

The Gradient Convergence Bound of Federated Multi-Agent Reinforcement Learning with Efficient Communication

Xing Xu¹, Rongpeng Li^{1*}, Zhifeng Zhao², and Honggang Zhang^{1,2}

¹Zhejiang University, ²Zhejiang Lab

{hsuxing, lirongpeng, honggangzhang}@zju.edu.cn, zhaozf@zhejianglab.com

Abstract—The paper considers independent reinforcement learning (IRL) for multi-agent collaborative decision-making in the paradigm of federated learning (FL). However, FL generates excessive communication overheads between agents and a remote central server, especially when it involves a large number of agents or iterations. Besides, due to the heterogeneity of independent learning environments, multiple agents may undergo asynchronous Markov decision processes (MDPs), which will affect the training samples and the model’s convergence performance. On top of the variation-aware periodic averaging (VPA) method and the policy-based deep reinforcement learning (DRL) algorithm (i.e., proximal policy optimization (PPO)), this paper proposes two advanced optimization schemes orienting to stochastic gradient descent (SGD): 1) A decay-based scheme gradually decays the weights of a model’s local gradients with the progress of successive local updates, and 2) By representing the agents as a graph, a consensus-based scheme studies the impact of exchanging a model’s local gradients among nearby agents from an algebraic connectivity perspective. This paper also provides novel convergence guarantees for both developed schemes, and demonstrates their superior effectiveness and efficiency in improving the system’s utility value through theoretical analyses and simulation results.

Index Terms—Independent Reinforcement Learning, Federated Learning, Consensus Algorithm, Communication Overheads

I. INTRODUCTION

With the development of wireless communication and advanced machine learning technologies in the past few years, a large amount of data has been generated by smart devices and can enable a variety of multi-agent systems, such as smart road traffic control [1], smart home energy management [2], and the deployment of unmanned aerial vehicles (UAVs) [3]–[5]. Through deep reinforcement learning (DRL), an intelligent agent can gradually improve the performance of its parameterized policy via the trial-and-error interaction with the environment [6]–[9]. However, directly applying DRL to multi-agent systems commonly faces several challenging problems, such as the non-stationary learning environment and the difficulty of reward assignment [10], [11]. As an alternative, independent reinforcement learning (IRL) is often employed in practical applications to alleviate the above-mentioned problems, where each agent undergoes an independent learning process with only self-related observations [12].

For each IRL agent, the training samples are obtained by going through a trajectory with the predefined terminal state

or a certain number of Markov state transitions from Markov decision processes (MDPs). With the obtained samples, an agent can improve its performance by updating its policy’s parameters along the gradient descent direction. In this paper, a policy-based DRL algorithm (i.e., proximal policy optimization (PPO) [8]) is applied to calibrate the loss function, by repetitively calculating the gradients [13]. Generally speaking, the performance of DRL is closely related to the amount and variety of obtained samples, since the more fully explored state space leads to more accurate estimation of the cumulative reward signal [10]. Meanwhile, for a multi-agent system with naturally distributed IRL agents, the locally calculated policy gradients need to be shared through a coordination channel. Therefore, in order to effectively improve the performance of DRL, we adopt a federated multi-agent reinforcement learning (FMARL) framework, by combining the stochastic gradient descent (SGD) in DRL and federated learning (FL) [14]–[16]. In particular, a central server in FL is leveraged by iteratively aggregating the policy gradients from multiple IRL agents and in turn providing updated policy parameters. Therefore, FL can facilitate the implementation of this coordination channel and significantly contribute to enriching the sample information of each IRL agent indirectly.

However, targeting this FMARL framework, there may be a large number of agents or policy iterations during the training phase. This naive implementation of FL may generate excessive communication overheads between agents and the central server. To alleviate this problem, periodic averaging has been naturally considered and popularly applied in many studies [17]–[19], in which agents are allowed to perform several local updates within a period before transmitting their local gradients to the central server, so as to reduce the frequency of information exchange. However, an increase in the number of local updates would influence the convergence performance. Therefore, appropriate optimization methods should be developed to better balance the reduction of communication overheads and the improvement of the convergence performance. Besides, considering the heterogeneity of independent learning environments, multiple IRL agents may spend various amounts of time on state transition processes, and perform different numbers of local updates in the same period under the periodic averaging method, thus possibly affecting the convergence performance as well. Therefore, variations in the number of

local updates under the periodic averaging method should be also considered.

Taking account of the model's error convergence bound with respect to the mainly required communication and computation overheads during the training phase, this paper proposes a system utility function-based metric to evaluate the effectiveness of different optimization methods for federated IRL (FIRL). Furthermore, in order to improve the system's utility value, this paper develops two new optimization methods (i.e., the decay-based and consensus-based methods) on top of the variation-aware periodic averaging (VPA) method. In particular, since the variance of model's gradients gradually increases along with the progress of local updating, the decay-based method utilizes a discrete periodic decay function to decrease the weights of a model's successive local gradients within one period, which also contrasts with the adoption of decaying learning rates for multiple epochs in [19]. We concretely provide a practical implementation of this decay-based method with an exponential decay function and theoretically demonstrate its superior effectiveness. On the other hand, the consensus-based method introduces the consensus algorithm [20] into both IRL and FL, where agents are allowed to exchange their local gradients with neighbors directly before performing local updates. Different from the spectral radius-based theoretical analyses in [21], this paper obtains the consensus-based method's error convergence bound dependent on the algebraic connectivity of the graph comprised by agents and their connections. We theoretically show that the consensus-based method can reduce the model's error convergence bound dramatically and speed up the learning process. Table I summarizes the key differences with highly related works. Finally, through an MARL simulation scenario, we demonstrate the superiority of these two developed methods in terms of the model's convergence performance and system's utility value.

In summary, the main contributions of this paper are summarized as follows.

- We propose an on-line FMARL framework, by improving the policy performance of distributed IRL agents through FL. Considering the possible excessive communication overheads and heterogeneous independent learning environments, we take the VPA method as the basis of systematic analysis.
- In order to measure and optimize the effectiveness of policy-based DRL (i.e., PPO) through general SGD [13], we put forward a system's utility function, which quantifies the convergence bound of the model's error reduction per unit of resource cost during the training phase, so as to reasonably evaluate the effectiveness of different optimization methods.
- On top of the systematic analysis in this paper, we propose some novel implementation means to realize the decay-based and consensus-based methods for FIRL, so as to more efficiently aggregate the gradients from multiple agents within each period. Specifically, an exponentially decaying function is applied to decrease the weights of successive local gradients, while the gradients

from multiple agents are merged together in a hierarchical manner. We demonstrate the superiority of both methods through theoretical analyses and numerical simulation results.

The remainder of this paper is mainly organized as follows. In Section II, we explain the related works and clarify the novelty of our work. In Section III, we present preliminaries of the periodic averaging method. In Section IV, we introduce the system model and formulate the optimization problem. In Section V, we consider the possible heterogeneity of FMARL and analyze the VPA method with different numbers of updates each period. In Section VI, we describe two optimization methods (i.e., the decay-based and consensus-based methods) and give their error convergence bounds. In Section VII, we introduce the MARL simulation scenario and present the corresponding results of the developed methods. In Section VIII, we conclude this paper with a summary.

II. RELATED WORKS

In recent years, distributed IRL [12] has been frequently studied and applied in many practical applications [29]–[31]. However, due to the heterogeneity of independent learning environments, the performance of IRL agents may vary significantly, even though agents are deployed in the same global environment and face exactly the same facilities. Several studies [8], [32], [33] have shown that experiences from homogeneous independent learning agents in the same multi-agent system can be collected and sampled together to obtain a shared model efficiently. Besides, the recent studies [14]–[16] have combined distributed IRL with FL to improve the involved agents' capability and collaboration efficiency. The combining of FL and DRL remains a hot research topic [22]–[24]. For example, [22] adopts a decoupled setting to combine DRL and FL. Wherein, [22] uses a hierarchical nested personalized FL for stratified UAVs swarms, while leverages DRL for swarm trajectory and learning duration design. [23] focuses on a digital-twin empowered industrial Internet-of-things (IoTs) scenario, and presents a single-agent DRL-based device selection for asynchronous FL. Contrary to these papers, our work is primarily oriented at the direct enhancement of decentralized multi-agent DRL by FL. Furthermore, [24] targets at the data heterogeneity issue of multi-agent DRL by imposing KL divergence-based penalty term. Instead, our work aims to solve learning inefficiency issue of multi-agent DRL by decay and consensus-based FL methods. Besides, we further quantify the effect of key parameters in FL on the convergence performance.

An on-line federated transfer reinforcement learning framework is introduced in [34] based on the deep deterministic policy gradient (DDPG) [35] algorithm for autonomous driving. However, this framework lacks the theoretical analysis for the performance of FL in DRL, while the interplay between deterministic actor network and the Q-function typically makes DDPG extremely difficult to stabilize and brittle to hyper-parameter settings [36]. Meanwhile, a federated DRL-based cooperative edge caching (FADE) framework is proposed in [37] to optimize the edge caching schemes in IoTs services.

TABLE I
SUMMARY OF THE DIFFERENCES WITH HIGHLY RELATED PAPERS.

Articles combining FL and DRL	
<p>[22]: <i>Decoupled</i> setting: hierarchical nested personalized FL for stratified UAVs swarms; DRL for swarm trajectory and learning duration design.</p> <p>[23]: <i>Single-agent</i> DRL-based device selection for asynchronous FL.</p> <p>[24]: Targeting at the <i>data heterogeneity</i> issue of multi-agent DRL by imposing KL divergence-based penalty term.</p>	<p>Ours: Integrated setting: <i>Direct enhancement</i> of decentralized <i>multi-agent</i> DRL by FL.</p> <p>Ours: Targeting at the <i>learning inefficiency</i> issue of multi-agent DRL by decay and consensus-based FL methods.</p>
Decay-based Method	
<p>[19]: Decaying the learning rate of <i>different epochs</i>.</p> <p>[25]: Assuming an <i>off-line</i> available set of data points and assigning <i>aggregation weights according to their sizes</i>. No consideration of the resource cost.</p>	<p>Ours: Decaying the weights of different mini-batches within <i>one period</i>.</p> <p>Ours: <i>On-line</i> DRL tasks and assigning time-decreasing <i>aggregation weights regardless of data sizes</i>.</p>
Consensus-based Method	
<p>[26]–[28]: <i>Disabled</i> central-server's functionality and only consensus-based aggregation of different agents' gradients.</p> <p>[21]: Convergence result with respect to the <i>spectral radius</i> of the network topology.</p>	<p>Ours: <i>Hierarchical</i> configuration with both the agent-server interaction and agents' mutual interaction.</p> <p>Ours: Convergence result with respect to the <i>algebraic connectivity</i> of the network topology.</p>

However, this framework ignores that the frequent information exchange between base stations (BSs) and UEs may generate excessive communication overheads. Moreover, FADE utilizes a value-based DRL method (i.e., deep Q-learning network (DQN) [38]) and theoretically analyzes the model's convergence in SGD based on some general classification loss functions. Nevertheless, the value-based DRL methods exhibit some instability issues in high-dimensional scenarios. Besides, the effect of deploying SGD in improving DRL under these functions is hard to characterize. Instead, this paper utilizes the periodic averaging method to further reduce the excessive communication overheads, and deploys the policy-based DRL method (i.e., PPO algorithm [8]) to effectively improve DRL by SGD [13].

FL is a parallelly distributed machine learning paradigm, aiming to train specific model through the samples distributed across different agents. Due to the protection requirement for data privacy and the restriction from communication bandwidth or delay [39], FL allows distributed agents to calculate their models' gradients locally, and then forward these gradients towards a central server to centrally update the model's parameters. However, the naive FL may generate excessive communication overheads between distributed agents and the central server, especially when there are a large number of agents or iterations [40]. Hence, the periodic averaging method has been proposed to make agents locally perform several updates to the model within a period before transmitting local gradients to the server [41]–[44]. Moreover, several quantization or sparsification methods have been also proposed to implement lossy compression for the model's local gradients that need to be transmitted [45]–[48]. Furthermore, FL with flexible device participation has been considered in [25], where devices (i.e., agents) are assumed to have different sample sizes during the training phase. However, [25] is not designed for on-line training tasks. Instead, [25] assumes devices already have available data points and could assign aggregation weights according to their sizes. On the contrary,

in this paper, the learning paradigm is oriented to on-line DRL tasks, and the discrepancy of agents in terms of sample-collecting efficiency and computing capability is additionally considered. Specifically, we reflect this discrepancy on the numbers of agents' local updates, and formulate our problem on the VPA method.

The theoretical analyses about the error convergence bound under the periodic averaging method have been provided in [19]. Moreover, to reduce this error convergence bound, [19] and [25] have proposed to decrease the learning rate over epochs and optimize the aggregation weights for agents with different sample sizes, respectively. Differently, the decay-based method in this paper aims to decrease the weights of successive gradients over local updates within one period. Note that each epoch includes several periods. Moreover, we additionally provide a practical implementation of the decay-based method through an exponential function, in which the distribution of agents' local updates is further considered. On the other hand, the consensus algorithm [20] has been applied in many decentralized averaging methods [26]–[28], which normally disable the central server's functionality but allow agents to directly exchange and average their parameters with neighbors. However, decentralized averaging methods will slow the convergence rate as the size of agents' network increases. A consensus-based FL algorithm has been proposed in [21] to obtain the training model in large-scale device-to-device (D2D)-enabled fog networks, and an upper bound of convergence with respect to the spectral radius (i.e., the largest eigenvalue of the built induced consensus matrix) of the network topology has been also derived. On the contrary, we derive the error convergence bound from the algebraic connectivity (i.e., the second smallest eigenvalue of the Laplace matrix). Together with our previous work [49], this bound additionally reflects the effect of local interaction metric (i.e., step size ϵ in the classical consensus algorithm [20]), which can be properly adjusted to guarantee the policy improvement of DRL. Finally, both the decay-based and consensus-based

TABLE II
MAIN NOTATIONS USED IN THIS PAPER.

Notation	Definition
π	Stochastic policy
s, s_t	Local state
\mathcal{S}	Local state space
a, a_t	Individual action
\mathcal{A}	Individual action space
K	Total number of iterations
η	Proper learning rate
\mathcal{L}	Loss function
ϕ_t	Markov state transition
r	Individual reward
\mathcal{R}	Reward function
ξ_k	Mini-batch
N	Total number of agents
m	Maximal number of selected agents
τ	Number of local updates within a period
T	Maximal length of an epoch
U	Number of epochs
P	Maximal length of a step
C_1	Communication overheads
C_2	Computation overheads
F	Empirical risk function
ψ_0	Resource cost under general settings
$\psi_0^{(C)}$	Resource cost under the consensus-based VPA method
ψ_1	Error bound of the learning model
$\psi_1^{(P)}$	Error bound under the periodic averaging method
$\psi_1^{(V)}$	Error bound under the VPA method
$\psi_1^{(D)}$	Error bound under the decay-based VPA method
$\psi_1^{(C)}$	Error bound under the consensus-based method
ψ_2	Expected error of the initial model
L	Lipschitz constant
β, σ^2	Non-negative constants
ν	Mean value of the number of local updates
w^2	Variance value of the number of local updates
λ	Decay constant
E	Total number of local interactions
G	Topology of agents' network
Ω_i	Set of neighbors
ϵ	Local interaction step size
\mathbf{La}	Laplace matrix
μ	Eigenvalue of Laplace matrix
W_1	Communication overheads in local interaction
W_2	Computation overheads in local interaction

methods are oriented to on-line MARL tasks, where theoretical analyses about the error convergence bound are based on general assumptions for the VPA method.

III. PRELIMINARIES

Main notations used in this paper are listed in Table II. We assume that each IRL agent maintains a DRL model with parameterized policy $\pi(s, a; \theta)$, where $s \in \mathcal{S}$ is sampled from the local state space, $a \in \mathcal{A}$ is selected from the individual action space, $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ is a stochastic policy, and $\theta \in \mathbb{R}^d$, where d denotes the dimension of parameters. According to stochastic policy gradient methods for the policy optimization process of DRL [7], the learning model's parameters θ can be updated by

$$\theta_{k+1} = \theta_k - \eta \nabla_{\theta_k} \mathcal{L}, \quad (1)$$

where k denotes the index of policy iteration and η is the learning rate with a reasonable step size. According to the PPO algorithm [8], $\mathcal{L}(\theta)$ is defined as

$\mathcal{L}(\theta) = \mathbb{E}_t \left[-J_t^{\text{Clip}}(\theta) + c_1 J_t^V(\theta) - c_2 S(s_t; \theta) \right]$, where c_1, c_2 are weighting factors and S is an entropy bonus. $J_t^V(\theta) = (V_\theta(s_t) - V^{\text{Target}}(s_t))^2$ is a squared-error loss, where V^{Target} denotes the target state value function. $J_t^{\text{Clip}}(\theta)$ denotes a clipped objective and is defined as $J_t^{\text{Clip}}(\theta) = \min \left(\frac{\pi_\theta(s_t, a_t)}{\pi_{\theta_{\text{old}}}(s_t, a_t)} A_t, \text{clip} \left(\frac{\pi_\theta(s_t, a_t)}{\pi_{\theta_{\text{old}}}(s_t, a_t)}, 1 - \zeta, 1 + \zeta \right) A_t \right)$, where ζ denotes the clipping parameter. $A_t = -V(s_t) + r_t + \gamma r_{t+1} + \dots + \gamma^{T-t} V(s_T)$ denotes the accumulated advantage across multiple steps, where γ is a discount factor. Here, \mathcal{L} denotes the loss function to be minimized, and facilitates the calculation for the gradients of objective functions in DRL. Commonly, SGD is used to optimize the loss function and [50] proves SGD could converge for any twice-differentiable loss function, regardless of its convexity. Moreover, our following derivations are based on common assumptions (e.g., L -smooth bounded objective functions). Therefore, the results in this paper can be easily extended to typical settings.

