + \frac{(1+\gamma)\cdot R_{\max}\cdot C_\psi^2}{(1-\gamma)^3}$ with L_ψ and C_ψ defined in Assumption 5.*

Lemma 3. *((Shen et al., 2020, Lemma 4)). Suppose Assumption 5 holds. The following holds:*

$$\|\nabla J(\boldsymbol{\theta})\| \leq L_v, \quad |J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2)| \leq L_v \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \ \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^{N\times D}, \ s \in \mathcal{S}, \tag{18}$$

where the constant $L_v := \frac{R_{\max}}{1-\gamma} \cdot C_\psi$, with C_ψ defined in Assumption 5.

Lemma 4. *((Shen et al., 2020, Lemma 1)) Suppose Assumption 6 holds. The following holds:*

$$d_{TV}(\mu_{\boldsymbol{\theta}}, d_{\boldsymbol{\theta}}) \leq 2 \left(\log_\tau \kappa^{-1} + \frac{1}{1-\tau} \right) (1-\gamma), \quad \forall\, \boldsymbol{\theta} \in \mathbb{R}^{N \times D} \tag{19}$$

where $\mu_{\boldsymbol{\theta}}(\cdot)$ is the stationary distribution of each state s under policy parameters $\boldsymbol{\theta}$ and $d_{\boldsymbol{\theta}}(\cdot)$ is the discounted visitation meausre $d_{\boldsymbol{\theta}}(s) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{P}^{\pi_{\boldsymbol{\theta}}}(s_t = s \mid s_0 \sim \eta)$.

With the technical assumptions in C, we could bound consensus errors over the iterations. Towards this end, let us provide some basic properties for the weight matrices. Based on Assumptions 1 - 2 which ensure the long-term connectivity and impose the underlying topology of the networked system, we can obtain the following condition (Nedic et al., 2009, Lemma 9):

$$\|W_t \cdots W_{t+B-1} \cdot Q \cdot \boldsymbol{\omega}\| \leq \eta \cdot \|Q \cdot \boldsymbol{\omega}\|, \tag{20}$$

when $\boldsymbol{\omega} := [\omega_1^T; \omega_2^T; \cdots; \omega_N^T] \in \mathbb{R}^{N \times K}$; and we define $Q := I - \frac{1}{N}\mathbf{1} \cdot \mathbf{1}^T \in \mathbb{R}^{N \times N}$. Further, the constant η in (20) is given by $\eta := \sqrt{1 - \frac{c}{2N^2}} \in (0,1)$, where constant c is defined in Assumption 2.

Based on the above property, we have the following bounds on various consensus errors. Please see Appendix F for the proof.

Lemma 5. *Based on Assumptions 1 - 4, there exist constants $L_B > 0$ and $\ell_p > 0$ such that the consensus errors $\|Q \cdot \boldsymbol{\omega}_t\| + \|Q \cdot \boldsymbol{\lambda}_t\|$ and $\|Q \cdot \boldsymbol{\theta}_t^s\|$ satisfy*

$$\|Q \cdot \boldsymbol{\omega}_t\| + \|Q \cdot \boldsymbol{\lambda}_t\| \leq \rho^t \cdot \frac{\|\boldsymbol{\omega}_0\| + \|\boldsymbol{\lambda}_0\|}{\eta} + \frac{2N \cdot L_b \cdot (\beta + \zeta)}{\eta \cdot (1-\rho)} \cdot \frac{1}{T^{\sigma_2}} \tag{21a}$$

$$\|Q \cdot \boldsymbol{\theta}_t^s\| \leq \rho^t \cdot \frac{\|\boldsymbol{\theta}_0^s\|}{\eta} + \frac{\ell_p \cdot \alpha}{\eta \cdot (1-\rho)} \cdot \frac{1}{T^{\sigma_1}} \tag{21b}$$

$$\frac{1}{T}\sum_{t=0}^{T-1} \Big(\|Q \cdot \boldsymbol{\omega}_t\| + \|Q \cdot \boldsymbol{\lambda}_t\| \Big) = \mathcal{O}\Big(T^{-\sigma_2} \Big), \tag{21c}$$

$$\frac{1}{T}\sum_{t=0}^{T-1} \Big(\|Q \cdot \boldsymbol{\lambda}_t\|^2 + \|Q \cdot \boldsymbol{\omega}_t\|^2 \Big) = \mathcal{O}(T^{-1}) + \mathcal{O}(T^{-2\sigma_2}) \tag{21d}$$

$$\frac{1}{T}\sum_{t=0}^{T-1} \|Q \cdot \boldsymbol{\theta}_t^s\| = \mathcal{O}\Big(T^{-\sigma_1} \Big), \ \frac{1}{T}\sum_{t=0}^{T-1} \|Q \cdot \boldsymbol{\theta}_t^s\|^2 = \mathcal{O}(T^{-1}) + \mathcal{O}(T^{-2\sigma_1}) \tag{21e}$$

where $\alpha_t := \frac{\alpha}{T^{\sigma_1}}$, $\beta_t := \frac{\beta}{T^{\sigma_2}}$ and $\zeta_t := \frac{\zeta}{T^{\sigma_2}}$ are three stepsizes; $\rho := \eta^{\frac{1}{B}}$.

Given the fixed policy parameter $\boldsymbol{\theta}$, solving the lower level problem of (8) is equivalent to solving the *centralized* policy evaluation problems, expressed in (6) and (7). Through the first-order optimality condition, it is easy to show that $w^*(\boldsymbol{\theta})$ satisfies the following condition:

$$A(\boldsymbol{\theta}) \cdot \omega^*(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} b_i(\boldsymbol{\theta}) \tag{22}$$

where we have defined:

$$A(\boldsymbol{\theta}) := \mathbb{E}_{s \sim \mu_{\boldsymbol{\theta}}(\cdot), s' \sim \mathcal{P}^{\pi_{\boldsymbol{\theta}}}(\cdot|s)} \Big[\phi(s)\big(\phi(s) - \gamma \cdot \phi(s')\big)^T \Big], \ \forall\, i \in \mathcal{N}, \tag{23a}$$

$$b_i(\boldsymbol{\theta}) := \mathbb{E}_{s \sim \mu_{\boldsymbol{\theta}}(\cdot), \boldsymbol{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[r_i(s,a) \cdot \phi(s) \right], \ \forall\, i \in \mathcal{N}. \tag{23b}$$

Under the full-rankness and bounded assumption of the feature matrices given in Assumption 4, we can apply (Tsitsiklis & Van Roy, 1997, Theorem 2), and show that $A(\boldsymbol{\theta})$ is a positive definite matrix for any fixed $\boldsymbol{\theta}$. Let us define

$$0 < \tilde{c}_{\min} \leq c_{\min}\big(A(\boldsymbol{\theta})\big), \quad 0 < c_{\max}\big(A(\boldsymbol{\theta})\big) \leq \tilde{c}_{\max}, \ \forall\, \boldsymbol{\theta} \in \mathbb{R}^{N \times D}, \tag{24}$$

where $c_{\min}\big(A(\boldsymbol{\theta})\big)$ and $c_{\max}\big(A(\boldsymbol{\theta})\big)$ are the minimum and maximum eigenvalue of $A(\boldsymbol{\theta})$; $\tilde{c}_{\min}$ and $\tilde{c}_{\max}$ are the lower bound and upper bound on the eigenvalues of $A(\boldsymbol{\theta})$. Then we have the following Lipschitz property of the optimal critic parameters.

Lemma 6. *((Shen et al., 2020, Proposition 2)) Suppose Assumptions 4,5,6 hold. Let $w^*(\boldsymbol{\theta})$ denote the optimal solution in* (7) *to approximate value function approximation. Then the following Lipschitz condition holds:*

$$\|\omega^*(\boldsymbol{\theta}_1) - \omega^*(\boldsymbol{\theta}_2)\| \leq L_\omega \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \quad \forall\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^{N\times D}, \tag{25}$$

where $L_\omega := 2 \cdot R_{\max} \cdot |\mathcal{A}| \cdot L_\pi \cdot \bigg(\tilde{c}_{\min}^{-1} + \tilde{c}_{\min}^{-2} \cdot (1+\gamma)\bigg)\bigg(1 + \log_\tau \kappa^{-1} + (1-\tau)^{-1}\bigg).$

E Discussion: Convergence Results

In this section, we discuss an extension of Theorem 1.

As a special case, when the agents do not share any policy parameters, that is, when $\boldsymbol{\theta} = \boldsymbol{\theta}^p$, the resulting algorithm reduces to the standard Decentralized AC algorithm and we name it as Coordinated Actor-Critic with no policy sharing (CAC-NPS), whose asymptotic convergence property has been analyzed in Zhang et al. (2018). The non-asymptotic convergence rate for this algorithm can be readily obtained from (a slightly modified versions of) Proposition 1 – 2. The following result states the convergence rate for CAC-NPS, and we refer the readers to Appendix H for detailed proof steps.

Corollary 2. *(Convergence of CAC-NPS Algorithm) Suppose Assumptions 1 - 6 hold. Consider applying Algorithm 1 to a problem with fully-personalized policy parameters, that is,* $\boldsymbol{\theta} := \cup_{i=1}^N \{\theta_i^p\}$. *Setting* $\sigma_1 = \frac{3}{5}$ *and* $\sigma_2 = \frac{2}{5}$*, then the following holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\bigg(\mathbb{E}\Big[\|\omega_{i,t} - \omega^*(\boldsymbol{\theta}_t)\|^2\Big] + \mathbb{E}\Big[\|\lambda_{i,t} - \lambda^*(\boldsymbol{\theta}_t)\|^2\Big]\bigg) = \mathcal{O}(T^{-\frac{2}{5}}) \tag{26}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\Big[\|\nabla_{\theta_i} J(\boldsymbol{\theta}_t)\|^2\Big] = \mathcal{O}(T^{-\frac{2}{5}}) + \mathcal{O}(\epsilon_{app} + \epsilon_{sp}). \tag{27}$$

The critic approximation error ϵ_{app} and the sampling error ϵ_{sp} are defined as follows (Wu et al., 2020; Shen et al., 2020):

$$\epsilon_{\text{app}} := \max_{\boldsymbol{\theta}} \sqrt{\mathop{\mathbb{E}}_{s\sim\mu_{\boldsymbol{\theta}}(\cdot)}\bigg[\Big(V_{\pi_{\boldsymbol{\theta}}}(s) - \phi(s)^T\omega^*(\boldsymbol{\theta})\Big)^2\bigg]} + \sqrt{\mathop{\mathbb{E}}_{\substack{s\sim\mu_{\boldsymbol{\theta}}(\cdot)\\ a\sim\pi_{\boldsymbol{\theta}}(\cdot)}}\bigg[\Big(\bar{r}(s,\boldsymbol{a}) - \varphi(s,\boldsymbol{a})^T\lambda^*(\boldsymbol{\theta})\Big)^2\bigg]} \tag{28}$$

$$\epsilon_{\text{sp}} := 4 \cdot R_{\max} \cdot C_\psi \cdot L_v \cdot \bigg(\log_\tau \kappa^{-1} + \frac{1}{1-\tau}\bigg) \tag{29}$$

where $\mu_{\boldsymbol{\theta}}(\cdot)$ is the stationary distribution of state under policy $\pi_{\boldsymbol{\theta}}$ and the transition kernel $\mathcal{P}(\cdot)$.

F Proof of Lemma 5

The proof of Lemma 5 is divided into two steps. In step 1, we first analyze the consensus error $\|Q\cdot\boldsymbol{\omega}_t\|$ and then extend the analysis results to $\|Q\cdot\boldsymbol{\lambda}_t\|$. In step 2, we further analyze the consensus error $\|Q\cdot\boldsymbol{\theta}_t^s\|$ in the shared part of policy parameters.

Step 1. Since the mixing matrix W_t is doubly stochastic so $W_t \cdot \mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is a column vector of all ones. We obtain that

$$W_t \cdot \boldsymbol{\omega}_t - \mathbf{1}\cdot\bar{\omega}_t^T = W_t \cdot \left(\boldsymbol{\omega}_t - \mathbf{1}\cdot\bar{\omega}_t^T\right).$$

By the definition of locally estimated TD error in (13), it follows that

$$\begin{aligned}\delta_{i,t} &:= r_{i,t} + \gamma\cdot\widehat{V}(s_{t+1};\omega_{i,t}) - \widehat{V}(s_t;\omega_{i,t})\\ &= r_{i,t} + \gamma\cdot\phi(s_{t+1})^T\omega_{i,t} - \phi(s_t)^T\omega_{i,t}\end{aligned} \tag{30}$$

where $\phi(\cdot) \in \mathbb{R}^K$ is the feature mapping for any state $s \in \mathcal{S}$. To perform the critic step according to equation (13), it holds that

$$\begin{aligned} \omega_{i,t} &= \Pi_{R_\omega}\bigg(\sum_{j\in\mathcal{N}_i(t)} W_{t-1}^{ij}\omega_{i,t-1} + \beta_t \cdot \delta_{i,t-1} \cdot \nabla_{\omega_i} \widehat{V}(s_{t-1}; \omega_{i,t-1}) \bigg), \\ &= \Pi_{R_\omega}\bigg(\sum_{j\in\mathcal{N}_i(t)} W_{t-1}^{ij}\omega_{i,t-1} + \beta_t \cdot \delta_{i,t-1} \cdot \phi(s_{t-1}) \bigg), \quad \forall i \in \mathcal{N} \end{aligned} \tag{31}$$

where $\nabla_{\omega_i} \widehat{V}(s_{t-1}; \omega_{i,t-1}) := \phi(s_{t-1})$ due to linear parameterization. Recall that $A(\boldsymbol{\theta}_t)$, $b_i(\boldsymbol{\theta}_t)$ are defined in (23), it holds that

$$\begin{aligned} b_i(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t)\cdot\omega_{i,t} &:= \mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}_t}(\cdot), \boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_t}(\cdot|s), s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})}\Big[\big(r_i(s,a) + \gamma\cdot\phi(s)^T\omega_{i,t} - \phi(s')^T\omega_{i,t}\big)\cdot\phi(s)\Big] \\ &= \mathbb{E}_{s_t\sim\mu_{\boldsymbol{\theta}_t}(\cdot), \boldsymbol{a}_t\sim\pi_{\boldsymbol{\theta}_t}(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,\boldsymbol{a}_t)}\Big[\delta_{i,t}\cdot\phi(s_t)\Big]. \end{aligned} \tag{32}$$

Hence, in (31) the estimated stochastic gradient at each iteration t is expressed as

$$\delta_{i,t}\cdot\phi(s_t) = \underbrace{\Big(\delta_{i,t}\cdot\phi(s_t) - [b_i(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t)\cdot\omega_{i,t}]\Big)}_{:=m_{i,t}} + \underbrace{\Big(b_i(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t)\cdot\omega_{i,t}\Big)}_{:=h_{i,t}} \tag{33}$$

where $h_{i,t}$ is the expectation of the estimated stochastic gradient $\delta_{i,t}\cdot\phi(s_t)$; $m_{i,t}$ denotes the deviation between $\delta_{i,t}\cdot\phi(s_t)$ and its expectation $h_{i,t}$.

