Research Article

D-(DP)²SGD: Decentralized Parallel SGD with Differential Privacy in Dynamic Networks

Yuan Yuan,¹ Zongrui Zou,¹ Dong Li,² Li Yan,² and Dongxiao Yu¹

¹School of Computer Science and Technology, Shandong University, Qingdao 266237, China

²Information & Telecommunications Company, State Grid Shandong Electric Power Company, Jinan 250000, China

Correspondence should be addressed to Dongxiao Yu; dxyu@sdu.edu.cn

Received 12 December 2020; Revised 5 February 2021; Accepted 2 March 2021; Published 24 March 2021

Academic Editor: Zhuojun Duan

Copyright © 2021 Yuan Yuan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Decentralized machine learning has been playing an essential role in improving training efficiency. It has been applied in many real-world scenarios, such as edge computing and IoT. However, in fact, networks are dynamic, and there is a risk of information leaking during the communication process. To address this problem, we propose a decentralized parallel stochastic gradient descent algorithm (D-(DP)²SGD) with differential privacy in dynamic networks. With rigorous analysis, we show that D-(DP)²SGD converges with a rate of $O(1/\sqrt{Kn})$ while satisfying ϵ -DP, which achieves almost the same convergence rate as previous works without privacy concern. To the best of our knowledge, our algorithm is the first known decentralized parallel SGD algorithm that can implement in dynamic networks and take privacy-preserving into consideration.

1. Introduction

Decentralized machine learning, as a modeling mechanism that allocates training tasks and computes resources to achieve a balance between training speed and accuracy, has demonstrated strong potential in various areas, especially for training large models on large datasets [1–3], such as ImageNet [4]. Typically, assume that there are n workers where each worker has its local data, decentralized machine learning problem is aimed at solving an empirical risk minimization problem as follows:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x), \quad (1)$$

where $f_i(x)$ is the local loss function at node i . The objective $f(x)$ can be rephrased as a linear combination of the local loss function $f_i(x)$. This formulation can be expressed as many popular decentralized learning models including deep learning [5], linear regression [6], and logistic regression [7].

In recent years, decentralized machine learning has attracted much attention to derive convergence solutions

while reducing communication costs [8, 9]. Previous works mainly study decentralized collaborative learning in a static network assumption. For example, decentralized parallel stochastic gradient descent (D-PSGD) is one of the fundamental methods in solving large-scale machine learning tasks in static networks [1]. In D-PSGD, all nodes compute the stochastic gradient using their local dataset and exchange the results with their neighbors iteratively. However, in fact, dynamicity has been an important feature for networks, especially for large-scale networks, such as IoT [10] and V2V networks [11, 12], as nodes in the network can move around and join or leave the network at any time. On the other hand, in large-scale networks, it is hard or even impossible to ensure every node is reliable [13, 14]. Consequently, during the collaborative learning process, it is unavoidable to face the risk of information leaking. Hence, when designing decentralized machine learning algorithms, it has been to consider the impact of the dynamicity of network topology and the demand for privacy preservation. However, to the best of our knowledge, there is no existing work taking both factors simultaneously into consideration. In this work, we focus on this missing piece in decentralized learning.

Specifically, based on differential privacy, we present a new dynamic decentralized stochastic gradient descent algorithm (D-(DP)²SGD), which offers a strong protection for local datasets of decentralized nodes. With rigorous analysis, we show that our proposed D-(DP)²SGD algorithm satisfies ϵ -DP and achieves the convergence rate of $O(1/\sqrt{Kn})$ when K is large enough. Empirically, we conduct experiments on CIFAR-10 datasets to accomplish image classification tasks. We conduct extensive experiments to evaluate the performance of our proposed algorithms.

The remainder of this paper is organized as follows. We present our survey on related work in Section 2. We then introduce our model, problem, and some useful preliminary knowledge in Section 3. Our algorithm, main results, and analysis are presented in Section 4, Section 5, and Section 6, respectively. Experimental results are illustrated in Section 7. The whole paper is concluded in Section 8.

2. Related Work

In this section, we introduce some closed related work.

2.1. Decentralized Parallel Stochastic Gradient Algorithms. Most existing work on decentralized parallel stochastic gradient focus on static networks in both synchronous and asynchronous settings [1, 15–18]. Under the synchronous setting, Lian et al. [1] illustrated the advantage of decentralized algorithms over centralized ones and showed that the proposed D-PSGD converges with a rate of $O(1/\sqrt{Kn})$ when K is large enough, where K is the number of iterations and n is the total number nodes in the network. Qureshi et al. [17] proposed an algorithm called S-ADDOPT, and it converges with a rate of $O(1/K)$.

Feyzmahdavian et al. [19] and Agarwal et al. [20] considered decentralized SGD in the asynchronous setting. They allowed workers to use stale weights to compute gradients in S-PSGD. Asynchronous algorithms avoid idling any worker to reduce the communication overhead, and it is very robust because it can still work well when part of the computing workers are down. For asynchronous algorithms, Lian et al. [16] proposed the asynchronous decentralized parallel SGD algorithm for convex optimization and showed AD-PSGD converges at $O(1/\sqrt{K})$. Then, Lian et al. [21] proposed the asynchronous decentralized parallel stochastic gradient descent algorithm for nonconvex optimization and showed the ergodic convergence rate is $O(1/\sqrt{K})$. They proved that the linear speedup is achievable when the number of workers is bounded by $\sqrt{K}$.

2.2. Differentially Privacy Decentralized Learning. Most existing work on differentially private decentralized learning focuses on the static network [22–25]. Our work combines decentralized learning and dynamic network in a DP setting. In contract, Lu et al. [24] proposed an asynchronous federated learning scheme with differential private for resource sharing in vehicular networks. Hsin-Pai et al. [26] proposed a new learning algorithm LEASGD (Leader-Follower Elastic Averaging Stochastic Gradient Descent), which is driven by a novel Leader-Follower topology and a differential privacy

model. And they provide a theoretical analysis of the convergence rate and the trade-off between the performance and privacy in the private setting. Based on the research in [16], Xu et al. [2] designed an algorithm on asynchronous decentralized parallel stochastic gradient descent algorithm with differential privacy (A(DP)²SGD). They showed that A(DP)²SGD converges at $O(1/\sqrt{K})$. For all of these reviewed papers, the study of decentralized parallel SGD for differential privacy in dynamic networks is still an open problem.

3. System Model and Problem Description

We consider a network consisting of n computational nodes (could be a machine or a GPU). At each iterate k , the network topology is denoted by a network $\mathcal{G}_k = (\mathcal{V}, \mathcal{E}_k)$, where $\mathcal{V}$ is the set of n computational nodes, $\mathcal{V} = \{1, 2, \dots, n\}$, and $\mathcal{E}_k \subset \mathcal{V} \times \mathcal{V}$ is the set of communication edges at iterate k . If there exists an edge from node i to node j at iterate k , then $(i, j) \in \mathcal{E}_k$. In a connected network, two nodes are neighbors if the node can be connected directly by an edge, i.e., the nodes can communicate with each other. The set of neighbors of node i at iterate k is denoted by $\mathcal{N}_k(i) = \{j \mid (i, j) \in \mathcal{E}_k\}$, and define $\mathcal{C}_k(i) = \mathcal{N}_k(i) \cup \{i\}$. We assume that the nodes keep unchanged, but the connection between nodes can be changed after every iteration. The network $\mathcal{G}_k$ is assumed to be strongly connected, i.e., for all nodes $i, j \in \mathcal{V}$, there exists a path from i to j at each iterate $k \geq 0$. Some frequently used notations are summarized in Table 1.

In a decentralized network, the data is stored at nodes, and each node is associated with a local loss function

$$f_i(x) = \mathbb{E}_{\xi \sim \mathcal{D}_i} F_i(x; \xi), \quad (2)$$

where $\mathcal{D}_i$ is a distribution associated with the local data at node i and ξ is a data sampled via $\mathcal{D}_i$.

In this work, we consider the following optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \mathbb{E}_{i \sim \mathcal{I}} f_i(x) = \mathbb{E}_{i \sim \mathcal{I}} \mathbb{E}_{\xi \sim \mathcal{D}_i} F_i(x; \xi), \quad (3)$$

where $\mathcal{I}$ is a uniform distribution of nodes.

Similarity, we gives an δ -approximation solution if

$$K^{-1} \left(\sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(\bar{x}_k)\|^2 \right) \leq \delta, \quad (4)$$

where $\bar{x}_k$ is the average local variable with all nodes at iterate k and K is the maximum iterations.

We next review the definition of differential privacy, which is originally proposed by Dwork [27].

Definition 1 (see [27] (Differential Privacy)). Given a $\epsilon \geq 0$, a randomized mechanism $\mathcal{M}$ with domain $\mathfrak{D}$ preserves ϵ -differential privacy if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for any adjacent datasets (Given two datasets $D = \{x_1, x_2, \dots, x_n\}$ and $D' = \{x'_1, x'_2, \dots, x'_n\}$, D and D' are adjacent if there exist $i \in [n]$ such that $x_i \neq x'_i$ and for $j \neq i$, $x_j = x'_j$, i.e., $\|D - D'\|_1 = 1$.) D, D'

TABLE 1: Frequently used notations.

Notation	Description
$\ \cdot\ $	The vector ℓ_2 norm or the matrix spectral norm depending on the argument
$\ \cdot\ _F$	The matrix Frobenius norm
$f(\cdot)$	The optimization function
$\nabla f(\cdot)$	The gradient of a function f
$\mathbf{1}_n$	The column vector in $\mathbb{R}^n$ with 1 for all elements
f^*	The optimal solution of problem
$\lambda_i(\cdot)$	The i th largest eigenvalue of a matrix
$\mathcal{G}_k$	The network topology at iterate k
$\mathcal{V}$	The set of nodes
$\mathcal{E}_k$	The set of edges at iterate k
K	The total number of iterations
$\mathcal{N}_k(i)$	The set of neighbors of node i at iterate k
$x_{k,i}$	The local variable in node i at iterate k
$\bar{x}_k$	The average local variable with all nodes at iterate k
$\xi_{k,i}$	The sampled training data in node i at iterate k
$\tilde{x}_{k,i}$	The perturbed variable in node i at iterate k
$\eta_{k,i}$	The Laplace noise drawn from $Lap(\varsigma)$
X_k	The concatenation of all local variables
$\tilde{X}_k$	The concatenation of all perturbed variables
η_k	The concatenation of all Laplace noises
ξ_k	The concatenation of all random samples
$\partial F(X_k, \xi_k)$	The concatenation of all stochastic gradients

$\in \mathfrak{D}$ such that:

$$\mathbb{P}[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\varepsilon) \mathbb{P}[\mathcal{M}(D') \in \mathcal{S}], \quad (5)$$

where $\text{Range}(\mathcal{M})$ is the output range of mechanism $\mathcal{M}$.