In addition, to stand in consistent with DRL [6], we define the sample used in the policy iteration as

$$\phi_t := \langle s_t, a_t, r_t, s_{t+1} \rangle, \quad (2)$$

where t denotes the time-stamp within an epoch, while an epoch has a predefined terminal state s_{terminal} or a certain number of sequential Markov state transitions ϕ_t . r_t is calculated by a reward function $\mathcal{R} : \mathcal{S} \rightarrow \mathbb{R}$. In order to speed up the training process and reduce the variance of gradients resulting from a single sample, a certain number of samples are picked out at each iteration to comprise a mini-batch. Therefore, the practical gradients used for training are written as

$$g(\theta_k; \xi_k) = \frac{1}{|\xi_k|} \sum_{\phi_t \in \xi_k} \nabla \mathcal{L}(\theta_k; \phi_t), \quad (3)$$

where ξ_k denotes the mini-batch at iteration k , and $|\xi_k|$ represents its size.

On the other hand, considering the multi-agent parallel training process in FL, we can obtain

$$\theta_{k+1} = \theta_k - \eta \frac{1}{m} \sum_{i=1}^m g(\theta_k; \xi_k^{(i)}), \quad (4)$$

where m denotes the maximal number of agents that transmit the local gradients to the central server at iteration k , and $\xi_k^{(i)}$ is the mini-batch from agent i . Furthermore, when the periodic averaging method is applied to aggregate local updates from agents, (4) can be updated as

$$\theta_{k+1}^{(i)} = \begin{cases} \frac{1}{m} \sum_{i=1}^m \left[\theta_k^{(i)} - \eta g(\theta_k^{(i)}; \xi_k^{(i)}) \right], & \text{if } k \bmod \tau = 0; \\ \theta_k^{(i)} - \eta g(\theta_k^{(i)}; \xi_k^{(i)}), & \text{otherwise,} \end{cases} \quad (5)$$

where τ represents the number of local updates in a period. For simplicity, we will use the notation $g(\theta_k^{(i)})$ to represent $g(\theta_k^{(i)}; \xi_k^{(i)})$ in the rest of this paper.

IV. SYSTEM MODEL AND PROBLEM FORMULATION

Fig. 1 illustrates how FMARL introduces FL into MARL. As depicted in Fig. 1, each IRL agent is designed to learn

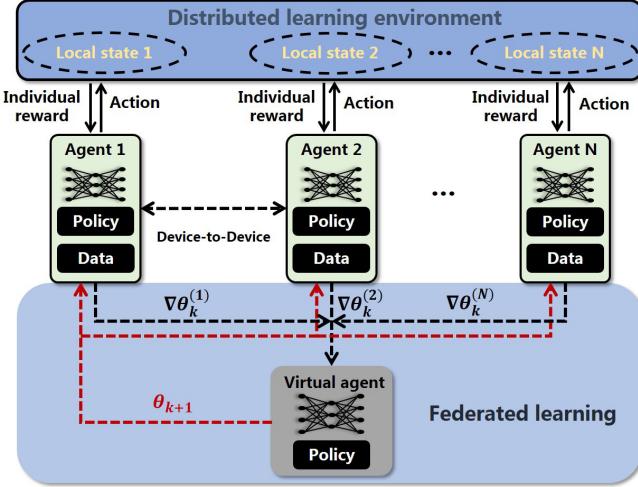


Fig. 1. Framework of FMARL.

independently through FL in the training phase, and operate autonomously in the execution phase via DRL. Within this multi-agent distributed learning scenario, the role of a conventional central server is played by a virtual agent, which can be deployed at a remote cloud center or some capable local agent. Similar to decentralized POMDP [51], we suppose that each agent can only observe a local state to characterize part of the global environment, and an individual reward will be returned to the agent immediately after an action being performed. Meanwhile, to facilitate the multi-agent collaboration, we allow agents to exchange their local gradients with nearby collaborators through the D2D communication whenever needed.

A. Local Updates

Without loss of generality, we assume that there are totally N agents involved in the multi-agent distributed learning scenario. Meanwhile, the maximal length of an epoch and the total number of epochs for training are denoted as T and U , respectively. To facilitate the convergence rate of DRL [52], an epoch is further uniformly divided into several steps, each of which generally contains a predefined number of sequential Markov state transitions (i.e., samples in (2)). In particular, each step includes P transitions as a mini-batch and a local update could be obtained correspondingly. Furthermore, considering that different agents may spend various amounts of time on state transition processes, the wall clock time spent by agent i to finish a step is denoted by a random variable x_i for $i = 1, 2, \dots, N$, where $x_i \in \mathbb{R}^+$. Meanwhile, these random variables satisfy $\mathbb{E}[x_1] \leq \mathbb{E}[x_2] \leq \dots \leq \mathbb{E}[x_N]$. Since the local update only happens after finishing the corresponding step, the number of local updates in a period (i.e., the interval between two successive periodic averaging) for each agent is determined by

$$\tau_i := \left\lfloor \tau \frac{\mathbb{E}[x_1]}{\mathbb{E}[x_i]} \right\rfloor, \quad (6)$$

where τ indicates the number of local updates that agent $i = 1$ can finish on average in a period and should be determined

beforehand. $\lfloor \cdot \rfloor$ is the round down operation and $\tau_i \in \mathbb{N}^+$. Here, the duration of a period is dynamically determined through the time spent by agent $i = 1$ to finish τ local updates. Therefore, for agent $i = 1$, an epoch approximately consists of $\frac{T}{P\tau}$ periods. The reason for this design is that agent $i = 1$ can experience the most transitions from the learning environment, and thus has the greatest weight on the training model. Moreover, in (6), the computation time for any local update is ignored, as it is commonly much less than the time required to finish a certain number of sequential Markov state transitions in DRL. Besides, in order to reduce the training time, only the agents already completing at least one local update at the end of a period need to transmit their local gradients to the virtual agent. Consistent with the notations in (4), we set the maximal number of these involved agents as m , where $m \leq N$. In the rest of this paper, we will focus most of our attention on these m agents. In Figs. 2(a) and (b), we provide an intuitive schematic of agents performing local updates without and with the periodic averaging method, respectively. It can be observed that the implementation of the periodic averaging method makes the numbers of local updates from different agents unsynchronized. Specifically, in (4) and (5), the agent i with a rather small τ_i may satisfy $\xi_k^{(i)} = \emptyset$ and $g(\theta_k^{(i)}) = \mathbf{0}$ at some iteration k .

B. Resource Cost

Considering the resources spent by agents during the training phase of FMARL, we basically suppose that the main communication overheads required for an agent to transmit its local gradients to the virtual agent are C_1 , while the main computation overheads to perform a local update are C_2 . Then, similar to the definition of resource cost in [53], the system's resource cost shall reflect the effect of key parameters (e.g., τ) in optimization methods on the resource cost, and can be intuitively formulated as

$$\psi_0 = \sum_{i=1}^m \left(\frac{C_1 TU}{\tau P} + \frac{C_2 \tau_i TU}{\tau P} \right). \quad (7)$$

It can be observed from (7) that a larger averaging period (i.e., $\tau \neq 1$) can reduce the communication overheads (i.e., the first item in (7)) required for the involved agents by τ times, while a decrease in the number of local updates (i.e., $\tau_i \leq \tau$) can reduce the computation overheads of each agent. However, (7) intentionally neglects the impact of τ on the performance. Therefore, in order to evaluate the model's performance improvement per unit of resource cost, we shall introduce some preliminary results on the stochastic policy gradient methods as well.

C. Convergence Bound of Stochastic Policy Gradient Methods

For any epoch, the optimization objective of policy iteration can be expressed as

$$\pi_\theta = \arg \min_{\theta \in \mathbb{R}^d} \mathbb{E}_t [\mathcal{L}(\pi_\theta; \phi_t)]. \quad (8)$$

According to (1) and the definition of \mathcal{L} , we accomplish the optimization objective in (8) by performing SGD for the

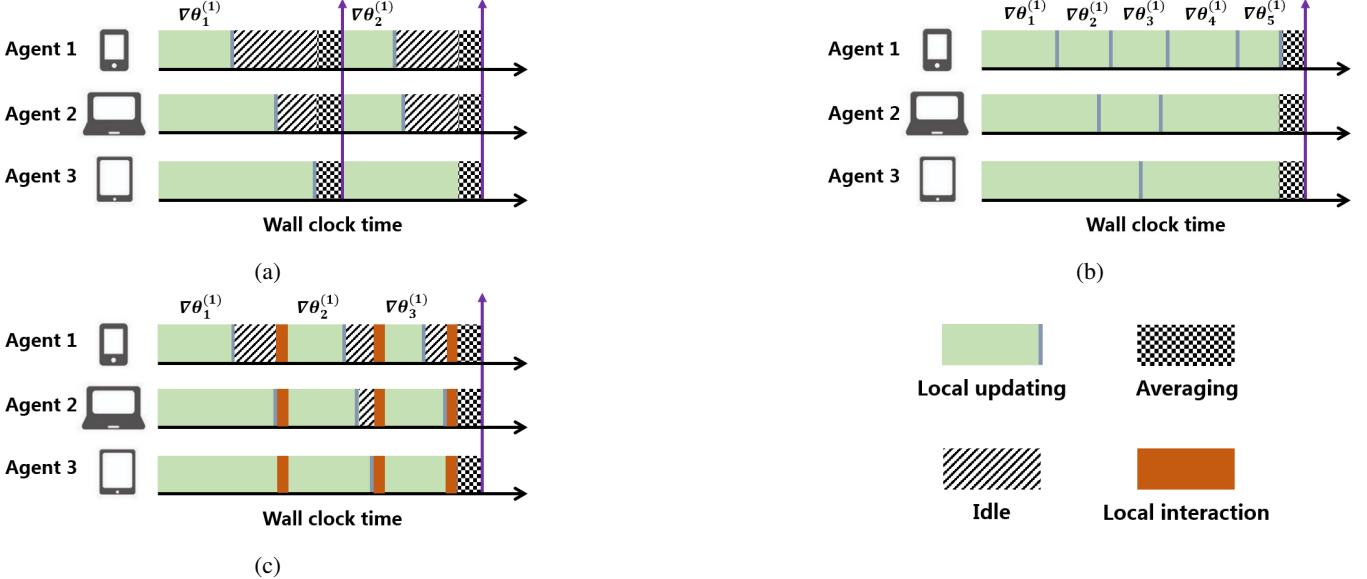


Fig. 2. Intuitive schematic of (a) FL without periodic averaging, (b) the periodic averaging method with $\tau = 5$, and (c) the consensus-based periodic averaging method with $\tau = 3$.

policy parameters θ . In the following discussions, we will replace the notation π_θ with θ for simplicity of representations. Furthermore, we define an empirical risk function as

$$F(\theta) := \frac{1}{|\xi|} \sum_{\phi \in \xi} \mathcal{L}(\theta; \phi). \quad (9)$$

In practical applications, since the empirical risk function $F(\theta)$ may be non-convex, the learning model's convergence may fall into a local minimum or saddle point. In our framework, similar to [42], [54], [55], the expected gradient norm is used as an indicator of the model's convergence to guarantee that it falls into a stationary point, that is

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}_k)\|^2 \right] \leq \psi_1, \quad (10)$$

where K denotes the expected total number of policy iterations, and $K = UT/P$. $\|\cdot\|$ denotes the ℓ_2 vector norm. $\bar{\theta}_k$ represents the average parameters of all agents at iteration k , which is also regarded as the final result at the end of each iteration. Notably, ψ_1 denotes the targeted error convergence bound, towards reducing which we can optimize the means to update the parameters $\bar{\theta}_k$ (or $\theta_k^{(i)}$, for $i \in \{1, 2, 3, \dots, m\}$). Moreover, $k \bmod \tau = 0$, and the model with $\bar{\theta}_k$ is maintained and updated by the virtual agent as

$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{k-1} I(\tau_i > y - t_0) g(\theta_y^{(i)}), \quad (11)$$

where $t_0 = z\tau$, $z \in \mathbb{N}$, and t_0 denotes the index of the iteration, at which the virtual agent performs the latest periodic averaging before iteration k . $I(\cdot)$ denotes the indicator function. Here, each agent transmits the accumulated gradients within a period (i.e., $\sum_{y=t_0}^{k-1} I(\tau_i > y - t_0) g(\theta_y^{(i)})$) to the virtual agent, so as to update the parameters $\bar{\theta}_k$ in a centralized manner. Similarly, we define $\bar{\theta}_0$ as the model's initial parameters

for all agents and obtain

$$\psi_2 := \mathbb{E} \left[\|\nabla F(\bar{\theta}_0)\|^2 \right], \quad (12)$$

where ψ_2 denotes the expected gradient norm of the initial model, which represents the initial convergence error and is gradually decreased by SGD.

D. Optimization Objective of FMARL

Inspired by the optimization objective in [53], which utilizes a predefined threshold as an upper bound to restrict the value of resource cost ψ_0 , we formulate the optimization objective of FMARL as

$$\begin{aligned} \max_{\bar{\theta}_k} \mathcal{U} &:= \frac{\psi_2 - \psi_1}{\psi_0}, \\ \text{s.t. } \psi_1 &\geq \mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}_k)\|^2 \right], \\ \psi_2 &= \mathbb{E} \left[\|\nabla F(\bar{\theta}_0)\|^2 \right], \\ \psi_0 &= \sum_{i=1}^m \left(\frac{C_1 TU}{\tau P} + \frac{C_2 \tau_i TU}{\tau P} \right), \end{aligned} \quad (13)$$

where \mathcal{U} denotes the system's utility function, and indicates the lower bound (rather than the exact value) of error reduction per unit of resource cost. Given the model's initial parameters $\bar{\theta}_0$, \mathcal{U} is jointly determined by the targeted error convergence bound ψ_1 and the resource cost ψ_0 . In order to increase \mathcal{U} , we can equivalently minimize ψ_1 or ψ_0 . Hence, we propose to primarily optimize the update means of $\bar{\theta}_k$ in two different optimization methods, aiming to further reduce the error convergence bound ψ_1 , but 1) keep the resource cost ψ_0 unchanged (i.e., the decay-based method), or 2) make the resource cost ψ_0 not increase too much (i.e., the consensus-based method).

V. VARIATION-AWARE PERIODIC AVERAGING METHOD

In this section, we derive the model's error convergence bound under the VPA method and highlight some useful observations. We consider the heterogeneity of independent learning environments faced by IRL agents from the perspective of the number of local updates τ . Note that one mini-batch consisting of P Markov transitions corresponds to one local update. Consistent with the theoretical analyses about the periodic averaging method in [19] and the situation where agents have different sample sizes in [25], we begin with relevant theoretical results under some common assumptions.

A. General Assumptions

In this paper, the learning model's error convergence bound is inferred under the following general assumptions, which are similar to those presented in previous studies for distributed SGD [42], [55].

Assumption 1 (A1)

1. (Smoothness): $\|\nabla F(\theta) - \nabla F(\theta')\| \leq L \|\theta - \theta'\|$;
2. (Lower bounded): $F(\theta) \geq F_{\inf}$;
3. (Unbiased gradients): $\mathbb{E}_{\theta|\xi} [g(\theta)] = \nabla F(\theta)$;
4. (Bounded variance): $\mathbb{E}_{\theta|\xi} [\|g(\theta) - \nabla F(\theta)\|^2] \leq \beta \|\nabla F(\theta)\|^2 + \sigma^2$,

where L represents the Lipschitz constant, which implies that the empirical risk function F is L -smooth [56]. F_{\inf} denotes the lower bound of F , and we suppose that it can be reached when the total number of iterations K is large enough, which satisfies $F(\bar{\theta}_K) = F_{\inf}$. In addition, β and σ^2 are both non-negative constants and inversely proportional to the size of mini-batch [42]. Condition 3 and 4 in A1 on the bias and variance of the mini-batch gradients are general for SGD methods [39]. Specifically, the variance of the mini-batch gradients is bounded by Condition 4 through the value that fluctuates with the exact gradients rather than through a constant as in previous studies [19], [39], which thus sets up a looser restriction. Note that A1 is the fundamental assumption that will be applied to other analyses in this paper.

