Differentially Private Gradient-Tracking-Based Distributed Stochastic Optimization over Directed Graphs

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

***Abstract*— This paper proposes a differentially private gradient-tracking-based distributed stochastic optimization algorithm over directed graphs. In particular, privacy noises are incorporated into each agent's state and tracking variable to mitigate information leakage, after which the perturbed states and tracking variables are transmitted to neighbors. We design two novel schemes for the step-sizes and the sampling number within the algorithm. The sampling parameter-controlled subsampling method employed by both schemes enhances the differential privacy level, and ensures a finite cumulative privacy budget even over infinite iterations. The algorithm achieves both almost sure and mean square convergence for nonconvex objectives. Furthermore, when nonconvex objectives satisfy the Polyak-Łojasiewicz condition, *Scheme (S1)* achieves a polynomial mean square convergence rate, and *Scheme (S2)* achieves an exponential mean square convergence rate. The trade-off between privacy and convergence is presented. The effectiveness of the algorithm and its superior performance compared to existing works are illustrated through numerical examples of distributed training on the benchmark datasets "MNIST" and "CIFAR-10".**

***Index Terms*— Differential privacy, distributed stochastic optimization, gradient-tracking, exponential mean square convergence rate, directed graphs.**

I. Introduction

DISTRIBUTED optimization allows cooperative agents to compute and update their state variables through inter-agent communication to obtain an optimal solution of a common optimization problem ([1]). Distributed stochastic optimization, a branch of distributed optimization, address scenarios where objectives are stochastic ([2]). This approach has found extensive applications across multiple domains, including distributed machine learning ([3]), cloud-based control systems ([4]), and the Internet of Things ([5]). While it is frequently utilized in distributed stochastic optimization because of its adaptability in communication-efficient methods ([6]) and simplicity in algorithm structure ([7]), stochastic gradient descent (SGD) does not guarantee the convergence over directed graphs ([8, Eq. 6]), and cannot achieve the exponential convergence rate ([9, Th. 2], [10, Eq. 2]). To address these issues, the gradient-tracking method has been proposed over undirected graphs ([11], [12]). By developing tracking variables to track global stochastic gradients, [11]–[13] initially achieve the exponential convergence rate. The convergence analysis is further extended to directed graphs in [14]–[17]. However, [14]–[16] prove convergence under the assumption that weight matrices are row- and column-stochastic, which is often difficult to be satisfied in various practical scenarios (see e.g. [4], [5]). [17] achieves the convergence by employing the two-time-scale step-sizes method, which removes the assumption that weight matrices are row- and column-stochastic, while requiring that the level sets of objectives are bounded.

When cooperative agents exchange information to address a distributed stochastic optimization problem, adversaries can infer stochastic gradients from the exchanged information, and further obtain agents' sensitive information through model inversion attacks ([18], [19]). To address this issue, various privacy-preserving techniques have been developed ([20]), such as homomorphic encryption ([21]), state decomposition ([22]), random coupling weights ([23]), uncoordinated step-sizes ([24]), network augmentation ([25]), and adding noises ([26]–[28]). Because of its simplicity of use and immunity to post-processing, differential privacy ([27], [28]) has attracted considerable interest and has been extensively applied in distributed optimization for both deterministic and stochastic objectives. When objectives are deterministic, based on the gradient-tracking method, differentially private distributed optimization has been well developed in [29]–[34]. Among others, [29]–[32], [34] have successfully achieved the finite cumulative differential privacy budget over infinite iterations. However, the difficulty caused by stochastic objectives makes the methods in [29]–[34] unsuitable to differentially private distributed stochastic optimization. In addition, to achieve convergence, (strongly) convex objectives ([29]–[32], [34]) and nonconvex objectives with the Polyak-Łojasiewicz condition ([33]) are required. However, these requirements may be hard

The work was supported by National Natural Science Foundation of China under Grants 62203045, 62433020 and T2293770. The material in this paper was not presented at any conference.

Jialong Chen is with the State Key Laboratory of Mathematical Sciences, 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 State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. (e-mail: jif@iss.ac.cn)

to be satisfied or verified in practice.

When objectives are stochastic, a method based on SGD has been proposed for differentially private distributed stochastic optimization. Some interesting works can be found in [35]–[41]. However, [35]–[41] only give the per-iteration differential privacy budget, and thus, cannot protect the sensitive information over infinite iterations. Fortunately, by sequentially acquiring data samples in distributed online learning ([42]), the time-varying sampling number method ([43]) and the sampling parameter-controlled subsampling method ([44]), the finite cumulative differential privacy budget over infinite iterations is given. In [42]–[44], the differential privacy is tailored for distributed SGD, and the convergence is given over undirected graphs. Although the gradient-tracking method has shown advantages over the distributed-SGD method regarding the convergence over directed graphs, to the best of our knowledge, differentially private gradient-tracking-based distributed stochastic optimization has not been studied yet. As a result, the differentially private distributed stochastic optimization based on the gradient-tracking method is a challenging issue, especially on how to achieve the finite cumulative differential privacy budget even over infinite iterations, the almost sure and mean square convergence for nonconvex objectives without the Polyak-Łojasiewicz condition, and the exponential mean square convergence rate.

Summarizing the discussion above, we propose a new differentially private gradient-tracking-based distributed stochastic optimization algorithm with two schemes of step-sizes and the sampling number over directed graphs. *Scheme (S1)* employs the polynomially decreasing step-sizes and the increasing sampling number with the maximum iteration number. *Scheme (S2)* employs constant step-sizes and the exponentially increasing sampling number with the maximum iteration number. The main contribution of this paper is as follows:

- The sampling parameter-controlled subsampling method is employed to enhance the differential privacy level of the algorithm. The algorithm with both schemes achieves the finite cumulative differential privacy budget even over infinite iterations. To the best of our knowledge, a finite cumulative differential privacy budget over infinite iterations is achieved in differentially private gradient-tracking-based distributed stochastic optimization for the first time.
- The almost sure and mean square convergence of the algorithm are given for nonconvex objectives without the Polyak-Łojasiewicz condition. Furthermore, when nonconvex objectives satisfy the Polyak-Łojasiewicz condition, the polynomial mean square convergence rate is achieved for *Scheme (S1)*, and the exponential mean square convergence rate is achieved for *Scheme (S2)*.
- Two schemes are shown to achieve the finite cumulative differential privacy budget over infinite iterations and mean square convergence simultaneously. For *Scheme (S1)*, the polynomial mean square convergence rate and the cumulative differential privacy budget are achieved simultaneously even over infinite iterations for general privacy noises, including decreasing, constant and increasing privacy noises. For *Scheme (S2)*, the exponential mean square convergence rate and the cumulative differential privacy budget are achieved simultaneously even over infinite iterations.

The remainder of this paper is organized as follows: Section II presents preliminaries and the problem formulation. Section III provides the algorithm with its convergence and privacy analysis. Section IV verifies the effectiveness of the algorithm through numerical examples of distributed training on the benchmark datasets "MNIST" and "CIFAR-10". Finally, Section V concludes the paper.

Notation. $\mathbb{R}$ and $\mathbb{R}^n$ denote the set of real numbers and n-dimensional Euclidean space, respectively. $\mathbf{1}_n$ denotes a n-dimensional vector whose elements are all 1, and $\|x\|$ denotes the standard Euclidean norm of a vector x. $X \sim \mathrm{Lap}(b)$ refers to a random variable that has a Laplacian distribution with the variance parameter $b > 0$, and the probability density function of the random variable X is given by $p(x; b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$. For a matrix $A \in \mathbb{R}^{n\times n}$, $A^\top$, ρ_A stand for its transpose and spectral radius, respectively. $\langle\cdot,\cdot\rangle$ denotes the inner product. $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a probability space, $\mathbb{P}(B)$ and $\mathbb{E}X$ stand for the probability of an event $B \in \mathcal{F}$ and the expectation of the random variable X, respectively. $\otimes$ denotes the Kronecker product of matrices. $\lfloor z \rfloor$ denotes the largest integer which is not larger than z. For a differentiable function $f(x)$, $\nabla f(x)$ denotes its gradient at the point x. For a vector $x = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$, the notation $\mathrm{diag}(x)$ denotes the diagonal matrix with diagonal elements being $x_1, x_2, \ldots, x_n$.

II. Preliminaries and problem formulation

A. Graph theory

In this paper, we consider a network of n agents which exchange the information over two different directed graphs $\mathcal{G}_R = (\mathcal{V}, \mathcal{E}_R)$ and $\mathcal{G}_C = (\mathcal{V}, \mathcal{E}_C)$. $\mathcal{V} = \{1, 2, \ldots, n\}$ is the set of all agents, and $\mathcal{E}_R$, $\mathcal{E}_C$ are sets of directed edges in $\mathcal{G}_R$, $\mathcal{G}_C$, respectively. In our gradient-tracking algorithm, agents exchange state variables over $\mathcal{G}_R$ and tracking variables over $\mathcal{G}_C$. Directed graphs $\mathcal{G}_R$ and $\mathcal{G}_C$ are induced by the weight matrix $R = (R_{ij})_{i,j=1,\ldots,n}$ and $C = (C_{ij})_{i,j=1,\ldots,n}$, respectively. Any element R_{ij} of R is either strictly positive if Agent i can receive Agent j's state variable, or 0, otherwise. The same property holds for any element C_{ij} of C. For any $i \in \mathcal{V}$, its in-neighbor and out-neighbor set of over $\mathcal{G}_R$ are defined as $\mathcal{N}^-_{R,i} = \{j \in \mathcal{V} : R_{ij} > 0, j \neq i\}$ and $\mathcal{N}^+_{R,i} = \{j \in \mathcal{V} : R_{ji} > 0, j \neq i\}$, respectively. Similarly, Agent i's in-neighbor and out-neighbor set over $\mathcal{G}_C$ are defined as $\mathcal{N}^-_{C,i}$ and $\mathcal{N}^+_{C,i}$, respectively. Denote the in-Laplacian matrix of R and the out-Laplacian matrix of C as $\mathcal{L}_1 = \mathrm{diag}(R\cdot\mathbf{1}_n) - R$ and $\mathcal{L}_2 = \mathrm{diag}(\mathbf{1}_n^\top C) - C$, respectively. Then, the assumption about directed graphs $\mathcal{G}_R$, $\mathcal{G}_C$ is given as follows:

Assumption 1: Let $\mathcal{G}_R$ and $\mathcal{G}_{C^\top}$ be directed graphs induced by nonnegative matrices R and $C^\top$, respectively. Then, both $\mathcal{G}_R$ and $\mathcal{G}_{C^\top}$ contain at least one spanning tree. Moreover, there exists at least one agent being a root of spanning trees in both $\mathcal{G}_R$ and $\mathcal{G}_{C^\top}$.

Remark 1: Directed graphs in Assumption 1 are more general than undirected connected graphs in [6], [10]–[13], [29], [33], [35]–[40], [42]–[44], directed graphs with stochastic weight matrices in [14]–[16], and strongly connected directed graphs in [17], [23]. In addition, by [45, Th. 3.8], Assumption

1 is a necessary condition for the consensus of Agents' state and tracking variables.

Based on Assumption 1, we have the following useful lemma for weight matrices R and C:

Lemma 1: ([1, Lemmas 1, 3]) Let α_K, β_K be positive constants such that $I_n-\alpha_K\mathcal{L}_1$ and $I_n-\beta_K\mathcal{L}_2$ are nonnegative matrices. If Assumption 1 holds, then we have the following statements:

(i) There exist unique nonnegative vectors $v_1,\, v_2\in\mathbb{R}^n$ such that $v_1^\top(I_n-\alpha_K\mathcal{L}_1)=v_1^\top$, $(I_n-\beta_K\mathcal{L}_2)v_2=v_2$, $v_1^\top\mathbf{1}_n=n$, $v_2^\top\mathbf{1}_n=n$, $v_1^\top v_2>0$.

(ii) There exist $r_{\mathcal{L}_1},\, r_{\mathcal{L}_2}>0$ such that the spectral radius of the matrices $I_n-\alpha_K\mathcal{L}_1-\frac{1}{n}\mathbf{1}_n v_1^\top$ and $I_n-\beta_K\mathcal{L}_2-\frac{1}{n}v_2\mathbf{1}_n^\top$ are $1-\alpha_K r_{\mathcal{L}_1}$ and $1-\beta_K r_{\mathcal{L}_2}$, respectively.

B. Problem formulation

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

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

where x is available to all agents, $\ell_i(x,\xi_i)$ is a local objective which is private to Agent i, ξ_i is a random variable drawn from an unknown probability distribution $\mathscr{D}_i$, and $\mathscr{D}_i$ is not required to be independent and identically distributed for any $i\in\mathcal{V}$. In practice, since the probability distribution $\mathscr{D}_i$ is difficult to obtain, it is usually replaced by the dataset $\mathcal{D}_i=\{\xi_{i,l}, l=1,\dots,D\}$. Then, (1) can be rewritten as the following empirical risk minimization problem:

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

To solve the problem (2), we need the following standard assumption:

Assumption 2: (i) For any $i\in\mathcal{V}$, there exists $L>0$ such that f_i is L-smooth, i.e., $\|\nabla f_i(x)-\nabla f_i(y)\|\le L\|x-y\|$, $\forall x,y\in\mathbb{R}^d$.

(ii) There exists $\sigma_g>0$ and a stochastic first-order oracle such that for any $i\in\mathcal{V}$, $x\in\mathbb{R}^d$ and λ_i generated by uniformly sampling ξ_i from $\mathcal{D}_i$, the stochastic first-order oracle returns a sampled gradient $g_i(x,\lambda_i)$ satisfying $\mathbb{E}[g_i(x,\lambda_i)]=\nabla f_i(x)$, $\mathbb{E}[\|g_i(x,\lambda_i)-\nabla f_i(x)\|^2]\le\sigma_g^2$.

Remark 2: Assumption 2(i) requires that each objective f_i has L-Lipschitz continuous gradients, which is commonly used (see e.g. [1], [7], [10]–[17], [23], [29], [31]–[33], [35]–[44]). Assumption 2(ii) requires that each sampled gradient $g_i(x,\lambda_i)$ is unbiased with a bounded variance σ_g^2, which is standard for distributed stochastic optimization (see e.g. [10]–[14], [16], [17], [35], [37], [39], [40], [42]–[44]).

Next, assumptions for the nonconvex and strongly convex global objective are respectively given as follows:

Assumption 3: There exists $x^*\in\mathbb{R}^d$ such that $F(x^*)$ is the global minimum of the nonconvex global objective $F(x)$. Moreover, the Polyak-Łojasiewicz condition holds, i.e., there exists $\mu>0$ such that $2\mu(F(x)-F(x^*))\le\|\nabla F(x)\|^2$, $\forall x\in\mathbb{R}^d$.

Remark 3: Assumption 3 requires the gradient $\nabla F(x)$ to grow faster than a quadratic function as we move away from the global minimum, which is commonly used (see e.g. [7], [16], [33], [36], [44]).

Remark 4: There exists functions that satisfy Assumptions 2, 3 simultaneously. We give two examples. One example is $l_i(x,\xi_i)=\frac{1}{2n}\|\mathbf{A}x-\mathbf{d}\|^2+\frac{\|x\|\xi_i}{1+\|x\|}$, where $x\in\mathbb{R}^d$, the matrix $\mathbf{A}\in\mathbb{R}^{m\times d}$ has full column rank, $\mathbf{d}\in\mathbb{R}^d$ is a constant vector, and $\xi_i\sim N(0,4)$ is a Gaussian noise. Denote $\rho_{\mathbf{A}},\Theta_{\mathbf{A}^\top\mathbf{A}}>0$ as the spectral radius of $\mathbf{A}$ and the minimum eigenvalue of $\mathbf{A}^\top\mathbf{A}$, respectively. Then, by [46, Th. 2] $f_i(x)=\frac{1}{2n}\|\mathbf{A}x-\mathbf{d}\|^2$ satisfies Assumption 2 with $L=\frac{\rho_{\mathbf{A}}^2}{2n}$, $\sigma_g=2$, and $F(x)$ satisfies Assumption 3 with $\mu=2\Theta_{\mathbf{A}^\top\mathbf{A}}^2$. Another example is $l_i(x,\xi_i)=x^2+(3+\xi_i)(\sin x)^2+2\xi_i\cos x$, where $x\in\mathbb{R}$, and $\xi_i\sim\mathrm{Lap}(\frac{1}{2})$ is a Laplacian noise. Then, by [46, Subsec. 2.2], $f_i(x)=x^2+3(\sin x)^2$ satisfies Assumption 2 with $L=8$, $\sigma_g=\frac{5}{2}$, and $F(x)$ satisfies Assumption 3 with $\mu=\frac{n}{32}$.

In practice, since finding the exact optimal solution is computationally expensive and time-consuming, suboptimal solutions within a given error $\varphi>0$ are often preferred. Inspired by [2], the φ-suboptimal solution and the oracle complexity are defined as follows:

Definition 1: Let $\varphi>0$, $K=0,1,\dots$, $x_K=[x_{1,K}^\top,\dots,x_{n,K}^\top]^\top$ be the output of an algorithm. Then, x_K is a φ-suboptimal solution if $\mathbb{E}\|\nabla F(x_{i,K+1})\|^2<\varphi$, $\forall i\in\mathcal{V}$.

Definition 2: Let $\varphi>0$, $N(\varphi)=\min\{K: x_K$ is a φ-suboptimal solution$\}$, and m_k be the sampling number at the k-th iteration. Then, the oracle complexity of the algorithm is $\sum_{k=0}^{N(\varphi)}m_k$.

C. Differential privacy

As shown in [38], [39], [42], 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 an agent within the network which has access to certain internal information (such as state variable $x_{i,k}$ and tracking variable $y_{i,k}$ of Agent i), follows the prescribed protocols and accurately computes iterative state and tracking correctly. However, it aims to infer the sensitive information of other agents.
- 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 agent. This enables the eavesdropper to infer the sensitive information of agents.

When cooperative agents exchange information to solve the empirical risk minimization problem (2), these two kinds of adversaries can use the model inversion attack ([18]) to infer sampled gradients, and further obtain the sensitive information in agents' data samples from sampled gradients ([19]). In order to provide the privacy protection for data samples, a symmetric binary relation called *adjacency relation* is defined as follows:

Definition 3: ([44]) Let $\mathcal{D}=\{\xi_{i,l}, i\in\mathcal{V}, l=1,\dots,D\}$, $\mathcal{D}'=\{\xi'_{i,l},\ i\in\mathcal{V}, l=1,\dots,D\}$ be two sets of data samples. If for a given $C>0$, 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 for any $x \in \mathbb{R}^r$,

$$\begin{cases} 0 < \|g_i(x, \xi_{i,l}) - g_i(x, \xi'_{i,l})\| \le C, & \text{if } i = i_0 \text{ and } l = l_0; \\ \|g_i(x, \xi_{i,l}) - g_i(x, \xi'_{i,l})\| = 0, & \text{if } i \neq i_0 \text{ or } l \neq l_0, \end{cases} \tag{3}$$

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

Remark 5: The boundary C characterizes the "closeness" of a pair of data samples ξ_{i_0,l_0}, ξ'_{i_0,l_0}. The larger the boundary C is, the larger the allowed magnitude of sampled gradients between adjacent datasets is. For more details, please refer to [44, Subsec. II-D].

Remark 6: Different from the adjacency relation defined in differentially private distributed optimization ([29]–[34]), Definition 3 is given by allowing one data sample of one agent to be different, which is commonly used in differentially private distributed stochastic optimization ([35]–[44]).

Next, the definition of differential privacy is given to show the privacy-preserving level of the algorithm:

Definition 4: ([28]) Let $\varepsilon \ge 0$ be the differential privacy budget. Then, the randomized algorithm $\mathcal{M}$ achieves ε-differential privacy for $\text{Adj}(\mathcal{D}, \mathcal{D}')$ if $\mathbb{P}(\mathcal{M}(\mathcal{D}) \in \mathcal{O}) \le e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathcal{D}') \in \mathcal{O})$ holds for any observation set $\mathcal{O} \subseteq \text{Range}(\mathcal{M})$.

Remark 7: As shown in [35]–[44], the differential privacy budget ε measure the similarity of the randomized algorithm $\mathcal{M}$'s output distributions under two adjacent datasets $\mathcal{D}$, $\mathcal{D}'$. The smaller the differential privacy budget ε is, the higher the differential privacy level is.

Remark 8: Both ε-differential privacy and (ε, δ)-differential privacy has been used in differentially private distributed stochastic optimization. ε-differential privacy is usually achieved by Laplacian noises, while (ε, δ)-differential privacy is usually achieved by Gaussian noises. To simplify the analysis, ε-differential privacy is used in this paper. If (ε, δ)-differential privacy is used, then the framework of the convergence and privacy analysis still holds.

Problem of interest: In this paper, we first aim to propose a new differentially private gradient-tracking-based algorithm for the problem (2) over directed graphs; then design schemes of step-sizes and the sampling number to enhance the differential privacy level, achieve the almost sure and mean square convergence for nonconvex objectives without the Polyak-Łojasiewicz condition, and further accelerate the convergence rate.

III. Main results

A. The proposed algorithm

In this subsection, we propose a differentially private gradient-tracking-based distributed stochastic optimization algorithm over directed graphs. Detailed steps are given in Algorithm 1.

For the convenience of the analysis, let $x_k = [x_{1,k}^\top, \dots, x_{n,k}^\top]^\top$ $y_k = [y_{1,k}^\top, \dots, y_{n,k}^\top]^\top$, $\zeta_k = [\zeta_{1,k}^\top, \dots, \zeta_{n,k}^\top]^\top$, $\eta_k = [\eta_{1,k}^\top, \dots, \eta_{n,k}^\top]^\top$, $g_k = [g_{1,k}^\top, \dots, g_{n,k}^\top]^\top$. Then, (4) and (6) can be written in the following compact form:

$$x_{k+1} = ((I_n - \alpha_K \mathcal{L}_1) \otimes I_d)x_k - \alpha_K(\mathcal{L}_1 \otimes I_d)\zeta_k - \gamma_K y_k, \tag{7}$$

$$y_{k+1} = ((I_n - \beta_K \mathcal{L}_2) \otimes I_d)y_k - \beta_K(\mathcal{L}_2 \otimes I_d)\eta_k + g_{k+1} - g_k. \tag{8}$$

Algorithm 1 Differentially private gradient-tracking-based distributed stochastic optimization algorithm over directed graphs

Initialization: $x_{i,0} \in \mathbb{R}^d$ for any $i \in \mathcal{V}$, m_K different data samples $\lambda_{i,0,1}, \dots, \lambda_{i,0,m_K}$ in $\mathcal{D}_i$, $y_{i,0} = g_{i,0} = \frac{1}{m_K}\sum_{l=1}^{m_K} g_i(x_{i,0}, \lambda_{i,0,l})$ for any $i \in \mathcal{V}$, weight matrices $R = (R_{ij})_{i,j=1,\dots,n}$, $C = (C_{ij})_{i,j=1,\dots,n}$, the maximum iteration number K, step-sizes $\alpha_K, \beta_K, \gamma_K$ and the sampling number m_K.

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

1: Agent i adds independent d-dimensional Laplacian noises $\zeta_{i,k}$, $\eta_{i,k}$ to its state variable $x_{i,k}$ and tracking variable $y_{i,k}$, respectively: $\breve{x}_{i,k} = x_{i,k} + \zeta_{i,k}$, $\breve{y}_{i,k} = y_{i,k} + \eta_{i,k}$, where each coordinate of $\zeta_{i,k}$, $\eta_{i,k}$ has the distribution $\text{Lap}(\sigma_k^{(\zeta)})$ and $\text{Lap}(\sigma_k^{(\eta)})$, respectively.

2: Agent i broadcasts its perturbed state variable $\breve{x}_{i,k}$ to all its out-neighbors in $\mathcal{N}_{R,i}^+$, and broadcasts its perturbed tracking variable $\breve{y}_{i,k}$ to all its out-neighbors in $\mathcal{N}_{C,i}^+$.

3: Agent i receives $\breve{x}_{j,k}$ from all its in-neighbors in $\mathcal{N}_{R,i}^-$ and $\breve{y}_{j,k}$ from all its in-neighbors in $\mathcal{N}_{C,i}^-$.