Recall that the subroutine to update critic parameters $\boldsymbol{\omega}_t$ in (13) is given below:

$$\widetilde{\boldsymbol{\omega}}_{t-1} = W_{t-1}\cdot\boldsymbol{\omega}_{t-1}, \quad \omega_{i,t} = \Pi_{R_\omega}\bigg(\widetilde{\omega}_{i,t-1} + \beta_{t-1}\cdot\delta_{i,t-1}\cdot\nabla_{\omega_i}\widehat{V}(s_{t-1};\omega_{i,t-1})\bigg), \ \forall i \in \mathcal{N}. \tag{34}$$

It can be decomposed using the following steps:

$$\widetilde{\omega}_{i,t-1} = \sum_{j\in\mathcal{N}_i(t-1)} W_{t-1}^{ij}\cdot\omega_{i,t-1} \tag{35a}$$

$$y_{i,t-1} \overset{(33)}{=} \widetilde{\omega}_{i,t-1} + \beta_{t-1}\cdot(h_{i,t-1} + m_{i,t-1}) \tag{35b}$$

$$e_{i,t-1} = y_{i,t-1} - \Pi_{R_\omega}(y_{i,t-1}) \tag{35c}$$

$$\omega_{i,t} = \Pi_{R_\omega}(y_{i,t-1}) = y_{i,t-1} - e_{i,t-1}. \tag{35d}$$

Express the above updates in matrix form, it holds that

$$\widetilde{\boldsymbol{\omega}}_{t-1} = W_{t-1}\cdot\boldsymbol{\omega}_{t-1} \tag{36a}$$

$$\boldsymbol{y}_{t-1} = \widetilde{\boldsymbol{\omega}}_{t-1} + \beta_{t-1}\cdot(\boldsymbol{h}_{t-1} + \boldsymbol{m}_{t-1}) \tag{36b}$$

$$\boldsymbol{e}_{t-1} = \boldsymbol{y}_{t-1} - \Pi_{R_\omega}(\boldsymbol{y}_{t-1}) \tag{36c}$$

$$\boldsymbol{\omega}_t = \boldsymbol{y}_{t-1} - \boldsymbol{e}_{t-1} \tag{36d}$$

where $\boldsymbol{y}_t, \boldsymbol{h}_t, \boldsymbol{m}_t, \boldsymbol{e}_t$ correspond to the collections of local vectors $\{y_{i,t}\}, \{h_{i,t}\}, \{m_{i,t}\}, \{e_{i,t}\}$. Recall the definitions $\boldsymbol{\omega}_t = [\omega_{1,t}^T; \omega_{2,t}^T; \cdots; \omega_{N,t}^T] \in \mathbb{R}^{N\times K}$ and $\bar{\omega}_t^T := \frac{1}{N}\mathbf{1}^T\boldsymbol{\omega}_t$, it follows

$$\begin{aligned} \bar{\omega}_t^T &= \frac{1}{N}\mathbf{1}^T\boldsymbol{\omega}_t \\ &\overset{(i)}{=} \frac{1}{N}\mathbf{1}^T\big(W_{t-1}\cdot\boldsymbol{\omega}_{t-1} + \beta_{t-1}\cdot(\boldsymbol{h}_{t-1} + \boldsymbol{m}_{t-1}) - \boldsymbol{e}_{t-1}\big) \\ &\overset{(ii)}{=} \bar{\omega}_{t-1}^T + \beta_{t-1}\cdot(\bar{h}_{t-1}^T + \bar{m}_{t-1}^T) - \bar{e}_{t-1}^T \end{aligned} \tag{37}$$

where $\bar{h}_{t-1}$, $\bar{m}_{t-1}$ and $\bar{e}_{t-1}$ are the averaged vectors of $\mathbf{h}_{t-1}$, $\mathbf{m}_{t-1}$ and $\mathbf{e}_{t-1}$ (as defined similarly as $\bar{\omega}_{t-1}$); (i) is from the subroutine (36); (ii) is from $\mathbf{1}^T W_{t-1} = \mathbf{1}^T$ due to double stochasticity in weight matrix W_{t-1}. Recall that we have defined $Q = I - \frac{1}{N}\mathbf{1}\cdot\mathbf{1}^T$, then it is clear that $Q\cdot\boldsymbol{\omega}$ indicates the consensus error. We can express such an error as follows:

$$
\begin{aligned}
Q\cdot\boldsymbol{\omega}_t &= \boldsymbol{\omega}_t - \mathbf{1}\bar{\omega}_t^T \\
&= W_{t-1}\cdot\boldsymbol{\omega}_{t-1} + \beta_{t-1}\cdot(\boldsymbol{h}_{t-1}+\boldsymbol{m}_{t-1}) - \boldsymbol{e}_{t-1} - [\mathbf{1}\bar{\omega}_{t-1}^T + \beta_{t-1}\cdot\mathbf{1}[\bar{h}_{t-1}+\bar{m}_{t-1}]^T - \mathbf{1}\bar{e}_{t-1}^T] \\
&= W_{t-1}\cdot\big(\boldsymbol{\omega}_{t-1} - \mathbf{1}\bar{\omega}_{t-1}^T\big) + \beta_{t-1}\cdot\big(\boldsymbol{h}_{t-1} - \mathbf{1}\bar{h}_{t-1}^T\big) + \beta_{t-1}\cdot\big(\boldsymbol{m}_{t-1} - \mathbf{1}\bar{m}_{t-1}^T\big) - \big(\boldsymbol{e}_{t-1} - \mathbf{1}\bar{e}_{t-1}^T\big) \\
&= W_{t-1}\cdot Q\cdot\boldsymbol{\omega}_{t-1} + \beta_{t-1}\cdot Q\cdot(\boldsymbol{h}_{t-1}+\boldsymbol{m}_{t-1}) - Q\cdot\boldsymbol{e}_{t-1} \qquad (38) \\
&= \big(\Pi_{k=0}^{t-1}W_k\big)\cdot Q\cdot\boldsymbol{\omega}_0 - \sum_{k=0}^{t-1}(\Pi_{\ell=k+1}^{t-1}W_\ell)\cdot Q\cdot\boldsymbol{e}_k + \sum_{k=0}^{t-1}\beta_k\cdot(\Pi_{\ell=k+1}^{t-1}W_\ell)\cdot Q\cdot(\boldsymbol{h}_k+\boldsymbol{m}_k)\,. \qquad (39)
\end{aligned}
$$

Then we can bound the norm of the consensus error using the following:

$$
\begin{aligned}
\|Q\cdot\boldsymbol{\omega}_t\| &\overset{(i)}{\le} \|(\Pi_{k=0}^{t-1}W_k)\cdot Q\cdot\boldsymbol{\omega}_0\| + \sum_{k=0}^{t-1}\|(\Pi_{\ell=k+1}^{t-1}W_\ell)\cdot Q\cdot\boldsymbol{e}_{k-1}\| \\
&\quad + \sum_{k=0}^{t-1}\beta_k\|(\Pi_{\ell=k+1}^{t-1}W_\ell)\cdot Q\cdot(\boldsymbol{h}_k+\boldsymbol{m}_k)\,\| \\
&\overset{(ii)}{\le} \eta^{\lfloor t/B\rfloor}\cdot\|\boldsymbol{\omega}_0\| + \sum_{k=0}^{t-1}\eta^{\lfloor (t-k-1)/B\rfloor}\cdot\|\boldsymbol{e}_k\| + \sum_{k=0}^{t-1}\eta^{\lfloor (t-k-1)/B\rfloor}\cdot\beta_k\cdot\|\boldsymbol{h}_k+\boldsymbol{m}_k\| \\
&\overset{(iii)}{\le} \frac{1}{\eta}\cdot\rho^t\cdot\|\boldsymbol{\omega}_0\| + \frac{1}{\eta}\sum_{k=0}^{t-1}\rho^{t-k-1}\cdot\|\boldsymbol{e}_k\| + \frac{1}{\eta}\sum_{k=0}^{t-1}\rho^{t-k-1}\cdot\beta_k\cdot\|\boldsymbol{h}_k+\boldsymbol{m}_k\| \qquad (40)
\end{aligned}
$$

where (ii) follows (20); in (iii) we utilize that $\eta^{\lfloor t/B\rfloor}\le\eta^{t/B-1}=\frac{1}{\eta}\cdot\rho^t$ where we define $\rho:=\eta^{\frac{1}{B}}$.

Next we bound $\|\boldsymbol{h}_k+\boldsymbol{m}_k\|$ and $\|\boldsymbol{e}_k\|$. We have

$$
\begin{aligned}
\|h_{i,t}\|+\|m_{i,t}\| &\overset{(33)}{=} \Big\|\delta_{i,t}\cdot\phi(s_t) - [b_i(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t)\cdot\omega_{i,t}]\Big\| + \Big\|b_i(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t)\cdot\omega_{i,t}\Big\| \\
&\le \|\delta_{i,t}\cdot\phi(s_t)\| + 2\cdot\|b_i(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t)\cdot\omega_{i,t}\| \\
&\overset{(23)}{=} \|\delta_{i,t}\cdot\phi(s_t)\| + 2\cdot\Big\|\mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}_t}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_t}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})}\Big[\big(r_i(s,a)+\gamma\cdot\phi(s)^T\omega_{i,t} - \phi(s')^T\omega_{i,t}\big)\cdot\phi(s)\Big]\Big\| \\
&\overset{(i)}{\le} \|\delta_{i,t}\cdot\phi(s_t)\| + 2\cdot\mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}_t}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_t}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})}\Big[\Big\|\big(r_i(s,a)+\gamma\cdot\phi(s)^T\omega_{i,t} - \phi(s')^T\omega_{i,t}\big)\cdot\phi(s)\Big\|\Big] \\
&\overset{(ii)}{\le} \|\delta_{i,t}\| + 2\cdot\mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}_t}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_t}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})}\Big[\Big\|\big(r_i(s,a)+\gamma\cdot\phi(s)^T\omega_{i,t} - \phi(s')^T\omega_{i,t}\big)\Big\|\Big] \\
&\overset{(30)}{=} \Big\|r_{i,t}+\gamma\cdot\phi(s_{t+1})^T\omega_{i,t} - \phi(s_t)^T\omega_{i,t}\Big\| \\
&\quad + 2\cdot\mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}_t}(\cdot),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}_t}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,\boldsymbol{a})}\Big[\Big\|\big(r_i(s,a)+\gamma\cdot\phi(s)^T\omega_{i,t} - \phi(s')^T\omega_{i,t}\big)\Big\|\Big] \\
&\overset{(iii)}{\le} 3R_{\max} + 3(1+\gamma)\cdot R_\omega \qquad (41)
\end{aligned}
$$

where (i) follows Jensen's inequality; (ii) follows Assumption 4 that $\|\phi(s)\|\le 1$ for any $s\in\mathcal{S}$; (iii) follows the fact that $|r_i(s,a)|\le R_{\max}$ and the critic parameter $\omega_{i,t}$ is constrained in a fixed region $\|\omega_{i,t}\|\le R_\omega$. For simplicity, in the following part we denote $L_b := 3R_{\max}+3(1+\gamma)\cdot R_\omega$.

Moreover, it holds that

$$
\|e_{i,t}\| \overset{(35)}{=} \|y_{i,t} - \Pi_{R_\omega}(y_{i,t})\| \overset{(i)}{\le} \|\widetilde{\omega}_{i,t} - y_{i,t}\| = \|\beta_t\cdot(h_{i,t}+m_{i,t})\,\| \overset{(ii)}{\le} \beta_t\cdot L_b, \qquad (42)
$$

where (i) follows from (16b) and (ii) follows from (41).

Recall that the stepsizes in critic steps are defined as $\beta_t := \frac{\beta}{T^{\sigma_2}}$, $\forall t$. Plugging (41) and (42) into (40), we get

$$\begin{aligned}\|Q\cdot\boldsymbol{\omega}_t\| &\leq \frac{1}{\eta}\cdot\rho^t\cdot\|\boldsymbol{\omega}_0\| + \frac{1}{\eta}\sum_{k=0}^{t-1}\rho^{t-k-1}\cdot\beta_k\cdot L_b\cdot N + \frac{1}{\eta}\sum_{k=0}^{t-1}\rho^{t-k-1}\cdot\beta_k\cdot L_b\cdot N \quad (43)\\ &= \frac{1}{\eta}\cdot\rho^t\cdot\|\boldsymbol{\omega}_0\| + \frac{2N\cdot L_b\cdot\beta}{\eta\cdot T^{\sigma_2}}\cdot\sum_{k=0}^{t-1}\rho^{t-k-1}\\ &\overset{(a)}{\leq} \frac{1}{\eta}\cdot\rho^t\cdot\|\boldsymbol{\omega}_0\| + \frac{2N\cdot L_b\cdot\beta}{\eta\cdot T^{\sigma_2}}\cdot\frac{1}{1-\rho} \quad (44)\end{aligned}$$

where (a) follows the fact that $\sum_{k=0}^{t-1}\rho^{t-k-1}\leq\frac{1}{1-\rho}$, $\forall t$.

Summing (44) from $t=0$ to $t=T-1$, we obtain

$$\begin{aligned}\sum_{t=0}^{T-1}\|Q\cdot\boldsymbol{\omega}_t\| &\leq \frac{\|\boldsymbol{\omega}_0\|}{\eta}\sum_{t=0}^{T-1}\rho^t + T\cdot\frac{2N\cdot L_b\cdot\beta}{\eta\cdot T^{\sigma_2}}\cdot\frac{1}{1-\rho}\\ &\overset{(i)}{\leq} \frac{\|\boldsymbol{\omega}_0\|}{\eta\cdot(1-\rho)} + \frac{2N\cdot L_b\cdot\beta}{\eta\cdot(1-\rho)}\cdot T^{1-\sigma_2}, \quad (45)\end{aligned}$$

where (i) is due to the fact that $\sum_{t=0}^{T-1}\rho^t\leq\frac{1}{1-\rho}$.

In summary, we obtain the following bound on the averaged consensus violation:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|Q\cdot\boldsymbol{\omega}_t\| \leq \frac{1}{T}\cdot\frac{\|\boldsymbol{\omega}_0\|}{\eta\cdot(1-\rho)} + + \frac{2N\cdot L_b\cdot\beta}{\eta\cdot(1-\rho)}\cdot T^{-\sigma_2}. \quad (46)$$

Extending above analysis steps on deriving a bound for the consensus error $\|Q\cdot\boldsymbol{\omega}_t\|$ in (45) to $\|Q\cdot\boldsymbol{\lambda}_t\|$, we can show that the following holds:

$$\|Q\cdot\boldsymbol{\lambda}_t\| \leq \frac{1}{\eta}\cdot\rho^t\cdot\|\boldsymbol{\lambda}_0\| + \frac{2N\cdot L_b\cdot\zeta}{\eta\cdot T^{\sigma_2}}\cdot\frac{1}{1-\rho} \quad (47)$$

where the stepsize ζ_t is defined as $\zeta_t := \frac{\zeta}{T^{\sigma_2}}$. Similar as (45), summing up (47) from $t=0$ to $t=T-1$, it holds that:

$$\sum_{t=0}^{T-1}\|Q\cdot\boldsymbol{\lambda}_t\| \leq \frac{\|\boldsymbol{\lambda}_0\|}{\eta\cdot(1-\rho)} + + \frac{2N\cdot L_b\cdot\zeta}{\eta\cdot(1-\rho)}\cdot T^{1-\sigma_2}. \quad (48)$$

In summary, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|Q\cdot\boldsymbol{\lambda}_t\| \leq \frac{1}{T}\cdot\frac{\|\boldsymbol{\omega}_0\|}{\eta\cdot(1-\rho)} + + \frac{2N\cdot L_b\cdot\zeta}{\eta\cdot(1-\rho)}\cdot T^{-\sigma_2}. \quad (49)$$

Summing up (46) and (49), we obtain the convergence rate of the consensus errors:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\|Q\cdot\boldsymbol{\omega}_t\| + \|Q\cdot\boldsymbol{\lambda}_t\|\Big) &= \frac{1}{T}\cdot\frac{\|\boldsymbol{\omega}_0\| + \|\boldsymbol{\lambda}_0\|}{\eta\cdot(1-\rho)} + \frac{1}{T^{\sigma_2}}\cdot\frac{2N\cdot L_b\cdot(\beta+\zeta)}{\eta\cdot(1-\rho)}\\ &= \mathcal{O}(T^{-\sigma_2}). \quad (50)\end{aligned}$$

Taking square on both side of (44) and applying Cauchy-Schwarz inequality, it holds that

$$\|Q\cdot\boldsymbol{\omega}_t\|^2 \leq \frac{2}{\eta^2}\cdot\rho^{2t}\cdot\|\boldsymbol{\omega}_0\|^2 + \frac{2}{T^{2\sigma_2}}\cdot\left(\frac{2N\cdot L_b\cdot\beta}{\eta\cdot(1-\rho)}\right)^2 \quad (51)$$

Summing (51) from $t=0$ to $t=T-1$, it holds that

$$\sum_{t=0}^{T-1}\|Q\cdot\boldsymbol{\omega}_t\|^2 \leq \frac{2\|\boldsymbol{\omega}_0\|^2}{\eta^2\cdot(1-\rho^2)} + 2\cdot\left(\frac{2N\cdot L_b\cdot\beta}{\eta\cdot(1-\rho)}\right)^2\cdot T^{1-2\sigma_2} \quad (52)$$

Extending above analysis steps on deriving a bound for the consensus error $\|Q \cdot \boldsymbol{\omega}_t\|^2$ in (52) to $\|Q \cdot \boldsymbol{\lambda}_t\|^2$, we can show that the following holds:

$$\sum_{t=0}^{T-1} \|Q \cdot \boldsymbol{\lambda}_t\|^2 \leq \frac{2\|\boldsymbol{\lambda}_0\|^2}{\eta^2 \cdot (1-\rho^2)} + 2 \cdot \left(\frac{2N \cdot L_b \cdot \zeta}{\eta \cdot (1-\rho)}\right)^2 \cdot T^{1-2\sigma_2} \tag{53}$$

Hence, adding (52) and (53), then divide both side by T, it holds that

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\bigg(\|Q \cdot \boldsymbol{\lambda}_t\|^2 + \|Q \cdot \boldsymbol{\omega}_t\|^2\bigg) &\leq \frac{2(\|\boldsymbol{\lambda}_0\|^2 + \|\boldsymbol{\omega}_0\|^2)}{T \cdot \eta^2 \cdot (1-\rho^2)} + \frac{8N^2 \cdot L_b^2 \cdot (\zeta^2 + \beta^2)}{\eta^2 \cdot (1-\rho)^2} \cdot T^{-2\sigma_2} \\ &= \mathcal{O}(T^{-1}) + \mathcal{O}(T^{-2\sigma_2}) \end{aligned} \tag{54}$$

This completes the proof for the first part. □

Step 2. In this part, we analyze the consensus errors for the shared policy parameters $\boldsymbol{\theta}^s$.