B. Convergence Bound of FMARL with Same τ

During the training phase of DRL, it is common to keep the learning rate η as a proper constant and decay it only when the performance saturates. Similarly, in the following discussions, we will investigate the model's error convergence bound under a fixed learning rate. We first discuss the situation where there is no difference in the number of local updates between the agents involved (i.e., $\forall i \in \{1, 2, 3, \dots, m\}$, $\tau_i = \tau$). The training algorithm of FMARL under the periodic averaging method is illustrated in Algorithm 1. By performing Algorithm 1, we suppose that the training samples obtained by agents and the learning model satisfy A1, and the total number of iterations K is large enough and divisible by τ . If the learning rate η satisfies

$$\eta L \left(\frac{\beta}{m} + 1 \right) - 1 + 2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1) \leq 0, \quad (14)$$

Algorithm 1 The training algorithm of FMARL.

Input: the model's initial parameters $\bar{\theta}_0$;

Output: the model's final average parameters $\bar{\theta}_k$;

```

1: Initialize entire environment, learning rate  $\eta$ , loss function  $\mathcal{L}$ , reward function  $\mathcal{R}$ , maximal length of an epoch  $T$ , total number of epochs for training  $U$ , maximal size of a mini-batch  $P$ , number of local updates  $\tau$  for agent  $i = 1$ , number of agents that need to transmit the model's local gradients  $m$ , and iteration index  $k$ ;

2: for epoch  $u = 1, 2, 3, \dots, U$  do
3:   for transition  $t = 0, 1, 2, \dots, T - 1$  do
4:     for agent  $i = 1, 2, 3, \dots, m$  do
5:       Calculate the input vector  $s_t$  according to the received local state from the environment;
6:       Select an action  $a_t$  according to  $\theta_k^{(i)}(s_t, a_t)$ ;
7:       Perform the selected action and receive the next state  $s_{t+1}$  from the environment;
8:       Calculate  $r_t$  according to the reward function  $\mathcal{R}$ ;
9:       Store this transition as  $\phi_t^{(i)}$ ;
10:      if  $t + 1 \bmod P == 0$  or  $t == T - 1$  then
11:        Form the mini-batch  $\xi_k^{(i)}$  by the stored transitions  $\phi_t^{(i)}$ ;
12:        Calculate the mini-batch gradients by
13:        
$$g(\theta_k^{(i)}) = \frac{1}{|\xi_k^{(i)}|} \sum_{\phi_t^{(i)} \in \xi_k^{(i)}} \nabla \mathcal{L}(\theta_k^{(i)}; \phi_t^{(i)})$$

14:        Perform the local update by
15:        
$$\theta_{k+1}^{(i)} = \theta_k^{(i)} - \eta g(\theta_k^{(i)})$$
;
16:        Clear the stored transitions  $\phi_t^{(i)}$ ;
17:        Store the mini-batch gradients  $g(\theta_k^{(i)})$ ;
18:         $k \leftarrow k + 1$ ;
19:      end if
20:      if  $k \bmod \tau == 0$  or reach the end of a period then
21:        Transmit all the accumulated gradients  $g(\theta_y^{(i)})$  to the virtual agent and receive the model's average parameters by
22:        
$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{k-1} \mathbb{I}(\tau_i > y - t_0) g(\theta_y^{(i)})$$
;
23:         $\theta_k^{(i)} \leftarrow \bar{\theta}_k$ ;
24:        Clear the stored transitions  $\phi_t^{(i)}$ ;
25:        Clear the accumulated gradients  $g(\theta_y^{(i)})$ ;
26:      end if
27:    end for
28:  end for
29: end for
30: Return the model's average parameters  $\bar{\theta}_k$ ;

```

then the expected gradient norm after K iterations is bounded by

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}_k)\|^2 \right] \leq \underbrace{\frac{2 [F(\bar{\theta}_0) - F_{\inf}]}{\eta K}}_{\psi_1^{(P)}} + \underbrace{\eta^2 L^2 \sigma^2 (\tau + 1)}_{\psi_1^{(P)}}. \quad (15)$$

(14) and (15) indicate a practical approach for (10) from the periodic averaging method. For the sake of distinction, we will use the notation $\psi_1^{(P)}$ to represent the bound under the periodic averaging method (i.e., relative to ψ_1). The corresponding proofs about (14) and (15) are provided in the Appendix B. *Remarks:* It can be observed that (14) indicates an upper bound for the learning rate η . In particular, we can get $\eta \in (0, \frac{1}{L(\tau+1)})$ when β in (14) is close to 0, which is practical when σ^2 is large [19], [39]. Since both L and τ are positive constants, general

settings of the learning rate η in DRL (e.g., 1.0×10^{-4}) could meet the requirement, and the performance impact of η will be further discussed in Section VII. Besides, $\psi_1^{(P)}$ shows that the model's error convergence bound is jointly influenced by several key parameters. Specifically, the first term of $\psi_1^{(P)}$ satisfies $\frac{2[F(\bar{\theta}_0) - F_{\inf}]}{\eta K} \geq \frac{1}{K} \sum_{k=0}^{K-1} [2\langle \nabla F(\bar{\theta}_k), \mathcal{G}_k \rangle - \eta L \|\mathcal{G}_k\|^2]$, where $\mathcal{G}_k = \frac{1}{m} \sum_{i=1}^m g(\theta_k^{(i)})$, and is affected by the calculated gradients $g(\theta_k^{(i)})$ as well as the Lipschitz constant L , which is related to the smooth properties of the empirical risk function F . Since $g(\theta_k^{(i)}) \rightarrow 0$ when $k \rightarrow K$, the first term of $\psi_1^{(P)}$ finally approaches 0. In addition, the second term of $\psi_1^{(P)}$ implies that an increase in the number of participating agents m can reduce $\psi_1^{(P)}$, but it comes at the expense of more resource cost and may reduce the system's utility value (see (7) and (13)). Furthermore, we can also conclude that although the periodic averaging method can reduce the communication overheads by many times, the increase of τ would enlarge the error bound and affect the convergence rate [40]. Note that the result in (15) is similar to those presented in existing works [19], [39], [42], but we still provide it here as the preliminary conclusion, so as to facilitate subsequent theoretical analyses.

C. Convergence Bound of FMARL with Different τ

Considering that different agents may spend various amounts of time finishing a step for each local update, we further derive the model's error convergence bound under the VPA method, where the variation comes from that the local updates from different agents are asynchronous, as illustrated in Fig. 2(b). Specifically, we make the following assumptions for the agents involved.

Assumption 2 (A2)

1. $\tau_i \in \{1, 2, 3, \dots, \tau\}$, for $i = 1, 2, 3, \dots, m$;
2. $\tau_i \geq \tau_{i+1}$, for $i = 1, 2, 3, \dots, m-1$;
3. $\sum_{i=1}^m I(\tau_i = \tau) \geq 1$;
4. $\frac{1}{m} \sum_{i=1}^m \tau_i = \bar{\tau}_i \xrightarrow{K \rightarrow \infty} \nu$;
5. $\frac{1}{m} \sum_{i=1}^m (\tau_i - \bar{\tau}_i)^2 \xrightarrow{K \rightarrow \infty} \omega^2$.

Note that Condition 1 and 2 in A2 on local updates are consistent with the previous definition of τ_i in (6) within the scope of m participating agents. Moreover, Condition 3 in A2 also conforms to the previous definition for the duration of a period, which is dynamically determined by the wall clock time required for some agent able to complete τ local updates. In addition, by Condition 4 and 5 in A2, we suppose that the numbers of local updates from m participating agents have the mean and variance value (i.e., ν and ω^2) when the total number of iterations K is large enough, so as to facilitate the convergence analysis.

Under the VPA method, the update rule of $\bar{\theta}_k$ at the virtual agent is determined by (11). Similarly, the update rule of $\theta_k^{(i)}$ within a period at each agent is determined by

$$\theta_k^{(i)} = \bar{\theta}_{t_0} - \eta \sum_{y=t_0}^{k-1} I(\tau_i > y - t_0) g(\theta_y^{(i)}). \quad (16)$$

The training algorithm of FMARL under the VPA method can be also illustrated by Algorithm 1. By performing Algorithm 1, we suppose that the learning model and the involved agents satisfy A1 and A2, while the total number of iterations K is large enough and divisible by τ . If the learning rate η satisfies (14), then the expected gradient norm after K iterations is bounded by

$$\begin{aligned} \mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}_k)\|^2 \right] &\leq \underbrace{\frac{2[F(\bar{\theta}_0) - F_{\inf}]}{\eta K}}_{\psi_1^{(V)}} + \frac{\eta L \sigma^2}{m} \\ &\quad + \underbrace{\frac{\eta^2 L^2 \sigma^2}{\tau} [-\nu^2 + (2\tau + 1)\nu - \omega^2]}_{\psi_1^{(V)}}. \end{aligned} \quad (17)$$

The corresponding proofs are provided in the Appendix C.

Remarks: Compared with the error convergence bound $\psi_1^{(P)}$ in (15), we can observe that the third term of $\psi_1^{(V)}$ in (17) reveals more details. Specifically, with the value of τ unchanged, the part in square brackets can be regarded as a quadratic function of the mean value ν . Furthermore, we can find that the maximal value of this quadratic function would be reached at $\nu = \tau + 1/2$. Since A2 ensures $1 < \nu \leq \tau$, we can conclude that the error convergence bound will increase monotonically as the mean value ν goes up, which is similar to the effect brought by the increase of τ . We can also observe that an increase in the variance value ω^2 can reduce the error convergence bound $\psi_1^{(V)}$. Besides, if $\nu = \tau$ and $\omega = 0$, the VPA method will reduce to the classical periodic averaging method. Compared with the convergence bound obtained in [25], the conclusion in (17) further considers the effect of the mean and variance value of τ on convergence performance. Next, we will take this VPA method as the basis of subsequent optimization methods.

VI. CONVERGENCE BOUND OF SGD CONSIDERING COMMUNICATION EFFICIENCY

A. Decay-based Method

For a single agent, since there may be a deviation between the distribution of obtained samples and that of real inputs, the direction of SGD can thus contain error. Through averaging the model's local gradients in FL, this error can be reduced by boosting many gradient descent directions from different agents. However, the frequency of this averaging process is greatly reduced in the periodic averaging method, and the error of SGD will be superposed continuously with the progress of local updates, resulting in the variance of subsequent local gradients gradually increasing. To solve this problem, we take advantage of a decay function to gradually decrease the weights of successive local gradients within each period, which satisfies the following assumptions.

Assumption 3 (A3)

1. $D(y) = D(y + \tau)$, and $y \in \mathbb{N}$;
2. $1 = D(y = t_0) \geq D(y = t_0 + 1) \geq D(y = t_0 + 2) \geq \dots \geq D(y = t_0 + \tau - 1) \geq 0$.

It can be observed that the decay function $D(y)$ is defined as a discrete periodic function, and is monotonically decreasing over a period of length τ . Under the decay-based method, $D(y)$ is used to decay the weight of mini-batch gradients when updating the model's parameters. Therefore, the update rules of $\theta_k^{(i)}$ and $\bar{\theta}_k$ over a period of length τ are expressed as

$$\theta_k^{(i)} = \bar{\theta}_{t_0} - \eta \sum_{y=t_0}^{k-1} \mathbb{I}(\tau_i > y - t_0) D(y) g(\theta_y^{(i)}), \quad (18)$$

$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{k-1} \mathbb{I}(\tau_i > y - t_0) D(y) g(\theta_y^{(i)}). \quad (19)$$

The training algorithm of FMARL under the decay-based method is almost the same as that illustrated in Algorithm 1, except that a common decay function needs to be provided for each agent in advance and the update rules of the model's parameters should follow (18) and (19). Then, we can draw the following theorem.

Theorem 1 (T1) Suppose the number of local updates for agent i is τ_i , and the model's training process follows Algorithm 1, where agents update their parameters according to (18) and (19). Under A1, A2, and A3, if the total number of iterations K is large enough and divisible by τ , and the learning rate η satisfies (14), then the upper bound of the expected gradient norm after K iterations satisfies

$$\psi_1^{(D)} \leq \psi_1^{(V)}. \quad (20)$$

where $\psi_1^{(D)}$ indicates the bound for the decay-based VPA method. The corresponding proofs are provided in the Appendix D.

Remarks: The gap between $\psi_1^{(D)}$ and $\psi_1^{(V)}$ depends on the applied decay function. On the other hand, considering the resource cost under the decay-based method, we can find that the proposed method can reduce the error convergence bound while maintaining the resource cost (i.e., ψ_0) unchanged, thus improving the system's utility value.

To further characterize the proposed decay-based method, we provide an example to highlight its advantages. In particular, we concretely define $D(y)$ as an exponential function, which is expressed by

$$D(y) := \lambda^{\frac{y}{\tau}}, \quad (21)$$

where $\lambda \in (0, 1]$ is a decay constant. To facilitate the convergence analysis, we also assume that the numbers of local updates from m participating agents are uniformly distributed within the domain (i.e., $\Pr(\tau_i = \tau_0) = 1/\tau$, for $\tau_0 \in \{1, 2, 3, \dots, \tau\}$). This will be practical when m is large and local agents are very diversified. Therefore, according to Condition 4 and 5 in A2, we can easily get $\nu = (1 + \tau)/2$ and $\omega^2 = (\tau - 1)^2/12$. We further can get the following corollary.

Corollary 1 (C1) Under the conditions in T1, suppose the numbers of agents' local updates are uniformly distributed within the domain, while the decay function is defined as (21). If the learning rate η satisfies (14), then the expected gradient

norm after K iterations is bounded by¹

$$\begin{aligned} \mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}_k)\|^2 \right] &\leq \underbrace{\frac{2[F(\bar{\theta}_0) - F_{\inf}]}{\eta K} + \frac{\eta L \sigma^2}{m}}_{\psi_1^{(P)}} \\ &+ \underbrace{\frac{2\eta^2 L^2 \sigma^2}{\tau} \left[\frac{\tau}{1-\lambda} - \frac{2\lambda}{(1-\lambda)^2} + \frac{\lambda(\lambda+1)(1-\lambda^\tau)}{\tau(1-\lambda)^3} \right]}_{\psi_1^{(D)}}. \end{aligned} \quad (22)$$

The corresponding proofs about (22) are provided in the Appendix E.

Remarks: It can be observed that with the value of τ unchanged, the part in square brackets of the third term of $\psi_1^{(D)}$ in (22) can be regarded as a function of λ . Through calculating its first-order derivative, we can figure out that this function is monotonically increasing as the value of λ increases. Moreover, the bound $\psi_1^{(D)}$ obtained in C1 will be equal to $\psi_1^{(P)}$ in (15) when $\lambda \rightarrow 0$ and thus $\tau \rightarrow 1$. And it will also approach the bound obtained in (17) when $\lambda \rightarrow 1$ and thus $D(y) \rightarrow 1$. By adjusting λ in this decay-based method, we can indirectly control the maximal number of samples that agents can employ in a period for training. In practical applications, λ can be set to a value that is slightly less than 1, such as $\lambda = 0.98$.

B. Consensus-based Method

To take full advantage of the multi-agent collective collaboration, we further propose to employ the consensus algorithm [20] to improve the local update of each agent through the D2D communication among nearby collaborators. Since the original objective of the consensus algorithm is to make all the distributed nodes in an ad-hoc network reach a consensus, it can be used to reduce the variance of the mini-batch gradients from a cluster of agents. Under the consensus-based method, we will use the notation $g(\theta_k^{(i)}, e)$ to distinguish different local interactions of agents towards optimizing $g(\theta_k^{(i)})$, where e denotes the index of local interactions, and $g(\theta_k^{(i)}, 0) = g(\theta_k^{(i)})$. In addition, to enable all the participating agents to reach a consensus successfully, we make the following assumption for the network of participating agents.

Assumption 4 (A4)

- The network of participating agents with topology G is a strongly connected undirected graph.