4: Agent i updates its state variable by

$$x_{i,k+1} = (1 - \alpha_K \sum_{j \in \mathcal{N}_{R,i}^-} R_{ij})x_{i,k} + \alpha_K \sum_{j \in \mathcal{N}_{R,i}^-} R_{ij}\breve{x}_{j,k} - \gamma_K y_{i,k}. \tag{4}$$

5: Agent i takes m_K different samples $\lambda_{i,k+1,1}, \dots, \lambda_{i,k+1,m_K}$ uniformly from $\mathcal{D}_i$ to generate sampled gradients $g_i(x_{i,k+1}, \lambda_{i,k+1,1}), \dots, g_i(x_{i,k+1}, \lambda_{i,k+1,m_K})$. Then, Agent i puts these data samples back into $\mathcal{D}_i$.

6: Agent i computes the averaged sampled gradient by

$$g_{i,k+1} = \frac{1}{m_K}\sum_{l=1}^{m_K} g_i(x_{i,k+1}, \lambda_{i,k+1,l}). \tag{5}$$

7: Agent i updates its tracking variable by

$$y_{i,k+1} = (1 - \beta_K \sum_{j \in \mathcal{N}_{C,i}^-} C_{ij})y_{i,k} + \beta_K \sum_{j \in \mathcal{N}_{C,i}^-} C_{ij}\breve{y}_{j,k} + g_{i,k+1} - g_{i,k}. \tag{6}$$

end for

Return $x_{1,K+1}, \dots, x_{n,K+1}$

B. Convergence analysis

In this subsection, we will give the convergence analysis of Algorithm 1. First, we give two different schemes of step-sizes and the sampling number for Algorithm 1:

Scheme (S1): For any $K = 0, 1, \dots,$

(I) step-sizes: $\alpha_K = \frac{a_1}{(K+1)^{p_\alpha}}$, $\beta_K = \frac{a_2}{(K+1)^{p_\beta}}$, $\gamma_K = \frac{a_3}{(K+1)^{p_\gamma}}$,

(II) the sampling number: $m_K = \lfloor a_4 K^{p_m} \rfloor + 1$,

where $a_1, a_2, a_3, a_4 > 0, p_\alpha, p_\beta, p_\gamma > 0, p_m \ge 0$.

Scheme (S2): For any $K = 0, 1, \dots,$

(I) step-sizes: $\alpha_K = \alpha, \beta_K = \beta, \gamma_K = \gamma$ are constants,

(II) the sampling number: $m_K = \lfloor p_m^K \rfloor + 1$,

where $\alpha, \beta, \gamma > 0$, $p_m \ge 0$.

To get the almost sure and mean square convergence of Algorithm 1, we need the following assumptions:

Assumption 4: Under *Scheme (S1)*, step-sizes α_K, β_K, γ_K, the sampling number m_K, and privacy noise parameters $\sigma_k^{(\zeta)} = (k+1)^{p_\zeta}$, $\sigma_k^{(\eta)} = (k+1)^{p_\eta}$ satisfy the following conditions:

$$\frac{1}{2}<p_\beta<p_\alpha<p_\gamma<1, 2p_\gamma-p_\alpha\geq 1, 2p_\alpha-2p_\zeta-p_\beta\geq 1,$$
$$2p_\alpha-p_\beta\geq 1, 2p_\beta-2p_\eta\geq 1, p_m-p_\beta\geq 1, a_3<\frac{n}{4(v_1^\top v_2)L},$$
$$a_1<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-}R_{ij}}\},\frac{1}{r_{\mathcal{L}_1}}\}, a_2<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^-}C_{ij}}\},\frac{1}{r_{\mathcal{L}_2}}\}.$$

Assumption 5: Under *Scheme (S2)*, step-sizes α,β,γ, the sampling number m_K, and privacy noise parameters $\sigma_k^{(\zeta)}=p_\zeta^K$, $\sigma_k^{(\eta)}=p_\eta^K$ satisfy the following conditions:

$$0<p_\zeta,p_\eta<1, p_m>1, \beta<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^-}C_{ij}}\},\frac{1}{r_{\mathcal{L}_2}}\},$$
$$\alpha<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-}R_{ij}}\},\frac{1}{r_{\mathcal{L}_1}},\frac{\sqrt{330n}(v_1^\top v_2)r_{\mathcal{L}_2}}{66n\|v_1\|\rho_{W_2}\rho_{\mathcal{L}_1}}\beta\},$$
$$\gamma<\min\{1,\frac{n}{15(v_1^\top v_2)L},Q_1\alpha,Q_2\beta\},$$

where

$$Q_1=\min\{\frac{\sqrt{n}r_{\mathcal{L}_1}}{8\rho_{W_1}\|v_2\|L},\frac{\sqrt{6n}r_{\mathcal{L}_1}}{24\rho_{W_1}\|v_2\|L}, \frac{\|v_1\|r_{\mathcal{L}_1}}{2\rho_{W_1}L}\sqrt{\frac{n\mu}{\|v_1\|^2\|v_2\|^2(16L^2+3\mu)}}\},$$
$$Q_2=\min\{\frac{\sqrt{2}r_{\mathcal{L}_2}}{6\rho_{W_2}L},\frac{r_{\mathcal{L}_1}r_{\mathcal{L}_2}}{24\rho_{W_1}\rho_{W_2}\rho_{\mathcal{L}_1}L},\frac{\sqrt{2n(v_1^\top v_2)^3}r_{\mathcal{L}_2}}{6\rho_{W_2}\|v_1\|\|v_2\|L}, \frac{(v_1^\top v_2)r_{\mathcal{L}_1}r_{\mathcal{L}_2}}{36\|v_1\|\|v_2\|\rho_{W_2}L}\sqrt{\frac{6}{4\rho_{W_1}^2\rho_{\mathcal{L}_1}^2+r_{\mathcal{L}_1}^2}},\frac{r_{\mathcal{L}_2}}{\rho_{W_2}L}\sqrt{\frac{v_1^\top v_2}{6n}}, \frac{(v_1^\top v_2)r_{\mathcal{L}_2}}{66\rho_{W_2}L}\sqrt{\frac{660\mu}{\|v_1\|^2\|v_2\|^2(16L^2+3\mu)}}\}.$$

Theorem 1: If Assumptions 1, 2, 4 hold under *Scheme (S1)* or Assumptions 1, 2, 5 hold under *Scheme (S2)*, then Algorithm 1 achieves the almost sure and mean square convergence, i.e., $\lim_{K\to\infty}\|\nabla F(x_{i,K+1})\|^2=0$ a.s., and $\lim_{K\to\infty}\mathbb{E}\|\nabla F(x_{i,K+1})\|^2=0,\forall i\in\mathcal{V}$.

Proof. See Appendix B. ■

Remark 9: Algorithm 1 achieves the almost sure and mean square convergence for nonconvex objectives without the Polyak-Łojasiewicz condition. The condition imposed on objectives is weaker than (strongly) convex objectives ([10]–[15], [17], [29]–[32], [34]) or the Polyak-Łojasiewicz condition ([16], [33]). Thus, Algorithm 1 has wider applicability than [10]–[17], [29]–[34].

The polynomial mean square convergence rate and the oracle complexity of Algorithm 1 with *Scheme (S1)* are given as follows:

Theorem 2: Under Assumptions 1-3 and 4, Algorithm 1 with *Scheme (S1)* achieves the following polynomial mean square convergence rate for any $i\in\mathcal{V}$:

$$\mathbb{E}\|\nabla F(x_{i,K+1})\|^2=O\left(\frac{1}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}}\right), \quad (9)$$

where $\theta=\min\{p_m-p_\beta, 2p_\alpha-2p_\zeta-p_\beta, 2p_\alpha-p_\beta, 2p_\beta-2p_\eta, 2p_\beta, 2p_\gamma-p_\beta+p_m\}$. Furthermore, for any $0<\varphi<\frac{1}{3}$, if $p_\alpha=1-\frac{2\varphi}{11}$, $p_\beta=\frac{2}{3}(1+\frac{3\varphi}{11})$, $p_\gamma=1-\frac{\varphi}{11}$, $p_m=\varphi$, $p_\zeta=p_\eta=\frac{2\varphi}{11}$, then the oracle complexity of Algorithm 1 with *Scheme (S1)* is $O(\varphi^{-\frac{3+3\varphi}{1-3\varphi}})$.

Proof. See Appendix C. ■

Remark 10: In Theorem 2, the polynomial mean square convergence rate is given for privacy noises with decreasing, constant (see e.g. [35], [36], [38], [41]), and increasing variances (see e.g. [42]–[44]). This is non-trivial even without considering the privacy protection. For example, let step-sizes $\alpha_K=\frac{1}{(K+1)^{0.96}}$, $\beta_K=\frac{1}{(K+1)^{0.7}}$, $\gamma_K=\frac{1}{(K+1)^{0.98}}$. Then, Theorem 2 holds as long as privacy noise parameters $\sigma_k^{(\zeta)}$, $\sigma_k^{(\eta)}$ have the increasing rate no more than $O(k^{0.6})$.

Remark 11: The key to achieving the polynomial mean square convergence rate without the assumption of bounded gradients is to use polynomially decreasing step-sizes and the increasing sampling number, which reduces the effect of stochastic gradient noises and privacy noises. This is different from [6], [7], [35]–[39], [41], [42], where the assumption of bounded gradients is required.

Remark 12: Theorem 2 shows that the oracle complexity of Algorithm 1 with *Scheme (S1)* is $O(\varphi^{-\frac{3+3\varphi}{1-3\varphi}})$. For example, if the error $\varphi=0.02$, then the oracle complexity is $O(10^5)$. This requirement for total number of data samples is acceptable since the oracle complexity of centralized quasi-Newton algorithm in [47] is also $O(10^5)$ to achieve the same accuracy as Algorithm 1 with *Scheme (S1)*.

Next, the exponential mean square convergence rate and the oracle complexity of Algorithm 1 with *Scheme (S2)* are given:

Theorem 3: Under Assumptions 1-3 and 5, Algorithm 1 with *Scheme (S2)* achieves the following exponential mean square convergence rate for any $i\in\mathcal{V}$:

$$\mathbb{E}\|\nabla F(x_{i,K+1})\|^2=O\left(\max\left\{\rho_{A_K},\frac{1}{p_m},p_\zeta^2,p_\eta^2\right\}^K\right).$$

Furthermore, for any $0<\varphi<\min\{1,\frac{n}{15(v_1^\top v_2)L},\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-}R_{ij}}\},\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^-}C_{ij}}\},\frac{1}{r_{\mathcal{L}_1}},\frac{1}{r_{\mathcal{L}_2}}\}$, if $\beta=\varphi$, $\alpha=\min\{\varphi,\frac{\sqrt{330n}(v_1^\top v_2)r_{\mathcal{L}_2}\varphi}{132n\|v_1\|\rho_{W_2}\rho_{\mathcal{L}_1}}\}$, $\gamma=\min\{\frac{1}{2},\frac{n}{30(v_1^\top v_2)L},\frac{Q_1\alpha}{2},\frac{Q_2\varphi}{2}\}$, $p_m=\min\{\frac{1}{\varphi},\frac{1}{\rho_{A_K}}\}$, $p_\zeta=p_\eta=\varphi$, then the oracle complexity of Algorithm 1 with *Scheme (S2)* is $O(\frac{1}{\varphi}\ln\frac{1}{\varphi})$.

Proof. See Appendix D. ■

Remark 13: By Theorems 2, 3, *Scheme (S2)* achieves the exponential mean square convergence rate, while *Scheme (S1)* and methods in [6], [7], [15], [17], [35]–[44] achieve the polynomial mean square convergence rate. For example, when the index of convergence rate is $\frac{1}{K+1}\sum_{k=0}^K\mathbb{E}(F(\bar{x}_k)-F(x^*))$, methods in [38], [39] achieve convergence rates of $O(\frac{1}{\sqrt{K}})$ and $O(1)$, respectively. Since the method in [38] is the same as the one in [48], by [48, Th. 2],the method in [38] achieves the convergence rate of $O(\frac{1}{\sqrt{K}})$. By [39, Th. 2], the method in [39] achieves the convergence rate of $O(1)$. Thus, *Scheme (S2)* is suitable for the scenario where the convergence rate is prioritized. However, by Theorem 1, *Scheme (S1)* achieves the almost sure and mean square convergence under decreasing, constant, and increasing privacy noises, while *Scheme (S2)* achieves the almost sure and mean square convergence only under decreasing privacy noises. This shows the trade-off of Algorithm 1 between the convergence rate and the added privacy noises.

Remark 14: When the global objective $F(x)$ is strongly convex (i.e., there exists $s>0$ such that $F(y)\geq F(x)+\langle\nabla F(x),y-x\rangle+\frac{s}{2}\|y-x\|^2,\forall x,y\in\mathbb{R}^d$), by [49, Lemma 6.9], we have $2s(F(x)-F(x^*))\leq\|\nabla F(x)\|^2$. Then Assumption

3 is satisfied with $\mu = s$, and thus, Theorems 2, 3 also hold for strongly convex objectives. Hence, we provides a general frame for Algorithm 1's convergence rate analysis under both nonconvex objectives with Polyak-Łojasiewicz conditions and strongly convex objectives.

Remark 15: The oracle complexity of *Scheme (S2)* shows that the sampling number required to achieve the desired accuracy is lower than existing works (see e.g. [14]). By Theorem 3, the oracle complexity of *Scheme (S2)* is $O(\frac{1}{\varphi}\ln\frac{1}{\varphi})$, which is smaller than the oracle complexity $O(\frac{1}{\varphi^2})$ of the gradient-tracking-based algorithm in [14]. For example, when the error $\varphi = 0.02$, $O(10^2)$ data samples are required in *Scheme (S2)*, while $O(10^3)$ data samples are required in the gradient-tracking-based algorithm in [14]. Moreover, the increasing sampling number in both *Schemes (S1)* and *(S2)* is feasible in machine learning scenarios, such as the speech recognition problem ([50]), the simulated annealing problem ([51]), and the noun-phrase chunking problem ([52]).

C. Privacy analysis

In the following, the definition of the sensitivity is provided to compute the cumulative differential privacy budget ε of Algorithm 1.

Definition 5 ([44]): For any $k = 0,\dots,K$, let $\mathcal{D}, \mathcal{D}'$ be two groups of adjacent sample sets, q be a mapping, and $\mathcal{D}_k = \{\lambda_{i,k,l}, i\in\mathcal{V}, l = 1,\dots,m_K\}$, $\mathcal{D}'_k = \{\lambda'_{i,k,l}, i\in\mathcal{V}, l = 1,\dots,m_K\}$ 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:

$$\Delta_k^q \triangleq \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|q(\mathcal{D}_k) - q(\mathcal{D}'_k)\|_1. \tag{10}$$

Remark 16: Definition 5 captures the magnitude by which a single agent's data sample can change the mapping q in the worst case. It is the key quantity showing how many noises should be added such that Algorithm 1 achieves ε_k-differential privacy at the k-th iteration. In Algorithm 1, the mapping $q(\mathcal{D}_k) = [x_k^\top, y_k^\top]^\top$, and the randomized algorithm $\mathcal{M}(\mathcal{D}_k) = [(x_k+\zeta_k)^\top, (y_k+\eta_k)^\top]^\top$.

The following lemma gives the sensitivity Δ_k^q of Algorithm 1 for any $k = 0,\dots,K$.

Lemma 2: The sensitivity of Algorithm 1 at the k-th iteration satisfies $\Delta_k^q = \|\Delta x_k\|_1 + \|\Delta y_k\|_1$, where $\|\Delta x_k\|_1$ and $\|\Delta y_k\|_1$ are given as follows:

$$\|\Delta x_k\|_1 \le \begin{cases} 0, & \text{if } k=0;\\ \gamma_K\sum_{l=0}^{k-1}|1-\alpha_K\sum_{j\in\mathcal{N}^-_{R,i_0}}R_{i_0j}|^{k-l-1}\|\Delta y_l\|_1, & \text{if } k=1,\dots,K,\end{cases}$$

$$\|\Delta y_k\|_1 \le \begin{cases} \frac{C}{m_K}, & \text{if } k=0;\\ \sum_{l=0}^{k-1}|1-\beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j}|^{l}\frac{2C}{m_K} + |1-\beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j}|^{k}\frac{C}{m_K}, & \text{if } k=1,\dots,K.\end{cases}$$

Proof: See Appendix E. ■

Lemma 3: For any $K = 0,1,\dots,$ Algorithm 1 achieves ε-differential privacy over K iterations, where $\varepsilon = \sum_{k=0}^K(\frac{\|\Delta x_k\|_1}{\sigma_k^{(\zeta)}} + \frac{\|\Delta y_k\|_1}{\sigma_k^{(\eta)}})$.

Proof. See Appendix F. ■

Theorem 4: For step-sizes $\alpha_K,\beta_K,\gamma_K$, the sampling number m_K satisfying *Scheme (S1)*, and privacy noise parameters $\sigma_k^{(\zeta)} = (k+1)^{p_\zeta}$, $\sigma_k^{(\eta)} = (k+1)^{p_\eta}$, if $0 < a_1\sum_{j\in\mathcal{N}^-_{R,i_0}}R_{i_0j} < 1$, $0 < a_2\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j} < 1$, $p_m - p_\beta - \max\{0, 1-p_\eta\} > 0$, $p_m + \min\{0, p_\gamma - p_\alpha - p_\beta\} - \max\{0, 1-p_\zeta\} > 0$ hold, then the cumulative privacy budget ε is finite even over infinite iterations.

Proof. First, we compute $\sum_{k=0}^K\frac{\|\Delta y_k\|_1}{\sigma_k^{(\eta)}}$. Since $0 < a_2\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j} < 1$, it can be seen that $0 < \beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j} < 1$. When $k = 0,1$, $\|\Delta y_k\|_1 = O(\frac{1}{(K+1)^{p_m}})$ by Lemma 2. When $2\le k\le K$, we have

$$\begin{aligned}\|\Delta y_k\|_1 &= O\left(\frac{|1-\beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j}|(1-|1-\beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j}|^k)}{m_K(1-|1-\beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j}|)}\right)\\ &= O\left(\frac{1}{(K+1)^{p_m-p_\beta}}\right).\end{aligned} \tag{11}$$

Then, for any $k = 0,\dots,K$, $\|\Delta y_k\|_1 = O(\frac{1}{(K+1)^{p_m-p_\beta}})$, and $\sum_{k=0}^K\frac{\|\Delta y_k\|_1}{\sigma_k^{(\eta)}}$ can be rewritten as

$$\begin{aligned}\sum_{k=0}^K\frac{\|\Delta y_k\|_1}{\sigma_k^{(\eta)}} &= \frac{1}{(K+1)^{p_m-p_\beta}}O\left(\sum_{k=1}^K\frac{1}{k^{p_\eta}}\right)\\ &= O\left(\frac{\ln(K+2)}{(K+1)^{p_m-p_\beta-\max\{0,1-p_\eta\}}}\right).\end{aligned}$$

Hence, if $p_m - p_\beta - \max\{0,1-p_\eta\} > 0$ holds, then $\sum_{k=0}^\infty\frac{\|\Delta y_k\|_1}{\sigma_k^{(\eta)}}$ is finite.

Next, we compute $\sum_{k=0}^K\frac{\|\Delta x_k\|_1}{\sigma_k^{(\zeta)}}$. Since $0 < a_1\sum_{j\in\mathcal{N}^-_{R,i_0}}R_{i_0j} < 1$, it can be seen that $0 < \alpha_K\sum_{j\in\mathcal{N}^-_{R,i_0}}R_{i_0j} < 1$. When $k = 0,1$, by Lemma 2, $\|\Delta x_k\|_1 = O(\frac{1}{(K+1)^{p_m}})$. When $k = 2,\dots,K$, by (11), we have

$$\begin{aligned}\|\Delta x_k\|_1 &\le \sum_{t=1}^{k-1}|1-\alpha_K\sum_{j\in\mathcal{N}^-_{R,i_0}}R_{i_0j}|^{k-t}\gamma_K\|\Delta y_{t-1}\|_1 + \gamma_K\|\Delta y_{k-1}\|_1\\ &= O\left(\frac{1}{(K+1)^{p_m+p_\gamma-p_\alpha-p_\beta}}\right).\end{aligned}$$

Then, for any $0\le k\le K$, $\|\Delta x_k\|_1 = O(\frac{1}{(K+1)^{p_m+\min\{0,p_\gamma-p_\alpha-p_\beta\}}})$, and thus, $\sum_{k=0}^K\frac{\|\Delta x_k\|_1}{\sigma_k^{(\zeta)}}$ can be rewritten as

$$\begin{aligned}\sum_{k=0}^K\frac{\|\Delta x_k\|_1}{\sigma_k^{(\zeta)}} &= \frac{1}{(K+1)^{p_m+\min\{0,p_\gamma-p_\alpha-p_\beta\}}}O\left(\sum_{k=1}^K\frac{1}{k^{p_\zeta}}\right)\\ &= O\left(\frac{\ln(K+2)}{(K+1)^{p_m+\min\{0,p_\gamma-p_\alpha-p_\beta\}-\max\{0,1-p_\zeta\}}}\right).\end{aligned}$$

If $p_m + \min\{0, p_\gamma - p_\alpha - p_\beta\} - \max\{0, 1-p_\zeta\} > 0$, then $\sum_{k=0}^\infty\frac{\|\Delta x_k\|_1}{\sigma_k^{(\zeta)}}$ is finite. Hence, this theorem is proved. ■

Theorem 5: For step-sizes $\alpha_K,\beta_K,\gamma_K$, the sampling number m_K satisfying *Scheme (S2)*, and privacy noise parameters $\sigma_k^{(\zeta)} = p_\zeta^K$, $\sigma_k^{(\eta)} = p_\eta^K$, $0 < p_\zeta, p_\eta < 1$, if $0 < \alpha_K\sum_{j\in\mathcal{N}^-_{R,i_0}}R_{i_0j} < 1$, $0 < \beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j} < 1$, $\frac{1}{p_m} < \min\{p_\zeta, p_\eta\}$ hold, then the cumulative privacy budget ε is finite even over infinite iterations.

Proof. By Lemma 2, it can be seen that

$$\sum_{k=0}^K\frac{\|\Delta x_k\|_1}{\sigma_k^{(\zeta)}} + \frac{\|\Delta y_k\|_1}{\sigma_k^{(\eta)}} = O\left(K\left(\frac{1}{p_m p_\zeta}\right)^K + K\left(\frac{1}{p_m p_\eta}\right)^K\right).$$

Hence, if $\frac{1}{p_m} < \min\{p_\zeta, p_\eta\}$, then $\sum_{k=0}^\infty\frac{\|\Delta x_k\|_1}{\sigma_k^{(\zeta)}} + \frac{\|\Delta y_k\|_1}{\sigma_k^{(\eta)}}$ is finite. Therefore, this theorem is proved. ■

Remark 17: Theorems 4 and 5 establish the sufficient condition for Algorithm 1 with *Schemes (S1)*, *(S2)* to achieve the finite cumulative differential privacy budget ε even over infinite iterations, respectively. This is different from [6], [7], [10]–[17] that do not consider the privacy protection, and [35]–[41] that achieve the infinite cumulative differential privacy budget ε over infinite iterations. Thus, compared to [35]–[41], Algorithm 1 with both *Schemes (S1)* and *(S2)* provides a higher differential privacy level.

D. Trade-off between privacy and convergence

Based on Theorems 2-5, the trade-off between the privacy and the convergence is given in the following corollary:

Corollary 1: (i) If Assumptions 1-3, 4, $0<\alpha_K\sum_{j\in\mathcal{N}^-_{R,i_0}}R_{i_0j}<1$, $0<\beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j}<1$, $p_m-p_\beta-\max\{0,1-p_\eta\}>0$, and $p_m+\min\{0,p_\gamma-p_\alpha-p_\beta\}-\max\{0,1-p_\zeta\}>0$ hold, then Algorithm 1 with *Scheme (S1)* achieves the polynomial mean square convergence rate and the finite cumulative differential privacy budget ε even over infinite iterations simultaneously. (ii) If Assumptions 1-3, 5, $0<\alpha_K\sum_{j\in\mathcal{N}^-_{R,i_0}}R_{i_0j}<1$, $0<\beta_K\sum_{j\in\mathcal{N}^-_{C,i_0}}C_{i_0j}<1$, and $\frac{1}{p_m}<\min\{p_\zeta,p_\eta\}$ hold, then Algorithm 1 with *Scheme (S2)* achieves the exponential mean square convergence rate and the finite cumulative differential privacy budget ε even over infinite iterations simultaneously.