Since the mixing matrix W_t is doubly stochastic which implies $W_t \cdot \mathbf{1} = \mathbf{1}$, we obtain that

$$W_t \cdot \boldsymbol{\theta}_t^s - \mathbf{1}\bar{\theta}_t^{s^T} = W_t \cdot \left(\boldsymbol{\theta}_t^s - \mathbf{1}\bar{\theta}_t^{s^T}\right)$$

where we have defined $\bar{\theta}_t^{s^T} := \frac{1}{N}\mathbf{1}^T\boldsymbol{\theta}_t^s$. Recall that the subroutine to update shared policy parameters in (9) - (10) is given below:

$$\theta_{i,t}^s := \sum_{j \in \mathcal{N}_i(t-1)} W_{t-1}^{ij}\theta_{j,t-1}^s + \alpha_{t-1} \cdot \widehat{\delta}_{i,t-1} \cdot \nabla_{\theta_i^s} \log \pi_i(a_{i,t-1}|s_{t-1}, \theta_{i,t-1}), \ \ \forall i \in \mathcal{N} \tag{55}$$

where we have defined

$$\begin{aligned} \widehat{\delta}_{i,t-1} &:= \widehat{r}(s_{t-1}, \boldsymbol{a}_{t-1}; \lambda_{i,t-1}) + \gamma \cdot \widehat{V}(s_t; \omega_{i,t-1}) - \widehat{V}(s_{t-1}; \omega_{i,t-1}) \\ &= \varphi(s_{t-1}, \boldsymbol{a}_{t-1})^T\lambda_{i,t-1} + \gamma \cdot \phi(s_{t-1})^T\omega_{i,t-1} - \phi(s_t)^T\omega_{i,t}. \end{aligned} \tag{56}$$

Then we define $g_{i,t-1} := \widehat{\delta}_{i,t-1} \cdot \nabla_{\theta_i^s} \log \pi_i(a_{i,t-1}|s_{t-1}, \theta_{i,t-1})$ and $\boldsymbol{g} := [g_1^T; g_2^T; \cdots; g_N^T]$, it holds that

$$\boldsymbol{\theta}_t^s := W_{t-1} \cdot \boldsymbol{\theta}_{t-1}^s + \alpha_{t-1} \cdot \boldsymbol{g}_{t-1}. \tag{57}$$

Recall $Q = I - \frac{1}{N}\mathbf{1} \cdot \mathbf{1}^T$, we analyze the consensus error $Q \cdot \boldsymbol{\theta}_t^s$ as below:

$$\begin{aligned} Q \cdot \boldsymbol{\theta}_t^s =& \boldsymbol{\theta}_t^s - \mathbf{1}\bar{\theta}_t^{s^T} \\ =& \bigg(W_{t-1} \cdot \boldsymbol{\theta}_{t-1}^s + \alpha_{t-1} \cdot \boldsymbol{g}_{t-1}\bigg) - \bigg(\mathbf{1}\bar{\theta}_{t-1}^{s^T} + \alpha_{t-1} \cdot \mathbf{1}\bar{g}_{t-1}^T\bigg) \\ =& W_{t-1} \cdot Q \cdot \boldsymbol{\theta}_{t-1}^s + \alpha_{t-1} \cdot Q \cdot \boldsymbol{g}_{t-1} \\ =& (\Pi_{k=0}^{t-1}W_k) \cdot Q \cdot \boldsymbol{\theta}_0^s + \sum_{k=0}^{t-1}\alpha_k \cdot (\Pi_{l=k+1}^{t-1}W_l) \cdot Q \cdot \boldsymbol{g}_k. \end{aligned} \tag{58}$$

By Assumptions 3 - 4, 5, the estimated stochastic gradient $\boldsymbol{g}_t = [g_{1,t}^T, g_{2,t}^T, \cdots, g_{N,t}^T]$ can be bounded as below:

$$\begin{aligned} \|\boldsymbol{g}_t\| &\overset{(i)}{\leq} \sum_{i=1}^N \|g_{i,t}\| \\ &= \sum_{i=1}^N \|\widehat{\delta}_{i,t-1} \cdot \nabla_{\theta_i^s} \log \pi_i(a_{i,t-1}|s_{t-1}, \theta_{i,t-1})\| \\ &\overset{(ii)}{\leq} C_\psi \cdot \sum_{i=1}^N \bigg\|\varphi(s_{t-1}, \boldsymbol{a}_{t-1})^T\lambda_{i,t-1} + \gamma \cdot \phi(s_{t-1})^T\omega_{i,t-1} - \phi(s_t)^T\omega_{i,t}\bigg\| \\ &\leq C_\psi \cdot \sum_{i=1}^N \bigg(\|\varphi(s_{t-1}, \boldsymbol{a}_{t-1})\| \cdot \|\lambda_{i,t-1}\| + \gamma\|\phi(s_{t-1})\| \cdot \|\omega_{i,t-1}\| + \|\phi(s_t)\| \cdot \|\omega_{i,t}\|\bigg) \\ &\overset{(iii)}{\leq} N \cdot C_\psi \cdot \bigg(R_\lambda + (1+\gamma) \cdot R_\omega\bigg) := \ell_p \end{aligned} \tag{59}$$

where (i) follows the definition $\boldsymbol{g}_t = [g_{1,t}^T, g_{2,t}^T, \cdots, g_{N,t}^T]$ and Triangle inequality; (ii) follows that $\|\nabla_{\theta_i^s} \log \pi_i(a_{i,t-1}|s_{t-1}, \theta_{i,t-1})\| \leq C_\psi$ in Assumption 5; (iii) is due to the assumptions that $\|\varphi(s_{t-1}, \boldsymbol{a}_{t-1})\| \leq 1$ and $\|\phi(s_{t-1})\| \leq 1$, as well as that approximation parameters are restricted in fixed regions, so $\|\omega_{i,t}\| \leq R_\omega$ and $\|\lambda_{i,t}\| \leq R_\lambda$; In the last equality, we have defined ℓ_p as

$$\ell_p := N \cdot C_\psi \cdot \bigg(R_\lambda + (1+\gamma) R_\omega \bigg). \tag{60}$$

Recall that the stepsizes in policy optimization are defined as $\alpha_t := \frac{\alpha}{T^{\sigma_1}}, \forall\, t$. Taking Frobenius norm on both side of (58), we have:

$$\begin{aligned}
\|Q \cdot \boldsymbol{\theta}_t^s\| &\overset{(i)}{\leq} \|(\Pi_{k=0}^{t-1} W_k) \cdot Q \cdot \boldsymbol{\theta}_0^s\| + \sum_{k=0}^{t-1} \alpha_k \|(\Pi_{l=k+1}^{t-1} W_l) \cdot Q \cdot \boldsymbol{g}_k\| \\
&\overset{(ii)}{\leq} \eta^{\lfloor t/\mathcal{B} \rfloor} \cdot \|\boldsymbol{\theta}_0^s\| + \sum_{k=0}^{t-1} \alpha_k \cdot \eta^{\lfloor (t-k-1)/\mathcal{B} \rfloor} \cdot \|\boldsymbol{g}_k\| \\
&\overset{(iii)}{\leq} \frac{1}{\eta} \cdot \rho^t \cdot \|\boldsymbol{\theta}_0^s\| + \frac{1}{\eta} \sum_{k=0}^{t-1} \alpha_k \cdot \rho^{t-k-1} \cdot \|\boldsymbol{g}_k\| \\
&\overset{(iv)}{\leq} \frac{1}{\eta} \cdot \rho^t \cdot \|\boldsymbol{\theta}_0^s\| + \frac{\ell_p}{\eta} \sum_{k=0}^{t-1} \rho^{t-k-1} \cdot \alpha_k \\
&= \frac{1}{\eta} \cdot \rho^t \cdot \|\boldsymbol{\theta}_0^s\| + \frac{\ell_p \cdot \alpha}{\eta \cdot T^{\sigma_1}} \cdot \sum_{k=0}^{t-1} \rho^{t-k-1} \\
&\leq \frac{1}{\eta} \cdot \rho^t \cdot \|\boldsymbol{\theta}_0^s\| + \frac{\ell_p \cdot \alpha}{\eta \cdot (1-\rho)} \cdot \frac{1}{T^{\sigma_1}}
\end{aligned} \tag{61}$$

where (ii) follows from (20); in (iii) we utilize that $\eta^{\lfloor t/B \rfloor} \leq \eta^{t/B-1} = \frac{1}{\eta} \cdot \rho^t$ where we have defined $\rho := \eta^{\frac{1}{B}}$; (iv) follows from (59). Summing (61) from $t = 0$ to $t = T-1$, it holds that:

$$\begin{aligned}
\sum_{t=0}^{T-1} \|Q \cdot \boldsymbol{\theta}_t^s\| &\leq \frac{\|\boldsymbol{\theta}_0^s\|}{\eta} \sum_{t=0}^{T-1} \rho^t + \frac{\ell_p \cdot \alpha_0}{\eta \cdot (1-\rho)} \cdot \frac{T}{T^{\sigma_1}} \\
&\leq \frac{\|\boldsymbol{\theta}_0^s\|}{\eta \cdot (1-\rho)} + \frac{\ell_p \cdot \alpha_0}{\eta \cdot (1-\rho)} \cdot T^{1-\sigma_1} \\
&= \mathcal{O}(T^{1-\sigma_1}).
\end{aligned} \tag{62}$$

Then dividing T on both side of (62), it holds that

$$\frac{1}{T} \sum_{t=0}^{T-1} \|Q \cdot \boldsymbol{\theta}_t^s\| \leq \mathcal{O}(T^{-\sigma_1}) \tag{63}$$

where consensus error converges to 0 as T goes to infinity.

Moreover, we can provide a bound for the averaged consensus error squared $\sum_{t=0}^{T-1} \|Q \cdot \boldsymbol{\theta}_t^s\|^2$. Taking square on both sides of (61) and summing from $t = 0$ to $t = T-1$. We obtain:

$$\begin{aligned}
\sum_{t=0}^{T-1} \|Q \cdot \boldsymbol{\theta}_t^s\|^2 &\overset{(61)}{\leq} \sum_{t=0}^{T-1} \left(\frac{1}{\eta} \cdot \rho^t \cdot \|\boldsymbol{\theta}_0^s\| + \frac{\ell_p \cdot \alpha}{\eta \cdot (1-\rho)} \cdot \frac{1}{T^{\sigma_1}} \right)^2 \\
&\overset{(i)}{\leq} \sum_{t=0}^{T-1} \left(\rho^{2t} \cdot \frac{2\|\boldsymbol{\theta}_0^s\|^2}{\eta^2} + \frac{2\ell_p^2 \cdot \alpha^2}{\eta^2 \cdot (1-\rho)^2} \cdot \frac{1}{T^{2\sigma_1}} \right) \\
&\overset{(ii)}{\leq} \frac{2\|\boldsymbol{\theta}_0^s\|^2}{\eta^2 \cdot (1-\rho^2)} + \frac{2\ell_p^2 \cdot \alpha^2}{\eta^2 \cdot (1-\rho)^2} \cdot T^{1-2\sigma_1}
\end{aligned} \tag{64}$$

where (i) follows from Cauchy–Schwarz inequality; (ii) follows from $\frac{2\|\boldsymbol{\theta}_0^s\|^2}{\eta^2} \cdot \sum_{t=0}^{T-1} \rho^{2t} \leq \frac{2\|\boldsymbol{\theta}_0^s\|^2}{\eta^2 \cdot (1-\rho^2)}$.

Then dividing T on both side of (64), it holds that

$$\frac{1}{T} \sum_{t=0}^{T-1} \|Q \cdot \boldsymbol{\theta}_t^s\|^2 \leq \mathcal{O}(T^{-1}) + \mathcal{O}(T^{-2\sigma_1}). \tag{65}$$

This completes the proof for the second step. □

G Proof of Proposition 1

Proof. In this proof, we show the convergence results of all approximation parameters $\boldsymbol{\omega}_t$ and $\boldsymbol{\lambda}_t$. We first analyze the convergence error $\|\bar{\omega}_t - \omega^*(\boldsymbol{\theta}_t)\|$ and then extend the results to $\|\bar{\lambda}_t - \lambda^*(\boldsymbol{\theta}_t)\|$. For simplicity, we write $\omega^*(\boldsymbol{\theta}_t)$ as ω_t^* and $\lambda^*(\boldsymbol{\theta}_t)$ as λ_t^*.

Denoting the expectation over data sampling procedures as $\mathbb{E}[\cdot]$, let us begin by bounding the error $\mathbb{E}[\|\bar{\omega}_{t+1} - \omega_{t+1}^*\|^2]$ as below:

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\omega}_{t+1} - \omega_{t+1}^*\|^2\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\|\bar{\omega}_t + \beta_t \cdot (\bar{h}_t + \bar{m}_t) - \bar{e}_t - \omega_t^* + \omega_t^* - \omega_{t+1}^*\|^2\right] \\
&= \mathbb{E}[\|\bar{\omega}_t - \omega_t^*\|^2] + \mathbb{E}[\|\beta_t \cdot (\bar{h}_t + \bar{m}_t) - \bar{e}_t + \omega_t^* - \omega_{t+1}^*\|^2] + 2\beta_t \cdot \mathbb{E}\left[\langle \bar{\omega}_t - \omega_t^*, \bar{h}_t + \bar{m}_t \rangle\right] \\
&\quad - 2\mathbb{E}[\langle \bar{\omega}_t - \omega_t^*, \bar{e}_t \rangle] + 2\mathbb{E}\left[\langle \bar{\omega}_t - \omega_t^*, \omega_t^* - \omega_{t+1}^* \rangle\right] \\
&\overset{(ii)}{\leq} \mathbb{E}[\|\bar{\omega}_t - \omega_t^*\|^2] + \underbrace{2\beta_t \cdot \mathbb{E}\left[\langle \bar{\omega}_t - \omega_t^*, \bar{h}_t + \bar{m}_t \rangle\right]}_{\text{term A}} + \underbrace{2\mathbb{E}[\|\omega_t^* - \omega_{t+1}^*\|^2] + 2\mathbb{E}\left[\langle \bar{\omega}_t - \omega_t^*, \omega_t^* - \omega_{t+1}^* \rangle\right]}_{\text{term B}} \\
&\quad + \underbrace{2\mathbb{E}[\|\beta_t \cdot (\bar{h}_t + \bar{m}_t) - \bar{e}_t\|^2]}_{\text{term C}} \underbrace{-2\mathbb{E}[\langle \bar{\omega}_t - \omega_t^*, \bar{e}_t \rangle]}_{\text{term D}}
\end{aligned} \tag{66}$$

where (i) follows (37); (ii) is from Cauchy-Schwarz inequality. Recall that $\bar{h}_t^T := \frac{1}{N}\mathbf{1}^T \boldsymbol{h}_t$, $\bar{m}_t^T := \frac{1}{N}\mathbf{1}^T \boldsymbol{m}_t$ and $\bar{e}_t^T := \frac{1}{N}\mathbf{1}^T \boldsymbol{e}_t$, where $\boldsymbol{h}_t$, $\boldsymbol{m}_t$ and $\boldsymbol{e}_t$ are defined in (36).