Note that the graph with undirected connections indicates that all the involved agents affect each other equally. Then, according to the consensus algorithm [20], we can obtain the following local interaction process of each agent

$$g(\theta_k^{(i)}, e+1) = g(\theta_k^{(i)}, e) + \epsilon \sum_{l \in \Omega_i} \left[g(\theta_k^{(l)}, e) - g(\theta_k^{(i)}, e) \right], \quad (23)$$

where Ω_i represents the neighbors set of agent i . Considering the network of participating agents, Ω_i can be regarded as the

¹For sake of simplicity of representation, we slightly abuse the notation of $\psi_1^{(D)}$ for both a general decay-based method and a specific exponential function-driven one.

set of the agents that are directly connected with agent i . ϵ denotes the local interaction step size, which plays a similar role as the learning rate η . Besides, $0 < \epsilon < 1/\Delta$, where Δ denotes the maximal degree of the graph and is defined by $\Delta := \max_i |\Omega_i| + 1$. Furthermore, the update rule of $\theta_k^{(i)}$ and $\bar{\theta}_k$ under the consensus-based method can be modified as

$$\theta_k^{(i)} = \bar{\theta}_{t_0} - \eta \sum_{y=t_0}^{k-1} g(\theta_y^{(i)}, e); \quad (24)$$

$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{k-1} g(\theta_y^{(i)}, e). \quad (25)$$

The training algorithm of FMARL under the consensus-based method is illustrated in Algorithm 2. An intuitive schematic of the model's training process is presented in Fig. 2(c). It can be observed that agents need to exchange their local gradients with nearby collaborators before performing the local update, and these exchanged gradients are also used to update the model's average parameters. Note that the update rules in (24) and (25) also conform to A2, since the delay in obtaining local gradients can only affect the value of $g(\theta_k^{(i)}, 0)$, which equals to 0 for some agents at the beginning of local interaction. Thus, although local interactions occur synchronously between neighboring agents, it is not necessary for an agent to wait for all the neighbors to finish a step in Algorithm 2, so as to reduce the training time. Furthermore, we can obtain the following theorem.

Theorem 2 (T2) Suppose the number of local updates for agent i is τ_i , and the model's training process follows Algorithm 2. Under A1, A2, and A4, if the total number of iterations K is large enough and divisible by τ , and the learning rate η satisfies (14), then the expected gradient norm after K iterations is bounded by

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}_k)\|^2 \right] \leq \underbrace{\frac{2[F(\bar{\theta}_0) - F_{\inf}]}{\eta K}}_{+ \eta^2 \sigma^2 L^2 (\tau + 1) [1 - \epsilon \mu_2(\mathbf{La})]^{2E}} + \underbrace{\frac{\eta L \sigma^2}{m}}_{\psi_1^{(C)}}, \quad (26)$$

where E represents the total number of local interactions before each local update. \mathbf{La} denotes the Laplace matrix of the graph G . $\mu_2(\mathbf{La})$ denotes the second smallest eigenvalue of \mathbf{La} , which is also called *algebraic connectivity* [20]. The corresponding proofs are presented in the Appendix F.

Remarks: Compared with the bound $\psi_1^{(P)}$ obtained in (15), we can find that the error convergence bound $\psi_1^{(C)}$ under the consensus-based method is additionally affected by the topological properties (i.e., algebraic connectivity) of the graph comprised by agents and their connections. Thus, it is completely different from the consensus-based FL algorithm in [21], which utilizes an upper bound on the spectral radius of a stochastic matrix for interaction. Moreover, the conclusion in (26) further considers the effect of local interaction step size ϵ , which can be properly adjusted to guarantee the policy improvement of DRL [49]. In addition, since $0 < \mu_2(\mathbf{La}) \leq \Delta$ and the equality holds only when G is a fully connected graph,

we can find that $0 < 1 - \epsilon \mu_2(\mathbf{La}) < 1$. Therefore, we can infer that the implementation of local interactions can reduce the error convergence bound greatly. In addition, a larger step size ϵ or a more densely connected network of agents can also help reduce this bound. In practical applications, a small number of local interactions can make this bound decrease dramatically, such as $E = 2$.

Algorithm 2 The training algorithm of FMARL under the consensus-based method.

Input: the model's initial parameters $\bar{\theta}_0$;
Output: the model's final average parameters $\bar{\theta}_k$;

- 1: **Initialize** entire environment, learning rate η , loss function \mathcal{L} , reward function \mathcal{R} , maximal length of an epoch T , total number of epochs for training U , maximal size of a mini-batch P , number of local updates τ for agent $i = 1$, number of agents that need to transmit the model's local gradients m , iteration index k , step size ϵ , and total number of local interactions E ;
- 2: **for** epoch $u = 1, 2, 3, \dots, U$ **do**
- 3: **for** transition $t = 0, 1, 2, \dots, T - 1$ **do**
- 4: **for** agent $i = 1, 2, 3, \dots, m$ **do**
- 5: Calculate the input vector s_t according to the received local state from the environment;
- 6: Select an action a_t according to $\theta_k^{(i)}(s_t, a_t)$;
- 7: Perform the selected action and receive the next state s_{t+1} from the environment;
- 8: Calculate r_t according to the reward function \mathcal{R} ;
- 9: Store this transition as $\phi_t^{(i)}$;
- 10: **if** $t + 1 \bmod P == 0$ **or** $t == T - 1$ **then**
- 11: Form the mini-batch $\xi_k^{(i)}$ by the stored transitions $\phi_t^{(i)}$;
- 12: Calculate the mini-batch gradients by

$$g(\theta_k^{(i)}) = \frac{1}{|\xi_k^{(i)}|} \sum_{\phi_t^{(i)} \in \xi_k^{(i)}} \nabla \mathcal{L}(\theta_k^{(i)}; \phi_t^{(i)});$$
- 13: $g(\theta_k^{(i)}, 0) \leftarrow g(\theta_k^{(i)})$;
- 14: Wait for other agents;
- 15: **for** interaction $e = 0, 1, 2, \dots, E - 1$ **do**
- 16: $g(\theta_k^{(i)}, e + 1) = g(\theta_k^{(i)}, e) + \epsilon \sum_{l \in \Omega_i} [g(\theta_k^{(l)}, e) - g(\theta_k^{(i)}, e)]$;
- 17: **end for**
- 18: Perform the local update by

$$\theta_{k+1}^{(i)} = \theta_k^{(i)} - \eta g(\theta_k^{(i)}, E);$$
- 19: Clear the stored transitions $\phi_t^{(i)}$;
- 20: Store the mini-batch gradients $g(\theta_k^{(i)}, E)$;
- 21: $k \leftarrow k + 1$;
- 22: **else**
- 23: $g(\theta_k^{(i)}, 0) \leftarrow \mathbf{0}$;
- 24: **end if**
- 25: **if** $k \bmod \tau == 0$ **or** reach the end of a period **then**
- 26: Transmit all the accumulated gradients $g(\theta_y^{(i)}, E)$ to the virtual agent and receive the model's average parameters by

$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{k-1} g(\theta_y^{(i)}, E);$$
- 27: $\theta_k^{(i)} \leftarrow \bar{\theta}_k$;
- 28: Clear the stored transitions $\phi_t^{(i)}$;
- 29: Clear the accumulated gradients $g(\theta_y^{(i)}, E)$;
- 30: **end if**
- 31: **end for**
- 32: **end for**
- 33: **end for**
- 34: **Return** the model's average parameters $\bar{\theta}_k$;

On the other hand, regarding the resource cost under the consensus-based method, we assume that the extra communication overheads required for an agent to exchange the mini-

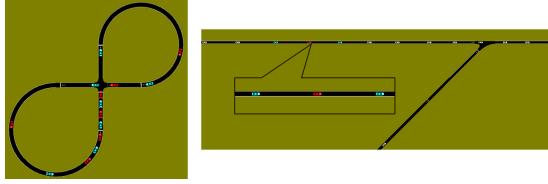


Fig. 3. Figure Eight (left) and Merge (right).

batch gradients with one of its neighbors are W_1 , and the computation overheads required to perform a local interaction are W_2 . Then the system's resource cost can be reformulated from (7) as

$$\psi_0^{(C)} = \sum_{i=1}^m \left[\frac{C_1 TU}{\tau P} + \frac{C_2 \tau_i TU}{\tau P} + |\Omega_i| (W_1 + W_2) \frac{ETU}{P} \right], \quad (27)$$

where $|\Omega_i|$ denotes the size of Ω_i . Here, the possible collision or interference issues in actual communication are omitted. Compared with ψ_0 defined in (7), it can be observed from (27) that the additional local interactions under the consensus-based method increase the resource cost. However, since the model's error convergence bound is reduced at the same time, the system's utility value may be improved in certain cases. For example, suppose that the overheads required for participating agents to implement the D2D communication are much less than those to transmit the same messages to the remote virtual agent, the effectiveness of the consensus-based method would be clearly demonstrated. In addition, participating agents can intentionally set up as a sparse graph (i.e., smaller $|\Omega_i|$) to decrease the number of local interactions between agents, so as to further reduce the resource cost.

VII. SIMULATION RESULTS

In this section, we provide the simulation results about the two proposed optimization methods. The simulation scenarios are taken from [57], which is released as a new benchmark in traffic control through creating DRL controllers for mixed-autonomy traffic. As illustrated in Fig. 3, two simulation scenarios in this benchmark are selected to verify the effectiveness and efficiency of the developed methods. For the first scenario ‘‘Figure Eight’’, there are totally 14 vehicles running circularly along a one-way lane that resembles the shape of ‘‘8’’. An intersection is located at the lane, and each vehicle must control its acceleration to pass through this intersection, so as to increase the average speed of the whole vehicle team. Note that slamming on the brakes will be forced on the vehicles that are about to crash, and the current epoch will be terminated once the collision occurs. We further modify the ‘‘Figure Eight’’ scenario to assign the related local state to each vehicle, including its position and speed as well as those of the vehicle ahead and behind. Depending on its local state, each vehicle needs to optimize its acceleration, which is a normalized continuous variable between -1 and 1 . All the involved vehicles are the same, and unless additional controllers are assigned, these vehicles are controlled by the underlying simulation of urban mobility

TABLE III
MAIN PARAMETERS IN THE SIMULATIONS.

Parameters	Value
Number of agents, m	7 (or 5)
Learning rate, η	10^{-4}
Length of epoch, T	1500
Number of epochs, U	500
Mini-batch size, P	250
Number of iterations, K	3000
Maximal local updates, τ	15
Step size in local interaction, ϵ	0.1
Discount factor in PPO [8]	0.9
Clipping parameter in PPO	0.2
Epoch parameter in PPO	4

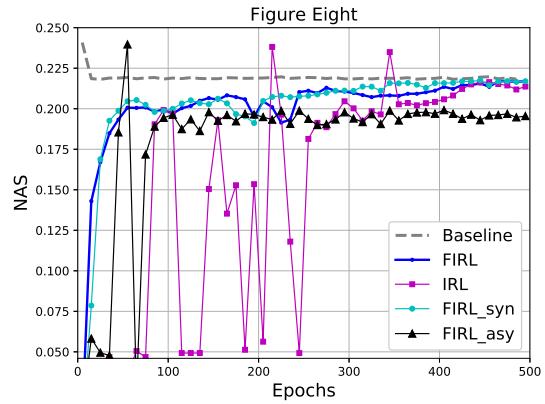


Fig. 4. Effectiveness of FL in IRL.

(SUMO) in the same mode, which is a widely used and open-source traffic simulation package [57]. In the ‘‘Figure Eight’’ scenario, half of the vehicles are empowered by the DRL-based controllers.

The simulation settings of the second scenario ‘‘Merge’’ are almost the same as those of the first scenario, except that the maximal speed and acceleration of each vehicle are increased. The ‘‘Merge’’ scenario simulates the intersection of a highway and a side lane, and each vehicle needs to control its acceleration to increase the average speed. Moreover, there are totally 50 vehicles in the second scenario, and 5 vehicles are randomly selected within each epoch to instantiate the DRL-based controllers. Note that we take the normalized average speed (NAS) of all vehicles at each iteration as the individual reward in each scenario, which is assigned to each training vehicle after its action being performed. Unless otherwise specified, the DRL-based controllers for each scenario are optimized through the PPO algorithm [8]. Main parameters in the simulations are listed in Table III.

In Fig. 4, we first present the simulation performance of naive IRL and FIRL (IRL with FL). Here, NAS denotes the normalized average speed of all vehicles during an entire testing epoch. The test is performed every 10 epochs and takes the average of five repetitions. We further take the performance of all vehicles controlled by the underlying SUMO as the optimal baseline. Fig. 4 shows that FL can clearly improve the performance of IRL in terms of training efficiency and stability, while verify the necessity of this

combination. Meanwhile, we test the effect of delay on FIRL in the method `FIRL_syn` and `FIRL_asy`. Specifically, both methods introduce the delay of 2 epochs (i.e., 300 seconds) to the up-link and down-link between each agent and the virtual agent. However, agents in `FIRL_syn` can update their parameters every iteration, while agents in `FIRL_asy` can only update their parameters every 2 epochs, which implies a worsening communication. As indicated in Fig. 4, a smaller updating frequency may reduce the effectiveness of FL in IRL.

Besides, we provide the performance of FIRL and `FIRL_D` with $\tau = 1 \sim 15$, $\lambda = 0.92$ under different η in Fig. 5. Note that the total number of policy iterations $K = UT/P$ and K is proportional to the number of epochs U . We can observe from Figs. 5(a) and (b) that with a proper η (e.g., $\eta = 10^{-3}$), the performance gradually converges with the increase of K . However, an over-trivial η may slow the convergence rate, as shown by the curves with $\eta = 10^{-5}$. On the other hand, since the model's gradients are calculated by the loss function in DRL, an overlarge step size of η would suffer from the catastrophic forgetting problem and make the model hard to converge [58], as shown by the curves with $\eta = 10^{-2}$. Therefore, the selection of η should balance the trade-off between convergence rate and convergence performance.

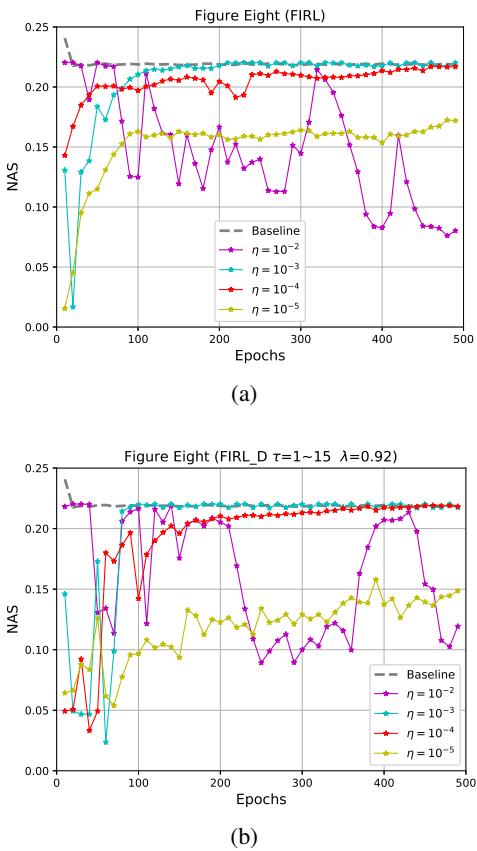


Fig. 5. Performance of (a) FIRL and (b) FIRL_D under different η .

In Fig. 6(a), we present the performance of FIRL with different local updates in a period. Here, the notation “ $\tau = 1 \sim 15$ ” denotes that the numbers of local updates from agents are uniformly distributed between 1 and 15, thus $\nu = 8$.

We set the maximal value of τ to 15, so as to highlight the performance improvement of the proposed methods on the classical periodic averaging method. It can be observed from Fig. 6(a) that the performance will decline as the number of local updates τ or its mean value ν increases, which is consistent with the conclusion in (17). In Fig. 6(b), we present the performance of FIRL under the decay-based method (i.e., `FIRL_D`), which realizes the practical implementation in C1 when the numbers of local updates from agents are uniformly distributed between 1 and 15. Fig. 6(b) demonstrates that `FIRL_D` can improve the model's convergence performance, and a smaller decay constant λ generates a faster convergence rate, which are consistent with our previous discussions in T1 and C1. In Fig. 6(c), we present the performance of FIRL under the consensus-based method (i.e., `FIRL_C`). On the premise of strong connectivity, the topology of agents' network with the algebraic connectivity $\mu_2 = 1.4384$ is constructed by 3 ~ 4 random connections from each agent to nearby collaborators, while these connections are increased to 4 ~ 6 when $\mu_2 = 2.5188$. Note that there is only one connection between any pair of agents. It can be observed from Fig. 6(c) that the model's convergence performance is improved when the local interactions are considered, even when various numbers of local updates are introduced, which are consistent with the conclusions in T2. In addition, the topology with either a larger μ_2 (i.e., more connections) or a greater E (i.e., more local interactions) generally obtains better performance. In Figs. 6(d) - (f), we further present the performance of `FIRL_C` in the “Merge” scenario under different policy gradient methods. Specifically, since vehicles in the “Merge” scenario are running along a highway, we construct the topology of agents' network by connecting any two adjacent DRL-based vehicles, thus $\mu_2 = 0.3820$. Besides, we use the trust region policy optimization (TRPO) method [7] and the Tsallis actor-critic (TAC) algorithm [9] as the corresponding policy gradient method in Fig. 6(e) and Fig. 6(f), respectively. We can observe from Figs. 6(d) - (f) that `FIRL_C` can stabilize the model's training process, meanwhile performs well across different policy gradient methods.

