

Differentially Private Distributed Nonconvex Stochastic Optimization with Quantized Communication

Jialong Chen, Jimin Wang, *Member, IEEE*, and Ji-Feng Zhang, *Fellow, IEEE*

Abstract—This paper proposes a new distributed nonconvex stochastic optimization algorithm that can achieve privacy protection, communication efficiency and convergence simultaneously. Specifically, each node adds general privacy noises to its local state to avoid information leakage, and then quantizes its noise-perturbed state before transmitting to improve communication efficiency. By using a subsampling method controlled through the sample-size parameter, the proposed algorithm reduces cumulative differential privacy parameters ϵ , δ , and thus enhances the differential privacy level, which is significantly different from the existing works. By using a two-time-scale step-sizes method, the mean square convergence for nonconvex cost functions is given. Furthermore, when the global cost function satisfies the Polyak-Łojasiewicz condition, the convergence rate and the oracle complexity of the proposed algorithm are given. In addition, the proposed algorithm achieves both the mean square convergence and finite cumulative differential privacy parameters ϵ , δ over infinite iterations as the sample-size goes to infinity. A numerical example of the distributed training on the “MNIST” dataset is given to show the effectiveness of the algorithm.

Index Terms—Differential privacy, distributed stochastic optimization, probabilistic quantization.

I. INTRODUCTION

DISTRIBUTED optimization is gaining more and more attraction due to its fundamental role in cooperative control, smart grids, sensor networks, and large-scale machine learning. In these applications, the problem can be formulated

The work was supported by National Natural Science Foundation of China under Grant 62203045, 62433020 and Grant T2293770. The material in this paper was not presented at any conference. Corresponding author: Jimin Wang.

Jialong Chen is with the Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, and also with the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China. (e-mail: chenjialong23@mails.ucas.ac.cn)

Jimin Wang is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, and also with the Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, Beijing 100083, China (e-mail: jimwang@ustb.edu.cn)

Ji-Feng Zhang is with the School of Automation and Electrical Engineering, Zhongyuan University of Technology, Zheng Zhou 450007; and also with the Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. (e-mail: jif@iss.ac.cn)

as a network of nodes cooperatively solve a common optimization problem through on-node computation and local communication [1]–[11]. As a branch of distributed optimization, distributed stochastic optimization focuses on finding optimal solutions for stochastic cost functions in a distributed manner. For example, distributed stochastic gradient descent (SGD) [6], [7], distributed SGD with quantized communication [8], SGD with gradient compression [9], [10], and distributed SGD with variable sample-size method [11] are given, respectively.

When nodes exchange information to solve a distributed stochastic optimization problem, there are two key issues worthy of attention. One is the leakage of the sensitive information concerning cost functions, and the other is the network bandwidth limitation. To solve the first issue, it is necessary to design some privacy-preserving techniques to protect the sensitive information in distributed stochastic optimization [12]. So far, various techniques have been employed such as homomorphic encryption [13], correlated noise based approach [14], structure techniques [15]–[17], differential privacy [18]–[23] and so on. Homomorphic encryption often incurs a communication and computation burden, while correlated noise based approach and structure techniques provide only limited privacy protection. Due to its simplicity and wide applicability in privacy protection, differential privacy has attracted a lot of attention and been used to solve privacy issues in distributed optimization. For example, distributed stochastic optimization algorithms with differential privacy are proposed in [24]–[33]. In distributed convex stochastic optimization with differential privacy, alternating direction method of multipliers with output perturbation [24], [26], distributed SGD with output perturbation [25], distributed SGD with quantized communication [27], zero-th order alternating direction method of multipliers with output perturbation [28], and distributed dual averaging with gradient perturbation [29] are given, respectively. In distributed nonconvex stochastic optimization with differential privacy, some valuable results have been given, such as distributed SGD with gradient perturbation [30], [31] and quantization enabled privacy protection [32], [33]. However, to prove the convergence and the differential privacy, the assumption of bounded gradients is required in [24]–[29], [31], [32]. What’s more, differential privacy is only given for each iteration in [24]–[33], leading to infinite cumulative differential privacy parameters ϵ , δ over infinite iterations.

To solve the second issue, a common method is to transmit quantized information instead of the raw information. The

examples include the adaptive quantizer [3], probabilistic quantizer [9], [27], [32], [33], uniform quantizer [34], logarithmic quantizer [35], zooming-in quantizer [36], and binary-valued quantizer [37]. All these quantizers can improve the communication efficiency. However, to our knowledge, the adaptive quantizer requires more network bandwidth than the probabilistic quantizer to reduce the quantization error; the uniform quantizer, logarithmic quantizer, zooming-in quantizer may bring difficulty to the convergence analysis; and the binary-valued quantizer requires specific probability distributions of noises in gradients.

Although privacy protection, communication efficiency and convergence are considered simultaneously in [27], differential privacy is only achieved for each iteration therein. Therefore, this paper will focus on how to design a privacy-preserving distributed nonconvex stochastic optimization algorithm that can enhance the differential privacy level while achieving communication efficiency and convergence simultaneously; and further show how the added privacy noises and the quantization error affect the convergence rate of the algorithm.

In this paper, we consider differentially private distributed nonconvex stochastic optimization with quantized communication. By using a subsampling and a two-time-scale step-sizes method, differential privacy, communication efficiency and convergence are obtained simultaneously. The main contributions of this paper are as follows:

- A subsampling method controlled through the sample-size parameter is proposed to enhance the differential privacy level. By using this subsampling method, cumulative differential privacy parameters ϵ , δ are reduced with guaranteed mean square convergence for general privacy noises. Furthermore, when the sample-size goes to infinity, the algorithm achieves both the mean square convergence and finite cumulative differential privacy parameters ϵ , δ over infinite iterations simultaneously.
- By using a two-time-scale step-sizes method, the mean square convergence of the algorithm for nonconvex cost functions is given without the assumption of bounded gradients. Under the Polyak-Łojasiewicz condition, the convergence rate of the algorithm for general privacy noises is provided, including decreasing, constant and increasing privacy noises.

The results in this paper are significantly different from those in existing works. A comparison with the state-of-the-art is as follows: Compared with [24]–[33], finite cumulative differential privacy parameters ϵ , δ are achieved over infinite iterations. Compared with [7], [10], [24]–[29], [31], [32], the convergence of the proposed algorithm is given without the assumption of bounded gradients. Compared with [9], [10], [28], [33], the convergence is achieved while achieving differential privacy. Compared with [6]–[11], [24]–[26], [28]–[31], privacy protection and communication efficiency are considered simultaneously in this paper.

This paper is organized as follows: Section II formulates the problem to be investigated. Section III presents the main results including the privacy, convergence and oracle complexity analysis of the algorithm. Section IV provides a numerical example of the distributed training of a convolutional neural

network on the “MNIST” dataset. Section V gives some concluding remarks.

Notation: \mathbb{R} and \mathbb{R}^r denote the set of all real numbers and r -dimensional Euclidean space, respectively. $\text{Range}(F)$ denotes the range of a mapping F , and $F \circ G$ denotes the composition of mappings F and G . For sequences $\{a_k\}_{k=1}^{\infty}$ and $\{b_k\}_{k=1}^{\infty}$, $a_k = O(b_k)$ means there exists $A_1 \geq 0$ such that $\limsup_{k \rightarrow \infty} |a_k/b_k| \leq A_1$. $\mathbf{1}_n$ represents an n -dimensional vector whose elements are all 1. A^\top stands for the transpose of the matrix A . We use the symbol $\|x\| = \sqrt{x^\top x}$ to denote the standard Euclidean norm of $x = [x_1, x_2, \dots, x_m]^\top$, and $\|A\|$ to denote the 2-norm of the matrix A . $\mathbb{P}(\mathcal{B})$ and $\mathbb{E}(X)$ refer to the probability of an event \mathcal{B} and the expectation of a random variable X , respectively. \otimes denotes the Kronecker product of matrices. $\lfloor z \rfloor$ denotes the largest integer no larger than z . For a vector $v = [v_1, v_2, \dots, v_n]^\top$, $\text{diag}(v)$ denotes the diagonal matrix with diagonal elements being v_1, v_2, \dots, v_n . For a differentiable function $f(x)$, $\nabla f(x)$ denotes its gradient at the point x .

II. PRELIMINARIES AND PROBLEM FORMULATION

A. Graph theory

Consider a network of n nodes which exchange information on an undirected and connected communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V} = \{1, 2, \dots, n\}$ is the set of all nodes, and \mathcal{E} is the set of all edges. An edge $e_{ij} \in \mathcal{E}$ if and only if Node i can receive the information from j . Different nodes in \mathcal{V} exchange information based on the weight matrix $\mathcal{A} = (a_{ij})_{1 \leq i, j \leq n}$, whose entry a_{ij} is either positive if $e_{ij} \in \mathcal{E}$, or 0, otherwise. The neighbor set of Node i is defined as $\mathcal{N}_i = \{j \in \mathcal{V} : a_{ij} > 0\}$, and the Laplacian matrix of \mathcal{A} is defined as $\mathcal{L} = \text{diag}(\mathcal{A}\mathbf{1}_n) - \mathcal{A}$. The assumption about the weight matrix \mathcal{A} is given as follows:

Assumption 1: The weight matrix \mathcal{A} is doubly stochastic, i.e., $\mathcal{A}\mathbf{1}_n = \mathbf{1}_n$, $\mathbf{1}_n^\top \mathcal{A} = \mathbf{1}_n^\top$.

Remark 1: Assumption 1 is standard and commonly used in undirected and connected communication graphs (see e.g. [3], [4], [6], [8], [14], [26]–[28], [30]–[32]). There are many examples satisfying Assumption 1 in practice, such as, the dynamic load balancing of distributed memory processors ([38]), the distributed estimation of sensor networks ([39]) and the distributed machine learning ([40]).

B. Distributed stochastic optimization

In this paper, the following distributed nonconvex stochastic optimization problem is considered:

$$\min_{x \in \mathbb{R}^r} F(x) = \min_{x \in \mathbb{R}^r} \frac{1}{n} \sum_{i=1}^n f_i(x), f_i(x) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [\ell_i(x, \xi_i)], \quad (1)$$

where x is available to all nodes, $\ell_i(x, \xi_i)$ is a local cost function which is private to Node i , and ξ_i is a random variable drawn from an unknown probability distribution \mathcal{D}_i . In practice, since the probability distribution \mathcal{D}_i is difficult to obtain, it is replaced by the dataset $\mathcal{D}_i = \{\xi_{i,l}, 1 \leq l \leq D\}$. Then, (1) can be rewritten as the following empirical risk minimization problem:

$$\min_{x \in \mathbb{R}^r} F(x) = \min_{x \in \mathbb{R}^r} \frac{1}{n} \sum_{i=1}^n f_i(x), f_i(x) = \frac{1}{D} \sum_{j=1}^D \ell_i(x, \xi_{i,j}). \quad (2)$$

To solve the empirical risk minimization problem (2), we need the following standard assumption.

Assumption 2: (i) For any node $i \in \mathcal{V}$, f_i has Lipschitz continuous gradients, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$, $\forall x, y \in \mathbb{R}^r$, where L is a positive constant.

(ii) Each cost function is bounded from below, i.e., $\min_{x \in \mathbb{R}^r} f_i(x) = f_i^* > -\infty$.

(iii) For any node $i \in \mathcal{V}$, $x \in \mathbb{R}^r$ and ζ_i uniformly sampled from \mathcal{D}_i , there exists a stochastic first-order oracle which returns a sampled gradient $g_i(x, \zeta_i)$ of $f_i(x)$. In addition, there exists $\sigma_g > 0$ such that each sampled gradient $g_i(x, \zeta_i)$ satisfies $\mathbb{E}[g_i(x, \zeta_i)] = \nabla f_i(x)$, $\mathbb{E}[\|g_i(x, \zeta_i) - \nabla f_i(x)\|^2] \leq \sigma_g^2$.

Remark 2: Assumption 2(i) is commonly used (see e.g. [5], [7], [8], [11], [24], [27], [30]–[33]). Assumption 2(ii) ensures the existence of the optimal solution. Assumption 2(iii) requires that each sampled gradient $g_i(x, \xi_{i,l})$ is unbiased with a bounded variance σ_g^2 (see e.g. [11], [27], [30], [32], [33]).

C. Quantized communication

Due to the network bandwidth limitation, the exchange of the uncompressed information brings communication burden. To address this, the probabilistic quantizer is used to quantize the exchanged information in this paper, which is a randomized mapping that maps an input to different values in a discrete set with some probability distribution, and satisfies the following assumption:

Assumption 3: The probabilistic quantizer $Q(x)$ is unbiased and its variance is bounded, which means there exists $\Delta > 0$, such that $\mathbb{E}(Q(x)|x) = x$ and $\mathbb{E}(|Q(x) - x|^2|x) \leq \Delta^2$.

Remark 3: Assumption 3 is standard and commonly used (see e.g. [8], [27]). Here is an example: Given $\Delta > 0$, the quantizer $Q(x)$ with the following probability distribution satisfies Assumption 3 by Lemma 1 of [41].

$$\begin{cases} \mathbb{P}(Q(x) = \Delta \lfloor \frac{x}{\Delta} \rfloor | x) = 1 - \frac{x}{\Delta} + \lfloor \frac{x}{\Delta} \rfloor; \\ \mathbb{P}(Q(x) = \Delta (\lfloor \frac{x}{\Delta} \rfloor + 1) | x) = \frac{x}{\Delta} - \lfloor \frac{x}{\Delta} \rfloor. \end{cases} \quad (3)$$

D. Differential privacy

As shown in [31], [32], there are two kinds of adversary models widely used in the privacy-preserving issue for distributed stochastic optimization:

- A *semi-honest* adversary. This kind of adversary is defined as a node within the network which has access to certain internal states (such as $x_{i,k}$ from Node i), follows the prescribed protocols and accurately computes iterative state correctly. However, it aims to infer the sensitive information of other nodes.
- An *eavesdropper*. This kind of adversary refers to an external adversary who has capability to wiretap and monitor all communication channels, allowing them to capture distributed messages from any node. This enables the eavesdropper to infer the sensitive information of internal nodes.

When solving the empirical risk minimization problem (2), the stochastic first-order oracle needs data samples to return sampled gradients. Meanwhile, the adversaries can infer the sensitive information of data samples from sampled gradients ([42]). In order to provide privacy protection for data samples, inspired by [21], [29], a symmetric binary relation called *adjacency relation* is defined as follows:

Definition 1: (Adjacency relation) Let $\mathcal{D} = \{\xi_{i,l}, i \in \mathcal{V}, 1 \leq l \leq D\}$, $\mathcal{D}' = \{\xi'_{i,l}, i \in \mathcal{V}, 1 \leq l \leq D\}$ be two sets of data samples. If for a given $C > 0$ and any $x \in \mathbb{R}^r$, there exists exactly one pair of data samples $\xi_{i_0, l_0}, \xi'_{i_0, l_0}$ in $\mathcal{D}, \mathcal{D}'$ such that

$$\begin{cases} \|g_i(x, \xi_{i,l}) - g_i(x, \xi'_{i,l})\| \leq C, & \text{if } i = i_0 \text{ and } l = l_0; \\ \|g_i(x, \xi_{i,l}) - g_i(x, \xi'_{i,l})\| = 0, & \text{if } i \neq i_0 \text{ or } l \neq l_0, \end{cases} \quad (4)$$

then \mathcal{D} and \mathcal{D}' are said to be adjacent, denoted by $\text{Adj}(\mathcal{D}, \mathcal{D}')$.

Remark 4: The boundary C characterizes the “closeness” of a pair of data samples $\xi_{i_0, l_0}, \xi'_{i_0, l_0}$. By (4), the larger the boundary C is, the larger the allowed magnitude of sampled gradients between adjacent datasets is, and thus the better the privacy protection level is. Furthermore, the boundary C is related to the distribution of the dataset. For example, as shown in Fig. 4 of Section IV, the boundary C is different for the “MNIST”, “CIFAR-10” and “CIFAR-100” dataset.

To give the privacy-preserving level of the algorithm, we adopt the definition of the (ϵ, δ) -differential privacy as follows:

Definition 2: [29] ((ϵ, δ) -differential privacy) Given $\epsilon, \delta > 0$, a randomized algorithm \mathcal{M} achieves the (ϵ, δ) -differential privacy for $\text{Adj}(\mathcal{D}, \mathcal{D}')$ if for any given observation set $T \subset \text{Range}(\mathcal{M})$, it holds that $\mathbb{P}(\mathcal{M}(\mathcal{D}) \in T) \leq e^\epsilon \mathbb{P}(\mathcal{M}(\mathcal{D}') \in T) + \delta$.

III. MAIN RESULT

A. The proposed algorithm

In this subsection, we give a differentially private distributed nonconvex stochastic optimization algorithm with quantized communication. The detailed implementation steps are given in Algorithm 1.