Informally, differential privacy means that the distributions over the outputs of the randomized algorithm should be nearly identical for two adjacent input datasets. The constant ε measures the privacy level of the randomized mechanism $\mathcal{M}$, i.e., a large ε implies a lower privacy level. Therefore, an appropriate constant ε should be chosen to balance the accuracy and the privacy level of the mechanism $\mathcal{M}$.

Then, we introduce the definition of sensitivity, which plays a key role in the design of differential privacy mechanisms.

Definition 2 (see [28] (Sensitivity)). The sensitivity of a function $f : \mathfrak{D} \rightarrow \mathbb{R}^d$ is defined as follows:

$$\Delta = \sup_{D, D' : \|D - D'\|_1 = 1} \|f(D) - f(D')\|_1. \quad (6)$$

The sensitivity of a mechanism $\mathcal{M}$ captures the magnitude by which a single individual's data can change the mechanism $\mathcal{M}$ in the worse case. Moreover, we will introduce the Laplace mechanism.

Definition 3 (see [27] (The Laplace mechanism)). Given any function $f : \mathfrak{D} \rightarrow \mathbb{R}^d$, the Laplace mechanism is defined as:

$$\mathcal{M}(D, f(\cdot), \varepsilon) = f(D) + \eta_{k,i}, \quad (7)$$

where $\eta_{k,i}$ are i.i.d. random variables drawn from $Lap(\varsigma)$. The variable of Laplace distribution is $2\varsigma^2$, where $\varsigma = \Delta/\varepsilon$ according to the property of differential privacy.

Throughout the paper, we adopt the following commonly used assumptions:

- (1) Lipschitzian gradient: all functions $f_i(\cdot)$'s are L -Lipschitzian gradients.
- (2) Unbiased estimation:

$$\begin{aligned}\mathbb{E}_{\xi \sim \mathcal{D}_i} \nabla F_i(x; \xi) &= \nabla f_i(x), \\ \mathbb{E}_{i \sim \mathcal{I}} \mathbb{E}_{\xi \sim \mathcal{D}_i} \nabla F_i(x; \xi) &= \nabla f(x).\end{aligned}\quad (8)$$

- (3) Bounded variance: assume the variance of stochastic gradient

$$\mathbb{E}_{\xi \sim \mathcal{D}_i} \mathbb{E}_{\xi \sim \mathcal{D}_i} \|\nabla F_i(x; \xi) - \nabla f(x)\|^2, \quad (9)$$

is bounded for any x with i sampled from the distribution $\mathcal{I}$ and ξ from the distribution $\mathcal{D}_i$. It implies there exist constants σ and ζ such that

$$\begin{aligned}\mathbb{E}_{\xi \sim \mathcal{D}_i} \|\nabla F_i(x; \xi) - \nabla f_i(x)\|^2 &\leq \sigma^2, \quad \forall i, \forall x, \\ \mathbb{E}_{i \sim \mathcal{I}} \|\nabla f_i(x) - \nabla f(x)\|^2 &\leq \zeta^2, \quad \forall x.\end{aligned}\quad (10)$$

- (4) Bounded subgradient: assume that the subgradient $\nabla F_i(x; \xi)$ of $F_i(x; \xi)$ is G_i -bounded for all $i \in \mathcal{I}$ and $x \in \mathbb{R}^d$, i.e., $\|\nabla F_i(x; \xi)\| \leq G_i$ and $G = \max_{i \in \mathcal{I}} G_i$.

4. Algorithm

The D-(DP)²SGD algorithm can be described as follows: each node maintains its own local variable $x_{k,i}$ and run the following steps.

- (i) Sample data: each node samples a training data $\xi_{k,i}$.
- (ii) Compute gradient: each node computes the stochastic gradient $\nabla F_i(x_{k,i}, \xi_{k,i})$ using the current local variable $x_{k,i}$ and the data $\xi_{k,i}$, where i is the node index and k is the iterate number.
- (iii) Add noise: random generate the Laplace noise $\eta_{k,i} \sim \text{Lap}(\varsigma)$ and add noise to the variable $x_{k,i}$, to get the perturbed variable $\tilde{x}_{k,i} = x_{k,i} + \eta_{k,i}$.
- (iv) Communication: send the perturbed variable $\tilde{x}_{k,i}$ and the degree $d_{k,i}$ to its neighbors; receive $\tilde{x}_{k,j}$ and $d_{k,j}$ from neighbors, where $j \in \mathcal{N}_k(i)$.
- (v) Determine the matrix W_k : determine the matrix W_k according to the local network topology, i.e.,

$$w_{k,ij} = \begin{cases} \frac{1}{\max\{d_{k,i}, d_{k,j}\}} & \text{if } i \neq j, \text{ and } j \in \mathcal{N}_k(i), \\ 1 - \sum_{m \in \mathcal{N}_k(i)} w_{k,im} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

where $w_{k,ij}$ describes how much node j can affect node i at iterate k .

- (vi) Weighted average: compute the weighted average by obtaining perturbed variable from neighbors and the matrix W_k : $x_{k+(1/2),i} = \sum_{j \in \mathcal{N}_k(i)} w_{k,ij} \tilde{x}_{k,j}$.
- (vii) Gradient update: each node updates its local variable using the weighted average and the local stochastic gradient $x_{k+1,i} = x_{k+(1/2),i} - \gamma \nabla F_i(x_{k,i}, \xi_{k,i})$.

From a global view, we define the concatenation of all local variables, perturbed variables, Laplace noises, random samples, and stochastic gradients by matrix $X_k \in \mathbb{R}^{d \times n}$, $\tilde{X}_k \in \mathbb{R}^{d \times n}$, and $\eta_k \in \mathbb{R}^{d \times n}$, vector $\xi_k \in \mathbb{R}^n$, and $\partial F(X_k, \xi_k)$, respectively:

$$\begin{aligned}X_k &= [x_{k,1}, \dots, x_{k,n}] \in \mathbb{R}^{d \times n}, \eta_k = [\eta_{k,1}, \dots, \eta_{k,n}] \in \mathbb{R}^{d \times n}, \\ \tilde{X}_k &= [\tilde{x}_{k,1}, \dots, \tilde{x}_{k,n}] \in \mathbb{R}^{d \times n}, \xi_k = [\xi_{k,1}, \dots, \xi_{k,n}] \in \mathbb{R}^n, \\ \partial F(X_k, \xi_k) &= [\nabla F_1(x_{k,1}; \xi_{k,1}), \nabla F_2(x_{k,2}; \xi_{k,2}), \dots, \nabla F_n(x_{k,n}; \xi_{k,n})] \in \mathbb{R}^{d \times n}.\end{aligned}\quad (12)$$

Then, the k th iterate of Algorithm 1 can be described as the following update

$$X_{k+1} = \tilde{X}_k W_k - \gamma \partial F(X_k, \xi_k), \quad (13)$$

i.e., $X_{k+1} = (X_k + \eta_k) W_k - \gamma \partial F(X_k, \xi_k)$.

5. Main Results

In this section, we present the main results, which guarantees the privacy and the convergence rate of our proposed algorithm.

Theorem 4. *Let assumptions hold. If $\eta_{k,i}$'s are i.i.d. random variables drawn according to the Laplace distribution with parameter ς , such that $\varsigma = \Delta/\epsilon$ for all $k \in \{0, \dots, K-1\}$ and $\epsilon > 0$. Then, our proposed algorithm guarantees ϵ -differential private.*

Theorem 5. *Let $B_1 = ((1/2) - ((18\gamma^2(n-1)L^2)/((1-\sqrt{\rho})^2 B_2)))$, $B_2 = (1 - ((36\gamma^2(n-1)L^2)/((1-\sqrt{\rho})^2)))$. Under the assumptions, we can get the convergence rate of Algorithm 1 as follows:*

$$\begin{aligned}& \frac{((1-\gamma L)/2) \sum_{k=0}^{K-1} \mathbb{E} \|\partial f(X_k) 1_n / n\|^2 + B_1 \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f((X_k 1_n)/n)\|^2}{K} \\ & \leq \frac{f(0) - f^*}{\gamma K} + \frac{4(1+2L)G^2 d\gamma}{\epsilon^2} + \frac{\gamma L}{n} \sigma^2 + \frac{24(n-1)L^2 G^2 \gamma^2}{\epsilon^2 (1-\sqrt{\rho})^2 B_2} \\ & \quad + \frac{2\gamma^2 L^2 (n-1) \sigma^2}{(1-\rho) B_2} + \frac{18\gamma^2 (n-1) L^2 \zeta^2}{(1-\sqrt{\rho})^2 B_2},\end{aligned}\quad (14)$$

where $((X_k 1_n)/n) = (1/n) \sum_{i=1}^n x_{k,i} = \bar{x}_k$.

This theorem characterizes the convergence of the average of all local optimization variables $x_{k,i}$. To take a closer

Initialization

Initial point $x_{0,i} = x_0 = 0$, step length γ , noise variance ς and number of iterations K

end

fork $k = 0, 1, \dots, K - 1$ in parallel for nodes $i \in \mathcal{V}$ **do**

Sample a training data $\xi_{k,i}$;

Compute the stochastic gradient $\nabla F_i(x_{k,i}, \xi_{k,i})$ using the current local variable $x_{k,i}$ and the data $\xi_{k,i}$;

Randomly generate the Laplace noise $\eta_{k,i} \sim \text{Lap}(\varsigma)$ and add noise to the variable $x_{k,i}$, to get the perturbed variable $\tilde{x}_{k,i}$: $\tilde{x}_{k,i} = x_{k,i} + \eta_{k,i}$;

Send the perturbed variable $\tilde{x}_{k,i}$ and its degree $d_{k,i}$ to its neighbors;

Receive $\tilde{x}_{k,j}$ and $d_{k,j}$ from its neighbors, $j \in \mathcal{N}_k(i)$;

Determine W_k according to Equation (11);

Compute the neighborhood weighted average by obtaining perturbed variables from neighbors: $x_{k+(1/2),i} = \sum_{j \in \mathcal{N}_k(i)} w_{k,ij} \tilde{x}_{k,j}$;

Update its local variable $x_{k+1,i} = x_{k+(1/2),i} - \gamma \nabla F_i(x_{k,i}, \xi_{k,i})$;

end

Output: $(1/n) \sum_{i=1}^n x_{K,i}$.