In Fig. 7, we select some representative methods in Figs. 4 and 6 to indicate the ratio of NAS to communication overhead (CO) with respect to epochs. Here, the ratio of NAS to CO denotes the performance improvement per unit of communication overhead and indicates the communication efficiency, where the parameter C_1 is set to 1 and W_1 is set to $1.0 \times 10^{-3}C_1$. A more detailed description about the communication overheads of these methods is provided in Table IV. It can be observed from Fig. 7(a) that compared with FIRL with $\tau = 1$, $\tau = 10$, and $\tau = 1 \sim 15$, `FIRL_D` and `FIRL_C` have higher starting points, which means their performance (i.e., NAS) increases faster when the communication overheads are small. Since the value of CO under this parameter setting is much greater than that of NAS, the methods with the same maximal range of τ converge to a similar level, such as FIRL with $\tau = 1 \sim 15$, `FIRL_D`, and `FIRL_C`. However, Fig. 7(b), a magnified image of the red box in Fig. 7(a), clearly shows that the proposed method `FIRL_D` and `FIRL_C` maintain better performance.

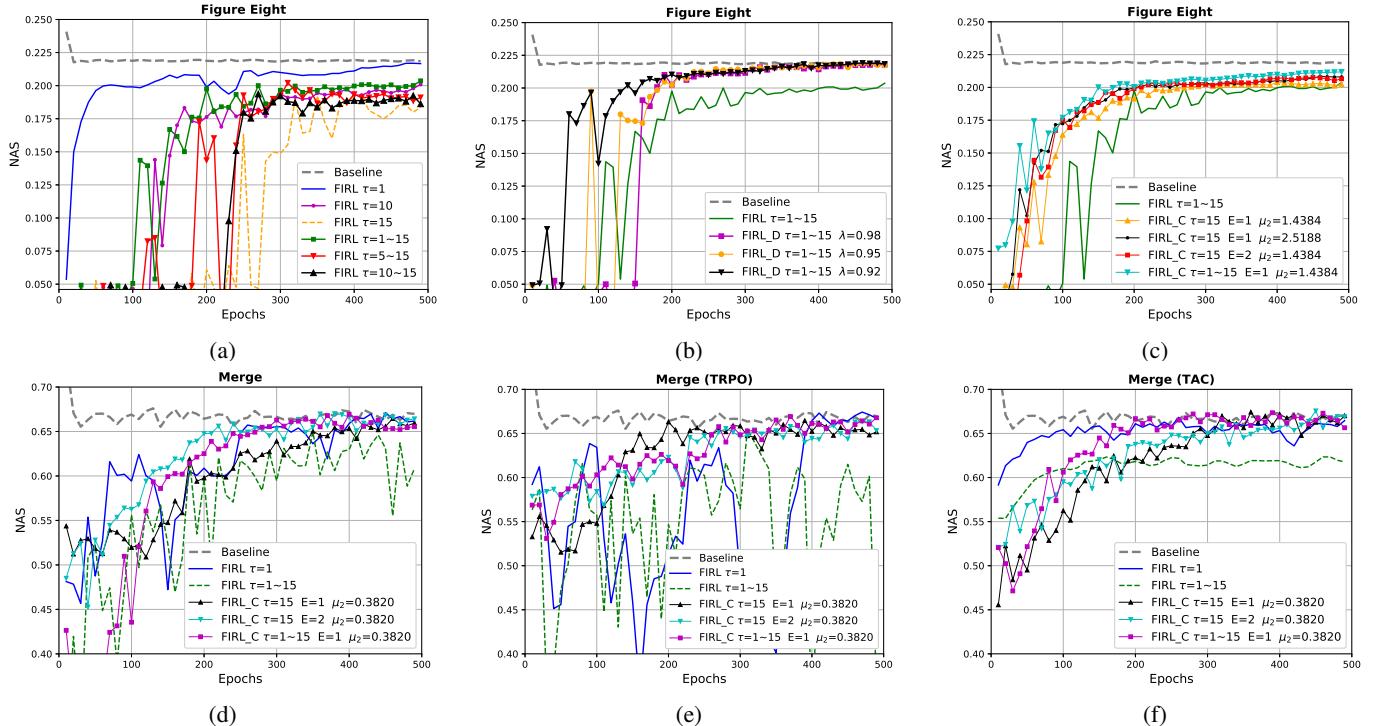


Fig. 6. Convergence performance of (a) the VPA method, (b) the decay-based method, (c) the consensus-based method, (d) the PPO algorithm, (e) the TRPO algorithm, and (f) the TAC algorithm.

TABLE IV
NUMERICAL SIMULATION RESULTS IN THE “FIGURE EIGHT” SCENARIO.

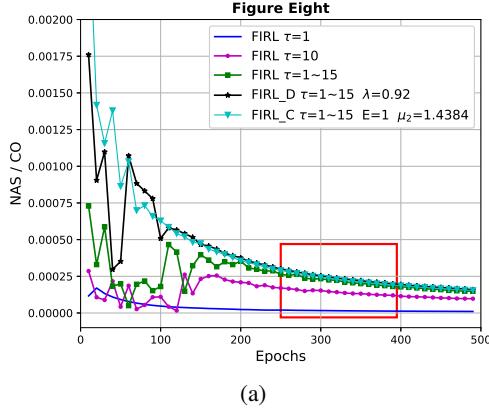
Methods	Local updates	Decay-based	Consensus-based	Communication overheads	Computation overheads	Expected gradient norm	Normalized utility value
FIRL	$\tau = 1$	None	None	$21000 C_1$	$21000 C_2$	1.5590	0.0
FIRL	$\tau = 10$	None	None	$2100 C_1$	$21000 C_2$	6.3421	0.5734
FIRL	$\tau = 15$	None	None	$1400 C_1$	$21000 C_2$	9.6069	0.7809
FIRL	$\tau = 10 \sim 15$	None	None	$1400 C_1$	$19000 C_2$	10.1892	0.7601
FIRL	$\tau = 5 \sim 15$	None	None	$1400 C_1$	$16200 C_2$	8.7182	0.8130
FIRL	$\tau = 1 \sim 15$	None	None	$1400 C_1$	$12600 C_2$	7.6476	0.8516
FIRL_D	$\tau = 1 \sim 15$	$\lambda = 0.98$	None	$1400 C_1$	$12600 C_2$	7.2782	0.8649
FIRL_D	$\tau = 1 \sim 15$	$\lambda = 0.95$	None	$1400 C_1$	$12600 C_2$	7.2537	0.8657
FIRL_D	$\tau = 1 \sim 15$	$\lambda = 0.92$	None	$1400 C_1$	$12600 C_2$	3.5090	1.0
FIRL_C	$\tau = 15$	None	$E = 1, \mu_2 = 1.4384$	$1400 C_1 + 78000 W_1$	$21000 C_2 + 78000 W_2$	3.6188	0.9336
FIRL_C	$\tau = 15$	None	$E = 1, \mu_2 = 2.5188$	$1400 C_1 + 96000 W_1$	$21000 C_2 + 96000 W_2$	2.1648	0.9688
FIRL_C	$\tau = 15$	None	$E = 2, \mu_2 = 1.4384$	$1400 C_1 + 156000 W_1$	$21000 C_2 + 156000 W_2$	2.8746	0.9023
FIRL_C	$\tau = 1 \sim 15$	None	$E = 1, \mu_2 = 1.4384$	$1400 C_1 + 78000 W_1$	$12600 C_2 + 78000 W_2$	3.6072	0.9346

In Table IV, we summarize the numerical simulation results in the “Figure Eight” scenario, so as to verify the effectiveness of the developed methods in improving the system’s utility value. In particular, the expected gradient norm is calculated upon a predetermined sample set, which is comprised by samples that are uniformly collected during the model’s training process when $\tau = 1$. Moreover, the expected gradient norm is calculated whenever the model’s average parameters (i.e., $\bar{\theta}_k$) are updated, and its final value is averaged across the entire training process. In Table IV, with the same communication and computation overheads as FIRL with $\tau = 1 \sim 15$, FIRL_D maintains a smaller expected gradient norm, thus having a greater system’s utility value. Besides, compared with FIRL with $\tau = 15$ or $\tau = 1 \sim 15$, FIRL_C requires the same amount of resource cost in FL, and although there are more resources required in local interactions, its expected gradient norm is smaller. Thus, the system’s utility value can

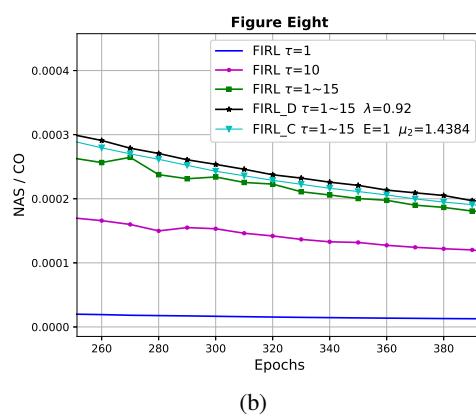
be also improved when W_1 and W_2 are small in FIRL_C. As an example, we give the system’s normalized utility value in Table IV under the conditions $C_1 = 1$, $W_1 = 1.0 \times 10^{-3}C_1$, $C_2 = 1.0 \times 10^{-4}C_1$, and $W_2 = C_2$. We can observe that compared with FIRL, FIRL_D and FIRL_C maintain higher utility values. Specifically, the utility value of FIRL_C will increase as the computation overheads decrease.

VIII. CONCLUSIONS

This paper has taken advantage of the paradigm of FL to improve the policy performance of IRL agents. Meanwhile, considering the excessive communication overheads generated between agents and a central server in FL as well as the heterogeneity of independent learning environments faced by IRL agents, this paper has built the framework of FMARL on the basis of the VPA method. Moreover, to reach a good balance between the system’s resource cost and the model’s



(a)



(b)

Fig. 7. (a) Ratio of NAS to communication overhead (CO) with respect to epochs in the “Figure Eight” scenario, and (b) magnified image in the red box of (a).

convergence performance, a novel utility function has been proposed to quantify the convergence bound of the model’s error reduction per unit of resource cost. Furthermore, to improve the system’s utility value, we have put forward two new optimization methods on top of the VPA method. By analyzing the theoretical convergence bounds and performing extensive simulations, both effectiveness and efficiency of the developed methods have been verified.

For future works, we plan to implement the proposed optimization methods in the real-world applications. In practice, when there are a large number of participating agents, multiple virtual agents may emerge simultaneously and their organization tends to be hierarchical, which means a more complex scenario and requires more careful considerations on the optimization methods. Moreover, we also plan to optimize the aggregation weights for agents with different sample sizes in FL, which promises an effective method to improve the system’s performance. Besides, it is also interesting to develop more smart means to determine the hyperparameter (e.g., η). Finally, whether the common shared learning model can be applied to heterogeneous tasks faced by different agents remains an open question that is worth exploring.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grants 62071425, in

part by the Zhejiang Key Research and Development Plan under Grant 2022C01093, in part by Huawei Cooperation Project, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR23F010005.

APPENDIX A PROOF PRELIMINARIES

In this section, we introduce some of the definitions and notations used across the Supplemental Material. In addition, we also introduce several important lemmas, on which the proofs of the proposed theorems are built. In particular, we define the average mini-batch gradients

$$\mathcal{G}_k = \frac{1}{m} \sum_{i=1}^m g(\theta_k^{(i)}), \quad (28)$$

and the average full-batch gradients

$$\mathcal{H}_k = \frac{1}{m} \sum_{i=1}^m \nabla F(\theta_k^{(i)}). \quad (29)$$

According to Condition 3 in A1, we can obtain

$$\mathbb{E}[\mathcal{G}_k] = \mathcal{H}_k. \quad (30)$$

In addition, we define the Frobenius norm for $\mathbf{A} \in M_n$ by

$$\|\mathbf{A}\|_F^2 = |\text{Tr}(\mathbf{A}\mathbf{A}^\top)| = \sum_{i=1}^n \sum_{j=1}^n |a_{i,j}|^2. \quad (31)$$

And the operator norm for $\mathbf{A} \in M_n$ is defined by

$$\|\mathbf{A}\|_{\text{op}} = \max_{\|x\|=1} \|\mathbf{A}x\| = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}, \quad (32)$$

where λ_{\max} denotes the maximal eigenvalue. Then we can infer that for real matrix $\mathbf{A} \in \mathbb{R}^{d \times m}$ and $\mathbf{B} \in \mathbb{R}^{m \times m}$, if \mathbf{B} is symmetric, then the following inequality holds

$$\|\mathbf{AB}\|_F \leq \|\mathbf{B}\|_{\text{op}} \|\mathbf{A}\|_F. \quad (33)$$

Besides, we use the notation $\mathbf{1}$ to represent the vector $[1, 1, 1, \dots, 1]_{1 \times m}^\top$, and $\mathbf{J} := \mathbf{1}\mathbf{1}^\top / (\mathbf{1}^\top \mathbf{1})$.

Lemma 1. *Under Condition 3 and 4 in A1, the variance of the average mini-batch gradients is bounded by*

$$\mathbb{E} \|\mathcal{G}_k - \mathcal{H}_k\|^2 \leq \frac{\beta}{m^2} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \frac{\sigma^2}{m}. \quad (34)$$

Proof. According to (28) and (29), we have

$$\begin{aligned} \mathbb{E} \|\mathcal{G}_k - \mathcal{H}_k\|^2 &= \mathbb{E} \left\| \frac{1}{m} \sum_{i=1}^m [g(\theta_k^{(i)}) - \nabla F(\theta_k^{(i)})] \right\|^2 \\ &= \frac{1}{m^2} \sum_{i=1}^m \mathbb{E} \left\| g(\theta_k^{(i)}) - \nabla F(\theta_k^{(i)}) \right\|^2 \\ &\quad + \frac{1}{m^2} \mathbb{E} \left[\sum_{s \neq i}^m \langle g(\theta_k^{(s)}) - \nabla F(\theta_k^{(s)}), g(\theta_k^{(i)}) - \nabla F(\theta_k^{(i)}) \rangle \right] \\ &= \frac{1}{m^2} \sum_{i=1}^m \mathbb{E} \left\| g(\theta_k^{(i)}) - \nabla F(\theta_k^{(i)}) \right\|^2 \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{m^2} \sum_{s \neq l}^m \left\langle \mathbb{E}_{\xi_k^{(s)} | \theta_k^{(s)}} \left[g(\theta_k^{(s)}) - \nabla F(\theta_k^{(s)}) \right], \mathbb{E}_{\xi_k^{(l)} | \theta_k^{(l)}} \left[g(\theta_k^{(l)}) - \nabla F(\theta_k^{(l)}) \right] \right\rangle \\
& \stackrel{(a)}{=} \frac{1}{m^2} \sum_{i=1}^m \mathbb{E} \left\| g(\theta_k^{(i)}) - \nabla F(\theta_k^{(i)}) \right\|^2 \\
& \stackrel{(b)}{\leq} \frac{1}{m^2} \sum_{i=1}^m \left[\beta \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \sigma^2 \right] \\
& = \frac{\beta}{m^2} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \frac{\sigma^2}{m},
\end{aligned}$$

where $\{\xi_k^{(s)}\}$ and $\{\xi_k^{(l)}\}$ are independent random variables. The equality (a) is due to that according to Condition 3 in A1, $\mathbb{E}_{\xi_k^{(s)} | \theta_k^{(s)}} \left[g(\theta_k^{(s)}) - \nabla F(\theta_k^{(s)}) \right]$ and $\mathbb{E}_{\xi_k^{(l)} | \theta_k^{(l)}} \left[g(\theta_k^{(l)}) - \nabla F(\theta_k^{(l)}) \right]$ turn to be 0. The inequality (b) comes from Condition 4 in A1. \square

Lemma 2. Under Condition 3 in A1, the expected inner product between the average mini-batch gradients and the full-batch gradients satisfies

$$\begin{aligned}
& \mathbb{E} [\langle \nabla F(\bar{\theta}_k), \mathcal{G}_k \rangle] \\
& = \frac{1}{2} \left\| \nabla F(\bar{\theta}_k) \right\|^2 + \frac{1}{2m} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\
& \quad - \frac{1}{2m} \sum_{i=1}^m \left\| \nabla F(\bar{\theta}_k) - \nabla F(\theta_k^{(i)}) \right\|^2.
\end{aligned} \tag{35}$$

Proof. According to the definition of \mathcal{G}_k , we have

$$\begin{aligned}
& \mathbb{E} [\langle \nabla F(\bar{\theta}_k), \mathcal{G}_k \rangle] = \mathbb{E} \left[\left\langle \nabla F(\bar{\theta}_k), \frac{1}{m} \sum_{i=1}^m g(\theta_k^{(i)}) \right\rangle \right] \\
& \stackrel{(a)}{=} \left\langle \nabla F(\bar{\theta}_k), \frac{1}{m} \sum_{i=1}^m \nabla F(\theta_k^{(i)}) \right\rangle \\
& = \frac{1}{m} \sum_{i=1}^m \langle \nabla F(\bar{\theta}_k), \nabla F(\theta_k^{(i)}) \rangle \\
& \stackrel{(b)}{=} \frac{1}{2m} \sum_{i=1}^m \left[\left\| \nabla F(\bar{\theta}_k) \right\|^2 + \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \right. \\
& \quad \left. - \left\| \nabla F(\bar{\theta}_k) - \nabla F(\theta_k^{(i)}) \right\|^2 \right] \\
& = \frac{1}{2} \left\| \nabla F(\bar{\theta}_k) \right\|^2 + \frac{1}{2m} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\
& \quad - \frac{1}{2m} \sum_{i=1}^m \left\| \nabla F(\bar{\theta}_k) - \nabla F(\theta_k^{(i)}) \right\|^2,
\end{aligned}$$

where the equality (a) comes from Condition 3 in A1, and the equality (b) is due to $\mathbf{a}^\top \mathbf{b} = \frac{1}{2} (\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2)$. \square