Algorithm 1 Differentially private distributed nonconvex stochastic optimization algorithm with quantized communication

Initialization: $x_{i,0} \in \mathbb{R}^r$ for any node $i \in \mathcal{V}$, weight matrix $(a_{ij})_{1 \leq i, j \leq n}$, iteration limit K , step-sizes $\hat{\alpha} = \frac{\alpha_1}{K^\alpha}$, $\hat{\beta} = \frac{\alpha_2}{K^\beta}$ and sample-size $\hat{\gamma} = \lfloor a_3 K^\gamma \rfloor + 1$.

for $k = 0, 1, 2, \dots, K$, **do**

- 1: Node i adds noise $d_{i,k}$ to $x_{i,k}$ and computes the quantized information $z_{i,k} = Q(x_{i,k} + d_{i,k}) = [Q(x_{i,k}^{(1)} + d_{i,k}^{(1)}), \dots, Q(x_{i,k}^{(r)} + d_{i,k}^{(r)})]^\top$ with the probabilistic quantizer in the form of (3), where $d_{i,k} \sim N(0, \sigma_k^2 I_r)$.
- 2: Node i broadcasts $z_{i,k}$ to its neighbors $j \in \mathcal{N}_i$, receives $z_{j,k}$ from its neighbors $j \in \mathcal{N}_i$, and aggregates the received information by

$$\tilde{x}_{i,k} = (1 - \hat{\beta})x_{i,k} + \hat{\beta} \sum_{j \in \mathcal{N}_i} a_{ij} z_{j,k}. \quad (5)$$

- 3: Node i takes $\hat{\gamma}$ different data samples $\zeta_{i,k,1}, \dots, \zeta_{i,k,\hat{\gamma}}$ uniformly from \mathcal{D}_i to generate sampled gradients $g_i(x_{i,k}, \zeta_{i,k,1}), \dots, g_i(x_{i,k}, \zeta_{i,k,\hat{\gamma}})$. Then, Node i puts these data samples back into \mathcal{D}_i .
- 4: Node i computes the averaged sampled gradient by

$$g_{i,k} = \frac{1}{\hat{\gamma}} \sum_{l=1}^{\hat{\gamma}} g_i(x_{i,k}, \zeta_{i,k,l}). \quad (6)$$

- 5: Node i updates its state by

$$x_{i,k+1} = \tilde{x}_{i,k} - \hat{\alpha} g_{i,k}. \quad (7)$$

end for

Remark 5: The subsampling method in Algorithm 1 can ensure that there are sufficient data samples to generate sampled gradients, even when each node only has one data sample. Specifically, let $\gamma = 0$, $a_3 = \frac{D}{2}$ for any $D \geq 1$. Then, the sample-size $\hat{\gamma} = \lfloor \frac{D}{2} \rfloor + 1 \leq D$. By this subsampling method, $\hat{\gamma}$ different data samples $\zeta_{i,0,1}, \dots, \zeta_{i,0,\hat{\gamma}}$ are drawn from the dataset \mathcal{D}_i to generate sampled gradients $g_i(x_{i,0}, \zeta_{i,0,1}), \dots, g_i(x_{i,0}, \zeta_{i,0,\hat{\gamma}})$ at the zero-th iteration. After sampled gradients are generated, these data samples are put back into the dataset \mathcal{D}_i to ensure that the dataset \mathcal{D}_i still has D data samples. Thus, $\hat{\gamma}$ data samples $\zeta_{i,1,1}, \dots, \zeta_{i,1,\hat{\gamma}}$ can be drawn from \mathcal{D}_i to generate sampled gradients $g_i(x_{i,1}, \zeta_{i,1,1}), \dots, g_i(x_{i,1}, \zeta_{i,1,\hat{\gamma}})$ at the first iteration. Therefore, by mathematical induction, it can be seen that there are sufficient data samples to run Algorithm 1 for any node $i \in \mathcal{V}$.

B. Privacy analysis

In this subsection, we will show the differential privacy analysis of Algorithm 1. Inspired by [21], we first provide the sensitivity of the algorithm, which helps us to analyze the differential privacy of the algorithm.

Definition 3: (Sensitivity) Given two groups of adjacent sample sets $\mathcal{D}, \mathcal{D}'$, and a mapping q . For any $0 \leq k \leq K$, let $\mathcal{D}_k = \{\zeta_{i,k,l}, i \in \mathcal{V}, 1 \leq l \leq \hat{\gamma}\}$, $\mathcal{D}'_k = \{\zeta'_{i,k,l}, i \in \mathcal{V}, 1 \leq l \leq \hat{\gamma}\}$ be the data samples taken from $\mathcal{D}, \mathcal{D}'$ at the k -th iteration, respectively. Define the sensitivity of q at the k -th iteration of Algorithm 1 as follows:

$$\Delta_k^q \triangleq \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|q(\mathcal{D}_k) - q(\mathcal{D}'_k)\|. \quad (8)$$

Remark 6: Definition 3 captures the magnitude by which one node's data sample can change the mapping q in the worst case. It is the key quantity showing how much noise should be added to achieve the (ϵ, δ) -differential privacy at the k -th iteration. In Algorithm 1, the mapping $q(\mathcal{D}_k) = x_{k+1} = [x_{1,k+1}^\top, \dots, x_{n,k+1}^\top]^\top$, the randomized mapping $\mathcal{M}(\mathcal{D}_k) = Q(q(\mathcal{D}_k) + d_{k+1}) = Q(x_{k+1} + d_{k+1}) = [Q(x_{1,k+1} + d_{1,k+1}), \dots, Q(x_{n,k+1} + d_{n,k+1})] = z_{k+1}$.

The following lemma gives the sensitivity Δ_k of Algorithm 1 for any $0 \leq k \leq K$.

Lemma 1: At the k -th iteration, the sensitivity of Algorithm 1 satisfies $\Delta_k^q \leq \frac{\hat{\alpha}C}{\hat{\gamma}} \left(\sum_{m=0}^k |1 - \hat{\beta}|^m \right)$.

Proof: When $k = 0$, (8) can be written as

$$\Delta_0^q = \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|q(\mathcal{D}_0) - q(\mathcal{D}'_0)\| = \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|x_1 - x'_1\|. \quad (9)$$

Note that the sensitivity is obtained by computing the maximum magnitude of the mapping q when changing one data sample. Then, observations (z_0, z_1, \dots, z_K) , $(z'_0, z'_1, \dots, z'_K)$ of Algorithm 1 between adjacent datasets $\mathcal{D}, \mathcal{D}'$ should be equal such that only the effect of changing one data sample is considered. This shows how much noise should be added such that the probability of $\mathcal{M}(\mathcal{D}) = t$ and the probability of $\mathcal{M}(\mathcal{D}') = t$ satisfy $\mathbb{P}(\mathcal{M}(\mathcal{D}) = t) \leq e^\epsilon \mathbb{P}(\mathcal{M}(\mathcal{D}') = t) + \delta$ for any $t \in T$ and observation set $T \subseteq \mathbb{R}^{nK}$. Thus, we have $\mathbb{P}(\mathcal{M}(\mathcal{D}) \in T) \leq e^\epsilon \mathbb{P}(\mathcal{M}(\mathcal{D}') \in T) + \delta$. Hence, $z_{j,k} = z'_{j,k}$ holds for any node $j \in \mathcal{N}_i$ and $0 \leq k \leq K$.

Since $x_{i,0} = x'_{i,0}$, $z_{i,0} = z'_{i,0}$ hold for any node $i \in \mathcal{V}$, by (5), $\tilde{x}_{i,0} = \tilde{x}'_{i,0}$ holds for any node $i \in \mathcal{V}$. Let $g_0 = [g_{1,0}^\top, \dots, g_{n,0}^\top]^\top$. Then, substituting (7) into (9) implies

$$\Delta_0^q = \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|\hat{\alpha}(g_0 - g'_0)\|. \quad (10)$$

By Definition 1, since \mathcal{D} and \mathcal{D}' are adjacent, there exists exactly one pair of data samples $\xi_{i_0, l_0}, \xi'_{i_0, l_0}$ in \mathcal{D} and \mathcal{D}' such that (4) holds. This implies that $g_{j,0} = g'_{j,0}$ holds for any node $j \neq i_0$. Thus, (10) can be rewritten as

$$\Delta_0^q = \hat{\alpha} \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|g_{i_0,0} - g'_{i_0,0}\|. \quad (11)$$

Since $\hat{\gamma}$ different data samples are taken uniformly from $\mathcal{D}, \mathcal{D}'$ respectively, there exists at most one pair of data samples $\zeta_{i_0,0,l_1}, \zeta'_{i_0,0,l_1}$ such that $\zeta_{i_0,0,l_1} = \xi_{i_0, l_0}$, $\zeta'_{i_0,0,l_1} = \xi'_{i_0, l_0}$. Thus, by (6), (11) can be rewritten as

$$\begin{aligned} \Delta_0^q &= \frac{\hat{\alpha}}{\hat{\gamma}} \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \left\| \sum_{l=1}^{\hat{\gamma}} (g_{i_0}(x_{i_0,0}, \zeta_{i_0,0,l}) - g_{i_0}(x_{i_0,0}, \zeta'_{i_0,0,l})) \right\| \\ &= \frac{\hat{\alpha}}{\hat{\gamma}} \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|g_{i_0}(x_{i_0,0}, \zeta_{i_0,0,l_1}) - g_{i_0}(x_{i_0,0}, \zeta'_{i_0,0,l_1})\| \\ &\leq \frac{\hat{\alpha}}{\hat{\gamma}} \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|g_{i_0}(x_{i_0,0}, \xi_{i_0, l_0}) - g_{i_0}(x_{i_0,0}, \xi'_{i_0, l_0})\| \leq \frac{\hat{\alpha}C}{\hat{\gamma}}. \end{aligned}$$

When $1 \leq k \leq K$, by (8) we have

$$\Delta_k^q = \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|x_{k+1} - x'_{k+1}\|. \quad (12)$$

Note that $x_{i,0} = x'_{i,0}$, $z_{i,k} = z'_{i,k}$ hold for any node $i \in \mathcal{V}$, $0 \leq k \leq K$, and $g_{j,m} = g'_{j,m}$ holds for any node $j \neq i_0$, $0 \leq m \leq k$. Then, by (5), $\tilde{x}_{j,k} = \tilde{x}'_{j,k}$ holds for any node $j \neq i_0$. Thus, by (7), $x_{j,k+1} = x'_{j,k+1}$ holds for any node $j \neq i_0$. Hence, (12) can be rewritten as

$$\Delta_k^q = \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|x_{i_0,k+1} - x'_{i_0,k+1}\|. \quad (13)$$

Then, substituting (5)-(7) into (13) implies

$$\begin{aligned} \Delta_k^q &= \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|(\tilde{x}_{i_0,k} - \tilde{x}'_{i_0,k}) - \hat{\alpha}(g_{i_0,k} - g'_{i_0,k})\| \\ &\leq \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|(1 - \hat{\beta})(x_{i_0,k} - x'_{i_0,k})\| \\ &\quad + \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \left\| \frac{\hat{\alpha}}{\hat{\gamma}} \sum_{l=1}^{\hat{\gamma}} (g_{i_0}(x_{i_0,k}, \zeta_{i_0,k,l}) - g_{i_0}(x_{i_0,k}, \zeta'_{i_0,k,l})) \right\|. \quad (14) \end{aligned}$$

Since \mathcal{D} and \mathcal{D}' are adjacent, there exists at most one pair of data samples $\zeta_{i_0,k,l_{k+1}}, \zeta'_{i_0,k,l_{k+1}}$ such that $\zeta_{i_0,k,l_{k+1}} = \xi_{i_0, l_0}$, $\zeta'_{i_0,k,l_{k+1}} = \xi'_{i_0, l_0}$. Then, (14) can be rewritten as

$$\begin{aligned} \Delta_k^q &\leq \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \left\| (1 - \hat{\beta})(x_{i_0,k} - x'_{i_0,k}) \right\| \\ &\quad + \frac{\hat{\alpha}}{\hat{\gamma}} \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|g_{i_0}(x_{i_0,k}, \xi_{i_0, l_0}) - g_{i_0}(x_{i_0,k}, \xi'_{i_0, l_0})\| \\ &\leq |1 - \hat{\beta}| \sup_{\text{Adj}(\mathcal{D}, \mathcal{D}')} \|x_{i_0,k} - x'_{i_0,k}\| + \frac{\hat{\alpha}C}{\hat{\gamma}}. \quad (15) \end{aligned}$$

By iteratively computing (15), this lemma is proved. \blacksquare

Next, we show that Algorithm 1 achieves the (ϵ_k, δ_k) -differential privacy at the k -th iteration for any $0 \leq k \leq K$.

Lemma 2: Given the mapping q and $0 < \epsilon_k < 1$, $\delta_k > 0$ for any $0 \leq k \leq K$. If there exists exactly one pair of data samples $\zeta_{i_0,k,l_{k+1}}, \zeta'_{i_0,k,l_{k+1}}$ in sample sets $\mathcal{D}_k, \mathcal{D}'_k$ such that (4) holds, then for any observation set $T^{(1)}, \dots, T^{(K)} \subset \mathbb{R}^{nr}$ and $1 \leq k \leq K - 1$, the randomized mapping $\mathcal{M}(\mathcal{D}_k) = Q(x_{k+1} + d_{k+1})$ satisfies

$$\begin{aligned} \mathbb{P}(\mathcal{M}(\mathcal{D}_0) \in T^{(1)}) &\leq e^{\epsilon_0} \mathbb{P}(\mathcal{M}(\mathcal{D}'_0) \in T^{(1)}) + \delta_0, \\ \mathbb{P}(\mathcal{M}(\mathcal{D}_k) \in T^{(k+1)} | \mathcal{M}(\mathcal{D}_0) \in T^{(1)}, \dots, \mathcal{M}(\mathcal{D}_{k-1}) \in T^{(k)}) \\ &\leq e^{\epsilon_k} \mathbb{P}(\mathcal{M}(\mathcal{D}'_k) \in T^{(k+1)} | \mathcal{M}(\mathcal{D}'_0) \in T^{(1)}, \dots, \mathcal{M}(\mathcal{D}'_{k-1}) \in T^{(k)}) + \delta_k, \end{aligned}$$

where $d_{k+1} = [d_{1,k+1}^\top, \dots, d_{n,k+1}^\top]^\top \sim N(0, \sigma_{k+1}^2 I_{nr})$ is a Gaussian noise with the variance $\sigma_{k+1}^2 = 4 \ln \left(\frac{1.25}{\delta_k} \right) \left(\frac{\Delta_k^q}{\epsilon_k} \right)^2$.

Proof. Note that the Gaussian noises d_1, d'_1 have the variance $\sigma_1^2 = 4 \ln \left(\frac{1.25}{\delta_0} \right) \left(\frac{\Delta_0^q}{\epsilon_0} \right)^2$. Then, by the Gaussian mechanism in Theorem A.1 of [19], $\mathbb{P}(x_1 + d_1 \in S^{(1)}) \leq e^{\epsilon_0} \mathbb{P}(x'_1 + d'_1 \in S^{(1)}) + \delta_0$ holds for any given observation set $S^{(1)} \in \mathbb{R}^r$. Let $T^{(1)} = Q(S^{(1)})$. Then, by the post-processing property in Proposition 2.1 of [19] we have $\mathbb{P}(\mathcal{M}(\mathcal{D}_0) \in T^{(1)}) \leq e^{\epsilon_0} \mathbb{P}(\mathcal{M}(\mathcal{D}'_0) \in T^{(1)}) + \delta_0$.

On the other hand, for any given observation set $S^{(k+1)} \in \mathbb{R}^r$ and $1 \leq k \leq K-1$, let $T^{(k+1)} = Q(S^{(k+1)})$. Then, since the Gaussian noises d_{k+1}, d'_{k+1} have the variance $\sigma_{k+1}^2 = 4 \ln \left(\frac{1.25}{\delta_k} \right) \left(\frac{\Delta_k^q}{\epsilon_k} \right)^2$, by the Gaussian mechanism in Theorem A.1 of [19] we have $\mathbb{P}(x_{k+1} + d_{k+1} \in S^{(k+1)} | z_1 \in T^{(1)}, \dots, z_k \in T^{(k)}) \leq e^{\epsilon_k} \mathbb{P}(x'_{k+1} + d'_{k+1} \in S^{(k+1)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)}) + \delta_k$. Thus, by the post-processing property in Proposition 2.1 of [19] we have $\mathbb{P}(z_{k+1} \in T^{(k+1)} | z_1 \in T^{(1)}, \dots, z_k \in T^{(k)}) \leq e^{\epsilon_k} \mathbb{P}(z'_{k+1} \in T^{(k+1)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)}) + \delta_k$. Therefore, this lemma is proved. ■

Lemma 3: Given $K \geq 1$ and $\varphi_k > 0$ for any $0 \leq k \leq K$. If $0 \leq y_k \leq 1$ holds for any $0 \leq k \leq K$, then $\prod_{k=0}^K (y_k + \varphi_k) - \prod_{k=0}^K y_k \leq \prod_{k=0}^K (1 + \varphi_k) - 1$.