Proof. By Theorems 2 and 4, Corollary 1(i) is proved. Then, by Theorems 3 and 5, Corollary 1(ii) is proved. ■

Remark 18: Scheme (S1) achieves the polynomial mean square convergence rate and the finite cumulative differential privacy budget ε over infinite iterations simultaneously under decreasing, constant and increasing privacy noises. For example, let p_α=0.987, p_β=0.69, $p_\gamma=0.997$, p_m=2. Then, conditions in Corollary 1(i) are satisfied as long as $-0.3<p_\zeta<0.15$, $-0.3<p_\eta<0.15$. *Scheme (S2)* achieves the exponential mean square convergence rate and the finite cumulative differential privacy budget ε over infinite iterations simultaneously under decreasing privacy noises. For example, let α=0.1, β=0.1, γ=0.01, p_m=1.1. Then, conditions in Corollary 1(ii) are satisfied as long as $0.91<p_\zeta<0.95$, $0.91<p_\eta<0.95$.

Remark 19: Corollary 1 shows the trade-off between privacy and the convergence rate in Algorithm 1. The smaller privacy noise parameters $\sigma_k^{(\zeta)}$, $\sigma_k^{(\eta)}$ are, the faster Algorithm 1 converges, while the larger the cumulative differential privacy budget ε is. Moreover, *Scheme (S1)* achieves the polynomial mean square convergence rate and finite cumulative differential privacy budget ε over infinite iterations under decreasing, constant, and increasing privacy noises, while *Scheme (S2)* achieves the exponential mean square convergence rate and finite cumulative differential privacy budget ε only for decreasing privacy noises. Then, the differential privacy level of *Scheme (S1)* is higher than the one of *Scheme (S2)*, while the convergence rate of *Scheme (S2)* is faster than the one of *Scheme (S1)*.

Remark 20: The parameter a_4 in the sampling number $m_K=\lfloor a_4K^{p_m}\rfloor+1$ affects both convergence rate and the cumulative privacy budget. Since by (98), $\mathbb{E}\|\nabla F(x_{i,K+1})\|^2=O(\frac{a_4+1}{a_4(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}})$ is decreasing with respect to a_4. Then, the larger the parameter a_4 is, the faster the convergence rate is. By Lemma 2, the larger the parameter a_4 is, the smaller the sensitivity Δ_k^q is, and then by Theorem 4, the smaller the cumulative privacy budget ε is.

Based on Corollary 1, we have the following corollary as the sampling number goes to infinity:

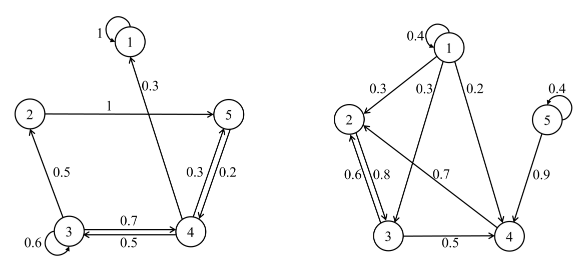
Corollary 2: Under the conditions of Corollary 1, Algorithm 1 with both *Schemes (S1)*, *(S2)* achieves the almost sure and mean square convergence and the finite cumulative differential privacy budget ε over infinite iterations simultaneously as the sampling number goes to infinity.

Remark 21: 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 [36], [40], [41] that compromise convergence accuracy to enable differential privacy, Algorithm 1 compromises the convergence rate and the sampling number (which are also utility metrics) instead. By Corollary 1, the larger privacy noise parameters $\sigma_k^{(\zeta)}$, $\sigma_k^{(\eta)}$ are, the slower the convergence rate is. By Corollary 2, the sampling number m_K is required to go to infinity when the convergence of Algorithm 1 and the finite cumulative privacy budget ε over infinite iterations are considered simultaneously. The ability to retain convergence accuracy makes our approach suitable for accuracy-critical scenarios.

IV. Numerical examples

In this section, we train the machine learning model ResNet18 ([53]) in a distributed manner with the benchmark datasets "MNIST" ([54]) and "CIFAR-10" ([55], [56]), respectively. Specifically, five agents cooperatively train ResNet18 over the directed graphs shown in Figs. 1(a) and 1(b), which satisfy Assumption 1. Then, each benchmark dataset is divided into two subsets for training and testing, respectively. The training dataset of each benchmark dataset is uniformly divided into 5 subsets, each of which can only be accessed by one agent to update its model parameters. The following three numerical experiments are given:

(a) the effect of privacy noises on Algorithm 1's convergence rate and differential privacy level;
(b) the comparison of Algorithm 1 with *Schemes (S1)*, *(S2)* between the convergence rate and the differential privacy level;
(c) the comparison between Algorithm 1 with *Schemes (S1)*, *(S2)* and methods in [36], [39], [40], [42]–[44] for the convergence rate and the differential privacy level.



(a) The directed graph $\mathcal{G}_{R_1}$ (b) The directed graph $\mathcal{G}_{C_1}$

Fig. 1: Topology structures of directed graphs $\mathcal{G}_R, \mathcal{G}_C$ induced by weight matrices R, C

A. Effect of privacy noises

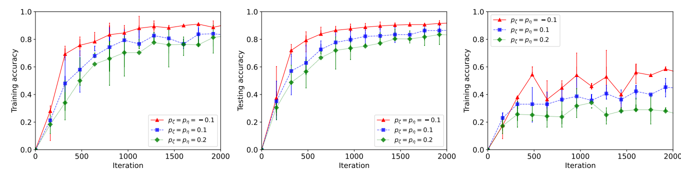
First, let step-sizes $\alpha_K = \frac{72}{2000^{0.987}} = 0.04$, $\beta_K = \frac{0.95}{2000^{0.69}} = 0.005$, $\gamma_K = \frac{98}{2000^{0.997}} = 0.05$, the sampling number $m_K = \lfloor 0.00007 \cdot 2000^{1.78} \rfloor + 1 = 53$, and privacy noise parameters $\sigma_k^{(\zeta)} = (k+1)^{p_\zeta}$, $\sigma_k^{(\eta)} = (k+1)^{p_\eta}$ with $p_\zeta, p_\eta = -0.1, 0.1, 0.2$ respectively in *Scheme (S1)*. Then, the training and testing accuracy on the benchmark datasets "MNIST" and "CIFAR-10" are given in Fig. 2(a)-2(d), from which one can see that the smaller privacy noise parameters $\sigma_k^{(\zeta)}$, $\sigma_k^{(\eta)}$ are, the faster Algorithm 1 converges. This is consistent with the convergence rate analysis in Theorem 2. Meanwhile, the cumulative differential privacy budget ε of Algorithm 1 is given in Fig. 2(e), from which one can see that that the smaller privacy noise parameters $\sigma_k^{(\zeta)}$, $\sigma_k^{(\eta)}$ are, the smaller the cumulative differential privacy budget ε is. This is consistent with the privacy analysis in Theorem 4, and thus consistent with the trade-off between the privacy and the convergence rate in Corollary 1.

Next, let step-sizes $\alpha_K = 0.1$, $\beta_K = 0.01$, $\gamma_K = 0.1$, the sampling number $m_K = \lfloor 1.002^{2000} \rfloor + 1 = 55$, and privacy noise parameters $\sigma_k^{(\zeta)} = {p_\zeta}^{2000}$, $\sigma_k^{(\eta)} = {p_\eta}^{2000}$ with $p_\zeta, p_\eta = 0.9994, 0.9996, 0.9998$ respectively in *Scheme (S2)*. Then, the training and testing accuracy on the benchmark datasets "MNIST" and "CIFAR-10" are given in Fig. 3(a)-3(d), from which one can see that the smaller privacy noise parameters $\sigma_k^{(\zeta)}$, $\sigma_k^{(\eta)}$ are, the faster Algorithm 1 converges. This is consistent with the convergence rate analysis in Theorem 3. Meanwhile, the cumulative differential privacy budget ε of Algorithm 1 is given in Fig. 3(e), from which one can see that that the smaller privacy noise parameters $\sigma_k^{(\zeta)}$, $\sigma_k^{(\eta)}$ are, the smaller the cumulative differential privacy budget ε is. This is consistent with the privacy analysis in Theorem 5, and thus, consistent with the trade-off between the privacy and the convergence rate in Corollary 1.

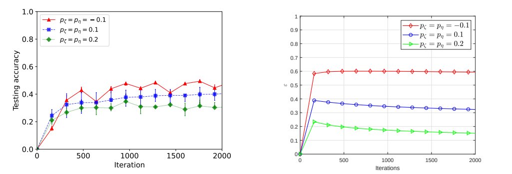
Remark 22: Due to the increasing sample size m_K, the cumulative differential privacy budget ε decreases in the later stages of the iterations in the numerical experiment. In *Scheme (S1)*, the sampling number $m_K = \lfloor 0.00007 \cdot K^{1.78} \rfloor + 1 = O(K^{1.78})$. By Theorem 4, the cumulative differential privacy budget $\varepsilon = O(\frac{\ln(K+2)}{(K+1)^{0.22}})$. Denote the function $\psi_1(t) = \frac{\ln(t+2)}{(t+1)^{0.22}}$. Then, it can be seen that the function $\psi_1(t)$ decreases when t satisfies $t+1 \leq 0.22(t+2)\ln(t+2)$, i.e., $t \geq 87.54$. Thus, the cumulative differential privacy budget ε decreases when the maximum iteration number $K \geq 88$. This result is consistent with Fig. 2(e). Similarly, in *Scheme (S2)*, the sampling number $m_K = \lfloor 1.002^K \rfloor + 1 = O(1.002^K)$. By Theorem 5, the cumulative differential privacy budget $\varepsilon = O(\frac{K}{1.0016^K})$. Denote the function $\psi_2(t) = \frac{t}{1.0016^t}$. Then, it can be seen that the function $\psi_2(t)$ decreases when $t \geq 625.49$. Thus, the cumulative differential privacy budget ε decreases when the maximum iteration number $K \geq 626$. This result is consistent with Fig. 3(e).

B. Comparison between Schemes (S1) and (S2)

In this subsection, the comparison of Algorithm 1 with *Schemes (S1)*, *(S2)* between the convergence rate and the differential privacy level is given. Let $p_\zeta, p_\eta = 0.1$ in *Scheme (S1)*, and $p_\zeta, p_\eta = 0.9996$ in *Scheme (S2)*. Then, from Fig. 4(a)-4(d) one can see that Algorithm 1 with *Scheme (S2)* converges faster than Algorithm 1 with *Scheme (S1)*, while from Fig. 4(e) one can see that the cumulative differential privacy budget ε of Algorithm 1 with *Scheme (S1)* is smaller than the cumulative differential privacy budget ε of Algorithm 1 with *Scheme (S1)*.

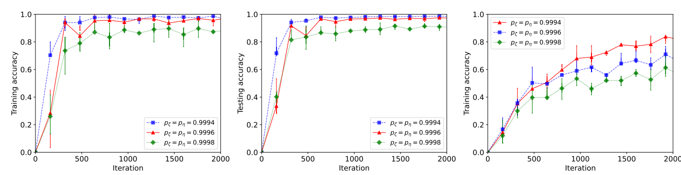


(a) Training accuracy on the "MNIST" dataset (b) Testing accuracy on the "MNIST" dataset (c) Training accuracy on the "CIFAR-10" dataset

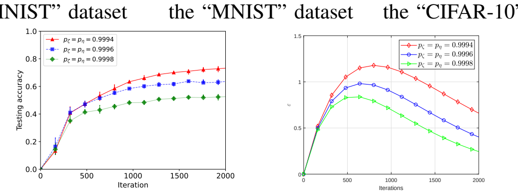


(d) Testing accuracy on the "CIFAR-10" dataset (e) Cumulative differential privacy budget ε

Fig. 2: Accuracy and cumulative differential privacy budget ε of Algorithm 1 with *Scheme (S1)* and $p_\zeta, p_\eta = -0.1, 0.1, 0.2$

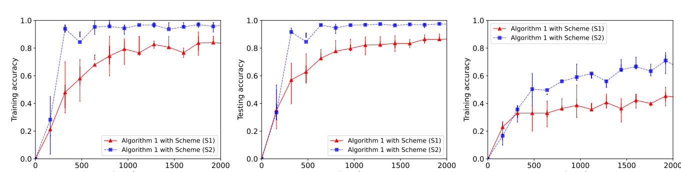


(a) Training accuracy on the "MNIST" dataset (b) Testing accuracy on the "MNIST" dataset (c) Training accuracy on the "CIFAR-10" dataset

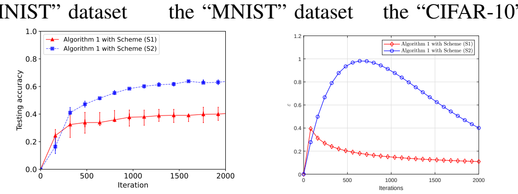


(d) Testing accuracy on the "CIFAR-10" dataset (e) Cumulative differential privacy budget ε

Fig. 3: Accuracy and cumulative differential privacy budget ε of Algorithm 1 with *Scheme (S2)* and $p_\zeta, p_\eta = 0.9994, 0.9996, 0.9998$



(a) Training accuracy on the "MNIST" dataset (b) Testing accuracy on the "MNIST" dataset (c) Training accuracy on the "CIFAR-10" dataset



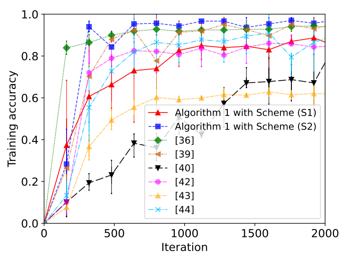
(d) Testing accuracy on the "CIFAR-10" dataset (e) Cumulative differential privacy budget ε

Fig. 4: Comparison of Algorithm 1 with *Schemes (S1)*, *(S2)* on accuracy and cumulative differential privacy budget ε

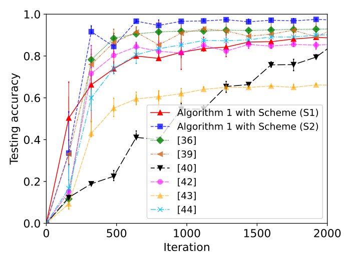
C. Comparison with methods in [36], [39], [40], [42]–[44]

Let $p_\zeta, p_\eta = 0.9996$ in *Scheme (S2)*, and iterations step-sizes $\alpha_K, \beta_K, \gamma_K$, the sampling number m_K, and privacy noise parameters $\sigma_k^{(\zeta)}, \sigma_k^{(\eta)}$ in *Scheme (S1)* and [36], [39], [40], [42]–[44] be the same as *Scheme (S2)* to ensure a fair comparison. Then, the comparison of the convergence rate and the differential privacy level between Algorithm 1 and the methods in [36], [39], [40], [42]–[44] is given in Fig. 5. From Fig. 5, one can see that Algorithm 1 with *Scheme (S2)* converges faster than methods in [36], [39], [40], [42]–[44].

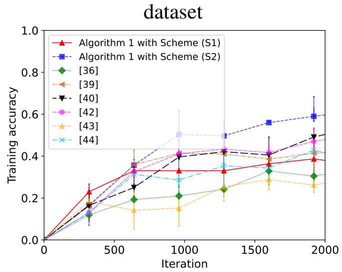
A comparison of the differential privacy level between Algorithm 1 and the methods in [36], [39], [40], [42]–[44] is given in Fig. 6. By Fig. 6(a), the cumulative differential privacy budget ε of Algorithm 1 with both *Schemes (S1)* and *(S2)* is smaller than the ones in [36], [40], [42]–[44]. By Fig. 6(b), [39] achieves the cumulative differential privacy budget $\delta = 1$ after 800 iterations, and thus, the one therein cannot protect sampled gradients after 800 iterations. Thus, Algorithm 1 with both *Schemes (S1)* and *(S2)* provides a higher differential privacy level than methods in [36], [39], [40], [42]–[44].



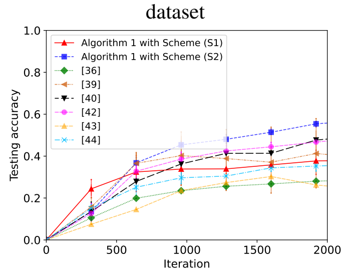
(a) Training accuracy on the “MNIST” dataset



(b) Testing accuracy on the “MNIST” dataset

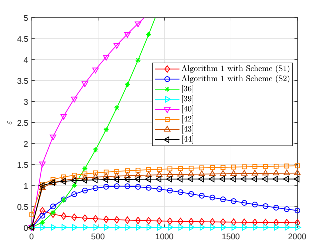


(c) Training accuracy on the “CIFAR-10” dataset

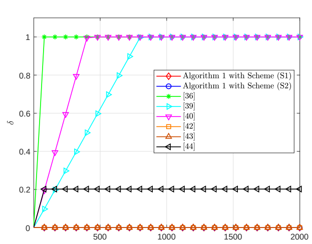


(d) Testing accuracy on the “CIFAR-10” dataset

Fig. 5: Comparison of accuracy on the benchmark datasets “MNIST” and “CIFAR-10”



(a) ε under same $\sigma_k^{(\zeta)}, \sigma_k^{(\eta)}$



(b) δ under same $\sigma_k^{(\zeta)}, \sigma_k^{(\eta)}$

Fig. 6: Comparison of cumulative differential privacy budgets ε and δ

V. Conclusion

In this paper, we have proposed a new differentially private gradient-tracking-based distributed stochastic optimization algorithm over directed graphs. Two novel schemes of step-sizes and the sampling number are given: *Scheme (S1)* uses polynomially decreasing step-sizes and the increasing sampling number with the maximum iteration number. *Scheme (S2)* uses constant step-sizes and the exponentially increasing sampling number with the maximum iteration number. By using the sampling parameter-controlled subsampling method, both schemes achieve the finite cumulative privacy budget even over infinite iterations, and thus, enhance the differential privacy level compared to the existing ones. By using the gradient-tracking method, the almost sure and mean square convergence of the algorithm is shown for nonconvex objectives over directed graphs with spanning trees. Further, when nonconvex objectives satisfy the Polyak-Łojasiewicz condition, the polynomial mean square convergence rate (*Scheme (S1)*) and the exponential mean square convergence rate (*Scheme (S2)*) are given, respectively. Furthermore, the oracle complexity of the algorithm, the trade-off between the privacy and the convergence are shown. Finally, numerical examples of distributed training on the benchmark datasets “MNIST” and “CIFAR-10” are given to show the effectiveness of the algorithm.

Appendix A
Useful Lemmas

For the convenience of the analysis, define

$$
\begin{aligned}
&\nabla f(x_k) = [\nabla f_1(x_{1,k})^\top, \ldots, \nabla f_n(x_{n,k})^\top]^\top,\\
&W_1 = I_n - \frac{1}{n}\mathbf{1}_n v_1^\top, W_2 = I_n - \frac{1}{n}v_2\mathbf{1}_n^\top, \bar{x}_k = \frac{1}{n}(v_1^\top \otimes I_d)x_k,\\
&\bar{y}_k = \frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)y_k, u_k^{(1)} = 2d\rho_{\mathcal{L}_1}^2\alpha_K^2(\sigma_k^{(\zeta)})^2 + \frac{\sigma_g^2}{m_K},\\
&u_k^{(2)} = \frac{(n(2+3r_{\mathcal{L}_2}\beta_K) + 6(1+r_{\mathcal{L}_2}\beta_K)\rho_{W_2}^2\|v_2\|^2\gamma_K^2L^2)\sigma_g^2}{r_{\mathcal{L}_2}m_K\beta_K}\\
&\quad + 2d\rho_{\mathcal{L}_2}^2\beta_K^2(\sigma_k^{(\eta)})^2 + \frac{4d(1+r_{\mathcal{L}_2}\beta_K)\rho_{\mathcal{L}_1}^2\rho_{W_2}^2\alpha_K^2(\sigma_k^{(\zeta)})^2}{r_{\mathcal{L}_2}\beta_K},\\
&u_k^{(3)} = \frac{(v_1^\top v_2)(n + (v_1^\top v_2)^2\gamma_K L)\gamma_K\sigma_g^2}{2n^2 m_K}, u_k = [u_k^{(1)}, u_k^{(2)}, u_k^{(3)}]^\top,\\
&\tilde{s} = [\tilde{s}_1, \tilde{s}_2, \tilde{s}_3]^\top = [\frac{n}{L^2}, \frac{2n^2(v_1^\top v_2)^2}{(2n+3(v_1^\top v_2)\gamma L)\|v_1\|^2}, \frac{2}{\mu}]^\top,\\
&\tilde{\mathbf{s}} = [\tilde{s}_1, \tilde{s}_2]^\top, \mathbf{u}_k = [u_k^{(1)}, u_k^{(2)}]^\top,\\
&A_K^{(11)} = 1 - r_{\mathcal{L}_1}\alpha_K + \frac{4(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2\rho_{W_1}^2\|v_2\|^2L^2}{nr_{\mathcal{L}_1}\alpha_K},\\
&A_K^{(12)} = \frac{2(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2\rho_{W_1}^2}{r_{\mathcal{L}_1}\alpha_K}, A_K^{(13)} = \frac{8(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2\rho_{W_1}^2\|v_2\|^2L}{r_{\mathcal{L}_1}\alpha_K},\\
&A_K^{(21)} = \frac{(1+r_{\mathcal{L}_2}\beta_K)(6\alpha_K^2\rho_{\mathcal{L}_1}^2 + \frac{12\|v_2\|^2\gamma_K^2L^2}{n})\rho_{W_2}^2L^2}{r_{\mathcal{L}_2}\beta_K},\\
&A_K^{(22)} = 1 - r_{\mathcal{L}_2}\beta_K + \frac{6(1+r_{\mathcal{L}_2}\beta_K)\rho_{W_2}^2\gamma_K^2L^2}{r_{\mathcal{L}_2}\beta_K},\\
&A_K^{(23)} = \frac{24(1+r_{\mathcal{L}_2}\beta_K)\rho_{W_2}^2\|v_2\|^2\gamma_K^2L^3}{r_{\mathcal{L}_2}\beta_K},\\
&A_K^{(31)} = \frac{(v_1^\top v_2)\gamma_K L^2(n+3(v_1^\top v_2)\gamma_K L)}{2n^3},\\
&A_K^{(32)} = \frac{(2n+3(v_1^\top v_2)\gamma_K L)\|v_1\|^2\gamma_K}{2n^2(v_1^\top v_2)},
\end{aligned}
$$

$$A_K^{(33)} = 1 - \frac{(v_1^\top v_2)\mu\gamma_K}{n} + \frac{3(v_1^\top v_2)^2\gamma_K^2 L}{2n^2},$$
$$V_k = [\mathbb{E}\|(W_1 \otimes I_d)x_k\|^2, \mathbb{E}\|(W_2 \otimes I_d)y_k\|^2, \mathbb{E}(F(x_k) - F(x^*))]^\top,$$
$$\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2]^\top = [A_K^{(31)}, A_K^{(32)}]^\top, \mathbf{b} = [\mathbf{b}_1, \mathbf{b}_2]^\top = [\frac{A_K^{(13)}}{2L}, \frac{A_K^{(23)}}{2L}]^\top,$$
$$\mathbf{V}_k = [\mathbb{E}\|(W_1 \otimes I_d)x_k\|^2, \mathbb{E}\|(W_2 \otimes I_d)y_k\|^2]^\top,$$
$$A_K = \begin{bmatrix} A_K^{(11)} & A_K^{(12)} & A_K^{(13)} \\ A_K^{(21)} & A_K^{(22)} & A_K^{(23)} \\ A_K^{(31)} & A_K^{(32)} & A_K^{(33)} \end{bmatrix}, \mathbf{M}_K = \begin{bmatrix} A_K^{(11)} & A_K^{(12)} \\ A_K^{(21)} & A_K^{(22)} \end{bmatrix},$$
$$D_K = \begin{bmatrix} A_K^{(11)} & A_K^{(12)} & A_K^{(13)} & 0 \\ A_K^{(21)} & A_K^{(22)} & A_K^{(23)} & 0 \\ A_K^{(31)} & A_K^{(32)} & 1 & -\frac{(v_1^\top v_2)\gamma_K}{2n} + \frac{3(v_1^\top v_2)^2\gamma_K^2 L}{2n^2} \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then, we give the following useful lemmas.