In the following, let us analyze each component in (66). First term A can be expressed below:

$$2\beta_t \cdot \mathbb{E}\bigg[\langle \bar{\omega}_t - \omega_t^*, \bar{h}_t + \bar{m}_t \rangle\bigg] \overset{(a)}{=} 2\beta_t \cdot \mathbb{E}\bigg[\langle \bar{\omega}_t - \omega_t^*, \mathbb{E}[\bar{h}_t + \bar{m}_t | \mathcal{F}_t] \rangle\bigg] \overset{(b)}{=} 2\beta_t \cdot \mathbb{E}\bigg[\langle \bar{\omega}_t - \omega_t^*, \bar{h}_t \rangle\bigg] \tag{67}$$

where $\mathcal{F}_t$ is the σ-algebra generated by $\mathcal{F}_t = \{\boldsymbol{\omega}_t, \boldsymbol{\theta}_t, \cdots, \boldsymbol{\omega}_0, \boldsymbol{\theta}_0\}$; (a) follows Tower rule in expectations; (b) is due to the fact that $\mathbb{E}[m_{i,t}|\mathcal{F}_t] = 0,\ \forall i \in \mathcal{N}$, which is from (32) and (33). Recall that in (23) we have defined:

$$A(\boldsymbol{\theta}) := \mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}}(\cdot), s'\sim\mathcal{P}^{\pi_{\boldsymbol{\theta}}}(\cdot|s)}\left[\phi(s) \cdot \big(\phi(s) - \gamma \cdot \phi(s')\big)^T\right],\ \forall\, i \in \mathcal{N}, \tag{68a}$$

$$b_i(\boldsymbol{\theta}) := \mathbb{E}_{s\sim\mu_{\boldsymbol{\theta}}(\cdot), \boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[r_i(s,a) \cdot \phi(s)\right],\ \forall\, i \in \mathcal{N}. \tag{68b}$$

Also in (22), it has shown that ω_t^* satisfies the condition

$$A(\boldsymbol{\theta}_t) \cdot \omega_t^* = \frac{1}{N}\sum_{i=1}^N b_i(\boldsymbol{\theta}_t) = \bar{b}(\boldsymbol{\theta}_t) \tag{69}$$

where we define $\bar{b}(\boldsymbol{\theta}_t) := \frac{1}{N}\sum_{i=1}^N b_i(\boldsymbol{\theta}_t)$. Recall that we have defined

$$\bar{h}_t := \frac{1}{N}\sum_{i=1}^N h_{i,t} = \frac{1}{N}\sum_{i=1}^N \bigg(b_i(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t) \cdot \omega_{i,t}\bigg) = \bar{b}(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t) \cdot \bar{\omega}_t,$$

then it holds that

$$\bar{h}_t = \bar{b}(\boldsymbol{\theta}_t) - A(\boldsymbol{\theta}_t) \cdot \bar{\omega}_t \overset{(69)}{=} A(\boldsymbol{\theta}_t) \cdot (\omega_t^* - \bar{\omega}_t) \tag{70}$$

Therefore, plugging (70) into (67), term A in (66) could be bounded as below:

$$2\beta_t \cdot \mathbb{E}\bigg[\langle \bar{\omega}_t - \omega_t^*, \bar{h}_t \rangle\bigg] = -2\beta_t \cdot \mathbb{E}\bigg[(\bar{\omega}_t - \omega_t^*)^T A(\boldsymbol{\theta}_t) \cdot (\bar{\omega}_t - \omega_t^*)\bigg] \overset{(i)}{\leq} -2\beta_t \cdot \tilde{c}_{\min} \cdot \mathbb{E}\bigg[\|\bar{\omega}_t - \omega_t^*\|^2\bigg] \tag{71}$$

where (i) is due to the fact that $A(\boldsymbol{\theta}_t)$ is a positive definite matrix and its minimum eigenvalue $c_{\min}(A(\boldsymbol{\theta}_t)) \geq \tilde{c}_{\min}$ in (24).

Second, term B can be bounded as below:

$$\begin{aligned}
&2\mathbb{E}[\|\omega_t^* - \omega_{t+1}^*\|^2] + 2\mathbb{E}\left[\langle \bar{\omega}_t - \omega_t^*, \omega_t^* - \omega_{t+1}^* \rangle\right] \\
&\overset{(a)}{\leq} 2\mathbb{E}[\|\omega_t^* - \omega_{t+1}^*\|^2] + 2\mathbb{E}\left[\|\bar{\omega}_t - \omega_t^*\| \cdot \|\omega_t^* - \omega_{t+1}^*\|\right] \\
&\overset{(b)}{\leq} 2L_\omega^2 \cdot \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|^2] + 2L_\omega \cdot \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\| \cdot \|\bar{\omega}_t - \omega_t^*\|]
\end{aligned} \tag{72}$$

where (a) follows Cauchy-Schwarz inequality; (b) is from the Lipschitz property (25) in Lemma 6. Recall the definition $\boldsymbol{\theta}_t := \{\boldsymbol{\theta}_t^s, \boldsymbol{\theta}_t^p\}$ and $Q = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$, we bound $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|$ as below:

$$\begin{aligned}
\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\| &\overset{(i)}{\leq} \|\boldsymbol{\theta}_t^s - \boldsymbol{\theta}_{t+1}^s\| + \|\boldsymbol{\theta}_t^l - \boldsymbol{\theta}_{t+1}^l\| \\
&\overset{(ii)}{\leq} \|\boldsymbol{\theta}_t^s - \mathbf{1}\bar{\theta}_t^s\| + \|\mathbf{1}\bar{\theta}_t^s - \mathbf{1}\bar{\theta}_{t+1}^s\| + \|\mathbf{1}\bar{\theta}_{t+1}^s - \boldsymbol{\theta}_{t+1}^s\| + \|\boldsymbol{\theta}_t^l - \boldsymbol{\theta}_{t+1}^l\| \\
&\overset{(iii)}{\leq} \|Q \cdot \boldsymbol{\theta}_t^s\| + \|Q \cdot \boldsymbol{\theta}_{t+1}^s\| + 2\alpha_t \cdot \ell_p \\
&\overset{(57)}{=} \|Q \cdot \boldsymbol{\theta}_t^s\| + \|Q \cdot (W_t \cdot \boldsymbol{\theta}_t^s + \alpha_t \cdot \boldsymbol{g}_t)\| + 2\alpha_t \cdot \ell_p \\
&\overset{(iv)}{\leq} \|Q \cdot \boldsymbol{\theta}_t^s\| + \|Q \cdot W_t \cdot \boldsymbol{\theta}_t^s\| + \alpha_t \cdot \|Q \cdot \boldsymbol{g}_t\| + 2\alpha_t \cdot \ell_p \\
&\overset{(v)}{\leq} \|Q \cdot \boldsymbol{\theta}_t^s\| + \|W_t \cdot Q \cdot \boldsymbol{\theta}_t^s\| + 3\alpha_t \cdot \ell_p \\
&\overset{(vi)}{\leq} 2\|Q \cdot \boldsymbol{\theta}_t^s\| + 3\alpha_t \cdot \ell_p.
\end{aligned} \tag{73}$$

where (i) and (ii) follow Triangle inequality; (iii) is due to the fact that estimated policy gradient in updating $\boldsymbol{\theta}_t$ is bounded by ℓ_p in (59); (iv) follows Cauchy-Schwarz inequality; in (v) we used the boudnedness of the gradient (59), the fact that eigenvalue value of Q is upper bounded bounded by 1, as well as the following:

$$Q \cdot W_t = \left(I - \frac{1}{N}\mathbf{1} \cdot \mathbf{1}^T\right) \cdot W_t = W_t - \frac{1}{N}\mathbf{1} \cdot \mathbf{1}^T = W_t \cdot \left(I - \frac{1}{N}\mathbf{1} \cdot \mathbf{1}^T\right) = W_t \cdot Q. \tag{74}$$

Additionally, (vi) is due to the fact that the eigenvalue of weight matrix W_t is bounded by 1. Then plugging the inequality (73) into (72), term B could be bounded as follows:

$$\begin{aligned}
&2\mathbb{E}[\|\omega_t^* - \omega_{t+1}^*\|^2] + 2\mathbb{E}\left[\langle \bar{\omega}_t - \omega_t^*, \omega_t^* - \omega_{t+1}^* \rangle\right] \\
&\overset{(72)}{\leq} 2L_\omega^2 \cdot \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|^2] + 2L_\omega \cdot \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\| \cdot \|\bar{\omega}_t - \omega_t^*\|] \\
&\overset{(i)}{\leq} 2L_\omega^2 \cdot \bigg(8\mathbb{E}\big[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\big] + 18\alpha_t^2 \cdot \ell_p^2\bigg) + 2L_\omega \cdot \mathbb{E}\bigg[\big(2\|Q \cdot \boldsymbol{\theta}_t^s\| + 3\alpha_t \cdot \ell_p\big) \cdot \|\bar{\omega}_t - \omega_t^*\|\bigg] \\
&= 16L_\omega^2 \cdot \mathbb{E}[\|Q \cdot \boldsymbol{\theta}_t^s\|^2] + 36L_\omega^2 \cdot \ell_p^2 \cdot \alpha_t^2 + 6L_\omega \cdot \ell_p \cdot \alpha_t \cdot \mathbb{E}[\|\bar{\omega}_t - \omega_t^*\|] \\
&\quad + 4L_\omega \cdot \mathbb{E}\bigg[\|Q \cdot \boldsymbol{\theta}_t^s\| \cdot \|\bar{\omega}_t - \omega_t^*\|\bigg]
\end{aligned} \tag{75}$$

where (i) follows (73) and the Cauchy-Schwarz inequality.

Recall the definition of $\bar{h}_t$ as $\bar{h}_t := \frac{1}{N}\sum_{i=1}^N h_{i,t}$, where $\bar{m}_t$ and $\bar{e}_t$ are also defined similarly. Then term C in (66) could be bounded as below:

$$\begin{aligned}
2\mathbb{E}[\|\beta_t(\bar{h}_t + \bar{m}_t) - \bar{e}_t\|^2] &\overset{(i)}{\leq} 4\beta_t^2 \cdot \mathbb{E}[\|\bar{h}_t + \bar{m}_t\|^2] + 4\mathbb{E}[\|\bar{e}_t\|^2] \\
&= 4\beta_t^2 \cdot \mathbb{E}\Big[\|\frac{1}{N}\sum_{i=1}^N \big(h_{i,t} + m_{i,t}\big)\|^2\Big] + 4\mathbb{E}\Big[\|\frac{1}{N}\sum_{i=1}^N e_{i,t}\|^2\Big] \\
&\overset{(ii)}{\leq} 4\beta_t^2 \cdot \mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^N \|h_{i,t} + m_{i,t}\|^2\Big] + 4\mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^N \|e_{i,t}\|^2\Big] \\
&\overset{(iii)}{\leq} 8\beta_t^2 \cdot L_b^2
\end{aligned} \tag{76}$$

where (i) follows Cauchy-Schwarz inequality; (ii) follows Jensen's inequality; (iii) is from the inequalities (41) and (42).

Recall that $\bar{e}_t := \frac{1}{N}\sum_{i=1}^N e_{i,t}$ and in (35) we have defined $y_{i,t} := \widetilde{\omega}_{i,t} + \beta_t \cdot (h_{i,t} + m_{i,t})$. Then term D in (66) can be bounded as below:

$$\begin{aligned}
& - 2\mathbb{E}\Big[\langle \bar{\omega}_t - \omega_t^*, \bar{e}_t\rangle\Big] \\
& = -\frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\langle \bar{\omega}_t - \omega_t^*, e_{i,t}\rangle\Big] \\
& = -\frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\langle \bar{\omega}_t - y_{i,t}, e_{i,t}\rangle\Big] - \frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\langle y_{i,t} - \omega_t^*, e_{i,t}\rangle\Big] \\
& \overset{(i)}{\leq} \frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\|\bar{\omega}_t - y_{i,t}\|\|e_{i,t}\|\Big] - \frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\langle y_{i,t} - \omega_t^*, y_{i,t} - \Pi_{R_\omega}[y_{i,t}]\rangle\Big] \\
& \overset{(ii)}{\leq} \frac{2\beta_t \cdot L_b}{N}\sum_{i=1}^N \mathbb{E}\Big[\|\bar{\omega}_t - y_{i,t}\|\Big] - \frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\|y_{i,t} - \Pi_{R_\omega}[y_{i,t}]\|^2\Big] \\
& \overset{(iii)}{\leq} \frac{2\beta_t \cdot L_b}{N}\sum_{i=1}^N \mathbb{E}\Big[\|\bar{\omega}_t - \sum_{j=1}^N W_t^{ij}\omega_{j,t} - \beta_t \cdot (h_{i,t} + m_{i,t})\|\Big] - \frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\|e_{i,t}\|^2\Big] \\
& \overset{(iv)}{\leq} \frac{2\beta_t \cdot L_b}{N}\sum_{i=1}^N \mathbb{E}\Big[\|\bar{\omega}_t - \sum_{j=1}^N W_t^{ij}\omega_{j,t}\|\Big] + \frac{2\beta_t \cdot L_b}{N}\sum_{i=1}^N \mathbb{E}\Big[\|\beta_t(h_{i,t} + m_{i,t})\|\Big] - \frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\|e_{i,t}\|^2\Big] \\
& \overset{(v)}{\leq} \frac{2\beta_t \cdot L_b}{N}\sum_{i=1}^N \mathbb{E}\Big[\|\bar{\omega}_t - \omega_{i,t}\|\Big] + 2\beta_t^2 \cdot L_b^2 - \frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\|e_{i,t}\|^2\Big] \\
& \overset{(vi)}{\leq} 2\beta_t \cdot L_b \cdot \mathbb{E}\Big[\|Q \cdot \omega_t\|\Big] + 2\beta_t^2 \cdot L_b^2 - \frac{2}{N}\sum_{i=1}^N \mathbb{E}\Big[\|e_{i,t}\|^2\Big]
\end{aligned} \tag{77}$$

where (i) follows Cauchy-Schwarz inequality and the definition $e_{i,t} := y_{i,t} - \Pi_{R_\omega}[y_{i,t}]$; (ii) is from (42) and the projection property (16a) in Lemma 1; (iii) follows the definition of $y_{i,t}$; (iv) follows Cauchy-Schwarz inequality; (v) is from the inequality (41) and that the eigenvalues of W_t is bounded by 1; (vi) is due to the fact that $\|\bar{\omega}_t - \omega_{i,t}\| \leq \sqrt{\sum_{i=1}^N \|\bar{\omega}_t - \omega_{i,t}\|^2} = \|Q \cdot \omega_t\| \; \forall \, i \in \mathcal{N}$.