Proof. Since the function $\prod_{k=0}^K (y_k + \varphi_k) - \prod_{k=0}^K y_k$ increases for any y_k satisfying $0 \leq y_k \leq 1$, we have $\prod_{k=0}^K (y_k + \varphi_k) - \prod_{k=0}^K y_k \leq (1 + \varphi_0) \prod_{k=1}^K (y_k + \varphi_k) - \prod_{k=1}^K y_k \leq (1 + \varphi_0)(1 + \varphi_1) \prod_{k=2}^K (y_k + \varphi_k) - \prod_{k=2}^K y_k \leq \dots \leq \prod_{k=0}^K (1 + \varphi_k) - 1$. Therefore, this lemma is proved. ■

Theorem 1: For any $K \geq 1, 0 \leq k \leq K$, let

$$\hat{\alpha} = \frac{a_1}{K^\alpha}, \hat{\beta} = \frac{a_2}{K^\beta}, \hat{\gamma} = \lfloor a_3 K^\gamma \rfloor + 1, \sigma_k = (k+1)^\sigma, \\ \delta_k = \frac{1}{(k+1)^\nu}, a_1, a_2, a_3 > 0.$$

If $0 < a_2 < K^\beta$ and $\nu > 0$, then Algorithm 1 achieves the (ϵ, δ) -differential privacy over finite iterations K , where

$$\epsilon = \sum_{k=0}^K \epsilon_k \leq \sum_{k=0}^K \frac{2C a_1 \sqrt{\ln(1.25(k+1)^\nu)}}{a_2 a_3 K^{\alpha+\gamma-\beta} (k+2)^\sigma}, \\ \delta = e^{\sum_{k=0}^K \epsilon_k} \left(\prod_{k=0}^K \left(1 + \frac{1}{(k+1)^\nu e^{\epsilon_k}} \right) - 1 \right). \quad (16)$$

Furthermore, if $\alpha + \gamma - \beta > \max\{1 - \sigma, 0\}$, $\nu > 1$, then Algorithm 1 achieves finite cumulative differential privacy parameters ϵ, δ over infinite iterations.

Proof. For $\text{Adj}(\mathcal{D}, \mathcal{D}')$ and any given observation set $T = \prod_{k=0}^K T^{(k)} \subseteq \text{Range}(\mathcal{M})$, by Lemma 2 we have

$$\frac{\mathbb{P}(\mathcal{M}(\mathcal{D}) \in T)}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in T)} = \frac{\mathbb{P}(z_1 \in T^{(1)}, \dots, z_K \in T^{(K)})}{\mathbb{P}(z'_1 \in T^{(1)}, \dots, z'_K \in T^{(K)})} \\ = \frac{\mathbb{P}(z_1 \in T^{(1)})}{\mathbb{P}(z'_1 \in T^{(1)})} \prod_{k=1}^{K-1} \frac{\mathbb{P}(z_{k+1} \in T^{(k+1)} | z_1 \in T^{(1)}, \dots, z_k \in T^{(k)})}{\mathbb{P}(z'_{k+1} \in T^{(k+1)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)})} \\ \leq \left(e^{\epsilon_0} + \frac{\delta_0}{\mathbb{P}(z'_1 \in T^{(1)})} \right) \cdot \\ \prod_{k=1}^{K-1} \left(e^{\epsilon_k} + \frac{\delta_k}{\mathbb{P}(z'_{k+1} \in T^{(k)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)})} \right), \quad (17)$$

where the differential privacy parameter $\epsilon_k = \frac{2\sqrt{\ln\left(\frac{1.25}{\delta_k}\right)\Delta_k^q}}{2C\hat{\alpha}\sqrt{\ln(1.25(k+1)^\nu)}(1-(1-\hat{\beta})^{k+1})} \frac{\sigma_{k+1}}{a_2 a_3 K^{\alpha+\gamma-\beta} (k+2)^\sigma} \leq \frac{2C a_1 \sqrt{\ln(1.25(k+1)^\nu)}}{a_2 a_3 K^{\alpha+\gamma-\beta} (k+2)^\sigma}$. Then, (17) can be rewritten as

$$\frac{\mathbb{P}(\mathcal{M}(\mathcal{D}) \in T)}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in T)} \leq \frac{(e^{\epsilon_0} \mathbb{P}(z'_1 \in T^{(1)}) + \delta_0)}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in T)} \\ \prod_{k=1}^{K-1} \left(e^{\epsilon_k} \mathbb{P}(z'_{k+1} \in T^{(k)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)}) + \delta_k \right) \\ = \frac{e^{\sum_{k=0}^K \epsilon_k}}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in T)} \left(\mathbb{P}(z'_1 \in T^{(1)}) + e^{-\epsilon_0} \delta_0 \right) \\ \prod_{k=1}^{K-1} \left(\mathbb{P}(z'_{k+1} \in T^{(k)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)}) + e^{-\epsilon_k} \delta_k \right) \\ = e^{\sum_{k=0}^K \epsilon_k} + \frac{e^{\sum_{k=0}^K \epsilon_k}}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in T)} \left(\mathbb{P}(z'_1 \in T^{(1)}) + e^{-\epsilon_0} \delta_0 \right) \\ \prod_{k=1}^{K-1} \left(\mathbb{P}(z'_{k+1} \in T^{(k)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)}) + e^{-\epsilon_k} \delta_k \right) \\ - \frac{e^{\sum_{k=0}^K \epsilon_k}}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in T)} \mathbb{P}(z'_1 \in T^{(1)}) \\ \prod_{k=1}^{K-1} \mathbb{P}(z'_{k+1} \in T^{(k)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)}). \quad (18)$$

Note that $0 \leq \mathbb{P}(z'_1 \in T^{(1)}) \leq 1$, $0 \leq \mathbb{P}(z'_{k+1} \in T^{(k+1)} | z'_1 \in T^{(1)}, \dots, z'_k \in T^{(k)}) \leq 1$ and $e^{-\epsilon_k} \delta_k > 0$. Then, by Lemma 3 (18) can be rewritten as

$$\frac{\mathbb{P}(\mathcal{M}(\mathcal{D}) \in T)}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in T)} \leq e^{\sum_{k=0}^K \epsilon_k} + \frac{e^{\sum_{k=0}^K \epsilon_k} (\prod_{k=0}^K (1 + e^{-\epsilon_k} \delta_k) - 1)}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in T)}.$$

Let $\epsilon = \sum_{k=0}^K \epsilon_k$, $\delta = e^{\sum_{k=0}^K \epsilon_k} (\prod_{k=0}^K (1 + e^{-\epsilon_k} \delta_k) - 1)$. Then by Definition 2, Algorithm 1 achieves the (ϵ, δ) -differential privacy, where the cumulative differential privacy parameter is $\epsilon \leq \sum_{k=0}^K \frac{2C a_1 \sqrt{\ln(1.25(k+1)^\nu)}}{a_2 a_3 K^{\alpha+\gamma-\beta} (k+2)^\sigma} \leq \sum_{k=0}^K \frac{2C a_1 \sqrt{\ln(1.25(K+1)^\nu)}}{a_2 a_3 K^{\alpha+\gamma-\beta} (K+2)^{\min\{0, \sigma\}}} = O\left(\frac{\sqrt{\ln(K+1)}}{K^{\alpha+\gamma-\beta-\max\{1-\sigma, 0\}}}\right)$.

Thus, if $\alpha + \gamma - \beta > \max\{1 - \sigma, 0\}$, then the cumulative differential privacy parameter $\lim_{K \rightarrow \infty} \sum_{k=0}^K \epsilon_k$ is finite. Note that $e^{\epsilon_k} \geq 1$ for any $0 \leq k \leq K$. Then, we have $\delta = e^{\sum_{k=0}^K \epsilon_k} (\prod_{k=0}^K (1 + \frac{1}{(k+1)^\delta e^{\epsilon_k}}) - 1) \leq e^{\sum_{k=0}^K \epsilon_k} (\prod_{k=0}^K (1 + \frac{1}{(k+1)^\delta}) - 1) \leq e^{\sum_{k=0}^{\infty} \epsilon_k} (\prod_{k=0}^{\infty} (1 + \frac{1}{(k+1)^\delta}) - 1) < \infty$. Hence, if $\nu > 1$, then the cumulative differential privacy parameter δ is finite. In this case, Algorithm 1 achieves finite cumulative differential privacy parameters ϵ, δ over infinite iterations. ■

Remark 7: Theorem 1 shows how step-size parameters α, β , the sample-size parameter γ and the privacy noise parameter σ_k affect cumulative differential privacy parameters ϵ, δ . As shown in (16), the larger the step-size parameter α , the sample-size parameter γ and the privacy noise parameter σ_k are, the smaller cumulative differential privacy parameters ϵ, δ are. In addition, the smaller the step-size parameter β is, the smaller cumulative differential privacy parameters ϵ, δ are.

Remark 8: By (16), the larger the sample-size $\hat{\gamma}$ is, the smaller cumulative differential privacy parameters ϵ, δ are. Then, the larger the sample-size $\hat{\gamma}$ is, the less privacy noises are required to achieve the same (ϵ, δ) -differential privacy, and thus the effect of privacy noises $d_{i,k}$ is reduced.

Remark 9: The sample-size $\hat{\gamma}$ is not required to go to infinity to achieve finite cumulative differential privacy parameters

ϵ , δ over infinite iterations. Specifically, let the sample-size parameter $\gamma = 0$. Then, the sample-size $\hat{\gamma}$ is constant. In this case, if $\alpha - \beta > \max\{1 - \sigma, 0\}$, $\nu > 1$, then Algorithm 1 can achieve finite cumulative differential privacy parameters ϵ , δ over infinite iterations. This shows the advantage over [24]–[33], since cumulative differential privacy parameters ϵ , δ go to infinity therein.

C. Convergence analysis

In this subsection, we will give the convergence analysis of Algorithm 1. First, we introduce an assumption on step-sizes, sample-size and the privacy noise parameter.

Assumption 4: For any $K \geq 1, 0 \leq k \leq K$, step-sizes $\hat{\alpha} = \frac{a_1}{K^\alpha}$, $\hat{\beta} = \frac{a_2}{K^\beta}$, the sample-size $\hat{\gamma} = \lfloor a_3 K^\gamma \rfloor + 1$ and the privacy noise parameter $\sigma_k = (k+1)^\sigma$ satisfy $a_1, a_2, a_3 > 0$, $2\alpha - \beta > 1$, $\frac{1}{2} + \max\{\sigma, 0\} < \beta < \alpha < 1$.

Next, we first provide the mean square convergence of Algorithm 1, and then show the convergence rate of Algorithm 1 for cost functions satisfying the Polyak-Łojasiewicz condition.

1) Mean square convergence:

Theorem 2: If Assumptions 1-4 hold, then for any node $i \in \mathcal{V}$ and $K \geq 1$, we have $\liminf_{K \rightarrow \infty} \mathbb{E} \|\nabla F(x_{i,K+1})\|^2 = 0$.

Proof. See Appendix B. ■

Remark 10: Note that the mean square convergence of Algorithm 1 is achieved without the assumption of bounded gradients. Then, this is different from [9], [10], [27], [28], [32], [33], where [27], [32] require the assumption of bounded gradients and [9], [10], [28], [33] do not achieve the mean square convergence. The key to achieving the mean square convergence is to ensure the bounded expectation of the gradient $\mathbb{E} \|\nabla F(\bar{x}_{K+1})\|^2$ by proving $\mathbb{E}(F(\bar{x}_{K+1}) - F^*)$ is bounded for any point x_K and $K \geq 1$ without the assumption of bounded gradients. As a result, the mean square convergence is achieved with a more general framework than [27], [32].

2) Convergence rate analysis:

Assumption 5: (Polyak-Łojasiewicz) The global cost function $F(x)$ satisfies the Polyak-Łojasiewicz condition, i.e., there exists $\mu > 0$ such that for any $x \in \mathbb{R}^r$, $2\mu(F(x) - F^*) \leq \|\nabla F(x)\|^2$.

Remark 11: Assumption 5 is commonly used (see e.g. [5], [7]), and means that the gradient $\nabla F(x)$ to grow faster than a quadratic function as the algorithm moves away from the optimal solution. Such functions exist, for example, $F(x) = x^2 + 3\sin^2 x$ is a nonconvex function satisfying Assumption 5 for any $0 < \mu < 0.3$. As shown in Theorem 2 of [43], Assumption 5 is more general than the convex cost functions assumed in [8], [11], [24]–[29].

Theorem 3: If Assumptions 1-5 hold, then for any node $i \in \mathcal{V}$, $K \geq 1$ and $1 \leq \psi \leq 2$, we have $\mathbb{E} \|\nabla F(x_{i,K+1})\|^\psi = O(K^{-\frac{\psi}{2} \min\{2\beta - 2 \max\{\sigma, 0\} - 1, 2\alpha - \beta - 1\}})$. Furthermore, when $\psi = 2$, $\mathbb{E}(F(x_{i,K+1}) - F^*) = O(K^{-\min\{2\beta - 2 \max\{\sigma, 0\} - 1, 2\alpha - \beta - 1\}})$, and the mean square convergence of Algorithm 1 is achieved as K goes to infinity, i.e., for any node $i \in \mathcal{V}$, $\lim_{K \rightarrow \infty} \mathbb{E} \|\nabla F(x_{i,K+1})\|^2 = 0$.

Proof. See Appendix C. ■

Remark 12: When the quantized information z_k is exchanged between neighboring nodes, the introduced quantization error e_k brings difficulty to the convergence analysis

of Algorithm 1. To combat this effect, the step-size $\hat{\beta}$ is introduced. The mean square convergence of Algorithm 1 is guaranteed by $\lim_{K \rightarrow \infty} \hat{\beta}^2 \Delta^2 = 0$. From this point of view, Algorithm 1 can also solve the adaptive quantization problem ([3]) and the probabilistic quantization problem ([4]). Moreover, from (56) it follows that the larger the quantization error Δ is, the larger $\theta_{k,2}$ is, and thus the slower the convergence rate is. Therefore, the probabilistic quantization does slow down the convergence rate of Algorithm 1.

Remark 13: The mean square convergence of Algorithm 1 is guaranteed for general privacy noises, including increasing, constant (see e.g. [25], [27]–[31]) and decreasing (see e.g. [24], [26]) privacy noises. This is non-trivial even without considering privacy protection problem. For example, let $\hat{\alpha} = \frac{1}{K^{0.9}}$, $\hat{\beta} = \frac{1}{K^{0.75}}$. Then, the convergence of Algorithm 1 holds as long as the privacy noise parameter σ_k has an increasing rate no more than $O(k^{0.25})$.

Remark 14: Note that by Theorem 2, the mean square convergence of Algorithm 1 holds for general cost functions, including convex and nonconvex cost functions. Then, when the global cost function is convex, Theorem 2 also holds. Furthermore, if the global cost function $F(x)$ is λ -strongly convex, i.e., there exists $\lambda > 0$ such that for any $x, y \in \mathbb{R}^r$, $F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\lambda}{2} \|y - x\|^2$, then by Lemma 6.9 in [45] we have $2\lambda(F(x) - F^*) \leq \|\nabla F(x)\|^2$, which means the global cost function $F(x)$ satisfies Assumption 5. Thus, Algorithm 1 achieves the same convergence rate as Theorem 3.

Remark 15: Note that distributed nonconvex stochastic optimization algorithms may converge to a saddle point instead of the desired global minimum. Then, the discussion of the avoidance of saddle points is necessary. Assumption 5 implies that each stationary point x^* of F satisfying $\nabla F(x^*) = 0$ is a global minimum of F , and thus guarantees the avoidance of saddle points discussed in [31]. Furthermore, compared with [7], [10], [24]–[29], [31], [32], Assumption 5 helps us to give the convergence rate of Algorithm 1 without the assumption of bounded gradients.

In practice, the time and number of running a distributed stochastic optimization algorithm are usually limited by various constraints, while selecting the best one from lots of running results is very time-consuming. To address this issue and guarantee the convergence of a single running result with any given probability, the following low-probability convergence rate of Algorithm 1 is given based on Theorem 3.

Corollary 1: Under Assumptions 1-5, for any node $i \in \mathcal{V}$, $K \geq 1$ and $0 < \delta^* < 1$, with probability at least $1 - \delta^*$, we have $F(x_{i,K+1}) - F^* = O\left(\frac{1}{K^{\min\{2\beta - 2 \max\{\sigma, 0\} - 1, 2\alpha - \beta - 1\}}}\right)$.

Proof. By Theorem 3, there exists $A_1 > 0$ such that for any node $i \in \mathcal{V}$, $\mathbb{E}(F(x_{i,K+1}) - F^*) \leq \frac{A_1}{K^{\min\{2\beta - 2 \max\{\sigma, 0\} - 1, 2\alpha - \beta - 1\}}}$. For any $0 < \delta^* < 1$, let $a = \frac{A_1}{\delta^* K^{\min\{2\beta - 2 \max\{\sigma, 0\} - 1, 2\alpha - \beta - 1\}}}$. Then, by Markov's inequality ([44]) we have

$$\mathbb{P}(F(x_{i,K+1}) - F^* > a) \leq \frac{\mathbb{E}(F(x_{i,K+1}) - F^*)}{a} \leq \delta^*. \quad (19)$$

Thus, by (19) we have $F(x_{i,K+1}) - F^* \leq \frac{A_1}{\delta^* K^{\min\{2\beta - 2 \max\{\sigma, 0\} - 1, 2\alpha - \beta - 1\}}}$ with probability at least $1 - \delta^*$. Therefore, this corollary is proved. ■

Remark 16: Corollary 1 guarantees the convergence of a single running result with probability at least $1 - \delta^*$, and thus avoids spending time on selecting the best one from lots of running results. Moreover, from Theorem 1, it follows that the low-probability convergence rate is affected by the failure probability δ^* . The larger the failure probability δ^* is, the faster the low-probability convergence rate is.