Lemma A.1: ([44, Lemma A.1]) If Assumption 2(i) holds for a function $h : \mathbb{R}^d \to \mathbb{R}$ with a global minimum $h(x^*)$, then following statements holds:
(i) $h(y) \le h(x) + \langle \nabla h(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$, $\forall x, y \in \mathbb{R}^d$.
(ii) $\|\nabla h(x)\|^2 \le 2L(h(x) - h(x^*))$, $\forall x \in \mathbb{R}^d$.

Lemma A.2: ([57, Cor. 8.1.29, Th. 8.4.4]) For any $n = 1, 2, \dots$, let $A \in \mathbb{R}^{n\times n}$ be a nonnegative matrix and $x \in \mathbb{R}^n$ be a positive vector. Then, following statements hold:
(i) If there exists $\rho > 0$ such that $Ax \le \rho x$, then $\rho_A \le \rho$.
(ii) If A is irreducible, then $\rho_A > 0$ and there exists a positive vector $y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$ such that $y^\top A = \rho_A y^\top$.

Lemma A.3: If Assumptions 1 and 2 hold, then the following inequality holds for any $k = 0, \dots, K$, $K = 0, 1, \dots$:

$$\begin{aligned}&\mathbb{E}\|(W_1 \otimes I_d)(x_{k+1} - x_k)\|^2 \\ &\le \frac{3n\rho_{\mathcal{L}_1}^2\alpha_K^2 + 6\rho_{W_1}^2\|v_2\|^2\gamma_K^2 L^2}{n}\mathbb{E}\|(W_1 \otimes I_d)x_k\|^2 \\ &+ 3\rho_{W_1}^2\gamma_K^2(\mathbb{E}\|(W_2 \otimes I_d)y_k\|^2 + 4\|v_2\|^2 L^2\mathbb{E}(F(\bar{x}_k) - F(x^*))) \\ &+ 2d\rho_{\mathcal{L}_1}^2\alpha_K^2(\sigma_k^{(\zeta)})^2 + \frac{3\rho_{W_1}^2\|v_2\|^2\gamma_K^2\sigma_g^2}{m_K}.\end{aligned} \tag{12}$$

Proof. By Assumption 1, Lemma 1 holds. Note that by Lemma 1(i), $\mathcal{L}_1 W_1 = W_1\mathcal{L}_1 = \mathcal{L}_1$. Then, multiplying $W_1 \otimes I_d$ on both sides of (7) implies

$$\begin{aligned}&(W_1 \otimes I_d)x_{k+1} \\ &= ((I_n - \alpha_K\mathcal{L}_1) \otimes I_d)(W_1 \otimes I_d)x_k - \alpha_K(\mathcal{L}_1 \otimes I_d)\zeta_k \\ &\quad - \gamma_K(W_1 \otimes I_d)y_k, \\ &= ((I_n - \alpha_K\mathcal{L}_1) \otimes I_d)(W_1 \otimes I_d)x_k - \alpha_K(\mathcal{L}_1 \otimes I_d)\zeta_k \\ &\quad - \gamma_K(W_1W_2 \otimes I_d)y_k - \frac{\gamma_K}{n}(W_1v_2\mathbf{1}_n^\top \otimes I_d)y_k.\end{aligned} \tag{13}$$

Rearranging (13) gives

$$\begin{aligned}&(W_1 \otimes I_d)(x_{k+1} - x_k) \\ &= -\alpha_K(\mathcal{L}_1W_1 \otimes I_d)x_k - \alpha_K(\mathcal{L}_1 \otimes I_d)\zeta_k \\ &\quad - \gamma_K(W_1W_2 \otimes I_d)y_k - \frac{\gamma_K}{n}(W_1v_2\mathbf{1}_n^\top \otimes I_d)y_k.\end{aligned} \tag{14}$$

Ttaking the mathematical expectation on the squared Euclidean norm of (14) implies

$$\begin{aligned}&\mathbb{E}\|(W_1 \otimes I_d)(x_{k+1} - x_k)\|^2 \\ &= \mathbb{E}\| - \alpha_K(\mathcal{L}_1W_1 \otimes I_d)x_k - \alpha_K(\mathcal{L}_1 \otimes I_d)\zeta_k \\ &\quad - \gamma_K(W_1W_2 \otimes I_d)y_k - \gamma_K(W_1v_2 \otimes I_d)\bar{y}_k\|^2.\end{aligned} \tag{15}$$

For any $k = 0, \dots, K$, let $\mathcal{F}_k = \sigma(\{x_k, y_k\})$. Then, since ζ_k is independent of $\mathcal{F}_k$ and follows the Laplacian distribution $\text{Lap}(\sigma_k^{(\zeta)})$, we have

$$\begin{aligned}&\mathbb{E}(\zeta_k|\mathcal{F}_k) = \mathbb{E}\zeta_k = 0, \\ &\mathbb{E}(\|\zeta_k\|^2|\mathcal{F}_k) = \mathbb{E}\|\zeta_k\|^2 = 2d(\sigma_k^{(\zeta)})^2.\end{aligned} \tag{16}$$

Then by (16), (15) can be rewritten as

$$\begin{aligned}&\mathbb{E}\|(W_1 \otimes I_d)(x_{k+1} - x_k)\|^2 \\ &\le \mathbb{E}\|\alpha_K(\mathcal{L}_1W_1 \otimes I_d)x_k + \gamma_K(W_1W_2 \otimes I_d)y_k \\ &\quad + \gamma_K(W_1v_2 \otimes I_d)\bar{y}_k\|^2 + 2d\rho_{\mathcal{L}_1}^2\alpha_K^2(\sigma_k^{(\zeta)})^2.\end{aligned} \tag{17}$$

Since for any $\mathbf{a}_1$, $\mathbf{a}_2$, $\dots$, $\mathbf{a}_m \in \mathbb{R}^d$, the following inequality holds:

$$\|\sum_{i=1}^m \mathbf{a}_i\|^2 \le m\sum_{i=1}^m \|\mathbf{a}_i\|^2. \tag{18}$$

Setting $m = 3$ in (18) and substituting (18) into (17) implies

$$\begin{aligned}&\mathbb{E}\|(W_1 \otimes I_d)(x_{k+1} - x_k)\|^2 \\ &\le 3\mathbb{E}\|\alpha_K(\mathcal{L}_1W_1 \otimes I_d)x_k\|^2 + 3\mathbb{E}\|\gamma_K(W_1W_2 \otimes I_d)y_k\|^2 \\ &\quad + 3\mathbb{E}\|\gamma_K(W_1v_2 \otimes I_d)\bar{y}_k\|^2 + 2d\rho_{\mathcal{L}_1}^2\alpha_K^2(\sigma_k^{(\zeta)})^2 \\ &\le 3\alpha_K^2\rho_{\mathcal{L}_1}^2\mathbb{E}\|(W_1 \otimes I_d)x_k\|^2 + 3\gamma_K^2\rho_{W_1}^2\mathbb{E}\|(W_2 \otimes I_d)y_k\|^2 \\ &\quad + 3\gamma_K^2\rho_{W_1}^2\|v_2\|^2\mathbb{E}\|\bar{y}_k\|^2 + 2d\rho_{\mathcal{L}_1}^2\alpha_K^2(\sigma_k^{(\zeta)})^2.\end{aligned} \tag{19}$$

By Assumption 2(ii), we have

$$\begin{aligned}&\mathbb{E}(g_k - \nabla f(x_k)) = \mathbb{E}((g_k - \nabla f(x_k))|\mathcal{F}_k) = 0, \\ &\mathbb{E}\|g_k - \nabla f(x_k)\|^2 = \mathbb{E}(\|g_k - \nabla f(x_k)\|^2|\mathcal{F}_k) \le \frac{n\sigma_g^2}{m_K}.\end{aligned} \tag{20}$$

Moreover, note that $\mathbf{1}_n^\top\mathcal{L}_2 = 0$. Then, for any $k = 0, \dots, K$, multiplying $\mathbf{1}_n^\top \otimes I_d$ on both sides of (8) and using $y_0 = g_0$ result in

$$\begin{aligned}\bar{y}_k &= \bar{y}_{k-1} + \frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)(g_k - g_{k-1}) \\ &= \frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)g_0 + \sum_{l=0}^{k-1}\frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)(g_{l+1} - g_l) \\ &= \frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)g_k.\end{aligned} \tag{21}$$

Thus, by (20), taking the mathematical expectation on the squared Euclidean norm of (21) implies

$$\begin{aligned}\mathbb{E}\|\bar{y}_k\|^2 &= \mathbb{E}\|\frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)(g_k - \nabla f(x_k) + \nabla f(x_k))\|^2 \\ &\le \frac{\sigma_g^2}{m_K} + \mathbb{E}\|\frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)\nabla f(x_k)\|^2.\end{aligned} \tag{22}$$

Since $\nabla f(x_k) = (\nabla f(x_k) - \nabla f((\mathbf{1}_n \otimes I_d)\bar{x}_k)) + \nabla f((\mathbf{1}_n \otimes I_d)\bar{x}_k)$, substituting (18) into $\|\frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)\nabla f(x_k)\|^2$ implies

$$\begin{aligned}&\|\frac{1}{n}(\mathbf{1}_n^\top \otimes I_d)\nabla f(x_k)\|^2 \\ &\le 2\|\frac{1}{n}\sum_{i=1}^n(\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k))\|^2 + 2\|\nabla F(\bar{x}_k)\|^2 \\ &\le \frac{2}{n}\sum_{i=1}^n\|\nabla f_i(x_{i,k}) - \nabla f_i(\bar{x}_k)\|^2 + 2\|\nabla F(\bar{x}_k)\|^2.\end{aligned} \tag{23}$$

By Assumption 2(i), it can be seen that

$$\begin{aligned}&\sum_{i=1}^{n}\|\nabla f_i(x_{i,k})-\nabla f_i(\bar{x}_k)\|^2\\ \leq& L^2\sum_{i=1}^{n}\|x_{i,k}-\bar{x}_k\|^2=L^2\|(W_1\otimes I_d)x_k\|^2. \end{aligned}\tag{24}$$

Substituting (23) and (24) into (22) results in

$$\mathbb{E}\|\bar{y}_k\|^2\leq\frac{2L^2}{n}\|(W_1\otimes I_d)x_k\|^2+2\mathbb{E}\|\nabla F(\bar{x}_k)\|^2+\frac{\sigma_g^2}{m_K}.\tag{25}$$

Note that by Lemma A.1(ii), $\|\nabla F(\bar{x}_k)\|^2\leq 2L(F(\bar{x}_k)-F(x^*))$. Then, (25) can be rewritten as

$$\mathbb{E}\|\bar{y}_k\|^2\leq\frac{2L^2}{n}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+4L\mathbb{E}(F(\bar{x}_k)-F(x^*))+\frac{\sigma_g^2}{m_K}.\tag{26}$$

Then, substituting (26) into (19) gives (12). Hence, this lemma is proved. ■

Lemma A.4: If Assumptions 1 and 2 hold, then the following inequality holds for any $k=0,\dots,K$, $K=0,1,\dots$:

$$\begin{aligned}&\mathbb{E}\|\frac{1}{n}(v_1^\top\otimes I_d)y_k\|^2\\ \leq&\frac{3\|v_1\|^2}{n^2}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2+\frac{(v_1^\top v_2)^2}{n^2}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2\\ &+\frac{3(v_1^\top v_2)^2L^2}{n^3}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+\frac{(v_1^\top v_2)^2\sigma_g^2}{n^2m_K}.\end{aligned}\tag{27}$$

Proof. By Assumptions 1, 2, (21) in Lemma A.3 holds. Then by (21), $\frac{1}{n}(v_1^\top\otimes I_d)y_k$ can be rewritten as

$$\begin{aligned}&\frac{1}{n}(v_1^\top\otimes I_d)y_k\\ =&\frac{1}{n}(v_1^\top\otimes I_d)(y_k-(v_2\otimes I_d)\bar{y}_k)+\frac{v_1^\top v_2}{n}\bar{y}_k\\ =&\frac{1}{n}(v_1^\top\otimes I_d)(y_k-(v_2\otimes I_d)\bar{y}_k)+\frac{v_1^\top v_2}{n}(\frac{1}{n}(\mathbf{1}_n^\top\otimes I_d)g_k)\\ =&\frac{1}{n}(v_1^\top\otimes I_d)(W_2\otimes I_d)y_k+\frac{v_1^\top v_2}{n^2}\sum_{i=1}^{n}(g_{i,k}-\nabla f_i(x_{i,k}))\\ &+\frac{v_1^\top v_2}{n^2}\sum_{i=1}^{n}(\nabla f_i(x_{i,k})-\nabla f_i(\bar{x}_k))+\frac{v_1^\top v_2}{n}\nabla F(\bar{x}_k).\end{aligned}\tag{28}$$

Thus, by (20) and (28), we have

$$\begin{aligned}&\mathbb{E}\|\frac{1}{n}(v_1^\top\otimes I_d)y_k\|^2\\ \leq&\mathbb{E}\|\frac{1}{n}(v_1^\top\otimes I_d)(W_2\otimes I_d)y_k+\frac{v_1^\top v_2}{n}\nabla F(\bar{x}_k)\\ &+\frac{v_1^\top v_2}{n^2}\sum_{i=1}^{n}(\nabla f_i(x_{i,k})-\nabla f_i(\bar{x}_k))\|^2+\frac{(v_1^\top v_2)^2\sigma_g^2}{n^2m_K}\end{aligned}\tag{29}$$

Setting $m=3$ in (18) and substituting (18) into (29) imply (27). Thus, this lemma is proved. ■

Lemma A.5: If Assumptions 1, 2, and $\alpha_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-}R_{ij}}\},\frac{1}{r_{\mathcal{L}_1}}\}$, $\beta_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^-}C_{ij}}\},\frac{1}{r_{\mathcal{L}_2}}\}$ hold, then the following inequality holds for any $k=0,\dots,K$:

$$\begin{aligned}&\mathbb{E}\|(W_1\otimes I_d)x_{k+1}\|^2\\ \leq& A_K^{(11)}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+A_K^{(12)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2\\ &+\frac{A_K^{(13)}}{2L}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2+u_k^{(1)}.\end{aligned}\tag{30}$$

Proof. By Assumption 1, (13) in Lemma A.3 holds. Then, taking the mathematical expectation on the squared Euclidean norm of (13) implies

$$\begin{aligned}&\mathbb{E}\|(W_1\otimes I_d)x_{k+1}\|^2\\ =&\mathbb{E}\|\left((I_n-\alpha_K\mathcal{L}_1)\otimes I_d\right)(W_1\otimes I_d)x_k-\alpha_K(\mathcal{L}_1\otimes I_d)\zeta_k\\ &-\gamma_K(W_1W_2\otimes I_d)y_k-\gamma_K(W_1v_2\otimes I_d)\bar{y}_k\|^2.\end{aligned}\tag{31}$$

Then, substituting (16) into (31) implies

$$\begin{aligned}&\mathbb{E}\|(W_1\otimes I_d)x_{k+1}\|^2\\ \leq&\mathbb{E}\left(\|\left((I_n-\alpha_K\mathcal{L}_1)\otimes I_d\right)(W_1\otimes I_d)x_k-\gamma_K(W_1W_2\otimes I_d)y_k\right.\\ &\left.-\gamma_K(W_1v_2\otimes I_d)\bar{y}_k\|^2\right)+2d\rho_{\mathcal{L}_1}^2\alpha_K^2\left(\sigma_k^{(\zeta)}\right)^2.\end{aligned}\tag{32}$$

Note that for any $\mathbf{a},\mathbf{b}\in\mathbb{R}^d, r>0$, the following Cauchy-Schwarz inequality ([58, Ex. 4(b)]) holds:

$$\|\mathbf{a}+\mathbf{b}\|^2\leq(1+r)\|\mathbf{a}\|^2+\left(1+\frac{1}{r}\right)\|\mathbf{b}\|^2.\tag{33}$$

Then, setting $r=r_{\mathcal{L}_1}\alpha_K$ in (33) and substituting (33) into (32) imply

$$\begin{aligned}&\mathbb{E}\|(W_1\otimes I_d)x_{k+1}\|^2\\ \leq&(1+r_{\mathcal{L}_1}\alpha_K)\mathbb{E}\|\left((I_n-\alpha_K\mathcal{L}_1)\otimes I_d\right)(W_1\otimes I_d)x_k\|^2\\ &+\left(1+\frac{1}{r_{\mathcal{L}_1}\alpha_K}\right)\mathbb{E}\|\gamma_K(W_1W_2\otimes I_d)y_k+\gamma_K(W_1v_2\otimes I_d)\bar{y}_k\|^2\\ &+2d\rho_{\mathcal{L}_1}^2\alpha_K^2\left(\sigma_k^{(\zeta)}\right)^2.\end{aligned}\tag{34}$$

By Assumption 1, Lemma 1 holds. By Lemma 1(i), since $v_1^\top\mathbf{1}_n=n$, we have $W_1^2=W_1$. Thus, it can be seen that $((I_n-\alpha_K\mathcal{L}_1)\otimes I_d)(W_1\otimes I_d)x_k=((I_n-\alpha_K\mathcal{L}_1-\frac{1}{n}\mathbf{1}_nv_1^\top)\otimes I_d)(W_1\otimes I_d)x_k$. Since $\alpha_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-}R_{ij}}\},\frac{1}{r_{\mathcal{L}_1}}\}$, by Lemma 1(ii), the spectral radius of $I_n-\alpha_K\mathcal{L}_1-\frac{1}{n}\mathbf{1}_nv_1^\top$ is $1-\alpha_Kr_{\mathcal{L}_1}$. Then, we have

$$\begin{aligned}&(1+r_{\mathcal{L}_1}\alpha_K)\|((I_n-\alpha_K\mathcal{L}_1)\otimes I_d)(W_1\otimes I_d)x_k\|^2\\ \leq&(1+r_{\mathcal{L}_1}\alpha_K)(1-r_{\mathcal{L}_1}\alpha_K)^2\|(W_1\otimes I_d)x_k\|^2\\ \leq&(1-r_{\mathcal{L}_1}\alpha_K)\|(W_1\otimes I_d)x_k\|^2.\end{aligned}\tag{35}$$

Substituting (35) into (34) implies

$$\begin{aligned}&\mathbb{E}\|(W_1\otimes I_d)x_{k+1}\|^2\\ \leq&(1-r_{\mathcal{L}_1}\alpha_K)\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+2d\rho_{\mathcal{L}_1}^2\alpha_K^2\left(\sigma_k^{(\zeta)}\right)^2\\ &+\frac{(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2}{r_{\mathcal{L}_1}\alpha_K}\mathbb{E}(\|(W_1W_2\otimes I_d)y_k+(W_1v_2\otimes I_d)\bar{y}_k\|^2).\end{aligned}\tag{36}$$

Setting $m=2$ in (18) and substituting (18) into (36) imply

$$\begin{aligned}&\mathbb{E}\|(W_1\otimes I_d)x_{k+1}\|^2\\ \leq&(1-r_{\mathcal{L}_1}\alpha_K)\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+2d\rho_{\mathcal{L}_1}^2\alpha_K^2\left(\sigma_k^{(\zeta)}\right)^2\\ &+\frac{2(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2}{r_{\mathcal{L}_1}\alpha_K}\mathbb{E}\|(W_1W_2\otimes I_d)y_k\|^2\\ &+\frac{2(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2}{r_{\mathcal{L}_1}\alpha_K}\mathbb{E}\|(W_1v_2\otimes I_d)\bar{y}_k\|^2\\ \leq&(1-r_{\mathcal{L}_1}\alpha_K)\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+2d\rho_{\mathcal{L}_1}^2\alpha_K^2\left(\sigma_k^{(\zeta)}\right)^2\\ &+\frac{2(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2\rho_{W_1}^2}{r_{\mathcal{L}_1}\alpha_K}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2\\ &+\frac{2(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2\rho_{W_1}^2\|v_2\|^2}{r_{\mathcal{L}_1}\alpha_K}\mathbb{E}\|\bar{y}_k\|^2.\end{aligned}\tag{37}$$

By Assumptions 1 and 2, (25) in Lemma A.3 holds. Then, substituting (25) into (37) implies (30). Thus, this lemma is proved. ■

Lemma A.6: If Assumptions 1, 2, and $\alpha_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^{-}}R_{ij}}\},\frac{1}{r_{\mathcal{L}_1}}\},\beta_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^{-}}C_{ij}}\},\ \frac{1}{r_{\mathcal{L}_2}}\}$ hold, then the following inequality holds for any $k=0,\dots,K$:

$$\begin{aligned}&\mathbb{E}\|(W_2\otimes I_d)y_{k+1}\|^2\\ \le&A_K^{(21)}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+A_K^{(22)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2\\ &+\frac{A_K^{(23)}}{2L}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2+u_k^{(2)}.\end{aligned}\tag{38}$$

Proof. By Assumption 1, Lemma 1 holds. Note that by Lemma 1(i), $\mathcal{L}_2W_2=W_2\mathcal{L}_2=\mathcal{L}_2$. Then, multiplying $W_2\otimes I_d$ on both sides of (8) leads to

$$\begin{aligned}(W_2\otimes I_d)y_{k+1}=&((I_n-\beta_K\mathcal{L}_2)\otimes I_d)(W_2\otimes I_d)y_k\\ &-\beta_K(\mathcal{L}_2\otimes I_d)\eta_k+(W_2\otimes I_d)(g_{k+1}-g_k)\end{aligned}\tag{39}$$

Thus, taking the mathematical expectation on the squared Euclidean norm of (39) implies

$$\begin{aligned}&\mathbb{E}\|(W_2\otimes I_d)y_{k+1}\|^2\\ =&\mathbb{E}\|((I_n-\beta_K\mathcal{L}_2)\otimes I_d)(W_2\otimes I_d)y_k-\beta_K(\mathcal{L}_2\otimes I_d)\eta_k\\ &+(W_2\otimes I_d)(g_{k+1}-g_k)\|^2.\end{aligned}\tag{40}$$

For any $k=0,\dots,K$, let $\mathcal{H}_k=\sigma(\{x_{k+1},y_k\})$. Then, since η_k is independent of $\mathcal{H}_k$ and follows the Laplacian distribution Lap$(\sigma_k^{(\eta)})$, we have

$$\begin{aligned}&\mathbb{E}(\eta_k|\mathcal{H}_k)=\mathbb{E}\eta_k=0,\\ &\mathbb{E}(\|\eta_k\|^2|\mathcal{H}_k)=\mathbb{E}\|\eta_k\|^2=2d(\sigma_k^{(\eta)})^2.\end{aligned}\tag{41}$$

Moreover, since $g_{k+1}-\nabla f(x_{k+1})$ is independent of $\mathcal{H}_k$, by Assumption 2(ii) we have

$$\begin{aligned}&\mathbb{E}(g_{k+1}-\nabla f(x_{k+1}))=\mathbb{E}((g_{k+1}-\nabla f(x_{k+1})|\mathcal{F}_k)=0,\\ &\mathbb{E}\|g_{k+1}-\nabla f(x_{k+1})\|^2=\mathbb{E}(\|g_{k+1}-\nabla f(x_{k+1})\|^2|\mathcal{F}_k)\le\frac{n\sigma_g^2}{m_K}.\end{aligned}\tag{42}$$

Then, substituting (41) and (42) into (40) implies

$$\begin{aligned}&\mathbb{E}\|(W_2\otimes I_d)y_{k+1}\|^2\\ \le&\mathbb{E}\|((I_n-\beta_K\mathcal{L}_2)\otimes I_d)(W_2\otimes I_d)y_k\\ &+(W_2\otimes I_d)(\nabla f(x_{k+1})-\nabla f(x_k)+\nabla f(x_k)-g_k)\|^2\\ &+2d\rho_{\mathcal{L}_2}^2\beta_K^2(\sigma_k^{(\eta)})^2+\frac{n\sigma_g^2}{m_K}.\end{aligned}\tag{43}$$