ALGORITHM 1: D-(DP)²SGD: Dynamic Decentralized Parallel Stochastic Gradient Descent.

look at this result, we choose an appropriate step length in Theorem 5 to obtain the following result:

Corollary 6. *Under the same assumptions as in Theorem 5, let $(G/\varepsilon) \leq \sqrt{((f(0) - f^* + L)\sigma)/(4(1 + 2L)dn)}$, by setting $\gamma = 1/(2L + \sigma\sqrt{K/n})$, we have the following convergence rate:*

$$\frac{\sum_{k=0}^{K-1} \mathbb{E} \|\nabla f((X_k I_n)/n)\|^2}{K} \leq \frac{8L(f(0) - f^*)}{K} + \frac{12(f(0) - f^* + L)\sigma}{\sqrt{Kn}}, \quad (15)$$

if the total number of iterates K is sufficiently large,

$$K \geq \max \left\{ \frac{16n^3(n-1)^2 L^4}{\sigma^6(f(0) - f^* + L)^2} \left(\frac{12G^2}{\varepsilon^2(1 - \sqrt{\rho})^2} + \frac{\sigma^2}{(1 - \rho)} + \frac{9\varsigma^2}{(1 - \sqrt{\rho})^2} \right)^2, \frac{144n(n-1)L^2}{\sigma^2(1 - \sqrt{\rho})^2} \right\}. \quad (16)$$

6. Result Analysis

In this section, we will give the analysis for privacy preservation and convergence rate of D-(DP)²SGD.

For convenience, we define

$$\partial f(X_k) = [\nabla f_1(x_{k,1}) \nabla f_2(x_{k,2}) \cdots \nabla f_n(x_{k,n})] \in \mathbb{R}^{d \times n}, \quad (17)$$

and

$$\Phi(k : s) = \begin{cases} W_k W_{k-1} \cdots W_s, & k \geq s, \\ I, & s = k + 1. \end{cases} \quad (18)$$

6.1. Privacy Analysis

Proof (Proof of Theorem 4). From the definition of sensitivity, we obtain

$$\|x_{k,i} - x'_{k,i}\|_1 \leq \Delta. \quad (19)$$

Note that $x_{k,i}$ and $x'_{k,i}$ are in space $\mathbb{R}^d$. According to the

definition of 1-norm, we have

$$\sum_{l=1}^d |x_{k,i}^l - x'_{k,i}{}^l| = \|x_{k,i} - x'_{k,i}\|_1 \leq \Delta, \quad (20)$$

where $x_{k,i}^l$ and $x'_{k,i}{}^l$ are the l th component of $x_{k,i}$ and $x'_{k,i}$, respectively.

Consider an output vectors $y_{k,i}$. Then, we follow from the property of Laplace distribution, we get

$$\begin{aligned} \mathbb{P}[\mathcal{M}(D_k) \in \mathcal{S}] &= \prod_{l=1}^d \frac{1}{2\varsigma} \exp \left(-\frac{|y_{k,i}^l - x_{k,i}^l|}{\varsigma} \right), \\ \mathbb{P}[\mathcal{M}(D'_k) \in \mathcal{S}] &= \prod_{l=1}^d \frac{1}{2\varsigma} \exp \left(-\frac{|y_{k,i}^l - x'_{k,i}{}^l|}{\varsigma} \right). \end{aligned} \quad (21)$$

Then,

$$\begin{aligned} \frac{\mathbb{P}[\mathcal{M}(D_k) \in \mathcal{S}]}{\mathbb{P}[\mathcal{M}(D'_k) \in \mathcal{S}]} &= \prod_{l=1}^d \exp \left(-\frac{|y_{k,i}^l - x_{k,i}^l| - |y_{k,i}^l - x'_{k,i}{}^l|}{\varsigma} \right) \\ &\leq \prod_{l=1}^d \exp \left(\frac{|x_{k,i}^l - x'_{k,i}{}^l|}{\varsigma} \right) = \exp \left(\frac{\|x_{k,i} - x'_{k,i}\|_1}{\varsigma} \right) \\ &\leq \exp \left(\frac{\Delta}{\varsigma} \right), \end{aligned} \quad (22)$$

where the first inequality comes from the triangle inequality, and the last inequality follows from (19).

Therefore, we know $\varepsilon = \Delta/\varsigma$, we can obtain

$$\frac{\mathbb{P}[\mathcal{M}(D) \in \mathcal{S}]}{\mathbb{P}[\mathcal{M}(D') \in \mathcal{S}]} \leq \exp(\varepsilon). \quad (23)$$

6.2. Convergence Rate Analysis. In order to obtain the result of convergence rate, we first give some lemmas.

Lemma 7. W_k is a symmetric doubly stochastic matrix.

Proof. From the definition of Equation (11), we can obtain that

$$w_{k,ij} \in [0, 1], \quad \forall i, j, \quad (24)$$

- (1) $w_{k,ij} \in [0, 1]$, for all i, j ;
- (2) $w_{k,ij} = w_{k,ji}$, for all i, j ;
- (3) $\sum_j w_{k,ij} = 1$, for all i .

Therefore, W_k is a symmetric doubly stochastic matrix.

Lemma 8. Define $\Phi(K-1, 0) = I$, where I is the identity matrix. Assume that there exists a $\rho \in [0, 1]$ such that

$$\max \{ |\lambda_2(\mathbb{E}[W_s^T W_s])|, |\lambda_n(\mathbb{E}[W_s^T W_s])| \} \leq \rho, \quad \forall s. \quad (25)$$

Then

$$\mathbb{E} \left\| \frac{I_n}{n} - \Phi(K-1, 0)e_i \right\|^2 \leq \frac{n-1}{n} \rho^K, \quad \forall K \geq 0. \quad (26)$$

Proof. Let $y_s = (1_n/n) - \Phi(s-1, 0)e_i$. We prove this lemma by induction. For $s=0$, $\|y_0\|^2 = \|(1_n/n) - e_i\|^2 = ((n-1)^2/n^2) + \sum_{i=1}^{n-1} (1/n^2) = ((n^2 - 2n + 1 + n - 1)/n^2) = (n-1)/n$.

We assume that $s=K$ hold, i.e., $\mathbb{E}\|y_K\|^2 \leq ((n-1)/n)\rho^K$. Then, for $s=K+1$, note that $y_{K+1} = W_K y_K$, then we have

$$\begin{aligned} \mathbb{E}\|y_{K+1}\|^2 &= \mathbb{E}\|W_K y_K\|^2 = \mathbb{E}\langle W_K y_K, W_K y_K \rangle = \mathbb{E}\langle y_K, W_K^T W_K y_K \rangle \\ &= \mathbb{E}\langle y_K, \mathbb{E}(W_K^T W_K) y_K \rangle. \end{aligned} \quad (27)$$

According to Lemma 7, $\mathbb{E}(W_K^T W_K)$ is symmetric and doubly stochastic. Then, 1_n is an eigenvector of $\mathbb{E}(W_K^T W_K)$, and 1 is the eigenvalue. According to the spectral theorem of Hermitian matrices, we can construct a basis of $\mathbb{R}^n$ composed by the eigenvectors of $\mathbb{E}(W_K^T W_K)$ starting from 1_n . From Equation (25), the magnitude of all other eigenvectors' associated eigenvalues should be smaller or equal to ρ . Note that y_K is orthogonal to 1_n , then we can find

$$\mathbb{E}\|y_{K+1}\|^2 \leq \rho \mathbb{E}\|y_K\|^2. \quad (28)$$

By induction, we complete the proof.

In Lemma 9, we give the bound of the sensitivity of our proposed algorithm.

Lemma 9. Under assumption 4, the sensitivity of the algorithm can be bounded as

$$\Delta \leq 2G\sqrt{d}\gamma, \quad (29)$$

where $G = \max_{i \in \mathcal{V}} G_i$ and d is the dimensionality of vectors.

Proof. Assume that D_k and D'_k are any two adjacent datasets at iterate k . Assume that $x_{k,i}$ and $x'_{k,i}$ be the executions for $\mathcal{M}(D_k)$ and $\mathcal{M}(D'_k)$, respectively. Then, from our proposed algorithm, we have

$$\begin{aligned} \|\mathcal{M}(D_k) - \mathcal{M}(D'_k)\|_1 &= \|x_{k+1,i} - x'_{k+1,i}\|_1 \\ &= \gamma \|\nabla F_i(x_{k,i}; \xi_{k,i}) - \nabla F_i(x_{k,i}; \xi'_{k,i})\|_1 \\ &\leq \sqrt{d}\gamma \|\nabla F_i(x_{k,i}; \xi_{k,i}) - \nabla F_i(x_{k,i}; \xi'_{k,i})\| \\ &\leq \sqrt{d}\gamma (\|\nabla F_i(x_{k,i}; \xi_{k,i})\| + \|\nabla F_i(x_{k,i}; \xi'_{k,i})\|), \end{aligned} \quad (30)$$

where the first inequality comes from the norm inequality and the last inequality is from the triangle inequality. From assumption 4, we have

$$\nabla F_i(x_{k,i}; \xi_{k,i}) \leq G_i \leq G. \quad (31)$$

Since we can choose the pair of adjacent datasets D_k, D'_k arbitrarily, and we can obtain

$$\Delta \leq 2G\sqrt{d}\gamma. \quad (32)$$

The lemma is obtained.