D. Trade-off between privacy and utility

Based on Theorems 1-3, the mean square convergence of Algorithm 1 as well as the differential privacy with finite cumulative differential privacy parameters ϵ , δ over infinite iterations can be established simultaneously, which is given in the following corollary:

Corollary 2: For any $0 \leq k \leq K$, let

$$\hat{\alpha} = \frac{a_1}{K^\alpha}, \quad \hat{\beta} = \frac{a_2}{K^\beta}, \quad \hat{\gamma} = \lfloor a_3 K^\gamma \rfloor + 1, \quad \sigma_k = (k+1)^\sigma, \\ \delta_k = \frac{1}{(k+1)^\nu}, \quad a_1, a_2, a_3 > 0.$$

If Assumptions 1-3, 5 hold, and $\nu > 1$, $\frac{1}{2} + \max\{\sigma, 0\} < \beta < \alpha < 1$, $\alpha + \gamma - \beta > \max\{1 - \sigma, 0\}$, $2\alpha - \beta > 1$, then Algorithm 1 achieves the mean square convergence and finite cumulative differential privacy parameters ϵ , δ over infinite iterations simultaneously as the sample-size $\hat{\gamma}$ goes to infinity. **Proof.** By Theorems 1-3, this corollary is proved. ■

Remark 17: Corollary 2 holds even when privacy noises have increasing variances. For example, when $\alpha = 1, \beta = 0.8, \sigma = 0.2, \gamma = 0.7, \nu = 1.5$, or $\alpha = 0.9, \beta = 0.6, \sigma = 0.05, \gamma = 0.8, \nu = 2$, the conditions of Corollary 2 hold. In this case, the differential privacy with finite cumulative privacy parameters ϵ , δ over infinite iterations as well as the mean square convergence can be established simultaneously.

Remark 18: The result of Corollary 2 does not contradict the trade-off between privacy and utility. In fact, to achieve differential privacy, Algorithm 1 incurs a compromise on the utility. However, different from [28], [33] which compromise convergence accuracy to enable differential privacy, Algorithm 1 compromises the convergence rate and the sample-size (which are also utility metrics) instead. From Corollary 2, it follows that the larger the privacy noise parameter σ_k is, the slower the mean square convergence rate is. Besides, the sample-size $\hat{\gamma}$ is required to go to infinity when the mean square convergence of Algorithm 1 and finite cumulative privacy parameters ϵ , δ over infinite iterations are considered simultaneously. The ability to retain convergence accuracy makes our approach suitable for accuracy-critical scenarios.

E. Oracle complexity

Since the subsampling method controlled through the sample-size parameter γ is employed in Algorithm 1, the total number of data samples to obtain an optimal solution is an issue worthy of attention. To show this, we give the definitions of η -optimal solutions and the oracle complexity as follows:

Definition 4: (η -optimal solution) Given $\eta > 0$, $x_K = [x_{1,K}^\top, \dots, x_{n,K}^\top]^\top$ is an η -optimal solution if for any node $i \in \mathcal{V}$, $\mathbb{E}|F(x_{i,K}) - F^*| < \eta$.

Definition 5: Given $\eta > 0$, the oracle complexity is the total number of data samples to obtain an η -optimal solution $\sum_{k=0}^{N(\eta)} \hat{\gamma}$, where $N(\eta) = \min\{K : x_K \text{ is an } \eta\text{-optimal solution}\}$.

Based on Theorem 3, Definitions 4 and 5, the oracle complexity of Algorithm 1 for obtaining an η -optimal solution is given as follows:

Theorem 4: Given $0 < \eta < \frac{1}{2}$, let $\alpha = 1 - \eta$, $\beta = \frac{1}{3} - \frac{2}{3}\eta$, $\sigma = \eta$, $\gamma = \eta$. Then, under Assumptions 1-3 and 5, the oracle complexity of Algorithm 1 is $O(\eta^{-\frac{3+3\eta}{1-2\eta}})$.

Proof. For the given $\eta > 0$, let the iteration limit in Algorithm 1 be $N(\eta)$. Then, we have $\hat{\gamma} = \lfloor a_3 N(\eta)^\eta \rfloor + 1 \leq a_3 N(\eta)^\eta + 1$.

Note that by Theorem 3, there exists $A_1 > 0$ such that

$$\mathbb{E}|F(x_{i,K+1}) - F^*| = \mathbb{E}(F(x_{i,K+1}) - F^*) \leq \frac{A_1}{K^{\frac{1}{3} - \frac{2}{3}\eta}}. \quad (20)$$

Then, when $K \geq \lfloor (\frac{A_1}{\eta})^{\frac{3}{1-2\eta}} \rfloor + 1 > (\frac{A_1}{\eta})^{\frac{3}{1-2\eta}}$, (20) can be rewritten as

$$\mathbb{E}|F(x_{i,K+1}) - F^*| \leq \frac{A_1}{K^{\frac{1}{3} - \frac{2}{3}\eta}} < \frac{A_1}{(\frac{A_1}{\eta})^{\frac{3}{1-2\eta}}} = \eta. \quad (21)$$

Thus, by (21) and Definition 4, x_{K+1} is an η -optimal solution.

Since $N(\eta)$ is the smallest integer such that $x_{N(\eta)}$ is an η -optimal solution, we have

$$N(\eta) \leq 1 + \min\{K : K \geq \lfloor (\frac{A_1}{\eta})^{\frac{3}{1-2\eta}} \rfloor + 1\} = \lfloor (\frac{A_1}{\eta})^{\frac{3}{1-2\eta}} \rfloor + 2. \quad (22)$$

Hence, by Definition 5 and (22), we have

$$\sum_{k=0}^{N(\eta)} \hat{\gamma} = (N(\eta) + 1)\hat{\gamma} \leq (N(\eta) + 1)(a_3 N(\eta)^\eta + 1) \\ = O(N(\eta)^{1+\eta}) = O\left(\eta^{-\frac{3+3\eta}{1-2\eta}}\right).$$

Therefore, this theorem is proved. ■

Remark 19: From Theorems 3 and 4, the faster the convergence rate is, the smaller the oracle complexity is. For example, if $\eta = 0.3$, then the total number of data samples to obtain an η -optimal solution is $O(10^5)$, which does not go to infinity. This requirement for the total number of data samples is acceptable since the computational cost of centralized stochastic gradient descent is $O(10^5)$ to achieve the same accuracy as Algorithm 1.

IV. NUMERICAL EXAMPLES

In this section, we verify the effectiveness and advantages of Algorithm 1 by the distributed training of a convolutional neural network (CNN) on the ‘‘MNIST’’ dataset ([46]). Specifically, five nodes cooperatively train a CNN using the ‘‘MNIST’’ dataset over a topology depicted in Fig. 1, which satisfies Assumption 1. Then, the ‘‘MNIST’’ dataset is divided into two subdatasets for training and testing, respectively. The training dataset is uniformly divided into 5 subdatasets consisting of 12000 binary images, and each of them can only be accessed by one agent to update its model parameters. In the following, the effect of the noise and the quantization on convergence, the differential privacy level, and the comparison with methods in [25]–[31] are presented for Algorithm 1, respectively.

A. Effect of the noise and the quantization on convergence

Let step-sizes $\hat{\alpha} = \frac{9.35}{2000^{0.9}} \approx 10^{-2}$, $\hat{\beta} = \frac{0.2}{2000^{0.7}} \approx 10^{-3}$, the sample-size $\hat{\gamma} = \lfloor 5.5 \cdot 10^{-4} \cdot 2000^{1.5} \rfloor + 1 = 50$, $\delta_k = \frac{1}{(k+1)^3}$, and the privacy noise parameter $\sigma_k = (k+5)^\sigma$ with $\sigma = -0.1, 0.1, 0.2$, respectively. The probabilistic quantizer is

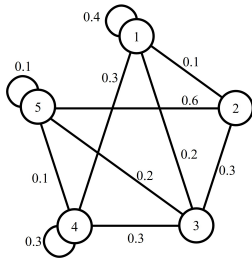


Fig. 1: Topology structure of the undirected graph

given in the form of (3) with $\Delta = 1, 5, 10$, respectively. Then, it can be seen that Assumptions 2-5 hold. The training and testing accuracy on the “MNIST” dataset are presented in Figs. 2 and 3, from which one can see that as iterations increase, the training and testing accuracy increase. More importantly, the smaller Δ and σ are, the faster Algorithm 1 converges, which is consistent with Theorem 3.

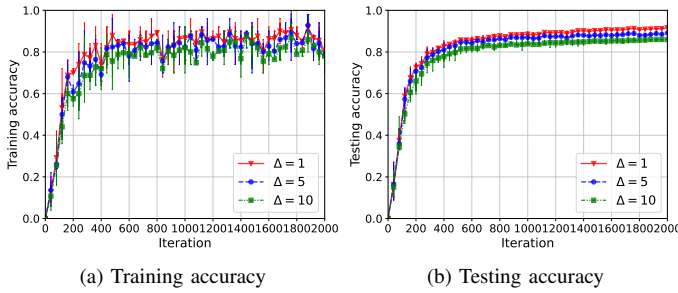


Fig. 2: Accuracy of Algorithm 1 with $\Delta = 1, 5, 10$

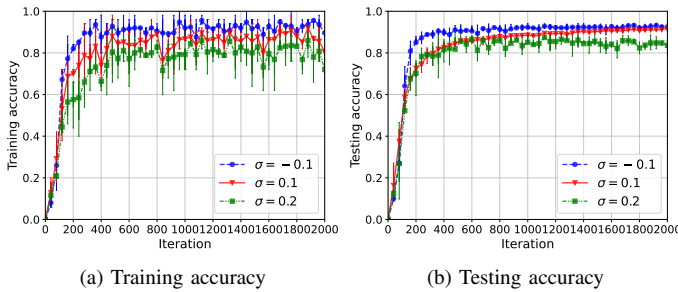


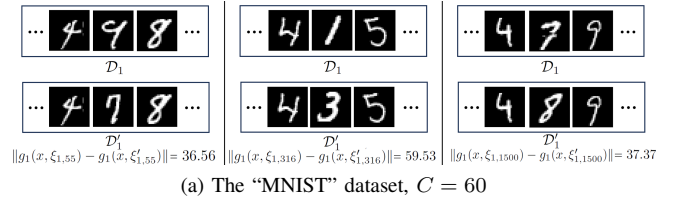
Fig. 3: Accuracy of Algorithm 1 with $\sigma = -0.1, 0.1, 0.2$

B. Differential privacy level

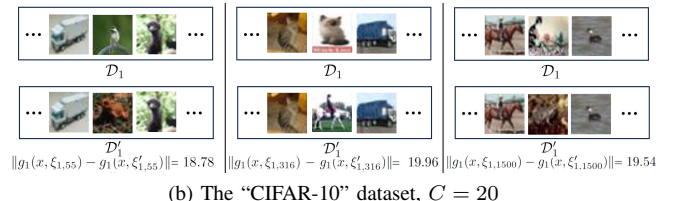
First, we show that the boundary C in Definition 1 is related to the distribution of the dataset. The boundary C is set as the maximum magnitude of sampled gradients when changing one data sample in the dataset, and the relation between the boundary C and the distribution of the dataset is given for the “MNIST”, “CIFAR-10”([47]) and “CIFAR-100”([48], [49]) dataset, respectively. For each dataset, we randomly change one data sample and compute the magnitude of sampled gradients. Due to the space limitation, only three examples are given for each dataset in Fig. 4. Fig. 4(a) shows that for the “MNIST” dataset, the magnitude of sampled gradients when respectively changing the 55th, 316th, 1500th data sample is 36.56, 59.53, 37.37, which is no more than the boundary $C = 60$. Similarly, Fig. 4(b) and 4(c) show that the magnitude of sampled gradients is no more than the boundary $C = 20$ and 19.5, respectively.

Then, based on the model inversion attack given in [42], we compare Algorithm 1 and the algorithms without privacy protection in [6], [7] to show that Algorithm 1 can prevent adversaries inferring the sensitive information from sampled gradients. A comparison of privacy protection between Algorithm 1 and distributed SGD on the “MNIST” dataset is presented in Fig. 5, from which one can see that adversaries cannot recover original handwritten digit images in Algorithm 1, while adversaries can completely recover original handwritten digit images in distributed SGD.

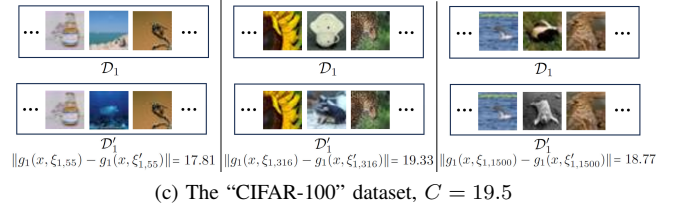
Next, the relationship of the cumulative differential privacy parameter ϵ over infinite iterations, the privacy noise parameter σ and sample-size parameter γ is presented in Fig. 6, from which one can see that as the privacy noise parameter σ and the sample-size parameter γ increase, the cumulative differential privacy parameter ϵ decrease. This is consistent with the privacy analysis in Subsection III-B. Moreover, in the first 2000 iterations, the cumulative differential privacy parameters $\epsilon = 0.7205$ and $\delta = 0.2021$, which is consistent with Theorem 1.



(a) The “MNIST” dataset, $C = 60$



(b) The “CIFAR-10” dataset, $C = 20$



(c) The “CIFAR-100” dataset, $C = 19.5$

Fig. 4: Different boundary C for different datasets

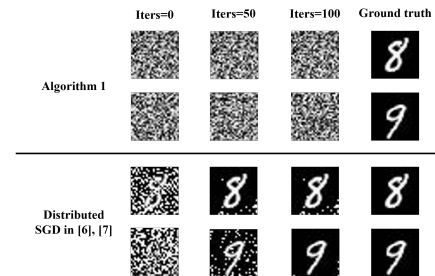


Fig. 5: Comparison of privacy protection between Algorithm 1 and distributed SGD in [6], [7]

C. Comparison with methods in [25]–[31]

Let $\Delta = 1, \sigma = 0.1$ in Algorithm 1. Then, the comparison of accuracy between Algorithm 1 and methods in [25]–[31] is

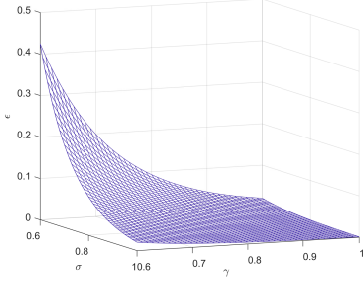


Fig. 6: Relationship of ϵ , σ and γ

presented in Fig. 7. To ensure a fair comparison, we set the same step-sizes in [25], [27], [31] as this paper, and the step-sizes in [26], [28]–[30] as chosen therein. In addition, we set sample-sizes in [25]–[31] as chosen therein. From Figs. 7(a) and 7(b), it can be seen that the convergence rate of Algorithm 1 is faster than [25]–[31].

When $C = 60$, the adjacency relation of this paper is equivalent to [25]–[31]. Then, we can compare the differential privacy level between Algorithm 1 and methods therein. The comparison of cumulative differential privacy parameters ϵ and δ is presented in Fig. 8. From Figs. 8(a) and 8(b) one can see that cumulative differential privacy parameters ϵ , δ of Algorithm 1 are bounded by finite constants over infinite iterations, while cumulative differential privacy parameters ϵ , δ in [25]–[31] go to infinity over infinite iterations. Based on the above discussions, Algorithm 1 not only converges, but also provides smaller cumulative differential privacy parameters ϵ , δ over infinite iterations.

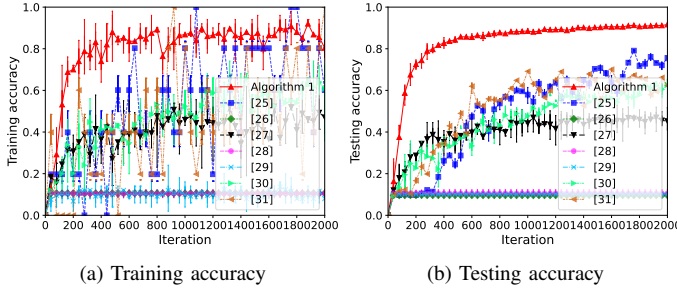


Fig. 7: Comparison of accuracy

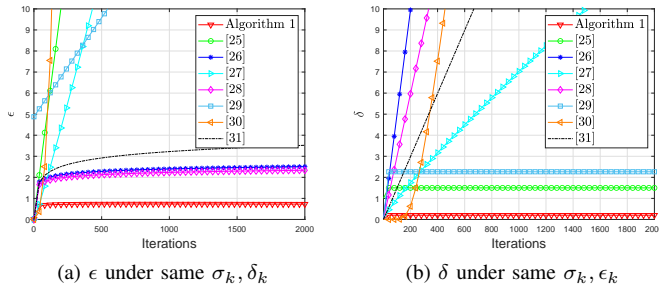


Fig. 8: Comparison of cumulative differential privacy parameters ϵ and δ

V. CONCLUSION

In this paper, we have proposed a differentially private distributed nonconvex stochastic optimization algorithm with quantized communication. In the proposed algorithm, general privacy noises are added to each node’s local states to

prevent information leakage, and then a probabilistic quantizer is employed on noise-perturbed states to improve the communication efficiency. By using the subsampling method controlled through the sample-size parameter, the differential privacy level of the algorithm is enhanced compared with the existing ones. By using the two-time-scale step-sizes method, the mean square convergence for nonconvex cost functions is given. Then, under the Polyak–Łojasiewicz condition, the mean square convergence rate and the oracle complexity of the algorithm are given. Meanwhile, the trade-off between the privacy and the utility is shown. Finally, a numerical example of the distributed training of CNN on the “MNIST” dataset is given to verify the effectiveness of the algorithm.