Then, setting $r=r_{\mathcal{L}_2}\beta_K$ in (33) and substituting (33) into (43) results in

$$\begin{aligned}&\mathbb{E}\|(W_2\otimes I_d)y_{k+1}\|^2\\ \le&(1+r_{\mathcal{L}_2}\beta_K)\mathbb{E}\|((I_n-\beta_K\mathcal{L}_2)\otimes I_d)(W_2\otimes I_d)y_k\|^2\\ &+\left(1+\frac{1}{r_{\mathcal{L}_2}\beta_K}\right)\mathbb{E}\|(W_2\otimes I_d)(\nabla f(x_{k+1})-\nabla f(x_k)\\ &+\nabla f(x_k)-g_k)\|^2+2d\rho_{\mathcal{L}_2}^2\beta_K^2\left(\sigma_k^{(\eta)}\right)^2+\frac{n\sigma_g^2}{m_K}.\end{aligned}\tag{44}$$

Setting $m=2$ in (18) and substituting (18), (20) into (44) implies

$$\begin{aligned}&\mathbb{E}\|(W_2\otimes I_d)y_{k+1}\|^2\\ \le&(1+r_{\mathcal{L}_2}\beta_K)\mathbb{E}\|((I_n-\beta_K\mathcal{L}_2)\otimes I_d)(W_2\otimes I_d)y_k\|^2\\ &+\frac{2(1+r_{\mathcal{L}_2}\beta_K)}{r_{\mathcal{L}_2}\beta_K}\mathbb{E}\|(W_2\otimes I_d)(\nabla f(x_{k+1})-\nabla f(x_k))\|^2\\ &+\frac{n(2+3r_{\mathcal{L}_2}\beta_K)\sigma_g^2}{r_{\mathcal{L}_2}\beta_Km_K}+2d\rho_{\mathcal{L}_2}^2\beta_K^2\left(\sigma_k^{(\eta)}\right)^2\\ \le&(1+r_{\mathcal{L}_2}\beta_K)\mathbb{E}\|((I_n-\beta_K\mathcal{L}_2)\otimes I_d)(W_2\otimes I_d)y_k\|^2\\ &+\frac{2(1+r_{\mathcal{L}_2}\beta_K)\rho_{W_2}^2}{r_{\mathcal{L}_2}\beta_K}\mathbb{E}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\\ &+\frac{n(2+3r_{\mathcal{L}_2}\beta_K)\sigma_g^2}{r_{\mathcal{L}_2}\beta_Km_K}+2d\rho_{\mathcal{L}_2}^2\beta_K^2\left(\sigma_k^{(\eta)}\right)^2\end{aligned}\tag{45}$$

By Assumption 2(i), it can be seen that

$$\begin{aligned}&\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2=\sum_{i=1}^n\|\nabla f_i(x_{i,k+1})-\nabla f_i(x_{i,k})\|^2\\ \le&L^2\sum_{i=1}^n\|x_{i,k+1}-x_{i,k}\|^2=L^2\|x_{k+1}-x_k\|^2.\end{aligned}$$

Thus, we have

$$\mathbb{E}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\le L^2\mathbb{E}\|x_{k+1}-x_k\|^2.\tag{46}$$

By Assumption 1, Lemma 1 holds. Then, rearranging (7) gives

$$\begin{aligned}&x_{k+1}-x_k\\ =&-\alpha_K(\mathcal{L}_1\otimes I_d)(W_1\otimes I_d)x_k-\alpha_K(\mathcal{L}_1\otimes I_d)\zeta_k\\ &-\gamma_K(W_2\otimes I_d)y_k-\gamma_K(v_2\otimes I_d)\bar{y}_k.\end{aligned}\tag{47}$$

Then, taking the mathematical expectation on the squared Euclidean norm of (47), setting m=3 in (18) and substituting (18) into $\mathbb{E}\|x_{k+1}-x_k\|^2$ imply

$$\begin{aligned}&\mathbb{E}\|x_{k+1}-x_k\|^2\\ \le&3\mathbb{E}\|\alpha_K(\mathcal{L}_1\otimes I_d)(W_1\otimes I_d)x_k\|^2+3\mathbb{E}\|\gamma_K(W_2\otimes I_d)y_k\|^2\\ &+3\mathbb{E}\|\gamma_K(v_2\otimes I_d)\bar{y}_k\|^2+2d\rho_{\mathcal{L}_1}^2\alpha_K^2\left(\sigma_k^{(\zeta)}\right)^2\\ \le&3\alpha_K^2\rho_{\mathcal{L}_1}^2\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+3\gamma_K^2\mathbb{E}\|(W_2\otimes I_d)y_k\|^2\\ &+3\|v_2\|^2\gamma_K^2\mathbb{E}\|\bar{y}_k\|^2+2d\rho_{\mathcal{L}_1}^2\alpha_K^2\left(\sigma_k^{(\zeta)}\right)^2.\end{aligned}\tag{48}$$

Substituting (25) and (48) into (46) leads to

$$\begin{aligned}&\mathbb{E}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\\ \le&\left(3\alpha_K^2\rho_{\mathcal{L}_1}^2+\frac{6\|v_2\|^2\gamma_K^2L^2}{n}\right)L^2\mathbb{E}\|(W_1\otimes I_d)x_k\|^2\\ &+3\gamma_K^2L^2\mathbb{E}\|(W_2\otimes I_d)y_k\|^2+6\|v_2\|^2\gamma_K^2L^2\mathbb{E}\|\nabla F(\bar{x}_k)\|^2\\ &+\frac{3\|v_2\|^2\gamma_K^2\sigma_g^2L^2}{m_K}+2d\rho_{\mathcal{L}_1}^2L^2\alpha_K^2\left(\sigma_k^{(\zeta)}\right)^2.\end{aligned}\tag{49}$$

By Lemma 1(i), since $v_2^\top\mathbf{1}_n=n$, we have $W_2^2=W_2$. Thus, it can be seen that $((I_n-\beta_K\mathcal{L}_2)\otimes I_d)(W_2\otimes I_d)y_k=((I_n-\beta_K\mathcal{L}_2-\frac{1}{n}v_2\mathbf{1}_n^\top)\otimes I_d)(W_2\otimes I_d)y_k$. Since $\beta_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^{-}}C_{ij}}\},\frac{1}{r_{\mathcal{L}_2}}\}$, by Lemma 1(ii), the spectral radius of $I_n-\beta_K\mathcal{L}_2-\frac{1}{n}v_2\mathbf{1}_n^\top$ is $1-\beta_Kr_{\mathcal{L}_2}$. Then, we have

$$\begin{aligned}&(1+r_{\mathcal{L}_2}\beta_K)\|((I_n-\beta_K\mathcal{L}_2)\otimes I_d)(W_2\otimes I_d)y_k\|^2\\ \le&(1+r_{\mathcal{L}_2}\beta_K)(1-\beta_Kr_{\mathcal{L}_2})^2\|(W_2\otimes I_d)y_k\|^2\\ \le&(1-r_{\mathcal{L}_2}\beta_K)\|(W_2\otimes I_d)y_k\|^2.\end{aligned}\tag{50}$$

Substituting (49) and (50) into (45) implies (38). Thus, this lemma is proved. ■

Lemma A.7: If Assumptions 1 and 2 hold, then the following inequalities hold for any $K = 0, 1, \ldots$:

$$\begin{aligned}
&\Big(\frac{(v_1^\top v_2)\gamma_K}{2n} - \frac{3(v_1^\top v_2)^2\gamma_K^2 L}{2n^2}\Big)\sum_{k=0}^{K}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2 \\
\leq&\mathbb{E}(F(\bar{x}_0)-F(x^*))+\frac{(K+1)(v_1^\top v_2)(n+(v_1^\top v_2)^2\gamma_K L)\gamma_K\sigma_g^2}{2n^2 m_K} \\
&+\frac{(v_1^\top v_2)\gamma_K L^2(n+3(v_1^\top v_2)\gamma_K L)}{2n^3}\sum_{k=0}^{K}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2 \\
&+\frac{(2n+3(v_1^\top v_2)\gamma_K L)\|v_1\|^2\gamma_K}{2n^2(v_1^\top v_2)}\sum_{k=0}^{K}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2. \qquad (51)
\end{aligned}$$

Proof. By Assumption 1, Lemma 1 holds. Then, multiplying $\frac{1}{n}(v_1^\top \otimes I_d)$ on both sides of (7) results in

$$\bar{x}_{k+1} = \bar{x}_k - \frac{\gamma_K}{n}(v_1^\top \otimes I_d)y_k. \qquad (52)$$

Thus, setting $y = \bar{x}_{k+1}$, $x = \bar{x}_k$ in Lemma A.1(i) and substituting (52) into Lemma A.1(i) gives

$$\begin{aligned}
F(\bar{x}_{k+1}) \leq& F(\bar{x}_k)+\langle\nabla F(\bar{x}_k),\bar{x}_{k+1}-\bar{x}_k\rangle+\frac{L}{2}\|\bar{x}_{k+1}-\bar{x}_k\|^2 \\
=&F(\bar{x}_k)-\gamma_K\langle\nabla F(\bar{x}_k),\frac{1}{n}(v_1^\top\otimes I_d)y_k\rangle \\
&+\frac{\gamma_K^2 L}{2}\|\frac{1}{n}(v_1^\top\otimes I_d)y_k\|^2. \qquad (53)
\end{aligned}$$

Note that $\frac{1}{n}(v_1^\top \otimes I_d)y_k = \frac{(v_1^\top v_2)}{n}\bar{y}_k + \frac{1}{n}((v_1^\top W_2)\otimes I_d)y_k$. Then, $-\gamma_K\langle\nabla F(\bar{x}_k), \frac{1}{n}(v_1^\top \otimes I_d)y_k\rangle = -\frac{(v_1^\top v_2)\gamma_K}{n}\langle\nabla F(\bar{x}_k), \bar{y}_k + \frac{1}{v_1^\top v_2}((v_1^\top W_2)\otimes I_d)y_k\rangle$. Since $\langle\mathbf{a},\mathbf{b}\rangle = \frac{\|\mathbf{a}\|^2+\|\mathbf{b}\|^2-\|\mathbf{a}-\mathbf{b}\|^2}{2}$ for any $\mathbf{a},\mathbf{b}\in\mathbb{R}^d$, it can be seen that

$$\begin{aligned}
&-\gamma_K\langle\nabla F(\bar{x}_k),\frac{1}{n}(v_1^\top\otimes I_d)y_k\rangle \\
=&\frac{(v_1^\top v_2)\gamma_K}{2n}\Big(\|\nabla F(\bar{x}_k)-\frac{1}{v_1^\top v_2}((v_1^\top W_2)\otimes I_d)y_k-\bar{y}_k\|^2 \\
&-\|\nabla F(\bar{x}_k)\|^2-\|\frac{1}{v_1^\top v_2}((v_1^\top W_2)\otimes I_d)y_k+\bar{y}_k\|^2\Big) \\
\leq&\frac{(v_1^\top v_2)\gamma_K}{2n}\|\nabla F(\bar{x}_k)-\frac{1}{v_1^\top v_2}((v_1^\top W_2)\otimes I_d)y_k-\bar{y}_k\|^2 \\
&-\frac{(v_1^\top v_2)\gamma_K}{2n}\|\nabla F(\bar{x}_k)\|^2. \qquad (54)
\end{aligned}$$

By Assumption 1, (21) in Lemma A.3 holds. Then, by (21) we have

$$\begin{aligned}
&\mathbb{E}\|\nabla F(\bar{x}_k)-\frac{1}{v_1^\top v_2}((v_1^\top W_2)\otimes I_d)y_k-\bar{y}_k\|^2 \\
=&\mathbb{E}\|(\nabla F(\bar{x}_k)-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k}))-\frac{1}{v_1^\top v_2}((v_1^\top W_2)\otimes I_d)y_k \\
&+\frac{1}{n}\sum_{i=1}^{n}(\nabla f_i(x_{i,k})-g_{i,k})\|^2. \qquad (55)
\end{aligned}$$

Then, setting $m = 2$ in (18), and substituting (18), (20) into (55) imply

$$\begin{aligned}
&\mathbb{E}\|\nabla F(\bar{x}_k)-\frac{1}{v_1^\top v_2}((v_1^\top W_2)\otimes I_d)y_k-\bar{y}_k\|^2 \\
\leq&\frac{2L^2}{n}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+\frac{2\|v_1\|^2}{n(v_1^\top v_2)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2+\frac{\sigma_g^2}{m_K}. \qquad (56)
\end{aligned}$$

By Assumptions 1 and 2, (27) in Lemma A.4 holds. Thus, substituting (27), (54), and (56) into (53) results in

$$\begin{aligned}
&\mathbb{E}F(\bar{x}_{k+1}) \\
\leq&\mathbb{E}F(\bar{x}_k)+\frac{(v_1^\top v_2)\gamma_K L^2(n+3(v_1^\top v_2)\gamma_K L)}{2n^3}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2 \\
&+\frac{(2n+3(v_1^\top v_2)\gamma_K L)\|v_1\|^2\gamma_K}{2n^2(v_1^\top v_2)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2 \\
&+\Big(-\frac{(v_1^\top v_2)\gamma_K}{2n}+\frac{3(v_1^\top v_2)^2\gamma_K^2 L}{2n^2}\Big)\mathbb{E}\|\nabla F(\bar{x}_k)\|^2 \\
&+\frac{(v_1^\top v_2)(n+(v_1^\top v_2)^2\gamma_K L)\gamma_K\sigma_g^2}{2n^2 m_K}. \qquad (57)
\end{aligned}$$

Rearranging (57) gives

$$\begin{aligned}
&\Big(\frac{(v_1^\top v_2)\gamma_K}{2n}-\frac{3(v_1^\top v_2)^2\gamma_K^2 L}{2n^2}\Big)\mathbb{E}\|\nabla F(\bar{x}_k)\|^2 \\
\leq&\mathbb{E}(F(\bar{x}_k)-F(\bar{x}_{k+1})) \\
&+\frac{(v_1^\top v_2)\gamma_K L^2(n+3(v_1^\top v_2)\gamma_K L)}{2n^3}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2 \\
&+\frac{(2n+3(v_1^\top v_2)\gamma_K L)\|v_1\|^2\gamma_K}{2n^2(v_1^\top v_2)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2 \\
&+\frac{(v_1^\top v_2)(n+(v_1^\top v_2)^2\gamma_K L)\gamma_K\sigma_g^2}{2n^2 m_K}. \qquad (58)
\end{aligned}$$

Summing (58) from 0 to K and using $F(x_{K+1}) \geq F(x^*)$ result in (51). Thus, this lemma is proved. ■

Lemma A.8: If Assumptions 1-3 and $\gamma_K < \frac{n}{4(v_1^\top v_2)L}$ hold, then the following inequality holds for any $k = 0, \ldots, K$:

$$\begin{aligned}
&\mathbb{E}(F(\bar{x}_{k+1})-F(x^*)) \\
\leq&A_K^{(31)}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+A_K^{(32)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2 \\
&+A_K^{(33)}\mathbb{E}(F(\bar{x}_k)-F(x^*))+u_k^{(3)}. \qquad (59)
\end{aligned}$$

Proof. By Assumptions 1 and 2, (57) in Lemma A.7 holds. Moreover, by Assumption 3 and $\gamma_K < \frac{n}{4(v_1^\top v_2)L}$, we have $(1-\frac{(v_1^\top v_2)\gamma_K}{2n}+\frac{3(v_1^\top v_2)^2\gamma_K^2 L}{2n^2})\|\nabla F(\bar{x}_k)\|^2 \leq (1-\frac{(v_1^\top v_2)\mu\gamma_K}{n}+\frac{3(v_1^\top v_2)^2\gamma_K^2 L}{2n^2})(F(\bar{x}_k)-F(x^*))$. Hence, subtracting $F(x^*)$ from both sides of (57) implies (59). Then, the lemma is proved. ■

Lemma A.9: If Assumptions 1-3 and $\alpha_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-}R_{ij}}\},\frac{1}{r_{\mathcal{L}_1}}\}$, $\beta_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^-}C_{ij}}\},\frac{1}{r_{\mathcal{L}_2}}\}$, $\gamma_K < \frac{n}{4(v_1^\top v_2)L}$ hold, then

$$\mathbb{E}V_{k+1} \leq A_K\mathbb{E}V_k + u_k. \qquad (60)$$

Proof. By Assumptions 1, 2 and $\alpha_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-}R_{ij}}\},\frac{1}{r_{\mathcal{L}_1}}\}$, $\beta_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^-}C_{ij}}\},\frac{1}{r_{\mathcal{L}_2}}\}$, Lemmas A.5 and A.6 hold. Thus, by Lemma A.1(ii), (30) and (38) can be rewritten as

$$\begin{aligned}
&\mathbb{E}\|(W_1\otimes I_d)x_{k+1}\|^2 \\
\leq&A_K^{(11)}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+A_K^{(12)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2 \\
&+A_K^{(13)}\mathbb{E}(F(\bar{x}_k)-F(x^*))+u_k^{(1)}, \qquad (61)\\
&\mathbb{E}\|(W_2\otimes I_d)y_{k+1}\|^2 \\
\leq&A_K^{(21)}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2+A_K^{(22)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2 \\
&+A_K^{(23)}\mathbb{E}(F(\bar{x}_k)-F(x^*))+u_k^{(2)}. \qquad (62)
\end{aligned}$$

Moreover, by Assumptions 1-3 and $\gamma_K < \frac{n}{4(v_1^\top v_2)L}$, Lemma A.8 holds. Then, (59) can be rewritten as

$$\begin{aligned}&\mathbb{E}(F(\bar{x}_{k+1}) - F(x^*))\\ \le& A_K^{(31)}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2 + A_K^{(32)}\mathbb{E}\|(W_2\otimes I_d)y_k\|^2\\ &+ A_K^{(33)}\mathbb{E}(F(\bar{x}_k) - F(x^*)) + u_k^{(3)}.\end{aligned} \tag{63}$$

Thus, combining (61)-(63) results in (60). ■

Appendix B
Proof of Theorem 1

We proceed with the following two cases for *Scheme (S1)* and *Scheme (S2)*.

Case 1. If Assumptions 1, 2, 4 holds under *Scheme (S1)*, then the proof of the almost sure and mean square convergence of Algorithm 1 is given in the following four steps:

Step 1. First, we prove that there exists $G_3 > 0$ such that for any $K = 0, 1, \ldots$, $\mathbb{E}(\mathbf{1}_3^\top V_K) \le G_3$. Let $\tilde{v} = [\tilde{v}_1, \tilde{v}_2, \tilde{v}_3]^\top$ be a positive vector. Then, by $p_\beta < p_\alpha < p_\gamma$ in Assumption 4, there exists a positive integer K_0 such that for any $K = K_0, K_0+1, \ldots$, the following inequality holds:

$$[\tilde{v}_1, \tilde{v}_2, \tilde{v}_3, 0]D_K \le (1+\frac{16\rho_{W_1}^2\gamma_K^2\|v_2\|^2L}{r_{\mathcal{L}_1}\alpha_K})[\tilde{v}_1, \tilde{v}_2, \tilde{v}_3, 0]. \tag{64}$$

By Assumptions 1, 2 and $\alpha_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-}R_{ij}}\}, \frac{1}{r_{\mathcal{L}_1}}\}$, $\beta_K<\min\{\min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^-}C_{ij}}\}, \frac{1}{r_{\mathcal{L}_2}}\}$, (57) in Lemma A.7 and (61), (62) in Lemma A.8 hold. Then, by (57), (61), (62), and (64), we have

$$\begin{aligned}\mathbb{E}(\tilde{v}^\top V_{k+1}) \le& [\tilde{v}_1, \tilde{v}_2, \tilde{v}_3, 0]D_K\begin{bmatrix}V_k\\ \mathbb{E}\|\nabla F(\bar{x}_k)\|^2\end{bmatrix} + \tilde{v}^\top u_k\\ \le& (1+\frac{16\rho_{W_1}^2\gamma_K^2\|v_2\|^2L}{r_{\mathcal{L}_1}\alpha_K})\mathbb{E}(\tilde{v}^\top V_k) + \tilde{v}^\top u_k.\end{aligned} \tag{65}$$

Let $\theta=\min\{p_m - p_\beta, 2p_\alpha - 2p_\zeta - p_\beta, 2p_\alpha - p_\beta, 2p_\beta - 2p_\eta, 2p_\beta, 2p_\gamma - p_\beta + p_m\}$. Then, by Assumption 4, $\tilde{v}^\top u_k = O(\frac{1}{(K+1)^\theta})$ holds for any $k = 0, \ldots, K$. Thus, iteratively computing (65) results in

$$\begin{aligned}\mathbb{E}(\tilde{v}^\top V_{K+1}) =& (1+\frac{16\rho_{W_1}^2\gamma_K^2\|v_2\|^2L}{r_{\mathcal{L}_1}\alpha_K})^{K+1}\mathbb{E}(\tilde{v}^\top V_0)\\ &+ O\left(\sum_{k=0}^{K}(1+\frac{16\rho_{W_1}^2\gamma_K^2\|v_2\|^2L}{r_{\mathcal{L}_1}\alpha_K})^k\frac{1}{(K+1)^\theta}\right).\end{aligned} \tag{66}$$

Since $2p_\gamma - p_\alpha \ge 1$ in Assumption 4, $\lim_{K\to\infty}(1+\frac{16\rho_{W_1}^2\gamma_K^2\|v_2\|^2L}{r_{\mathcal{L}_1}\alpha_K})^{K+1} < \infty$. Then, there exists $G_1 > 0$ such that for any $K = 0, 1, \ldots$, $(1+\frac{16\rho_{W_1}^2\gamma_K^2\|v_2\|^2L}{r_{\mathcal{L}_1}\alpha_K})^{K+1} \le G_1$. Thus, (66) can be rewritten as

$$\begin{aligned}\mathbb{E}(\tilde{v}^\top V_{K+1}) \le& G_1\mathbb{E}(\tilde{v}^\top V_0) + O\left(\sum_{k=0}^{K}\frac{1}{(K+1)^\theta}\right)\\ =& G_1\mathbb{E}(\tilde{v}^\top V_0) + O\left(\frac{1}{(K+1)^{\theta-1}}\right).\end{aligned} \tag{67}$$

By $2p_\gamma - p_\alpha \ge 1$, $2p_\alpha - 2p_\zeta - p_\beta \ge 1$, $2p_\alpha - p_\beta \ge 1$, $2p_\beta \ge 1$, $2p_\beta - 2p_\eta \ge 1$, $p_m - p_\beta \ge 1$ in Assumption 4, we have $\theta - 1 \ge 0$. Thus, there exists $G_2 > 0$ such that for any for any $K = K_0, K_0+1, \ldots$, $\mathbb{E}(\tilde{v}^\top V_{K+1}) \le G_2$. Let $G_3=(\tilde{v}_1+\tilde{v}_2+\tilde{v}_3)\max\{\mathbb{E}(\tilde{v}^\top V_0, \mathbb{E}(\tilde{v}^\top V_1)), \ldots, \mathbb{E}(\tilde{v}^\top V_{K_0-1}), G_2\}$. Then, for any $K = 0, 1, \ldots$, $\mathbb{E}(\mathbf{1}_3^\top V_K) \le G_3$ holds.