From Lemma 9, we know that the learning rate γ , the dimensionality of vectors d , the maximal bound of subgradient G , and the privacy level ϵ have an effect on the magnitude of the added random noise. Based on Lemma 9, we next provide the bound of the noise.

Lemma 10. We give the following inequalities:

$$\mathbb{E} \left\| \frac{\eta_s I_n}{n} \right\|^2 \leq \frac{8G^2 d \gamma^2}{\epsilon^2} \text{ and } \mathbb{E}\|\eta_s\|^2 \leq \frac{8nG^2 d \gamma^2}{\epsilon^2}. \quad (33)$$

Proof. According to the property of the Laplace mechanism $\mathbb{E}\|\eta_{s,i}\|^2 \leq 2\varsigma^2$ and $\varsigma = \Delta/\epsilon$, we obtain a bound on η_s in the following:

$$\begin{aligned} \mathbb{E} \left\| \frac{\eta_s I_n}{n} \right\|^2 &\leq \mathbb{E} \left\| \frac{\sum_{i=1}^n \eta_{s,i}}{n} \right\|^2 \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}\|\eta_{s,i}\|^2 \leq \frac{1}{n} \cdot n \cdot 2\varsigma^2 = 2\varsigma^2 \stackrel{(19)}{\leq} 2 \left(\frac{2G\sqrt{d}\gamma}{\epsilon} \right)^2 \\ &= \frac{8G^2 d \gamma^2}{\epsilon^2}. \end{aligned} \quad (34)$$

And we can get a bound on η_s .

$$\begin{aligned}\mathbb{E}\|\eta_s\|^2 &\leq \mathbb{E}\|\eta_s\|_F^2 \leq \sum_{i=1}^n \mathbb{E}\|\eta_{s,i}\|^2 \leq n \cdot 2\zeta^2 \leq 2n \left(\frac{2G\sqrt{d}\gamma}{\varepsilon} \right)^2 \\ &= \frac{8nG^2d\gamma^2}{\varepsilon^2}.\end{aligned}\quad (35)$$

Lemma 11 (see [1]). *Under assumption 1, the following inequality holds:*

$$\mathbb{E}\|\partial f(X_s)\|^2 \leq \sum_{h=1}^n 3\mathbb{E}L^2 \left\| \frac{\sum_{i'=1}^n X_{s,i'} - x_{s,h}}{n} \right\|^2 + 3n\zeta^2 + 3\mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2, \quad \forall s \in K. \quad (36)$$

The proof of this lemma can be found in the full version of [1]. And we define $\text{Dis}_{k,i}$ as the squared distance of the local optimization variable on node i from the averaged local optimization variable on all nodes at iterate k , i.e., $\text{Dis}_{k,i} = \mathbb{E} \left\| \left(\left(\sum_{i'=1}^n x_{k,i'} \right) / n \right) - x_{k,i} \right\|^2$. In the following, we will present the bound of $\text{Dis}_{k,i}$.

Lemma 12. *Under the definition of $\text{Dis}_{k,i}$, we can get:*

$$\begin{aligned}\text{Dis}_{k,i} &\leq \frac{48(n-1)G^2d\gamma^2}{\varepsilon^2(1-\sqrt{\rho})^2} + 2\gamma^2 \left(\frac{2(n-1)\sigma^2}{1-\rho} + \frac{18(n-1)\zeta^2}{(1-\sqrt{\rho})^2} + \frac{6(n-1)}{n} \right. \\ &\quad \cdot \sum_{s=0}^{k-1} \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) \\ &\quad \left. + \frac{6(n-1)}{n} \sum_{s=0}^{k-1} \sum_{h=1}^n \mathbb{E}L^2 \text{Dis}_{s,h} \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) \right).\end{aligned}\quad (37)$$

Proof. According to the update method of X_k , we split $\text{Dis}_{k,i}$ into two terms:

$$\begin{aligned}\text{Dis}_{k,i} &= \mathbb{E} \left\| \frac{\sum_{i'=1}^n x_{k,i'}}{n} - x_{k,i} \right\|^2 = \mathbb{E} \left\| \frac{X_k 1_n}{n} - X_k e_i \right\|^2 \\ &= \mathbb{E} \left\| \frac{(X_{k-1} + \eta_{k-1}) W_{k-1} 1_n - \gamma \partial F(X_{k-1}; \xi_{k-1}) 1_n}{n} \right. \\ &\quad \left. - ((X_{k-1} + \eta_{k-1}) W_{k-1} e_i - \gamma \partial F(X_{k-1}; \xi_{k-1}) e_i) \right\|^2 \\ &= \mathbb{E} \left\| \frac{X_{k-1} 1_n + \eta_{k-1} 1_n - \gamma \partial F(X_{k-1}; \xi_{k-1}) 1_n}{n} \right. \\ &\quad \left. - (X_{k-1} W_{k-1} e_i + \eta_{k-1} W_{k-1} e_i - \gamma \partial F(X_{k-1}; \xi_{k-1}) e_i) \right\|^2 \\ &= \mathbb{E} \left\| \frac{X_0 1_n}{n} + \frac{\sum_{s=0}^{k-1} \eta_s 1_n}{n} - \frac{\sum_{s=0}^{k-1} \gamma \partial F(X_s; \xi_s) 1_n}{n} \right. \\ &\quad \left. - \left(X_0 \Phi(k-1, 0) e_i + \sum_{s=0}^{k-1} \eta_s \Phi(k-1, s) e_i - \gamma \partial F(X_{k-1}, \xi_{k-1}) e_i \right) \right. \\ &\quad \left. - \sum_{s=0}^{k-2} \gamma \partial F(X_s; \xi_s) \Phi(k-1, s+1) e_i \right\|^2\end{aligned}$$

$$\begin{aligned}&\stackrel{(18)}{=} \mathbb{E} \left\| X_0 \left(\frac{1_n}{n} - \Phi(k-1, 0) e_i \right) + \sum_{s=0}^{k-1} \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right. \\ &\quad \left. - \gamma \sum_{s=0}^{k-1} \partial F(X_s; \xi_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1) e_i \right) \right\|^2 \\ &= \mathbb{E} \left\| \sum_{s=0}^{k-1} \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) - \gamma \sum_{s=0}^{k-1} \partial F(X_s; \xi_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1) e_i \right) \right\|^2 \\ &= 2\mathbb{E} \underbrace{\left\| \sum_{s=0}^{k-1} \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right\|^2}_{A_1} \\ &\quad + 2\gamma^2 \mathbb{E} \underbrace{\left\| \sum_{s=0}^{k-1} \partial F(X_s; \xi_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1) e_i \right) \right\|^2}_{A_2},\end{aligned}\quad (38)$$

where the seventh equation comes from $x_{0,i} = x_0 = 0$ for $\forall i$.

Firstly, we split A_1 into two terms,

$$\begin{aligned}A_1 &= \mathbb{E} \left\| \sum_{s=0}^{k-1} \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right\|^2 \\ &= \underbrace{\sum_{s=0}^{k-1} \mathbb{E} \left\| \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right\|^2}_{=A_3} \\ &\quad + \underbrace{\sum_{s \neq s'}^{k-1} \mathbb{E} \left\langle \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right), \eta_{s'} \left(\frac{1_n}{n} - \Phi(k-1, s') e_i \right) \right\rangle}_{=A_4}.\end{aligned}\quad (39)$$

To bound A_1 , we first bound A_3 and A_4 :

$$\begin{aligned}A_3 &= \sum_{s=0}^{k-1} \mathbb{E} \left\| \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right\|^2 \\ &\leq \sum_{s=0}^{k-1} \mathbb{E} \|\eta_s\|^2 \left\| \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right\|^2 \stackrel{(33)}{\leq} \sum_{s=0}^{k-1} \frac{8nG^2d\gamma^2}{\varepsilon^2} \cdot \frac{n-1}{n} \rho^{k-s} \\ &\leq \frac{8(n-1)G^2d\gamma^2}{\varepsilon^2} \frac{1}{1-\rho}.\end{aligned}\quad (40)$$

Moreover, A_4 can be bound as follows:

$$\begin{aligned}A_4 &= \sum_{s \neq s'}^{k-1} \mathbb{E} \left\langle \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right), \eta_{s'} \left(\frac{1_n}{n} - \Phi(k-1, s') e_i \right) \right\rangle \\ &\leq \sum_{s \neq s'}^{k-1} \mathbb{E} \left\| \eta_s \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right\| \left\| \eta_{s'} \left(\frac{1_n}{n} - \Phi(k-1, s') e_i \right) \right\| \\ &\leq \sum_{s \neq s'}^{k-1} \mathbb{E} \|\eta_s\| \left\| \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right\| \left\| \eta_{s'} \right\| \left\| \left(\frac{1_n}{n} - \Phi(k-1, s') e_i \right) \right\| \\ &\leq \sum_{s \neq s'}^{k-1} \mathbb{E} \frac{\|\eta_s\|^2}{2} \left\| \left(\frac{1_n}{n} - \Phi(k-1, s) e_i \right) \right\|^2 \\ &\quad + \sum_{s \neq s'}^{k-1} \mathbb{E} \frac{\|\eta_{s'}\|^2}{2} \left\| \left(\frac{1_n}{n} - \Phi(k-1, s') e_i \right) \right\|^2 \\ &\stackrel{(26)}{\leq} \sum_{s \neq s'}^{k-1} \mathbb{E} \left(\frac{\|\eta_s\|^2}{2} + \frac{\|\eta_{s'}\|^2}{2} \right) \frac{n-1}{n} \rho^{k-\frac{s+s'}{2}}\end{aligned}$$

$$\begin{aligned}
&= \frac{n-1}{n} \sum_{s \neq s'} \mathbb{E} \|\eta_s\|^2 \rho^{k-((s+s')/2)} \\
&\stackrel{(33)}{\leq} 2 \frac{n-1}{n} \cdot \frac{8nG^2 d\gamma^2}{\varepsilon^2} \sum_{s=0}^{k-1} \sum_s'_{s \neq s'} k-1 \rho^{k-((s+s')/2)} \\
&\leq 2 \frac{n-1}{n} \cdot \frac{8nG^2 d\gamma^2}{\varepsilon^2} \sum_{s=0}^{k-1} \frac{\rho^{(k-s+1)/2} (1-\rho^{(k+1)/2})}{1-\sqrt{\rho}} \\
&\leq \frac{16(n-1)G^2 d\gamma^2}{\varepsilon^2} \frac{(1-\rho^{(k-1)/2})(\rho-\rho^{(k+1)/2})}{(1-\sqrt{\rho})^2} \\
&\leq \frac{16(n-1)G^2 d\gamma^2}{\varepsilon^2} \frac{1}{(1-\sqrt{\rho})^2}.
\end{aligned} \tag{41}$$

Then, plugging A_3 and A_4 into A_1 ,

$$\begin{aligned}
A_1 &= \frac{8(n-1)G^2 d\gamma^2}{\varepsilon^2} \frac{1}{1-\rho} + \frac{16(n-1)G^2 d\gamma^2}{\varepsilon^2} \frac{1}{(1-\sqrt{\rho})^2} \\
&= \frac{8(n-1)G^2 d\gamma^2}{\varepsilon^2} \left(\frac{1}{1-\rho} + \frac{2}{(1-\sqrt{\rho})^2} \right) \leq \frac{24(n-1)G^2 d\gamma^2}{\varepsilon^2 (1-\sqrt{\rho})^2}.
\end{aligned} \tag{42}$$

where the last inequality comes from the fact that $(1/(1-\rho)) \leq (1/((1-\sqrt{\rho})^2))$.