Lemma 3. Under Condition 3 and 4 in A1, the squared norm

of the average mini-batch gradients is bounded by

$$\mathbb{E} \|\mathcal{G}_k\|^2 \leq \left(\frac{\beta}{m^2} + \frac{1}{m} \right) \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \frac{\sigma^2}{m}. \tag{36}$$

Proof. According to the definition of \mathcal{G}_k , we have

$$\begin{aligned}
\mathbb{E} \|\mathcal{G}_k\|^2 & = \mathbb{E} \|\mathcal{G}_k - \mathbb{E}[\mathcal{G}_k]\|^2 + \|\mathbb{E}[\mathcal{G}_k]\|^2 \\
& = \mathbb{E} \|\mathcal{G}_k - \mathcal{H}_k\|^2 + \|\mathcal{H}_k\|^2 \\
& \stackrel{(a)}{\leq} \frac{\beta}{m^2} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \frac{\sigma^2}{m} + \left\| \frac{1}{m} \sum_{i=1}^m \nabla F(\theta_k^{(i)}) \right\|^2 \\
& \stackrel{(b)}{\leq} \frac{\beta}{m^2} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \frac{\sigma^2}{m} + \frac{1}{m} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\
& = \left(\frac{\beta}{m^2} + \frac{1}{m} \right) \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \frac{\sigma^2}{m},
\end{aligned}$$

where the inequality (a) is due to Lemma 1, and the inequality (b) comes from the convexity of the vector norm and Jensen's inequality

$$\left\| \sum_{i=1}^m \mathbf{a}_i \right\|^2 \leq m \sum_{i=1}^m \|\mathbf{a}_i\|^2.$$

\square

Lemma 4. Under A1, if the total number of iterations K is large enough, and the learning rate η satisfies

$$\eta L \left(\frac{\beta}{m} + 1 \right) - 1 \leq 0, \tag{37}$$

then the expected gradient norm after K iterations is bounded by

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \left\| \nabla F(\bar{\theta}_k) \right\|^2 \right] & \leq \frac{2 [F(\bar{\theta}_0) - F_{\inf}]}{\eta K} + \frac{\eta L \sigma^2}{m} \\
& + \underbrace{\frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2}_{(a)}.
\end{aligned} \tag{38}$$

Proof. According to the Lipschitz continuous gradient assumption, we obtain

$$\begin{aligned}
& \mathbb{E} [F(\bar{\theta}_{k+1})] - \mathbb{E} [F(\bar{\theta}_k)] \\
& \leq \mathbb{E} [\langle \nabla F(\bar{\theta}_k), \bar{\theta}_{k+1} - \bar{\theta}_k \rangle] + \frac{L}{2} \mathbb{E} \left\| \bar{\theta}_{k+1} - \bar{\theta}_k \right\|^2 \\
& \stackrel{(a)}{=} -\eta \mathbb{E} [\langle \nabla F(\bar{\theta}_k), \mathcal{G}_k \rangle] + \frac{\eta^2 L}{2} \mathbb{E} \|\mathcal{G}_k\|^2 \\
& \stackrel{(b)}{\leq} -\frac{\eta}{2} \left\| \nabla F(\bar{\theta}_k) \right\|^2 - \frac{\eta}{2m} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\
& \quad + \frac{\eta}{2m} \sum_{i=1}^m \left\| \nabla F(\bar{\theta}_k) - \nabla F(\theta_k^{(i)}) \right\|^2 \\
& \quad + \frac{\eta^2 L}{2} \left(\frac{\beta}{m^2} + \frac{1}{m} \right) \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \frac{\eta^2 \sigma^2 L}{2m},
\end{aligned} \tag{39}$$

where the equality (a) comes from $\bar{\theta}_{k+1} = \bar{\theta}_k - \eta \mathcal{G}_k$, and the inequality (b) is based on Lemma 2 and 3. We apply Condition 1 in A1 to (39) and rearrange the expression to get

$$\begin{aligned} \|\nabla F(\bar{\theta}_k)\|^2 &\leq \frac{2[\mathbb{E}[F(\bar{\theta}_k)] - \mathbb{E}[F(\bar{\theta}_{k+1})]]}{m} + \frac{\eta L \sigma^2}{m} \\ &+ \left[\eta L \left(\frac{\beta}{m^2} + \frac{1}{m} \right) - \frac{1}{m} \right] \sum_{i=1}^m \|\nabla F(\theta_k^{(i)})\|^2 \\ &+ \frac{L^2}{m} \sum_{i=1}^m \|\bar{\theta}_k - \theta_k^{(i)}\|^2. \end{aligned}$$

By superposing and averaging the expression over all iterations K , and based on Condition 2 in A1, we obtain

$$\begin{aligned} \mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}_k)\|^2 \right] &\leq \frac{2[F(\bar{\theta}_0) - F_{\inf}]}{\eta K} + \frac{\eta L \sigma^2}{m} \\ &+ \left[\eta L \left(\frac{\beta}{m} + 1 \right) - 1 \right] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \quad (40) \\ &+ \frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2. \end{aligned}$$

According to (40), if $\eta L \left(\frac{\beta}{m} + 1 \right) - 1 \leq 0$, then

$$\begin{aligned} \mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}_k)\|^2 \right] &\leq \frac{2[F(\bar{\theta}_0) - F_{\inf}]}{\eta K} + \frac{\eta L \sigma^2}{m} \\ &+ \frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2. \end{aligned}$$

have

$$\begin{aligned} &\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \\ &= \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} g(\theta_y^{(i)}) - \sum_{y=t_0}^{t_0+j-1} \frac{1}{m} \sum_{i=1}^m g(\theta_y^{(i)}) \right\|^2 \\ &= \eta^2 \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} G_y - \sum_{y=t_0}^{t_0+j-1} G_y \mathbf{J} \right\|_{\mathbf{F}}^2 \\ &= \eta^2 \mathbb{E} \|Y_j - Y_j \mathbf{J}\|_{\mathbf{F}}^2 \\ &= \eta^2 \mathbb{E} \|Y_j(\mathbf{I} - \mathbf{J})\|_{\mathbf{F}}^2 \\ &\stackrel{(a)}{\leq} \eta^2 \mathbb{E} \|Y_j\|_{\mathbf{F}}^2 \|\mathbf{I} - \mathbf{J}\|_{\text{op}}^2 \\ &= \eta^2 \mathbb{E} \|Y_j\|_{\mathbf{F}}^2, \quad (43) \end{aligned}$$

where the inequality (a) comes from (33). Based on (43), we obtain

$$\begin{aligned} &\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} g(\theta_y^{(i)}) \right\|^2 \\ &= \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} (g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)})) + \sum_{y=t_0}^{t_0+j-1} \nabla F(\theta_y^{(i)}) \right\|^2 \\ &\stackrel{(a)}{\leq} 2\eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} (g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)})) \right\|^2 \\ &+ 2\eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \nabla F(\theta_y^{(i)}) \right\|^2 \\ &\stackrel{(b)}{\leq} 2\eta^2 \sum_{i=1}^m \left[\sum_{y=t_0}^{t_0+j-1} \mathbb{E} \left\| g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)}) \right\|^2 \right. \\ &\quad \left. + \sum_{y \neq q} \mathbb{E} \left\langle g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)}), g(\theta_q^{(i)}) - \nabla F(\theta_q^{(i)}) \right\rangle \right] \\ &+ 2\eta^2 j \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{E} \left\| \nabla F(\theta_y^{(i)}) \right\|^2 \\ &\stackrel{(c)}{\leq} 2\eta^2 \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \left(\beta \left\| \nabla F(\theta_y^{(i)}) \right\|^2 + \sigma^2 \right) \\ &+ 2\eta^2 j \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \left\| \nabla F(\theta_y^{(i)}) \right\|^2, \end{aligned}$$

where the inequality (a) and (b) come from Jensen's inequality, and according to Condition 3 in A1, the term $\sum_{y \neq q} \mathbb{E} \left\langle g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)}), g(\theta_q^{(i)}) - \nabla F(\theta_q^{(i)}) \right\rangle$ turns to be 0. The inequality (c) comes from Condition 4 in A1. By superposing the expression over a period of τ , we get

$$\begin{aligned} &\sum_{j=1}^{\tau} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq \eta^2 \sigma^2 m \tau (\tau + 1) \\ &+ [2\eta^2 \tau \beta + \eta^2 \tau (\tau + 1)] \sum_{i=1}^m \sum_{y=t_0}^{t_0+\tau-1} \left\| \nabla F(\theta_y^{(i)}) \right\|^2. \end{aligned}$$

By substituting (41) and (42) into $\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$, we

$$\theta_k^{(i)} = \bar{\theta}_{t_0} - \eta \sum_{y=t_0}^{t_0+j-1} g(\theta_y^{(i)}); \quad (41)$$

$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} g(\theta_y^{(i)}). \quad (42)$$

By further superposing the expression over all iterations, we get

$$\begin{aligned} & \sum_{t_0=0}^{K-\tau} \sum_{j=1}^{\tau} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq K \eta^2 \sigma^2 m (\tau + 1) \\ & + [2\eta^2 \tau \beta + \eta^2 \tau (\tau + 1)] \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2. \end{aligned}$$

Note that $k = t_0 + j$ and $t_0 = z\tau$, where $z \in \mathbb{N}$. Finally, we get

$$\begin{aligned} & \frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq \eta^2 \sigma^2 L^2 (\tau + 1) \\ & + [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2. \end{aligned}$$

By substituting the above inequality into Lemma 4, we obtain

$$\begin{aligned} \mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \left\| \nabla F(\bar{\theta}_k) \right\|^2 \right] & \leq \frac{2 [F(\bar{\theta}_0) - F_{\inf}]}{\eta K} + \frac{\eta L \sigma^2}{m} \\ & + \eta^2 \sigma^2 L^2 (\tau + 1) + [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1) \\ & + \eta L \left(\frac{\beta}{m} + 1 \right) - 1] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \nabla F(\theta_k^{(i)}) \right\|^2. \end{aligned}$$

If $2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1) + \eta L \left(\frac{\beta}{m} + 1 \right) - 1 \leq 0$, then we can get the following conclusion

$$\begin{aligned} \mathbb{E} \left[\frac{1}{K} \sum_{k=0}^{K-1} \left\| \nabla F(\bar{\theta}_k) \right\|^2 \right] & \leq \frac{2 [F(\bar{\theta}_0) - F_{\inf}]}{\eta K} + \frac{\eta L \sigma^2}{m} \\ & + \eta^2 \sigma^2 L^2 (\tau + 1). \end{aligned}$$

□

APPENDIX C

PROOF OF THE RESULT IN SECTION V-C (I.E., (17))

Since the conclusion is also based on A1, we can find that the additional A2 only affects $\sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$ while keeping the other terms in the conclusion of Lemma 4 unchanged. Therefore, we directly begin our proofs from Lemma 4 to provide the corresponding bound for $\frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$. In particular, we define the m -dimensional vector $M_y := [\mathbf{I}(\tau_1 > y - t_0), \mathbf{I}(\tau_2 > y - t_0), \mathbf{I}(\tau_3 > y - t_0), \dots, \mathbf{I}(\tau_m > y - t_0)]$, and the matrix $P_y := M_y \odot G_y$, where \odot denotes the Hadamard product. Moreover, for $k = t_0 + j$, the summation of P_y over a length of j can be expressed by $Q_j := \sum_{y=t_0}^{t_0+j-1} P_y$.

Proof. Since the update rule of the model's parameters in a period is expressed by

$$\theta_k^{(i)} = \bar{\theta}_{t_0} - \eta \sum_{y=t_0}^{t_0+j-1} \mathbf{I}(\tau_i > y - t_0) g(\theta_y^{(i)}); \quad (44)$$

$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbf{I}(\tau_i > y - t_0) g(\theta_y^{(i)}). \quad (45)$$

By substituting (44) and (45) into $\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$, we obtain

$$\begin{aligned} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 & = \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbf{I}(\tau_i > y - t_0) g(\theta_y^{(i)}) \right\|^2 \\ & - \sum_{y=t_0}^{t_0+j-1} \frac{1}{m} \sum_{i=1}^m \mathbf{I}(\tau_i > y - t_0) g(\theta_y^{(i)}) \right\|^2 \\ & = \eta^2 \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} P_y - \sum_{y=t_0}^{t_0+j-1} P_y \mathbf{J} \right\|_{\mathbf{F}}^2 \\ & = \eta^2 \mathbb{E} \|Q_j - Q_j \mathbf{J}\|_{\mathbf{F}}^2 \\ & = \eta^2 \mathbb{E} \|Q_j (\mathbf{I} - \mathbf{J})\|_{\mathbf{F}}^2 \\ & \leq \eta^2 \mathbb{E} \|Q_j\|_{\mathbf{F}}^2 \|\mathbf{I} - \mathbf{J}\|_{\text{op}}^2 \\ & = \eta^2 \mathbb{E} \|Q_j\|_{\mathbf{F}}^2 \\ & = \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbf{I}(\tau_i > y - t_0) g(\theta_y^{(i)}) \right\|^2 \\ & = \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbf{I}(\tau_i > y - t_0) (g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)})) \right. \\ & \left. + \sum_{y=t_0}^{t_0+j-1} \mathbf{I}(\tau_i > y - t_0) \nabla F(\theta_y^{(i)}) \right\|^2 \\ & \leq 2\eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbf{I}(\tau_i > y - t_0) (g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)})) \right\|^2 \\ & + 2\eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbf{I}(\tau_i > y - t_0) \nabla F(\theta_y^{(i)}) \right\|^2 \\ & \leq 2\eta^2 \sum_{i=1}^m \left[\sum_{y=t_0}^{t_0+j-1} \mathbb{E} \left\| \mathbf{I}(\tau_i > y - t_0) (g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)})) \right\|^2 \right. \\ & \left. + \sum_{y \neq q} \mathbb{E} \left\langle \mathbf{I}(\tau_i > y - t_0) (g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)})) \right. \right. \\ & \left. \left. I(\tau_i > q - t_0) (g(\theta_q^{(i)}) - \nabla F(\theta_q^{(i)})) \right\rangle \right] \\ & + 2\eta^2 j \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{E} \left\| \mathbf{I}(\tau_i > y - t_0) \nabla F(\theta_y^{(i)}) \right\|^2. \end{aligned} \quad (46)$$

Since A2 does not affect Condition 3 in A1, the term $\sum_{y \neq q} \mathbb{E} \left\langle \mathbf{I}(\tau_i > y - t_0) (g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)})) \right. I(\tau_i > q - t_0) (g(\theta_q^{(i)}) - \nabla F(\theta_q^{(i)})) \right\rangle$ turns to be 0. Moreover, by applying Condition 4 in A1 to (46), we obtain

$$\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$$

$$\begin{aligned} & \underbrace{\sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) \left[\beta \left\| \nabla F(\theta_y^{(i)}) \right\|^2 + \sigma^2 \right]}_{(a)} \\ & + \underbrace{2\eta^2 j \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) \left\| \nabla F(\theta_y^{(i)}) \right\|^2}_{(b)}. \end{aligned} \quad (47)$$

By superposing the expression over a period of τ , we find

$$\begin{aligned} \sum_{j=1}^{\tau} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 & \leq 2\eta^2 \sigma^2 \sum_{i=1}^m \sum_{j=1}^{\tau} \min\{\tau_i, j\} + \\ & [2\eta^2 \tau \beta + \eta^2 \tau (\tau + 1)] \sum_{i=1}^m \sum_{y=t_0}^{t_0+\tau-1} \mathbb{I}(\tau_i > y - t_0) \left\| \nabla F(\theta_y^{(i)}) \right\|^2. \end{aligned}$$

By further superposing the expression over all iterations, we finally obtain

$$\begin{aligned} & \sum_{t_0=0}^{K-\tau} \sum_{j=1}^{\tau} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq [2\eta^2 \tau \beta + \eta^2 \tau (\tau + 1)] \\ & \sum_{i=1}^m \sum_{t_0=0}^{K-\tau} \sum_{y=t_0}^{t_0+\tau-1} \mathbb{I}(\tau_i > y - t_0) \left\| \nabla F(\theta_y^{(i)}) \right\|^2 \\ & + 2\eta^2 \sigma^2 \sum_{i=1}^m \sum_{t_0=0}^{K-\tau} \sum_{j=1}^{\tau} \min\{\tau_i, j\} \\ & = [2\eta^2 \tau \beta + \eta^2 \tau (\tau + 1)] \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{I}(\tau_i > k \bmod \tau) \\ & \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + 2\eta^2 \sigma^2 \frac{K}{\tau} \sum_{i=1}^m \sum_{j=1}^{\tau} \min\{\tau_i, j\}. \end{aligned}$$

Therefore, the term (a) of (38) in Lemma 4 can be bounded by

$$\begin{aligned} & \frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \\ & \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{I}(\tau_i > k \bmod \tau) \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\ & + \frac{1}{m\tau} 2\eta^2 L^2 \sigma^2 \sum_{i=1}^m \sum_{j=1}^{\tau} \min\{\tau_i, j\} \\ & \leq [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\ & + \underbrace{\frac{\eta^2 L^2 \sigma^2}{\tau} \frac{1}{m} \sum_{i=1}^m (\tau_i + 2\tau \tau_i - \tau_i^2)}_{(a)}. \end{aligned} \quad (48)$$

Based on Condition 4 and 5 in A2, we can estimate the expectation of the term (a) in (48) when the total number of iterations K is large enough. We have

$$\frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$$

$$\begin{aligned} & \leq [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\ & + \frac{\eta^2 \sigma^2 L^2}{\tau} [-\nu^2 + (2\tau + 1)\nu - w^2]. \end{aligned}$$

By substituting the above inequality into the term (a) of (38) in Lemma 4, we prove the conclusion. \square

APPENDIX D PROOF OF THEOREM 1

According to A3, we define the $d \times m$ -dimensional matrix $G'_y := [D(y)g(\theta_y^{(1)}), D(y)g(\theta_y^{(2)}), D(y)g(\theta_y^{(3)}), \dots, D(y)g(\theta_y^{(m)})]$, and $P'_y := M_y \odot G'_y$. Then, the summation of P'_y over a length of j can be expressed by $Q'_j := \sum_{y=t_0}^{t_0+j-1} P'_y$. Since the conclusion in Theorem 1 is based on A1, we begin our proofs directly from providing the bound of $\frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$ in Lemma 4.