Step 2: At this step, we prove that for any $i \in \mathcal{V}$, $\lim_{K\to\infty}\|(W_1\otimes I_d)x_{K+1}\|^2=0$ a.s., $\lim_{K\to\infty}\mathbb{E}\|(W_1\otimes I_d)x_{K+1}\|^2=0$. By **Step 1**, there exists $G_3 > 0$ such that for any $K = 0, 1, \ldots$, $\mathbb{E}\|(W_1\otimes I_d)x_K\|^2\le G_3$, $\mathbb{E}\|(W_2\otimes I_d)y_K\|^2 \le G_3$, $\mathbb{E}(F(\bar{x}_K) - F(x^*))\le G_3$. Then, substituting these inequalities into (61) gives

$$\begin{aligned}&\mathbb{E}\|(W_1\otimes I_d)x_{k+1}\|^2\\ \le& (1-r_{\mathcal{L}_1}\alpha_K)\mathbb{E}\|(W_1\otimes I_d)x_k\|^2 + u_k^{(1)}\\ &+ \frac{2(1+r_{\mathcal{L}_1}\alpha_K)\gamma_K^2\rho_{W_1}^2G_3}{r_{\mathcal{L}_1}\alpha_K}(1+4\|v_2\|^2L+\frac{2\|v_2\|^2L^2}{n})\end{aligned} \tag{68}$$

Thus, similar to **Step 4** in the proof of [44, Th. 2], it can be seen that $\lim_{K\to\infty}\mathbb{E}\|(W_1\otimes I_d)x_{K+1}\|^2 = 0$. By [59, Th. 4.2.3], $\|(W_1\otimes I_d)x_{K+1}\|^2$ converges in probability to 0, which means $\lim_{K\to\infty}\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\|^2 \ge \delta_1) = 0$ for any given $\delta_1 > 0$. Hence, by Fatou Lemma ([59, Th. 4.2.2(ii)]), we have

$$\begin{aligned}&\mathbb{P}(\liminf_{K\to\infty}\{\|(W_1\otimes I_d)x_{K+1}\|^2\ge\delta_1\})\le\liminf_{K\to\infty}\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\|^2\ge\delta_1)\\ =& \lim_{K\to\infty}\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\|^2\ge\delta_1) = 0.\end{aligned} \tag{69}$$

Moreover, for any given $\delta_1, \delta_2 > 0$, note that

$$\begin{aligned}&\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\| \ge \delta_1 + \delta_2, \|(W_1\otimes I_d)x_K\| \le \delta_1)\\ \le& \mathbb{P}(\|(W_1\otimes I_d)(x_{K+1} - x_K)\| \ge \delta_2).\end{aligned} \tag{70}$$

Then, by Markov inequality ([59, Th. 4.1.1]) and (7), (70) can be rewritten as

$$\begin{aligned}&\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\| \ge \delta_1 + \delta_2, \|(W_1\otimes I_d)x_K\| \le \delta_1)\\ \le& \frac{\mathbb{E}\|-\alpha_K(\mathcal{L}_1W_1\otimes I_d)x_K-\alpha_K(\mathcal{L}_1\otimes I_d)\zeta_K-\gamma_K(W_1\otimes I_d)y_K\|^2}{\delta_2^2}.\end{aligned}$$

By Lemma A.3, setting $k = K$ in (12) and substituting (12) into the inequality above imply

$$\begin{aligned}&\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\| \ge \delta_1 + \delta_2, \|(W_1\otimes I_d)x_K\| \le \delta_1)\\ \le& \frac{3n\rho_{\mathcal{L}_1}^2\alpha_K^2 + 6\rho_{W_1}^2\|v_2\|^2\gamma_K^2L^2}{n\delta_2^2}\mathbb{E}\|(W_1\otimes I_d)x_K\|^2\\ &\frac{3\rho_{W_1}^2\gamma_K^2}{\delta_2^2}(\mathbb{E}\|(W_2\otimes I_d)y_K\|^2+4\|v_2\|^2L^2\mathbb{E}(F(\bar{x}_K)-F(x^*)))\\ &+ \frac{2d\rho_{\mathcal{L}_1}^2\alpha_K^2(\sigma_k^{(\zeta)})^2}{\delta_2^2} + \frac{3\rho_{W_1}^2\|v_2\|^2\gamma_K^2\sigma_g^2}{\delta_2^2m_K}.\end{aligned} \tag{71}$$

Since $\frac{1}{2}<p_\beta<p_\alpha<p_\gamma<1$, $2p_\alpha-2p_\zeta-p_\beta\ge1$, $2p_\alpha-p_\beta\ge1$ in Assumption 4 and there exists $G_3>0$ such that $\mathbb{E}\|(W_1\otimes I_d)x_K\|^2 \le G_3$, $\mathbb{E}\|(W_2\otimes I_d)y_K\|^2\le G_3$, $\mathbb{E}(F(\bar{x}_K)-F(x^*))\le G_3$, we have $\sum_{K=0}^{\infty}\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\| \ge \delta_1 + \delta_2, \|(W_1\otimes I_d)x_K\| \le \delta_1) < \infty$. By Borel-Cantelli Lemma ([59, Lemma 2.2.2]), $\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\| \ge \delta_1 + \delta_2, \|(W_1\otimes I_d)x_K\| \le \delta_1, \text{ i.o.}) = 0$. Hence, we have

$$\begin{aligned}&\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\| > \delta_1, \|(W_1\otimes I_d)x_K\| \le \delta_1, \text{ i.o.})\\ =& \mathbb{P}(\bigcup_{l=1}^{\infty}\{\|(W_1\otimes I_d)x_{K+1}\| \ge \delta_1+\frac{1}{l}, \|(W_1\otimes I_d)x_K\| < \delta_1, \text{i.o.}\})\\ \le& \sum_{l=1}^{\infty}\mathbb{P}(\|(W_1\otimes I_d)x_{K+1}\| \ge \delta_1+\frac{1}{l}, \|(W_1\otimes I_d)x_K\| < \delta_1, \text{i.o.})\\ =& 0.\end{aligned} \tag{72}$$

By (69), (72), and Barndorff-Nielsen Lemma ([60, Th. 2.1.2]), $\mathbb{P}(\|(W_1\otimes I_d)x_K\|>\delta_1,\ \text{i.o.})=0$. Hence, by [59, Lemma 2.2.1], $\lim_{K\to\infty}\|(W_1\otimes I_d)x_K\|=0$ a.s., and thus, $\lim_{K\to\infty}\|(W_1\otimes I_d)x_{K+1}\|^2=0$ a.s..

Step 3: At this step, we prove that $\lim_{K\to\infty}\|\nabla F(\bar{x}_{K+1})\|^2=0$ a.s., $\lim_{K\to\infty}\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2=0$. Similar to **Step 6** in the proof of [44, Th. 2], it can be seen that $\liminf_{K\to\infty}\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2=0$. Then, by Markov inequality, for any $K=0,1,\ldots$ and $\delta_1>0$ we have

$$\begin{aligned}&\mathbb{P}(\bigcap_{l=K}^{\infty}\{\|\nabla F(\bar{x}_{l+1})\|>\delta_1\})=\mathbb{P}(\inf_{l\geq K}\|\nabla F(\bar{x}_{l+1})\|>\delta_1)\\ \leq&\frac{\mathbb{E}(\inf_{l\geq K}\|\nabla F(\bar{x}_{l+1})\|)^2}{\delta_1^2}=\frac{\mathbb{E}(\inf_{l\geq K}\|\nabla F(\bar{x}_{l+1})\|^2)}{\delta_1^2}\\ \leq&\frac{\inf_{l\geq K}\mathbb{E}\|\nabla F(\bar{x}_{l+1})\|^2}{\delta_1^2}\leq\frac{\liminf_{K\to\infty}\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2}{\delta_1^2}\\ =&0.\end{aligned}\tag{73}$$

Thus, by the definition of the limit inferior and (73), it can be seen that

$$\begin{aligned}&\mathbb{P}(\liminf_{K\to\infty}\{\|\nabla F(\bar{x}_{K+1})\|>\delta_1\})=\mathbb{P}(\bigcup_{K=0}^{\infty}\bigcap_{l=K}^{\infty}\{\|\nabla F(\bar{x}_{l+1})\|>\delta_1\})\\ \leq&\sum_{K=0}^{\infty}\mathbb{P}(\bigcap_{l=K}^{\infty}\{\|\nabla F(\bar{x}_{l+1})\|>\delta_1\})=0.\end{aligned}\tag{74}$$

Moreover, for any given $\delta_1,\delta_2>0$, note that by Assumption 2(i) and (52) we have

$$\begin{aligned}&\mathbb{P}(\|\nabla F(\bar{x}_{K+1})\|\geq\delta_1+\delta_2,\|\nabla F(\bar{x}_K)\|\leq\delta_1)\\ \leq&\mathbb{P}(\|\nabla F(\bar{x}_{K+1})-\nabla F(\bar{x}_K)\|\geq\delta_2)\\ \leq&\mathbb{P}(\gamma_K L\|\frac{1}{n}(v_1^\top\otimes I_d)y_K\|\geq\delta_2).\end{aligned}\tag{75}$$

By Markov inequality and Lemma A.4, setting $k=K$ in (27) and substituting (27) into (75) imply

$$\begin{aligned}&\mathbb{P}(\|\nabla F(\bar{x}_{K+1})\|\geq\delta_1+\delta_2,\|\nabla F(\bar{x}_K)\|\leq\delta_1)\\ \leq&\frac{\gamma_K^2L^2\mathbb{E}\|\frac{1}{n}(v_1^\top\otimes I_d)y_K\|^2}{\delta_2^2}\\ \leq&\frac{3\|v_1\|^2\gamma_K^2L^2}{n^2\delta_2^2}\mathbb{E}\|(W_2\otimes I_d)y_K\|^2+\frac{6(v_1^\top v_2)^2\gamma_K^2L^3}{n^2\delta_2^2}\mathbb{E}(F(\bar{x}_K)-F(x^*))\\ &+\frac{3(v_1^\top v_2)^2\gamma_K^2L^4}{n^3\delta_2^2}\mathbb{E}\|(W_1\otimes I_d)x_K\|^2+\frac{(v_1^\top v_2)^2\gamma_K^2L^2\sigma_g^2}{n^2m_K\delta_2^2}.\end{aligned}\tag{76}$$

Since by **Step 1**, there exists $G_3>0$ such that $\mathbb{E}\|(W_1\otimes I_d)x_K\|^2\leq G_3$, $\mathbb{E}\|(W_2\otimes I_d)y_K\|^2\leq G_3$, $\mathbb{E}(F(\bar{x}_K)-F(x^*))\leq G_3$. Thus, by $p_\gamma>\frac{1}{2}$ in Assumption 4, we have $\sum_{K=0}^{\infty}\mathbb{P}(\|\nabla F(\bar{x}_{K+1})\|\geq\delta_1+\delta_2,\|\nabla F(\bar{x}_K)\|\leq\delta_1)<\infty$. By Borel-Cantelli Lemma, $\mathbb{P}(\|\nabla F(\bar{x}_{K+1})\|\geq\delta_1+\delta_2,\|\nabla F(\bar{x}_K)\|\leq\delta_1,\ \text{i.o.})=0$. Hence, we have

$$\begin{aligned}&\mathbb{P}(\|\nabla F(\bar{x}_{K+1})\|>\delta_1,\|\nabla F(\bar{x}_K)\|\leq\delta_1,\ \text{i.o.})\\ =&\mathbb{P}(\bigcup_{l=1}^{\infty}\{\|\nabla F(\bar{x}_{K+1})\|\geq\delta_1+\frac{1}{l},\|\nabla F(\bar{x}_K)\|<\delta_1,\ \text{i.o.}\})\\ \leq&\sum_{l=1}^{\infty}\mathbb{P}(\|\nabla F(\bar{x}_{K+1})\|\geq\delta_1+\frac{1}{l},\|\nabla F(\bar{x}_K)\|<\delta_1,\ \text{i.o.})\\ =&0.\end{aligned}\tag{77}$$

By (74), (77), and Barndorff-Nielsen Lemma, $\mathbb{P}(\|\nabla F(\bar{x}_K)\|>\delta_1,\ \text{i.o.})=0$. Hence, by [59, Lemma 2.2.1], $\lim_{K\to\infty}\|\nabla F(\bar{x}_K)\|=0$ a.s., and thus, $\lim_{K\to\infty}\|\nabla F(\bar{x}_{K+1})\|^2=0$ a.s..

Furthermore, by **Step 1** and Lemma A.1(ii), $\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2\leq 2LG_3$ for any $K=0,1,\ldots$. Then, by [59, Th. 4.2.1], $\{\|\nabla F(\bar{x}_{K+1})\|^2,K=0,1,\ldots\}$ are uniformly integrable. Thus, by [59, Th. 4.2.3], $\lim_{K\to\infty}\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2=0$.

Step 4: At this step, we prove that $\lim_{K\to\infty}\|\nabla F(x_{i,K+1})\|^2=0$ a.s., $\lim_{K\to\infty}\mathbb{E}\|\nabla F(x_{i,K+1})\|^2=0$ for any $i\in\mathcal{V}$. By Assumption 2(i), the following inequality holds for any $i\in\mathcal{V}$:

$$\begin{aligned}&\|\nabla F(x_{i,K+1})\|^2\\ =&\|\nabla F(\bar{x}_{K+1})+\nabla F(x_{i,K+1})-\nabla F(\bar{x}_{K+1})\|^2\\ \leq&2\|\nabla F(\bar{x}_{K+1})\|^2+2\|\nabla F(x_{i,K+1})-\nabla F(\bar{x}_{K+1})\|^2\\ \leq&2\|\nabla F(\bar{x}_{K+1})\|^2+2L^2\|x_{i,K+1}-\bar{x}_{K+1}\|^2\\ \leq&2\|\nabla F(\bar{x}_{K+1})\|^2+2L^2\|(W_1\otimes I_d)x_{K+1}\|^2.\end{aligned}\tag{78}$$

Then, by **Steps 2** and **3**, we have $\lim_{K\to\infty}\|\nabla F(x_{i,K+1})\|^2=0$ a.s., $\lim_{K\to\infty}\mathbb{E}\|\nabla F(x_{i,K+1})\|^2=0$ for any $i\in\mathcal{V}$. Therefore, the almost sure and mean square convergence of Algorithm 1 with *Scheme (S1)* is proved.

Case 2. If Assumptions 1, 2, 5 holds under *Scheme (S2)*, then the proof of the almost sure and mean square convergence of Algorithm 1 is given in the following three steps:

Step 1: First, we give the upper bound of $\sum_{k=0}^{K+1}\mathbf{V}_k$. Since step-sizes $\alpha_K=\alpha,\beta_K=\beta,\gamma_K=\gamma$ are constants under *Scheme (S2)*, the matrix $\mathbf{M}_K$ is a constant matrix. Then, by Lemmas A.5 and A.6, we have

$$\mathbf{V}_{k+1}\leq\mathbf{M}_K\mathbf{V}_k+\mathbf{b}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2+\mathbf{u}_k.\tag{79}$$

Iteratively computing (79) results in $\mathbf{V}_{k+1}\leq\mathbf{M}_K^{k+1}\mathbf{V}_0+\sum_{l=0}^{k}\mathbf{M}_K^{k-l}(\mathbf{b}\mathbb{E}\|\nabla F(\bar{x}_l)\|^2+\mathbf{u}_l)$. Thus, summing the inequality above from 0 to $K+1$ gives

$$\begin{aligned}\sum_{k=0}^{K+1}\mathbf{V}_k\leq&(\sum_{k=0}^{K+1}\mathbf{M}_K^k)\mathbf{V}_0+\sum_{k=0}^{K}\sum_{l=0}^{k}\mathbf{M}_K^{k-l}(\mathbf{b}\mathbb{E}\|\nabla F(\bar{x}_l)\|^2+\mathbf{u}_l)\\ \leq&(\sum_{k=0}^{\infty}\mathbf{M}_K^k)(\mathbf{V}_0+\sum_{k=0}^{K}(\mathbf{b}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2+\mathbf{u}_k)).\end{aligned}\tag{80}$$

By $\beta<\frac{1}{r_{\mathcal{L}_2}}$, $\alpha<\min\{\frac{1}{r_{\mathcal{L}_1}},\frac{\sqrt{330n}(v_1^\top v_2)r_{\mathcal{L}_2}\beta}{66n\|v_1\|\rho_{W_2}\rho_{\mathcal{L}_1}}\}$, $\gamma<\min\{\frac{n}{15(v_1^\top v_2)L},\frac{\sqrt{3n}r_{\mathcal{L}_1}\alpha}{6\rho_{W_1}\|v_2\|L},\frac{\sqrt{55}(v_1^\top v_2)r_{\mathcal{L}_2}\beta}{33\|v_1\|\|v_2\|\rho_{W_2}L}\}$ in Assumption 5, we have $\mathbf{M}_K\tilde{\mathbf{s}}<\tilde{\mathbf{s}}$. By Lemma A.2(i), $\rho_{\mathbf{M}_K}<1$ holds. Thus, by Gelfand formula ([57, Cor. 5.6.16]), $I_2-\mathbf{M}_K$ is invertible and its inverse matrix is $(I_2-\mathbf{M}_K)^{-1}=\sum_{k=0}^{\infty}\mathbf{M}_K^k$. Hence, (80) can be rewritten as

$$\sum_{k=0}^{K+1}\mathbf{V}_k\leq(I_2-\mathbf{M}_K)^{-1}(\mathbf{V}_0+\sum_{k=0}^{K}(\mathbf{b}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2+\mathbf{u}_k)).\tag{81}$$

Step 2: At this step, we prove that $\lim_{K\to\infty}\|\nabla F(\bar{x}_{K+1})\|^2=0$ a.s., $\lim_{K\to\infty}\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2=0$. By Lemma A.7, substituting (81) into (51) implies

$$\Big(\frac{(v_1^\top v_2)\gamma_K}{2n}-\frac{3(v_1^\top v_2)^2\gamma_K^2L}{2n^2}\Big)\sum_{k=0}^{K}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2$$
$$\leq\mathbb{E}(F(\bar{x}_0)-F(x^*))+\frac{(K{+}1)(v_1^\top v_2)(n+(v_1^\top v_2)^2\gamma_K L)\gamma_K\sigma_g^2}{2n^2m_K}$$
$$+\mathbf{c}^\top\sum_{k=0}^{K}\mathbf{V}_k$$
$$\leq\mathbb{E}(F(\bar{x}_0)-F(x^*))+\frac{(K+1)(n^2+(v_1^\top v_2)^2\gamma L)\gamma\sigma_g^2}{2n^2m_K}$$
$$+\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{V}_0+\sum_{k=0}^{K}\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{u}_k$$
$$+\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{b})\sum_{k=0}^{K}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2. \tag{82}$$

Rearranging (82) gives

$$\Big(\frac{(v_1^\top v_2)\gamma}{2n}-\frac{3(v_1^\top v_2)^2\gamma^2L}{2n^2}-\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{b}\Big)\sum_{k=0}^{K}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2$$
$$\leq\mathbb{E}(F(\bar{x}_0)-F(x^*))+\frac{(K+1)(n^2+(v_1^\top v_2)^2\gamma L)\gamma\sigma_g^2}{2n^2m_K}$$
$$+\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{V}_0+\sum_{k=0}^{K}\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{u}_k. \tag{83}$$

By $\gamma<\min\{1,\frac{\sqrt{n}r_{\mathcal{L}_1}\alpha}{8\rho_{W_1}\|v_2\|L},\frac{\sqrt{2}r_{\mathcal{L}_2}\beta}{6\rho_{W_2}L},\frac{r_{\mathcal{L}_1}r_{\mathcal{L}_2}\beta}{24\rho_{W_1}\rho_{W_2}\rho_{\mathcal{L}_1}L}\}$ in Assumption 5, we have

$$\det(I_2-\mathbf{M}_K)=A_K^{(11)}A_K^{(22)}-A_K^{(12)}A_K^{(21)}>\frac{5}{6}r_{\mathcal{L}_1}r_{\mathcal{L}_2}\alpha\beta. \tag{84}$$

Since $\gamma<\min\{1,\frac{\sqrt{6n}r_{\mathcal{L}_1}\alpha}{24\rho_{W_1}\|v_2\|L},\frac{\sqrt{2n(v_1^\top v_2)^3}r_{\mathcal{L}_2}\beta}{6\rho_{W_2}\|v_1\|\|v_2\|L},\frac{\sqrt{2}(v_1^\top v_2)r_{\mathcal{L}_2}\beta}{6\rho_{W_2}\|v_1\|\|v_2\|L},\frac{(v_1^\top v_2)r_{\mathcal{L}_1}r_{\mathcal{L}_2}\beta}{36\|v_1\|\|v_2\|\rho_{W_2}L}\sqrt{\frac{6}{4\rho_{W_1}^2\rho_{\mathcal{L}_1}^2+r_{\mathcal{L}_1}^2}},\frac{\sqrt{2n(v_1^\top v_2)^3}r_{\mathcal{L}_2}\beta}{6\rho_{W_2}\|v_1\|\|v_2\|L}\}$ in Assumption 5, by (84), we have

$$\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{b}=\frac{1}{\det(I_2-\mathbf{M}_K)}(\mathbf{c}_1\mathbf{b}_1(1-A_K^{(22)})+\mathbf{c}_2\mathbf{b}_1A_K^{(21)}$$
$$+\mathbf{c}_1\mathbf{b}_2A_K^{(12)}+\mathbf{c}_2\mathbf{b}_2(1-A_K^{(11)}))$$
$$<\frac{2(v_1^\top v_2)}{5n}\gamma. \tag{85}$$

By $\gamma<\frac{n}{15(v_1^\top v_2)L}$ in Assumption 5, we have $\frac{3(v_1^\top v_2)^2\gamma^2L}{2n^2}<\frac{(v_1^\top v_2)\gamma}{10n}$. Thus, combining this inequality and (85) leads to $\frac{(v_1^\top v_2)\gamma}{2n}-\frac{3(v_1^\top v_2)^2\gamma^2L}{2n^2}-\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{b}>\frac{(v_1^\top v_2)\gamma}{2n}-\frac{(v_1^\top v_2)\gamma}{10n}-\frac{2(v_1^\top v_2)\gamma}{5n}=0$. Moreover, since $m_K=\lfloor p_m^K\rfloor+1$ and the definition of $\mathbf{u}_k$, there exists $G_4>0$ such that for any $K=0,1,\dots$, $\mathbb{E}(F(\bar{x}_0)-F(x^*))+\frac{(K+1)(v_1^\top v_2)(n+(v_1^\top v_2)^2\gamma L)\gamma\sigma_g^2}{2n^2m_K}+\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{V}_0+\sum_{k=0}^{K}\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{u}_k\leq G_4$. Then, for any $K=0,1,\dots$, by (83) we have

$$\sum_{k=0}^{K}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2\leq\frac{G_4}{\frac{(v_1^\top v_2)\gamma}{2n}-\frac{3(v_1^\top v_2)^2\gamma^2L}{2n^2}-\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{b}}. \tag{86}$$

Since the series in (86) is uniformly bounded for any $K=0,1,\dots$, we have $\lim_{K\to\infty}\mathbb{E}\|\nabla F(\bar{x}_{K+1})\|^2=0$. Then, by the monotone convergence theorem ([59, Th. 4.2.2(i)]), $\mathbb{E}\sum_{K=0}^{\infty}\|\nabla F(\bar{x}_K)\|^2$ converges, which implies $\sum_{K=0}^{\infty}\|\nabla F(\bar{x}_K)\|^2$ converges almost surely. Thus, $\lim_{K\to\infty}\|\nabla F(\bar{x}_{K+1})\|^2=0$, a.s..

Step 3: At this step, we prove that $\lim_{K\to\infty}\|\nabla F(\bar{x}_{i,K+1})\|^2=0$ a.s., $\lim_{K\to\infty}\mathbb{E}\|\nabla F(\bar{x}_{i,K+1})\|^2=0$ for any $i\in\mathcal{V}$. By (81) and (86), the following inequality holds for any $K=0,1,\dots$:

$$\sum_{k=0}^{K+1}\mathbb{E}\|(W_1\otimes I_d)x_k\|^2$$
$$\leq\frac{\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}(\mathbf{V}_0+\sum_{k=0}^{K}(\mathbf{b}\mathbb{E}\|\nabla F(\bar{x}_k)\|^2+\mathbf{u}_k))}{\min\{\mathbf{c}_1,\mathbf{c}_2\}}$$
$$\leq\frac{(\frac{(v_1^\top v_2)\gamma}{2n}-\frac{3(v_1^\top v_2)^2\gamma^2L}{2n^2})G_4}{\min\{\mathbf{c}_1,\mathbf{c}_2\}(\frac{(v_1^\top v_2)\gamma}{2n}-\frac{3(v_1^\top v_2)^2\gamma^2L}{2n^2}-\mathbf{c}^\top(I_2-\mathbf{M}_K)^{-1}\mathbf{b})}. \tag{87}$$

Since the series in (87) is uniformly bounded for any $K=0,1,\dots$, we have $\lim_{K\to\infty}\mathbb{E}\|(W_1\otimes I_d)x_{K+1}\|^2=0$. Then, by the monotone convergence theorem, $\mathbb{E}\sum_{K=0}^{\infty}\|(W_1\otimes I_d)x_K\|^2$ converges, which implies $\sum_{K=0}^{\infty}\|(W_1\otimes I_d)x_K\|^2$ converges almost surely. Hence, $\lim_{K\to\infty}\|(W_1\otimes I_d)x_{K+1}\|^2=0$, a.s.. Therefore, by (78), the almost sure and mean square convergence of Algorithm 1 with *Scheme (S2)* is proved. ■

Appendix C
Proof of Theorem 2

Let $0<\Gamma<1$ and $\omega_K=\Gamma\min\{r_{\mathcal{L}_1}\alpha_K,r_{\mathcal{L}_2}\beta_K,\frac{(v_1^\top v_2)\mu\gamma_K}{n}\}$. Then, the following four steps are given to prove Theorem 1.