Moreover, we split A_2 into two terms:

$$\begin{aligned}
A_2 &= \mathbb{E} \left\| \sum_{s=0}^{k-1} \partial F(X_s; \xi_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\|^2 \\
&\leq 2 \mathbb{E} \left\| \sum_{s=0}^{k-1} (\partial F(X_s; \xi_s) - \partial f(X_s)) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\|^2 \\
&\quad \underbrace{\hspace{10em}}_{=A_5} \\
&\quad + 2 \mathbb{E} \left\| \sum_{s=0}^{k-1} \partial f(X_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\|^2. \\
&\quad \underbrace{\hspace{10em}}_{=A_6}
\end{aligned} \tag{43}$$

We give an upper bound A_5 as follows:

$$\begin{aligned}
A_5 &= \mathbb{E} \left\| \sum_{s=0}^{k-1} (\partial F(X_s; \xi_s) - \partial f(X_s)) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\|^2 \\
&= \sum_{s=0}^{k-1} \mathbb{E} \left\| (\partial F(X_s; \xi_s) - \partial f(X_s)) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\|^2 \\
&\leq \sum_{s=0}^{k-1} \mathbb{E} \|\partial F(X_s; \xi_s) - \partial f(X_s)\|^2 \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\|^2 \\
&\leq \sum_{s=0}^{k-1} \mathbb{E} \|\partial F(X_s; \xi_s) - \partial f(X_s)\|_F^2 \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\|^2 \\
&\leq n\sigma^2 \sum_{s=0}^{k-1} \frac{n-1}{n} \rho^{k-s-1} = \frac{(n-1)\sigma^2}{1-\rho},
\end{aligned} \tag{44}$$

where the last second inequality comes from Lemma 8 and assumption 3.

For A_6 , we will give the following upper bound:

$$\begin{aligned}
A_6 &= \mathbb{E} \left\| \sum_{s=0}^{k-1} \partial f(X_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\|^2 \\
&= \underbrace{\sum_{s=0}^{k-1} \mathbb{E} \left\| \partial f(X_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\|^2}_{=A_7} \\
&\quad + \underbrace{\sum_{s \neq s'}^{k-1} \mathbb{E} \left\langle \partial f(X_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right), \partial f(X_{s'}) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\rangle}_{=A_8}.
\end{aligned} \tag{45}$$

To bound A_6 , we first bound A_7 and A_8 , for A_7 :

$$\begin{aligned}
A_7 &= \sum_{s=0}^{k-1} \mathbb{E} \left\| \partial f(X_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\|^2 \\
&\leq \sum_{s=0}^{k-1} \mathbb{E} \|\partial f(X_s)\|^2 \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\|^2 \\
&\leq 3 \sum_{s=0}^{k-1} \sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\|^2 \\
&\quad + 3n\zeta^2 \sum_{s=0}^{k-1} \frac{n-1}{n} \rho^{k-s-1} + 3 \sum_{s=0}^{k-1} \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \\
&\quad \cdot \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\|^2 \\
&\leq \frac{3(n-1)}{n} \sum_{s=0}^{k-1} \left(\sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} + \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \right) \rho^{k-s-1} \\
&\quad + \frac{3(n-1)\zeta^2}{1-\rho},
\end{aligned} \tag{46}$$

where the last inequality comes from Lemmas 8 and 11.

Then, we bound A_8 as follows:

$$\begin{aligned}
A_8 &= \sum_{s \neq s'}^{k-1} \mathbb{E} \left\langle \partial f(X_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right), \partial f(X_{s'}) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\rangle \\
&\leq \sum_{s \neq s'}^{k-1} \mathbb{E} \left\| \partial f(X_s) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\| \left\| \partial f(X_{s'}) \left(\frac{1_n}{n} - \Phi(k-1, s+1)e_i \right) \right\| \\
&\leq \sum_{s \neq s'}^{k-1} \mathbb{E} \|\partial f(X_s)\| \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\| \left\| \partial f(X_{s'}) \right\| \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\| \\
&\leq \sum_{s \neq s'}^{k-1} \mathbb{E} \frac{\|\partial f(X_s)\|^2}{2} \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\| \\
&\quad + \sum_{s \neq s'}^{k-1} \mathbb{E} \frac{\|\partial f(X_{s'})\|^2}{2} \left\| \frac{1_n}{n} - \Phi(k-1, s+1)e_i \right\| \\
&\stackrel{(26)}{\leq} \sum_{s \neq s'}^{k-1} \mathbb{E} \left(\frac{\|\partial f(X_s)\|^2}{2} + \frac{\|\partial f(X_{s'})\|^2}{2} \right) \frac{n-1}{n} \rho^{k-((s+s')/2)-1} \\
&= \frac{n-1}{n} \sum_{s \neq s'}^{k-1} \mathbb{E} (\|\partial f(X_s)\|^2) \rho^{k-((s+s')/2)-1} \\
&\stackrel{(36)}{\leq} \underbrace{\frac{3(n-1)}{n} \sum_{s \neq s'}^{k-1} \left(\sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} + \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \right) \rho^{k-((s+s')/2)-1}}_{=A_9} \\
&\quad + \underbrace{\sum_{s \neq s'}^{k-1} 3(n-1)\zeta^2 \rho^{k-((s+s')/2)-1}}_{=A_{10}},
\end{aligned} \tag{47}$$

where A_{10} can be bounded by ζ and ρ :

$$\begin{aligned} A_{10} &= 6(n-1)\zeta^2 \sum_{s>s'}^{k-1} \rho^{k-((s+s')/2)-1} = 6(n-1)\zeta^2 \sum_{s=0}^{k-1} \sum_{s'=s+1}^{k-1} \rho^{k-((s+s')/2)-1} \\ &= 6(n-1)\zeta^2 \frac{(\rho^{k/2}-1)(\rho^{k/2}-\sqrt{\rho})}{(\sqrt{\rho}-1)^2(\sqrt{\rho}+1)} \leq 6(n-1)\zeta^2 \frac{1}{(1-\sqrt{\rho})^2}, \end{aligned} \quad (48)$$

and we give a bound of A_9 :

$$\begin{aligned} A_9 &= \frac{3(n-1)}{n} \sum_{s \neq s'}^{k-1} \left(\sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} + \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \right) \rho^{k-((s+s')/2)-1} \\ &= \frac{6(n-1)}{n} \sum_{s=0}^{k-1} \left(\sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} + \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \right) \sum_{s'=s+1}^{k-1} \rho^{2k-(s+s')-2} \\ &\leq \frac{6(n-1)}{n} \sum_{s=0}^{k-1} \left(\sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} + \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \right) \frac{\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}}. \end{aligned} \quad (49)$$

Then, we plug A_9 , A_{10} into A_8 , plug A_8 , A_7 into A_6 . We can yield the upper bound for A_6 .

$$\begin{aligned} A_6 &\leq \frac{6(n-1)}{n} \sum_{s=0}^{k-1} \left(\sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} + \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \right) \frac{\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \frac{6(n-1)\zeta^2}{(1-\sqrt{\rho})^2} \\ &\quad + \frac{3(n-1)}{n} \sum_{s=0}^{k-1} \left(\sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} + \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \right) \rho^{k-s-1} + \frac{3(n-1)\zeta^2}{1-\rho} \\ &\leq \frac{3(n-1)}{n} \sum_{s=0}^{k-1} \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) \\ &\quad + \frac{3(n-1)}{n} \sum_{s=0}^{k-1} \sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) + \frac{9(n-1)\zeta^2}{(1-\sqrt{\rho})^2}. \end{aligned} \quad (50)$$

Then, we plug A_5 and A_6 into A_2 yielding the upper bound of A_2 .

$$\begin{aligned} A_2 &\leq \frac{2(n-1)\sigma^2}{1-\rho} + \frac{6(n-1)}{n} \sum_{s=0}^{k-1} \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) \\ &\quad + \frac{6(n-1)}{n} \sum_{s=0}^{k-1} \sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) + \frac{18(n-1)\zeta^2}{(1-\sqrt{\rho})^2}. \end{aligned} \quad (51)$$

Finally, we can describe $\text{Dis}_{k,i}$ as follows:

$$\begin{aligned} \text{Dis}_{k,i} &= 2A_1 + 2\gamma^2 A_2 \leq \frac{48(n-1)G^2 d\gamma^2}{\varepsilon^2(1-\sqrt{\rho})^2} + 2\gamma^2 \left(\frac{2(n-1)\sigma^2}{1-\rho} \right. \\ &\quad + \frac{18(n-1)\zeta^2}{(1-\sqrt{\rho})^2} + \frac{6(n-1)}{n} \sum_{s=0}^{k-1} \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \\ &\quad \cdot \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) + \frac{6(n-1)}{n} \sum_{s=0}^{k-1} \sum_{h=1}^n \mathbb{E} L^2 \text{Dis}_{s,h} \\ &\quad \cdot \left. \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) \right). \end{aligned} \quad (52)$$

The lemma is obtained.