Proof. Since the update rule of the model's parameters in a period can be expressed by

$$\theta_k^{(i)} = \bar{\theta}_{t_0} - \eta \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D(y) g(\theta_y^{(i)}); \quad (49)$$

$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D(y) g(\theta_y^{(i)}). \quad (50)$$

By substituting (49) and (50) into $\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$, we obtain

$$\begin{aligned} & \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \\ & = \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D(y) g(\theta_y^{(i)}) \right. \\ & \quad \left. - \sum_{y=t_0}^{t_0+j-1} \frac{1}{m} \sum_{i=1}^m \mathbb{I}(\tau_i > y - t_0) D(y) g(\theta_y^{(i)}) \right\|^2 \\ & = \eta^2 \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} P'_y - \sum_{y=t_0}^{t_0+j-1} P'_y \mathbf{J} \right\|_{\mathbb{F}}^2 \\ & = \eta^2 \mathbb{E} \left\| Q'_j - Q'_j \mathbf{J} \right\|_{\mathbb{F}}^2 \\ & = \eta^2 \mathbb{E} \left\| Q'_j (\mathbf{I} - \mathbf{J}) \right\|_{\mathbb{F}}^2 \\ & \leq \eta^2 \mathbb{E} \left\| Q'_j \right\|_{\mathbb{F}}^2 \left\| \mathbf{I} - \mathbf{J} \right\|_{\text{op}}^2 \\ & = \eta^2 \mathbb{E} \left\| Q'_j \right\|_{\mathbb{F}}^2 \\ & = \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D(y) g(\theta_y^{(i)}) \right\|^2 \\ & = \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D(y) (g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)})) \right\|^2 \end{aligned}$$

$$\begin{aligned}
& + \sum_{y=t_0}^{t_0+j-1} \mathbb{E} \left\| \sum_{i=1}^m \mathbb{I}(\tau_i > y - t_0) D(y) \nabla F(\theta_y^{(i)}) \right\|^2 \\
& \leq 2\eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D(y) \right. \\
& \quad \left. \left(g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)}) \right) \right\|^2 \\
& + 2\eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D(y) \nabla F(\theta_y^{(i)}) \right\|^2 \\
& \leq 2\eta^2 \sum_{i=1}^m \left[\sum_{y=t_0}^{t_0+j-1} \mathbb{E} \left\| \mathbb{I}(\tau_i > y - t_0) D(y) \right. \right. \\
& \quad \left. \left. \left(g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)}) \right) \right\|^2 \right. \\
& \quad \left. + \sum_{y \neq q} \mathbb{E} \left\langle \mathbb{I}(\tau_i > y - t_0) D(y) \left(g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)}) \right), \right. \right. \\
& \quad \left. \left. \mathbb{I}(\tau_i > q - t_0) D(q) \left(g(\theta_q^{(i)}) - \nabla F(\theta_q^{(i)}) \right) \right\rangle \right] \\
& + 2\eta^2 j \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{E} \left\| \mathbb{I}(\tau_i > y - t_0) D(y) \nabla F(\theta_y^{(i)}) \right\|^2. \tag{51}
\end{aligned}$$

Since $D(y)$ is independent from the mini-batch gradients and according to Condition 3 in A1, we can find that the term $\sum_{y \neq q} \mathbb{E} \left\langle \mathbb{I}(\tau_i > y - t_0) D(y) \left(g(\theta_y^{(i)}) - \nabla F(\theta_y^{(i)}) \right), \mathbb{I}(\tau_i > q - t_0) D(q) \left(g(\theta_q^{(i)}) - \nabla F(\theta_q^{(i)}) \right) \right\rangle$ turns to be 0. By further applying Condition 4 in A1 to (51), we can find

$$\begin{aligned}
& \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \\
& \leq 2\eta^2 \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D^2(y) \left[\beta \left\| \nabla F(\theta_y^{(i)}) \right\|^2 + \sigma^2 \right] \\
& + 2\eta^2 j \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D^2(y) \left\| \nabla F(\theta_y^{(i)}) \right\|^2 \\
& \stackrel{(a)}{\leq} 2\eta^2 \underbrace{\sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) \left[\beta \left\| \nabla F(\theta_y^{(i)}) \right\|^2 + \sigma^2 \right]}_{(a)} \\
& + 2\eta^2 j \underbrace{\sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) \left\| \nabla F(\theta_y^{(i)}) \right\|^2}_{(b)}, \tag{52}
\end{aligned}$$

where the inequality (a) is due to $D^2(y) \leq 1$. Notice that the terms (a) and (b) in (52) are the same as those in (47), which comprise the bound of the variation-aware periodic averaging method. Therefore, the model's error convergence bound under the decay-based method is decreased, and Theorem 1 is proved. \square

APPENDIX E PROOF OF COROLLARY 1

Since the conclusion is given as an example of the decay-based method, we can directly begin our proofs on the basis of the proof of Theorem 1. We define the summation of $D^2(y)$ over a length of j as $Z(j) := \sum_{y=t_0}^{t_0+j-1} D^2(y)$, then $Z(j) \leq j$. According to the definition of $D(y)$, we can get $Z(j) = \frac{1-\lambda^j}{1-\lambda}$.

Proof. According to the proof of Theorem 1, we obtain

$$\begin{aligned}
& \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \\
& \leq 2\eta^2 \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D^2(y) \left[\beta \left\| \nabla F(\theta_y^{(i)}) \right\|^2 + \sigma^2 \right] \\
& + 2\eta^2 j \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D^2(y) \left\| \nabla F(\theta_y^{(i)}) \right\|^2.
\end{aligned}$$

By superposing the expression over a period of τ , we find

$$\begin{aligned}
& \sum_{j=1}^\tau \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq [2\eta^2 \tau \beta + \eta^2 \tau (\tau + 1)] \\
& \sum_{i=1}^m \sum_{y=t_0}^{t_0+\tau-1} \mathbb{I}(\tau_i > y - t_0) D^2(y) \left\| \nabla F(\theta_y^{(i)}) \right\|^2 \\
& + 2\eta^2 \sigma^2 \sum_{i=1}^m \sum_{j=1}^\tau \sum_{y=t_0}^{t_0+j-1} \mathbb{I}(\tau_i > y - t_0) D^2(y) \\
& = [2\eta^2 \tau \beta + \eta^2 \tau (\tau + 1)] \sum_{i=1}^m \sum_{y=t_0}^{t_0+\tau-1} \mathbb{I}(\tau_i > y - t_0) \\
& D^2(y) \left\| \nabla F(\theta_y^{(i)}) \right\|^2 + 2\eta^2 \sigma^2 \sum_{i=1}^m \sum_{j=1}^\tau \min \{Z(\tau_i), Z(j)\}.
\end{aligned}$$

By further superposing the expression over all iterations, we finally obtain

$$\begin{aligned}
& \sum_{t_0=0}^{K-\tau} \sum_{j=1}^\tau \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq [2\eta^2 \tau \beta + \eta^2 \tau (\tau + 1)] \\
& \sum_{i=1}^m \sum_{t_0=0}^{K-\tau} \sum_{y=t_0}^{t_0+\tau-1} \mathbb{I}(\tau_i > y - t_0) D^2(y) \left\| \nabla F(\theta_y^{(i)}) \right\|^2 \\
& + 2\eta^2 \sigma^2 \frac{K}{\tau} \sum_{i=1}^m \sum_{j=1}^\tau \min \{Z(\tau_i), Z(j)\}.
\end{aligned}$$

Therefore, the bound of the term (a) of (38) in Lemma 4 can be expressed by

$$\begin{aligned}
& \frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \\
& \leq [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\
& + \frac{2\eta^2 L^2 \sigma^2}{m\tau} \sum_{i=1}^m \sum_{j=1}^\tau \min \{Z(\tau_i), Z(j)\}
\end{aligned}$$

$$\begin{aligned}
&= [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\
&\quad + \frac{2\eta^2 L^2 \sigma^2}{m\tau} \sum_{i=1}^m \left[\sum_{j=1}^{\tau_i} Z(j) + (\tau - \tau_i) Z(\tau_i) \right] \\
&= [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 + \\
&\quad \frac{2\eta^2 L^2 \sigma^2}{\tau} \underbrace{\frac{1}{m} \sum_{i=1}^m \left[\frac{(1-\lambda)\tau_i - \lambda(1-\lambda^{\tau_i})}{(1-\lambda)^2} + (\tau - \tau_i) \frac{1-\lambda^{\tau_i}}{1-\lambda} \right]}_{(a)}.
\end{aligned} \tag{53}$$

Next, we estimate the expectation of the term (a) in (53) when the total number of iterations K is large enough. Since we suppose that the numbers of local updates from different agents are uniformly distributed across the domain, that is $\frac{1}{m} \sum_{i=1}^m \tau_i \xrightarrow{K \rightarrow \infty} \frac{1+\tau}{2}$, we thus can get $\frac{1}{m} \sum_{i=1}^m \lambda^{\tau_i} \xrightarrow{K \rightarrow \infty} \frac{\lambda(1-\lambda^\tau)}{\tau(1-\lambda)}$, and $\frac{1}{m} \sum_{i=1}^m \tau_i \lambda^{\tau_i} \xrightarrow{K \rightarrow \infty} \frac{\lambda(1-\lambda^\tau)}{\tau(1-\lambda)^2} - \frac{\lambda^{\tau+1}}{1-\lambda}$. By substituting these conclusions into (53), we get

$$\begin{aligned}
&\frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \\
&\leq [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\
&\quad + \frac{2\eta^2 L^2 \sigma^2}{\tau} \left[\frac{\tau}{1-\lambda} - \frac{2\lambda}{(1-\lambda)^2} + \frac{\lambda(1+\lambda)(1-\lambda^\tau)}{\tau(1-\lambda)^3} \right].
\end{aligned}$$

By substituting the above inequality into the term (a) of (38) in Lemma 4, we get the conclusion. \square

APPENDIX F PROOF OF THEOREM 2

Since the conclusion in Theorem 2 is based on A1 and A2, we directly begin our proofs from presenting the bound of $\frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$ in Lemma 4. Similarly, we define the $d \times m$ -dimensional matrix $G_{y,e} := [g(\theta_y^{(1)}, e), g(\theta_y^{(2)}, e), g(\theta_y^{(3)}, e), \dots, g(\theta_y^{(m)}, e)]$, and $Y_{j,e} := \sum_{y=t_0}^{t_0+j-1} G_{y,e}$. According to the consensus algorithm, we can infer that $G_{y,0} = G_y$, and $G_{y,e+1} = G_{y,e} \mathbf{P}$. Furthermore, we can conclude that $G_{y,e} = G_y \mathbf{P}^e$ and $Y_{j,e} = Y_j \mathbf{P}^e$, where e denotes the power exponent. Here, the matrix $\mathbf{P} := \mathbf{I} - \epsilon \mathbf{La}$. For the network of agents, the Laplace matrix \mathbf{La} is determined by

$$a_{i,l} = \begin{cases} |\Omega_i|, & i = l; \\ -1, & i \neq l, l \in \Omega_i; \\ 0, & \text{otherwise}, \end{cases} \tag{54}$$

where $a_{i,l}$ is the element in \mathbf{La} . In addition, we define μ as the eigenvalue of \mathbf{La} . Then, according to A4, we can obtain that $\mu_{min}(\mathbf{La}) = 0$, and the number of eigenvalues that are equal to 0 is 1.

Proof. Since the update rule of the model's parameters in a

period can be expressed by

$$\bar{\theta}_k^{(i)} = \bar{\theta}_{t_0} - \eta \sum_{y=t_0}^{t_0+j-1} g(\theta_y^{(i)}, e); \tag{55}$$

$$\bar{\theta}_k = \bar{\theta}_{t_0} - \eta \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} g(\theta_y^{(i)}, e). \tag{56}$$

By substituting (55) and (56) into $\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2$, we obtain

$$\begin{aligned}
&\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \\
&= \eta^2 \sum_{i=1}^m \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} g(\theta_y^{(i)}, e) - \frac{1}{m} \sum_{i=1}^m \sum_{y=t_0}^{t_0+j-1} g(\theta_y^{(i)}, e) \right\|^2 \\
&= \eta^2 \mathbb{E} \left\| \sum_{y=t_0}^{t_0+j-1} G_{y,e} - \sum_{y=t_0}^{t_0+j-1} G_{y,e} \mathbf{J} \right\|_{\mathbf{F}}^2 \\
&= \eta^2 \mathbb{E} \|Y_{j,e} - Y_{j,e} \mathbf{J}\|_{\mathbf{F}}^2 \\
&= \eta^2 \mathbb{E} \|Y_{j,e}(\mathbf{I} - \mathbf{J})\|_{\mathbf{F}}^2 \\
&= \eta^2 \mathbb{E} \|Y_j \mathbf{P}^e(\mathbf{I} - \mathbf{J})\|_{\mathbf{F}}^2 \\
&\leq \eta^2 \mathbb{E} \|Y_j\|_{\mathbf{F}}^2 \|\mathbf{P}^e(\mathbf{I} - \mathbf{J})\|_{\text{op}}^2.
\end{aligned} \tag{57}$$

According to the operator norm defined in (32), we find

$$\begin{aligned}
\|\mathbf{P}^e(\mathbf{I} - \mathbf{J})\|_{\text{op}}^2 &= \lambda_{\max} [(\mathbf{P}^e(\mathbf{I} - \mathbf{J}))^\top (\mathbf{P}^e(\mathbf{I} - \mathbf{J}))] \tag{58} \\
&= \lambda_{\max} [(\mathbf{I} - \mathbf{J}) \mathbf{P}^{2e} (\mathbf{I} - \mathbf{J})]. \tag{59}
\end{aligned}$$