Step 1: First, we prove that there exists a positive integer K_0 such that for any $K=K_0,K_0+1,\dots$,

$$\rho_{A_K}\leq 1-\omega_K. \tag{88}$$

Since $2p_\alpha-2p_\zeta-p_\beta\geq 1$, $\frac{1}{2}<p_\beta<p_\alpha<p_\gamma<1$, and $2p_\gamma-p_\alpha\geq 1$ in Assumption 4, there exists a positive vector $\tilde{u}\in\mathbb{R}^3$ and a positive integer K_0 such that for any $K=K_0,K_0+1,\dots$, the following inequality holds:

$$A_K\tilde{u}\leq(1-\omega_K)\tilde{u}. \tag{89}$$

Then, by (89) and Lemma A.2(i), (88) holds for any $K=K_0,K_0+1,\dots$.

Step 2: At this step, we prove that there exists a positive vector $\tilde{t}=[\tilde{t}_1,\tilde{t}_2,\tilde{t}_3]^\top$ such that for any $K=0,1,\dots$, $\mathbb{E}(\tilde{t}^\top V_{K+1})=O(\frac{1}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}})$. Note that for any $K=K_0,K_0+1,\dots$, (88) holds. Then, by Lemma A.2(ii), there exists a positive vector $\tilde{t}=[\tilde{t}_1,\tilde{t}_2,\tilde{t}_3]^\top$ such that $\tilde{t}^\top A_K=\rho_{A_K}\tilde{t}^\top\leq(1-\omega_K)\tilde{t}^\top$. Moreover, by Assumptions 1-4, (60) in Lemma A.8 holds. Then, multiplying $\tilde{t}^\top$ on both sides of (60) implies

$$\mathbb{E}(\tilde{t}^\top V_{k+1})\leq\tilde{t}^\top A_K\mathbb{E}V_k+\tilde{t}^\top u_k$$
$$\leq(1-\omega_K)\mathbb{E}(\tilde{t}^\top V_k)+\tilde{t}^\top u_k. \tag{90}$$

By Assumption 4, $\tilde{t}^\top u_k=O(\frac{a_4+1}{a_4(K+1)^\theta})$ holds for any $k=0,\dots,K$. Thus, iteratively computing (90) results in

$$\mathbb{E}(\tilde{t}^\top V_{K+1})=(1-\omega_K)^{K+1}\mathbb{E}(\tilde{t}^\top V_0)+O\left(\sum_{k=0}^{K}(1-\omega_K)^k\frac{a_4+1}{a_4(K+1)^\theta}\right)$$
$$=(1-\omega_K)^{K+1}\mathbb{E}(\tilde{t}^\top V_0)+O\left(\frac{a_4+1}{a_4\omega_K(K+1)^\theta}\right). \tag{91}$$

By the definition of ω_K, it can be seen that

$$O\left(\frac{1}{\omega_K(K+1)^{\theta}}\right) = O\left(\frac{1}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}}\right),$$
$$(1-\omega_K)^{K+1} = \exp\left((K+1)\ln(1-\omega_K)\right)$$
$$\le \exp\left(-(K+1)\omega_K\right) = \exp\left(-O\left((K+1)^{1-\max\{p_\alpha,p_\beta,p_\gamma\}}\right)\right)$$
$$= o\left(\frac{1}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}}\right). \tag{92}$$

By (92), we have $\mathbb{E}(\tilde{t}^\top V_{K+1})=O(\frac{1}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}})$ for any $K=K_0, K_0+1, \dots$. Thus, there exists $S_0>0$ such that $\mathbb{E}(\tilde{t}^\top V_{K+1})\le\frac{S_0}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}}$. Let $S=\max\{\mathbb{E}(\tilde{t}^\top V_1), 2^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}\mathbb{E}(\tilde{t}^\top V_2),\dots,(K_0-1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}\mathbb{E}(\tilde{t}^\top V_{K_0-1}),S_0\}$. Then, for any $K=0,1,\dots,$ we have $\mathbb{E}(\tilde{t}^\top V_{K+1})\le\frac{S}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}}$, which leads to

$$\mathbb{E}(\tilde{t}^\top V_{K+1}) = O\left(\frac{a_4+1}{a_4(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}}\right). \tag{93}$$

Step 3: At this step, we prove that for any $i\in\mathcal{V}$ and $K=0,1,\dots,$ $\mathbb{E}\|\nabla F(x_{i,K+1})\|^2=O(\frac{1}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}})$. By Lemma A.1(i), we have

$$\begin{aligned}&F(x_{i,K+1})-F(\bar{x}_{K+1})\\&\le\langle\nabla F(\bar{x}_{K+1}),x_{i,K+1}-\bar{x}_{K+1}\rangle+\frac{L}{2}\|\bar{x}_{K+1}-x_{i,K+1}\|^2.\end{aligned} \tag{94}$$

Note that $\langle\mathbf{a},\mathbf{b}\rangle\le\frac{\|\mathbf{a}\|^2+\|\mathbf{b}\|^2}{2}$ for any $\mathbf{a},\mathbf{b}\in\mathbb{R}^d$. Then, (94) can be rewritten as

$$\begin{aligned}&F(x_{i,K+1})-F(\bar{x}_{K+1})\\&\le\frac{\|\nabla F(\bar{x}_{K+1})\|^2+\|\bar{x}_{K+1}-x_{i,K+1}\|^2}{2}+\frac{L}{2}\|\bar{x}_{K+1}-x_{i,K+1}\|^2\\&=\frac{L+1}{2}\|\bar{x}_{K+1}-x_{i,K+1}\|^2+\frac{\|\nabla F(\bar{x}_{K+1})\|^2}{2}.\end{aligned} \tag{95}$$

By Lemma A.1(ii), $\|\nabla F(\bar{x}_{K+1})\|^2\le2L(F(\bar{x}_{K+1})-F(x^*))$. Substituting it into (95) gives $F(x_{i,K+1})-F(\bar{x}_{K+1})\le\frac{L+1}{2}\|\bar{x}_{K+1}-x_{i,K+1}\|^2+L(F(\bar{x}_{K+1})-F(x^*))$. Thus, we have

$$\begin{aligned}&F(x_{i,K+1})-F(\bar{x}_{K+1})\\&\le\frac{L+1}{2}\sum_{i=1}^{n}\|\bar{x}_{K+1}-x_{i,K+1}\|^2+L(F(\bar{x}_{K+1})-F(x^*))\\&=\frac{L+1}{2}\|(W_1\otimes I_d)x_{K+1}\|^2+L(F(\bar{x}_{K+1})-F(x^*)).\end{aligned} \tag{96}$$

Then, by (96) it can be seen that

$$\begin{aligned}&F(x_{i,K+1})-F(x^*)\\&=(F(x_{i,K+1})-F(\bar{x}_{K+1}))+(F(\bar{x}_{K+1})-F(x^*))\\&\le\frac{L+1}{2}\|(W_1\otimes I_d)x_{K+1}\|^2+(L+1)(F(\bar{x}_{K+1})-F(x^*))\\&\le(L+1)\left(\mathbf{1}_3^\top\mathbb{E}V_{K+1}\right)=O\left(\mathbb{E}(\tilde{t}^\top V_{K+1})\right).\end{aligned} \tag{97}$$

Thus, combining (93) and (97) gives $\mathbb{E}(F(x_{i,K+1})-F(x^*))=O\left(\frac{a_4+1}{a_4(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}}\right)$. By Lemma A.1(ii), we have

$$\begin{aligned}\mathbb{E}\|\nabla F(x_{i,K+1})\|^2&\le2L\mathbb{E}(F(x_{i,K+1})-F(x^*))\\&=O(\frac{a_4+1}{a_4(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}})\\&=O(\frac{1}{(K+1)^{\theta-\max\{p_\alpha,p_\beta,p_\gamma\}}}).\end{aligned} \tag{98}$$

Hence, the polynomial mean square convergence rate is achieved.

Step 4: At this step, we prove that the oracle complexity of Algorithm 1 with *Scheme (S1)* is $O(\varphi^{-\frac{3+3\varphi}{1-3\varphi}})$ for any $0<\varphi<\frac{1}{3}$. When $p_\alpha=1-\frac{2\varphi}{11}$, $p_\beta=\frac{2}{3}(1+\frac{3\varphi}{11})$, $p_\gamma=1-\frac{\varphi}{11}$, $p_m=\varphi$, $p_\zeta=p_\eta=\frac{2\varphi}{11}$ for any $0<\varphi<\frac{1}{3}$, by Step 3, $\mathbb{E}\|\nabla F(x_{i,K+1})\|^2=O(\frac{1}{(K+1)^{\frac{1}{3}-\varphi}})$ holds for any $i\in\mathcal{V}$ and $K=0,1,\dots$. Thus, there exists $\Phi>0$ such that the following inequality holds:

$$\mathbb{E}\|\nabla F(x_{i,K+1})\|^2\le\frac{\Phi}{(K+1)^{\frac{1}{3}-\varphi}}. \tag{99}$$

Let $K=\lfloor(\frac{\Phi}{\varphi})^{\frac{3}{1-3\varphi}}\rfloor$. Then, by (99) we have

$$\mathbb{E}\|\nabla F(x_{i,K+1})\|^2\le\frac{\Phi}{(K+1)^{\frac{1}{3}-\varphi}}<\frac{\Phi}{(\frac{\Phi}{\varphi})^{(\frac{1}{3}-\varphi)\frac{3}{1-3\varphi}}}=\varphi. \tag{100}$$

Thus, by (100) and Definition 1, x_{K+1} is a φ-suboptimal solution. Since $N(\varphi)$ is the smallest integer such that $x_{N(\varphi)}$ is a φ-suboptimal solution, we have

$$N(\varphi)\le\lfloor(\frac{\Phi}{\varphi})^{\frac{3}{1-3\varphi}}\rfloor+1. \tag{101}$$

Since $m_K=\lfloor a_4K^\varphi\rfloor+1=\lfloor a_4\lfloor(\frac{\Phi}{\varphi})^{\frac{3}{1-3\varphi}}\rfloor^\varphi\rfloor+1$, by Definition 2 and (101), the oracle complexity of Algorithm 1 with *Scheme (S1)* is given as follows:

$$\begin{aligned}\sum_{k=0}^{N(\varphi)}m_K&=(N(\varphi)+1)(\lfloor a_4\lfloor(\frac{\Phi}{\varphi})^{\frac{3}{1-3\varphi}}\rfloor^\varphi\rfloor+1)\\&\le(\lfloor(\frac{\Phi}{\varphi})^{\frac{3}{1-3\varphi}}\rfloor+2)(a_4\lfloor(\frac{\Phi}{\varphi})^{\frac{3}{1-3\varphi}}\rfloor+1)\\&=O(\varphi^{-\frac{3+3\varphi}{1-3\varphi}}).\end{aligned}$$

Therefore, this theorem is proved. ■

APPENDIX D
PROOF OF THEOREM 3

The following two steps are given to prove Theorem 2.

Step 1: First, we prove that Algorithm 1 with *Scheme (S2)* achieves the exponential mean square convergence rate. By $\beta<\frac{1}{r_{\mathcal{L}_2}}$, $\alpha<\min\{\frac{1}{r_{\mathcal{L}_1}},\frac{\sqrt{n}r_{\mathcal{L}_2}}{6\rho_{W_1}\rho_{\mathcal{L}_1}\|v_1\|}\beta\}$, $\gamma<\min\{\frac{n}{15(v_1^\top v_2)L},Q_1\alpha,Q_2\beta\}$ in Assumption 5, we have $A_K\tilde{s}<\tilde{s}$. By Lemma A.2(i), we have $\rho_{A_K}<1$. Thus, by Lemma A.2(ii), there exists a positive vector $\tilde{\mathbf{r}}=[\tilde{\mathbf{r}}_1,\tilde{\mathbf{r}}_2,\tilde{\mathbf{r}}_3]^\top$ such that $\tilde{\mathbf{r}}^\top A_K=\rho_{A_K}\tilde{\mathbf{r}}^\top$. Moreover, by Assumptions 1-3, 5, (60) in Lemma A.8 holds. Then, multiplying $\tilde{\mathbf{r}}^\top$ on both sides of (60) implies that for any $k=0,\dots,K$,

$$\mathbb{E}(\tilde{\mathbf{r}}^\top V_{k+1})\le\tilde{\mathbf{r}}^\top A_K\mathbb{E}V_k+\tilde{\mathbf{r}}^\top u_k=\rho_{A_K}\mathbb{E}(\tilde{\mathbf{r}}^\top V_k)+\tilde{\mathbf{r}}^\top u_k. \tag{102}$$

Iteratively computing (102) gives

$$\mathbb{E}\tilde{\mathbf{r}}^\top V_{K+1}\le\rho_{A_K}^{K+1}\mathbb{E}\tilde{\mathbf{r}}^\top V_K+\sum_{k=0}^{K}\rho_{A_K}^{K-k}\tilde{\mathbf{r}}^\top u_k. \tag{103}$$

By Assumption 5, $\tilde{\mathbf{r}}^\top u_k=O(p_m^{-K}+p_\zeta^{2K}+p_\eta^{2K})$ for any $k=0,\dots,K$. Then, (103) can be rewritten as

$$\begin{aligned}\mathbb{E}\tilde{\mathbf{r}}^\top V_{K+1}&=\rho_{A_K}^{K+1}\mathbb{E}\tilde{\mathbf{r}}^\top V_0+O(\max\{\rho_{A_K},\frac{1}{p_m}\}^K)\\&\quad+O(\max\{\rho_{A_K},p_\zeta^2\}^K)+O(\max\{\rho_{A_K},p_\eta^2\}^K)\\&=O(\max\{\rho_{A_K},\frac{1}{p_m},p_\zeta^2,p_\eta^2\}^K).\end{aligned} \tag{104}$$

As shown in Step 3 of Appendix C, $F(x_{i,K+1})-F(x^*)=O\left(\mathbb{E}(\tilde{\mathbf{r}}^\top V_{K+1})\right)$. Hence, by (104) and Lemma A.1(ii), the exponential mean convergence rate of Algorithm 1 is achieved.

Step 2: Next, we prove that the oracle complexity of Algorithm 1 with *Scheme (S2)* is $O(\frac{1}{\varphi}\ln\frac{1}{\varphi})$. For any $0<\varphi<$

$\min\{1, \frac{n^2}{15(v_1^\top v_2)^2 L}, \min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{R,i}^-} R_{ij}}\}, \min_{i\in\mathcal{V}}\{\frac{1}{\sum_{j\in\mathcal{N}_{C,i}^-} C_{ij}}\}, \frac{1}{r_{\mathcal{L}_1}}, \frac{1}{r_{\mathcal{L}_2}}\}$, let $\beta = \varphi$, $\alpha=\min\{\varphi, \frac{\sqrt{330n}(v_1^\top v_2) r_{\mathcal{L}_2}\varphi}{132n\|v_1\|\rho_{W_2}\rho_{\mathcal{L}_1}}\}$, $\gamma=\min\{\frac{1}{2}, \frac{n}{30(v_1^\top v_2)L}, \frac{Q_1\alpha}{2}, \frac{Q_2\varphi}{2}\}$, $p_m=\min\{\frac{1}{\varphi}, \frac{1}{\rho_{A_K}}\}$, $p_\zeta= p_\eta=\varphi$. Then, by Theorem 2, there exists $\Phi > 0$ such that for any $i \in \mathcal{V}$, $K = 0, 1, \ldots,$

$$\mathbb{E}\|\nabla F(x_{i,K+1})\|^2 \leq \Phi \max\{\rho_{A_K}, \varphi\}^K. \tag{105}$$

Let $K=\lfloor \max\{\frac{\ln\varphi-\ln\Phi}{\ln\varphi}, \frac{\ln\varphi-\ln\Phi}{\ln\Phi}\}\rfloor+1$. Then, by (105) we have $\mathbb{E}\|\nabla F(x_{i,K+1})\|^2 < \varphi$. By Definition 1, x_{K+1} is a φ-suboptimal solution. Thus, by the definition of $N(\varphi)$, we have

$$N(\varphi) \leq \lfloor \max\{\frac{\ln\varphi-\ln\Phi}{\ln\varphi}, \frac{\ln\varphi-\ln\Phi}{\ln\Phi}\}\rfloor+2. \tag{106}$$

Since $m_K = \lfloor \min\{\frac{1}{\varphi}, \frac{1}{\rho_{A_K}}\}^K\rfloor+1$. Thus, by Definition 2 and (106), the oracle complexity of Algorithm 1 with *Scheme (S2)* is given as follows:

$$\begin{aligned}
&\sum_{k=0}^{N(\varphi)} m_K \\
=&(N(\varphi)+1)(\lfloor \min\{\frac{1}{\varphi}, \frac{1}{\rho_{A_K}}\}^{\lfloor\max\{\frac{\ln\varphi-\ln\Phi}{\ln\varphi}, \frac{\ln\varphi-\ln\Phi}{\ln\Phi}\}\rfloor+1}\rfloor+1) \\
\leq&(\lfloor \max\{\frac{\ln\varphi-\ln\Phi}{\ln\varphi}, \frac{\ln\varphi-\ln\Phi}{\ln\Phi}\}\rfloor+3)\cdot \\
&\quad(\min\{\frac{1}{\varphi}, \frac{1}{\rho_{A_K}}\}^{\max\{\frac{\ln\varphi-\ln\Phi}{\ln\varphi}, \frac{\ln\varphi-\ln\Phi}{\ln\Phi}\}+1}+1) \\
=&O(\frac{1}{\varphi}\ln\frac{1}{\varphi}).
\end{aligned}$$

Therefore, this theorem is proved. ■

Appendix E
Proof of Lemma 2

The following two steps are given to prove Lemma 2.

Step 1: We compute $\|\Delta y_k\|_1$ for any $k = 0, \ldots, K$. When $k = 0$, by Definition 5, we have

$$\|\Delta y_0\|_1 = \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|y_0 - y_0'\|_1 = \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|g_0 - g_0'\|_1. \tag{107}$$

Since $\mathcal{D}, \mathcal{D}'$ are adjacent, by Definition 3, there exists exactly one pair of data samples $\xi_{i_0,l_0}, \xi'_{i_0,l_0}$ such that (3) holds. This implies that $g_{j,k} = g'_{j,k}$ holds for any agent $j \neq i_0$ and $k = 0, \ldots, K$. Thus, (107) can be rewritten as

$$\|\Delta y_0\|_1 = \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|g_{i_0,0} - g'_{i_0,0}\|_1. \tag{108}$$

Note that m_K different data samples are taken uniformly from $\mathcal{D}$, $\mathcal{D}'$, respectively. Then, there exists at most one pair of data samples $\lambda_{i_0,0,l_1}, \lambda'_{i_0,0,l_1}$ such that $\lambda_{i_0,0,l_1} = \xi_{i_0,l_0}$, $\lambda'_{i_0,0,l_1} = \xi'_{i_0,l_0}$. Thus, by (108) and (5) we have

$$\begin{aligned}
&\|\Delta y_0\|_1 \\
=& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \left\| \frac{1}{m_K}\sum_{l=1}^{m_K} (g_{i_0}(x_{i_0,0}, \lambda_{i_0,0,l}) - g_{i_0}(x_{i_0,0}, \lambda'_{i_0,0,l})) \right\|_1 \\
=& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \left\| \frac{1}{m_K} (g_{i_0}(x_{i_0,0}, \lambda_{i_0,0,l_1}) - g_{i_0}(x_{i_0,0}, \lambda'_{i_0,0,l_1})) \right\|_1 \\
\leq& \frac{1}{m_K} \left\| g_{i_0}(x_{i_0,0}, \xi_{i_0,l_0}) - g_{i_0}(x_{i_0,0}, \xi'_{i_0,l_0}) \right\|_1 \leq \frac{C}{m_K}.
\end{aligned} \tag{109}$$

When $k = 1$, by Definition 5, we have

$$\begin{aligned}
\|\Delta y_1\|_1 =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|y_1 - y_1'\|_1 \\
=& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \sum_{i=1}^n \|(1-\beta_K \sum_{j\in\mathcal{N}_{C,i}^-} C_{ij})(y_{i,0} - y'_{i,0}) \\
& - \beta_K \sum_{j\in\mathcal{N}_{C,i}^-} C_{ij}(\breve{y}_{j,0} - \breve{y}'_{j,0}) \\
& + (g_{i,1} - g'_{i,1}) + (g_{i,0} - g'_{i,0})\|_1.
\end{aligned} \tag{110}$$

Note that the sensitivity is obtained by computing the maximum magnitude of the mapping q when changing one data sample. Then, observations $(x_0, y_0, \ldots, x_K, y_K)$, $(x_0', y_0', \ldots, x_K', y_K')$ 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. Thus, $\breve{x}_{j,k} = \breve{x}'_{j,k}$, $\breve{y}_{j,k} = \breve{y}'_{j,k}$ holds for any agent $j \in \mathcal{N}_{R,i}^- \cup \mathcal{N}_{C,i}^-$ and $k = 0, \ldots, K$. Then, (110) can be rewritten as

$$\begin{aligned}
\|\Delta y_1\|_1 =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|y_1 - y_1'\|_1 \\
=& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \sum_{i=1}^n \|(1-\beta_K \sum_{j\in\mathcal{N}_{C,i}^-} C_{ij})(y_{i,0} - y'_{i,0}) \\
& + (g_{i,1} - g'_{i,1}) + (g_{i,0} - g'_{i,0})\|_1.
\end{aligned} \tag{111}$$

Since $y_{j,0} = y'_{j,0}$, $g_{j,0} = g'_{j,0}$, $g_{j,1} = g'_{j,1}$ hold for any agent $j \neq i_0$, by (109), (111) can be rewritten as

$$\begin{aligned}
\|\Delta y_1\|_1 \leq& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|(1-\beta_K \sum_{j\in\mathcal{N}_{C,i_0}^-} C_{i_0 j})(y_{i_0,0} - y'_{i_0,0})\|_1 \\
& + \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|g_{i_0,1} - g'_{i_0,1}\|_1 + \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|g_{i_0,0} - g'_{i_0,0}\|_1.
\end{aligned} \tag{112}$$

Note that $\mathcal{D}, \mathcal{D}'$ are adjacent. Then, there exists at most one pair of data samples $\lambda_{i_0,1,l_2}, \lambda'_{i_0,1,l_2}$ such that $\lambda_{i_0,1,l_2} = \xi_{i_0,l_0}$, $\lambda'_{i_0,1,l_2} = \xi'_{i_0,l_0}$. Hence, (112) can be rewritten as

$$\begin{aligned}
\|\Delta y_1\|_1 \leq& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|(1-\beta_K \sum_{j\in\mathcal{N}_{C,i_0}^-} C_{i_0 j})(y_{i_0,0} - y'_{i_0,0})\|_1 + \frac{2C}{m_K} \\
=& |1 - \beta_K \sum_{j\in\mathcal{N}_{C,i_0}^-} C_{i_0 j}| \|\Delta y_0\|_1 + \frac{2C}{m_K} \\
\leq& |1 - \beta_K \sum_{j\in\mathcal{N}_{C,i_0}^-} C_{i_0 j}| \frac{C}{m_K} + \frac{2C}{m_K}.
\end{aligned} \tag{113}$$

When $k = 2, \ldots, K$, by Definition 5, we have

$$\begin{aligned}
\|\Delta y_k\|_1 =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|y_k - y_k'\|_1 \\
=& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \sum_{i=1}^n \|(1-\beta_K \sum_{j\in\mathcal{N}_{C,i}^-} C_{ij})(y_{i,k-1} - y'_{i,k-1}) \\
& - \beta_K \sum_{j\in\mathcal{N}_{C,i}^-} C_{ij}(\breve{y}_{j,k-1} - \breve{y}'_{j,k-1}) \\
& + (g_{i,k} - g'_{i,k}) + (g_{i,k-1} - g'_{i,k-1})\|_1.
\end{aligned} \tag{114}$$