Based on these lemmas above, we prove Theorem 5 subsequently.

Proof (Proof of Theorem 5). We start from $f((X_{k+1} 1_n)/n)$:

$$\begin{aligned} \mathbb{E} f \left(\frac{X_{k+1} 1_n}{n} \right) &= \mathbb{E} f \left(\frac{(X_k + \eta_k) W_k 1_n}{n} - \gamma \frac{\partial F(X_k; \xi_k) 1_n}{n} \right) \\ &= \mathbb{E} f \left(\frac{X_k W_k 1_n}{n} + \frac{\eta_k W_k 1_n}{n} - \gamma \frac{\partial F(X_k; \xi_k) 1_n}{n} \right) \\ &\stackrel{\text{Lemma 7}}{=} \mathbb{E} f \left(\frac{X_k 1_n}{n} + \frac{\eta_k 1_n}{n} - \gamma \frac{\partial F(X_k; \xi_k) 1_n}{n} \right) \\ &\leq \mathbb{E} f \left(\frac{X_k 1_n}{n} \right) + \mathbb{E} \left\langle \nabla f \left(\frac{X_k 1_n}{n} \right), \frac{\eta_k 1_n}{n} - \gamma \frac{\partial f(X_k) 1_n}{n} \right\rangle \\ &\quad + \frac{L}{2} \mathbb{E} \left\| \frac{\eta_k 1_n}{n} - \gamma \frac{\partial F(X_k; \xi_k) 1_n}{n} \right\|^2 \\ &\leq \mathbb{E} f \left(\frac{X_k 1_n}{n} \right) + \mathbb{E} \left\langle \nabla f \left(\frac{X_k 1_n}{n} \right), \frac{\eta_k 1_n}{n} \right\rangle \\ &\quad - \mathbb{E} \left\langle \nabla f \left(\frac{X_k 1_n}{n} \right), \gamma \frac{\partial f(X_k) 1_n}{n} \right\rangle + L \mathbb{E} \left\| \frac{\sum_{i=1}^n \eta_{k,i}}{n} \right\|^2 \\ &\quad + \gamma^2 L \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla F_i(x_{k,i}; \xi_{k,i})}{n} \right\|^2, \end{aligned} \quad (53)$$

where the last step comes from $\|a - b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.

Then, we split the second term according to $2\langle a, b \rangle \leq \|a\|^2 + \|b\|^2$,

$$\begin{aligned} \mathbb{E} \left\langle \nabla f \left(\frac{X_k 1_n}{n} \right), \frac{\eta_k 1_n}{n} \right\rangle &\leq \frac{1}{2} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) \right\|^2 + \frac{1}{2} \mathbb{E} \left\| \frac{\eta_k 1_n}{n} \right\|^2 \\ &= \frac{1}{2} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) \right\|^2 + \frac{1}{2} \mathbb{E} \left\| \frac{\sum_{i=1}^n \eta_{k,i}}{n} \right\|^2. \end{aligned} \quad (54)$$

We split the last term of (53) into two terms because of

$$\begin{aligned} \mathbb{E}_{\xi_{k,i}} \nabla F_i(x_{k,i}; \xi_{k,i}) &= \nabla f_i(x_{k,i}), \\ \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla F_i(x_{k,i}; \xi_{k,i})}{n} \right\|^2 &= \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla F_i(x_{k,i}; \xi_{k,i}) - \sum_{i=1}^n \nabla f_i(x_{k,i})}{n} \right\|^2 \\ &\quad + \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla f_i(x_{k,i})}{n} \right\|^2. \end{aligned} \quad (55)$$

According to (54) and (55), (53) can be expressed as:

$$\begin{aligned} \mathbb{E} f \left(\frac{X_{k+1} 1_n}{n} \right) &\leq \mathbb{E} f \left(\frac{X_k 1_n}{n} \right) - \gamma \mathbb{E} \left\langle \nabla f \left(\frac{X_k 1_n}{n} \right), \frac{\partial f(X_k) 1_n}{n} \right\rangle \\ &\quad + \frac{1}{2} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) \right\|^2 + \frac{1+2L}{2} \mathbb{E} \left\| \frac{\sum_{i=1}^n \eta_{k,i}}{n} \right\|^2 \\ &\quad + \gamma^2 L \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla F_i(x_{k,i}; \xi_{k,i}) - \sum_{i=1}^n \nabla f_i(x_{k,i})}{n} \right\|^2 \\ &\quad + \gamma^2 L \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla f_i(x_{k,i})}{n} \right\|^2. \end{aligned} \quad (56)$$

FIGURE 1: The convergence rate with different privacy budgets.

We can bound the second last term using σ :

$$\begin{aligned}
 & \gamma^2 L \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla F_i(x_{k,i}; \xi_{k,i}) - \sum_{i=1}^n \nabla f_i(x_{k,i})}{n} \right\|^2 \\
 &= \frac{\gamma^2 L}{n^2} \sum_{i=1}^n \mathbb{E} \|\nabla F_i(x_{k,i}; \xi_{k,i}) - \nabla f_i(x_{k,i})\|^2 + \frac{2\gamma^2 L}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \mathbb{E} \langle \nabla F_i(x_{k,i}; \xi_{k,i}) \\
 &\quad - \nabla f_i(x_{k,i}), \nabla F_{i'}(x_{k,i'}; \xi_{k,i'}) - \nabla f_{i'}(x_{k,i'}) \rangle = \frac{\gamma^2 L}{n^2} \sum_{i=1}^n \mathbb{E} \|\nabla F_i(x_{k,i}; \xi_{k,i}) \\
 &\quad - \nabla f_i(x_{k,i})\|^2 + \frac{2\gamma^2 L}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \mathbb{E} \langle \nabla F_i(x_{k,i}; \xi_{k,i}) \\
 &\quad - \nabla f_i(x_{k,i}), \mathbb{E}_{\xi_{k,i'}} \nabla F_{i'}(x_{k,i'}; \xi_{k,i'}) - \nabla f_{i'}(x_{k,i'}) \rangle \\
 &= \frac{\gamma^2 L}{n^2} \sum_{i=1}^n \mathbb{E} \|\nabla F_i(x_{k,i}; \xi_{k,i}) - \nabla f_i(x_{k,i})\|^2 \leq \frac{\gamma^2 L}{n} \sigma^2,
 \end{aligned} \tag{57}$$

where the last step is true because of assumption 3.

Then, it follows from (56):

$$\begin{aligned}
 \mathbb{E} f\left(\frac{X_{k+1} \mathbf{1}_n}{n}\right) &\leq \mathbb{E} f\left(\frac{X_k \mathbf{1}_n}{n}\right) - \gamma \mathbb{E} \left\langle \nabla f\left(\frac{X_k \mathbf{1}_n}{n}\right), \frac{\partial f(X_k) \mathbf{1}_n}{n} \right\rangle \\
 &\quad + \frac{1}{2} \mathbb{E} \left\| \nabla f\left(\frac{X_k \mathbf{1}_n}{n}\right) \right\|^2 + \frac{1+2L}{2} \mathbb{E} \left\| \frac{\sum_{i=1}^n \eta_{k,i}}{n} \right\|^2 \\
 &\quad + \frac{\gamma^2 L}{n} \sigma^2 + \gamma^2 L \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla f_i(x_{k,i})}{n} \right\|^2 \\
 &\stackrel{(33)}{\leq} \mathbb{E} f\left(\frac{X_k \mathbf{1}_n}{n}\right) - \frac{\gamma - 2\gamma^2 L}{2} \mathbb{E} \left\| \frac{\partial f(X_k) \mathbf{1}_n}{n} \right\|^2 \\
 &\quad - \frac{\gamma - 1}{2} \mathbb{E} \left\| \nabla f\left(\frac{X_k \mathbf{1}_n}{n}\right) \right\|^2 + \frac{4(1+2L)G^2 d \gamma^2}{\varepsilon^2} \\
 &\quad + \underbrace{\frac{\gamma^2 L}{n} \sigma^2 + \frac{\gamma}{2} \mathbb{E} \left\| \nabla f\left(\frac{X_k \mathbf{1}_n}{n}\right) - \frac{\partial f(X_k) \mathbf{1}_n}{n} \right\|^2}_{=X},
 \end{aligned} \tag{58}$$

where the last step comes from $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$.

We then bound the equation X:

$$\begin{aligned}
 X &= \mathbb{E} \left\| \nabla f\left(\frac{X_k \mathbf{1}_n}{n}\right) - \frac{\partial f(X_k) \mathbf{1}_n}{n} \right\|^2 \\
 &= \mathbb{E} \left\| \frac{\sum_{i=1}^n \nabla f_i\left(\left(\frac{\sum_{i'=1}^n x_{k,i'}}{n}\right)\right)}{n} - \frac{\sum_{i=1}^n \nabla f_i(x_{k,i})}{n} \right\|^2 \\
 &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left\| \nabla f_i\left(\frac{\sum_{i'=1}^n x_{k,i'}}{n}\right) - \nabla f_i(x_{k,i}) \right\|^2 \\
 &\stackrel{\text{Assumption 1}}{\leq} \underbrace{\frac{L^2}{n} \sum_{i=1}^n \mathbb{E} \left\| \frac{\sum_{i'=1}^n x_{k,i'}}{n} - x_{k,i} \right\|^2}_{=Dis_{k,i}},
 \end{aligned} \tag{59}$$

where the first inequality comes from $\|\sum_{i=1}^n a_i\|^2 \leq n \sum_{i=1}^n \|a_i\|^2$.

According to Equation (37) in Lemma 12, we have the bound of $Dis_{k,i}$. Then, we will bound its average M_k on all nodes as follows:

$$\begin{aligned}
 \mathbb{E} M_k &= \frac{\mathbb{E} \sum_{i=1}^n Dis_{k,i}}{n} = \frac{48(n-1)G^2 d \gamma^2}{\varepsilon^2 (1-\sqrt{\rho})^2} + \frac{4\gamma^2(n-1)\sigma^2}{1-\rho} \\
 &\quad + \frac{36\gamma^2(n-1)\zeta^2}{(1-\sqrt{\rho})^2} + 12(n-1)\gamma^2 L^2 \sum_{s=0}^{k-1} \mathbb{E} M_k \\
 &\quad \cdot \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) + \frac{12(n-1)}{n} \gamma^2 \sum_{s=0}^{k-1} \mathbb{E} \left\| \nabla f\left(\frac{X_s \mathbf{1}_n}{n}\right) \mathbf{1}_n^T \right\|^2 \\
 &\quad \cdot \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right).
 \end{aligned} \tag{60}$$

FIGURE 2: Comparison between dynamic networks and static networks.