Then we plug in the above derived bounds on terms A-D (inequalities (71), (75), (76) and (77)) into (66), and obtain:

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{\omega}_{t+1} - \omega_{t+1}^*\|^2\right] \\
&\leq (1 - 2\beta_t \cdot \tilde{c}_{\min}) \cdot \mathbb{E}\left[\|\bar{\omega}_t - \omega_t^*\|^2\right] + 16L_\omega^2 \cdot \mathbb{E}\left[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\right] + 36L_\omega^2 \cdot \ell_p^2 \cdot \alpha_t^2 + 6L_\omega \cdot \ell_p \cdot \alpha_t \cdot \mathbb{E}\left[\|\bar{\omega}_t - \omega_t^*\|\right] \\
&\quad + 4L_\omega \cdot \mathbb{E}\left[\|Q \cdot \boldsymbol{\theta}_t^s\| \cdot \|\bar{\omega}_t - \omega_t^*\|\right] + 2\beta_t \cdot L_b \cdot \mathbb{E}\left[\|Q \cdot \boldsymbol{\omega}_t\|\right] + 10\beta_t^2 \cdot L_b^2 - \frac{2}{N}\sum_{i=1}^{N} \mathbb{E}\left[\|e_{i,t}\|^2\right].
\end{aligned} \tag{78}
$$

In (78), we have already obtain a bound of the distance between averaged parameter $\bar{\omega}_t$ and its optimal solution ω_t^*. Then we could utilize the inequality (78) to derive the convergence rate for the averaged parameters $\bar{\omega}_t$. Towards this end, we first rearrange the inequality (78), divide both sides by $2\beta_t \cdot \tilde{c}_{\min}$ and then sum it from $t = 0$ to $T - 1$, and obtain:

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{E}\left[\|\bar{\omega}_t - \omega_t^*\|^2\right] \\
&\leq \underbrace{\sum_{t=0}^{T-1} \frac{1}{2\beta_t \cdot \tilde{c}_{\min}} \left(\mathbb{E}\left[\|\omega_t^* - \bar{\omega}_t\|^2\right] - \mathbb{E}\left[\|\omega_{t+1}^* - \bar{\omega}_{t+1}\|^2\right]\right)}_{\text{term } \mathrm{I}_1} + \underbrace{\frac{8L_\omega^2}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \frac{1}{\beta_t} \cdot \mathbb{E}\left[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\right]}_{\text{term } \mathrm{I}_2} \\
&\quad + \underbrace{\sum_{t=0}^{T-1} \frac{1}{\tilde{c}_{\min}} \left(18L_w^2 \cdot \ell_p^2 \cdot \frac{\alpha_t^2}{\beta_t} + 5L_b^2 \cdot \beta_t\right)}_{\text{term } \mathrm{I}_3} + \underbrace{\frac{3L_\omega \cdot \ell_p}{\tilde{c}_{\min}} \sum_{t=0}^{T-1} \frac{\alpha_t}{\beta_t} \cdot \mathbb{E}\left[\|\bar{\omega}_t - \omega_t^*\|\right]}_{\text{term } \mathrm{I}_4} \\
&\quad + \underbrace{\frac{2L_\omega}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \frac{1}{\beta_t} \cdot \mathbb{E}\left[\|Q \cdot \boldsymbol{\theta}_t^s\| \cdot \|\bar{\omega}_t - \omega_t^*\|\right]}_{\text{term } \mathrm{I}_5} + \underbrace{\frac{L_b}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \mathbb{E}\left[\|Q \cdot \boldsymbol{\omega}_t\|\right]}_{\text{term } \mathrm{I}_6}.
\end{aligned} \tag{79}
$$

Recall the fixed stepsizes $\alpha_t := \frac{\alpha}{T^{\sigma_1}}$ and $\beta_t := \frac{\beta}{T^{\sigma_2}}$ $\forall t$. We can bound each term in (79) as below. First, term I_1 can be bounded as:

$$
\begin{aligned}
I_1 &:= \sum_{t=0}^{T-1} \frac{T^{\sigma_2}}{2\beta \cdot \tilde{c}_{\min}} \cdot \left(\mathbb{E}\left[\|\omega_t^* - \bar{\omega}_t\|^2\right] - \mathbb{E}\left[\|\omega_{t+1}^* - \bar{\omega}_{t+1}\|^2\right]\right) \\
&= \frac{T^{\sigma_2}}{2\beta \cdot \tilde{c}_{\min}} \cdot \left(\mathbb{E}\left[\|\omega_0^* - \bar{\omega}_0\|^2\right] - \mathbb{E}\left[\|\omega_T^* - \bar{\omega}_T\|^2\right]\right) \\
&\overset{(a)}{\leq} \frac{T^{\sigma_2} \cdot 4R_\omega^2}{2\beta \cdot \tilde{c}_{\min}} = \mathcal{O}(T^{\sigma_2})
\end{aligned} \tag{80}
$$

where the inequality (a) follows $\mathbb{E}\big[\|\omega_t^* - \bar{\omega}_t\|^2\big] \leq 2\mathbb{E}\big[\|\omega_t^*\|^2 + \|\bar{\omega}_t\|^2\big] = 4R_\omega^2$, since both ω_t^* and $\bar{\omega}_t$ are in a fixed region with radius R_ω.

Second, term I_2 in (79) can be bounded as:

$$
I_2 := \frac{8L_\omega^2}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \frac{1}{\beta_t} \cdot \mathbb{E}\left[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\right] \overset{(i)}{=} \frac{8L_\omega^2 \cdot T^{\sigma_2}}{\tilde{c}_{\min} \cdot \beta} \cdot \sum_{t=0}^{T-1} \mathbb{E}\left[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\right] \overset{(64)}{=} \mathcal{O}(T^{1-2\sigma_1+\sigma_2}) \tag{81}
$$

where (i) follows the definition $\beta_t = \frac{\beta}{T^{\sigma_2}}$, $\forall t$.

Third, term I_3 in (79) can be bounded as:

$$\begin{aligned}
I_3 &:= \sum_{t=0}^{T-1} \frac{1}{\tilde{c}_{\min}} \cdot \left(18L_w^2 \cdot \ell_p^2 \cdot \frac{\alpha_t^2}{\beta_t} + 5L_b^2 \cdot \beta_t\right) \\
&\overset{(a)}{=} \frac{18L_w^2 \cdot \ell_p^2}{\tilde{c}_{\min}} \cdot \frac{\alpha^2}{\beta} \cdot T \cdot T^{\sigma_2 - 2\sigma_1} + \frac{5L_b^2 \cdot \beta}{\tilde{c}_{\min}} \cdot T \cdot T^{-\sigma_2} \\
&= \frac{18L_w^2 \cdot \ell_p^2}{\tilde{c}_{\min}} \cdot \frac{\alpha^2}{\beta} \cdot T^{1+\sigma_2 - 2\sigma_1} + \frac{5L_b^2 \cdot \beta}{\tilde{c}_{\min}} \cdot T^{1-\sigma_2} \\
&= \mathcal{O}(T^{1-\sigma_2}) + \mathcal{O}\left(T^{1+\sigma_2-2\sigma_1}\right)
\end{aligned} \tag{82}$$

where (a) follows the definitions $\alpha_t := \frac{\alpha}{T^{\sigma_1}}$ and $\beta_t := \frac{\beta}{T^{\sigma_2}}$ for any iteration t.

Next, we can bound the term I_4 in (79) as below:

$$\begin{aligned}
I_4 &= \frac{3L_\omega \cdot \ell_p}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \frac{\alpha_t}{\beta_t} \cdot \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|\Big] \\
&\overset{(i)}{\le} \frac{3L_\omega \cdot \ell_p}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \left(\sqrt{\frac{\alpha_t^2}{\beta_t^2}} \cdot \sqrt{\mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big]}\right) \\
&\overset{(ii)}{\le} \frac{3L_\omega \cdot \ell_p}{\tilde{c}_{\min}} \cdot \sqrt{\sum_{t=0}^{T-1} \frac{\alpha_t^2}{\beta_t^2}} \cdot \sqrt{\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big]} \\
&\overset{(iii)}{=} \frac{3L_\omega \cdot \ell_p \cdot \alpha_0^2}{\tilde{c}_{\min} \cdot \beta_0^2} \cdot \sqrt{T \cdot T^{2\sigma_2 - 2\sigma_1}} \cdot \sqrt{\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big]} \\
&= \frac{3L_\omega \cdot \ell_p \cdot \alpha_0^2}{\tilde{c}_{\min} \cdot \beta_0^2} \cdot \sqrt{T^{1+2\sigma_2 - 2\sigma_1}} \cdot \sqrt{\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big]} \\
&\overset{(iv)}{=} \sqrt{C_4 \cdot T^{1+2\sigma_2 - 2\sigma_1}} \cdot \sqrt{\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big]}
\end{aligned} \tag{83}$$

where (i) is by Jensen's inequality; (ii) follows Cauchy-Schwarz inequality; (iii) follows the definition of stepsizes α_t and β_t; in equality (iv) we define $C_4 := \left(\frac{3L_\omega \cdot \ell_p \cdot \alpha_0^2}{\tilde{c}_{\min} \cdot \beta_0^2}\right)^2$.

Next, the term I_5 in (79) can be bounded as below:

$$\begin{aligned}
I_5 &:= \frac{2L_\omega}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \frac{1}{\beta_t} \cdot \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\| \cdot \|\bar{\omega}_t - \omega_t^*\|\Big] \\
&\overset{(i)}{\le} \frac{2L_\omega}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \sqrt{\frac{1}{\beta_t^2} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\Big] \cdot \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big]} \\
&\overset{(ii)}{\le} \frac{2L_\omega}{\tilde{c}_{\min}} \cdot \sqrt{\sum_{t=0}^{T-1} \frac{1}{\beta_t^2} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\Big]} \cdot \sqrt{\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big]} \\
&\overset{(iii)}{=} \sqrt{\frac{C_5 \cdot T^{2\sigma_2}}{\beta^2} \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\Big]} \cdot \sqrt{\sum_{t=0}^{T-1} \mathbb{E}[\|\bar{\omega}_t - \omega_t^*\|^2]}
\end{aligned} \tag{84}$$

where (i) and (ii) follows Cauchy-Schwarz inequality; (iii) follows the stepsize $\beta_t := \frac{\beta}{T^{\sigma_2}}$ and we define the constant $C_5 := \left(\frac{2L_\omega}{\tilde{c}_{\min}}\right)^2$.

Finally, we can bound the last term I_6 in (79) as below:

$$I_6 := \frac{L_b}{\tilde{c}_{\min}} \cdot \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\omega}_t\|\Big] \overset{(46)}{=} \mathcal{O}\Big(T^{1-\sigma_2}\Big). \tag{85}$$

Then we can revisit (79) to obtain the exact convergence rate. Let us rearrange (79) as below:

$$\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big] \leq \underbrace{(I_1 + I_2 + I_3 + I_6)}_{:=\text{term } \mathrm{K}_1} + (I_4 + I_5). \tag{86}$$

For the terms $I_4 + I_5$, we utilize (83) and (84) to obtain that

$$\begin{aligned} I_4 + I_5 &\leq \bigg(\sqrt{C_4 \cdot T^{1+2\sigma_2-2\sigma_1}} + \sqrt{C_5 \cdot \frac{T^{2\sigma_2}}{\beta^2} \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\Big]}\bigg) \cdot \sqrt{\sum_{t=0}^{T-1} \mathbb{E}[\|\bar{\omega}_t - \omega_t^*\|^2]} \\ &\overset{(a)}{\leq} \sqrt{\underbrace{2C_4 \cdot T^{1+2\sigma_2-2\sigma_1} + \frac{2C_5 \cdot T^{2\sigma_2}}{\beta^2} \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\Big]}_{:=\text{term } \mathrm{K}_2}} \cdot \sqrt{\underbrace{\sum_{t=0}^{T-1} \mathbb{E}[\|\bar{\omega}_t - \omega_t^*\|^2]}_{:=\text{term } \mathrm{K}_3}} \end{aligned} \tag{87}$$

where (a) follows Cauchy-Schwarz inequality.

With terms K_1, K_2 and K_3 defined as above, we can plug (87) into (86), and obtain:

$$K_3 \leq K_1 + \sqrt{K_2 \cdot K_3} \implies \big(\sqrt{K_3} - \frac{1}{2}\sqrt{K_2}\big)^2 \leq K_1 + \frac{1}{4}K_2 \implies \sqrt{K_3} \leq \frac{1}{2}\sqrt{K_2} + \sqrt{K_1 + \frac{1}{4}K_2}$$

$$K_3 \leq \bigg(\frac{1}{2}\sqrt{K_2} + \sqrt{K_1 + \frac{1}{4}K_2}\bigg)^2 \overset{(a)}{\leq} \frac{1}{2}K_2 + 2K_1 + \frac{1}{2}K_2 = 2K_1 + K_2 \tag{88}$$

where (a) is due to Cauchy-Schwarz inequality.

Combining the inequalities (80), (81), (82) and (85), the convergence rate of term K_1 in (88) could be expressed as below:

$$\begin{aligned} K_1 &:= I_1 + I_2 + I_3 + I_6 \\ &= \mathcal{O}\big(T^{\sigma_2}\big) + \mathcal{O}\big(T^{1-2\sigma_1+\sigma_2}\big) + \mathcal{O}\big(T^{1-\sigma_2}\big) + \mathcal{O}\big(T^{1+\sigma_2-2\sigma_1}\big) + \mathcal{O}\big(T^{1-\sigma_2}\big). \end{aligned} \tag{89}$$

Moreover, the convergence rate of term K_2 could be bounded as below:

$$\begin{aligned} K_2 &:= 2C_4 \cdot T^{1+2\sigma_2-2\sigma_1} + \frac{2C_5 \cdot T^{2\sigma_2}}{\beta^2} \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|^2\Big] \\ &\overset{(a)}{=} \mathcal{O}\big(T^{1+2\sigma_2-2\sigma_1}\big) + \mathcal{O}\big(T^{-1+2\sigma_2}\big) \end{aligned} \tag{90}$$

where (a) follows the inequality (64).

Plugging (87) - (90) into (86), dividing T on both sides, we can obtain the convergence rate of $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\bar{\omega}_t - \omega_t^*\|^2\right]$ as below:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\bar{\omega}_t - \omega_t^*\|^2\Big] = \mathcal{O}(T^{-1+\sigma_2}) + \mathcal{O}(T^{-\sigma_2}) + \mathcal{O}\left(T^{\sigma_2-2\sigma_1}\right) + \mathcal{O}(T^{-2\sigma_1+2\sigma_2}) + \mathcal{O}(T^{-2+2\sigma_2}). \tag{91}$$

When the fixed stepsizes ζ_t and β_t are in the same order ($\zeta_t := \frac{\zeta}{T^{\sigma_2}}$ and $\beta_t := \frac{\beta}{T^{\sigma_2}}$ for any iteration t), the analysis of parameters $\boldsymbol{\omega}_t$ can directly extend to establish the convergence results of parameters $\boldsymbol{\lambda}_t$. Then, it holds that $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\bar{\lambda}_t - \lambda_t^*\|^2\big]$ has the same convergence rate as $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\bar{\omega}_t - \omega_t^*\|^2\big]$.

Finally, combining the convergence results for consensus errors in (21c), we obtain the convergence of approximation parameters as below:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left(\mathbb{E}\left[\|\omega_{i,t}-\omega_t^*\|^2\right]+\mathbb{E}\left[\|\lambda_{i,t}-\lambda_t^*\|^2\right]\right)\\
&\overset{(a)}{\leq}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left(2\mathbb{E}\left[\|\omega_{i,t}-\bar{\omega}_t\|^2\right]+2\mathbb{E}\left[\|\bar{\omega}_t-\omega_t^*\|^2\right]+2\mathbb{E}\left[\|\lambda_{i,t}-\bar{\lambda}_t\|^2\right]+2\mathbb{E}\left[\|\bar{\lambda}_t-\lambda_t^*\|^2\right]\right)\\
&=\frac{2}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left(\mathbb{E}\left[\|\bar{\omega}_t-\omega_t^*\|^2\right]+\mathbb{E}\left[\|\bar{\lambda}_t-\lambda_t^*\|^2\right]\right)+\frac{2}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|Q\cdot\boldsymbol{\omega}_t\|^2\right]+\mathbb{E}\left[\|Q\cdot\boldsymbol{\lambda}_t\|^2\right]\right)\\
&\overset{(b)}{=}\mathcal{O}(T^{-1+\sigma_2})+\mathcal{O}(T^{-\sigma_2})+\mathcal{O}\left(T^{\sigma_2-2\sigma_1}\right)+\mathcal{O}(T^{-2\sigma_1+2\sigma_2})+\mathcal{O}(T^{-2+2\sigma_2})+\mathcal{O}(T^{-2\sigma_2}).
\end{aligned}
\tag{92}
$$

where (a) follows Cauchy-Schwarz inequality; (b) follows (54) and (91). □

H Proof of Proposition 2

Proof. In this proposition, we will analyze the convergence of actor in algorithm CAC.

With linear approximations, recall that in (11) we have defined:

$$
\widehat{\delta}_{i,t}=\varphi(s_t,\boldsymbol{a}_t)^T\lambda_{i,t}+\gamma\cdot\phi(s_{t+1})^T\omega_{i,t}-\phi(s_t)^T\omega_{i,t}.
$$

Due to the facts that feature vectors are assumed to be bounded in Assumption 4 and the approximation parameters $\lambda_{i,t}$ and $\omega_{i,t}$ are restricted in fixed regions, we have

$$
\|\widehat{\delta}_{i,t}\|\leq R_\lambda+(1+\gamma)R_\omega,\ \forall\, i\in\mathcal{N}
\tag{93}
$$

For simplicity, let us define $C_\delta:=R_\lambda+(1+\gamma)R_\omega$.