Since $g_{j,l} = g'_{j,l}$, $y_{j,l} = y'_{j,l}$, $\breve{y}_{j,l} = \breve{y}'_{j,l}$ hold for any agent $j \neq i_0$ and $l = 0, \ldots, k-1$, (114) can be rewritten as

$$\begin{aligned}
\|\Delta y_k\|_1 =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|(1-\beta_K \sum_{j\in\mathcal{N}_{C,i_0}^-} C_{i_0 j})(y_{i_0,k-1} - y'_{i_0,k-1})\|_1 \\
& + \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|g_{i_0,k} - g'_{i_0,k}\|_1 + \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|g_{i_0,k-1} - g'_{i_0,k-1}\|_1.
\end{aligned} \tag{115}$$

Note that $\mathcal{D}, \mathcal{D}'$ are adjacent. Then, there exists at most one pair of data samples $\lambda_{i,k,l_{k+1}}, \lambda'_{i,k,l_{k+1}}$ such that $\lambda_{i,k,l_{k+1}} =$

ξ_{i,l_0}, $\lambda'_{i,k,l_{k+1}} = \xi'_{i,l_0}$. Hence, (115) can be rewritten as

$$\|\Delta y_k\|_1 \le |1-\beta_K \sum_{j\in\mathcal{N}^-_{C,i_0}} C_{i_0 j}| \|\Delta y_{k-1}\|_1 + \frac{2C}{m_K}. \tag{116}$$

Iteratively computing (116) implies

$$\|\Delta y_k\|_1 \le \sum_{l=0}^{k-1} |1-\beta_K \sum_{j\in\mathcal{N}^-_{C,i_0}} C_{i_0 j}|^l \frac{2C}{m_K} + |1-\beta_K \sum_{j\in\mathcal{N}^-_{C,i_0}} C_{i_0 j}|^k \frac{C}{m_K}. \tag{117}$$

Step 2: Next, we compute $\|\Delta x_k\|_1$ for any $k = 0,\dots,K$. When $k = 0$, since the initial value $x_{i,0} = x'_{i,0}$ for any $i \in \mathcal{V}$, we have $\|\Delta x_0\|_1 = 0$. When $k = 1$, by Definition 5, we have

$$\begin{aligned}\|\Delta x_1\|_1 =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|x_1 - x'_1\|_1 \\ =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \sum_{i=1}^n \|(1-\alpha_K \sum_{j\in\mathcal{N}^-_{R,i}} R_{ij})(x_{i,0}-x'_{i,0}) \\ &- \alpha_K \sum_{j\in\mathcal{N}^-_{R,i}} R^{(1)}_{ij}(\breve{x}_{j,0}-\breve{x}'_{j,0}) - \gamma_K (y_{i,0}-y'_{i,0})\|_1\end{aligned} \tag{118}$$

Note that the initial value $x_{i,0} = x'_{i,0}$ and $\breve{x}_{j,0} = \breve{x}'_{j,0}$ for any $i \in \mathcal{V}$, $j \in \mathcal{N}^-_{R,i}$. Then, by (109), (118) can be rewritten as

$$\|\Delta x_1\|_1 = \gamma_K \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|y_{i_0,0} - y'_{i_0,0}\|_1 \le \frac{\gamma_K C}{m_K}. \tag{119}$$

When $k = 2,\dots,K$, by Definition 5, we have

$$\begin{aligned}\|\Delta x_k\|_1 =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|x_k - x'_k\|_1 \\ =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \sum_{i=1}^n \|(1-\alpha_K \sum_{j\in\mathcal{N}^-_{R,i}} R_{ij})(x_{i,k-1}-x'_{i,k-1}) \\ &- \alpha_K \sum_{j\in\mathcal{N}^-_{R,i}} R_{ij}(\breve{x}_{j,k-1}-\breve{x}'_{j,k-1}) - \gamma_K (y_{i,k-1}-y'_{i,k-1})\|_1\end{aligned} \tag{120}$$

Since $x_{i,0} = x'_{i,0}$, $y_{j,l} = y'_{j,l}$ hold for any $i \in \mathcal{V}$, $j \in \mathcal{N}^-_{R,i}$, $l = 0,\dots,k-1$, $x_{j,l} = x'_{j,l}$ holds for any agent $j \ne i_0$. Thus, (120) can be rewritten as

$$\begin{aligned}\|\Delta x_k\|_1 =& \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|(1-\alpha_K \sum_{j\in\mathcal{N}^-_{R,i_0}} R_{i_0 j})(x_{i_0,k-1}-x'_{i_0,k-1}) \\ &- \gamma_K (y_{i_0,k-1}-y'_{i_0,k-1})\|_1.\end{aligned} \tag{121}$$

Note that $\sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|y_{i_0,k} - y'_{i_0,k}\|_1 = \|\Delta y_k\|_1$ for any $k = 0,\dots,K$. Then, (121) can be rewritten as

$$\begin{aligned}\|\Delta x_k\|_1 \le& |1-\alpha_K \sum_{j\in\mathcal{N}^-_{R,i_0}} R_{i_0 j}| \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|x_{i_0,k-1}-x'_{i_0,k-1}\|_1 \\ &+ \gamma_K \sup_{\mathrm{Adj}(\mathcal{D},\mathcal{D}')} \|y_{i_0,k-1} - y'_{i_0,k-1}\|_1 \\ =& |1-\alpha_K \sum_{j\in\mathcal{N}^-_{R,i_0}} R_{i_0 j}| \|\Delta x_{k-1}\|_1 + \gamma_K \|\Delta y_{k-1}\|_1.\end{aligned} \tag{122}$$

Iteratively computing (122) implies

$$\|\Delta x_k\|_1 \le \gamma_K \sum_{l=0}^{k-1} |1-\alpha_K \sum_{j\in\mathcal{N}^-_{R,i_0}} R_{i_0 j}|^{k-l-1} \|\Delta y_l\|_1. \tag{123}$$

Therefore, by (109), (113), (117), (119) and (123), this lemma is proved. ■

Appendix F
Proof of Lemma 3

For any observation set $\mathcal{O} \subseteq \mathbb{R}^{2n(K+1)d}$, let $\mathcal{K}_{\mathcal{D},\mathcal{O}} = \{(\zeta_0,\eta_0,\dots,\zeta_K,\eta_K)\colon \mathcal{M}(\mathcal{D})\in\mathcal{O}\}$, $\mathcal{K}_{\mathcal{D}',\mathcal{O}} = \{(\zeta'_0,\eta'_0,\dots,\zeta'_K,\eta'_K)\colon \mathcal{M}(\mathcal{D}')\in\mathcal{O}\}$ be sets of all possible state and tracking variables under the observation set $\mathcal{O}$ for adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$, respectively. Then, for any $(\zeta_0,\eta_0,\dots,\zeta_K,\eta_K)\in\mathcal{K}_{\mathcal{D},\mathcal{O}}$ there exists a unique $(\zeta'_0,\eta'_0,\dots,\zeta'_K,\eta'_K)\in\mathcal{K}_{\mathcal{D}',\mathcal{O}}$ such that $(\breve{x}_0,\breve{y}_0,\dots,\breve{x}_K,\breve{y}_K) = (\breve{x}'_0,\breve{y}'_0,\dots,\breve{x}'_K,\breve{y}'_K)$. Thus, we can define a bijection $\mathcal{B} : \mathcal{K}_{\mathcal{D},\mathcal{O}} \to \mathcal{K}_{\mathcal{D}',\mathcal{O}}$ such that $\mathcal{B}((\zeta_0,\eta_0,\dots,\zeta_K,\eta_K)) = (\zeta'_0,\eta'_0,\dots,\zeta'_K,\eta'_K)$ satisfies

$$\begin{aligned}&(x_0+\zeta_0, y_0+\eta_0,\dots,x_K+\zeta_K, y_K+\eta_K) \\ =&(\breve{x}_0,\breve{y}_0,\dots,\breve{x}_K,\breve{y}_K) = (\breve{x}'_0,\breve{y}'_0,\dots,\breve{x}'_K,\breve{y}'_K) \\ =&(x'_0+\zeta'_0, y'_0+\eta'_0,\dots,x'_K+\zeta'_K, y'_K+\eta'_K).\end{aligned} \tag{124}$$

Let $x^{(q)}_{i,k}$, $y^{(q)}_{i,k}$, $\zeta^{(q)}_{i,k}$, $\eta^{(q)}_{i,k}$, $x^{(q)\prime}_{i,k}$, $y^{(q)\prime}_{i,k}$, $\zeta^{(q)\prime}_{i,k}$, $\eta^{(q)\prime}_{i,k}$ be the q-th coordinate of $x_{i,k}$, $y_{i,k}$, $\zeta_{i,k}$, $\eta_{i,k}$, $x'_{i,k}$, $y'_{i,k}$, $\zeta'_{i,k}$, $\eta'_{i,k}$, respectively. Then, by (124), the following holds for any $k = 0,\dots,K$, $i = 1,\dots,n$, $q = 1,\dots,d$:

$$\begin{aligned}x^{(q)}_{i,k} - x^{(q)\prime}_{i,k} =& \zeta^{(q)\prime}_{i,k} - \zeta^{(q)}_{i,k}, \\ y^{(q)}_{i,k} - y^{(q)\prime}_{i,k} =& \eta^{(q)\prime}_{i,k} - \eta^{(q)}_{i,k}.\end{aligned} \tag{125}$$

Note that probability density functions of $(\zeta_0,\eta_0,\dots,\zeta_K,\eta_K)$ and $(\zeta'_0,\eta'_0,\dots,\zeta'_K,\eta'_K)$ are given as follows, respectively:

$$\begin{aligned}p(\zeta,\eta) =& \prod_{k=0}^K \prod_{i=1}^n \prod_{q=1}^d p(\zeta^{(q)}_{i,k};\sigma^{(\zeta)}_k) p(\eta^{(q)}_{i,k};\sigma^{(\eta)}_k), \\ p(\zeta',\eta') =& \prod_{k=0}^K \prod_{i=1}^n \prod_{q=1}^d p(\zeta^{(q)\prime}_{i,k};\sigma^{(\zeta)}_k) p(\eta^{(q)\prime}_{i,k};\sigma^{(\eta)}_k).\end{aligned} \tag{126}$$

Then, by (126), $\frac{p(\zeta,\eta)}{p(\mathcal{B}(\zeta,\eta))}$ can be rewritten as

$$\begin{aligned}&\frac{p(\zeta,\eta)}{p(\mathcal{B}(\zeta,\eta))} \\ =& \prod_{k=0}^K \prod_{i=1}^n \prod_{q=1}^d \frac{p(\zeta^{(q)}_{i,k};\sigma^{(\zeta)}_k) p(\eta^{(q)}_{i,k};\sigma^{(\eta)}_k)}{p(\zeta^{(q)\prime}_{i,k};\sigma^{(\zeta)}_k) p(\eta^{(q)\prime}_{i,k};\sigma^{(\eta)}_k)} \\ =& \prod_{k=0}^K \prod_{i=1}^n \prod_{q=1}^d \exp\left(\frac{|\zeta^{(q)\prime}_{i,k}| - |\zeta^{(q)}_{i,k}|}{\sigma^{(\zeta)}_k}\right) \exp\left(\frac{|\eta^{(q)\prime}_{i,k}| - |\eta^{(q)}_{i,k}|}{\sigma^{(\eta)}_k}\right) \\ \le& \prod_{k=0}^K \prod_{i=1}^n \prod_{q=1}^d \exp\left(\frac{|\zeta^{(q)\prime}_{i,k} - \zeta^{(q)}_{i,k}|}{\sigma^{(\zeta)}_k}\right) \exp\left(\frac{|\eta^{(q)\prime}_{i,k} - \eta^{(q)}_{i,k}|}{\sigma^{(\eta)}_k}\right).\end{aligned} \tag{127}$$

Substituting (125) into (127) implies

$$\begin{aligned}\frac{p(\zeta,\eta)}{p(\mathcal{B}(\zeta,\eta))} \le& \prod_{k=0}^K \prod_{i=1}^n \prod_{q=1}^d \exp\left(\frac{|x^{(q)}_{i,k} - x^{(q)\prime}_{i,k}|}{\sigma^{(\zeta)}_k}\right) \exp\left(\frac{|y^{(q)}_{i,k} - y^{(q)\prime}_{i,k}|}{\sigma^{(\eta)}_k}\right) \\ =& \prod_{k=0}^K \exp\left(\frac{\|x_k - x'_k\|_1}{\sigma^{(\zeta)}_k}\right) \exp\left(\frac{\|y_k - y'_k\|_1}{\sigma^{(\eta)}_k}\right) \\ =& \exp\left(\sum_{k=0}^K \left(\frac{\|\Delta x_k\|_1}{\sigma^{(\zeta)}_k} + \frac{\|\Delta y_k\|_1}{\sigma^{(\eta)}_k}\right)\right).\end{aligned} \tag{128}$$

Let $\varepsilon = \sum_{k=0}^{K}(\frac{\|\Delta x_k\|_1}{\sigma_k^{(\zeta)}} + \frac{\|\Delta y_k\|_1}{\sigma_k^{(\eta)}})$. Then, by (128) we have

$$\begin{aligned}&\frac{\mathbb{P}(\mathcal{M}(\mathcal{D}) \in \mathcal{O})}{\mathbb{P}(\mathcal{M}(\mathcal{D}') \in \mathcal{O})} = \frac{\int_{\mathcal{K}_{\mathcal{D},\mathcal{O}}} p(\zeta,\eta)\mathrm{d}\zeta\mathrm{d}\eta}{\int_{\mathcal{K}_{\mathcal{D}',\mathcal{O}}} p(\zeta',\eta')\mathrm{d}\zeta'\mathrm{d}\eta'}\\ =&\frac{\int_{\mathcal{K}_{\mathcal{D},\mathcal{O}}} p(\zeta,\eta)\mathrm{d}\zeta\mathrm{d}\eta}{\int_{\mathcal{K}_{\mathcal{D}',\mathcal{O}}} p(\mathcal{B}(\zeta,\eta))\mathrm{d}\zeta'\mathrm{d}\eta'} = \frac{\int_{\mathcal{K}_{\mathcal{D},\mathcal{O}}} p(\zeta,\eta)\mathrm{d}\zeta\mathrm{d}\eta}{\int_{\mathcal{B}^{-1}(\mathcal{K}_{\mathcal{D}',\mathcal{O}})} p(\mathcal{B}(\zeta,\eta))\mathrm{d}\zeta\mathrm{d}\eta}\\ =&\frac{\int_{\mathcal{K}_{\mathcal{D},\mathcal{O}}} p(\zeta,\eta)\mathrm{d}\zeta\mathrm{d}\eta}{\int_{\mathcal{K}_{\mathcal{D},\mathcal{O}}} p(\mathcal{B}(\zeta,\eta))\mathrm{d}\zeta\mathrm{d}\eta} \le \frac{e^{\varepsilon}\int_{\mathcal{K}_{\mathcal{D},\mathcal{O}}} p(\mathcal{B}(\zeta,\eta))\mathrm{d}\zeta\mathrm{d}\eta}{\int_{\mathcal{K}_{\mathcal{D},\mathcal{O}}} p(\mathcal{B}(\zeta,\eta))\mathrm{d}\zeta\mathrm{d}\eta} = e^{\varepsilon}.\end{aligned}$$

Therefore, this lemma is proved. ■

REFERENCES

[1] S. Pu, W. Shi, J. Xu, and A. Nedić, "Push-pull gradient methods for distributed optimization in networks," *IEEE Trans. Autom. Control*, vol. 66, no. 1, pp. 1–16, 2020.

[2] N. Bhavsar and L. A. Prashanth, "Nonasymptotic bounds for stochastic optimization with biased noisy gradient oracles," *IEEE Trans. Autom. Control*, vol. 68, no. 3, pp. 1628–1641, 2023.

[3] O. Shamir and N. Srebro, "Distributed stochastic optimization and learning," in *52nd Annu. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, 2014, pp. 850–857.

[4] T. Chen, A. G. Marques, and G. B. Giannakis, "DGLB: distributed stochastic geographical load balancing over cloud networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 7, pp. 1866–1880, 2016.

[5] M. Wang, C. Xu, X. Chen, H. Hao, L. Zhong, and D. O. Wu, "Design of multipath transmission control for information-centric Internet of Things: a distributed stochastic optimization framework," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9475–9488, 2019.

[6] 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.

[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] P. Xie, K. You, R. Tempo, S. Song, and C. Wu,"Distributed convex optimization with inequality constraints over time-varying unbalanced digraphs," *IEEE Trans. Autom. Control*, vol. 63, no. 12, pp. 4331–4337, 2018.

[9] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, 2018.

[10] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with accelerated convergence," *IEEE Trans. Signal Process.*, vol. 68, pp. 6255–6271, Oct. 2020.

[11] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Math. Program.*, vol. 187, no. 1, pp. 409–457, 2021.

[12] A. Koloskova, T. Lin, and S. U. Stich, "An improved analysis of gradient tracking for decentralized machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 11422–11435.

[13] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *IEEE 58th Conf. Decis. Control*, Nice, France, 2019, pp. 8353–8358.

[14] 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.

[15] S. Zhao and Y. Liu, "Asymptotic properties of $\mathcal{S}$-$\mathcal{AB}$ method with diminishing step-size," *IEEE Trans. Autom. Control*, vol. 69, no. 5, pp. 3222-3229, 2024.

[16] Y. Chen, A. Hashemi, and H. Vikalo, "Accelerated distributed stochastic non-convex optimization over time-varying directed networks," *IEEE Trans. Autom. Control*, vol. 70, no. 4, pp. 2196-2211, 2025.

[17] Y. Wang and T. Başar, "Gradient-tracking-based distributed optimization with guaranteed optimality under noisy information sharing," *IEEE Trans. Autom. Control*, vol. 68, no. 8, pp. 4796–4811, 2023.

[18] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures", in *Conf. Comput. Commun. Secur.*, Denver, CO, USA, 2015, pp. 1322–1333.

[19] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. Adv. Neural Inf. Process Syst.*, Vancouver, Canada, 2019, vol. 32, pp. 14774–14784.

[20] 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.

[21] Y. Lu and M. Zhu, "Privacy preserving distributed optimization using homomorphic encryption," *Automatica*, vol. 96, pp. 314–325, 2018.

[22] Y. Wang, "Privacy-preserving average consensus via state decomposition," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4711–4716, 2019.

[23] H. Gao, Y. Wang, and A. Nedić, "Dynamics based privacy preservation in decentralized optimization," *Automatica*, vol. 151, 2023, Art. no. 110878.

[24] Y. Wang and A. Nedić, "Decentralized gradient methods with time-varying uncoordinated stepsizes: convergence analysis and privacy design," *IEEE Trans. Autom. Control*, vol. 69, no. 9, pp. 5352–5367, 2024.

[25] G. Ramos, A. P. Aguiarz, S. Karx, and S. Pequito, "Privacy preserving average consensus through network augmentation," *IEEE Trans. Autom. Control*, vol. 69, no. 10, pp. 6907–6919, 2024.

[26] Y. Mo and R. M. Murray, "Privacy preserving average consensus," *IEEE Trans. Autom. Control*, vol. 62, no. 2, pp. 753–765, 2017.

[27] 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.

[28] J. Le Ny and G. J. Pappas, "Differentially private filtering," *IEEE Trans. Autom. Control*, vol. 59, no. 2, pp. 341–354, 2014.

[29] T. Ding, S. Zhu, J. He, C. Chen, and X. Guan, "Differentially private distributed optimization via state and direction perturbation in multiagent systems," *IEEE Trans. Autom. Control*, vol. 67, no. 2, pp. 722–737, 2022.

[30] Y. Xuan and Y. Wang, "Gradient-tracking based differentially private distributed optimization with enhanced optimization accuracy," *Automatica*, vol. 155, 2023, Art. no. 111150.

[31] Y. Wang and A. Nedić, "Tailoring gradient methods for differentially private distributed optimization," *IEEE Trans. Autom. Control*, vol. 69, no. 2, pp. 872–887, 2024.

[32] L. Huang, J. Wu, D. Shi, S. Dey, and L. Shi, "Differential privacy in distributed optimization with gradient tracking," *IEEE Trans. Autom. Control*, vol. 69, no. 9, pp. 5727–5742, 2024.

[33] A. Xie, X. Yi, X. Wang, M. Cao, and X. Ren, "Differentially private and communication-efficient distributed nonconvex optimization algorithms," *Automatica*, vol. 177, 2025, Art. no. 112338.

[34] W. Huo, X. Chen, L. Huang, K. H. Johansson, and L. Shi, "Differentially private dual gradient tracking for distributed resource allocation," 2024, *arXiv:2403.18275*.

[35] 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, 2021, vol. 35, no. 8, pp. 7219–7227.

[36] Y. Kang, Y. Liu, B. Niu, and W. Wang, "Weighted distributed differential privacy ERM: convex and non-convex," *Comput. Secur.*, vol. 106, 2021, Art. no. 102275.

[37] 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.

[38] Y. Wang and T. Başar, "Decentralized nonconvex optimization with guaranteed privacy and accuracy," *Automatica*, vol. 150, 2023, Art. no. 110858.

[39] 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.

[40] 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.

[41] C. Liu, K. H. Johansson, and Y. Shi, "Distributed empirical risk minimization with differential privacy," *Automatica*, vol. 162, 2024, Art. no. 111514.

[42] Z. Chen and Y. Wang, "Locally differentially private distributed online learning with guaranteed optimality," *IEEE Trans. Autom. Control*, vol. 70, no. 4, pp. 2521–2536, 2025.

[43] J. Wang and J. F. Zhang, "Differentially private distributed stochastic optimization with time-varying sample sizes," *IEEE Trans. Autom. Control*, vol. 69, no. 9, pp. 6341–6348, 2024.

[44] J. Chen, J. Wang, and J. F. Zhang, "Differentially private distributed stochastic nonconvex optimization with quantized communication," 2024, *arXiv:2403.18254v2*.

[45] W. Ren and R. W. Beard, "Consensus seeking in multiagent systems under dynamically changing interaction topologies," *IEEE Trans. Autom. control*, vol. 50, no. 5, pp. 655–661, 2005.

[46] 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.

[47] A. Jalilzadeh, A. Nedić, U. V. Shanbhag, and F. Yousefian, "A variable sample-size stochastic quasi-Newton method for smooth and nonsmooth stochastic convex optimization", *Math. Operations Res.*, vol. 47, no. 1, pp. 690–719, 2022.

[48] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *Int. Conf. Mach. Learn.*, Vienna, Austria, 2020, pp. 5381–5393.

[49] S. Bubeck, "Convex optimization: algorithms and complexity," *Found. Trends Theor. Comput. Sci.*, vol. 8, nos. 3-4, pp. 231–358, 2015.

[50] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, "Sample size selection in optimization methods for machine learning," *Math. Program.*, vol. 134, no. 1, pp. 127–155, 2012.

[51] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," in *Proc. Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018, pp. 1–11.

[52] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. 1380–1405, 2012.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[54] 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/

[55] A. Krizhevsky, V. Nair, and G. Hinton, 2009, "Canadian Institute for Advanced Research, 10 classes," Department of Computer Science of University of Toronto. [Online]. Avalable: http://www.cs.toronto.edu/kriz/cifar.html

[56] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, CA, 2009. [Online]. Avalable: http://www.cs.utoronto.ca/kriz/learning-features-2009-TR.pdf

[57] R. A. Horn and C. R. Johnson, "Positive and nonnegative matrices," in *Matrix analysis*, Cambridge, U.K.: Cambridge University Press, 2012, ch. 8, sec. 4, pp. 529–545.

[58] V. A. Zorich, "Integration," in *Mathematical analysis I*, Berlin, German: Springer-Verlag, 2015, ch. 6, sec. 2, pp. 349–360.

[59] 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.

[60] T. K. Chandra, "Extensions of the first BCL," in *The Borel-Cantelli lemma*, New York, NY, USA: Springer-Verlag, 2012, ch. 2, sec. 1, pp. 51–53.