Summing it from $k = 0$ to $K - 1$, we can get the following result:

$$\begin{aligned}
 \sum_{k=0}^{K-1} \mathbb{E} M_k &\leq \frac{48(n-1)G^2 d \gamma^2 K}{\varepsilon^2 (1-\sqrt{\rho})^2} + \frac{4\gamma^2(n-1)\sigma^2 K}{1-\rho} + \frac{36\gamma^2(n-1)\zeta^2 K}{(1-\sqrt{\rho})^2} \\
 &\quad + 12(n-1)\gamma^2 L^2 \sum_{k=0}^{K-1} \sum_{s=0}^{k-1} \mathbb{E} M_k \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) \\
 &\quad + \frac{12(n-1)}{n} \gamma^2 \sum_{k=0}^{K-1} \sum_{s=0}^{k-1} \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \left(\frac{2\sqrt{\rho}^{k-s-1}}{1-\sqrt{\rho}} + \rho^{k-s-1} \right) \\
 &\leq \frac{48(n-1)G^2 d \gamma^2 K}{\varepsilon^2 (1-\sqrt{\rho})^2} + \frac{4\gamma^2(n-1)\sigma^2 K}{1-\rho} + \frac{36\gamma^2(n-1)\zeta^2 K}{(1-\sqrt{\rho})^2} \\
 &\quad + 12(n-1)\gamma^2 L^2 \sum_{k=0}^{K-1} \mathbb{E} M_k \left(\frac{1}{1-\rho} + \frac{2}{(1-\sqrt{\rho})^2} \right) \\
 &\quad + \frac{12(n-1)}{n} \gamma^2 \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f \left(\frac{X_s 1_n}{n} \right) 1_n^T \right\|^2 \left(\frac{1}{1-\rho} + \frac{2}{(1-\sqrt{\rho})^2} \right) \\
 &\leq \frac{48(n-1)G^2 d \gamma^2 K}{\varepsilon^2 (1-\sqrt{\rho})^2} + \frac{4\gamma^2(n-1)\sigma^2 K}{1-\rho} + \frac{36\gamma^2(n-1)\zeta^2 K}{(1-\sqrt{\rho})^2} \\
 &\quad + \frac{36(n-1)\gamma^2 L^2}{(1-\sqrt{\rho})^2} \sum_{k=0}^{K-1} \mathbb{E} M_k + \frac{36\gamma^2(n-1)}{(1-\sqrt{\rho})^2 n} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) 1_n^T \right\|^2.
 \end{aligned} \tag{61}$$

We then can get the bound of the summation of $\mathbb{E} M_k$'s from $k = 0$ to $K - 1$:

$$\begin{aligned}
 \left(1 - \frac{36\gamma^2(n-1)L^2}{(1-\sqrt{\rho})^2} \right) \sum_{k=0}^{K-1} \mathbb{E} M_k &\leq \frac{48(n-1)G^2 d \gamma^2 K}{\varepsilon^2 (1-\sqrt{\rho})^2} + \frac{4\gamma^2(n-1)\sigma^2 K}{1-\rho} \\
 &\quad + \frac{36\gamma^2(n-1)\zeta^2 K}{(1-\sqrt{\rho})^2} \\
 &\quad + \frac{36\gamma^2(n-1)}{(1-\sqrt{\rho})^2 n} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) 1_n^T \right\|^2.
 \end{aligned} \tag{62}$$

Rearranging the term, it can be obtained that

$$\begin{aligned}
 \sum_{k=0}^{K-1} \mathbb{E} M_k &\leq \frac{48(n-1)G^2 d \gamma^2 K}{\varepsilon^2 (1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \\
 &\quad + \frac{4\gamma^2(n-1)\sigma^2 K}{(1-\rho) \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \\
 &\quad + \frac{36\gamma^2(n-1)\zeta^2 K}{(1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \\
 &\quad + \frac{36\gamma^2(n-1)}{(1-\sqrt{\rho})^2 n \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \\
 &\quad \cdot \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) 1_n^T \right\|^2.
 \end{aligned} \tag{63}$$

Plugging the bound of M_k into X :

$$X \leq \frac{L^2}{n} \sum_{i=1}^n \mathbb{E} Dis_{k,i} = L^2 \mathbb{E} M_k. \tag{64}$$

Then, we bound the $\mathbb{E} f((X_{k+1} 1_n)/n)$ by the above bound,

$$\begin{aligned}
 \mathbb{E} f \left(\frac{X_{k+1} 1_n}{n} \right) &\leq \mathbb{E} f \left(\frac{X_k 1_n}{n} \right) - \frac{\gamma - 2\gamma^2 L}{2} \mathbb{E} \left\| \frac{\partial f(X_k) 1_n}{n} \right\|^2 - \frac{\gamma - 1}{2} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) \right\|^2 \\
 &\quad + \frac{4(1+2L)G^2 d \gamma^2}{\varepsilon^2} + \frac{\gamma^2 L \sigma^2}{n} + \frac{\gamma}{2} X \\
 &\leq \mathbb{E} f \left(\frac{X_k 1_n}{n} \right) - \frac{\gamma - 2\gamma^2 L}{2} \mathbb{E} \left\| \frac{\partial f(X_k) 1_n}{n} \right\|^2 - \frac{\gamma - 1}{2} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) \right\|^2 \\
 &\quad + \frac{4(1+2L)G^2 d \gamma^2}{\varepsilon^2} + \frac{\gamma^2 L \sigma^2}{n} + \frac{\gamma}{2} L^2 \mathbb{E} M_k.
 \end{aligned} \tag{65}$$

Summing from $k = 0$ to $k = K - 1$:

$$\begin{aligned}
& \frac{\gamma - 2\gamma^2 L}{2} \sum_{k=0}^{K-1} \mathbb{E} \left\| \frac{\partial f(X_k) 1_n}{n} \right\|^2 + \frac{\gamma - 1}{2} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) \right\|^2 \\
& \leq f(0) - f^* + \frac{4(1+2L)G^2 d \gamma^2 K}{\varepsilon^2} + \frac{\gamma^2 K L \sigma^2}{n} + \frac{\gamma}{2} L^2 \sum_{k=0}^{K-1} \mathbb{E} M_k \\
& \stackrel{(61)}{\leq} f(0) - f^* + \frac{4(1+2L)G^2 d \gamma^2 K}{\varepsilon^2} + \frac{\gamma^2 K L \sigma^2}{n} \\
& \quad + \frac{\gamma}{2} L^2 \left(\frac{48(n-1)G^2 d \gamma^2 K}{\varepsilon^2 (1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \right) \\
& \quad + \frac{\gamma}{2} L^2 \left(\frac{4\gamma^2(n-1)\sigma^2 K}{(1-\rho) \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \right) \\
& \quad + \frac{\gamma}{2} L^2 \left(\frac{36\gamma^2(n-1)\zeta^2 K}{(1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \right)
\end{aligned}$$

$$\begin{aligned}
& + \frac{\gamma}{2} L^2 \left(\frac{36\gamma^2(n-1)}{(1-\sqrt{\rho})^2 n \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \right) \\
& \cdot \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) 1_n^T \right\|^2 = f(0) - f^* + \frac{4(1+2L)G^2 d \gamma^2 K}{\varepsilon^2} + \frac{\gamma^2 K L \sigma^2}{n} \\
& + \frac{24(n-1)L^2 G^2 \gamma^2 K}{\varepsilon^2 (1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \\
& + \frac{2\gamma^3 L^2 (n-1)\sigma^2}{(1-\rho) \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)}^K \\
& + \frac{18\gamma^3 L^2 (n-1)\zeta^2}{(1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)}^K \\
& + \frac{18\gamma^3 L^2 (n-1)}{(1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f \left(\frac{X_k 1_n}{n} \right) \right\|^2.
\end{aligned} \tag{66}$$

Rearranging the terms, we get the following result:

$$\begin{aligned}
& (1 - \gamma L/2) \sum_{k=0}^{K-1} \mathbb{E} \left\| \partial f(X_k) 1_n / n \right\|^2 + \left((1/2) - \left((18\gamma^2(n-1)L^2) / (1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right) \right) \right) \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f((X_k 1_n)/n) \right\|^2 \\
& \leq \frac{f(0) - f^*}{\gamma K} + \frac{4(1+2L)G^2 d \gamma}{\varepsilon^2} + \frac{\gamma L}{n} \sigma^2 + \frac{24(n-1)L^2 G^2 \gamma^2}{\varepsilon^2 (1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} + \frac{2\gamma^2 L^2 (n-1)\sigma^2}{(1-\rho) \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)} \\
& \quad + \frac{18\gamma^2(n-1)L^2 \zeta^2}{(1-\sqrt{\rho})^2 \left(1 - \left((36\gamma^2(n-1)L^2) / \left((1-\sqrt{\rho})^2 \right) \right) \right)}.
\end{aligned} \tag{67}$$

This completes the proof.