APPENDIX A USEFUL LEMMAS

Lemma A.1: If Assumption 2(i) holds for a function $h : \mathbb{R}^r \rightarrow \mathbb{R}$, and $\min_{x \in \mathbb{R}^r} h(x) = h^* > -\infty$, then the following results hold: (i) For any $x, y \in \mathbb{R}^r$, $h(y) \leq h(x) + \langle \nabla h(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$; (ii) For any $x \in \mathbb{R}^r$, $\|\nabla h(x)\|^2 \leq 2L(h(x) - h^*)$.

Proof. Lemma A.1(i) is directly from Lemma 3.4 of [45]. To prove Lemma A.1(ii), by (3.5) in [45], we have $\|\nabla h(x)\|^2 \leq 2L(h(x) - h(x - \frac{1}{L}\nabla h(x))) \leq 2L(h(x) - h^*)$. ■

Lemma A.2: If for any node $i \in \mathcal{V}$, Assumption 2(i) holds for the cost function $f_i(x)$, then Assumption 2(i) holds for the global cost function $F(x)$.

Proof. Note that by Assumption 2(i), $f_i(x)$ has Lipschitz continuous gradients for any node $i \in \mathcal{V}$. Then, for any $x, y \in \mathbb{R}^r$, we have $\|\nabla F(x) - \nabla F(y)\| = \|\frac{1}{n} \sum_{i=1}^n (\nabla f_i(x) - \nabla f_i(y))\| \leq \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(y)\| \leq \frac{L}{n} \sum_{i=1}^n \|x - y\| = L\|x - y\|$. Therefore, this lemma is proved. ■

APPENDIX B PROOF OF THEOREM 2

To provide an explanation of our results clearly, define

$$\begin{aligned} \nabla f^k &\triangleq [\nabla f_1(x_{1,k})^\top, \nabla f_2(x_{2,k})^\top, \dots, \nabla f_n(x_{n,k})^\top]^\top, \\ \nabla f(\bar{x}_k) &\triangleq [\nabla f_1(\bar{x}_k)^\top, \nabla f_2(\bar{x}_k)^\top, \dots, \nabla f_n(\bar{x}_k)^\top]^\top, \\ W &\triangleq I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top, \quad Y_k \triangleq (W \otimes I_r) x_k, \\ e_k &\triangleq z_k - x_k - d_k, \quad w_k \triangleq g^k - \nabla f^k, \\ \bar{x}_k &\triangleq \frac{1}{n} (\mathbf{1}_n^\top \otimes I_r) x_k, \quad \bar{w}_k \triangleq \frac{1}{n} (\mathbf{1}_n^\top \otimes I_r) w_k, \\ \overline{\nabla f^k} &\triangleq \frac{1}{n} (\mathbf{1}_n^\top \otimes I_r) \nabla f^k = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_{i,k}). \end{aligned}$$

Then, we can express (7) for all nodes in a compact form as follows:

$$\begin{aligned} x_{k+1} &= ((I_n - \hat{\beta}\mathcal{L}) \otimes I_r) x_k - \hat{\alpha} \nabla f^k \\ &\quad + \hat{\beta} (\mathcal{A} \otimes I_r) (e_k + d_k) - \hat{\alpha} w_k. \end{aligned} \quad (23)$$

Next, the following six steps are given to prove Theorem 2.

Step 1: We first consider the term $\|Y_k\|^2$. Note that $W(I_n - \hat{\beta}\mathcal{L}) = (I_n - \hat{\beta}\mathcal{L})W$. Then, multiplying both sides of (23) by $W \otimes I_r$ gives

$$\begin{aligned} Y_{k+1} &= ((I_n - \hat{\beta}\mathcal{L}) \otimes I_r) Y_k - \hat{\alpha} (W \otimes I_r) \nabla f^k \\ &\quad + \hat{\beta} (AW \otimes I_r) (e_k + d_k) - \hat{\alpha} (W \otimes I_r) w_k. \end{aligned} \quad (24)$$

Since $d_{i,k} \sim N(0, \sigma_k^2 I_r)$, we have

$$\mathbb{E}(d_k) = 0, \quad (25)$$

$$\mathbb{E}\|d_k\|^2 = nr\sigma_k^2. \quad (26)$$

Since $w_k = g^k - \nabla f^k$, $g^k = [g_{1,k}^\top, \dots, g_{n,k}^\top]^\top$ and $g_{i,k} = \frac{1}{\hat{\gamma}} \sum_{l=1}^{\hat{\gamma}} g_i(x_{i,k}, \zeta_{i,k,l})$, by Assumption 2(iii) we have

$$\mathbb{E}w_k = \mathbb{E}g^k - \nabla f^k = 0, \quad (27)$$

$$\mathbb{E}\|w_k\|^2 = \mathbb{E}\|g^k - \nabla f^k\|^2 \leq \frac{n\sigma_g^2}{\hat{\gamma}}. \quad (28)$$

Since $e_k = z_k - x_k - d_k = [(z_{1,k} - x_{1,k} - d_{1,k})^\top, \dots, (z_{n,k} - x_{n,k} - d_{n,k})^\top]^\top$, $z_{i,k} = [Q(x_{i,k}^{(1)} + d_{i,k}^{(1)}), \dots, Q(x_{i,k}^{(r)} + d_{i,k}^{(r)})]^\top$, by Assumption 3 we have

$$\mathbb{E}e_k = \mathbb{E}(z_k - x_k - d_k) = 0, \quad (29)$$

$$\mathbb{E}\|e_k\|^2 = \mathbb{E}\|z_k - x_k - d_k\|^2 \leq nr\Delta^2. \quad (30)$$

By (25), (27) and (29), taking mathematical expectation of $\|Y_{k+1}\|^2$ leads to

$$\begin{aligned} & \mathbb{E}\|Y_{k+1}\|^2 \\ &= \mathbb{E}\left\| \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_d \right) Y_k - \hat{\alpha}(W \otimes I_d) \nabla f^k \right. \\ & \quad \left. + \hat{\beta}(\mathcal{A}W \otimes I_d)(e_k + n_k) - \hat{\alpha}(W \otimes I_d)w_k \right\|^2 \\ &= \mathbb{E}\left\| \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_d \right) Y_k - \hat{\alpha}(W \otimes I_d) \nabla f^k \right\|^2 \\ & \quad + \hat{\beta}^2 \mathbb{E}\left\| (\mathcal{A}W \otimes I_d)(n_k + e_k) \right\|^2 + \hat{\alpha}^2 \mathbb{E}\| (W \otimes I_d)w_k \|^2 \\ & \quad + 2\mathbb{E}\left\langle \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_d \right) Y_k - \hat{\alpha}(W \otimes I_d) \nabla f^k, (\mathcal{A}W \otimes I_d)(n_k + e_k) \right\rangle \\ & \quad + 2\mathbb{E}\left\langle \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_d \right) Y_k - \hat{\alpha}(W \otimes I_d) \nabla f^k, (W \otimes I_d)w_k \right\rangle \\ & \quad + 2\mathbb{E}\langle (\mathcal{A}W \otimes I_d)(n_k + e_k), (W \otimes I_d)w_k \rangle \\ &= \mathbb{E}\left\| \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_d \right) Y_k - \hat{\alpha}(W \otimes I_d) \nabla f^k \right\|^2 \\ & \quad + \hat{\beta}^2 \mathbb{E}\left\| (\mathcal{A}W \otimes I_d)(n_k + e_k) \right\|^2 \\ & \quad + \hat{\alpha}^2 \mathbb{E}\| (W \otimes I_d)w_k \|^2. \end{aligned} \quad (31)$$

For any $0 \leq k \leq K$, let $\mathcal{F}_k = \sigma(x_k, n_k)$. Then, by the law of total expectation, we have

$$\begin{aligned} & \mathbb{E}\langle ((\mathcal{A}W)^\top \mathcal{A}W) \otimes I_d n_k, e_k \rangle \\ &= \mathbb{E}(\mathbb{E}(\langle ((\mathcal{A}W)^\top \mathcal{A}W) \otimes I_d n_k, e_k \rangle | \mathcal{F}_k)) \\ &= \mathbb{E}\langle ((\mathcal{A}W)^\top \mathcal{A}W) \otimes I_d n_k, \mathbb{E}(e_k | \mathcal{F}_k) \rangle \\ &= \mathbb{E}\langle ((\mathcal{A}W)^\top \mathcal{A}W) \otimes I_d n_k, 0 \rangle = 0. \end{aligned} \quad (32)$$

Thus, substituting equation (32) into equation (31) implies

$$\begin{aligned} \mathbb{E}\|Y_{k+1}\|^2 &= \mathbb{E}\left\| \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_d \right) Y_k - \hat{\alpha}(W \otimes I_d) \nabla f^k \right\|^2 \\ & \quad + \hat{\beta}^2 \mathbb{E}\left(\|(\mathcal{A}W \otimes I_d)n_k\|^2 + \|(\mathcal{A}W \otimes I_d)e_k\|^2 \right) \\ & \quad + 2\mathbb{E}\langle ((\mathcal{A}W)^\top \mathcal{A}W) \otimes I_d n_k, e_k \rangle + \hat{\alpha}^2 \mathbb{E}\| (W \otimes I_d)w_k \|^2 \\ &= \mathbb{E}\left\| \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_d \right) Y_k - \hat{\alpha}(W \otimes I_d) \nabla f^k \right\|^2 \\ & \quad + \hat{\beta}^2 \mathbb{E}\left(\|(\mathcal{A}W \otimes I_d)n_k\|^2 + \|(\mathcal{A}W \otimes I_d)e_k\|^2 \right) \\ & \quad + \hat{\alpha}^2 \mathbb{E}\| (W \otimes I_d)w_k \|^2. \end{aligned} \quad (33)$$

Note that $\|Ax\| \leq \|A\|\|x\|$ for any $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$. Then, by $\|W\| = 1$, substituting (26), (28) and (30) into (33) implies

$$\begin{aligned} \mathbb{E}\|Y_{k+1}\|^2 &\leq \mathbb{E}\left\| \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_r \right) Y_k - \hat{\alpha}(W \otimes I_r) \nabla f^k \right\|^2 \\ & \quad + nr\hat{\beta}^2(\Delta^2 + \sigma_k^2) + \frac{n\hat{\alpha}^2\sigma_g^2}{\hat{\gamma}}. \end{aligned} \quad (34)$$

Furthermore, for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^r$, the following Cauchy-Schwarz inequality ([50]) holds: $\|\mathbf{a} + \mathbf{b}\|^2 \leq (1 + \rho_{\mathcal{L}}\hat{\beta})\|\mathbf{a}\|^2 + (1 + \frac{1}{\rho_{\mathcal{L}}\hat{\beta}})\|\mathbf{b}\|^2$. This together with (34) gives

$$\begin{aligned} \mathbb{E}\|Y_{k+1}\|^2 &\leq (1 + \rho_{\mathcal{L}}\hat{\beta}) \mathbb{E}\left\| \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_r \right) Y_k \right\|^2 \\ & \quad + \left(1 + \frac{1}{\rho_{\mathcal{L}}\hat{\beta}} \right) \mathbb{E}\| \hat{\alpha}(W \otimes I_r) \nabla f^k \|^2 \\ & \quad + \frac{n\hat{\alpha}^2\sigma_g^2}{\hat{\gamma}} + nr\hat{\beta}^2(\Delta^2 + \sigma_k^2). \end{aligned} \quad (35)$$

Denote $\rho_{\mathcal{L}} > 0$ as the second smallest eigenvalue of \mathcal{L} . Then, by Courant-Fischer's Theorem ([51]) we have

$$\left\| \left((I_n - \hat{\beta}\mathcal{L}) \otimes I_r \right) Y_k \right\|^2 \leq (1 - \rho_{\mathcal{L}}\hat{\beta})^2 \|Y_k\|^2. \quad (36)$$

Thus, substituting (36) into (35) and noticing $\|W\| = 1$, one can get

$$\begin{aligned} & \mathbb{E}\|Y_{k+1}\|^2 \\ &\leq (1 + \rho_{\mathcal{L}}\hat{\beta})(1 - \rho_{\mathcal{L}}\hat{\beta})^2 \mathbb{E}\|Y_k\|^2 + nr\hat{\beta}^2(\Delta^2 + \sigma_k^2) \\ & \quad + \frac{1 + \rho_{\mathcal{L}}\hat{\beta}}{\rho_{\mathcal{L}}\hat{\beta}} \mathbb{E}\| \hat{\alpha}(W \otimes I_r) \nabla f^k \|^2 + \frac{n\hat{\alpha}^2\sigma_g^2}{\hat{\gamma}} \\ &\leq (1 + \rho_{\mathcal{L}}\hat{\beta})(1 - \rho_{\mathcal{L}}\hat{\beta})^2 \mathbb{E}\|Y_k\|^2 + nr\hat{\beta}^2(\Delta^2 + \sigma_k^2) \\ & \quad + \frac{(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2}{\rho_{\mathcal{L}}\hat{\beta}} \mathbb{E}\| \nabla f^k \|^2 + \frac{n\hat{\alpha}^2\sigma_g^2}{\hat{\gamma}} \\ &= (1 + \rho_{\mathcal{L}}\hat{\beta})(1 - \rho_{\mathcal{L}}\hat{\beta})^2 \mathbb{E}\|Y_k\|^2 + nr\hat{\beta}^2(\Delta^2 + \sigma_k^2) \\ & \quad + \frac{(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2}{\rho_{\mathcal{L}}\hat{\beta}} \mathbb{E}\| \nabla f^k - \nabla f(\bar{x}_k) + \nabla f(\bar{x}_k) \|^2 + \frac{n\hat{\alpha}^2\sigma_g^2}{\hat{\gamma}}. \end{aligned} \quad (37)$$

Note that for any $m \geq 1$ and $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m \in \mathbb{R}^r$, the following inequality holds:

$$\|\mathbf{a}_1 + \mathbf{a}_2 + \dots + \mathbf{a}_m\|^2 \leq m(\|\mathbf{a}_1\|^2 + \|\mathbf{a}_2\|^2 + \dots + \|\mathbf{a}_m\|^2). \quad (38)$$

Then, by letting $m = 2$ in (38), $\|\nabla f^k - \nabla f(\bar{x}_k) + \nabla f(\bar{x}_k)\|^2$ in (37) can be rewritten as

$$\begin{aligned} & \left\| \nabla f^k - \nabla f(\bar{x}_k) + \nabla f(\bar{x}_k) \right\|^2 \\ &\leq 2\left\| \nabla f^k - \nabla f(\bar{x}_k) \right\|^2 + 2\left\| \nabla f(\bar{x}_k) \right\|^2 \\ &= 2\sum_{i=1}^n \left\| \nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k) \right\|^2 + 2\sum_{i=1}^n \left\| \nabla f_i(\bar{x}_k) \right\|^2. \end{aligned} \quad (39)$$

By Assumption 2(i) we have $\|\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)\| \leq L\|x_{i,k} - \bar{x}_k\|$. Then, $\sum_{i=1}^n \|\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)\|^2$ can be rewritten as

$$\sum_{i=1}^n \|\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)\|^2 \leq L^2 \sum_{i=1}^n \|x_{i,k} - \bar{x}_k\|^2 = L^2 \|Y_k\|^2. \quad (40)$$

By Assumption 2(ii) and Lemma A.1(ii), $\|\nabla f_i(\bar{x}_k)\|^2 \leq 2L(f_i(\bar{x}_k) - f_i^*)$, we have

$$\sum_{i=1}^n \|\nabla f_i(\bar{x}_k)\|^2 \leq 2L \sum_{i=1}^n (f_i(\bar{x}_k) - f_i^*). \quad (41)$$

Thus, substituting (40) and (41) into (39) gives

$$\begin{aligned} & \|\nabla f^k - \nabla f(\bar{x}_k) + \nabla f(\bar{x}_k)\|^2 \\ & \leq 2L^2 \|Y_k\|^2 + 4L \left(\sum_{i=1}^n f_i(\bar{x}_k) - f_i^* \right). \end{aligned} \quad (42)$$

Note that by Assumption 2(ii), each cost function $f_i(x)$ has the global minimum f_i^* . Then, the global cost function $F(x)$ has the global minimum $F^* = \min_{x \in \mathbb{R}^r} F(x)$. Let $M^* = F^* - \frac{1}{n} \sum_{i=1}^n f_i^*$. Then, (42) can be rewritten as $\|\nabla f^k - \nabla f(\bar{x}_k) + \nabla f(\bar{x}_k)\|^2 \leq 2L^2 \|Y_k\|^2 + 4L(\sum_{i=1}^n f_i(\bar{x}_k) - f_i^*) = 2L^2 \|Y_k\|^2 + 4nL(F(\bar{x}_k) - F^*) + 4nLM^*$. This together with (37) implies

$$\begin{aligned} \mathbb{E}\|Y_{k+1}\|^2 & \leq \left(1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{2(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}} \right) \mathbb{E}\|Y_k\|^2 \\ & + \frac{4n(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L}{\rho_{\mathcal{L}}\hat{\beta}} \mathbb{E}(F(\bar{x}_k) - F^*) + \frac{n\hat{\alpha}^2 \sigma_g^2}{\hat{\gamma}} \\ & + nr\hat{\beta}^2(\Delta^2 + \sigma_k^2) + \frac{4n(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LM^*}{\rho_{\mathcal{L}}\hat{\beta}}. \end{aligned} \quad (43)$$