Recall that for each agent i, we have denoted its local policy parameters as $\theta_i:=\{\theta_i^s,\theta_i^p\}$ where θ_i^s is the shared policy parameter and θ_i^p is the personalized policy parameter. Moreover, for each policy optimization step, the update in shared policy parameters θ_i^s is given by:

$$
\theta_{i,t+1}^s:=\sum_{j=1}^{N}W_{ij}\cdot\theta_{j,t}^s+\alpha_t\cdot\widehat{\delta}_{i,t}\cdot\psi_{\theta_i^s}(s_t,a_{i,t})
\tag{94}
$$

where we have defined the score function $\psi_{\theta_i^s}(s_t,a_{i,t}):=\nabla_{\theta^s}\log\pi(a_{i,t}\mid s_t,\theta_{i,t})$. Therefore, the update of the *average* of the shared policy parameter $\bar{\theta}_t^s:=\frac{1}{N}\sum_{i=1}^{N}\theta_{i,t}^s$ is given below:

$$
\bar{\theta}_{t+1}^s:=\bar{\theta}_t^s+\frac{\alpha_t}{N}\sum_{i=1}^{N}\widehat{\delta}_{i,t}\cdot\psi_{\theta_i^s}(s_t,a_{i,t}).
\tag{95}
$$

We further define $\bar{\boldsymbol{\theta}}_t:=\cup_{i=1}^{N}\{\bar{\theta}_{i,t}\}$ and $\bar{\theta}_{i,t}:=\{\bar{\theta}_t^s,\theta_{i,t}^p\}$. Also recall we have defined $\boldsymbol{\theta}_t:=\cup_{i=1}^{N}\{\theta_{i,t}\}$ and $\theta_{i,t}:=\{\theta_{i,t}^s,\theta_{i,t}^p\}$. In the analysis, we start from analyzing objective value $J(\bar{\boldsymbol{\theta}}_t)$ at each iteration t. According to L_J-Lipschitz of policy gradient in Lemma 2, it holds that

$$\begin{aligned}
&J(\bar{\boldsymbol{\theta}}_{t+1})\\
&\overset{(17)}{\geq} J(\bar{\boldsymbol{\theta}}_t) + \langle \nabla J(\bar{\boldsymbol{\theta}}_t), \bar{\boldsymbol{\theta}}_{t+1} - \bar{\boldsymbol{\theta}}_t \rangle - \frac{L_J}{2} \|\bar{\boldsymbol{\theta}}_{t+1} - \bar{\boldsymbol{\theta}}_t\|^2\\
&= J(\bar{\boldsymbol{\theta}}_t) + \langle \nabla J(\bar{\boldsymbol{\theta}}_t) - \nabla J(\boldsymbol{\theta}_t), \bar{\boldsymbol{\theta}}_{t+1} - \bar{\boldsymbol{\theta}}_t \rangle + \langle \nabla J(\boldsymbol{\theta}_t), \bar{\boldsymbol{\theta}}_{t+1} - \bar{\boldsymbol{\theta}}_t \rangle - \frac{L_J}{2} \|\bar{\boldsymbol{\theta}}_{t+1} - \bar{\boldsymbol{\theta}}_t\|^2\\
&= J(\bar{\boldsymbol{\theta}}_t) + \sum_{i=1}^N \langle \nabla_{\theta_i} J(\bar{\boldsymbol{\theta}}_t) - \nabla_{\theta_i} J(\boldsymbol{\theta}_t), \bar{\theta}_{i,t+1} - \bar{\theta}_{i,t} \rangle + \sum_{i=1}^N \langle \nabla_{\theta_i} J(\boldsymbol{\theta}_t), \bar{\theta}_{i,t+1} - \bar{\theta}_{i,t} \rangle - \frac{L_J}{2} \sum_{i=1}^N \|\bar{\theta}_{i,t+1} - \bar{\theta}_{i,t}\|^2\\
&\geq J(\bar{\boldsymbol{\theta}}_t) - \sum_{i=1}^N \|\nabla_{\theta_i} J(\bar{\boldsymbol{\theta}}_t) - \nabla_{\theta_i} J(\boldsymbol{\theta}_t)\| \cdot \|\bar{\theta}_{i,t+1} - \bar{\theta}_{i,t}\| + \sum_{i=1}^N \langle \nabla_{\theta_i} J(\boldsymbol{\theta}_t), \bar{\theta}_{i,t+1} - \bar{\theta}_{i,t} \rangle - \frac{L_J}{2} \sum_{i=1}^N \|\bar{\theta}_{i,t+1} - \bar{\theta}_{i,t}\|^2\\
&\overset{(i)}{\geq} J(\bar{\boldsymbol{\theta}}_t) - N \cdot \alpha_t \cdot \ell_p \cdot L_J \cdot \|\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\| + \sum_{i=1}^N \langle \nabla_{\theta_i} J(\boldsymbol{\theta}_t), \bar{\theta}_{i,t+1} - \bar{\theta}_{i,t} \rangle - \frac{\alpha_t^2 \cdot \ell_p^2 \cdot L_J \cdot N}{2}\\
&\overset{(ii)}{=} J(\bar{\boldsymbol{\theta}}_t) - N \cdot \alpha_t \cdot \ell_p \cdot L_J \cdot \|\boldsymbol{\theta}_t^s - \mathbf{1} \cdot \bar{\theta}_t^{s^T}\| + \sum_{i=1}^N \langle \nabla_{\theta_i} J(\boldsymbol{\theta}_t), \bar{\theta}_{i,t+1} - \bar{\theta}_{i,t} \rangle - \frac{\alpha_t^2 \cdot \ell_p^2 \cdot L_J \cdot N}{2}
\end{aligned} \tag{96}$$

where (i) follows (17) and $\|\bar{\theta}_{i,t+1} - \bar{\theta}_{i,t}\| \leq \alpha_t \cdot \ell_p$ where the constant ℓ_p is defined in (59); (ii) follows the fact that $\|\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\| = \|\boldsymbol{\theta}_t^s - \mathbf{1} \cdot \bar{\theta}_t^{s^T}\|$ since we have defined $\bar{\boldsymbol{\theta}}_t := \cup_{i=1}^N \{\bar{\theta}_{i,t}\} = \cup_{i=1}^N \{\bar{\theta}_t^s, \theta_{i,t}^p\}$. Therefore, it follows that

$$\begin{aligned}
&J(\bar{\boldsymbol{\theta}}_{t+1})\\
&\overset{(96)}{\geq} J(\bar{\boldsymbol{\theta}}_t) - N \cdot \alpha_t \cdot \ell_p \cdot L_J \cdot \|\boldsymbol{\theta}_t^s - \mathbf{1} \cdot \bar{\theta}_t^{s^T}\| + \sum_{i=1}^N \langle \nabla_{\theta_i} J(\boldsymbol{\theta}_t), \bar{\theta}_{i,t+1} - \bar{\theta}_{i,t} \rangle - \frac{\alpha_t^2 \cdot \ell_p^2 \cdot L_J \cdot N}{2}\\
&\overset{(a)}{=} J(\bar{\boldsymbol{\theta}}_t) - N \cdot \alpha_t \cdot \ell_p \cdot L_J \cdot \|Q \cdot \boldsymbol{\theta}_t^s\| + \alpha_t \sum_{i=1}^N \left\langle \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \frac{1}{N} \sum_{j=1}^N \widehat{\delta}_{j,t} \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \right\rangle\\
&\quad + \alpha_t \sum_{i=1}^N \left\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \widehat{\delta}_{i,t} \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t}) \right\rangle - \frac{\alpha_t^2 \cdot \ell_p^2 \cdot L_J \cdot N}{2}\\
&= J(\bar{\boldsymbol{\theta}}_t) + \frac{\alpha_t}{N} \sum_{i,j=1}^N \left\langle \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \widehat{\delta}_{j,t} \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \right\rangle + \alpha_t \sum_{i=1}^N \left\langle \nabla_{\theta_i^p} J(\theta_t), \widehat{\delta}_{i,t} \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t}) \right\rangle\\
&\quad - N \cdot \alpha_t \cdot \ell_p \cdot L_J \cdot \|Q \cdot \boldsymbol{\theta}_t^s\| - \frac{N \cdot L_J \cdot \alpha_t^2 \cdot \ell_p^2}{2}
\end{aligned} \tag{97}$$

where (a) follows the fact that the update of $\bar{\theta}_{i,t} := \{\bar{\theta}_t^s, \theta_{i,t}^p\}$ could be decomposed as below:

$$\begin{aligned}
\bar{\theta}_{t+1}^s - \bar{\theta}_t^s &= \frac{\alpha_t}{N} \sum_{j=1}^N \widehat{\delta}_{j,t} \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}),\\
\theta_{i,t+1}^p - \theta_{i,t}^p &= \alpha_t \cdot \widehat{\delta}_{i,t} \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t}).
\end{aligned}$$

Taking expectation on both sides of inequality (97), we have:

$$\begin{aligned}
\mathbb{E}[J(\bar{\boldsymbol{\theta}}_{t+1})] \geq \mathbb{E}[J(\bar{\boldsymbol{\theta}}_t)] &+ I_{1,1} + I_{1,2} + I_{2,1} + I_{2,2}\\
&- N \cdot \alpha_t \cdot \ell_p \cdot L_J \cdot \mathbb{E}\big[\|Q \cdot \boldsymbol{\theta}_t^s\|\big] - \frac{N \cdot L_J \cdot \alpha_t^2 \cdot \ell_p^2}{2}
\end{aligned} \tag{99}$$

where we have defined:

$$
\begin{aligned}
I_{1,1} &:= \frac{\alpha_t}{N} \sum_{i,j=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \Big(\widehat{\delta}_{j,t} - \widehat{\delta}_t^*\Big) \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \Big\rangle\Big] \\
I_{1,2} &:= \frac{\alpha_t}{N} \sum_{i,j=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \widehat{\delta}_t^* \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \Big\rangle\Big] \\
I_{2,1} &:= \alpha_t \sum_{i=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \Big(\widehat{\delta}_{i,t} - \widehat{\delta}_t^*\Big) \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t}) \Big\rangle\Big] \\
I_{2,2} &:= \alpha_t \sum_{i=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \widehat{\delta}_t^* \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t}) \Big\rangle\Big] \\
\widehat{\delta}_t^* &:= \varphi(s_t, \boldsymbol{a}_t)^T \lambda_t^* + \gamma \cdot \phi(s_{t+1})^T \omega_t^* - \phi(s_t)^T \omega_t^*,
\end{aligned}
\tag{100}
$$

and these terms satisfy the following relations:

$$
\begin{aligned}
I_{1,1} + I_{1,2} &= \frac{\alpha_t}{N} \sum_{i,j=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \widehat{\delta}_{j,t} \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \Big\rangle\Big] \\
I_{2,1} + I_{2,2} &= \alpha_t \sum_{i=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \widehat{\delta}_{i,t} \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t}) \Big\rangle\Big].
\end{aligned}
$$

Below we analyze each term on the rhs of (99). Towards this end, let us define $\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} := \frac{1}{N}\sum_{i=1}^{N} \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t)$. We have the following:

$$
\begin{aligned}
I_{1,1} &= \frac{\alpha_t}{N} \sum_{i,j=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \Big(\widehat{\delta}_{j,t} - \widehat{\delta}_t^*\Big) \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \Big\rangle\Big] \\
&= -\alpha_t \cdot \sum_{j=1}^{N} \cdot \mathbb{E}\Big[\Big\langle \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \varphi(s_t, \boldsymbol{a}_t)^T (\lambda_t^* - \lambda_{j,t}) \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \Big\rangle\Big] \\
&\quad - \alpha_t \cdot \sum_{j=1}^{N} \cdot \mathbb{E}\Big[\Big\langle \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), (\gamma \cdot \phi(s_{t+1}) - \phi(s_t))^T (\omega_t^* - \omega_{j,t}) \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \Big\rangle\Big] \\
&\overset{(i)}{\geq} -\alpha_t \cdot C_\psi \cdot \sum_{j=1}^{N} \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\| \cdot \|\lambda_t^* - \lambda_{i,t}\|\Big] - (1+\gamma) \cdot \alpha_t \cdot C_\psi \cdot \sum_{j=1}^{N} \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\| \cdot \|\omega_t^* - \omega_{i,t}\|\Big] \\
&\geq -2\alpha_t \cdot C_\psi \cdot \sum_{j=1}^{N} \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\| \cdot \Big(\|\omega_t^* - \omega_{i,t}\| + \|\lambda_t^* - \lambda_{i,t}\|\Big)\Big] \\
&= -2\alpha_t \cdot C_\psi \cdot \sum_{j=1}^{N} \sqrt{\Big(\mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\| \cdot \Big(\|\omega_t^* - \omega_{i,t}\| + \|\lambda_t^* - \lambda_{i,t}\|\Big)\Big]\Big)^2} \\
&\overset{(ii)}{\geq} -2\alpha_t \cdot C_\psi \cdot \sum_{j=1}^{N} \sqrt{\mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big] \cdot \mathbb{E}\Big[\Big(\|\omega_t^* - \omega_{i,t}\| + \|\lambda_t^* - \lambda_{i,t}\|\Big)^2\Big]} \\
&\overset{(iii)}{\geq} -4\alpha_t \cdot C_\psi \cdot \sqrt{N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big]} \cdot \sqrt{\sum_{j=1}^{N} \Big(\mathbb{E}\Big[\|\omega_t^* - \omega_{i,t}\|^2\Big] + \mathbb{E}\Big[\|\lambda_t^* - \lambda_{i,t}\|^2\Big]\Big)}
\end{aligned}
\tag{101}
$$

where (i) is due to the facts that all feature vectors are bounded (cf. Assumption 4), and the score functions are bounded as $\|\psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\| \leq C_\psi$ (cf. (14a)); (ii) and (iii) follow from the Cauchy-Schwarz inequality.

Here, we define the TD error

$$\delta_t := Q_{\pi_{\theta_t}}(s_t, \boldsymbol{a}_t) - V_{\pi_{\boldsymbol{\theta}_t}}(s_t) = \bar{r}_t + \gamma V_{\pi_{\boldsymbol{\theta}_t}}(s_{t+1}) - V_{\pi_{\boldsymbol{\theta}_t}}(s_t) \tag{102}$$

and $\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} := \frac{1}{N}\sum_{i=1}^N \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t)$, then the term $I_{1,2}$ can be decomposed as below:

$$\begin{aligned}
I_{1,2} &:= \frac{\alpha_t}{N} \cdot \sum_{i,j=1}^N \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \widehat{\delta}_t^* \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\Big\rangle\Big] \\
&= \alpha_t \cdot \sum_{j=1}^N \mathbb{E}\Big[\Big\langle \frac{1}{N}\sum_{i=1}^N \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \widehat{\delta}_t^* \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\Big\rangle\Big] \\
&= \alpha_t \cdot (1-\gamma) \cdot N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big] + \alpha_t \cdot \sum_{j=1}^N \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, (\widehat{\delta}_t^* - \delta_t) \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\Big\rangle\Big] \\
&\quad + \alpha_t \cdot \sum_{j=1}^N \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \delta_t \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\Big\rangle\Big].
\end{aligned} \tag{103}$$

Then we are able to bound each term above:

$$\begin{aligned}
&\alpha_t \cdot \sum_{j=1}^N \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, (\widehat{\delta}_t^* - \delta_t) \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\Big\rangle\Big] \\
&\geq -\alpha_t \cdot \sum_{j=1}^N \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\| \cdot \|\psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\| \cdot \big|\widehat{\delta}_t^* - \delta_t\big|\Big] \\
&\overset{(i)}{\geq} -\alpha_t \cdot L_v \cdot C_\psi \cdot N \cdot \mathbb{E}\Big[\Big|\Big(\varphi(s_t, \boldsymbol{a}_t)^T \lambda_t^* - \bar{r}_t\Big) + \gamma\Big(\phi(s_{t+1})^T \omega_t^* - V_{\pi_{\boldsymbol{\theta}_t}}(s_{t+1})\Big) - \Big(\phi(s_t)^T \omega_t^* - V_{\pi_{\boldsymbol{\theta}_t}}(s_t)\Big)\Big|\Big] \\
&\overset{(ii)}{\geq} -\alpha_t \cdot L_v \cdot C_\psi \cdot N \cdot \Big(\mathbb{E}\Big[|\varphi(s_t, a_t)^T \lambda_t^* - \bar{r}_t|\Big] + \gamma \mathbb{E}\Big[|\phi(s_{t+1})^T \omega_t^* - V_{\pi_{\boldsymbol{\theta}_t}}(s_{t+1})|\Big] + \mathbb{E}\Big[|V_{\pi_{\boldsymbol{\theta}_t}}(s_t) - \phi(s_t)^T \omega_t^*|\Big]\Big) \\
&\overset{(iii)}{\geq} -\alpha_t \cdot L_v \cdot C_\psi \cdot N \cdot \Big(\sqrt{\mathbb{E}\Big[\big(\varphi(s_t, \boldsymbol{a}_t)^T \lambda_t^* - \bar{r}_t\big)^2\Big]} + \gamma\sqrt{\mathbb{E}\Big[\big(\phi(s_{t+1})^T \omega_t^* - V_{\pi_{\boldsymbol{\theta}_t}}(s_{t+1})\big)^2\Big]} \\
&\quad + \sqrt{\mathbb{E}\Big[\big(V_{\pi_{\boldsymbol{\theta}_t}}(s_t) - \phi(s_t)^T \omega_t^*\big)^2\Big]}\Big) \\
&\overset{(iv)}{\geq} -2N \cdot \alpha_t \cdot L_v \cdot C_\psi \cdot \epsilon_{app}
\end{aligned} \tag{104}$$

where (i) follows (14a), (18), definition of $\widehat{\delta}_t^*$ in (100) and definition of δ_t in (102); (ii) follows the triangle inequality; (iii) follows Jensen's inequality; (iv) follows the definition of approximation error in (28).