Proof (Proof of Corollary 6). Substituting $\gamma = 1/(2L + \sigma\sqrt{K/n})$ into the result of Theorem 5 and removing the first term of the RHS on the LHS, we can obtain that

$$\begin{aligned}
& \frac{B_1 \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f((X_k 1_n)/n) \right\|^2}{K} \leq \frac{(2L + \sigma\sqrt{K/n})(f(0) - f^*)}{K} + \frac{4(1+2L)G^2 d}{(2L + \sigma\sqrt{K/n})\varepsilon^2} \\
& \quad + \frac{L\sigma^2}{2nL + \sigma\sqrt{Kn}} + \left(\frac{2(n-1)L^2}{(2L + \sigma\sqrt{K/n})^2 B_2} \right) \\
& \quad \cdot \left(\frac{12G^2}{\varepsilon^2 (1-\sqrt{\rho})^2} + \frac{\sigma^2}{(1-\rho)} + \frac{9\zeta^2}{(1-\sqrt{\rho})^2} \right) \\
& \leq \frac{2L(f(0) - f^*)}{K} + \frac{(f(0) - f^*)\sigma}{\sqrt{Kn}} + \frac{4(1+2L)G^2 d}{\sigma\sqrt{K/n}\varepsilon^2} \\
& \quad + \frac{L\sigma^2}{\sigma\sqrt{Kn}} + \left(\frac{2(n-1)L^2}{(\sigma\sqrt{K/n})^2 B_2} \right) \\
& \quad \cdot \left(\frac{12G^2}{\varepsilon^2 (1-\sqrt{\rho})^2} + \frac{\sigma^2}{(1-\rho)} + \frac{9\zeta^2}{(1-\sqrt{\rho})^2} \right).
\end{aligned} \tag{68}$$

Let $(G/\varepsilon) \leq \sqrt{((f(0) - f^*) + L)\sigma / (4(1+2L)d n)}$, and we show B_1 and B_2 are approximately constants when (16) is

satisfied.

$$\begin{aligned}
B_1 &= \left(\frac{1}{2} - \frac{18\gamma^2(n-1)L^2}{(1-\sqrt{\rho})^2 B_2} \right), \\
B_2 &= \left(1 - \frac{36\gamma^2(n-1)L^2}{(1-\sqrt{\rho})^2} \right).
\end{aligned} \tag{69}$$

Note that

$$\begin{aligned}
\gamma^2 &\leq \frac{(1-\sqrt{\rho})^2}{72(n-1)L^2} \Rightarrow B_2 \geq \frac{1}{2}, \\
\gamma^2 &\leq \frac{(1-\sqrt{\rho})^2}{144(n-1)L^2} \Rightarrow B_1 \geq \frac{1}{4}.
\end{aligned} \tag{70}$$

Since $\gamma^2 \leq (n/(\sigma^2 K))$, as long as we have

$$\frac{n}{\sigma^2 K} \leq \frac{(1-\sqrt{\rho})^2}{144(n-1)L^2}, \tag{71}$$

$B_2 \geq 1/2$ and $B_1 \geq 1/4$ will be satisfied. Then, we can obtain $K \geq (144n(n-1)L^2)/(\sigma^2(1-\sqrt{\rho})^2)$.

Let $K \geq ((16n^3(n-1)^2L^4)/(\sigma^6(f(0) - f^* + L)^2))$
 $((12G^2/(\varepsilon^2(1 - \sqrt{\rho})^2)) + (\sigma^2/(1 - \rho)) + (9\zeta^2/((1 - \sqrt{\rho})^2)))^2$,
 then

$$\frac{\sum_{k=0}^{K-1} \mathbb{E} \|\nabla f((X_k 1_n)/n)\|^2}{K} \leq \frac{8L(f(0) - f^*)}{K} + \frac{12(f(0) - f^* + L)\sigma}{\sqrt{Kn}}. \quad (72)$$

7. Experiments

In this section, we perform extensive simulations to evaluate our proposed algorithm. In particular, we compare the convergence rate of our proposed algorithm with the best-known D-PSGD algorithm give in [1], in the settings of different privacy budgets, different number of nodes and different extents of dynamicity.

In our experiments, we evaluate our proposed algorithm on image classification tasks. We train ResNet model on CIFAR-10 dataset. Moreover, we use MPI to implement the communication scheme. We run our experiments on a CPU server cluster. Each server has 32 cores, which is an Intel Xeon E5-2620 v4 @ 2.10GHz cluster. Each server is equipped with 128G memory.

7.1. Impact of Privacy Budgets. We evaluate the import of privacy budget on the convergence of our proposed algorithm in a dynamic network. Here, we run our algorithm in a dynamic network with different number of nodes under different privacy budgets. The results are illustrated in Figure 1, where the privacy budget ε is set to 10^{-1} , 10^{-2} , and 10^{-3} , respectively. It can be seen that the smaller ε is, the slower the learning converges. This is because a smaller privacy budget means more noise are added, which affects the convergence speed of the algorithm.

7.2. Impact of Dynamicity. We compare our algorithm with the best-known D-PSGD algorithm (with privacy protection) in a static network. Due to D-PSGD is applied to a ring structure, we set the expected degree in the dynamic network as 2. Here, the privacy budget is set to 10^{-1} , and the number of nodes is 4 and 8. From Figure 2, we can find that our proposed algorithm can reach the same convergence rate in dynamic networks as the D-PSGD algorithm (with privacy protection) in static networks.

8. Conclusion

We presented D-(DP)²SGD, a decentralized parallel stochastic gradient descent algorithm with privacy preservation in dynamic networks. With theoretical analysis and extensive experiments, it shows that our proposed algorithm can achieve the same convergence rate as the best know previous work in static networks without considering privacy issue. Based on this work, it is meaningful to further devise privacy-preserving algorithms in an asynchronous dynamic environment.

Data Availability

1. The CIFAR-10 data used to support the findings of this study have been deposited in <http://www.cs.toronto.edu/~kriz/cifar.html>. 2. The software code used to support the findings of this study have been deposited in <https://github.com/zongruisdu/D-DPDP-SGD>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partially supported by the National Key R&D Program of China with grant no. 2019YFB2102600 and NSFC (No. 61971269).

References

- [1] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 5330–5340, Long Beach, CA, USA, December 2017.
- [2] J. Xu, W. Zhang, and F. Wang, "A(dp)2sgd: asynchronous decentralized parallel stochastic gradient descent with differential privacy," 2020, <https://arxiv.org/abs/2008.09246>.
- [3] M. Zhu and Q. Chen, "Big data image classification based on distributed deep representation learning model," *IEEE Access*, vol. 8, pp. 133890–133904, 2020.
- [4] W. Zhang, S. Gupta, X. Lian, and J. Liu, "Staleness-aware async-sgd for distributed deep learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*, pp. 2350–2356, New York, NY, USA, July 2016.
- [5] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [6] W. Xu, H. Peng, X. Zeng, F. Zhou, X. Tian, and X. Peng, "A hybrid modelling method for time series forecasting based on a linear regression model and deep learning," *Applied Intelligence*, vol. 49, no. 8, pp. 3002–3015, 2019.
- [7] S. Yang, B. Ren, X. Zhou, and L. Liu, "Parallel distributed logistic regression for vertical federated learning without third-party coordinator," 2019, <https://arxiv.org/abs/1911.09824>.
- [8] S. H. Alsamhi, O. Ma, and M. S. Ansari, "Convergence of machine learning and robotics communication in collaborative assembly: mobility, connectivity and future perspectives," *Journal of Intelligent & Robotic Systems*, vol. 98, no. 3–4, pp. 541–566, 2020.
- [9] A. Elgabri, J. Park, A. S. Bedi, M. Bennis, and V. Aggarwal, "Communication efficient framework for decentralized machine learning," in *54th Annual Conference on Information Sciences and Systems, CISS 2020*, pp. 1–5, Princeton, NJ, USA, March 2020.
- [10] S. Y. Hashemi and F. S. Aliee, "Fuzzy, dynamic and trust based routing protocol for iot," *Journal of Network and Systems Management*, vol. 28, no. 4, pp. 1248–1278, 2020.

- [11] Z. Cai, Z. Xu, and Y. Jiguo, "A differential-private framework for urban traffic flows estimation via taxi companies," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6492–6499, 2019.
- [12] K. Lim and K. M. Tuladhar, "LIDAR: lidar information based dynamic V2V authentication for roadside infrastructure-less vehicular networks," in *16th IEEE Annual Consumer Communications & Networking Conference, CCNC 2019*, pp. 1–6, Las Vegas, NV, USA, January 2019.
- [13] S. Chen, Y. Tao, D. Yu, F. Li, B. Gong, and X. Cheng, "Privacy-preserving collaborative learning for multiarmed bandits in iot," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3276–3286, 2021.
- [14] Z. Xu, Z. Cai, J. Yu, C. Wang, and Y. Li, "Follow but no track: privacy preserved profile publishing in cyber-physical social systems," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1868–1878, 2017.
- [15] S. Chen, Y. Tao, D. Yu, F. Li, and B. Gong, "Distributed learning dynamics of multi-armed bandits for edge intelligence," *Journal of Systems Architecture*, vol. 114, p. 101919, 2020.
- [16] X. Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous decentralized parallel stochastic gradient descent," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pp. 3049–3058, Stockholmsmässan, Stockholm, Sweden, July 2018.
- [17] M. I. Qureshi, R. Xin, S. Kar, and U. A. Khan, "S-ADDOPT: decentralized stochastic first-order optimization over directed graphs," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 953–958, 2021.
- [18] Y. Yuan, F. Li, D. Yu, J. Zhao, J. Yu, and X. Cheng, "Distributed social learning with imperfect information," *IEEE Transactions on Network Science and Engineering*, 2020.
- [19] H. R. Feyzmahdavian, A. Aytekin, and M. Johansson, "An asynchronous mini-batch algorithm for regularized stochastic optimization," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3740–3754, 2016.
- [20] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011*, pp. 873–881, Granada, Spain, 2011.
- [21] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pp. 2737–2745, Montreal, Quebec, Canada, December 2015.
- [22] Z. Cai and Z. He, "Trading private range counting over big iot data," in *39th IEEE International Conference on Distributed Computing Systems, ICDCS 2019*, pp. 144–153, Dallas, TX, USA, July 2019.
- [23] Z. Cai and Z. Xu, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2020.
- [24] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Differentially private asynchronous federated learning for mobile edge computing in urban informatics," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2134–2143, 2020.
- [25] Y. Zhou and S. Tang, "Differentially private distributed learning," *INFORMS Journal on Computing*, vol. 32, no. 3, pp. 779–789, 2020.
- [26] H.-P. Cheng, P. Yu, H. Hu et al., "LEASGD: an efficient and privacy-preserving decentralized algorithm for distributed learning," 2018, <https://arxiv.org/abs/1811.11124>.
- [27] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [28] J. Zhu, C. Xu, J. Guan, and D. O. Wu, "Differentially private distributed online algorithms over time-varying directed networks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 1, pp. 4–17, 2018.