Step 2: We next focus on the term $F(\bar{x}_k) - F^*$. Multiplying both sides of (23) by $\frac{1}{n}(\mathbf{1}_n \otimes I_r)$ implies

$$\bar{x}_{k+1} = \bar{x}_k - \hat{\alpha} \overline{\nabla f^k} - \hat{\alpha} \bar{w}_k + \frac{\hat{\beta}}{n} (\mathbf{1}_n^\top \otimes I_r) (e_k + d_k). \quad (44)$$

Then by (44) and Lemma A.1(i), we can derive that

$$\begin{aligned} & F(\bar{x}_{k+1}) - F^* \\ & \leq (F(\bar{x}_k) - F^*) + \frac{L}{2} \|\bar{x}_{k+1} - \bar{x}_k\|^2 + \langle \nabla F(\bar{x}_k), \bar{x}_{k+1} - \bar{x}_k \rangle \\ & = (F(\bar{x}_k) - F^*) + \frac{L}{2} \|\hat{\alpha} \overline{\nabla f^k} - \frac{\hat{\beta}}{n} (\mathbf{1}_n^\top \otimes I_r) (e_k + d_k) \\ & \quad + \hat{\alpha} \bar{w}_k\|^2 - \langle \nabla F(\bar{x}_k), -\frac{\hat{\beta}}{n} (\mathbf{1}_n^\top \otimes I_r) (e_k + d_k) \\ & \quad + \hat{\alpha} \overline{\nabla f^k} + \hat{\alpha} \bar{w}_k \rangle. \end{aligned} \quad (45)$$

By (25), (27) and (29), taking mathematical expectation of (45) gives

$$\begin{aligned} & \mathbb{E}(F(\bar{x}_{k+1}) - F^*) \\ & \leq \mathbb{E}(F(\bar{x}_k) - F^*) - \hat{\alpha} \mathbb{E} \langle \nabla F(\bar{x}_k), \overline{\nabla f^k} \rangle \\ & \quad + \frac{L}{2} \mathbb{E} \|\hat{\alpha} \overline{\nabla f^k} - \frac{\hat{\beta}}{n} (\mathbf{1}_n^\top \otimes I_r) (e_k + d_k) + \hat{\alpha} \bar{w}_k\|^2 \\ & = \mathbb{E}(F(\bar{x}_k) - F^*) - \hat{\alpha} \mathbb{E} \langle \nabla F(\bar{x}_k), \overline{\nabla f^k} \rangle \\ & \quad + \frac{\hat{\beta}^2 L}{2n^2} \mathbb{E} (\|(\mathbf{1}_n^\top \otimes I_r) e_k\|^2 + \|(\mathbf{1}_n^\top \otimes I_r) d_k\|^2) \\ & \quad + \frac{\hat{\alpha}^2 L}{2} \mathbb{E} \|\overline{\nabla f^k}\|^2 + \frac{\hat{\alpha}^2 L}{2} \mathbb{E} \|\bar{w}_k\|^2. \end{aligned} \quad (46)$$

Note that $\|(\mathbf{1}_n^\top \otimes I_r) d_k\|^2 = n \|\sum_{i=1}^n d_{i,k}\|^2 \leq n^2 \|d_k\|^2$, $\|(\mathbf{1}_n^\top \otimes I_r) e_k\|^2 = n \|\sum_{i=1}^n e_{i,k}\|^2 \leq n^2 \|e_k\|^2$, $\|\bar{w}_k\|^2 = \|\frac{1}{n} \sum_{i=1}^n e_{i,k}\|^2 \leq \frac{1}{n} \sum_{i=1}^n \|e_{i,k}\|^2$. Then, by (26), (28) and (30), (46) can be rewritten as

$$\begin{aligned} & \mathbb{E}(F(\bar{x}_{k+1}) - F^*) \\ & \leq \mathbb{E}(F(\bar{x}_k) - F^*) - \hat{\alpha} \mathbb{E} \langle \nabla F(\bar{x}_k), \overline{\nabla f^k} \rangle \\ & \quad + \frac{\hat{\alpha}^2 L}{2} \mathbb{E} \|\overline{\nabla f^k}\|^2 + \frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}}. \end{aligned} \quad (47)$$

Note that $\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{2} \|\mathbf{a}\|^2 + \frac{1}{2} \|\mathbf{b}\|^2 - \frac{1}{2} \|\mathbf{a} - \mathbf{b}\|^2$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^r$. Then, $-\hat{\alpha} \langle \nabla F(\bar{x}_k), \overline{\nabla f^k} \rangle$ in (47) can be rewritten as

$$\begin{aligned} & -\hat{\alpha} \langle \nabla F(\bar{x}_k), \overline{\nabla f^k} \rangle \\ & = -\frac{\hat{\alpha}}{2} \|\nabla F(\bar{x}_k)\|^2 - \frac{\hat{\alpha}}{2} \|\overline{\nabla f^k}\|^2 + \frac{\hat{\alpha}}{2} \|\nabla F(\bar{x}_k) - \overline{\nabla f^k}\|^2 \\ & \leq -\frac{\hat{\alpha}}{2} \|\nabla F(\bar{x}_k)\|^2 + \frac{\hat{\alpha}}{2} \|\nabla F(\bar{x}_k) - \overline{\nabla f^k}\|^2. \end{aligned} \quad (48)$$

Let $m = n$ in (38). Then, $\|\nabla F(\bar{x}_k) - \overline{\nabla f^k}\|^2$ in (48) can be rewritten as

$$\begin{aligned} \|\nabla F(\bar{x}_k) - \overline{\nabla f^k}\|^2 & = \left\| \frac{1}{n} \sum_{i=1}^n (\nabla f_i(\bar{x}_k) - \nabla f_i(x_{i,k})) \right\|^2 \\ & \leq \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\bar{x}_k) - \nabla f_i(x_{i,k})\|^2. \end{aligned} \quad (49)$$

Thus, by (40), (49) can be rewritten as

$$\|\nabla F(\bar{x}_k) - \overline{\nabla f^k}\|^2 \leq \frac{L^2}{n} \|Y_k\|^2. \quad (50)$$

Substituting (48) and (50) into (47) implies

$$\begin{aligned} & \mathbb{E}(F(\bar{x}_{k+1}) - F^*) \\ & \leq \mathbb{E}(F(\bar{x}_k) - F^*) - \frac{\hat{\alpha}}{2} \mathbb{E} \|\nabla F(\bar{x}_k)\|^2 + \frac{\hat{\alpha} L^2}{2n} \mathbb{E} \|Y_k\|^2 \\ & \quad + \frac{\hat{\alpha}^2 L}{2} \mathbb{E} \|\overline{\nabla f^k} - \nabla F(\bar{x}_k) + \nabla F(\bar{x}_k)\|^2 \\ & \quad + \frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}}. \end{aligned} \quad (51)$$

Furthermore, by letting $m = 2$ in (38) and using (50), $\|\overline{\nabla f^k} - \nabla F(\bar{x}_k) + \nabla F(\bar{x}_k)\|^2$ in (51) can be rewritten as

$$\begin{aligned} & \|\overline{\nabla f^k} - \nabla F(\bar{x}_k) + \nabla F(\bar{x}_k)\|^2 \\ & \leq 2 \|\overline{\nabla f^k} - \nabla F(\bar{x}_k)\|^2 + 2 \|\nabla F(\bar{x}_k)\|^2 \\ & \leq \frac{2L^2}{n} \|Y_k\|^2 + 2 \|\nabla F(\bar{x}_k)\|^2. \end{aligned} \quad (52)$$

By letting $m = n$ in (38) and using (41), $\|\nabla F(\bar{x}_k)\|^2$ in (52) can be rewritten as

$$\begin{aligned} \|\nabla F(\bar{x}_k)\|^2 & \leq \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\bar{x}_k)\|^2 \\ & \leq \frac{2L}{n} \sum_{i=1}^n (f_i(\bar{x}_k) - f_i^*) \\ & = 2L(F(\bar{x}_k) - F^*) + 2LM^*. \end{aligned} \quad (53)$$

Thus, substituting (52)-(53) into (51) implies

$$\begin{aligned} & \mathbb{E}(F(\bar{x}_{k+1}) - F^*) \\ & \leq (1 + 2\hat{\alpha}^2 L^2) \mathbb{E}(F(\bar{x}_k) - F^*) \\ & \quad - \frac{\hat{\alpha}}{2} \mathbb{E} \|\nabla F(\bar{x}_k)\|^2 + \frac{\hat{\alpha} L^2 (1 + 2\hat{\alpha} L)}{2n} \mathbb{E} \|Y_k\|^2 \\ & \quad + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}} + \frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + 2\hat{\alpha}^2 L^2 M^* \\ & \leq (1 + 2\hat{\alpha}^2 L^2) \mathbb{E}(F(\bar{x}_k) - F^*) \\ & \quad + \frac{\hat{\alpha} L^2 (1 + 2\hat{\alpha} L)}{2n} \mathbb{E} \|Y_k\|^2 + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}} \end{aligned}$$

$$+ \frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + 2\hat{\alpha}^2 L^2 M^*. \quad (54)$$

Let

$$\theta_1 = \max\left\{1 + 2\hat{\alpha}^2 L^2 + \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L}{\rho_{\mathcal{L}}\hat{\beta}}, \right. \\ \left. 1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{\hat{\alpha} L^2(1 + 2\hat{\alpha}L)}{2n} + \frac{2(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}}\right\}, \quad (55)$$

$$\theta_{k,2} = \frac{(L+2)nr\hat{\beta}^2}{2} (\Delta^2 + \sigma_k^2) + \frac{\hat{\alpha}^2 \sigma_g^2 (2n+L)}{2\hat{\gamma}} \\ + 2\hat{\alpha}^2 L^2 M^* + \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LM^*}{\rho_{\mathcal{L}}\hat{\beta}}. \quad (56)$$

Then, summing (43) and (54) implies

$$\mathbb{E}(\|Y_{k+1}\|^2 + F(\bar{x}_{k+1}) - F^*) \\ \leq \theta_1 \mathbb{E}(\|Y_k\|^2 + F(\bar{x}_k) - F^*) + \theta_{k,2}. \quad (57)$$

By iteratively computing (57), the following inequality holds for any $0 \leq k \leq K$:

$$\mathbb{E}(\|Y_{k+1}\|^2 + F(\bar{x}_{k+1}) - F^*) \\ \leq \theta_1^{k+1} \mathbb{E}(\|Y_0\|^2 + F(\bar{x}_0) - F^*) + \sum_{m=0}^k \theta_1^{k-m} \theta_{m,2}. \quad (58)$$

Step 3: At this step, we prove that there exists $G_1 \geq 0$ such that $\mathbb{E}(F(\bar{x}_k) - F^*) \leq G_1$ holds for any $K \geq 1$ and $0 \leq k \leq K+1$. Note that $2\hat{\alpha}L^2 = O(\frac{1}{K^{2\alpha}})$ and $\frac{4n(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L}{\rho_{\mathcal{L}}\hat{\beta}} = O(\frac{1}{K^{2\alpha-\beta}})$ holds for any $K \geq 1$. Then, by $2\alpha - \beta > 1$ in Assumption 4, it can be seen that for any $K \geq 1$,

$$\left(1 + 2\hat{\alpha}^2 L^2 + \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L}{\rho_{\mathcal{L}}\hat{\beta}}\right)^{K+1} \\ = \left(1 + O\left(\frac{1}{K^{2\alpha-\beta}}\right)\right)^{K+1} \\ = \exp\left((K+1) \ln\left(1 + O\left(\frac{1}{K^{2\alpha-\beta}}\right)\right)\right) \\ = \exp\left(O\left(\frac{1}{K^{2\alpha-\beta-1}}\right)\right) < \infty. \quad (59)$$

Note that by $\beta < \alpha$ in Assumption 4, there exists $K_0 > 0$ such that for any $K \geq K_0$, $1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{\hat{\alpha}L^2(1+2\hat{\alpha}L)}{2n} + \frac{2(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}} \leq 1 - \frac{\rho_{\mathcal{L}}\hat{\beta}}{2}$ holds. Then, it can be seen that for any $K \geq K_0$,

$$\left(1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{\hat{\alpha}L^2(1+2\hat{\alpha}L)}{2n} + \frac{2(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}}\right)^{K+1} \\ \leq \left(1 - \frac{\rho_{\mathcal{L}}\hat{\beta}}{2}\right)^{K+1} \\ = \exp\left((K+1) \ln\left(1 - \frac{\rho_{\mathcal{L}}\hat{\beta}}{2}\right)\right) \\ \leq \exp\left(-\frac{\rho_{\mathcal{L}}\hat{\beta}}{2} \left(1 + \frac{1}{K}\right) K^{1-\beta}\right) \\ \leq \exp\left(-\frac{\rho_{\mathcal{L}}\hat{\beta}}{2} K_0^{1-\beta}\right) < \infty. \quad (60)$$

Thus, for any $K \geq 1$, we have

$$\left(1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{\hat{\alpha}L^2(1+2\hat{\alpha}L)}{2n} + \frac{2(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}}\right)^{K+1} \\ \leq \max\left\{\exp\left(-\frac{\rho_{\mathcal{L}}\hat{\beta}}{2} K_0^{1-\beta}\right), \right. \\ \left. 1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{a_1 L^2(2a_1 L + 1)}{2n} + \frac{2(1+\rho_{\mathcal{L}}a_2)a_1^2 L^2}{\rho_{\mathcal{L}}a_2}, \dots, \right. \\ \left. 1 - \frac{\rho_{\mathcal{L}}a_2}{K_0^\beta} + \frac{a_1 L^2(2a_1 L + K_0^\alpha)}{2n K_0^{2\alpha}} + \frac{2(K_0^\beta + \rho_{\mathcal{L}}a_2)a_1^2 L^2}{\rho_{\mathcal{L}}a_2 K_0^{2\alpha}}\right\} < \infty. \quad (61)$$

Hence, for any $K \geq 1$, (59) together with (61) implies

$$1 < \theta_1^{K+1} < \infty. \quad (62)$$

When $\sigma \leq 0$, σ_k is decreasing, and then $\sigma_k \leq \sigma_0$ for any $0 \leq k \leq K$. When $\sigma > 0$, σ_k is increasing, and then $\sigma_k \leq \sigma_K$ for any $0 \leq k \leq K$. As a result, $\sigma_k \leq \max\{\sigma_0, \sigma_K\}$ for any $0 \leq k \leq K$. Hence, by the definition of $\theta_{k,2}$ in (56), $\theta_{k,2} \leq \max\{\theta_{0,2}, \theta_{K,2}\}$ for any $0 \leq k \leq K$. This helps us to obtain that

$$\sum_{m=0}^K \theta_1^{K-m} \theta_{m,2} \leq \sum_{m=0}^K \theta_1^{K+1} K \theta_{m,2} \\ \leq K \max\{\theta_{0,2}, \theta_{K,2}\} \theta_1^{K+1}. \quad (63)$$

Note that

$$\max\{\theta_{0,2}, \theta_{K,2}\} = \frac{(L+2)nr\hat{\beta}^2}{2} (\Delta^2 + \max\{\sigma_0^2, \sigma_K^2\}) \\ + \frac{\hat{\alpha}^2 \sigma_g^2 (2n+L)}{2\hat{\gamma}} + 2\hat{\alpha}^2 L^2 M^* \\ + \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LM^*}{\rho_{\mathcal{L}}\hat{\beta}} \\ = O\left(\frac{1}{K^{2\beta-2\max\{\sigma,0\}}} + \frac{1}{K^{2\alpha-\beta}}\right). \quad (64)$$

Then, by $2\alpha - \beta > 1$ and $\frac{1}{2} + \max\{\sigma, 0\} < \beta$ in Assumption 4, substituting (64) into (63) implies

$$\sum_{m=0}^K \theta_1^{K-m} \theta_{m,2} = O\left(\frac{1}{K^{2\beta-2\max\{\sigma,0\}-1}} + \frac{1}{K^{2\alpha-\beta-1}}\right) < \infty. \quad (65)$$

Thus, for any $K \geq 1$ and $0 \leq k \leq K$, by (58), (62) and (65) we have

$$\mathbb{E}(\|Y_{k+1}\|^2 + F(\bar{x}_{k+1}) - F^*) \\ \leq \theta_1^{k+1} \mathbb{E}(\|Y_0\|^2 + F(\bar{x}_0) - F^*) + \sum_{m=0}^k \theta_1^{k-m} \theta_{m,2} \\ \leq \theta_1^{K+1} \mathbb{E}(\|Y_0\|^2 + F(\bar{x}_0) - F^*) + \sum_{m=0}^K \theta_1^{K-m} \theta_{m,2} < \infty.$$

Hence, there exists $G_1 \geq 0$ such that $\mathbb{E}(F(\bar{x}_k) - F^*) \leq G_1$ holds for any $K \geq 1$ and $0 \leq k \leq K+1$.