Recall the definition $Q_{\pi_{\boldsymbol{\theta}}}(s, \boldsymbol{a}) := \mathbb{E}\big[\sum_{t=0}^\infty \gamma^t r(s_t, \boldsymbol{a}_t) | s_0 = s, \boldsymbol{a}_0 = \boldsymbol{a}\big]$. According to Agarwal et al. (2019), policy gradient could be expressed as below

$$\nabla J(\boldsymbol{\theta}) := \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{\boldsymbol{\theta}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\Big[\big(Q_{\pi_{\boldsymbol{\theta}}}(s, \boldsymbol{a}) - V_{\pi_{\boldsymbol{\theta}}}(s)\big) \cdot \psi_{\boldsymbol{\theta}}(s, \boldsymbol{a})\Big] \tag{105}$$

where $d_{\boldsymbol{\theta}}(\cdot)$ denotes the discounted visitation measure $d_{\boldsymbol{\theta}}(s) := \mathbb{E}_{s_0\sim\eta}\big[(1-\gamma)\sum_{t=0}^\infty \gamma^t \mathcal{P}^{\pi_{\boldsymbol{\theta}}}(s_t = s|s_0)\big]$. According to the policy gradient as shown in (105), it holds that:

$$\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} := \frac{1}{1-\gamma} \mathbb{E}_{s_t\sim d_{\boldsymbol{\theta}_t}(\cdot), \boldsymbol{a}_t\sim\pi_{\boldsymbol{\theta}_t}(\cdot|s_t)}\Big[\big(Q_{\pi_{\boldsymbol{\theta}_t}}(s_t, \boldsymbol{a}_t) - V_{\pi_{\boldsymbol{\theta}_t}}(s_t)\big) \cdot \big(\frac{1}{N}\sum_{j=1}^N \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\big)\Big]. \tag{106}$$

Therefore, the third term in (103) could be bounded as

$$
\begin{aligned}
&\alpha_t \cdot \sum_{j=1}^{N} \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \delta_t \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} \Big\rangle\Big] \\
&= \alpha_t \cdot N \cdot \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \delta_t \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\big) - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} \Big\rangle\Big] \\
&\overset{(i)}{=} \alpha_t \cdot N \cdot \mathbb{E}\Big[\mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \delta_t \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\big) - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} \Big\rangle \Big| \boldsymbol{\theta}_t\Big]\Big] \\
&= \alpha_t \cdot N \cdot \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \Big(\mathbb{E}_{s_t \sim \mu_{\boldsymbol{\theta}_t}(\cdot), \boldsymbol{a}_t \sim \pi_{\boldsymbol{\theta}_t}(\cdot|s_t)}\Big[\big(Q_{\pi_{\boldsymbol{\theta}_t}}(s_t, \boldsymbol{a}_t) - V_{\pi_{\boldsymbol{\theta}_t}}(s_t)\big) \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\big) \Big| \boldsymbol{\theta}_t\Big] \\
&\quad - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\Big)\Big\rangle\Big] \\
&\overset{(ii)}{\geq} -\alpha_t \cdot L_v \cdot N \cdot \Big\| \mathbb{E}_{s_t \sim \mu_{\boldsymbol{\theta}_t}(\cdot), \boldsymbol{a}_t \sim \pi_{\boldsymbol{\theta}_t}(\cdot|s_t)}\Big[\big(Q_{\pi_{\boldsymbol{\theta}_t}}(s_t, \boldsymbol{a}_t) - V_{\pi_{\boldsymbol{\theta}_t}}(s_t)\big) \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\big)\Big] \\
&\quad - \mathbb{E}_{s_t \sim d_{\boldsymbol{\theta}_t}(\cdot), \boldsymbol{a}_t \sim \pi_{\boldsymbol{\theta}_t}(\cdot|s_t)}\Big[\big(Q_{\pi_{\boldsymbol{\theta}_t}}(s_t, \boldsymbol{a}_t) - V_{\pi_{\boldsymbol{\theta}_t}}(s_t)\big) \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\big)\Big] \Big\| \\
&\overset{(iii)}{\geq} -\alpha_t \cdot L_v \cdot N \cdot \sup_{s, \boldsymbol{a}} \Big\| \big(Q_{\pi_{\boldsymbol{\theta}_t}}(s_t, \boldsymbol{a}_t) - V_{\pi_{\boldsymbol{\theta}_t}}(s_t)\big) \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\big) \Big\| \cdot d_{TV}(\mu_{\boldsymbol{\theta}_t}, d_{\boldsymbol{\theta}_t}) \\
&\overset{(iv)}{\geq} -4\alpha_t \cdot N \cdot L_v \cdot C_\psi \cdot R_{\max} \cdot \left(\log_\tau \kappa^{-1} + \frac{1}{1-\tau}\right) = -\alpha_t \cdot N \cdot \epsilon_{sp} \qquad (107)
\end{aligned}
$$

where (i) follows the tower rule; (ii) follows (18) and (106); (iii) is due to distribution mismatch between the policy gradient in (106) and its estimator; the sampling error ϵ_{sp} is defined in (29) and the inequality (iv) is due to the facts that $Q_{\pi_{\boldsymbol{\theta}_t}}(s_t, \boldsymbol{a}_t) \leq \frac{R_{\max}}{1-\gamma}$, $V_{\pi_{\boldsymbol{\theta}_t}}(s_t) \leq \frac{R_{\max}}{1-\gamma}$, $\|\psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t})\| \leq C_\psi$ and the distribution mismatch inequality (19) in Lemma 4.

By plugging the inequalities (104) - (107) into (103), we obtain the following:

$$
\begin{aligned}
I_{1,2} &:= \frac{\alpha_t}{N}\sum_{i,j=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^s} J(\boldsymbol{\theta}_t), \widehat{\delta}_t^* \cdot \psi_{\theta_j^s}(s_t, a_{j,t}; \theta_{j,t}) \Big\rangle\Big] \\
&\geq -2N \cdot \alpha_t \cdot L_v \cdot C_\psi \cdot \epsilon_{app} - \alpha_t \cdot N \cdot \epsilon_{sp} + \alpha_t \cdot (1-\gamma) \cdot N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big] \qquad (108)
\end{aligned}
$$

Moreover, by adding (101) and (108), we obtain $I_1 := I_1^1 + I_1^2$ as below:

$$
\begin{aligned}
I_1 &\geq -2N \cdot \alpha_t \cdot L_v \cdot C_\psi \cdot \epsilon_{app} - \alpha_t \cdot N \cdot \epsilon_{sp} + \alpha_t \cdot (1-\gamma) \cdot N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big] \\
&\quad - 4\alpha_t \cdot C_\psi \cdot \sqrt{N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big]} \cdot \sqrt{\sum_{j=1}^{N}\Big(\mathbb{E}\Big[\|\omega_t^* - \omega_{i,t}\|^2\Big] + \mathbb{E}\Big[\|\lambda_t^* - \lambda_{i,t}\|^2\Big]\Big)}. \qquad (109)
\end{aligned}
$$

In order to bound term I_2, we are able to further analyze $I_{2,1}$ and $I_{2,1}$ as below:

$$\begin{aligned}
I_{2,1} &:= \alpha_t \cdot \sum_{i=1}^N \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \Big(\widehat{\delta}_{i,t} - \widehat{\delta}_t^*\Big)\psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t})\Big\rangle\Big] \\
&= -\alpha_t \cdot \sum_{i=1}^N \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \big(\gamma\phi(s_{t+1}) - \phi(s_t)\big)^T \big(\omega_t^* - \omega_{i,t}\big) \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t})\Big\rangle\Big] \\
&\quad - \alpha_t \cdot \sum_{i=1}^N \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \varphi(s_t, \boldsymbol{a}_t)^T(\lambda_t^* - \lambda_{i,t}) \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t})\Big\rangle\Big] \\
&\overset{(i)}{\geq} -2\alpha_t \cdot C_\psi \cdot \sum_{i=1}^N \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\| \cdot \Big(\|\omega_t^* - \omega_{i,t}\| + \|\lambda_t^* - \lambda_{i,t}\|\Big)\Big] \\
&= -2\alpha_t \cdot C_\psi \cdot \sum_{i=1}^N \sqrt{\Big(\mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\| \cdot \Big(\|\omega_t^* - \omega_{i,t}\| + \|\lambda_t^* - \lambda_{i,t}\|\Big)\Big]\Big)^2} \\
&\overset{(ii)}{\geq} -2\alpha_t \cdot C_\psi \cdot \sum_{i=1}^N \sqrt{\mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big]} \cdot \sqrt{\mathbb{E}\Big[(\|\omega_t^* - \omega_{i,t}\| + \|\lambda_t^* - \lambda_{i,t}\|)^2\Big]} \\
&\overset{(iii)}{\geq} -4\alpha_t \cdot C_\psi \cdot \sqrt{\sum_{i=1}^N \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big]} \cdot \sqrt{\sum_{i=1}^N \mathbb{E}\Big[\|\omega_t^* - \omega_{i,t}\|^2 + \|\lambda_t^* - \lambda_{i,t}\|^2\Big]} \qquad (110)
\end{aligned}$$

where (i) is due to the facts that all feature vectors are bounded by Assumption 4 and the score functions are bounded as : $\|\psi_{\theta_i^p}(s_t, a_{j,t}; \theta_{i,t})\| \leq C_\psi$ in (14a); (ii) and (iii) are from Cauchy–Schwarz inequality. Moreover, for the term $I_{2,2}$ in (99), we can further express it as:

$$\begin{aligned}
I_{2,2} &:= \alpha_t \sum_{i=1}^N \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \widehat{\delta}_t^* \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t})\Big\rangle\Big] \\
&= \alpha_t \cdot (1-\gamma) \sum_{i=1}^N \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big] + \alpha_t \sum_{i=1}^N \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \Big(\widehat{\delta}_t^* - \delta_t\Big) \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t})\Big\rangle\Big] \\
&\quad + \alpha_t \sum_{i=1}^N \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \delta_t \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t}) - (1-\gamma)\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\Big\rangle\Big] \qquad (111)
\end{aligned}$$

where we have defined $\delta_t := \bar{r}_t + \gamma V_{\pi_{\boldsymbol{\theta}_t}}(s_{t+1}) - V_{\pi_{\boldsymbol{\theta}_t}}(s_t)$.

For each term above in (111), it holds that

$$\begin{aligned}
&\alpha_t \sum_{i=1}^N \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \Big(\widehat{\delta}_t^* - \delta_t\Big) \cdot \psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t})\Big\rangle\Big] \\
&\geq -\alpha_t \sum_{i=1}^N \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\| \cdot \|\psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t})\| \cdot \big|\widehat{\delta}_t^* - \delta_t\big|\Big] \\
&\overset{(i)}{\geq} -\alpha_t \cdot L_v \cdot C_\psi \cdot N \cdot \mathbb{E}\Big[\Big|\Big(\varphi(s_t, \boldsymbol{a}_t)^T\lambda_t^* - \bar{r}_t\Big) + \gamma\Big(\phi(s_{t+1})^T\omega_t^* - V_{\pi_{\boldsymbol{\theta}_t}}(s_{t+1})\Big) - \Big(\phi(s_t)^T\omega_t^* - V_{\pi_{\boldsymbol{\theta}_t}}(s_t)\Big)\Big|\Big] \\
&\overset{(ii)}{\geq} -\alpha_t \cdot L_v \cdot C_\psi \cdot N \cdot \Big(\mathbb{E}[|\varphi(s_t, a_t)^T\lambda_t^* - r_t|] + \gamma\mathbb{E}\Big[|\phi(s_{t+1})^T\omega_t^* - V_{\pi_{\boldsymbol{\theta}_t}}(s_{t+1})|\Big] + \mathbb{E}\Big[|V_{\pi_{\boldsymbol{\theta}_t}}(s_t) - \phi(s_t)^T\omega_t^*|\Big]\Big) \\
&\overset{(iii)}{\geq} -\alpha_t \cdot L_v \cdot C_\psi \cdot N \cdot \Big(\sqrt{\mathbb{E}\Big[\big(\varphi(s_t, \boldsymbol{a}_t)^T\lambda_t^* - \bar{r}_t\big)^2\Big]} + \gamma\sqrt{\mathbb{E}\Big[\big(\phi(s_{t+1})^T\omega_t^* - V_{\pi_{\boldsymbol{\theta}_t}}(s_{t+1})\big)^2\Big]} \\
&\quad + \sqrt{\mathbb{E}\Big[\big(V_{\pi_{\boldsymbol{\theta}_t}}(s_t) - \phi(s_t)^T\omega_t^*\big)^2\Big]}\Big) \\
&\overset{(iv)}{\geq} -2N \cdot \alpha_t \cdot L_v \cdot C_\psi \cdot \epsilon_{app} \qquad (112)
\end{aligned}$$

where (i) follows (18) and $\|\psi_{\theta_i^p}(s_t, a_{i,t}; \theta_{i,t})\| \leq C_\psi$; (ii) follows from the triangle inequality; (iii) follows Jensen's inequality; (iv) follows the definition of approximation error in (28).