We define ζ as the eigenvalue of \mathbf{P} . Since \mathbf{P} is a real symmetric matrix, it thus can be decomposed by $\mathbf{P} = \mathbf{Q} \Lambda_P \mathbf{Q}^\top$, where \mathbf{Q} is an orthogonal matrix and $\text{diag}(\Lambda_P) = \{\zeta_1, \zeta_2, \zeta_3, \dots, \zeta_m\}$. Similarly, we also have $\mathbf{I} - \mathbf{J} = \mathbf{Q} \Lambda_0 \mathbf{Q}^\top$, and $\text{diag}(\Lambda_0) = \{0, 1, 1, \dots, 1\}$. Then, we can infer that

$$(\mathbf{I} - \mathbf{J}) \mathbf{P}^{2e} (\mathbf{I} - \mathbf{J}) = \mathbf{Q} \Lambda \mathbf{Q}^\top, \tag{60}$$

where $\text{diag}(\Lambda) = \{0, \zeta_2^{2e}, \zeta_3^{2e}, \dots, \zeta_m^{2e}\}$. Since $\mathbf{P} = \mathbf{I} - \epsilon \mathbf{La}$, the corresponding eigenvalue ζ and μ satisfy

$$\zeta_i = 1 - \epsilon \mu_i, \tag{61}$$

where i denotes the index of eigenvalue. Since \mathbf{La} is a real symmetric matrix, $\mu_i \geq 0$. According to A4, $\mu_{min} = 0$, thus $\zeta_{max} = 1$ and $\zeta_i \leq 1$. Besides, since $\zeta_1 \geq \zeta_2 \geq \dots \geq \zeta_m$, we can observe that $\zeta_1 = 1$, and $\lambda_{\max} [(\mathbf{I} - \mathbf{J}) \mathbf{P}^{2e} (\mathbf{I} - \mathbf{J})] = \zeta_2^{2e}$. Moreover, we can obtain

$$\|\mathbf{P}^e(\mathbf{I} - \mathbf{J})\|_{\text{op}}^2 = [1 - \epsilon \mu_2(\mathbf{La})]^{2e}, \tag{62}$$

where μ_2 is the second smallest eigenvalue of \mathbf{La} . By substituting (62) into (57), we can further obtain

$$\sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq \underbrace{\eta^2 \mathbb{E} \|Y_j\|_{\mathbf{F}}^2}_{(a)} [1 - \epsilon \mu_2(\mathbf{La})]^{2e}. \tag{63}$$

We can observe that the term (a) of (63) also emerges in (43)

in the proof of Appendix B. Therefore, we can directly get

$$\begin{aligned} & \frac{L^2}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \mathbb{E} \left\| \bar{\theta}_k - \theta_k^{(i)} \right\|^2 \leq \eta^2 \sigma^2 L^2 (\tau + 1) [1 - \epsilon \mu_2(\mathbf{La})]^{2e} \\ & + [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] [1 - \epsilon \mu_2(\mathbf{La})]^{2e} \\ & \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2 \\ & \leq \eta^2 \sigma^2 L^2 (\tau + 1) [1 - \epsilon \mu_2(\mathbf{La})]^{2e} \\ & + [2\eta^2 L^2 \tau \beta + \eta^2 L^2 \tau (\tau + 1)] \frac{1}{mK} \sum_{k=0}^{K-1} \sum_{i=1}^m \left\| \nabla F(\theta_k^{(i)}) \right\|^2. \end{aligned}$$

By substituting the above equality into the term (a) of (38) in Lemma 4, we get the conclusion in Theorem 2. \square

REFERENCES

- [1] A. Ata, M. A. Khan, S. Abbas, M. S. Khan, and G. Ahmad, "Adaptive IoT Empowered Smart Road Traffic Congestion Control System Using Supervised Machine Learning Algorithm," *The Computer Journal*, vol. 64, no. 11, pp. 1672–1679, May 2020.
- [2] A. R. Al-Ali, I. A. Zualkernan, M. Rashid, R. Gupta, and M. Alikarar, "A smart home energy management system using IoT and big data analytics approach," *IEEE Trans. on Consumer Electronics*, vol. 63, no. 4, pp. 426–434, November 2017.
- [3] J. Schwarzkopf, I. Zacarias, A. L. C. Bazzan, L. H. Moreira, and E. P. D. Freitas, "Solving task allocation problem in multi Unmanned Aerial Vehicles systems using Swarm intelligence," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 10–20, June 2018.
- [4] H. V. D. Parunak, S. A. Brueckner, and J. Sauter, "Digital Pheromones for Coordination of Unmanned Vehicles," in *Environments for Multi-Agent Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 246–263.
- [5] M. G. C. A. Cimino, A. Lazzeri, and G. Vaglini, "Combining stigmergic and flocking behaviors to coordinate swarms of drones performing target search," in *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Corfu, Greece, July 2015.
- [6] M. Volodymyr, K. Koray, S. David, A. A. Rusu, V. Joel, M. G. Bellemare, G. Alex, R. Martin, A. K. Fidjeland, and O. Georg, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, February 2015.
- [7] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust Region Policy Optimization," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, Lille, France, July 2015, pp. 1889–1897.
- [8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," July 2017, arXiv. [Online]. Available: <https://arxiv.org/abs/1707.06347>.
- [9] K. Lee, S. Kim, S. Lim, S. Choi, and S. Oh, "Tsallis Reinforcement Learning: A Unified Framework for Maximum Entropy Reinforcement Learning," February 2019, arXiv. [Online]. Available: <https://arxiv.org/abs/1902.00137v1>.
- [10] M. L. Littman, "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, vol. 4, no. 1, p. 237–285, 1996.
- [11] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, November 2017.
- [12] M. Tan, "Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents," in *Readings in Agents*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, p. 487–494.
- [13] R. Sutton and A. Barto, "Reinforcement Learning: An Introduction," *IEEE Trans. on Neural Networks*, vol. 9, no. 5, pp. 1054–1054, September 1998.
- [14] X. Xu, R. Li, Z. Zhao, and H. Zhang, "Stigmergic Independent Reinforcement Learning for Multiagent Collaboration," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4285–4299, 2022.
- [15] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing Federated Learning on Non-IID Data with Reinforcement Learning," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, Toronto, ON, Canada, July 2020, pp. 1698–1707.
- [16] H. Cha, J. Park, H. Kim, M. Bennis, and S. L. Kim, "Proxy Experience Replay: Federated Distillation for Distributed Reinforcement Learning," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 94–101, May 2020.
- [17] P. Chaudhari, C. Baldassi, R. Zecchina, S. Soatto, A. Talwalkar, and A. Oberman, "Parle: parallelizing stochastic gradient descent," July 2017, arXiv. [Online]. Available: <https://arxiv.org/abs/1707.00424>.
- [18] V. Smith, S. Forte, C. Ma, M. Takac, M. I. Jordan, and M. Jaggi, "Co-CoA: A General Framework for Communication-Efficient Distributed Optimization," *Journal of Machine Learning Research*, vol. 18, pp. 1–49, November 2016.
- [19] H. Yu, S. Yang, and S. Zhu, "Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, Hawaii, USA, July 2019, pp. 5693–5700.
- [20] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and Cooperation in Networked Multi-Agent Systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, March 2007.
- [21] S. Hosseinalipour, S. S. Azam, C. G. Brinton, N. Michelusi, V. Aggarwal, D. J. Love, and H. Dai, "Multi-Stage Hybrid Federated Learning over Large-Scale D2D-Enabled Fog Networks," January 2022, arXiv. [Online]. Available: <https://arxiv.org/abs/2007.09511>.
- [22] S. Wang, S. Hosseinalipour, M. Gorlatova, C. G. Brinton, and M. Chiang, "UAV-assisted Online Machine Learning over Multi-Tiered Networks: A Hierarchical Nested Personalized Federated Learning Approach," *IEEE Trans. on Network and Service Management (Early Access)*, pp. 1–36, October 2022.
- [23] W. Yang, W. Xiang, Y. Yang, and P. Cheng, "Optimizing Federated Learning With Deep Reinforcement Learning for Digital Twin Empowered Industrial IoT," *IEEE Trans. on Industrial Informatics*, vol. 19, no. 2, pp. 1884–1893, 2023.
- [24] Z. Xie and S. Song, "FedKL: Tackling Data Heterogeneity in Federated Reinforcement Learning by Penalizing KL Divergence," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 4, pp. 1227–1242, 2023.
- [25] Y. Ruan, X. Zhang, S. Liang, and C. Joe-Wong, "Towards Flexible Device Participation in Federated Learning," February 2021, arXiv. [Online]. Available: <https://arxiv.org/abs/2006.06954>.
- [26] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, "Collaborative Deep Learning in Fixed Topology Networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, December 2017, p. 5906–5916.
- [27] H. Xing, O. Simeone, and S. Bi, "Decentralized Federated Learning via SGD over Wireless D2D Networks," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Atlanta, GA, USA, May 2020, pp. 1–5.
- [28] X. Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous Decentralized Parallel Stochastic Gradient Descent," September 2018, arXiv. [Online]. Available: <https://arxiv.org/abs/1710.06952>.
- [29] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proceedings of 11th International Conference on Machine Learning*, New Brunswick, NJ, USA, July 1994, pp. 157–163.
- [30] K. Shah and M. Kumar, "Distributed Independent Reinforcement Learning (DIRL) Approach to Resource Management in Wireless Sensor Networks," in *2007 IEEE International Conference on Mobile Adhoc and Sensor Systems*, Pisa, Italy, October 2007, pp. 1–9.
- [31] L. Busoniu, R. Babuska, and B. D. Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning," *IEEE Trans. on Systems, Man, and Cybernetics, Part C*, vol. 38, no. 2, pp. 156–172, March 2008.
- [32] V. Mnih, A. P. Badia, M. Mirza, A. Graves, and K. Kavukcuoglu, "Asynchronous Methods for Deep Reinforcement Learning," June 2016, arXiv. [Online]. Available: <https://arxiv.org/abs/1602.01783>.
- [33] G. Sartoretti, Y. Wu, W. Paivine, and T. K. S. Kumar, "Distributed Reinforcement Learning for Multi-robot Decentralized Collective Construction," in *Distributed Autonomous Robotic Systems*. Cham: Springer International Publishing, 2019, pp. 35–49.
- [34] X. Liang, Y. Liu, T. Chen, M. Liu, and Q. Yang, "Federated Transfer Reinforcement Learning for Autonomous Driving," October 2019, arXiv. [Online]. Available: <https://arxiv.org/abs/1910.06001>.
- [35] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic Policy Gradient Algorithms," in *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, June 2014, p. 387–395.
- [36] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft Actor-Critic Algorithms and Applications," January 2019, arXiv. [Online]. Available: <https://arxiv.org/abs/1812.05905>.

- [37] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, "Federated Deep Reinforcement Learning for Internet of Things with Decentralized Cooperative Edge Caching," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9441–9455, April 2020.
- [38] H. van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-learning," December 2015, arXiv. [Online]. Available: <https://arxiv.org/abs/1509.06461>.
- [39] R. Hu, Y. Gong, and Y. Guo, "Concentrated Differentially Private and Utility Preserving Federated Learning," September 2020, arXiv. [Online]. Available: <https://arxiv.org/abs/2003.13761>.
- [40] P. Kairouz, H. B. McMahan, and B. Avent, "Advances and Open Problems in Federated Learning," March 2021, arXiv. [Online]. Available: <https://arxiv.org/abs/1912.04977>.
- [41] J. Konecný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated Learning: Strategies for Improving Communication Efficiency," October 2017, arXiv. [Online]. Available: <https://arxiv.org/abs/1610.05492v2>.
- [42] J. Wang and G. Joshi, "Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms," January 2019, arXiv. [Online]. Available: <https://arxiv.org/abs/1808.07576>.
- [43] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, HI, USA, April 2018, pp. 63–71.
- [44] F. Haddadpour, M. Mahdi Kamani, M. Mahdavi, and V. R. Cadambe, "Local SGD with Periodic Averaging: Tighter Analysis and Adaptive Synchronization," May 2020, arXiv. [Online]. Available: <https://arxiv.org/abs/1910.13598>.
- [45] J. Wu, W. Huang, J. Huang, and T. Zhang, "Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, Stockholmässan, Stockholm Sweden, July 2018, pp. 5325–5333.
- [46] H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright, "ATOMO: Communication-efficient Learning via Atomic Sparsification," in *Advances in Neural Information Processing Systems*, vol. 31, Montréal, Canada, December 2018, pp. 9850–9861.
- [47] S. Shi, Q. Wang, X. Chu, B. Li, Y. Qin, R. Liu, and X. Zhao, "Communication-Efficient Distributed Deep Learning with Merged Gradient Sparsification on GPUs," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, Toronto, ON, Canada, July 2020, pp. 406–415.
- [48] H. Wang, Z. Qu, S. Guo, X. Gao, R. Li, and B. Ye, "Intermittent Pulling With Local Compensation for Communication-Efficient Distributed Learning," *IEEE Trans. on Emerging Topics in Computing*, vol. 10, no. 2, pp. 779–791, 2022.
- [49] X. Xu, R. Li, Z. Zhao, and H. Zhang, "Trustable Policy Collaboration Scheme for Multi-Agent Stigmergic Reinforcement Learning," *IEEE Communications Letters*, vol. 26, no. 4, pp. 823–827, January 2022.
- [50] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient Descent Only Converges to Minimizers," in *29th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 49, Columbia University, New York, USA, June 2016, pp. 1246–1257.
- [51] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, International Convention Centre, Sydney, Australia, August 2017, pp. 2681–2690.
- [52] G. Tesuaro, "Temporal difference learning and TD-gammon," *Communications of the ACM*, vol. 38, no. 3, pp. 58–68, 1995.
- [53] R. Hu, Y. Guo, E. P. Ratazzi, and Y. Gong, "Differentially Private Federated Learning for Resource-Constrained Internet of Things," March 2020, arXiv. [Online]. Available: <https://arxiv.org/abs/2003.12705v1>.
- [54] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization," in *Advances in Neural Information Processing Systems*, vol. 28, Montreal, Canada, December 2015, pp. 2737–2745.
- [55] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization Methods for Large-Scale Machine Learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, May 2018.
- [56] X. Zhou, "On the Fenchel Duality between Strong Convexity and Lipschitz Continuous Gradient," March 2018, arXiv. [Online]. Available: <https://arxiv.org/abs/1803.06573>.
- [57] E. Vinitsky, A. Kreidieh, L. L. Flem, N. Kheterpal, K. Jang, C. Wu, F. Wu, R. Liaw, E. Liang, and A. M. Bayen, "Benchmarks for reinforcement learning in mixed-autonomy traffic," in *Proceedings of the 2nd Conference on Robot Learning*, vol. 87, Zurich, Switzerland, October 2018, pp. 399–409.
- [58] S. Kakade and J. Langford, "Approximately Optimal Approximate Reinforcement Learning," in *Proceedings of the 19th International Conference on Machine Learning*, Sydney, Australia, 2002, pp. 267–274.



Xing Xu received the B.E. degree in Communication Engineering and the Ph.D. degree in Information and Communication Engineering from Huazhong University of Science and Technology and Zhejiang University, respectively. His research interests include collective intelligence, deep reinforcement learning, data analysis, and artificial intelligence.



Rongpeng Li is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He was a Research Engineer with the Wireless Communication Laboratory, Huawei Technologies Company, Ltd., Shanghai, China, from August 2015 to September 2016. He was a Visiting Scholar with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K., from February 2020 to August 2020. His research interest currently focuses on networked intelligence for communications evolving (NICE). He received the Wu Wenjun Artificial Intelligence Excellent Youth Award in 2021. He serves as an Editor for *China Communications*.



Zhifeng Zhao received the B.E. degree in computer science, the M.E. degree in communication and information systems, and the Ph.D. degree in communication and information systems from the PLA University of Science and Technology, Nanjing, China, in 1996, 1999, and 2002, respectively. From 2002 to 2004, he acted as a Post-Doctoral Researcher with Zhejiang University, Hangzhou, China, where his researches were focused on multimedia next-generation networks (NGNs) and softswitch technology for energy efficiency. From 2005 to 2006, he acted as a Senior Researcher with the PLA University of Science and Technology, where he performed research and development on advanced energy-efficient wireless router, *ad-hoc* network simulator, and cognitive mesh networking test-bed. From 2006 to 2019, he was an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University. Currently, he is with the Zhejiang Lab, Hangzhou as the Chief Engineering Officer. His research areas include software defined networks (SDNs), wireless network in 6G, computing networks, and collective intelligence. He is the Symposium Co-Chair of ChinaCom 2009 and 2010. He is the Technical Program Committee (TPC) Co-Chair of the 10th IEEE International Symposium on Communication and Information Technology (ISCIT 2010).



Honggang Zhang was an Honorary Visiting Professor with the University of York, York, U.K., and an International Chair Professor of excellence with the Université Européenne de Bretagne and Supélec, France. He is the Chief Managing Editor of Intelligent Computing, a Science Partner Journal, as well as a Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He has coauthored and edited two books: *Cognitive Communications: Distributed Artificial Intelligence (DAI), Regulatory Policy & Economics, Implementation* (John Wiley & Sons) and *Green Communications: Theoretical Fundamentals, Algorithms and Applications* (CRC Press), respectively. His research interests include cognitive radio and networks, green communications, mobile computing, machine learning, artificial intelligence, and the Internet of Intelligence (IoI). He is a co-recipient of the 2021 IEEE Communications Society Outstanding Paper Award and the 2021 IEEE INTERNET OF THINGS JOURNAL (IoT-J) Best Paper Award. He was the leading Guest Editor for the Special Issues on Green Communications of the *IEEE Communications Magazine*. He served as a Series Editor for the *IEEE Communications Magazine* (Green Communications and Computing Networks Series) from 2015 to 2018 and the Chair of the Technical Committee on Cognitive Networks of the IEEE Communications Society from 2011 to 2012. He is the Associate Editor-in-Chief of *China Communications*.