Step 4: At this step, we prove $\lim_{K \rightarrow \infty} \mathbb{E}\|Y_{K+1}\|^2 = 0$ for any $K \geq 1$. By Step 3, since there exists $G_1 \geq 0$ such that $\mathbb{E}(F(\bar{x}_k) - F^*) \leq G_1$ holds for any $K \geq 1$ and $0 \leq k \leq K+1$,

by (43) we have

$$\begin{aligned} \mathbb{E}\|Y_{k+1}\|^2 &\leq \left(1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{2(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}}\right) \mathbb{E}\|Y_k\|^2 \\ &\quad + \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LG_1}{\rho_{\mathcal{L}}\hat{\beta}} + \frac{n\hat{\alpha}^2\sigma_g^2}{\hat{\gamma}} \\ &\quad + nr\hat{\beta}^2(\Delta^2 + \sigma_k^2) + \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LM^*}{\rho_{\mathcal{L}}\hat{\beta}}. \end{aligned} \quad (66)$$

Let

$$\begin{aligned} \theta_3 &= 1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{2(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}}, \quad (67) \\ \theta_{k,4} &= \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LG_1}{\rho_{\mathcal{L}}\hat{\beta}} + \frac{n\hat{\alpha}^2\sigma_g^2}{\hat{\gamma}} \\ &\quad + nr\hat{\beta}^2(\Delta^2 + \sigma_k^2) + \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LM^*}{\rho_{\mathcal{L}}\hat{\beta}}. \end{aligned} \quad (68)$$

Then, substituting (67) and (68) into (66) and iteratively computing (66) gives

$$\mathbb{E}\|Y_k\|^2 \leq \theta_3^{k+1} \mathbb{E}\|Y_0\|^2 + \sum_{m=0}^k \theta_3^{k-m} \theta_{m,4}. \quad (69)$$

Note that by the definition of θ_3 in (67) and $2\alpha - \beta > 1$, $0 < \beta < \alpha < 1$ in Assumption 4, we have

$$\frac{1}{1 - \theta_3} = \frac{1}{\rho_{\mathcal{L}}\hat{\beta} - \frac{2(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}}} = O(K^\beta), \quad (70)$$

and

$$\begin{aligned} \max\{\theta_{0,4}, \theta_{K,4}\} &= \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LG_1}{\rho_{\mathcal{L}}\hat{\beta}} + \frac{n\hat{\alpha}^2\sigma_g^2}{\hat{\gamma}} \\ &\quad + nr\hat{\beta}^2(\Delta^2 + \max\{\sigma_K^2, \sigma_0^2\}) + \frac{4n(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 LM^*}{\rho_{\mathcal{L}}\hat{\beta}} \\ &= O\left(\frac{1}{K^{2\alpha - \beta}} + \frac{1}{K^{2\beta - 2\max\{\sigma, 0\}}}\right). \end{aligned} \quad (71)$$

Moreover, by the definition of $\theta_{k,4}$ in (68), $\theta_{k,4} \leq \max\{\theta_{0,4}, \theta_{K,4}\}$ for any $0 \leq k \leq K$. Then, it follows from (70) and (71) that

$$\begin{aligned} \sum_{m=0}^K \theta_3^{K-m} \theta_{m,4} &\leq \max\{\theta_{0,4}, \theta_{K,4}\} \sum_{m=0}^K \theta_3^{K-m} \\ &= \max\{\theta_{0,4}, \theta_{K,4}\} \frac{1 - \theta_3^{K+1}}{1 - \theta_3} = O\left(\frac{\max\{\theta_{0,4}, \theta_{K,4}\}}{1 - \theta_3}\right) \\ &= O\left(\frac{1}{K^{2\alpha - 2\beta}} + \frac{1}{K^{\beta - 2\max\{\sigma, 0\}}}\right). \end{aligned} \quad (72)$$

Meanwhile, by (60) we have

$$\begin{aligned} \theta_3^{K+1} &\leq \left(1 - \rho_{\mathcal{L}}\hat{\beta} + \frac{\hat{\alpha}L^2(1 + 2\hat{\alpha}L)}{2n} + \frac{2(1 + \rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2 L^2}{\rho_{\mathcal{L}}\hat{\beta}}\right)^{K+1} \\ &= O\left(\left(1 - \frac{\rho_{\mathcal{L}}\hat{\beta}}{2}\right)^{K+1}\right) \\ &= O\left(\exp\left((K+1) \ln\left(1 - \frac{\rho_{\mathcal{L}}\hat{\beta}}{2}\right)\right)\right) \\ &= O\left(\exp\left(-\frac{\rho_{\mathcal{L}}\hat{\alpha}^2}{2} K^{1-\beta}\right)\right). \end{aligned} \quad (73)$$

Let $k = K$ in (69). Then, substituting (73) and (72) into (69) implies $\mathbb{E}\|Y_{K+1}\|^2 \leq O\left(\exp\left(-\frac{\rho_{\mathcal{L}}\hat{\alpha}^2}{2} K^{1-\beta}\right)\right) + O\left(\frac{1}{K^{2\alpha - 2\beta}} + \frac{1}{K^{\beta - 2\max\{\sigma, 0\}}}\right) = O\left(\frac{1}{K^{2\alpha - 2\beta}} + \frac{1}{K^{\beta - 2\max\{\sigma, 0\}}}\right)$. Hence, we have $\lim_{K \rightarrow \infty} \mathbb{E}\|Y_{K+1}\|^2 = 0$.

Step 5: At this step, we give the estimation of $\sum_{k=0}^K \mathbb{E}\|Y_k\|^2$ for any $K \geq 1$. Note that summing (69) from $k = 0$ to K gives $\sum_{k=0}^K \mathbb{E}\|Y_k\|^2 \leq \sum_{k=0}^K \theta_3^{k+1} \mathbb{E}\|Y_0\|^2 + \sum_{k=0}^K \sum_{m=0}^k \theta_3^{k-m} \theta_{m,4}$. Then, it follows from (70) that

$$\sum_{k=0}^K \theta_3^{k+1} = \frac{\theta_3(1 - \theta_3^{K+1})}{1 - \theta_3} = O(K^\beta). \quad (74)$$

Moreover, by (70) and (71), we have

$$\begin{aligned} \sum_{k=0}^K \sum_{m=0}^k \theta_3^{k-m} \theta_{m,4} &\leq \max\{\theta_{0,4}, \theta_{K,4}\} \sum_{k=0}^K \sum_{m=0}^k \theta_3^{k-m} \\ &= \max\{\theta_{0,4}, \theta_{K,4}\} \sum_{k=0}^K \frac{1 - \theta_3^{k+1}}{1 - \theta_3} = O\left(\frac{K \max\{\theta_{0,4}, \theta_{K,4}\}}{1 - \theta_3}\right) \\ &= O\left(\frac{1}{K^{2\alpha - 2\beta - 1}} + \frac{1}{K^{\beta - 2\max\{\sigma, 0\} - 1}}\right). \end{aligned} \quad (75)$$

Hence, substituting (74) and (75) into (69) implies

$$\sum_{k=0}^K \mathbb{E}\|Y_k\|^2 = O\left(K^\beta + \frac{1}{K^{2\alpha - 2\beta - 1}} + \frac{1}{K^{\beta - 2\max\{\sigma, 0\} - 1}}\right). \quad (76)$$

Step 6: Finally, we prove $\liminf_{K \rightarrow \infty} \mathbb{E}\|\nabla F(x_{i,K+1})\|^2 = 0$ for any node $i \in \mathcal{V}$. From Step 3, since there exists $G_1 \geq 0$ such that $\mathbb{E}(F(\bar{x}_k) - F^*) \leq G_1$ holds for any $K \geq 1$ and $0 \leq k \leq K + 1$, by Lemma A.1(ii) we have

$$\mathbb{E}\|\nabla F(\bar{x}_k)\|^2 \leq 2L\mathbb{E}(F(\bar{x}_k) - F^*) \leq 2LG_1. \quad (77)$$

Then, substituting (52) and (77) into (51) implies

$$\begin{aligned} &\mathbb{E}(F(\bar{x}_{k+1}) - F^*) \\ &\leq \mathbb{E}(F(\bar{x}_k) - F^*) - \frac{\hat{\alpha}}{2} \mathbb{E}\|\nabla F(\bar{x}_k)\|^2 \\ &\quad + \frac{\hat{\alpha}L^2(1 + 2\hat{\alpha}L)}{2n} \mathbb{E}\|Y_k\|^2 + \frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) \\ &\quad + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}} + 2\hat{\alpha}^2 L^2 G_1. \end{aligned} \quad (78)$$

Note that (78) can be rewritten as

$$\begin{aligned} &\frac{\hat{\alpha}}{2} \mathbb{E}\|\nabla F(\bar{x}_k)\|^2 \\ &\leq \mathbb{E}(F(\bar{x}_k) - F(\bar{x}_{k+1})) + \frac{\hat{\alpha}L^2(1 + 2\hat{\alpha}L)}{2n} \mathbb{E}\|Y_k\|^2 \\ &\quad + \frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}} + 2\hat{\alpha}^2 L^2 G_1. \end{aligned} \quad (79)$$

Then, since $F(x^*) \leq F(x)$ holds for any $x \in \mathbb{R}^r$, iteratively computing (79) implies

$$\frac{\hat{\alpha}}{2} \sum_{k=0}^K \mathbb{E}\|\nabla F(\bar{x}_k)\|^2 \leq \mathbb{E}(F(\bar{x}_0) - F(\bar{x}_{K+1}))$$

$$\begin{aligned}
& + \frac{\hat{\alpha}L^2(1+2\hat{\alpha}L)}{2n} \sum_{k=0}^K \mathbb{E}\|Y_k\|^2 \\
& + \sum_{k=0}^K \left(\frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}} + 2\hat{\alpha}^2 L^2 G_1 \right) \\
\leq & \mathbb{E}(F(\bar{x}_0) - F(x^*)) + \frac{\hat{\alpha}L^2(1+2\hat{\alpha}L)}{2n} \sum_{k=0}^K \mathbb{E}\|Y_k\|^2 \\
& + \sum_{k=0}^K \left(\frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}} + 2\hat{\alpha}^2 L^2 G_1 \right). \quad (80)
\end{aligned}$$

By $\frac{1}{2} + \max\{\sigma, 0\} < \beta$ and $2\alpha - \beta > 1$ in Assumption 4, we have

$$\begin{aligned}
& \sum_{k=0}^K \left(\frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}} + 2\hat{\alpha}^2 L^2 G_1 \right) \\
= & O \left(\sum_{k=0}^K \left(\frac{1}{K^{2\beta-2\max\{\sigma, 0\}}} + \frac{1}{K^{2\alpha}} \right) \right) \\
= & O \left(\frac{1}{K^{2\beta-2\max\{\sigma, 0\}-1}} + \frac{1}{K^{2\alpha-1}} \right) < \infty. \quad (81)
\end{aligned}$$

Note that $\alpha > \beta$, $2\alpha - \beta > 1$ and $\frac{1}{2} + \max\{\sigma, 0\} < \beta$ in Assumption 4. Then, we have $3\alpha - 2\beta - 1 = (2\alpha - \beta - 1) + (\alpha - \beta) > 0$, $\alpha + \beta - 2\max\{\sigma, 0\} - 1 > 2\beta - 2\max\{\sigma, 0\} - 1 > 0$. Thus, substituting (76) and (81) into (80) implies

$$\begin{aligned}
& \hat{\alpha} \sum_{k=0}^K \mathbb{E}\|\nabla F(\bar{x}_k)\|^2 \\
\leq & + O \left(\frac{1}{K^{\alpha-\beta}} + \frac{1}{K^{3\alpha-2\beta-1}} + \frac{1}{K^{\alpha+\beta-2\max\{\sigma, 0\}-1}} \right) \\
& + O \left(\frac{1}{K^{2\beta-2\max\{\sigma, 0\}-1}} + \frac{1}{K^{2\alpha-1}} \right) \\
& + 2\mathbb{E}(F(\bar{x}_0) - F(x^*)) < \infty. \quad (82)
\end{aligned}$$

Next, we prove $\liminf_{K \rightarrow \infty} \mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2 = 0$ by contradiction. Suppose there exists $G_2 > 0$ such that $\liminf_{K \rightarrow \infty} \mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2 = G_2 > 0$. Then, there exists $K_1 > 0$ such that $\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2 \geq G_2$ holds for any $K \geq K_1$. Thus, for any $K \geq K_1$ we have

$$\begin{aligned}
\hat{\alpha} \sum_{k=0}^K \mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2 & \geq \hat{\alpha} \sum_{k=K_1}^K \mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2 \\
& \geq \hat{\alpha}(K - K_1)G_2 = O(K^{1-\alpha}). \quad (83)
\end{aligned}$$

Note that when K goes to infinity, $\hat{\alpha} \sum_{k=0}^K \mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2$ goes to infinity since the right hand side of (83) goes to infinity, which contradicts with (82). Then, we have $\liminf_{K \rightarrow \infty} \mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2 = 0$. Moreover, for any node $i \in \mathcal{V}$, we have

$$\begin{aligned}
& \mathbb{E}\|\nabla F(x_{i,K+1})\|^2 \\
= & \mathbb{E}\|\nabla F(x_{i,K+1}) - \nabla F(\bar{x}_{K+1}) + \nabla F(\bar{x}_{K+1})\|^2 \\
\leq & 2\mathbb{E}\|\nabla F(x_{i,K+1}) - \nabla F(\bar{x}_{K+1})\|^2 + 2\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2 \\
\leq & 2L^2 \mathbb{E}\|x_{i,K+1} - \bar{x}_{K+1}\|^2 + 2\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2 \\
\leq & 2L^2 \mathbb{E}\|Y_{K+1}\|^2 + 2\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2. \quad (84)
\end{aligned}$$

Therefore, by $\lim_{K \rightarrow \infty} \mathbb{E}\|Y_{K+1}\|^2 = 0$ in Step 3, $\liminf_{K \rightarrow \infty} \mathbb{E}\|\nabla F(x_{i,K+1})\|^2 = 0$ holds for any node $i \in \mathcal{V}$. ■

APPENDIX C PROOF OF THEOREM 3

If Assumption 5 holds, then (54) can be rewritten as

$$\begin{aligned}
\mathbb{E}(F(\bar{x}_{k+1}) - F^*) & \leq (1 - \mu\hat{\alpha} + 2\hat{\alpha}^2 L^2) \mathbb{E}(F(\bar{x}_k) - F^*) \\
& + \frac{\hat{\alpha}L^2(1+2\hat{\alpha}L)}{2n} \mathbb{E}\|Y_k\|^2 + \frac{\hat{\alpha}^2 \sigma_g^2 L}{2\hat{\gamma}} \\
& + \frac{\hat{\beta}^2 nrL}{2} (\Delta^2 + \sigma_k^2) + 2\hat{\alpha}^2 L^2 M^*. \quad (85)
\end{aligned}$$

For any $i \in \mathcal{V}$, by Lemma A.1(i), we have

$$\begin{aligned}
F(x_{i,K+1}) - F(\bar{x}_{K+1}) & \leq \langle \nabla F(\bar{x}_{K+1}), x_{i,K+1} - \bar{x}_{K+1} \rangle \\
& + \frac{L}{2} \|\bar{x}_{K+1} - x_{i,K+1}\|^2. \quad (86)
\end{aligned}$$

Note that $\langle \mathbf{a}, \mathbf{b} \rangle \leq \|\mathbf{a}\| \|\mathbf{b}\| \leq \frac{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2}{2}$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^r$. Then, (86) can be rewritten as

$$\begin{aligned}
& F(x_{i,K+1}) - F(\bar{x}_{K+1}) \\
\leq & \frac{\|\nabla F(\bar{x}_{K+1})\|^2 + \|\bar{x}_{K+1} - x_{i,K+1}\|^2}{2} + \frac{L}{2} \|\bar{x}_{K+1} - x_{i,K+1}\|^2 \\
= & \frac{L+1}{2} \|\bar{x}_{K+1} - x_{i,K+1}\|^2 + \frac{\|\nabla F(\bar{x}_{K+1})\|^2}{2}. \quad (87)
\end{aligned}$$

By Lemma A.1(ii) we have $\|\nabla F(\bar{x}_{K+1})\|^2 \leq 2L(F(\bar{x}_{K+1}) - F^*)$. This together with (87) gives $F(x_{i,K+1}) - F(\bar{x}_{K+1}) \leq \frac{L+1}{2} \|\bar{x}_{K+1} - x_{i,K+1}\|^2 + L(F(\bar{x}_{K+1}) - F^*)$. Thus, we have

$$\begin{aligned}
& F(x_{i,K+1}) - F(\bar{x}_{K+1}) \\
\leq & \frac{L+1}{2} \sum_{i=1}^n \|\bar{x}_{K+1} - x_{i,K+1}\|^2 + L(F(\bar{x}_{K+1}) - F^*) \\
= & \frac{L+1}{2} \|Y_{K+1}\|^2 + L(F(\bar{x}_{K+1}) - F^*). \quad (88)
\end{aligned}$$