Similarly as the derivation in (107), we can further bound the third term in (111) as below:

$$\alpha_t \sum_{i=1}^{N} \mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \delta_t \cdot \psi_{\theta_i^l}(s_t, a_{i,t}; \theta_{i,t}) - (1-\gamma)\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t) \Big\rangle\Big] \geq -\alpha_t \cdot N \cdot \epsilon_{sp}. \tag{113}$$

Plugging the inequalities (112) - (113) into (111), then it holds that:

$$I_{2,2} \geq \alpha_t \cdot (1-\gamma) \cdot \sum_{i=1}^{N} \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big] - \alpha_t \cdot N \cdot \epsilon_{sp} - 2\alpha_t \cdot N \cdot L_v \cdot C_\psi \cdot \epsilon_{app}. \tag{114}$$

Adding (110) and (114), it follows that:

$$\begin{aligned} I_{2,1} + I_{2,2} \geq\ & \alpha_t \cdot (1-\gamma) \cdot \sum_{i=1}^{N} \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big] - \alpha_t \cdot N \cdot \epsilon_{sp} - 2\alpha_t \cdot N \cdot L_v \cdot C_\psi \cdot \epsilon_{app} \\ & - 4\alpha_t \cdot C_\psi \cdot \sqrt{\sum_{i=1}^{N} \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big]} \cdot \sqrt{\sum_{i=1}^{N} \mathbb{E}\Big[(\|\omega_t^* - \omega_{i,t}\| + \|\lambda_t^* - \lambda_{i,t}\|)^2\Big]} \end{aligned} \tag{115}$$

Denote $P_t := \sum_{i=1}^{N} \mathbb{E}\big[\|\omega_t^* - \omega_{i,t}\|^2 + \|\lambda_t^* - \lambda_{i,t}\|^2\big]$. Then we plug the inequalities (109) and (115) into (99), it holds that

$$\begin{aligned} \mathbb{E}[J(\bar{\boldsymbol{\theta}}_{t+1})] \geq\ & \mathbb{E}[J(\bar{\boldsymbol{\theta}}_t)] - N \cdot \alpha_t \cdot \ell_p \cdot L_J \cdot E\big[\|Q \cdot \boldsymbol{\theta}_t^s\|\big] - \frac{N \cdot L_J \cdot \alpha_t^2 \cdot \ell_p^2}{2} \\ & + \alpha_t \cdot (1-\gamma) \cdot N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big] - 2N \cdot \alpha_t \cdot L_v \cdot C_\psi \cdot \epsilon_{app} - \alpha_t \cdot N \cdot \epsilon_{sp} \\ & + \alpha_t \cdot (1-\gamma) \cdot \sum_{i=1}^{N} \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big] - 2N \cdot \alpha_t \cdot L_v \cdot C_\psi \cdot \epsilon_{app} - \alpha_t \cdot N \cdot \epsilon_{sp} \\ & - 4\alpha_t \cdot C_\psi \cdot \left(\sqrt{N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big]} + \sqrt{\sum_{i=1}^{N} \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big]}\right) \cdot \sqrt{P_t}. \end{aligned} \tag{116}$$

Denote a constant $C_s := N \cdot \ell_p \cdot L_J$. Rearrange inequality (116) and apply Cauchy-Schwarz inequality, it holds that

$$\begin{aligned} & N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}_s} J(\boldsymbol{\theta}_t)}\|^2\Big] + \sum_{i=1}^{N} \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big] \\ & \leq \frac{1}{\alpha_t \cdot (1-\gamma)} \cdot \mathbb{E}\Big[J(\bar{\boldsymbol{\theta}}_{t+1}) - J(\bar{\boldsymbol{\theta}}_t)\Big] + \frac{1}{1-\gamma}\left(\frac{\alpha_t \cdot N \cdot L_J \cdot \ell_p^2}{2} + 4N \cdot L_v \cdot C_\psi \cdot \epsilon_{app} + 2N \cdot \epsilon_{sp}\right) \\ & + \frac{C_s}{1-\gamma} \cdot \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|\Big] + \frac{8C_\psi}{1-\gamma} \cdot \sqrt{N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big] + \sum_{i=1}^{N} \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big]} \cdot \sqrt{P_t} \end{aligned} \tag{117}$$

Then we can denote $G_t := N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big] + \sum_{i=1}^N \mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big]$. Through summing the inequality (117) from $t = 0$ to $T-1$ and divide T on both side, it holds that

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} G_t \\
&\leq \underbrace{\frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{\alpha_t(1-\gamma)} \cdot \Big(\mathbb{E}\big[J(\bar{\boldsymbol{\theta}}_{t+1})\big] - \mathbb{E}\big[J(\bar{\boldsymbol{\theta}}_t)\big]\Big) + \frac{C_s}{T(1-\gamma)} \cdot \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|\Big] + \frac{N L_J \ell_p^2}{2T(1-\gamma)} \cdot \sum_{t=0}^{T-1} \alpha_t}_{\text{term } \mathrm{M}_1} \\
&\quad + \underbrace{\frac{1}{1-\gamma}\Big(4N \cdot L_v \cdot C_\psi \cdot \epsilon_{app} + 2N \cdot \epsilon_{sp}\Big)}_{\text{term } \mathrm{M}_2} + \frac{8C_\psi}{T(1-\gamma)} \cdot \sum_{t=0}^{T-1} \sqrt{G_t} \cdot \sqrt{P_t}
\end{aligned}
\tag{118}
$$

Then (118) could be expressed as below:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} G_t &\overset{(118)}{\leq} M_1 + M_2 + \frac{8C_\psi}{T \cdot (1-\gamma)} \cdot \sum_{t=0}^{T-1} \sqrt{G_t} \cdot \sqrt{P_t} \\
&\overset{(a)}{\leq} M_1 + M_2 + \frac{8C_\psi}{1-\gamma} \cdot \sqrt{\frac{1}{T}\sum_{t=0}^{T-1} G_t} \cdot \sqrt{\frac{1}{T}\sum_{t=0}^{T-1} P_t}
\end{aligned}
\tag{119}
$$

where (a) follows Cauchy-Schwarz inequality. Then we define $C_g := \frac{8C_\psi}{1-\gamma}$, $B_1 := \frac{1}{T}\sum_{t=0}^{T-1} G_t$, $B_2 := M_1 + M_2$ and $B_3 := \frac{1}{T}\sum_{t=0}^{T-1} P_t$. The inequality (119) could be expressed as below:

$$
\begin{aligned}
&B_1 \leq B_2 + C_g \cdot \sqrt{B_1 \cdot B_3} \implies \Big(\sqrt{B_1} - \frac{C_g}{2} \cdot \sqrt{B_3}\Big)^2 \leq B_2 + \frac{C_g^2}{4} \cdot B_3 \\
&\sqrt{B_1} \leq \sqrt{B_2 + \frac{C_g^2}{4} \cdot B_3} + \frac{C_g}{2} \cdot \sqrt{B_3} \implies B_1 \leq 2B_2 + C_g^2 \cdot B_3
\end{aligned}
\tag{120}
$$

Then we are able to analyze the convergence rate of each term in (118). Recall the fixed stepsize $\alpha_t = \frac{\alpha}{T^{\sigma_1}}$. We bound each component in M_1 defined above.

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{\alpha_t(1-\gamma)} \cdot \Big(\mathbb{E}\big[J(\bar{\boldsymbol{\theta}}_{t+1})\big] - \mathbb{E}\big[J(\bar{\boldsymbol{\theta}}_t)\big]\Big) \\
&\overset{(a)}{\leq} \frac{T^{\sigma_1}}{\alpha \cdot T \cdot (1-\gamma)} \cdot \mathbb{E}\big[J(\bar{\boldsymbol{\theta}}_T)\big] \\
&\overset{(b)}{\leq} \frac{T^{\sigma_1}}{\alpha_{T-1} \cdot T \cdot (1-\gamma)} \cdot \frac{R_{\max}}{1-\gamma} = \mathcal{O}\Big(T^{\sigma_1 - 1}\Big)
\end{aligned}
\tag{121}
$$

where (a) follows $\alpha_t = \frac{\alpha}{T^{\sigma_1}}$; (b) is due to the fact that each reward is bounded by $R_{\max}$ and $J(\boldsymbol{\theta}) \leq \sum_{t=0}^{\infty} \gamma^t \cdot R_{\max} = \frac{R_{\max}}{1-\gamma}$ for any $\boldsymbol{\theta} \in R^{N \times D}$.

For the second and third terms in M_1, it holds that

$$
\frac{C_s}{(1-\gamma) \cdot T} \cdot \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Q \cdot \boldsymbol{\theta}_t^s\|\Big] \overset{(63)}{=} \mathcal{O}\Big(T^{-\sigma_1}\Big)
\tag{122}
$$

$$
\frac{N \cdot L_J \cdot \ell_p^2}{(1-\gamma) \cdot T} \cdot \sum_{t=0}^{T-1} \alpha_t = \frac{N \cdot L_J \cdot \ell_p^2}{(1-\gamma)} \cdot \frac{\alpha}{T^{\sigma_1}} = \mathcal{O}\Big(T^{-\sigma_1}\Big)
\tag{123}
$$

Combining (121) - (123), we obtain that the convergence rate of term M_1 in (118) could be expressed as below:

$$
M_1 = \mathcal{O}\Big(T^{\sigma_1 - 1}\Big) + \mathcal{O}\Big(T^{-\sigma_1}\Big)
\tag{124}
$$

Then according to (118) and (120), it holds that

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left(N\cdot\mathbb{E}\left[\|\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)}\|^2\right]+\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta_i^p}J(\boldsymbol{\theta}_t)\|^2\right]\right)\\ \leq&2M_1+2M_2+\frac{C_g^2}{T}\sum_{i=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[\|\omega_t^*-\omega_{i,t}\|^2+\|\lambda_t^*-\lambda_{i,t}\|^2\right].\end{aligned}$$

By applying in the convergence results of approximation parameters in (92), we obtain that

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left(N\cdot\mathbb{E}\left[\|\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)}\|^2\right]+\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta_i^p}J(\boldsymbol{\theta}_t)\|^2\right]\right)\\ &\overset{(92)}{\leq}2M_1+2M_2+\frac{2}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left(\mathbb{E}\left[\|\bar{\omega}_t-\omega_t^*\|^2\right]+\mathbb{E}\left[\|\bar{\lambda}_t-\lambda_t^*\|^2\right]\right)+\frac{2}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\|Q\cdot\boldsymbol{\omega}_t\|^2\right]+\mathbb{E}\left[\|Q\cdot\boldsymbol{\lambda}_t\|^2\right]\right)\\ &=\mathcal{O}(\epsilon_{app}+\epsilon_{sp})+\mathcal{O}(T^{\sigma_1-1})+\mathcal{O}(T^{-\sigma_1})+\mathcal{O}(T^{-1+\sigma_2})+\mathcal{O}(T^{-\sigma_2})+\mathcal{O}\left(T^{\sigma_2-2\sigma_1}\right)\\ &\quad+\mathcal{O}(T^{-2\sigma_1+2\sigma_2})+\mathcal{O}(T^{-2+2\sigma_2})+\mathcal{O}(T^{-2\sigma_2}).\end{aligned}\tag{125}$$

To optimize the convergence rate, we choose $\alpha_t=\mathcal{O}(\frac{1}{T^{0.6}}),\beta_t=\mathcal{O}(\frac{1}{T^{0.4}})$ so that $\sigma_1=\frac{3}{5}$ and $\sigma_2=\frac{2}{5}$, then plug them into (125). Therefore, the convergence rate of the actor is:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(N\cdot\mathbb{E}\left[\|\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)}\|^2\right]+\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla_{\theta_i^p}J(\boldsymbol{\theta}_t)\|^2\right]\right)=\mathcal{O}(T^{-\frac{2}{5}})+\mathcal{O}(\epsilon_{app}+\epsilon_{sp}).\tag{126}$$

This completes the proof. □

I Detailed Analysis for Double Sampling Procedures

For the CAC algorithm with double sampling procedures in Algorithm 2, we generate two different tuples $x_t:=(s_t,\boldsymbol{a}_t,s_{t+1})$ and $\tilde{x}_t:=(\tilde{s}_t,\tilde{\boldsymbol{a}}_t,\tilde{s}_{t+1})$ in each iteration t to perform critic step and actor step. In x_t, the state s_t is sampled from the stationary distribution $\mu_{\boldsymbol{\theta}_t}(\cdot)$. In $\tilde{x}_t$, the state $\tilde{s}_t$ is sampled from the discounted visitation measure $d_{\boldsymbol{\theta}_t}(\cdot)$ where $d_{\boldsymbol{\theta}}(\tilde{s}):=(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\cdot\mathcal{P}^{\pi_{\boldsymbol{\theta}}}(s_t=\tilde{s}\mid s_0\sim\eta)$.

Then the difference between double sampling procedures and single sampling procedure comes from bounding the term $I_{1,2}$ in (103) and $I_{2,2}$ in (111). With double samples x_t and $\tilde{x}_t$ at each iteration t, the sampling error ϵ_{sp} could be avoided in analyzing $I_{1,2}$ and $I_{2,2}$

With tuple $\tilde{x}_t$ to perform actor step, the component $I_{1,2}$ in (103) could be expressed as

$$\begin{aligned}I_{1,2}:=&\alpha_t\cdot(1-\gamma)\cdot N\cdot\mathbb{E}\left[\|\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)}\|^2\right]+\alpha_t\cdot\sum_{j=1}^{N}\mathbb{E}\left[\left\langle\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)},(\widehat{\delta}_t^*-\delta_t)\cdot\psi_{\theta_j^s}(\tilde{s}_t,\tilde{a}_{j,t};\theta_{j,t})\right\rangle\right]\\ &+\alpha_t\cdot\sum_{j=1}^{N}\mathbb{E}\left[\left\langle\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)},\delta_t\cdot\psi_{\theta_j^s}(\tilde{s}_t,\tilde{a}_{j,t};\theta_{j,t})-(1-\gamma)\cdot\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)}\right\rangle\right]\end{aligned}$$

Following the inequality (104), the second term in $I_{1,2}$ above could be bounded. Then it holds that

$$\begin{aligned}I_{1,2}\geq&-2N\cdot\alpha_t\cdot L_v\cdot C_\psi\cdot\epsilon_{app}+\alpha_t\cdot(1-\gamma)\cdot N\cdot\mathbb{E}\left[\|\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)}\|^2\right]\\ &+\alpha_t\cdot\sum_{j=1}^{N}\mathbb{E}\left[\left\langle\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)},\delta_t\cdot\psi_{\theta_j^s}(\tilde{s}_t,\tilde{a}_{j,t};\theta_{j,t})-(1-\gamma)\cdot\overline{\nabla_{\boldsymbol{\theta}^s}J(\boldsymbol{\theta}_t)}\right\rangle\right]\end{aligned}$$

For the third term in $I_{1,2}$, it holds that

$$
\begin{aligned}
&\alpha_t \cdot \sum_{j=1}^{N} \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \delta_t \cdot \psi_{\theta_j^s}(\tilde{s}_t, \tilde{a}_{j,t}; \theta_{j,t}) - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} \Big\rangle\Big] \\
&= \alpha_t \cdot N \cdot \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \delta_t \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(\tilde{s}_t, \tilde{a}_{j,t}; \theta_{j,t})\big) - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} \Big\rangle\Big] \\
&\overset{(i)}{=} \alpha_t \cdot N \cdot \mathbb{E}\Big[\mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \delta_t \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(\tilde{s}_t, \tilde{a}_{j,t}; \theta_{j,t})\big) - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)} \Big\rangle\Big|\boldsymbol{\theta}_t\Big]\Big] \\
&\overset{(ii)}{=} \alpha_t \cdot N \cdot \mathbb{E}\Big[\Big\langle \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}, \Big(\mathbb{E}_{\tilde{s}_t \sim d_{\boldsymbol{\theta}_t}(\cdot), \tilde{\boldsymbol{a}}_t \sim \pi_{\boldsymbol{\theta}_t}(\cdot|\tilde{s}_t)}\Big[\big(Q_{\pi_{\boldsymbol{\theta}_t}}(\tilde{s}_t, \tilde{\boldsymbol{a}}_t) - V_{\pi_{\boldsymbol{\theta}_t}}(\tilde{s}_t)\big) \cdot \big(\frac{1}{N}\sum_{j=1}^{N} \psi_{\theta_j^s}(\tilde{s}_t, \tilde{a}_{j,t}; \theta_{j,t})\big)\Big] \\
&\quad - (1-\gamma) \cdot \overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\Big)\Big\rangle\Big] \\
&= 0
\end{aligned}
$$

where (i) follows Tower rule; (ii) follows the policy gradient expression in (106).

Therefore, we bound term $I_{1,2}$ as below:

$$
I_{1,2} \geq -2N \cdot \alpha_t \cdot L_v \cdot C_\psi \cdot \epsilon_{app} + \alpha_t \cdot (1-\gamma) \cdot N \cdot \mathbb{E}\Big[\|\overline{\nabla_{\boldsymbol{\theta}^s} J(\boldsymbol{\theta}_t)}\|^2\Big]
$$

Moreover, the component $I_{2,2}$ in (111) is expressed as below:

$$
\begin{aligned}
I_{2,2} :=& \alpha_t(1-\gamma)\sum_{i=1}^{N}\mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big] + \alpha_t \sum_{i=1}^{N}\mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \Big(\widehat{\delta}_t^* - \delta_t\Big) \cdot \psi_{\theta_i^p}(\tilde{s}_t, \tilde{a}_{i,t}; \theta_{i,t})\Big\rangle\Big] \\
&+ \alpha_t \sum_{i=1}^{N}\mathbb{E}\Big[\Big\langle \nabla_{\theta_i^p} J(\boldsymbol{\theta}_t), \delta_t \cdot \psi_{\theta_i^p}(\tilde{s}_t, \tilde{a}_{i,t}; \theta_{i,t}) - (1-\gamma)\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\Big\rangle\Big] \qquad (127)
\end{aligned}
$$

Similarly following the same steps in analyzing term $I_{1,2}$, the sampling error ϵ_{sp} could be avoided due to double samples in each iteration t. Hence, it holds that

$$
I_{2,2} \geq \alpha_t \cdot (1-\gamma) \cdot \sum_{i=1}^{N}\mathbb{E}\Big[\|\nabla_{\theta_i^p} J(\boldsymbol{\theta}_t)\|^2\Big] - 2\alpha_t \cdot N \cdot L_v \cdot C_\psi \cdot \epsilon_{app}
$$

Following remaining analysis steps in Proposition 2, we obtain the convergence analysis for CAC with double sampling procedures in Algorithm 2. We are able to avoid the sampling error ϵ_{sp} at the cost of utilizing one more sample at each iteration. Therefore, we are able to present the results in Corollary 1.