Furthermore, for any $i \in \mathcal{V}$, by (88), we have

$$\begin{aligned}
& F(x_{i,K+1}) - F^* \\
= & (F(x_{i,K+1}) - F(\bar{x}_{K+1})) + (F(\bar{x}_{K+1}) - F^*) \\
\leq & \frac{L+1}{2} \|Y_{K+1}\|^2 + (L+1)(F(\bar{x}_{K+1}) - F^*) \\
\leq & (L+1) (\|Y_{K+1}\|^2 + (F(\bar{x}_{K+1}) - F^*)). \quad (89)
\end{aligned}$$

Let

$$\begin{aligned}
\theta_5 & = \max\{1 - \mu\hat{\alpha} + 2\hat{\alpha}^2 L^2, \\
& 1 - \rho_L \hat{\beta} + \frac{\hat{\alpha}L^2(1+2\hat{\alpha}L)}{2n} + \frac{2(1+\rho_L \hat{\beta})\hat{\alpha}^2 L^2}{\rho_L \hat{\beta}}\}. \quad (90)
\end{aligned}$$

Then, substituting (56) and (90) into (85) implies

$$\mathbb{E}(F(\bar{x}_{k+1}) - F^*) \leq \theta_5 \mathbb{E}(F(\bar{x}_k) - F^*) + \theta_{k,2}. \quad (91)$$

Thus, iteratively computing (91) gives

$$\begin{aligned}
\mathbb{E}(F(x_{i,K+1}) - F^*) & \leq \theta_5^{K+1} (L+1) \mathbb{E}(\|Y_0\|^2 + F(\bar{x}_0) - F^*) \\
& + (L+1) \sum_{m=0}^K \theta_5^{K-m} \theta_{m,2}. \quad (92)
\end{aligned}$$

By Assumption 4, we have $0 < \theta_5 < 1$. Since $\ln(1-x) \leq -x$ for any $x < 1$, we can obtain that $\theta_5^{K+1} = \exp((K+1)\ln(1-(1-\theta_5))) \leq \exp(-(K+1)(1-\theta_5))$. Substituting (55) into the inequality above implies

$$\begin{aligned} & \theta_5^{K+1} \\ & \leq \max\left\{\exp(-(K+1)\mu\hat{\alpha}+(K+1)(2\hat{\alpha}^2L^2+\frac{4n(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2L}{\rho_{\mathcal{L}}\hat{\beta}})), \right. \\ & \left. \exp(-(K+1)\hat{\beta}+(K+1)\frac{2(1+\rho_{\mathcal{L}}\hat{\beta})\hat{\alpha}^2L^2}{\rho_{\mathcal{L}}\hat{\beta}})\right\}. \end{aligned} \quad (93)$$

Note that $2\alpha - \beta > 1$ and $\alpha > \beta$ in Assumption 4. Then, (93) can be rewritten as

$$\begin{aligned} \theta_5^{K+1} & = O\left(\max\{\exp(-(K+1)\mu\hat{\alpha}), \exp(-(K+1)\rho_{\mathcal{L}}\hat{\beta})\}\right) \\ & = O\left(\max\{\exp(-\mu a_1 K^{1-\alpha}), \exp(-\rho_{\mathcal{L}} a_2 K^{1-\beta})\}\right). \end{aligned} \quad (94)$$

Moreover, by (65) we have

$$\begin{aligned} \sum_{m=0}^K \theta_5^{K-m} \theta_{m,2} & = O\left(\frac{1}{K^{2\beta-2\max\{\sigma,0\}-1}} + \frac{1}{K^{2\alpha-\beta-1}}\right) \\ & = O\left(\frac{1}{K^{\min\{2\beta-2\max\{\sigma,0\}-1, 2\alpha-\beta-1\}}}\right). \end{aligned} \quad (95)$$

Hence, by substituting (94) and (95) into (92), we have

$$\mathbb{E}(F(x_{i,K+1}) - F^*) = O\left(\frac{1}{K^{\min\{2\beta-2\max\{\sigma,0\}-1, 2\alpha-\beta-1\}}}\right). \quad (96)$$

Note that by Lemma A.1(ii), we have

$$\|\nabla F(x_{i,K+1})\|^2 \leq 2L(F(x_{i,K+1}) - F^*). \quad (97)$$

Then, taking the mathematical expectation on (97) and substituting (96) into (97) imply

$$\begin{aligned} \mathbb{E}\|\nabla F(x_{i,K+1})\|^2 & \leq 2L\mathbb{E}(F(x_{i,K+1}) - F^*) \\ & = O\left(\frac{1}{K^{\min\{2\beta-2\max\{\sigma,0\}-1, 2\alpha-\beta-1\}}}\right). \end{aligned} \quad (98)$$

Note that for any $1 \leq \psi \leq 2$, the function $x^{\frac{\psi}{2}}$ is concave in x . Then, by Jensen's inequality ([44]) we have $\mathbb{E}\|\nabla F(x_{i,K+1})\|^\psi = \mathbb{E}\left(\|\nabla F(x_{i,K+1})\|^2\right)^{\frac{\psi}{2}} \leq \left(\mathbb{E}\|\nabla F(x_{i,K+1})\|^2\right)^{\frac{\psi}{2}}$. Thus, substituting (98) into it implies $\mathbb{E}\|\nabla F(x_{i,K+1})\|^\psi = O\left(\frac{1}{K^{\frac{\psi}{2}\min\{2\beta-2\max\{\sigma,0\}-1, 2\alpha-\beta-1\}}}\right)$. ■

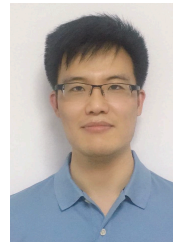
REFERENCES

- [1] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Trans. Autom. Control*, vol. 57, no. 1, pp. 151–164, 2011.
- [2] M. Zhu and S. Martínez, "An approximate dual subgradient algorithm for multi-agent non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 6, pp. 1534–1539, 2013.
- [3] T. T. Doan, S. T. Maguluri, and J. Romberg, "Fast convergence rates of distributed subgradient methods with adaptive quantization," *IEEE Trans. Autom. Control*, vol. 66, no. 5, pp. 2191–2205, 2021.
- [4] T. T. Doan, S. T. Maguluri, and J. Romberg, "Convergence rates of distributed gradient methods under random quantization: a stochastic approximation approach," *IEEE Trans. Autom. Control*, vol. 66, no. 10, pp. 4469–4484, 2021.
- [5] R. Xin, U. A. Khan, and S. Kar, "A fast randomized incremental gradient method for decentralized nonconvex optimization," *IEEE Trans. Autom. Control*, vol. 67, no. 10, pp. 5150–5165, 2022.
- [6] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, "Collaborative deep learning in fixed topology networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, vol. 30, 2017, pp. 5904–5914.
- [7] K. Lu, H. Wang, H. Zhang, and L. Wang, "Convergence in high probability of distributed stochastic gradient descent algorithms," *IEEE Trans. Autom. Control*, vol. 69, no. 4, pp. 2189–2204, 2024.
- [8] A. Reiszadeh, H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Robust and communication-efficient collaborative learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, vol. 32, 2019, pp. 8388–8399.
- [9] Z. Zhang, Y. Zhang, D. Guo, S. Zhao, and X. Zhu, "Communication-efficient federated continual learning for distributed learning system with non-iid data," *Sci. China Inf. Sci.*, vol. 66, no. 2, 2023, Art. no. 122102.
- [10] K. Ge, Y. Zhang, Y. Fu, Z. Lai, X. Deng, and D. Li, "Accelerate distributed deep learning with cluster-aware sketch quantization," *Sci. China Inf. Sci.*, vol. 66, no. 6, 2023, Art. no. 162102.
- [11] J. Lei, P. Yi, J. Chen, and Y. Hong, "Distributed variable sample-size stochastic optimization with fixed step-sizes," *IEEE Trans. Autom. Control*, vol. 67, no. 10, pp. 5630–5637, 2022.
- [12] J. F. Zhang, J. W. Tan, and J. Wang, "Privacy security in control systems," *Sci. China Inf. Sci.*, vol. 64, no. 7, 2021, Art. no. 176201.
- [13] Y. Lu and M. Zhu, "Privacy preserving distributed optimization using homomorphic encryption," *Automatica*, vol. 96, pp. 314–325, 2018.
- [14] Y. L. Mo and R. M. Murray, "Privacy preserving average consensus," *IEEE Trans. Autom. Control*, vol. 62, no. 2, pp. 753–765, 2017.
- [15] Y. Lou, L. Yu, S. Wang, and P. Yi, "Privacy preservation in distributed subgradient optimization algorithms," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2154–2165, 2018.
- [16] Y. Wang, "Privacy-preserving average consensus via state decomposition," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4711–4716, 2019.
- [17] Y. Lu and M. Zhu, "On privacy preserving data release of linear dynamic networks," *Automatica*, vol. 115, 2020, Art. no. 108839.
- [18] J. Le Ny and G. J. Pappas, "Differentially private filtering," *IEEE Trans. Autom. Control*, vol. 59, no. 2, pp. 341–354, 2014.
- [19] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.
- [20] X. K. Liu, J. F. Zhang, and J. Wang, "Differentially private consensus algorithm for continuous-time heterogeneous multi-agent systems," *Automatica*, vol. 122, 2020, Art. no. 109283.
- [21] J. Wang, J. F. Zhang, and X. He, "Differentially private distributed algorithms for stochastic aggregative games," *Automatica*, vol. 142, 2022, Art. no. 110440.
- [22] X. Chen, C. Wang, Q. Yang, T. Hu, and C. Jiang, "Locally differentially private high-dimensional data synthesis," *Sci. China Inf. Sci.*, vol. 66, no. 1, 2023, Art. no. 112101.
- [23] J. Wang, J. Ke, and J. F. Zhang, "Differentially private bipartite consensus over signed networks with time-varying noises," *IEEE Trans. Autom. Control*, vol. 69, no. 9, pp. 5788–5803, 2024.
- [24] X. Zhang, M. M. Khalili, and M. Liu, "Improving the privacy and accuracy of ADMM-based distributed algorithms," in *Int. Conf. Mach. Learn.*, Stockholm, Sweden, 2018, pp. 5796–5805.
- [25] C. Li, P. Zhou, L. Xiong, Q. Wang, and T. Wang, "Differentially private distributed online learning," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 8, pp. 1440–1453, 2018.
- [26] Z. Huang, R. Hu, Y. Guo, E. Chan-Tin, and Y. Gong, "DP-ADMM: ADMM-based distributed learning with differential privacy," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1002–1012, 2020.
- [27] J. Ding, G. Liang, J. Bi, and M. Pan, "Differentially private and communication efficient collaborative learning," in *Proc. AAAI Conf. Artif. Intell.*, Palo Alto, CA, USA, vol. 35, no. 8, 2021, pp. 7219–7227.
- [28] C. Gratton, N. K. D. Venkatesowda, R. Arablouci, and S. Werner, "Privacy-preserved distributed learning with zeroth-order optimization," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 265–279, 2022.
- [29] C. Liu, K. H. Johansson, and Y. Shi, "Distributed empirical risk minimization with differential privacy," *Automatica*, vol. 162, 2024, Art. no. 111514.
- [30] J. Xu, W. Zhang, and F. Wang, "A (DP)² SGD: asynchronous decentralized parallel stochastic gradient descent with differential privacy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8036–8047, 2022.
- [31] Y. Wang and T. Başar, "Decentralized nonconvex optimization with guaranteed privacy and accuracy," *Automatica*, vol. 150, 2023, Art. no. 110858.

- [32] Y. Wang and T. Başar, "Quantization enabled privacy protection in decentralized stochastic optimization," *IEEE Trans. Autom. Control*, vol. 68, no. 7, pp. 4038–4052, 2023.
- [33] G. Yan, T. Li, K. Wu, and L. Song, "Killing two birds with one stone: quantization achieves privacy in distributed learning," *Digit. Signal Process.*, vol. 146, 2024, Art. no. 104353.
- [34] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 446–472, 1948.
- [35] R. Carli, F. Bullo, and S. Zampieri, "Quantized average consensus via dynamic coding/decoding schemes," *Int. J. Robust Nonlinear Control*, vol. 20, no. 2, pp. 156–175, 2010.
- [36] T. Li, M. Fu, L. Xie, and J. F. Zhang, "Distributed consensus with limited communication data rate," *IEEE Trans. Autom. Control*, vol. 56, no. 2, pp. 279–292, 2011.
- [37] Y. Zhao, T. Wang, and W. Bi, "Consensus protocol for multiagent systems with undirected topologies and binary-valued communications," *IEEE Trans. Autom. Control*, vol. 64, no. 1, pp. 206–221, 2019.
- [38] G. Cybenko, "Dynamic load balancing for distributed memory multiprocessors," *J. Parallel Distrib. Comput.*, vol. 7, no. 2, pp. 279–301, 1989.
- [39] D. Blatt and A. Hero, "Distributed maximum likelihood estimation for sensor networks," in *Int. Conf. Acoust. Speech Signal Process.*, Montreal, Canada, vol. 3, 2004, pp. 929–932.
- [40] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, vol. 30, 2017, pp. 5330–5340.
- [41] T. C. Aysal, M. J. Coates, and M. G. Rabbat, "Distributed average consensus with dithered quantization," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4905–4918, 2008.
- [42] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, vol. 32, 2019, pp. 14774–14784.
- [43] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Proc. Mach. Learn. Knowl. Discov. Databases Euro. Conf.*, Riva del Garda, Italy, 2016, pp. 795–811.
- [44] Y. S. Chow and H. Teicher, "Integration in a probability space," in *Probability theory: independence, interchangeability, martingales*. New York, NY, USA: Springer-Verlag, 2012, ch. 4, sec. 1, pp. 84–92.
- [45] S. Bubeck, "Convex optimization: algorithms and complexity," *Found. Trends Theor. Comput. Sci.*, vol. 8, nos. 3–4, pp. 231–357, 2015.
- [46] Y. LeCun, C. Cortes, and C. J. C. Burges, 1998, "The MNIST database of handwritten digits," National Institute of Standards and Technology. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [47] A. Krizhevsky, V. Nair, and G. Hinton, 2009, "Canadian Institute for Advanced Research, 10 classes," Department of Computer Science of University of Toronto. [Online]. Available: <http://www.cs.toronto.edu/kriz/cifar.html>
- [48] A. Krizhevsky, V. Nair, and G. Hinton, 2009, "Canadian Institute for Advanced Research, 100 classes," Department of Computer Science of University of Toronto. [Online]. Available: <http://www.cs.toronto.edu/kriz/cifar.html>
- [49] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, CA, 2009. [Online]. Available: <http://www.cs.utoronto.ca/kriz/learning-features-2009-TR.pdf>
- [50] V. A. Zorich, "Integration," in *Mathematical analysis I*, Berlin, German: Springer-Verlag, 2015, ch. 6, sec. 2, pp. 349–360.
- [51] R. A. Horn and C. R. Johnson, "Hermitian matrices, symmetric matrices, and congruences," in *Matrix analysis*, Cambridge, U.K.: Cambridge University Press, 2012, ch. 4, sec. 2, pp. 234–239.



Jialong Chen received the B.S. degree in Beihang University, Beijing, China, in 2023. He is currently working toward the Ph.D. degree majoring in system theory at Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing, China. His research interest is privacy and security in stochastic systems.



Jimin Wang received the B.S. degree in mathematics from Shandong Normal University, China, in 2012 and the Ph.D. degree from School of Mathematics, Shandong University, China, in 2018. From May 2017 to May 2018, he was a joint Ph.D. student with the School of Electrical Engineering and Computing, The University of Newcastle, Australia. From July 2018 to December 2020, he was a postdoctoral researcher in the Institute of Systems Science (ISS), Chinese Academy of Sciences (CAS), China. He

is currently an associate professor in the School of Automation and Electrical Engineering, University of Science and Technology Beijing. His current research interests include privacy and security in cyber-physical systems, stochastic systems and networked control systems.

He is a member of the IEEE CSS Technical Committee on Security and Privacy, the IEEE CSS Technical Committee on Networks and Communication Systems, the IFAC Technical Committee 1.5 on Networked Systems. He was a recipient of Shandong University's excellent doctoral dissertation.



Ji-Feng Zhang received the B.S. degree in mathematics from Shandong University, China, in 1985 and the Ph.D. degree from the Institute of Systems Science (ISS), Chinese Academy of Sciences (CAS), China, in 1991. Since 1985, he has been with the ISS, CAS. Now he is also with the School of Automation and Electrical Engineering, Zhongyuan University of Technology. His current research interests include system modeling, adaptive control, stochastic systems, and multi-agent systems.

He is an IEEE Fellow, IFAC Fellow, CAA Fellow, CSIAM Fellow, member of the European Academy of Sciences and Arts, and Academician of the International Academy for Systems and Cybernetic Sciences. He received the second prize of the State Natural Science Award of China in 2010 and 2015, respectively. He was a Vice-Chair of the IFAC Technical Board, member of the Board of Governors, IEEE Control Systems Society; Convenor of Systems Science Discipline, Academic Degree Committee of the State Council of China; Vice-President of the Systems Engineering Society of China, the Chinese Mathematical Society, and the Chinese Association of Automation. He has served as Editor-in-Chief, Deputy Editor-in-Chief, Senior Editor or Associate Editor for more than 10 journals, including *Science China Information Sciences*, *National Science Review*, *IEEE Transactions on Automatic Control*, and *SIAM Journal on Control and Optimization* etc.